Chapter 5

Statistical Analysis

> It is easy to lie with statistics, but easier to lie without them.
>
> – Frederick Mosteller

I will confess that I have never had a truly satisfying conversation with a statistician. This is not completely for want of trying. Several times over the years I have taken problems of interest to statisticians, but always came back with answers like "You can't do it that way" or "But it's not independent," instead of hearing "Here is the way you can handle it."

To be fair, these statisticians generally did not appreciate talking with me, either. Statisticians have been thinking seriously about data for far longer than computer scientists, and have many powerful methods and ideas to show for it. In this chapter, I will introduce some of these important tools, like the definitions of certain fundamental distributions and tests for statistical significance. This chapter will also introduce Baysian analysis, a way to rigorously assess how new data should affect our previous estimates of future events.

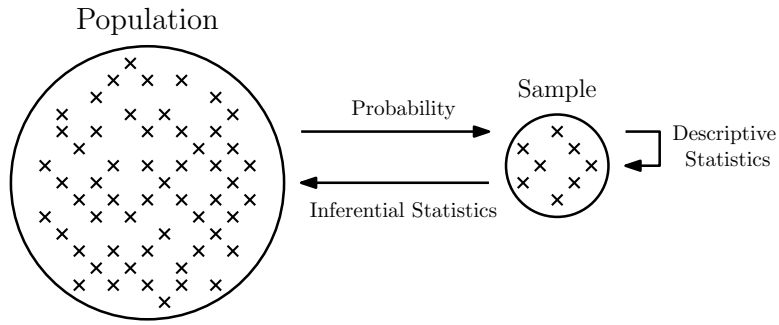

Figure 5.1: The central dogma of statistics: analysis of a small random sample enables drawing rigorous inferences about the entire population.

S.S. Skiena, *The Data Science Design Manual,*
Texts in Computer Science, https://doi.org/10.1007/978-3-319-55444-0_5

Figure 5.1 illustrates the process of statistical reasoning. There is an underlying population of possible things that we can potentially observe. Only a relatively small subset of them are actually sampled, ideally at random, meaning that we can observe properties of the sampled items. Probability theory describes what properties our sample should have, given the properties of the underlying population. But *statistical inference* works the other way, where we try to deduce what the full population is like given analysis of the sample.

Ideally, we will learn to think like a statistician: enough so as to remain vigilant and guard against overinterpretation and error, while retaining our confidence to play with data and take it where it leads us.

5.1 Statistical Distributions

Every variable that we observe defines a particular frequency distribution, which reflects how often each particular value arises. The unique properties of variables like height, weight, and IQ are captured by their distributions. But the shapes of these distributions are themselves not unique: to a great extent, the world's rich variety of data appear only in a small number of classical forms.

These classical distributions have two nice properties: (1) they describe shapes of frequency distributions that arise often in practice, and (2) they can often be described mathematically using closed-form expressions with very few parameters. Once abstracted from specific data observations, they become *probability distributions*, worthy of independent study.

Familiarity with the classical probability distributions is important. They arise often in practice, so you should be on the look out for them. They give us a vocabulary to talk about what our data looks like. We will review the most important statistical distributions (binomial, normal, Poisson, and power law) in the sections to follow, emphasizing the properties that define their essential character.

Note that your observed data does not necessarily arise from a particular theoretical distribution just because its shape is similar. Statistical tests can be used to rigorously prove whether your experimentally-observed data reflects samples drawn from a particular distribution.

But I am going to save you the trouble of actually running any of these tests. I will state with high confidence that your real-world data *does not* precisely fit any of the famous theoretical distributions.

Why is that? Understand that the world is a complicated place, which makes measuring it a messy process. Your observations will probably be drawn from multiple sample populations, each of which has a somewhat different underlying distribution. Something funny generally happens at the tails of any observed distribution: a sudden burst of unusually high or low values. Measurements will have errors associated with them, sometimes in weird systematic ways.

But that said, understanding the basic distributions is indeed very important. Each classical distribution is classical for a reason. Understanding these reasons tells you a lot about observed data, so they will be reviewed here.

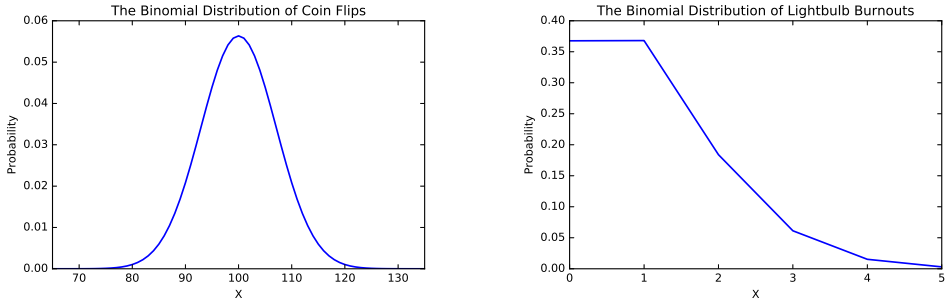

Figure 5.2: The binomial distribution can be used to model the distribution of heads in 200 coin tosses with $p = 0.5$ (left), and the number of blown lightbulbs in 1000 events with failure probability $p = 0.001$ (right).

5.1.1 The Binomial Distribution

Consider an experiment consisting of identical, independent trials which have two possible outcomes P_1 and P_2, with the respective probabilities of p and $q = (1-p)$. Perhaps your experiment is flipping fair coins, where the probability of heads ($p = 0.5$) is the same as getting tails ($q = 0.5$). Perhaps it is repeatedly turning on a light switch, where the probability of suddenly discovering that you must change the bulb ($p = 0.001$) is much less than that of seeing the light ($q = 0.999$).

The *binomial distribution* reports the probability of getting exactly x P_1 events in the course of n independent trials, in no particular order. Independence is important here: we are assuming the probability of failure of a bulb has no relation to how many times it has previously been used. The pdf for the binomial distribution is defined by:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

There are several things to observe about the binomial distribution:

- *It is discrete*: Both arguments to the binomial distribution (n and x) must be integers. The smoothness of Figure 5.2 (left) is an illusion, because $n = 200$ is fairly large. There is no way of getting 101.25 heads in 200 coin tosses.

- *You probably can explain the theory behind it*: You first encountered the binomial distribution in high school. Remember Pascal's triangle? To end up with exactly x heads in n flips in a particular sequence occurs with probability $p^x(1-p)^{(n-x)}$, for each of the $\binom{n}{x}$ distinct flip sequences.

- *It is sort of bell-shaped*: For a fair coin ($p = 0.5$), the binomial distribution is perfectly symmetrical, with the mean in the middle. This is not true

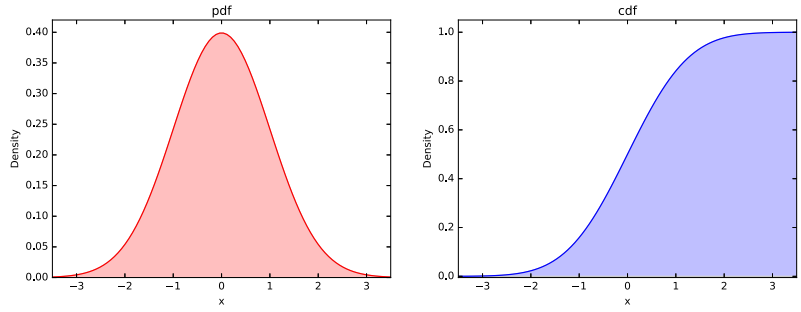

Figure 5.3: The probability density function (pdf) of the normal distribution (left) with its corresponding cumulative density function (cdf) on right.

in the lightbulb case: if we only turn on the bulb $n = 1000$ times, the most likely number of failures will be zero. This rings just half the bell in Figure 5.2. That said, as $n \to \infty$ we will get a symmetric distribution peaking at the mean.

- *It is defined using only two parameters*: All we need are values of p and n to completely define a given binomial distribution.

Many things can be reasonably modeled by the binomial distribution. Recall the variance in the performance of a $p = 0.300$ hitter discussed in Section 2.2.3. There the probability of getting a hit with each trial was $p = 0.3$, with $n = 500$ trials per season. Thus the number of hits per season are drawn from a binomial distribution.

Realizing that it was a binomial distribution meant that we really didn't have to use simulation to construct the distribution. Properties like the expected number of hits $\mu = np = 500 \times 0.3 = 150$ and its standard deviation $\sigma = \sqrt{npq} = \sqrt{500 \times 0.3 \times 0.7} = 10.25$ simply fall out of closed-form formulas that you can look up when needed.

5.1.2 The Normal Distribution

A great many natural phenomenon are modeled by bell-shaped curves. Measured characteristics like height, weight, lifespan, and IQ all fit the same basic scheme: the bulk of the values lie pretty close to the mean, the distribution is symmetric, and no value is too extreme. In the entire history of the world, there has never been either a 12-foot-tall man or a 140-year-old woman.

The mother of all bell-shaped curves is the *Gaussian* or *normal distribution*, which is completely parameterized by its mean and standard deviation:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Figure 5.3 shows the pdf and cdf of the normal distribution. There are several things to note:

- *It is continuous*: The arguments to the normal distribution (mean μ and standard deviation σ) are free to arbitrary real numbers, with the lone constraint that $\sigma > 0$.

- *You probably can't explain where it comes from*: The normal distribution is a generalization of the binomial distribution, where $n \to \infty$ and the degree of concentration around the mean is specified by the parameter σ. Take your intuition here from the binomial distribution, and trust that Gauss got his calculations right: the great mathematician worked out the normal distribution for his Ph.D. dissertation. Or consult any decent statistics book if you are really curious to see where it comes from.

- *It truly is bell-shaped*: The Gaussian distribution is the platonic example of a bell-shaped curve. Because it operates on a continuous variable (like height) instead of a discrete count (say, the number of events) it is perfectly smooth. Because it goes infinitely in both directions, there is no truncation of the tails at either end. The normal distribution is a theoretical construct, which helps explain this perfection.

- *It is also defined using only two parameters*: However, these are different parameters than the binomial distribution! The normal distribution is completely defined by its central point (given by the mean μ) and its spread (given by the standard deviation σ). They are the only knobs we can use to tweak the distribution.

What's Normal?

An amazing number of naturally-occurring phenomenon are modeled by the normal distribution. Perhaps the most important one is measurement error. Every time you measure your weight on a bathroom scale, you will get a somewhat different answer, even if your weight has not changed. Sometimes the scale will read high and other times low, depending upon room temperature and the warping of the floor. Small errors are more likely than big ones, and slightly high is just as likely as slightly low. Experimental error is generally normally distributed as *Gaussian noise*.

Physical phenomenon like height, weight, and lifespan all have bell-shaped distributions, by similar arguments. Yet the claim that such distributions are *normal* is usually made too casually, without precisely specifying the underlying population. Is human height normally distributed? Certainly not: men and women have different mean heights and associated distributions. Is male height normally distributed? Certainly not: by including children in the mix and shrinking senior citizens you again have the sum of several different underlying distributions. Is the height of adult males in the United States normal? No, probably not even then. There are non-trivial populations with growth disorders

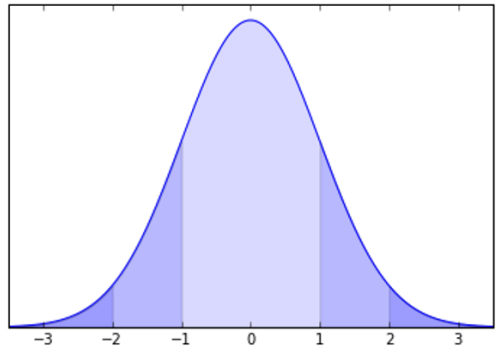

Figure 5.4: The normal distribution implies tight bounds on the probability of lying far from the mean. 68% of the values must lie within one sigma of the mean, and 95% within 2σ, and 99.7% within 3σ.

like dwarfism and acromegaly, that leave bunches of people substantially shorter and taller than could be explained by the normal distribution.

Perhaps the most famous bell-shaped but non-normal distribution is that of daily returns (percentage price movements) in the financial markets. A big market crash is defined by a large percentage price drop: on October 10, 1987, the Dow Jones average lost 22.61% of its value. Big stock market crashes occur with much greater frequency than can be accurately modeled by the normal distribution. Indeed, every substantial market crash wipes out a certain number of quants who assumed normality, and inadequately insured against such extreme events. It turns out that the *logarithm* of stock returns proves to be normally distributed, resulting in a distribution with far fatter tails than normal.

Although we must remember that bell-shaped distributions are not always normal, making such an assumption is a reasonable way to start thinking in the absence of better knowledge.

5.1.3 Implications of the Normal Distribution

Recall that the mean and standard deviation together always roughly characterize any frequency distribution, as discussed in Section 2.2.4. But they do a spectacularly good job of characterizing the normal distribution, because they *define* the normal distribution.

Figure 5.4 illustrates the famous 68%–95%–99% rule of the normal distribution. Sixty-eight percent of the probability mass must lie within the region $\pm 1\sigma$ of the mean. Further, 95% of the probability is within 2σ, and 99.7% within 3σ.

This means that values far from the mean (in terms of σ) are vanishingly rare in any normally distributed variable. Indeed the term *six sigma* is used to

connote quality standards so high that defects are incredibly rare events. We want plane crashes to be six sigma events. The probability of a 6σ event on the normal distribution is approximately 2 parts per billion.

Intelligence as measured by IQ is normally distributed, with a mean of 100 and standard deviation $\sigma = 15$. Thus 95% of the population lies within 2σ of the mean, from 70 to 130. This leaves only 2.5% of people with IQs above 130, and another 2.5% below 70. A total of 99.7% of the mass lies within 3σ of the mean, i.e. people with IQs between 55 and 145.

So how smart is the smartest person in the world? If we assume a population of 7 billion people, the probability of a randomly-selected person being smartest is approximately 1.43×10^{-10}. This is about the same probability of a single sample lying more than 6.5σ from the mean. Thus the smartest person in the world should have an IQ of approximately 197.5, according to this reckoning.

The degree to which you accept this depends upon how strongly you believe that IQ really is normally distributed. Such models are usually in grave danger of breaking down at the extremes. Indeed, by this model there is almost the same probability of there being someone dumb enough to earn a negative score on an IQ test.

5.1.4 Poisson Distribution

The *Poisson distribution* measures the frequency of intervals between rare events. Suppose we model human lifespan by a sequence of daily events, where there is a small but constant probability $1 - p$ that one happens to stop breathing today. A lifespan of exactly n days means successfully breathing for each of the first $n - 1$ days and then forever breaking the pattern on the nth day. The probability of living exactly n days is given by $Pr(n) = p^{n-1}(1 - p)$, yielding an expected lifespan

$$\mu = \sum_{k=0}^{\infty} k \cdot Pr(k).$$

The Poisson distribution basically follows from this analysis, but takes a more convenient argument than p. Instead it is based on μ, the average value of the distribution. Since each p defines a particular value of μ, these parameters are in some sense equivalent, but the average is much easier to estimate or measure. The Poisson distribution yields the very simple closed form:

$$Pr(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Once you start thinking the right way, many distributions begin to look Poisson, because they represent intervals between rare events.

Recall the binomial distribution lightbulb model from the previous section. This made it easy to compute the expected number of changes in Figure 5.2 (right), but not the lifespan distribution, which is Poisson. Figure 5.5 plots the associated Poisson distribution for $\mu = 1/p = 1000$, which shows that we should

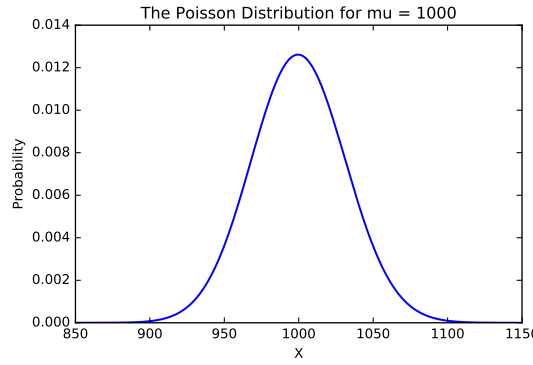

Figure 5.5: The lifespan distribution of lightbulbs with an expected life of $\mu = 1000$ hours, as modeled by a Poisson distribution.

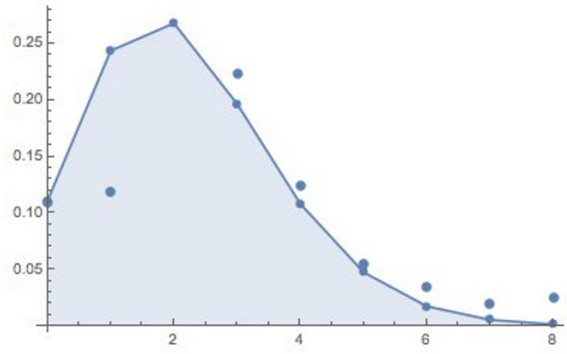

Figure 5.6: The observed fraction of families with x kids (isolated points) is accurately modeled by Poisson distribution, defined by an average of $\mu = 2.2$ children per family (polyline).

expect almost all bulbs to glow for between 900 and 1100 hours before the dying of the light.

Alternately, suppose we model the number of children by a process where the family keeps having children until after one too many tantrums, bake sales, or loads of laundry, a parent finally cracks. *"That's it! I've had enough of this. No more!"*

Under such a model, family size should be modeled as a Poisson distribution, where every day there is a small but non-zero probability of a breakdown that results in shutting down the factory.

How well does the "I've had it" model work to predict family size? The polygonal line in Figure 5.6 represents the Poisson distribution with the parameter $\lambda = 2.2$, meaning families have an average of 2.2 kids. The points represent the fraction of families with k children, drawn from the 2010 U.S. General Social

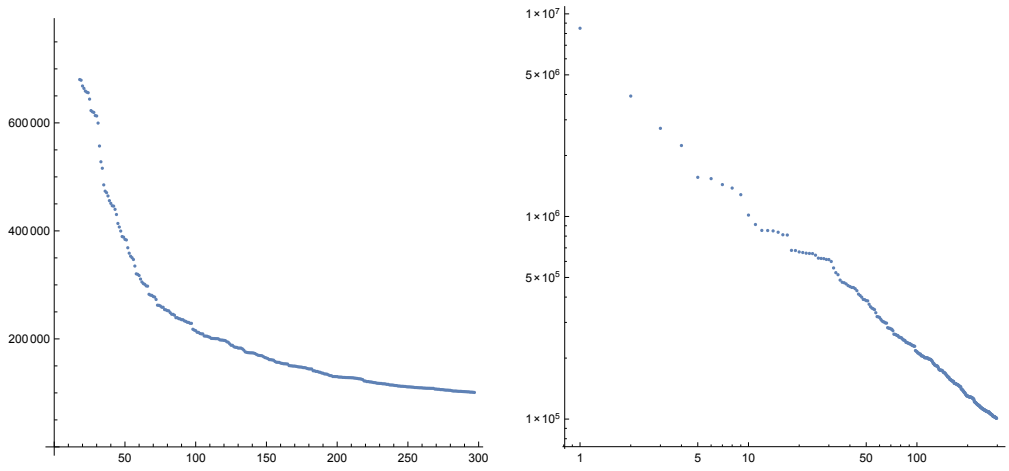

Figure 5.7: The population of U.S. cities by decreasing rank (left). On the right is the same data, now including the very largest cities, but plotted on a log-log scale. That they sit on a line is indicative of a power law distribution.

Survey (GSS).

There is excellent agreement over all family sizes except $k = 1$, and frankly, my personal experience suggests there are more singleton kids than this data set represents. Together, knowing just the mean and the formula for Poisson distribution enables us to construct a reasonable estimate of the real family-size distribution.

5.1.5 Power Law Distributions

Many data distributions exhibit much longer tails than could be possible under the normal or Poisson distributions. Consider, for example, the population of cities. There were exactly 297 U.S. cities in 2014 with populations greater than 100,000 people, according to Wikipedia. The population of the kth largest city, for $1 \leq k \leq 297$ is presented in Figure 5.7 (left). It shows that a relatively small number of cities have populations wildly dominating the rest. Indeed, the seventeen largest cities have populations so large they have been clipped off this plot so that we can see the rest.

These cities have a mean population of 304,689, with a ghastly standard deviation of 599,816. Something is wrong when the standard deviation is so large relative to the mean. Under a normal distribution, 99.7% of the mass lies within 3σ of the mean, thus making it unlikely that any of these cities would have a population above 2.1 million people. Yet Houston has a population of 2.2 million people, and New York (at 8.4 million people) is more than 13σ above the mean! City populations are clearly *not* normally distributed. In fact, they observe a different distribution, called a *power law*.

For a given variable X defined by a power law distribution,

$$P(X = x) = cx^{-\alpha}$$

This is parameterized by two constants: the exponent α and normalization constant c.

Power law distributions require some thinking to properly parse. The total probability defined by this distribution is the area under the curve:

$$A = \int_{x=-\infty}^{\infty} cx^{-\alpha} = c \int_{x=-\infty}^{\infty} x^{-\alpha}$$

The particular value of A is defined by the parameters α and c. The normalization constant c is chosen specifically for a given α to make sure that $A = 1$, as demanded by the laws of probability. Other than that, c is of no particular importance to us.

The real action happens with α. Note that when we double the value of the input (from x to $2x$), we decrease the probability by a factor of $f = 2^{-\alpha}$. This looks bad, but for any given α it is just a constant. So what the power law is really saying is that the probability of a $2x$-sized event is 2^{α} times less frequent than an x-sized event, for all x.

Personal wealth is well modeled by a power law, where $f \approx 0.2 = 1/5$. This means that over a large range, if Z people have x dollars, then $Z/5$ people have $2x$ dollars. One fifth as many people have \$200,000 than have \$100,000. If there are 625 people in the world worth \$5 billion, then there should be approximately 125 multi-billionaires each worth \$10 billion. Further, there should be 25 super-billionaires each worth \$20 billion, five hyper-billionaires at the \$40 billion level, and finally a single Bill Gates worth \$80 billion.

Power laws define the "80/20" rules which account for all the inequality of our world: the observation that the top 20% of the A gets fully 80% of the B. Power laws tend to arise whenever the rich get richer, where there is an increasing probability you will get more based on what you already have. Big cities grow disproportionately large because more people are attracted to cities when they are big. Because of his wealth, Bill Gates gets access to much better investment opportunities than I do, so his money grows faster than mine does.

Many distributions are defined by such preferential growth or attachment models, including:

- *Internet sites with x users*: Websites get more popular because they have more users. You are more likely to join Instagram or Facebook because your friends have already joined Instagram or Facebook. Preferential attachment leads to a power law distribution.

- *Words used with a relative frequency of x*: There is a long tail of millions of words like *algorist* or *defenestrate*[1] that are rarely used in the English

[1] *Defenestrate* means "to throw someone out a window."

language. On the other hand, a small set of words like *the* are used wildly more often than the rest.

Zipf's law governs the distribution of word usage in natural languages, and states that the kth most popular word (as measured by frequency rank) is used only $1/k$th as frequently as the most popular word. To gauge how well it works, consider the ranks of words based on frequencies from the English Wikipedia below:

Rank	Word	Count	Rank	Word	Count	Rank	Word	Count
1	the	25131726	1017	build	41890	10021	glances	2767
110	even	415055	2017	essential	21803	20026	ecclesiastical	881
212	men	177630	3018	sounds	13867	30028	zero-sum	405
312	least	132652	4018	boards	9811	40029	excluded	218
412	police	99926	5018	rage	7385	50030	sympathizes	124
514	quite	79205	6019	occupied	5813	60034	capon	77
614	include	65764	7020	continually	4650	70023	fibs	49
714	knowledge	57974	8020	delay	3835	80039	conventionalized	33
816	set	50862	9021	delayed	3233	90079	grandmom	23
916	doctor	46091	10021	glances	2767	100033	slum-dwellers	17

It should be convincing that frequency of use drops rapidly with rank: recall that *grandmom* is only a slang form of *grandma*, not the real McCoy.

Why is this a power law? A word of rank $2x$ has a frequency of $F_{2x} \sim F_1/2x$, compared to $F_x \sim F_1/x$. Thus halving the rank doubles the frequency, and this corresponds to the power law with $\alpha = 1$.

What is the mechanism behind the evolution of languages that lead to this distribution? A plausible explanation is that people learn and use words because they hear other people using them. Any mechanism that favors the already popular leads to a power law.

- *Frequency of earthquakes of magnitude x*: The Richter scale for measuring the strength of earthquakes is logarithmic, meaning a 5.3 quake is ten times stronger than a 4.3 scale event. Adding one to the magnitude multiplies the strength by a factor of ten.

  With such a rapidly increasing scale it makes sense that bigger events are rarer than smaller ones. I cause a 0.02 magnitude quake every time I flush a toilet. There are indeed billions of such events each day, but larger quakes get increasingly rare with size. Whenever a quantity grows in a potentially unbounded manner but the likelihood it does diminishes exponentially, you get a power law. Data shows this is as true of the energy released by earthquakes as it is with the casualties of wars: mercifully the number of conflicts which kill x people decreases as a power law.

Learn to keep your eyes open for power law distributions. You will find them everywhere in our unjust world. They are revealed by the following properties:

- *Power laws show as straight lines on log value, log frequency plots*: Check out the graph of city populations in Figure 5.7 (right). Although there are some gaps at the edges where data gets scarce, by and large the points lie

neatly on a line. This is *the* main characteristic of a power law. By the way, the slope of this line is determined by α, the constant defining the shape of the power law distribution.

- *The mean does not make sense*: Bill Gates alone adds about $250 to the wealth of the average person in the United States. This is weird. Under a power law distribution there is a very small but non-zero probability that someone will have infinite wealth, so what does this do to the mean? The median does a much better job of capturing the bulk of such distributions than the observed mean.

- *The standard deviation does not make sense*: In a power law distribution, the standard deviation is typically as large or larger than the mean. This means that the distribution is very poorly characterized by μ and σ, while the power law provides a very good description in terms of α and c.

- *The distribution is scale invariant*: Suppose we plotted the populations of the 300th through 600th largest U.S. cities, instead of the top 300 as in Figure 5.7 (left). The shape would look very much the same, with the population of the 300th largest city towering over the tail. Any exponential function is *scale invariant*, because it looks the same at any resolution. This is a consequence of it being a straight line on a log-log plot: any subrange is a straight line segment, which has the same parameters in its window as the full distribution.

Take-Home Lesson: Be on the lookout for power law distributions. They reflect the inequalities of the world, which means that they are everywhere.

5.2 Sampling from Distributions

Sampling points from a given probability distribution is a common operation, one which it is pays to know how to do. Perhaps you need test data from a power law distribution to run a simulation, or to verify that your program operates under extreme conditions. Testing whether your data in fact fits a particular distribution requires something to compare it against, and that should generally be properly-generated synthetic data drawn from the canonical distribution.

There is a general technique for sampling from any given probability distribution, called *inverse transform sampling*. Recall that we can move between the probability density function P and the cumulative density function C by integration and differentiation. We can move back and forth between them because:

$$P(k = X) = C'(k) = C(X \leq k + \delta) - C(X \leq k), \text{ and}$$

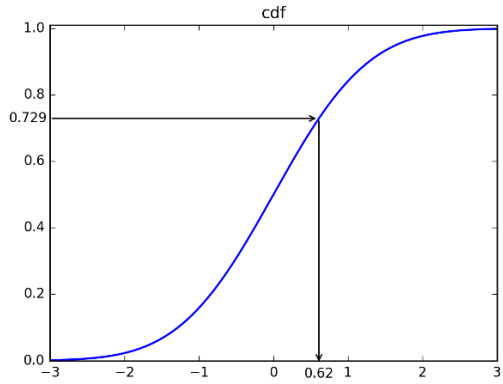

Figure 5.8: The inverse transform sampling method enables us to convert a random number generated uniformly from $[0, 1]$ (here 0.729) to a random sample drawn from any distribution, given its cdf.

$$C(X \leq k) = \int_{x=-\infty}^{k} P(X = x).$$

Suppose I want to sample a point from this possibly very complicated distribution. I can use a uniform random number generator to select a value p in the interval $[0, \ldots, 1]$. We can interpret p as a probability, and use it as an index on the cumulative distribution C. Precisely, we report the exact value of x such that $C(X \leq x) = p$.

Figure 5.8 illustrates the approach, here sampling from the normal distribution. Suppose $p = 0.729$ is the random number selected from our uniform generator. We return the x value such that $y = 0.729$, so $x = 0.62$ as per this cdf.

If you are working with a popular probability distribution in a well-supported language like Python, there is almost certainly a library function to generate random samples already available. So look for the right library before you write your own.

5.2.1 Random Sampling beyond One Dimension

Correctly sampling from a given distribution becomes a very subtle problem once you increase the number of dimensions. Consider the task of sampling points uniformly from within a circle. Think for a moment about how you might do this before we proceed.

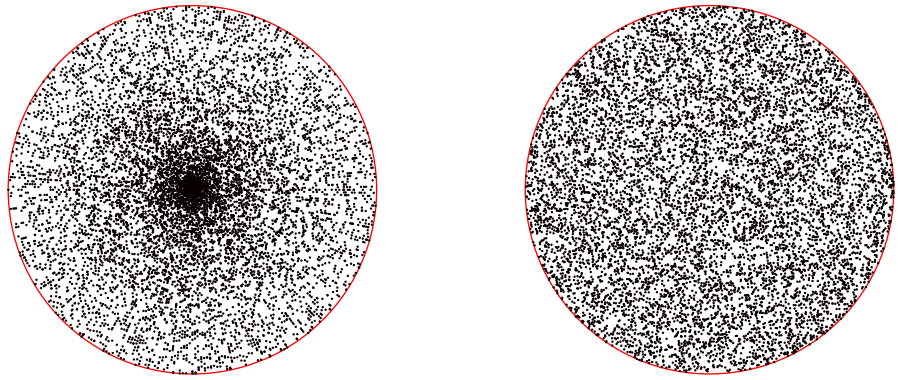

Figure 5.9: Randomly generating 10,000 points by angle-radius pairs clearly oversamples near the origin of the circle (left). In contrast, Monte Carlo sampling generates points uniformly within the circle (right).

The clever among you may hit upon the idea of sampling the angle and distance from the center independently. The angle that any sampled point must make with respect to the origin and positive x-axis varies between 0 and 2π. The distance from the origin must be a value between 0 and r. Select these coordinates uniformly at random and you have a random point in the circle.

This method is clever, but wrong. Sure, any point so created must lie within the circle. But the points are not selected with uniform frequency. This method will generate points where half of them will lie within a distance of at most $r/2$ from the center. But most of the area of the circle is farther from the center than that! Thus we will oversample near the origin, at the expense of the mass near the boundary. This is shown by Figure 5.9 (left), a plot of 10,000 points generated using this method.

A dumb technique that proves correct is *Monte Carlo sampling*. The x and y coordinates of every point in the circle range from $-r$ to r, as do many points outside the circle. Thus sampling these values uniformly at random gives us a point which lies in a bounding box of the circle, but not always within the circle itself. This can be easily tested: is the distance from (x, y) to the origin at most r, i.e. is $\sqrt{x^2 + y^2} \leq r$? If yes, we have found a random point in the circle. If not, we toss it out and try again. Figure 5.9 (right) plots 10,000 points constructed using this method: see how uniformly they cover the circle, without any obvious places of over- or under-sampling.

The efficiency here depends entirely upon the ratio of the desired region volume (the area of the circle) to the volume of the bounding box (the area of a square). Since 78.5% of this bounded box is occupied by the circle, less than two trials on average suffice to find each new circle point.

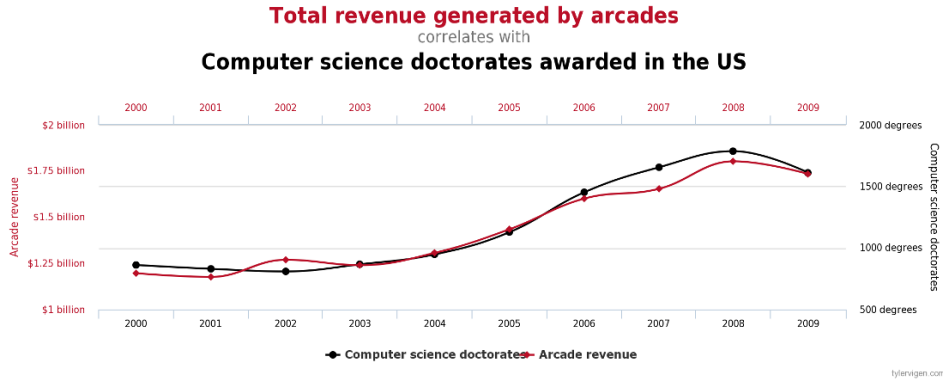

Figure 5.10: Correlation vs. causation: the number of Computer Science Ph.Ds awarded each year in the United States strongly correlates with video/pinball arcade revenue. (from [Vig15])

5.3 Statistical Significance

Statisticians are largely concerned with whether observations on data are significant. Computational analysis will readily find a host of patterns and correlations in any interesting data set. But does a particular correlation reflect a real phenomena, as opposed to just chance? In other words, when is an observation really *significant*?

Sufficiently strong correlations on large data sets may seem to be "obviously" meaningful, but the issues are often quite subtle. For one thing, *correlation does not imply causation.* Figure 5.10 convincingly demonstrates that the volume of advanced study in computer science correlates with how much video games are being played. I'd like to think I have driven more people to algorithms than Nintendo, but maybe this is just the same thing? The graphs of such spurious correlations literally fill a book [Vig15], and a very funny one at that.

The discipline of statistics comes into its own in making subtle distinctions about whether an observation is meaningful or not. The classical example comes from medical statistics, in determining the efficacy of drug treatments. A pharmaceutical company conducts an experiment comparing two drugs. Drug A cured 19 of 34 patients. Drug B cured 14 of 21 patients. Is drug B really better than drug A? FDA approval of new drugs can add or subtract billions from the value of drug companies. But can you be sure that a new drug represents a real improvement? How do you tell?

5.3.1 The Significance of Significance

Statistical significance measures our confidence that there is a genuine difference between two given distributions. This is important. But statistical significance does not measure the importance or magnitude of this difference. For large

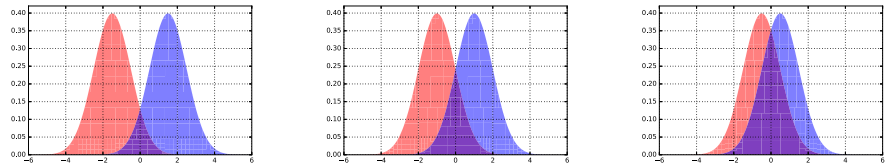

Figure 5.11: Pairs of normal distributions with the same variance, but decreasing difference in their means from left to right. As the means get closer, the greater overlap between the distributions makes it harder to tell them apart.

enough sample sizes, extremely small differences can register as highly significant on statistical tests.

For example, suppose I get suckered into betting tails on a coin which comes up heads 51% of the time, instead of the 50% we associate with a fair coin. After 100 tosses of a fair coin, I would expect to see 51% or more heads 46.02% of the time, so I have absolutely no grounds for complaint when I do. After 1,000 tosses, the probability of seeing at least 510 heads falls to 0.274. By 10,000 tosses, the probability of seeing so many heads is only 0.0233, and I should start to become suspicious of whether the coin is fair. After 100,000 tosses, the probability of fairness will be down to 1.29×10^{-10}, so small that I must issue a formal complaint, even if I thought my opponent to be a gentleman.

But here is the thing. Although it is now crystal clear that I had been tricked into using a biased coin, the consequences of this act are not substantial. For almost any issue in life worth flipping over, I would be willing to take the short side of the coin, because the stakes are just not high enough. At $1 bet per flip, my expected loss even after 100,000 tosses would only be $1,000 bucks.

Significance tells you how unlikely it is that something is due to chance, but not whether it is important. We really care about *effect size*, the magnitude of difference between the two groups. We informally categorize a *medium*-level effect size as visible to the naked eye by a careful observer. On this scale, *large* effects pop out, and *small* effects are not completely trivial [SF12]. There are several statistics which try to measure the effect size, including:

- *Cohen's d*: The importance of the difference between two means μ and μ' depends on the absolute magnitude of the change, but also the natural variation of the distributions as measured by σ or σ'. This effect size can be measured by:

  $$d = (|\mu - \mu'|)/\sigma.$$

  A reasonable threshold for a small effect size is > 0.2, medium effect > 0.5, and large effect size > 0.8.

- *Pearson's correlation coefficient r*: Measures the degree of linear relationship between two variables, on a scale from -1 to 1. The thresholds for effect sizes are comparable to the mean shift: small effects start at ± 0.2,

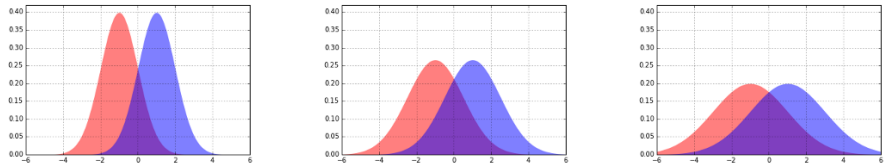

Figure 5.12: Pairs of normal distributions with the same difference in their means but increasing variance, from left to right. As the variance increases, there becomes greater overlap between the distributions, making it harder to tell them apart.

medium effects about ± 0.5, and large effect sizes require correlations of ± 0.8.

- *The coefficient of variation* r^2: The square of the correlation coefficient reflects the proportion of the variance in one variable that is explained by the other. The thresholds follow from squaring those above. Small effects explain at least 4% of the variance, medium effects $\geq 25\%$ and large effect sizes at least 64%.

- *Percentage of overlap*: The area under any single probability distribution is, by definition, 1. The area of intersection between two given distributions is a good measure of their similarity, as shown in Figure 5.11. Identical distributions overlap 100%, while disjoint intervals overlap 0%. Reasonable thresholds are: for small effects 53% overlap, medium effects 67% overlap, and large effect sizes 85% overlap.

Of course, any sizable effect which is not statistically significant is inherently suspect. The CS study vs. video game play correlation in Figure 5.10 was so high ($r = 0.985$) that the effect size would be huge, were the number of sample points and methodology sound enough to support the conclusion.

Take-Home Lesson: Statistical significance depends upon the number of samples, while the effect size does not.

5.3.2 The T-test: Comparing Population Means

We have seen that large mean shifts between two populations suggest large effect sizes. But how many measurements do we need before we can safely believe that the phenomenon is real. Suppose we measure the IQs of twenty men and twenty women. Does the data show that one group is smarter, on average? Certainly the sample means will differ, at least a bit, but is this difference significant?

The t-test evaluates whether the population means of two samples are different. This problem commonly arises in *AB testing*, associated with evaluating

whether a product change makes a difference in performance. Suppose you show one group of users version A, and another group version B. Further, suppose you measure a system performance value for each user, such as the number of times they click on ads or the number of stars they give it when asked about the experience. The t-test measures whether the observed difference between the two groups is significant.

Two means differ significantly if:

- *The mean difference is relatively large*: This makes sense. One can conclude that men weigh more than women on average fairly easily, because the effect size is so large. According to the Center for Disease Control[2] the average American male weighed 195.5 pounds in 2010, whereas the average American woman weighted 166.2 pounds. This is huge. Proving that a much more subtle difference, like IQ, is real requires much more evidence to be equally convincing.

- *The standard deviations are small enough*: This also makes sense. It is easy to convince yourself that men and women have, on average, the same number of fingers because the counts we observe are very tightly bunched around the mean: $\{10, 10, 10, 10, 9, 10 \ldots\}$. The equal-finger count hypothesis would require much more evidence if the numbers jumped around a lot. I would be reluctant to commit to a true distributional average of $\mu = 10$ if I what I observed was {3, 15, 6, 14, 17, 5}.

- *The number of samples are large enough*: This again makes sense. The more data I see, the more solidly I become convinced that the sample will accurately represent its underlying distribution. For example, men undoubtedly have *fewer* fingers on average than women, as a consequence of more adventures with power tools.[3] But it would require a very large number of samples to observe and validate this relatively rare phenomenon.

The t-test starts by computing a *test statistic* on the two sets of observations. Welch's t-statistic is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

where $\bar{x}_i$, σ_i, and n_i are the mean, standard deviation, and population size of sample i, respectively.

Let us parse this equation carefully. The numerator is the difference between the means, so the bigger this difference, the bigger the value of the t-statistic. The standard deviations are in the denominator, so the smaller that σ_i is, the bigger the value of the t-statistic. If this is confusing, recall what happens when you divide x by a number approaching zero. Increasing the sample sizes n_i

[2] `http://www.cdc.gov/nchs/fastats/obesity-overweight.htm`

[3] This observation alone may be sufficient to resolve the gender–IQ relationship, without the need for additional statistical evidence.

also makes the denominator smaller, so the larger n_i is, the bigger the value of the t-statistic. In all cases, the factors that make us more confident in there being a real difference between the two distributions increases the value of the t-statistic.

Interpreting the meaning of a particular value of the t-statistic comes from looking up a number in an appropriate table. For a desired *significance level* α and number of *degrees of freedom* (essentially the sample sizes), the table entry specifies the value v that the t-statistic t must exceed. If $t > v$, then the observation is significant to the α level.

Why Does This Work?

Statistical tests like the t-test often seem like voodoo to me, because we look up a number from some magic table and treat it like gospel. *The oracle has spoken: the difference is significant!* Of course there is real mathematics behind significance testing, but the derivation involves calculus and strange functions (like the gamma function $\Gamma(n)$, a real numbered generalization of factorials). These complex calculations are why the convention arose to look things up in a precomputed table, instead of computing it yourself.

You can find derivations of the relevant formulae in any good statistics book, if you are interested. These tests are based on ideas like random sampling. We have seen how the mean and standard deviation constrain the shape of any underlying probability distribution. Getting a sample average very far from the mean implies bad luck. Randomly picking values several standard deviations away from the population mean is very unlikely, according to the theory. This makes it more likely that observing such a large difference is the result of drawing from a different distribution.

Much of the technicality here is a consequence of dealing with subtle phenomenon and small data sets. Historically, observed data was a very scarce resource, and it remains so in many situations. Recall our discussion of drug efficacy testing, where someone new must die for every single point we collect. The big data world you will likely inhabit generally features more observations (everybody visiting our webpage), lower stakes (do customers buy more when you show them a green background instead of a blue background?), and perhaps smaller effect sizes (how big an improvement do we really need to justify changing the background color?).

5.3.3 The Kolmogorov-Smirnov Test

The t-test compares two samples drawn from presumably normal distributions according to the distance between their respective means. Instead, the *Kolmogorov-Smirnov* (KS) test compares the cumulative distribution functions (cdfs) of the two sample distributions and assesses how similar they are.

This is illustrated in Figure 5.13. The cdfs of the two different samples are plotted on the same chart. If the two samples are drawn from the same distribution, the ranges of x values should largely overlap. Further, since both

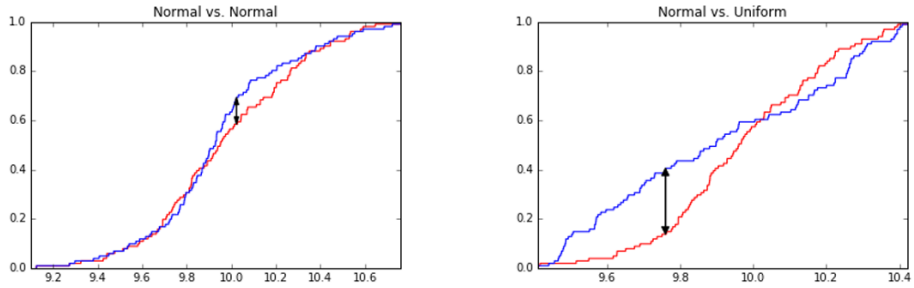

Figure 5.13: The Kolmogorov-Smirnov test quantifies the difference between two probability distributions by the maximum y-distance gap between the two cumulative distribution functions. On the left, two samples from the same normal distribution. On the right, comparison of samples from uniform and normal distributions drawn over the same x-range.

distributions are represented as cdfs, the y-axis represents cumulative probability from 0 to 1. Both functions increase monotonically from left to right, where $C(x)$ is the fraction of the sample $\leq x$.

We seek to identify the value of x for which the associated y values of the two cdfs differ by as much as possible. The distance $D(C_1, C_2)$ between the distributions C_1 and C_2 is the difference of the y values at this critical x, formally stated as

$$D(C_1, C_2) = \max_{-\infty \leq x \leq \infty} |C_1(x) - C_2(x)|$$

The more substantially that two sample distributions differ at some value, the more likely it is that they were drawn from different distributions. Figure 5.13 (left) shows two independent samples from the same normal distribution. Note the resulting tiny gap between them. In contrast, Figure 5.13 (right) compares a sample drawn from a normal distribution against one drawn from the uniform distribution. The KS-test is not fooled: observe the big gaps near the tails, where we would expect to see it.

The KS-test compares the value of $D(C_1, C_2)$ against a particular target, declaring that two distributions differ at the significance level of α when:

$$D(C_1, C_2) > c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where $c(\alpha)$ is a constant to look up in a table.

The function of the sample sizes has some intuition behind it. Assume for simplicity that both samples have the same size, n. Then

$$\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = \sqrt{\frac{2n}{n^2}} = \sqrt{\frac{2}{n}}$$

The quantity $\sqrt{n}$ arises naturally in sampling problems, such as the standard deviation of the binomial distribution. The expected difference between the number of heads and tails in n coin flips is on the order of $\sqrt{n}$. In the context of the KS-test, it similarly reflects the expected deviation when two samples should be considered the same. The KS-test reflects what is happening in the meat of the distribution, where a robust determination can be made.

I like the Kolmogorov-Smirnov test. It provides pictures of the distributions that I can understand, that identify the weakest point in the assumption that they are identical. This test has fewer technical assumptions and variants than the t-test, meaning we are less likely to make a mistake using it. And the KS-test can be applied to many problems, including testing whether points are drawn from a normal distribution.

Normality Testing

When plotted, the normal distribution yields a bell-shaped curve. But not every bell-shaped distribution is normal, and it is sometimes important to know the difference.

There exist specialized statistical tests for testing the normality of a given distributional samples f_1. But we can use the general KS-test to do the job, provided we can identify a meaningful f_2 to compare f_1 to.

This is exactly why I introduced random sampling methods in Section 5.2. Using the cumulative distribution method we described, statistically-sound random samples of n points can be drawn from any distribution that you know the cdf of, for any n. For f_2, we should pick a meaningful number of points to compare against. We can use $n_2 = n_1$, or perhaps a somewhat larger sample if n_1 is very small. We want to be sure that we are capturing the shape of the desired distribution with our sample.

So if we construct our random sample for f_2 from the normal distribution, the KS-test should not be able to distinguish f_1 from f_2 if f_1 also comes from a normal distribution on the same μ and σ.

One word of caution. A sufficiently-sensitive statistical test will probably reject the normality of just about *any* observed distribution. The normal distribution is an abstraction, and the world is a complicated place. But looking at the plot of the KS-test shows you exactly where the deviations are happening. Are the tails too fat or too lean? Is the distribution skewed? With this understanding, you can decide whether the differences are big enough to matter to you.

5.3.4 The Bonferroni Correction

It has long been the convention in science to use $\alpha = 0.05$ as the cutoff between statistical significance and irrelevance. A statistical significance of 0.05 means there is a probability of 1/20 that this result would have come about purely by chance.

This is not an unreasonable standard when collecting data to test a challenging hypothesis. Betting on a horse at 20 to 1 odds and winning your bet is a noteworthy accomplishment. Unless you simultaneously placed bets on millions of other horses. Then bragging about the small number of 20-1 bets where you actually won would be misleading, to say the least.

Thus fishing expeditions which test millions of hypotheses must be held to higher standards. *This* is the fallacy that created the strong but spurious correlation between computer science Ph.Ds and video game activity in Figure 5.10. It was discovered in the course of comparing thousands of time series against each other, and retaining only the most amusing-sounding pairs which happened to show a high correlation score.

The *Bonferroni correction*[4] provides an important balance in weighing how much we trust an apparently significant statistical result. It speaks to the fact that how you found the correlation can be as important as the strength of the correlation itself. Someone who buys a million lottery tickets and wins once has much less impressive mojo going than the fellow who buys a single ticket and wins.

The Bonferroni correction states that when testing n different hypotheses simultaneously, the resulting p-value must rise to a level of α/n, in order to be considered as significant at the α level.

As with any statistical test, there lurk many subtleties in properly applying the correction. But the big principle here is important to grasp. Computing people are particularly prone to running large-scale comparisons of all things against all things, or hunting for unusual outliers and patterns. After all, once you've written the analysis program, why not run it on all your data? Presenting only the best, cherry-picked results makes it easy to fool other people. The Bonferroni correction is the way to prevent you from fooling yourself.

5.3.5 False Discovery Rate

The Bonferroni correction safeguards us from being too quick to accept the significance of a lone successful hypothesis among many trials. But often when working with large, high-dimensional data we are faced with a different problem. Perhaps all m of the variables correlate (perhaps weakly) with the target variable. If n is large enough, many of these correlations will be statistically significant. Have we really made so many important discoveries?

The *Benjamini-Hochberg* procedure for minimizing *false discovery rate* (FDR) procedure gives a very simple way to draw the cutoff between interesting and uninteresting variables based on significance. Sort the variables by the strength of their p-value, so the more extreme variables lie on the left, and the least significant variables on the right. Now consider the ith ranked variable in this ordering. We accept the significance of this variable at the α level if

$$\forall_{j=1}^{i} \ (p_j \leq \frac{j}{m}\alpha)$$

[4] I have always thought that "The Bonferroni Correction" would make a fabulous title for an action movie. Dwayne Johnson as Bonferroni?

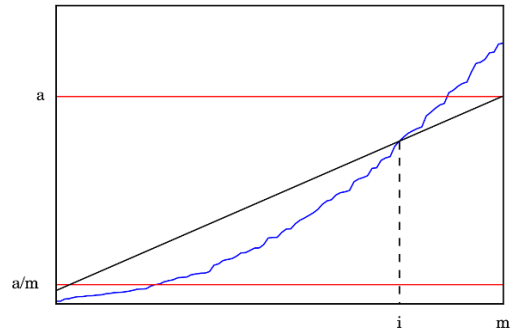

Figure 5.14: The Benjamini-Hochberg procedure minimizes false discovery rate, by accepting p-values only when $p_i \leq \alpha i/m$. The blue curve shows the sorted p-values, and the diagonal defines the cutoff for when such a p-value is significant.

The situation is illustrated in Figure 5.14. The p-values are sorted in increasing order from left to right, as denoted by the irregular blue curve. If we were accepting all p-values less than α, we accept too many. This is why Bonferroni developed his correction. But requiring all p-values to meet the standard of the Bonferroni correction (where the curve crosses α/m) is too stringent.

The Benjamini-Hochberg procedure recognizes that if many values are really significant to a certain standard, a certain fraction of them should be significant to a much higher standard. The diagonal line in Figure 5.14 appropriately enforces this level of quality control.

5.4 War Story: Discovering the Fountain of Youth?

It had been a beautiful wedding. We were very happy for Rachel and David, the bride and groom. I had eaten like a king, danced with my lovely wife, and was enjoying a warm post-prime rib glow when it hit me that something was off. I looked around the room and did a double-take. Somehow, for the first time in many years, I had become younger than most of the people in the crowd.

This may not seem like a big deal to you, but that is because you the reader probably *are* younger than most people in many settings. But trust me, there will come a time when you notice such things. I remember when I first realized that I had been attending college at the time when most of my students were being born. Then they started being born when I was in graduate school. Today's college students were not only born after I became a professor, but after I got tenure here. So how could I be younger than most of the people at this wedding?

There were two possibilities. Either it was by chance that so many older

people entered the room, or there was a reason explaining this phenomenon, This is why statistical significance tests and p-values were invented, to aid in distinguishing something from nothing.

So what was the probability that I, then at age 54, would have been younger than most of the 251 people at Rachel's wedding? According to Wolfram Alpha (more precisely, the 2008–2012 American Community Survey five-year estimates), there were 309.1 million people in the United States, of whom 77.1 million were age 55 or older. Almost exactly 25% of the population is older than I am as I write these words.

The probability that the majority of 251 randomly selected Americans would be older than 55 years is thus given by:

$$p = \sum_{i=126}^{251} \binom{251}{i} (1 - .75)^i (0.75)^{(251-i)} = 8.98 \times 10^{-18}$$

This probability is impossibly small, comparable to pulling a fair coin out of your pocket and having it come up heads 56 times in a row. This could not have been the result of a chance event. There had to be a reason why I was junior to most of this crowd, and the answer wasn't that I was getting any younger.

When I asked Rachel about it, she mentioned that, for budgetary reasons, they decided against inviting children to the wedding. This seemed like it could be a reasonable explanation. After all, this rule excluded 73.9 million people under the age of eighteen from attending the wedding, thus saving billions of dollars over what it would have cost to invite them all. The fraction f of people younger than me who are not children works out to $f = 1 - (77.1/(309.1 - 73.9)) = 0.672$. This is substantially larger than 0.5, however. The probability of my being younger than the median in a random sample drawn from this cohort is:

$$p = \sum_{i=126}^{251} \binom{251}{i} (1 - .0672)^i (0.672)^{(251-i)} = 9.118 \times 10^{-9}$$

Although this is much larger than the previous p-value, it is still impossibly small: akin to tossing off 27 straight heads on your fair coin. Just forbidding children was not nearly powerful enough to make me young again.

I went back to Rachel and made her fess up. It turns out her mother had an unusually large number of cousins growing up, and she was exceptionally good at keeping in touch with *all* of them. Recall Einstein's Theory of Relativity, where $E = mc^2$ denotes that everyone is my mother's cousin, twice removed. *All* of these cousins were invited to the wedding. With Rachel's family outpopulating the groom's unusually tiny clan, this cohort of senior cousins came to dominate the dance floor.

Indeed, we can compute the number of the older-cousins (c) that must be invited to yield a 50/50 chance that I would be younger than the median guest, assuming the rest of the 251 guests were selected at random. It turns out that

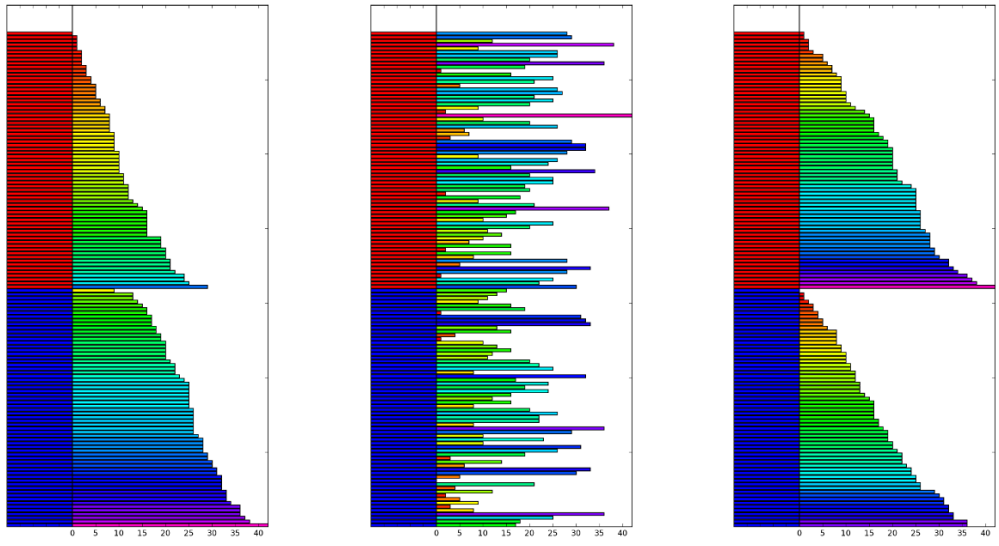

Figure 5.15: Permutation tests reveal the significance of the correlation between gender and the height (left). A random assignment of gender to height (center) results in a substantially different distribution of outcomes when sorted (right), validating the significance of the original relationship.

$c = 65$ single cousins (or 32.5 married pairs) suffice, once the children have been excluded ($f = 0.672$).

The moral here is that it is important to compute the probability of any interesting observation before declaring it to be a miracle. Never stop with a partial explanation, if it does not reduce the surprise to plausible levels. There likely is a genuine phenomenon underlying any sufficiently rare event, and ferreting out what it is makes data science exciting.

5.5 Permutation Tests and P-values

Traditional statistical significance tests prove quite effective at deciding whether two samples are in fact drawn from the same distribution. However, these tests must be properly performed in order to do their job. Many standard tests have subtleties like the issues of one- vs. two-sided tests, distributional assumptions, and more. Performing these tests correctly requires care and training.

Permutation tests allow a more general and computationally idiot-proof way to establish significance. If your hypothesis is supported by the data, then

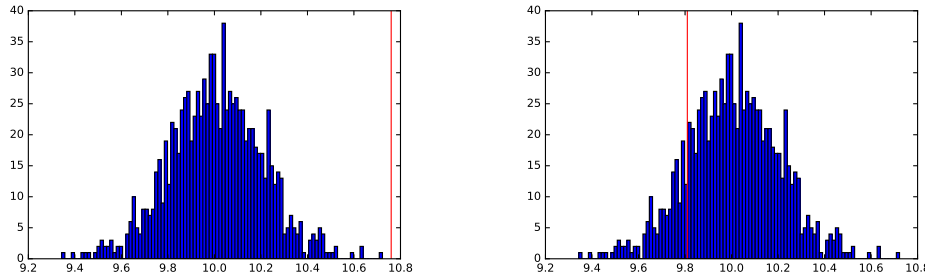

Figure 5.16: Permutation tests score significance by the position of the score on actual data against a distribution of scores produced by random permutations. A position on the extreme tail (left) is highly significant, but one within the body of the distribution (left) is uninteresting.

randomly shuffled data sets should be less likely to support it. By conducting many trials against randomized data, we can establish exactly how unusual the phenomenon is that you are testing for.

Consider Figure 5.15, where we denote the independent variable (gender: male or female) and dependent variable (say, height) using colors. The original outcome color distribution (left) looks starkly different between men and women, reflective of the genuine differences in height. But how unusual is this difference? We can construct a new data set by randomly assigning gender to the original outcome variables (center). Sorting within each group makes clear that the pseudo-male/female distribution of outcomes is now much more balanced than in the original data (right). This demonstrates that gender *was* indeed a significant factor in determining height, a conclusion we would come to believe even more strongly after it happens again and again over 1,000 or 1,000,000 trials.

The rank of the test statistic on the real data among the distribution of statistic values from random permutations determines the significance level or *p-value*. Figure 5.16 (left) shows what we are looking for. The real value lies on the very right of the distribution, attesting to significant. In the figure on right, the real value lies in the squishy middle of the distribution, suggesting no effect.

Permutation tests require that you develop a statistic which reflects your hypothesis about the data. The correlation coefficient is a reasonable choice if you want to establish an important relationship between a specific pair of variables. Ideally the observed correlation in the real data will be stronger than in any random permutation of it. To validate the gender–height connection, perhaps our statistic could be the difference in the average heights of men and women. Again, we hope this proves larger in the real data than in most random permutations of it.

Be creative in your choice of statistic: the power of permutation tests is

that they can work with pretty much anything you can come up with to prove your case. It is best if your statistic minimizes the chance of ties, since you are honor-bound to count all ties against your hypothesis.

> *Take-Home Lesson*: Permutation tests give you the probability of your data given your hypothesis, namely that the statistic will be an outlier compared to the random sample distribution. This is not quite the same as proving your hypothesis given the data, which is the traditional goal of statistical significance testing. But it is much better than nothing.

The significance score or p-value of a permutation test depends upon how many random tries are run. Always try to do at least 1,000 random trials, and more if it is feasible. The more permutations you try, the more impressive your significance p-value can be, at least up to a point. If the given input is in fact the best of all $k!$ permutations, the most extreme p-value you can get is $1/k!$, no matter how many random permutations you try. Oversampling will inflate your denominator without increasing your true confidence one bit.

> *Take-Home Lesson*: P-values are computed to increase *your* confidence that an observation is real and interesting. This only works when you do the permutation test honestly, by performing experiments that can provide a fair measure of surprise.

5.5.1 Generating Random Permutations

Generating random permutations is another important sampling problem that people often botch up. The two algorithms below both use sequences of random swaps to scramble up the initial permutation $\{1, 2, \ldots, n\}$.

But ensuring that all $n!$ permutations are generated uniformly at random is a tricky business. Indeed only one of these algorithms gets it right. Is it this one,

for $i = 1$ to n do $a[i] = i$;
for $i = 1$ to $n - 1$ do $swap[a[i], a[Random[i, n]]$;

or is it this:

for $i = 1$ to n do $a[i] = i$;
for $i = 1$ to $n - 1$ do $swap[a[i], a[Random[1, n]]$;

Think about this carefully: the difference here is very subtle. It is so subtle you might not even notice it in the code. The critical difference is the 1 or i in the call to Random. One of these algorithms is right and one of these algorithms is wrong. If you think you can tell, convincingly explain why one works and the other one doesn't.

If you really must know, the first algorithm is the correct one. It picks a random element from 1 to n for the first position, then leaves it alone and

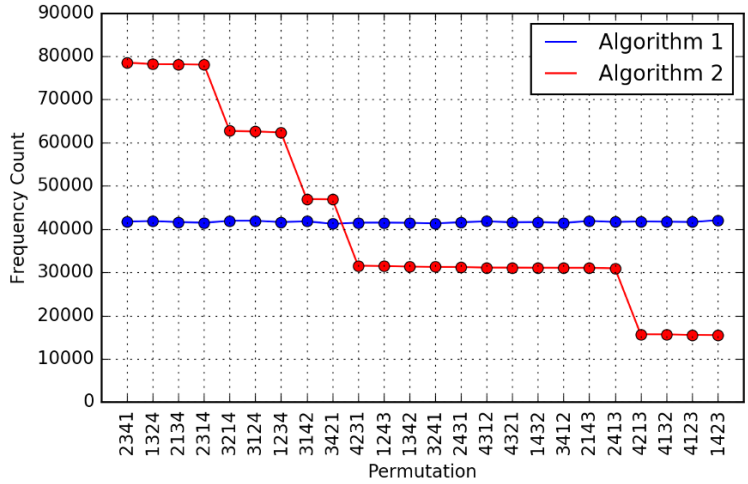

Figure 5.17: The generation frequency of all 4! = 24 permutations using two different algorithms. Algorithm 1 generates them with uniform frequency, while algorithm 2 is substantially biased.

recurs on the rest. It generates permutations uniformly at random. The second algorithm gives certain elements a better chance to end up first, showing that the distribution is not uniform.

But if you can't prove this theoretically, you can use the idea of a permutation test. Implement both of the algorithms, and perform 1,000,000 runs of each, constructing random permutations of, say, $n = 4$ elements. Count how often each algorithm generates each one of the 4! = 24 distinct permutations. The results of such an experiment are shown in Figure 5.17. Algorithm 1 proves incredibly steady, with a standard deviation of 166.1 occurrences. In contrast, there is an eight-fold difference between the most and least frequent permutations under algorithm 2, with $\sigma = 20,923.9$.

The moral here is that random generation can be very subtle. And that Monte Carlo-type experiments like permutation tests can eliminate the need for subtle reasoning. Verify, then trust.

5.5.2 DiMaggio's Hitting Streak

One of baseball's most amazing records is Joe DiMaggio's 56-game hitting streak. The job of a batter is to get hits, and they receive perhaps four chances every game to get one. Even very good hitters often fail.

But back in 1941, Joe DiMaggio succeeded in getting hits in 56 straight games, a truly amazing accomplishment. No player in the seventy-five years since then has come close to this record, nor any one before him.

But how unusual was such a long streak in the context of his career? DiMag-

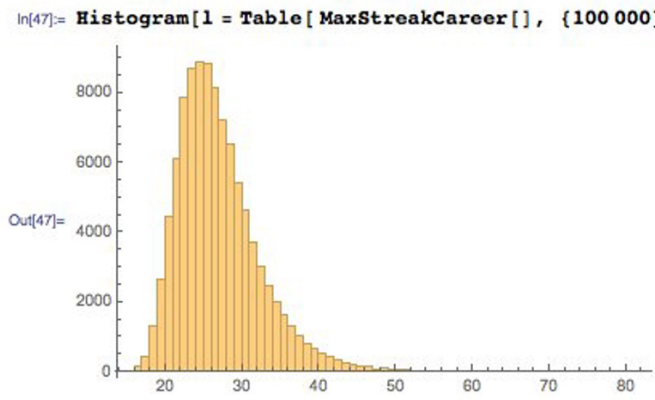

Figure 5.18: The distribution of longest hitting streaks, over 100,000 simulated careers. DiMaggio's actual 56-game hitting streak stands at the very tail of this distribution, thus demonstrating the difficulty of this feat.

gio played 1736 games, with 2214 hits in 6821 at bats. Thus he should get hits in roughly $1 - (1 - (2214/6821))^4 = 79.2\%$ of his games with four at bats. What is the probability that someone with his skill level could manage such a consecutive game streak in the course of their career?

For those of you tired of my baseball analogies, let's put this in another context. Suppose you are a student who averages a grade of 90 on tests. You are a very good student to be sure, but not perfect. What are the chances you could have a hot streak where you scored above 90 on ten straight tests? What about twenty straight? Could you possibly ace 56 tests in a row?[5] If such a long streak happened, would that mean that you had taken your studies to another level, or did you just get lucky?

So when DiMaggio had his hitting streak, was it just an expected consequence of his undisputed skills and consistency, or did he just get lucky? He was one of the very best hitters of his or any time, an all-star every season of his thirteen-year career. But we also know that DiMaggio got lucky from time to time. After all, he *was* married to the movie star Marilyn Monroe.

To resolve this question, we used random numbers to simulate when he got hits over a synthetic "career" of 1736 games. Each game, the simulated Joe received four chances to hit, and succeeded with a probability of $p = (2214/6821) = 0.325$. We could then identify the longest hitting streak over the course of this simulated career. By simulating 100,000 DiMaggio careers, we get a frequency distribution of streaks that can put the rarity of his accomplishment in context, getting a p-value in the process.

The results are shown in Figure 5.18. In only 44 of 100,000 simulated careers ($p = 0.00044$) did DiMaggio manage a streak of at least 56 games. Thus the length is quite out of line with what would be expected from him. The second

[5]Not if you are taking one of my classes, I tell you.

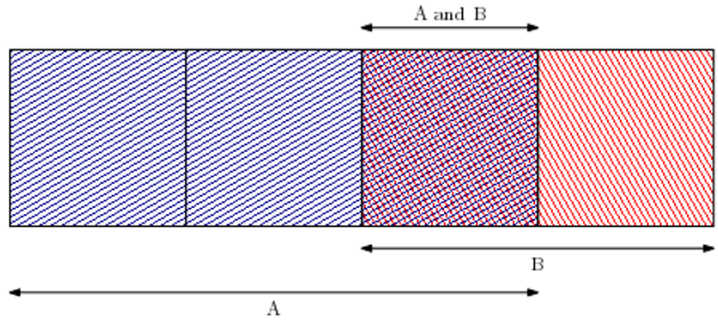

Figure 5.19: Bayes' Theorem in action.

longest streak of any major league hitter is only 44 games, so it is out of line with everyone else as well. But he also once hit in 61 straight games at a lower level of competition, so he seems to have had an extraordinary capacity for consistency.

Hitting streaks can be thought of as runs between games without hits, and so can be modeled using a Poisson distribution. But Monte Carlo simulations provide answers without detailed mathematics. Permutation tests give us insight with minimal knowledge and intellectual effort.

5.6 Bayesian Reasoning

The conditional probability $P(A|B)$ measures the likelihood of event A given knowledge that event B has occurred. We will rely on conditional probability throughout this book, because it lets us update our confidence in an event in response to fresh evidence, like observed data.

Bayes' Theorem is an important tool for working with conditional probabilities, because it lets us turn the conditionals around:

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

With Bayes theorem, we can convert the question of $P(outcome|data)$ to $P(data|outcome)$, which is often much easier to compute. In some sense, Bayes' theorem is just a consequence of algebra, but it leads to a different way of thinking about probability.

Figure 5.19 illustrates Bayes' theorem in action. The event space consists of picking one of four blocks. The complex events A and B represent sub-ranges of blocks, where $P(A) = 3/4$ and $P(B) = 2/4 = 1/2$. By counting blocks from the figure, we can see that $P(A|B) = 1/2$ and $P(B|A) = 1/3$. These also follow

directly from Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{(1/3)\cdot(3/4)}{(1/2)} = 1/2$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{(1/2)\cdot(1/2)}{(3/4)} = 1/3$$

Bayesian reasoning reflects how a prior probability $P(A)$ is updated to give the posterior probability $P(A|B)$ in the face of a new observation B, according to the ratio of the likelihood $P(B|A)$ and the marginal probability $P(B)$. The *prior* probability $P(A)$ reflects our initial assumption about the world, to be revised based on the additional evidence B.

Bayesian reasoning is an important way to look at the world. Walking into Rachel and David's wedding, my prior assumption was that the age distribution would reflect that of the world at large. But my confidence weakened with every elderly cousin I encountered, until it finally crumbled.

We will use Bayesian reasoning to build classifiers in Section 11.1. But keep this philosophy in mind as you analyze data. You should come to each task with a prior conception of what the answers should be, and then revise in accordance with statistical evidence.

5.7 Chapter Notes

Every data scientist should take a good elementary statistics course. Representative texts include Freedman [FPP07] and James et al. [JWHT13]. Wheelan [Whe13] is a gentler introduction, with Huff [Huf10] the classic treatise on how best to lie with statistics.

Donoho [Don15] presents a fascinating history of data science from the vantage point of a statistician. It makes an effective case that most of the major principles of today's data science were originally developed by statisticians, although they were not quickly embraced by the discipline at large. Modern statisticians have begun having much more satisfying conversations with computer scientists on these matters as interests have mutually converged.

Vigen [Vig15] presents an amusing collection of spurious correlations drawn from a large number of interesting time series. Figure 5.10 is representative, and is reprinted with permission.

It has been demonstrated that the size of American families is reasonably well fit by a Poisson distribution. In fact, an analysis of household size distributions from 104 countries suggests that the "I've had enough" model works around the world [JLSI99].

5.8 Exercises

Statistical Distributions

5-1. *[5]* Explain which distribution seems most appropriate for the following phenomenon: binomial, normal, Poisson, or power law?

(a) The number of leaves on a fully grown oak tree.

(b) The age at which people's hair turns grey.

(c) The number of hairs on the heads of 20-year olds.

(d) The number of people who have been hit by lightning x times.

(e) The number of miles driven before your car needs a new transmission.

(f) The number of runs a batter will get, per cricket over.

(g) The number of leopard spots per square foot of leopard skin.

(h) The number of people with exactly x pennies sitting in drawers.

(i) The number of apps on people's cell phones.

(j) The daily attendance in Skiena's data science course.

5-2. *[5]* Explain which distribution seems most appropriate for the following *The Quant Shop* phenomenon: binomial, normal, Poisson, or power law?

(a) The beauty of contestants at the Miss Universe contest.

(b) The gross of movies produced by Hollywood studios.

(c) The birth weight of babies.

(d) The price of art works at auction.

(e) The amount of snow New York will receive on Christmas.

(f) The number of teams that will win x games in a given football season.

(g) The lifespans of famous people.

(h) The daily price of gold over a given year.

5-3. *[5]* Assuming that the relevant distribution is normal, estimate the probability of the following events:

(a) That there will be 70 or more heads in the next hundred flips of a fair coin?

(b) That a randomly selected person will weight over 300 lbs?

5-4. *[3]* The average on a history exam was 85 out of 100 points, with a standard deviation of 15. Was the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

5-5. *[5]* Facebook data shows that 50% of Facebook users have a hundred or more friends. Further, the average user's friend count is 190. What do these findings say about the shape of the distribution of number of friends of Facebook users?

Significance Testing

5-6. *[3]* Which of the following events are likely independent and which are not?

(a) Coin tosses.

(b) Basketball shots.

(c) Party success rates in presidential elections.

5-7. *[5]* The 2010 American Community Survey estimates that 47.1% of women aged 15 years and over are married.

(a) Randomly select three women between these ages. What is the probability that the third woman selected is the only one that is married?

(b) What is the probability that all three women are married?

(c) On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(d) If the proportion of married women was actually 30%, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does decreasing the probability of an event affect the mean and standard deviation of the wait time until success?

Permutation Tests and P-values

5-8. *[5]* Prove that the permutation generation algorithm of page 147 generates permutations correctly, meaning uniformly at random.

5-9. *[5]* Obtain data on the heights of m men and w women.

(a) Use a t-test to establish the significance of whether the men are on average taller than the women.

(b) Perform a permutation test to establish the same thing: whether the men are on average taller than the women.

Implementation Projects

5-10. *[5]* In sporting events, good teams tend to come back and win. But is this because they know how to win, or just because after a long-enough game the better team will generally prevail? Experiment with a random coin flip model, where the better team has a probability of $p > 0.5$ of outplaying the other over a single period.

For games of n periods, how often does the better team win, and how often does it come from behind, for a given probability p? How does this compare to statistics from real sports?

5-11. *[8]* February 2 is Groundhog Day in the United States, when it is said that six more weeks of winter follows if the groundhog sees its shadow. Taking whether it is sunny on February 2 as a proxy for the groundhog's input, is there any predictive power to this tradition? Do a study based on weather records, and report the accuracy of the beast's forecasts along with its statistical significance

Interview Questions

5-12. *[3]* What is conditional probability?

5-13. *[3]* What is Bayes Theorem? And why is it useful in practice?

5-14. *[8]* How would you improve a spam detection algorithm that uses a naive Bayes classifier?

5-15. *[5]* A coin is tossed ten times and the results are two tails and eight heads. How can you tell whether the coin is fair? What is the p-value for this result?

5-16. *[8]* Now suppose that ten coins are each tossed ten times, for a total of 100 tosses. How would you test whether the coins are fair?

5-17. *[8]* An ant is placed on an infinitely long twig. The ant can move one step backward or one step forward with the same probability, during discrete time steps. What is the probability that the ant will return to its starting point after $2n$ steps?

5-18. *[5]* You are about to get on a plane to Seattle. Should you bring an umbrella? You call three random friends of yours who live there and ask each independently whether it is raining. Each of friends has a 2/3 chance of telling you the truth and a 1/3 chance of lying. All three friends tell you that it is raining. What is the probability that it is actually raining in Seattle?

Kaggle Challenges

5-19. Decide whether a car bought at an auction is a bad buy.
`https://www.kaggle.com/c/DontGetKicked`

5-20. Forecast the demand of a product in a given week.
`https://www.kaggle.com/c/grupo-bimbo-inventory-demand`

5-21. How much rain we will get in the next hour?
`https://www.kaggle.com/c/how-much-did-it-rain`