# Chapter 11: Basic Statistical Concepts and Techniques[[1]](#footnote-1)

## I. Introduction

In this chapter, the goal is to equip ourselves to understand, analyze, and criticize arguments using statistics. Such arguments are extremely common; they’re also frequently manipulative and/or fallacious. As Mark Twain once said, “There are three kinds of lies: lies, damned lies, and statistics.” Misused statistics are very convincing lies, because for some reason or another, putting numbers into an inductive argument makes it look more official, and we’re more willing to trust the argument. It is possible, however, with a minimal understanding of some basic statistical concepts and techniques, along with an awareness of the various ways these are commonly misused (intentionally or not), to see the “lies” for what they are: bad arguments that shouldn’t persuade us. We will first provide a foundation of basic statistical knowledge, and then we will look at various statistical fallacies.

## II. Basic Statistical Knowledge

Here we’ll look over averages, testing hypotheses, and statistical sampling. The goal here is not to deep-dive into the math behind these concepts; this is a logic book, not a math book. The goal, rather, is to help non-mathematicians be less deceived when confronted with arguments using statistics.

### A. Averages: Mean, Median, Mode

The word ‘average’ is slippery: it can be used to refer to one of three things: the arithmetic mean, the median, or the mode of a set of values. The mean, median, and mode for the same set are often different, and when this is the case, use of the word ‘average’ is equivocal. A clever person can use this fact to her rhetorical advantage. We hear the word ‘average’ thrown around quite a bit in arguments: the average family has such-and-such an income, the average student carries such-and-such in student loan debt, and so on. It is important to recognize that there is not an actual family which is The Average One—averages are collective properties, not individual ones. That is, they are a property of the group as a whole—all families, all students. There may be a family in the group whose income matches the average—or there may not be one. However, audiences are supposed to take this fictional average entity to be representative of all (or at least most) of the actual members of the group. So, depending on the conclusion she’s trying to convince people of, the person making the argument will choose between mean, median, and mode, picking the number that best serves her rhetorical purpose. It’s important, therefore, for the critical listener to ask, every time the word ‘average’ is used, “What type of average are they talking about? What’s the difference between the three averages for this group? How would using another type of average affect the argument?”

To define our terms: a **mean** is calculated by adding all the values together, and dividing by the total number of values. So, if we’re finding the average age of a group of people, to find the mean, you add all the ages together, and divide by the number of people in the group. A **median** is the exact middle number of a group. To find the median age, we’d line the group members up from youngest to oldest (not physically lining up the people, unless you really want to), and find the age of the person in the middle. If there is an even number of people in the group, and thus no member who is in the exact middle, you find the two members in the middle, add their ages together and divide by 2, to find the age in the middle of them. A **mode** refers to the most common value in a set. So, the modal age would be the most popular age.

A simple example can demonstrate both how to calculate the three different averages, and how an unscrupulous person can use the different results to their advantage.[[2]](#footnote-2) Suppose I run a masonry contracting business on the side—Logical Constructions (a wholly owned subsidiary of LogiCorp). Including myself, 22 people work at Logical Constructions. This is how much they’re paid per year:

$350,000 for me (I’m the boss)

$75,000 each for two foremen

$70,000 for my accountant

$50,000 each for five stone masons

$30,000 for the office secretary

$25,000 each for two apprentices

$20,000 each for ten laborers.

To calculate the mean salary at Logical Constructions, we add up all the individual salaries (my $350,000, $75,000 twice since there are two foremen, $70,000 once, $50,000 five times since there are five stone masons, and so on). This gives us $1,100,000. We next divide by the number of employees—there are 22 employees total. The result is $50,000.

To calculate the median salary, we put all the individual salaries in numerical order (ten entries of $20,000 for the laborers, then two entries of $25,000 for the apprentices, and so on) and find the middle number—or, as is the case with our set, which has an even number of entries, the mean of the middle two numbers. The middle two numbers are both $25,000, so the median salary is $25,000. The median tells me that half of my employees get that much or more, and half of my employees get that much or less.

The mode is often the easiest to calculate; here it is the most common salary. There are more laborers than any other type of employee, so the mode is what each of them get: $20,000.

Now, you may have noticed, a lot of my workers don’t get paid particularly well. Suppose one day, as I’m driving past our construction site (in the back of my limo, naturally), I notice some union organizers commiserating with my laborers during their lunch break. They’re trying to convince my employees to bargain collectively for higher wages. Now we have a debate: should the workers at Logical Constructions be paid more? I take one side of the issue; the workers and organizers take the other. In the course of making our arguments, we might both refer to the “average worker” at Logical Constructions. I’ll want to do so in a way that makes it appear that this mythical worker is doing pretty well, and so we don’t need to change anything; the organizers will want to do so in such a way that makes it appear that the average worker isn’t doing very well at all. If you were the boss, would you present the mean, median, or mode as your average? What if you were the union organizers?

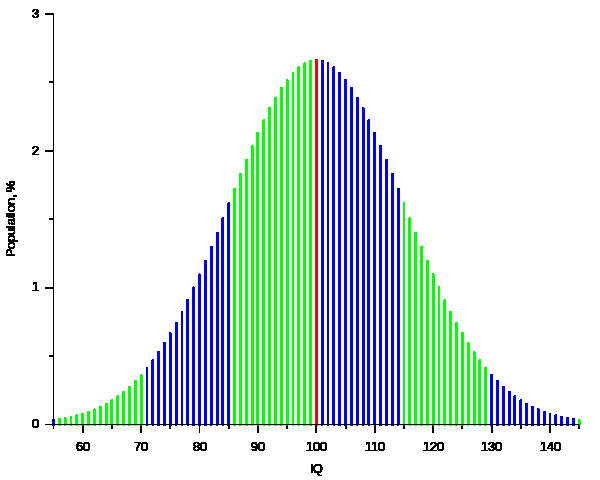
In this case, the mean is higher, so I will use it: “The average worker at Logical Constructions makes $50,000 per year. That’s a pretty good wage!” My opponents, the union organizers, will counter, using either the median, or even better, the mode: “The average worker at Logical Constructions makes a mere $20,000 per year. Try raising a family on such a pittance!”

The mean is very different from the other two averages because my salary is so much higher than everyone else’s. An outlier (someone significantly different from the rest of the group) will pull the mean toward them. A single outlier, however, will not affect the mean very much, and won’t affect the mode at all.

So, a lot hangs on which sense of ‘average’ we pick. This is true in lots of real-life circumstances. For example, household income in the United States is distributed much as salaries are at my fictional Logical Constructions company: those at the top of the range fare much better than those at the bottom.[[3]](#footnote-3) Because the minority at the top have salaries so much higher than everyone else, that pulls the mean up, so it’s higher than the median or the mode. In 2014, the mean household income in the U.S. was $72,641. The median, however, was $53,657. That’s a big difference! “The average family makes about $72,000 per year” sounds a lot better than “The average family makes about $53,000 per year.”

### B. Normal Distributions: Standard Deviation, Confidence Intervals

If you gave IQ tests to a whole bunch of people, and then graphed the results on a histogram or bar chart—so that every time you saw a particular score, the bar for that score would get higher— you’d end up with a picture like this:



This kind of distribution is called a “normal” or “Gaussian” distribution[[4]](#footnote-4); because of its shape, it’s also often called a “bell curve.” Many phenomena in nature are (approximately) distributed along a bell curve: height, blood pressure, motions of individual molecules in a collection, lifespans of industrial products, measurement errors, and so on. And even when traits are not normally distributed, it can be useful to treat them as if they were. This is because the bell curve provides an extremely convenient starting point for making certain inferences. Because the curve is symmetrical, the mean is the same as the median. Because the highest point is in the middle, the mode is also the same as the median. A bell curve is thus convenient because one can know everything about such a curve by specifying two of its features: its mean and its standard deviation.

We already understand the mean. Let’s get a grip on standard deviation. We don’t need to learn how to calculate it (though that can be done); we just want a qualitative (as opposed to quantitative) understanding of what it signifies. Roughly, it’s a measure of the spread of the data represented on the curve; it’s a way of indicating how far, on average, values tend to stray from the mean. An example can make this clear. Consider two cities: Milwaukee, Wisconsin, and San Diego, California. These two cities are different in a variety of ways, not least in the kind of weather their residents experience. Setting aside precipitation, let’s focus just on temperature. If you recorded the high temperatures every day in each town over a long period of time and made a histogram for each (with temperatures on the x-axis, number of days on the y-axis), you’d get two very different-looking curves. Maybe something like these:

Two histogram charts, unlabelled. On the left, the curve is low, wide, and rises and falls gently. It is labelled Milwaukee. On the right, a curve labelled San Diego is narrow, high, and rises and falls sharply.

The average high temperatures for the two cities—the peaks of the curves—would of course be different: San Diego is warmer on average than Milwaukee. But the range of temperatures experienced in Milwaukee is much greater than that in San Diego: some days in Milwaukee, the high temperature is below zero, while on some days in the summer it’s over 100°F. San Diego, on the other hand, is basically always perfect: right around 70° or so.[[5]](#footnote-5) While both of them have normal distributions, the standard deviation of temperatures in Milwaukee is much greater than in San Diego. This is reflected in the shapes of the respective bell curves. Milwaukee’s is shorter and wider, with a non-trivial number of days at the temperature extremes and a wide spread for all the other days. San Diego’s is taller and narrower, with temperatures hovering in a tight range all year, and hence more days at each temperature recorded (which explains the relative heights of the curves).

When you encounter a standard deviation, the take-away is that a small number means there is very little variety in the group, like San Diego temperatures. A large standard deviation means a lot of variety in the group, like Milwaukee.

If we’re dealing with a normal distribution, once we know the mean and standard deviation, we know everything we need to know about it. There are three very useful facts about these curves that can be stated in terms of the mean and standard deviation (SD). As a matter of mathematical fact, 68.3% of the population depicted on the curve (whether they’re people with certain IQs, days on which certain temperatures were reached, measurements with a certain amount of error) falls within a range of one standard deviation on either side of the mean. So, for example, the mean IQ is 100; the standard deviation is 15. It follows that 68.3% of people have an IQ between 85 and 115—15 points (one SD) on either side of 100 (the mean). Another fact: 95.4% of the population depicted on a bell curve will fall within a range two standard deviations from the mean. So, 95.4% of people have an IQ between 70 and 130—30 points (2 SDs) on either side of 100. Finally, 99.7% of the population falls within three standard deviations of the mean; 99.7% of people have IQs between 55 and 145. These ranges are called confidence intervals.[[6]](#footnote-6) They are convenient reference points commonly used in statistical inference.[[7]](#footnote-7)

This only works if we’re dealing with a normal distribution pattern, though. Yearly income in the U.S., or the yearly salary of employees at my fictional company, do not follow a normal distribution. In these cases, it’s often much more helpful to know the median and mode than it is to know the mean. The median gives us a number and tells us that half of the population is at or above that number, and half of the population is at or below it. The mode tells us the most common number in the group. Looking at all three types of average, in addition to the standard deviation, gives us a much broader picture of the population as a whole.

### C. Statistical Inference: Hypothesis Testing

If we start with knowledge of the properties of a given normal distribution, we can test claims about the world to which that information is relevant. Starting with a bell curve—information of a general nature—we can draw conclusions about particular hypotheses. These are conclusions of inductive arguments; they are not certain, but more or less probable. When we use knowledge of normal distributions to draw them, we can be precise about how probable they are. This is inductive logic.

The basic pattern of the kinds of inferences we’re talking about is this: one formulates a hypothesis, then runs an experiment to test it; the test involves comparing the results of that experiment to what is known (some normal distribution); depending on how well the results of the experiment comport with what would be expected given the background knowledge represented by the bell curve, we draw a conclusion about whether or not the hypothesis is true.

Though they are applicable in a very wide range of contexts, it’s perhaps easiest to explain the patterns of reasoning we’re going to examine using examples from medicine. These kinds of cases are vivid; they aid in understanding by making the consequences of potential errors more real. Also, in these cases the hypotheses being tested are relatively simple: claims about individuals’ health—whether they’re healthy or sick, whether they have some condition or don’t—as opposed to hypotheses dealing with larger populations and measurements of their properties. Examining these simpler cases will allow us to see more clearly the underlying patterns of reasoning that cover all such instances of hypothesis testing, and to gain familiarity with the vocabulary statisticians use in their work.

The knowledge we start with is how some trait relevant to the particular condition is distributed in the population generally—a bell curve.[[8]](#footnote-8) The experiment we run is to measure the relevant trait in the individual whose health we’re assessing. The result of a comparison with the result of this measurement and the known distribution of the trait tells us something about whether or not the person is healthy. Suppose we start with information about how a trait is distributed among people who are healthy. Hematocrit, for example, is a measure of how much of a person’s blood is taken up by red blood cells, expressed as a percentage of total blood volume. Lower hematocrit levels are associated with anemia; higher levels are associated with dehydration, certain kinds of tumors, and other disorders. Among healthy men, the mean hematocrit level is 47%, with a standard deviation of 3.5%. We can draw the curve, noting the boundaries of the confidence intervals:

The chart is labelled "Hematocrit Levels, Healthy Men." It shows a bell curve, highest in the middle, with gentle slopes down both sides. The numbers across the bottom are: 36.5, 40, 43.5, 47, 50.5, 54, 57.5. 47 is the exact middle, and where the graph is the highest.

Because of the fixed mathematical properties of the bell curve, we know that 68.3% of healthy men have hematocrit levels between 43.5% and 50.5%; 95.4% of them are between 40% and 54%; and 99.7% of them are between 36.5% and 57.5%.

Let’s consider a man whose health we’re interested in evaluating. Call him Larry. We take a sample of Larry’s blood and measure the hematocrit level. We compare it to the values on the curve to see if there might be some reason to be concerned about Larry’s health. Remember, the curve tells us the levels of hematocrit for healthy men; we want to know if Larry’s one of them. The hypothesis we’re testing is that Larry’s healthy. Statisticians often refer to the hypothesis under examination in such tests as the “null hypothesis”—a default assumption, something we’re inclined to believe unless we discover evidence against it. Anyway, we’re measuring Larry’s hematocrit; what kind of result should he be hoping for? Clearly, he’d like to be as close to the middle, fat part of the curve as possible; that’s where most of the healthy people are. The further away from the average healthy person’s level of hematocrit he strays, the more he’s worried about his health. That’s how these tests work: if the result of the experiment (measuring Larry’s hematocrit) is sufficiently close to the mean, we have no reason to reject the null hypothesis (that Larry’s healthy); if the result is far away, we do have reason to reject it.

How far away from the mean is too far away? It depends. A typical cutoff is two standard deviations from the mean—the 95.4% confidence interval.[[9]](#footnote-9) That is, if Larry’s hematocrit level is below 40% or above 54%, then we might say we have reason to doubt the null hypothesis that Larry is healthy. The language statisticians use for such a result—say, for example, if Larry’s hematocrit came in at 38%—is to say that it’s “statistically significant.” In addition, they specify the level at which it’s significant—an indication of the confidence-interval cutoff that was used. In this case, we’d say Larry’s result of 38% is statistically significant at the .05 level. (95% = .95; 1 - .95 = .05) Either Larry is unhealthy (anemia, most likely), or he’s among the (approximately) 5% of healthy people who fall outside of the two standard-deviation range. If he came in at a level even further from the mean—say, 36%—we would say that this result is significant at the .003 level (99.7% = .997; 1 - .997 = .003). That would give us all the more reason to doubt that Larry is healthy.

So, when we’re designing a medical test like this, the crucial decision to make is where to set the cutoff. Again, typically that’s the 95% confidence interval. If a result falls outside that range, the person tests “positive” for whatever condition we’re on the lookout for. (Of course, a “positive” result is hardly positive news—in the sense of being something you want to hear). But these sorts of results are not conclusive: it may be that the null hypothesis (this person is healthy) is true, and that they’re simply one of the relative rare 5% who fall on the outskirts of the curve. In such a case, we would say that the test has given the person a “false positive” result: the test indicates sickness when in fact there is none. Statisticians refer to this kind of mistake as “type I error.” We could reduce the number of mistaken results our test gives by changing the confidence levels at which we give a positive result. Returning to the concrete example above: suppose Larry has a hematocrit level of 38%, but that he is not in fact anemic; since 38% is outside of the two standard-deviation range, our test would give Larry a false positive result if we used the 95% confidence level. However, if we raised the threshold of statistical significance to the three standard-deviation level of 99.7%, Larry would not get flagged for anemia; there would be no false positive, no type I error.

So, we should always use the wider range on these kinds of tests to avoid false positives, right? Not so fast. There’s another kind of mistake we can make: false negatives, also called “type II errors.” Increasing our range increases our risk of this second kind of foul-up. Down there at the skinny end of the curve there are relatively few healthy people. Sick people are the ones who generally have measurements in that range; they’re the ones we’re trying to catch. When we issue a false negative, we’re missing them. A false negative occurs when the test tells you there’s no reason to doubt the null hypothesis (that you’re healthy), when as a matter of fact you are sick. If we increase our range from two to three standard deviations—from the 95% level to the 99.7% level—we will avoid giving a false positive result to Larry, who is healthy despite his low 38% hematocrit level. But we will end up giving false reassurance to some anemic people who have levels similar to Larry’s; someone who has a level of 38% and is sick will get a false negative result if we only flag those outside the 99.7% confidence interval (36.5% - 57.5%).

This is a perennial dilemma in medical screening: how best to strike a balance between the two types of errors—between needlessly alarming healthy people with false positive results and failing to detect sickness in people with false negative results. The terms clinicians use to characterize how well diagnostic tests perform along these two dimensions are “sensitivity” and “specificity.” A highly sensitive test will catch a large number of cases of sickness—it has a high rate of true positive results. Of course, this comes at the cost of increasing the number of false positive results as well. A test with a high level of specificity will have a high rate of true negative results— correctly identifying healthy people as such. The cost of increased specificity, though, is an increase in the number of false negative results—sick people that the test misses. Since every false positive is a missed opportunity for a true negative, increasing sensitivity comes at the cost of decreasing specificity. And since every false negative is a missed true positive, increasing specificity comes at the cost of decreasing sensitivity. A final bit of medical jargon: a screening test is accurate to the degree that it is both sensitive and specific.

Given sufficiently thorough information about the distributions of traits among healthy and sick populations, clinicians can rig their diagnostic tests to be as sensitive or specific as they like. But since those two properties pull in opposite directions, there are limits to degree of accuracy that is possible. And depending on the particular case, it may be desirable to sacrifice specificity for more sensitivity, or vice versa.

To see how a screening test might be rigged to maximize sensitivity, let’s consider an abstract hypothetical example. Suppose we knew the distribution of a certain trait among the population of people suffering from a certain disease. (Contrast this with our starting point above: knowledge of the distribution among healthy individuals). This kind of knowledge is common in medical contexts: various so-called biomarkers—gene mutations, proteins in the blood, etc.—are known to be indicative of certain conditions; often, one can know how such markers are distributed among people with the condition. Again, keeping it abstract and hypothetical, suppose we know that among people who suffer from Disease X, the mean level of a certain biomarker β for the disease is 20, with a standard deviation of 3. We can sum up this knowledge with a curve:

The title reads "Beta levels, people with disease x". The graph is a standard bell curve, with the highest point being the center, and a gentle slope down both sides. The numbers on the chart are 11, 14, 17, 20, 23, 26, 29, with 20 being the middle, where the highest point on the graph hits.

Now, suppose Disease X is very serious indeed. It would be a benefit to public health if we were able to devise a screening test that could catch as many cases as possible—a test with a high sensitivity. Given the knowledge we have about the distribution of β among patients with the disease, we can make our test as sensitive as we like. We know, as a matter of mathematical fact, that 68.3% percent of people with the disease have β-levels between 17 and 23; 95.4% of people with the disease have levels between 14 and 26; 99.7% have levels between 11 and 29. Given these facts, we can devise a test that will catch 99.7% of cases of Disease X like so: measure the level of biomarker β in people, and if they have a value between 11 and 29, they get a positive test result; a positive result is indicative of disease. This will catch 99.7% of cases of the condition, because the range chosen is three standard deviations on either side of the mean, and that range contains 99.7% of unhealthy people; if we flag everybody in that range, we will catch 99.7% of cases. Of course, we’ll probably end up catching a whole lot of healthy people as well if we cast our net this wide; we’ll get a lot of false positives. We could correct for this by making our test less sensitive, say by lowering the threshold for a positive test to the two standard-deviation range of 14 – 26. We would now only catch 95.4% of cases of sickness, but we would reduce the number of healthy people given false positives; instead, they would get true negative results, increasing the specificity of our test.

Notice that the way we used the bell curve in our hypothetical test for Disease X was different from the way we used the bell curve in our test of hematocrit levels above. In that case, we flagged people as potentially sick when they fell outside of a range around the mean; in the new case, we flagged people as potentially sick when they fell inside a certain range. This difference corresponds to the differences in the two populations the respective distributions represent: in the case of hematocrit, we started with a curve depicting the distribution of a trait among healthy people; in the second case, we started with a curve telling us about sick people. In the former case, sick people will tend to be far from the mean; in the latter, they’ll tend to cluster closer.

### D. Statistical Inference: Sampling

When we were testing hypotheses, our starting point was knowledge about how traits were distributed among a large population—e.g., hematocrit levels among healthy men. We now ask a pressing question: how do we acquire such knowledge? How do we figure out how things stand with a very large population? The difficulty is that it’s usually impossible to check every member of the population. Instead, we have to make an inference. This inference involves sampling: instead of testing every member of the population, we test a small portion of the population—a sample— and infer from its properties to the properties of the whole. Reasoning from part of a group to a group as a whole is a *generalization*. Reasoning from statistical samples is a simple inductive argument:

The sample has property X.

Therefore, the general population has property X.

The argument is clearly inductive: the premise does not guarantee the truth of the conclusion; it merely makes it more probable. As was the case in hypothesis testing, we can be precise about the probabilities involved, and our probabilities come from the good-old bell curve.

Let’s take a simple example.[[10]](#footnote-10) Suppose we were trying to discover the percentage of men in the general population; we survey 100 people, and it turns out there are 55 men in our sample. So, the proportion of men in our sample is .55 (or 55%). We’re trying to make an inference from this premise to a conclusion about the proportion of men in the general population. What’s the probability that the proportion of men in the general population is .55? This isn’t exactly the question we want to answer in these sorts of cases, though. Rather, we ask, what’s the probability that the true proportion of men in the general population is in some range on either side of .55? We can give a precise answer to this question, and the answer depends on the size of the range you’re considering in a familiar way.

Given that our sample’s proportion of men is .55, it is relatively more likely that the true proportion in the general population is close to that number, less likely that it’s far away. For example, it’s more likely, given the result of our survey, that in fact 50% of the population is men than it is that only 45% are men. And it’s still less likely that only 40% are men. The same pattern holds in the opposite direction: it’s more likely that the true percentage of men is 60% than 65%. Generally speaking, the further away from our survey results we go, the less probable it is that we have the true value for the general population. The drop off in probabilities described takes the form of a bell curve:

The label says "proportion of men in the population." There's a label on the left that says "probability" with an arrow pointing up. The graph is a normal bell curve. The numbers on the graph at the bottom are .40, .45, .50, .55, .60, .65, .70, with .55 at the center where the peak of the curve is.

The standard deviation of .05 is a function of our sample size of 100. We can use the usual confidence intervals—again, with 2 standard deviations, 95.4% being standard practice—to interpret the findings of our survey: we’re pretty sure—to the tune of 95%—that the general population is between 45% and 65% male.

That’s a pretty wide range. Our result is not that impressive (especially considering the fact that we know the actual number is very close to 50%). But that’s the best we can do given the limitations of our survey. The main limitation, of course, was the size of our sample: 100 people just isn’t very many. We could narrow the range within which we’re 95% confident if we increased our sample size; doing so would likely (though not certainly) give us a proportion in our sample closer to the true value of (approximately) .5. The relationship between the sample size and the width of the confidence intervals is a purely mathematical one. As sample size goes up, standard deviation goes down—the curve narrows.

The pattern of reasoning on display in our toy example is the same as that used in sampling generally. Perhaps the most familiar instances of sampling in everyday life are public opinion surveys. Rather than trying to determine the proportion of people in the general population who are men (not a real mystery), opinion pollsters try to determine the proportion of a given population who, say, intend to vote for a certain candidate, or approve of the job the president is doing, or believe in Bigfoot. Pollsters survey a sample of people on the question at hand, and end up with a result: 29% of Americans believe in Bigfoot, for example.[[11]](#footnote-11) But the headline number, as we have seen, doesn’t tell the whole story. 29% of the sample (in this case, about 1,000 Americans) reported believing in Bigfoot; it doesn’t follow with certainty that 29% of the general population (all Americans) have that belief. Rather, the pollsters have some degree of confidence (again, 95% is standard) that the actual percentage of Americans who believe in Bigfoot is in some range around 29%. You may have heard the “margin of error” mentioned in connection with such surveys. This phrase refers to the very range we’re talking about. In the survey about Bigfoot, the margin of error is 3%. That’s the distance from the mean (the 29% found in the sample) and the ends of the two standard-deviation confidence interval—the range in which we’re 95% sure the true value lies. Again, this range is just a mathematical function of the sample size: if the sample size is around 100, the margin of error is about 10% (see the toy example above: 2 SDs = .10); if the sample size is around 400, you get that down to 5%; at 600, you’re down to 4%; at around 1,000, 3%; to get down to 2%, you need around 2,500 in the sample, and to get down to 1%, you need 10,000.[[12]](#footnote-12) So the real upshot of the Bigfoot survey result is something like this: somewhere between 26% and 32% of Americans believe in Bigfoot, and we’re 95% sure that’s the correct range; or, to put it another way, we used a method for determining the true proportion of Americans who believe in Bigfoot that can be expected to determine a range in which the true value actually falls 95% of the time, and the range that resulted from our application of the method on this occasion was 26% - 32%.

That last sentence, we must admit, would make for a pretty lousy newspaper headline (“29% of Americans believe in Bigfoot!” is much sexier), but it’s the most honest presentation of what the results of this kind of sampling exercise actually show. Sampling gives us a range, which will be wider or narrower depending on the size of the sample, and not even a guarantee that the actual value is within that range. That’s the best we can do; these are inductive, not deductive, arguments.

Finally, on the topic of sampling, we should acknowledge than in actual practice, polling is hard. The mathematical relationships between sample size and margin of error/confidence that we’ve noted all hold in the abstract, but real-life polls can have errors that go beyond these theoretical limitations on their accuracy. As the 2016 U.S. presidential election—and the so-called “Brexit” vote in the United Kingdom that same year, and many, many other examples throughout the history of public opinion polling—showed us, polls can be systematically in error. The kinds of facts we’ve been stating—that with a sample size of 600, a poll has a margin of error of 4% at the 95% confidence level—hold only on the assumption that there’s a systematic relationship between the sample and the general population it’s meant to represent; namely, that the sample is representative.

A representative sample mirrors the general population; in the case of people, this means that the sample and the general population have the same demographic make-up—same percentage of old people and young people, white people and people of color, rich people and poor people, etc., etc. Polls whose samples are not representative are likely to misrepresent the feature of the population they’re trying to capture.

Suppose I wanted to find out what percentage of the U.S. population thinks favorably of Donald Trump. If I asked 1,000 people in, say, rural Oklahoma, I’d get one result; if I asked 1,000 people in midtown Manhattan, I’d get a much different result. Neither of those two samples is representative of the population of the United States as a whole. To get such a sample, I’d have to be much more careful about whom I surveyed. A famous example from the history of public polling illustrates the difficulties here rather starkly: in the 1936 U.S. presidential election, the contenders were Republican Alf Landon of Kansas, and the incumbent President Franklin D. Roosevelt. A (now-defunct) magazine, Literary Digest conducted a poll with 2.4 million (!) participants, and predicted that Landon would win in a landslide. Instead, he lost in a landslide; FDR won the second of his four presidential elections. What went wrong? With a sample size so large, the margin of error would be tiny. The problem was that their sample was not representative of the American population. They chose participants randomly from three sources: (a) their list of subscribers; (b) car registration forms; and (c) telephone listings. The problem with this selection procedure is that all three groups tended to be wealthier than average. This was 1936, during the depths of the Great Depression. Most people didn’t have enough disposable income to subscribe to magazines, let alone have telephones or own cars. The survey therefore over-sampled Republican voters and got a skewed result. Even a large and seemingly random sample can lead one astray. This is what makes polling so difficult: finding representative samples is hard.[[13]](#footnote-13)

Other practical difficulties with polling are worth noting. First, the way your polling question is worded can make a big difference in the results you get. The framing of an issue—the words used to specify a particular policy or position—can have a dramatic effect on how a relatively uninformed person will feel about it. If you wanted to know the American public’s opinion on whether or not it’s a good idea to tax the transfer of wealth to the heirs of people whose holdings are more than $5.5 million or so, you’d get one set of responses if you referred to the policy as an “estate tax,” a different set of responses if you referred to it as an “inheritance tax,” and a still different set if you called it the “death tax.” A poll of Tennessee residents found that 85% opposed “Obamacare,” while only 16% opposed “Insure Tennessee” (they’re the same thing, of course).[[14]](#footnote-14)

Even slight changes in the wording of questions can alter the results of an opinion poll. This is why the polling firm Gallup hasn’t changed the wording of its presidential-approval question since the 1930s. They always ask: “Do you approve or disapprove of the way [name of president] is handling his job as President?” A deviation from this standard wording can produce different results. The polling firm Ipsos found that its polls were more favorable than others’ for the president. They traced the discrepancy to the different way they worded their question, giving an additional option: “Do you approve, disapprove, or have mixed feelings about the way Barack Obama is handling his job as president?”[[15]](#footnote-15)

Another difficulty with polling is that some questions are harder to get reliable data about than others, simply because they involve topics about which people tend to be untruthful. Asking someone whether they approve of the job the president is doing is one thing; asking them whether or not they’ve ever cheated on their taxes, say, is quite another. They’re probably not shy about sharing their opinion on the former question; they’ll be much more reluctant to be truthful on the latter (assuming they’ve ever fudged things on tax returns). There are lots of things it would be difficult to discover for this reason: how often people floss, how much alcohol they drink, whether or not they exercise, their sexual habits, and so on. Sometimes this reluctance to share the truth about oneself is quite consequential: some experts think that the reason polls failed to predict the election of Donald Trump as president of the United States in 2016 was that some of his supporters were “shy”—unwilling to admit that they supported the controversial candidate.[[16]](#footnote-16) They had no such qualms in the voting booth, however.

Finally, who’s asking the question—and the context in which it’s asked—can make a big difference. People may be more willing to answer questions in the relative anonymity of an online poll, slightly less willing in the somewhat more personal context of a telephone call, and still less forthcoming in a face-to-face interview. Pollsters use all of these methods to gather data, and the results vary accordingly. Of course, these factors become especially relevant when the question being polled is a sensitive one, or something about which people tend not to be honest or forthcoming. To take an example: the best way to discover how often people truly floss is probably with an anonymous online poll. People would probably be more likely to lie about that over the phone, and still more likely to do so in a face-to-face conversation. The absolute worst source of data on that question, perversely, would probably be from the people who most frequently ask it: dentists and dental hygienists. Every time you go in for a cleaning, they ask you how often you brush and floss; and if you’re like most people, you lie, exaggerating the assiduity with which you attend to your dental-health maintenance (“I brush after every meal and floss twice a day, honest.”).

As was the case with hypothesis testing, the logic of statistical sampling is relatively clear. Things get murky, again, when straightforward abstract methods confront the confounding factors involved in real-life application.

## VI. How to Lie with Statistics[[17]](#footnote-17)

The basic grounding in fundamental statistical concepts and techniques provided in the last section gives us the ability to understand and analyze statistical arguments. Since real-life examples of such arguments are so often manipulative and misleading, our aim in this section is to build on the foundation of the last by examining some of the most common statistical fallacies—the bad arguments and deceptive techniques used to try to bamboozle us with numbers.

### 1. Impressive Numbers without Context

I’m considering buying a new brand of shampoo. The one I’m looking at promises “85% more body.” That sounds great to me (I’m pretty bald; I can use all the extra body I can get). But before I make my purchase, maybe I should consider the fact that the shampoo bottle doesn’t answer this simple follow-up question: 85% more body than what? The bottle does mention that the formulation inside is “new and improved.” So maybe it’s 85% more body than the unimproved shampoo? Or possibly they mean that their shampoo gives hair 85% more body than their competitors’. Which competitor, though? The one that does the best at giving hair more body? The one that does the worst? The average of all the competing brands? Or maybe it’s 85% more body than something else entirely. I once had a high school teacher who advised me to massage my scalp for 10 minutes every day to prevent baldness (I didn’t take the suggestion; maybe I should have). Perhaps this shampoo produces 85% more body than daily 10-minute massages. Or maybe it’s 85% more body than never washing your hair at all. And just what is “body” anyway? How is it quantified and measured? Did they take high-precision calipers and systematically gauge the widths of hairs? Or is it more a function of coverage—hairs per square inch of scalp surface area?

The sad fact is, answers to these questions are not forthcoming. The claim that the shampoo will give my hair 85% more body sounds impressive, but without some additional information for me to contextualize that claim, I have no idea what it means. This is a classic rhetorical technique: throw out a large number to impress your audience, without providing the context necessary for them to evaluate whether or not your claim is actually all that impressive. Usually, on closer examination, it isn’t. Advertisers and politicians use this technique all the time.

In the spring of 2009, the economy was in really bad shape (the fallout from the financial crisis that began in the fall of the year before was still being felt; stock market indices didn’t hit their bottom until March 2009, and the unemployment rate was still on the rise). Barack Obama, the newly inaugurated president at the time, wanted to send the message to the American people that he got it: households were cutting back on their spending because of the recession, and so the government would do the same thing. After his first meeting with his cabinet (the Secretaries of Defense, State, Energy, etc.), he held a press conference in which he announced that he had ordered each of them to cut $100 million from their agencies’ budgets. He had a great line to go with the announcement: “$100 million there, $100 million here—pretty soon, even here in Washington, it adds up to real money.” Funny. And impressive-sounding. $100 million is a hell of a lot of money! At least, it’s a hell of a lot of money to me. I’ve got—give me a second while I check—$64 in my wallet right now. I wish I had $100 million. But of course my personal finances are the wrong context in which to evaluate the president’s announcement. He’s talking about cutting from the federal budget; that’s the context. How big is that? In 2009, it was a little more the $3 trillion. There are fifteen departments that the members of the cabinet oversee. The cut Obama ordered amounted to $1.5 billion, then. That’s .05% of the federal budget. That number’s not sounding as impressive now that we put it in the proper context.

2009 provides another example of this technique. Opponents of the Affordable Care Act (“Obamacare”) complained about the length of the bill: they repeated over and over that it was 1,000 pages long. That complaint dovetailed nicely with their characterization of the law as a boondoggle and a government takeover of the healthcare system. 1,000 pages sure sounds like a lot of pages. That’s up there with notoriously long books like *War and Peace*, *Les Miserables*, and *Infinite Jest*. It’s long for a book, but is it a lot of pages for a piece of federal legislation? Well, it’s big, but certainly not unprecedented. That year’s stimulus bill was about the same length. President Bush’s 2007 budget bill was just shy of 1,500 pages. His No Child Left Behind bill clocks in at just shy of 700. The fact is, major pieces of legislation have a lot of pages. The Affordable Care Act was not especially unusual.

### 2. Misunderstanding Error

As we discussed, built in to the logic of sampling is a margin of error. It is true of measurement generally that random error is unavoidable: whether you’re measuring length, weight, velocity, or whatever, there are inherent limits to the precision and accuracy with which our instruments can measure things. Measurement errors are built into the logic of scientific practice generally; they must be accounted for. Failure to do so—or intentionally ignoring error—can produce misleading reports of findings.

This is particularly clear in the case of public opinion surveys. As we saw, the results of such polls are not the precise percentages that are often reported, but rather ranges of possible percentages (with those ranges only being reliable at the 95% confidence level, typically). And so to report the results of a survey, for example, as “29% of Americans believe in Bigfoot,” is a bit misleading since it leaves out the margin of error and the confidence level. A worse sin is committed (quite commonly) when comparisons between percentages are made and the margin of error is omitted. This is typical in politics, when the levels of support for two contenders for an office are being measured. A typical newspaper headline might report something like this: “Smith Surges into the Lead over Jones in Latest Poll, 44% to 43%.” This is a sexy headline: it’s likely to sell papers (or, nowadays, generate clicks), both to (happy) Smith supporters and (alarmed) Jones supporters. But it’s misleading: it suggests a level of precision, a definitive result, that the data simply do not support. Let’s suppose that the margin of error for this hypothetical poll was 3%. What the survey results actually tell us, then, is that (at the 95% confidence level) the true level of support for Smith in the general population is somewhere between 41% and 47%, while the true level of support for Jones is somewhere between 40% and 46%. Those data are consistent with a Smith lead, to be sure; but they also allow for a commanding 46% to 41% lead for Jones. The best we can say is that it’s slightly more likely that Smith’s true level of support is higher than Jones’s (at least, we’re pretty sure; 95% confidence interval and all). When differences are smaller than the margin of error (really, twice the margin of error when comparing two numbers), they just don’t mean very much. That’s a fact that headline-writers typically ignore. This gives readers a misleading impression about the certainty with which the state of the election can be known.

Early in their training, scientists learn that they cannot report values that are smaller than the error attached to their measurements. If you weigh some substance, say, and then run an experiment in which it’s converted into a gas, you can plug your numbers into the ideal gas law and punch them into your calculator, but you’re not allowed to report all the numbers that show up after the decimal place. The number of so-called “significant digits” (or sometimes “figures”) you can use is constrained by the size of the error in your measurements. If you can only know the original weight to within .001 grams, for example, then even though the calculator spits out .4237645, you can only report a result using three significant digits—.424 after rounding.

The more significant digits you report, the more precise you imply your measurement is. This can have the rhetorical effect of making your audience easier to persuade. Precise numbers are impressive; they give people the impression that you really know what you’re talking about, that you’ve done some serious quantitative analytical work. Suppose I ask 1,000 college students how much sleep they got last night. I add up all the numbers and divide by 1,000, and my calculator gives me 7.037 hours. If I went around telling people that I’d done a study that showed that the average college student gets 7.037 hours of sleep per night, they’d be pretty impressed: my research methods were so thorough that I can report sleep times down to the thousandths of an hour. They’ve probably got a mental picture of my laboratory, with elaborate equipment hooked up to college students in beds, measuring things like rapid eye movement and breathing patterns to determine the precise instants at which sleep begins and ends. But I have no such laboratory. I just asked a bunch of people. Ask yourself: how much sleep did you get last night? I got about 9 hours (it’s the weekend). The key word in that sentence is ‘about.’ Could it have been a little bit more or less than 9 hours? Could it have been 9 hours and 15 minutes? 8 hours and 45 minutes? Sure. The error on any person’s report of how much they slept last night is bound to be something like a quarter of an hour. That means that I’m not entitled to those 37 thousandths of an hour that I reported from my little survey. The best I can do is say that the average college student gets about 7 hours of sleep per night, plus or minus 15 minutes or so. 7.037 is precise, but the precision of that figure is spurious (not genuine, false).

Ignoring the error attached to measurements can have profound real-life effects. Consider the 2000 U.S. presidential election. George W. Bush defeated Al Gore that year, and it all came down to the state of Florida, where the final margin of victory (after recounts were started, then stopped, then started again, then finally stopped by order of the Supreme Court of the United States) was 327 votes. There were about 6 million votes cast in Florida that year. The margin of 327 is about .005% of the total. Here’s the thing: counting votes is a measurement like any other; there is an error attached to it. You may remember that in many Florida counties, they were using punch-card ballots, where voters indicate their preference by punching a hole through a perforated circle in the paper next to their candidate’s name. Sometimes, the circular piece of paper—a so-called “chad”—doesn’t get completely detached from the ballot, and when that ballot gets run through the vote-counting machine, the chad ends up covering the hole and a non-vote is mistakenly registered. Other types of vote-counting methods—even hand-counting—have their own error. And whatever method is used, the error is going to be greater than the .005% margin that decided the election. As one prominent mathematician put it, “We’re measuring bacteria with a yardstick.”[[18]](#footnote-18) That is, the instrument we’re using (counting, by machine or by hand) is too crude to measure the size of the thing we’re interested in (the difference between Bush and Gore). He suggested they flip a coin to decide Florida. It’s simply impossible to know who won that election.

In 2011, newly elected Wisconsin Governor Scott Walker, along with his allies in the state legislature, passed a budget bill that had the effect, among other things, of cutting the pay of public sector employees by a pretty significant amount. There was a lot of uproar. People who were against the bill made their case in various ways. One of the lines of attack was economic: depriving so many Wisconsin residents of so much money would damage the state’s economy and cause job losses (state workers would spend less, which would hurt local businesses’ bottom lines, which would cause them to lay off their employees). One newspaper story at the time quoted a professor of economics who claimed that the Governor’s bill would cost the state 21,843 jobs.[[19]](#footnote-19) Not 21, 844 jobs; it’s not that bad. Only 21,843. This number sounds impressive; it’s very precise. But of course that precision is spurious. Estimating the economic effects of public policy is an extremely uncertain business. I don’t know what kind of model this economist was using to make his estimate, but whatever it was, it’s impossible for its results to be reliable enough to report that many significant digits. My guess is that at best the 2 in 21,843 has any meaning at all.

### 3. Tricky Percentages

Statistical arguments are full of percentages, and there are lots of ways you can fool people with them. The key to not being fooled by such figures, usually, is to keep in mind what it’s a percentage of. Inappropriate, shifting, or strategically chosen numbers can give you misleading percentages.

When the numbers are very small, using percentages instead of fractions is misleading. Johns Hopkins Medical School, when it opened in 1893, was one of the few medical schools that allowed women to enroll.[[20]](#footnote-20) In those benighted times, people worried about women enrolling in schools with men for a variety of silly reasons. One of them was the fear that the impressionable young ladies would fall in love with their professors and marry them. Absurd, right? Well, maybe not: in the first class to enroll at the school, 33% of the women did indeed marry their professors! The sexists were apparently right. That figure sounds impressive, until you learn that the denominator is 3. Three women enrolled at Johns Hopkins that first year, and one of them married her anatomy professor. Using the percentage rather than the fraction exaggerates in a misleading way. Another made up example: I live in a relatively safe little town. If I saw a headline in my local newspaper that said “Armed Robberies are Up 100% over Last Year” I would be quite alarmed. That is, until I realized that last year there was one armed robbery in town, and this year there were two. That is a 100% increase, but using the percentage of such a small number is misleading.

You can fool people by changing the number you’re taking a percentage of mid-stream. Suppose you’re an employee at my aforementioned LogiCorp. You evaluate arguments for $10.00 per hour. One day, I call all my employees together for a meeting. The economy has taken a turn for the worse, I announce, and we’ve got fewer arguments coming in for evaluation; business is slowing. I don’t want to lay anybody off, though, so I suggest that we all share the pain: I’ll cut everybody’s pay by 20%; but when the economy picks back up, I’ll make it up to you. So you agree to go along with this plan, and you suffer through a year of making a mere $8.00 per hour evaluating arguments. But when the year is up, I call everybody together and announce that things have been improving and I’m ready to set things right: starting today, everybody gets a 20% raise. First a 20% cut, now a 20% raise; we’re back to where we were, right? Wrong. I changed numbers mid- stream. When I cut your pay initially, I took twenty percent of $10.00, which is a reduction of $2.00. When I gave you a raise, I gave you twenty percent of your reduced pay rate of $8.00 per hour. That’s only $1.60. Your final pay rate is a mere $9.60 per hour.

Often, people make a strategic decision about what number to take a percentage of, choosing the one that gives them a more impressive-sounding, rhetorically effective figure. Suppose I, as the CEO of LogiCorp, set an ambitious goal for the company over the next year: I propose that we increase our productivity from 800 arguments evaluated per day to 1,000 arguments per day. At the end of the year, we’re evaluating 900 arguments per day. We didn’t reach our goal, but we did make an improvement. In my annual report to investors, I proclaim that we were 90% successful. That sounds good; 90% is really close to 100%. But it’s misleading. I chose to take a percentage of 1,000: 900 divided by 1,000 give us 90%. But is that the appropriate way to measure the degree to which we met the goal? I wanted to increase our production from 800 to 1,000; that is, I wanted a total increase of 200 arguments per day. How much of an increase did we actually get? We went from 800 up to 900; that’s an increase of 100. Our goal was 200, but we only got up to 100. In other words, we only got to 50% of our goal. That doesn’t sound as good.

Another case of strategic choices. Opponents of abortion rights might point out that 97% of gynecologists in the United States have had patients seek abortions. This creates the impression that there’s an epidemic of abortion-seeking, that it happens regularly. Someone on the other side of the debate might point out that only 1.25% of women of childbearing age get an abortion each year. That’s hardly an epidemic. Each of the participants in this debate has chosen a convenient number to take a percentage of. For the anti-abortion activist, that is the number of gynecologists. It’s true that 97% have patients who seek abortions; only 14% of them actually perform the procedure, though. The 97% exaggerates the prevalence of abortion (to achieve a rhetorical effect). For the pro-choice activist, it is convenient to take a percentage of the total number of women of childbearing age. It’s true that a tiny fraction of them get abortions in a given year; but we have to keep in mind that only a small percentage of those women are pregnant in a given year. As a matter of fact, among those that actually get pregnant, something like 17% have an abortion. The 1.25% minimizes the prevalence of abortion (again, to achieve a rhetorical effect).

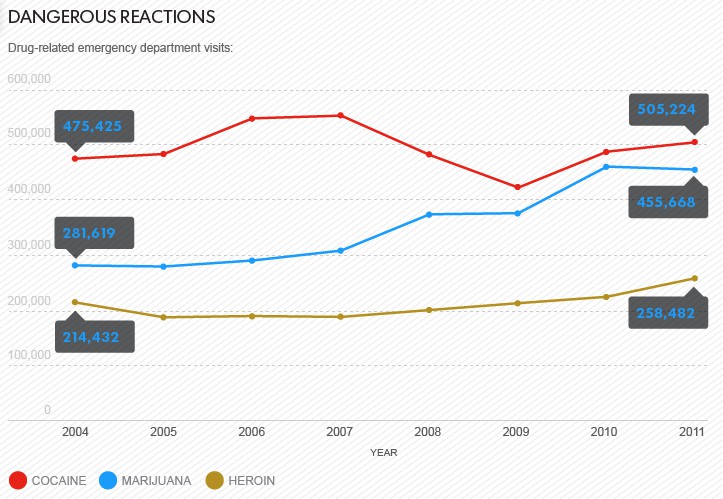
### 4. The Base-Rate Fallacy

The base rate is the frequency with which some kind of event occurs, or some kind of phenomenon is observed. When we ignore this information, or forget about it, we commit a fallacy and make mistakes in reasoning.

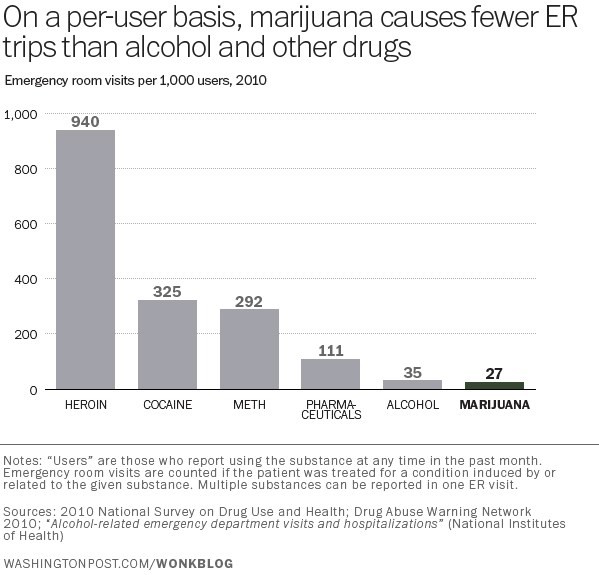
Most car accidents occur in broad daylight, at low speeds, and close to home. So does that mean I’m safer if I drive really fast, at night, in the rain, far away from my house? Of course not. Then why are there more accidents in the former conditions? The base rates: much more of our driving time is spent at low speeds, during the day, and close to home; relatively little of it is spent driving fast at night, in the rain and far from home.

Consider a woman formerly known as Mary (she changed her name to Moon Flower). She’s a committed pacifist, vegan, and environmentalist; she volunteers with Green Peace; her favorite exercise is yoga. Which is more probable: that she’s a best-selling author of new-age, alternative- medicine, self-help books—or that she’s a waitress? If you answered that she’s more likely to be a best-selling author of self-help books, you fell victim to the base-rate fallacy. Granted, Moon Flower fits the stereotype of the kind of person who would be the author of such books perfectly. Nevertheless, it’s far more probable that a person with those characteristics would be a waitress than a best-selling author. Why? Base rates. There are far, far (far!) more waitresses in the world than best-selling authors (of new-age, alternative-medicine, self-help books). The base rate of waitressing is higher than that of best-selling authorship by many orders of magnitude.

Sometimes people will ignore base rates on purpose to try to fool you. Did you know that marijuana is more dangerous than heroin? Neither did I. But look at this chart:



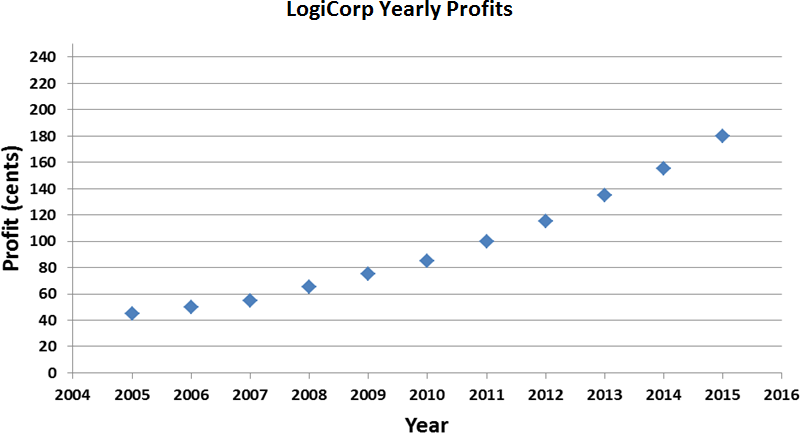
That graphic was published in a story in USA Today under the headline “Marijuana poses more risks than many realize.”[[21]](#footnote-21) The chart/headline combo create an alarming impression: if so many more people are going to the emergency room because of marijuana, it must be more dangerous than I realized. Look at that: more than twice as many emergency room visits for pot than heroin; it’s almost as bad as cocaine! Or maybe not. What this chart ignores is the base rates of marijuana, cocaine, and heroin use in the population. Far (far!) more people use marijuana than use heroin or cocaine. A truer measure of the relative dangers of the various drugs would be the number of emergency room visits per user. That gives you a far different chart:[[22]](#footnote-22)

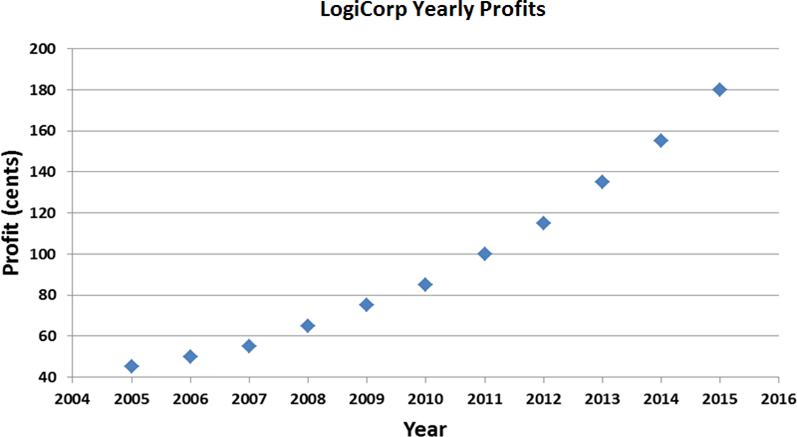


### 5. Lying with Pictures

Speaking of charts, they are another tool that can be used (abused) to make dubious statistical arguments. We often use charts and other pictures to graphically convey quantitative information. But we must take special care that our pictures accurately depict that information. There are all sorts of ways in which graphical presentations of data can distort the actual state of affairs and mislead our audience.

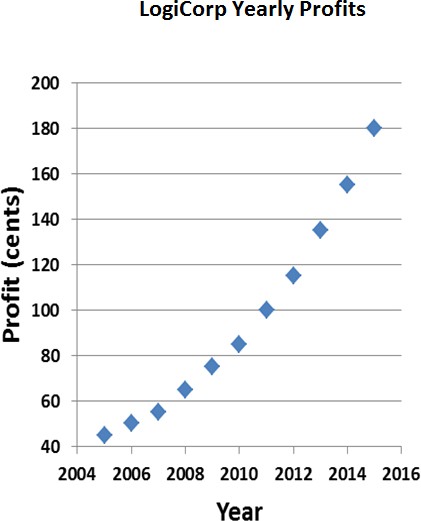
Consider, once again, my fictional company, LogiCorp. Business has been improving lately, and I’m looking to get some outside investors so I can grow even more quickly. So I decide to go on that TV show Shark Tank. You know, the one with Mark Cuban and panel of other rich people, where you make a presentation to them and they decide whether or not your idea is worth investing in. Anyway, I need to plan a persuasive presentation to convince one of the sharks to give me a whole bunch of money for LogiCorp. I’m going to use a graph to impress them with company’s potential for future growth. Here’s a graph of my profits over the last decade:

Not bad. But not great, either. The positive trend in profits is clearly visible, but it would be nice if I could make it look a little more dramatic. I’ll just tweak things a bit:

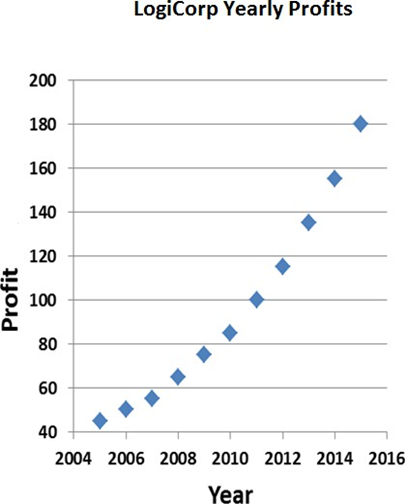


Better. All I did was adjust the y-axis. No reason it has to go all the way down to zero and up to 240. Now the upward slope is accentuated; it looks like LogiCorp is growing more quickly.

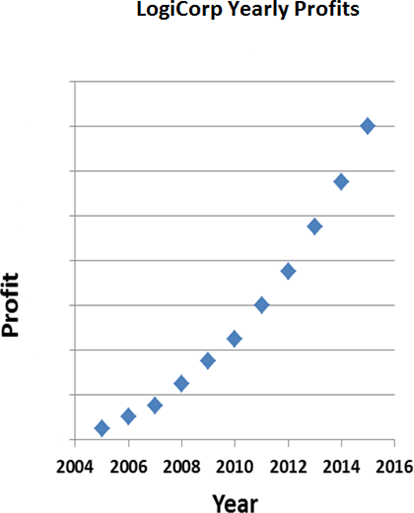
But I think I can do even better. Why does the x-axis have to be so long? If I compressed the graph horizontally, my curve would slope up even more dramatically:



Now that’s explosive growth! The sharks are gonna love this. Well, that is, as long as they don’t look too closely at the chart. Profits on the order of $1.80 per year aren’t going to impress a billionaire like Mark Cuban. But I can fix that:

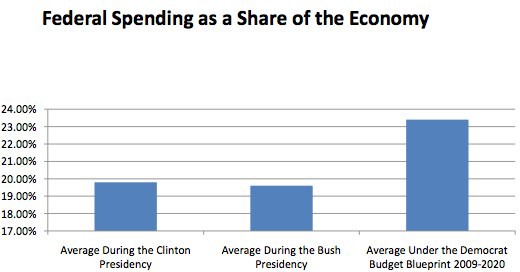


There. For all those sharks know, profits are measured in the millions of dollars. Of course, for all my manipulations, they can still see that profits have increased 400% over the decade. That’s pretty good, of course, but maybe I can leave a little room for them to mentally fill in more impressive numbers:

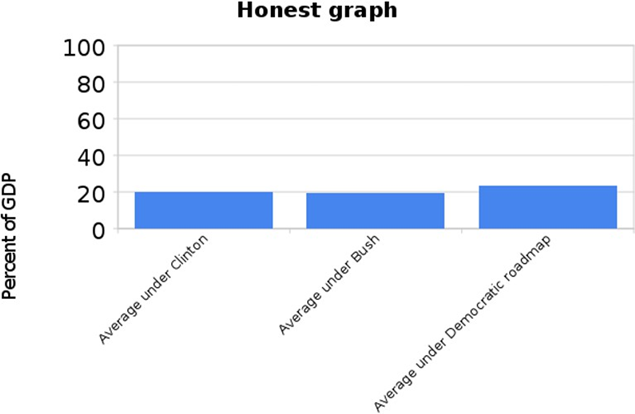


That’s the one. Soaring profits, and it looks like they started close to zero and went up to—well, we can’t really tell. Maybe those horizontal lines go up in increments of 100, or 1,000. LogiCorp’s profits could be unimaginably high.

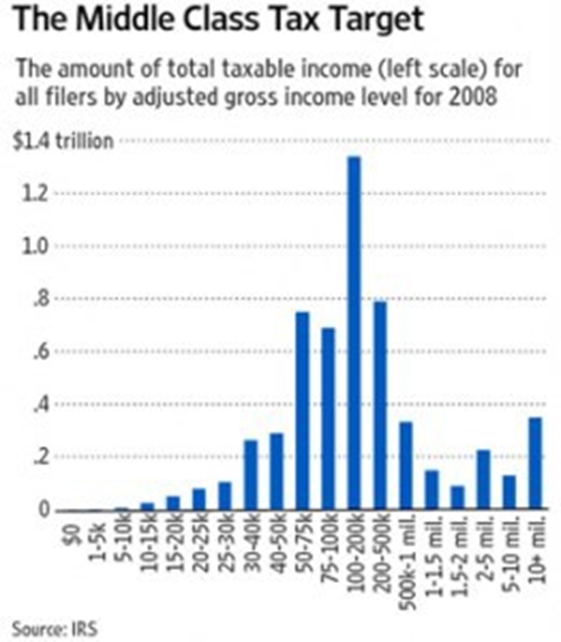
People manipulate the y-axis of charts for rhetorical effect all the time. In their “Pledge to America” document of 2010, the Republican Party promised to pursue various policy priorities if they were able to achieve a majority in the House of Representatives (which they did). They included the following chart in that diagram to illustrate that government spending was out of control:



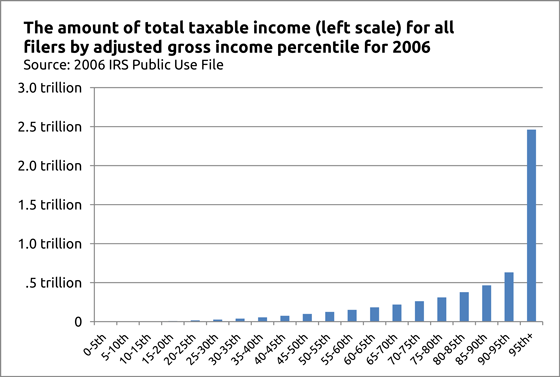
Writing for New Republic, Alexander Hart pointed out that the Republicans’ graph, by starting the y-axis at 17% and only going up to 24%, exaggerates the magnitude of the increase. That bar on the right is more than twice as big as the other two, but federal spending hadn’t doubled. He produced the following alternative presentation of the data[[23]](#footnote-23):



One can make mischief on the x-axis, too. In an April 2011 editorial entitled “Where the Tax Money Is”, The Wall Street Journal made the case that President Obama’s proposal to raise taxes on the rich was a bad idea.[[24]](#footnote-24) If he was really serious about raising revenue, he would have to raise taxes on the middle class, since that’s where most of the money is. To back up that claim, they produced this graph:



This one is subtle. What they present has the appearance of a histogram, but it breaks one of the rules for such charts: each of the bars has to represent the same portion of the population. That’s not even close to the case here. To get their tall bars in the middle of the income distribution, the Journal’s editorial board groups together incomes between $50 and $75 thousand, $75 and $100 thousand, then $100 and $200 thousand, and so on. There are far (far!) more people (or probably households; that’s how these data are usually reported) in those income ranges than there are in, say, the range between $20 and $25 thousand, or $5 to $10 million—and yet those ranges get their own bars, too. That’s just not how histograms work. Each bar in an income distribution chart would have to contain the same number of people (or households). When you produce such a histogram, you see what the distribution really looks like (these data are from a different tax year, but the basic shape of the graph didn’t change during the interim):



The lesson: don’t just glance at a chart and come away with a (potentially false) picture of the data. Charts can be manipulated to present whatever picture a person wants. People try to fool you in so many different ways. The only defense is a little logic, and a whole lot of skepticism. Be vigilant!

1. This chapter is based on *Fundamental Methods of Logic*, by Matthew Knachel. [↑](#footnote-ref-1)
2. Inspiration for this example, as with much that follows, comes from Darrell Huff, *How to Lie with Statistics*, New York: Norton, 1954. [↑](#footnote-ref-2)
3. In 2014, the richest fifth of American households accounted for over 51% of income; the poorest fifth, 3%. [↑](#footnote-ref-3)
4. “Gaussian” because the great German mathematician Carl Friedrich Gauss made a study of such distributions in the early 19th century (in connection with their relationship to errors in measurement). [↑](#footnote-ref-4)
5. This is an exaggeration, of course, but not much of one. The average high in San Diego in January is 65°; in July, it’s 75°. Meanwhile, in Milwaukee, the average high in January is 29°, while in July it’s 80°. [↑](#footnote-ref-5)
6. Pick a person at random. How confident are you that they have an IQ between 70 and 130? 95.4%, that’s how confident. [↑](#footnote-ref-6)
7. As a matter of fact, in current practice, other confidence intervals are more often used: 90%, (exactly) 95%, 99%, etc. These ranges lie on either side of the mean within non-whole-number multiples of the standard deviation. For example, the exactly-95% interval is 1.96 SDs to either side of the mean. The convenience of calculators and spreadsheets to do our math for us makes these confidence intervals more practical. But we’ll stick with the 68.3/95.4/99.7 intervals for simplicity’s sake. [↑](#footnote-ref-7)
8. [↑](#footnote-ref-8)
9. Actually, the typical level is now exactly 95%, or 1.96 standard deviations from the mean. From now on, we’re just going to pretend that the 95.4% and 95% levels are the same thing. [↑](#footnote-ref-9)
10. I am indebted for this example in particular (and for much background on the presentation of statistical reasoning in general) to John Norton, 1998, *How Science Works*, New York: McGraw-Hill, pp. 12.14 – 12.15. [↑](#footnote-ref-10)
11. Here’s an actual survey with that result:

    http://angusreidglobal.com/wp-content/uploads/2012/03/2012.03.04\_Myths.pdf [↑](#footnote-ref-11)
12. Interesting mathematical fact: these relationships hold no matter how big the general population from which you’re sampling (as long as it’s above a certain threshold). It could be the size of the population of Wisconsin or the population of China: if your sample is 600 Wisconsinites, your margin of error is 4%; if it’s 600 Chinese people, it’s still 4%. This is counterintuitive, but true—at least, in the abstract. We’re omitting the very serious difficulty that arises in actual polling (which we will discuss in a minute): finding the right 600 Wisconsinites or Chinese people to make your survey reliable; China will present more difficulty than Wisconsin due to the size of the population. [↑](#footnote-ref-12)
13. It’s even harder than this paragraph makes it out to be. It’s usually impossible for a sample—the people you’ve talked to on the phone about the president or whatever—to mirror the demographics of the population exactly. So pollsters have to weight the responses of certain members of their sample more than others to make up for these discrepancies. This is more art than science. Different pollsters, presented with the exact same data, will make different choices about how to weight things, and will end up reporting different results. See this fascinating piece for an example: http://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks- about.html?\_r=0 [↑](#footnote-ref-13)
14. Source: http://www.nbcnews.com/politics/elections/rebuke-tennessee-governor-koch-group-shows-its-power- n301031 [↑](#footnote-ref-14)
15. http://spotlight.ipsos-na.com/index.php/news/is-president-obama-up-or-down-the-effect-of-question-wording-on- levels-of-presidential-support/ [↑](#footnote-ref-15)
16. See here, for example: https://www.washingtonpost.com/news/monkey-cage/wp/2016/12/13/why-the-polls- missed-in-2016-was-it-shy-trump-supporters-after-all/?utm\_term=.f20212063a9c [↑](#footnote-ref-16)
17. The title of this section, a lot of the topics it discusses, and even some of the examples it uses, are taken from Darrell Huff’s book, *How to Lie with Statistics*. [↑](#footnote-ref-17)
18. John Paulos, “We’re Measuring Bacteria with a Yardstick,” November 22, 2000, The New York Times. [↑](#footnote-ref-18)
19. Steven Verburg, “Study: Budget Could Hurt State’s Economy,” March 20, 2011, Wisconsin State Journal. [↑](#footnote-ref-19)
20. Not because the school’s administration was particularly enlightened. They could only open with the financial support of four wealthy women who made this a condition for their donations. [↑](#footnote-ref-20)
21. Liz Szabo, “Marijuana poses more risks than many realize,” July 27, 2014, USA Today. The following link opens in a new window: [Marijuana article](http://www.usatoday.com/story/news/nation/2014/07/27/risks-of-marijuana/10386699/?sf29269095=1) [↑](#footnote-ref-21)
22. From German Lopez, “Marijuana sends more people to the ER than heroin. But that's not the whole story.” August 2, 2014, Vox.com. The following link opens in a new window: [second marijuana article](http://www.vox.com/2014/8/2/5960307/marijuana-legalization-heroin-USA-Today). [↑](#footnote-ref-22)
23. Alexander Hart, “Lying With Graphs, Republican Style (Now Featuring 50% More Graphs),” December 22, 2010, New Republic. Link opens in new window: [Lying with Graphs article](https://newrepublic.com/article/77893/lying-graphs-republican-style). [↑](#footnote-ref-23)
24. The article is at the following link, which opens a new page: [Wall Street Journal article](http://www.wsj.com/articles/SB10001424052748704621304576267113524583554). [↑](#footnote-ref-24)