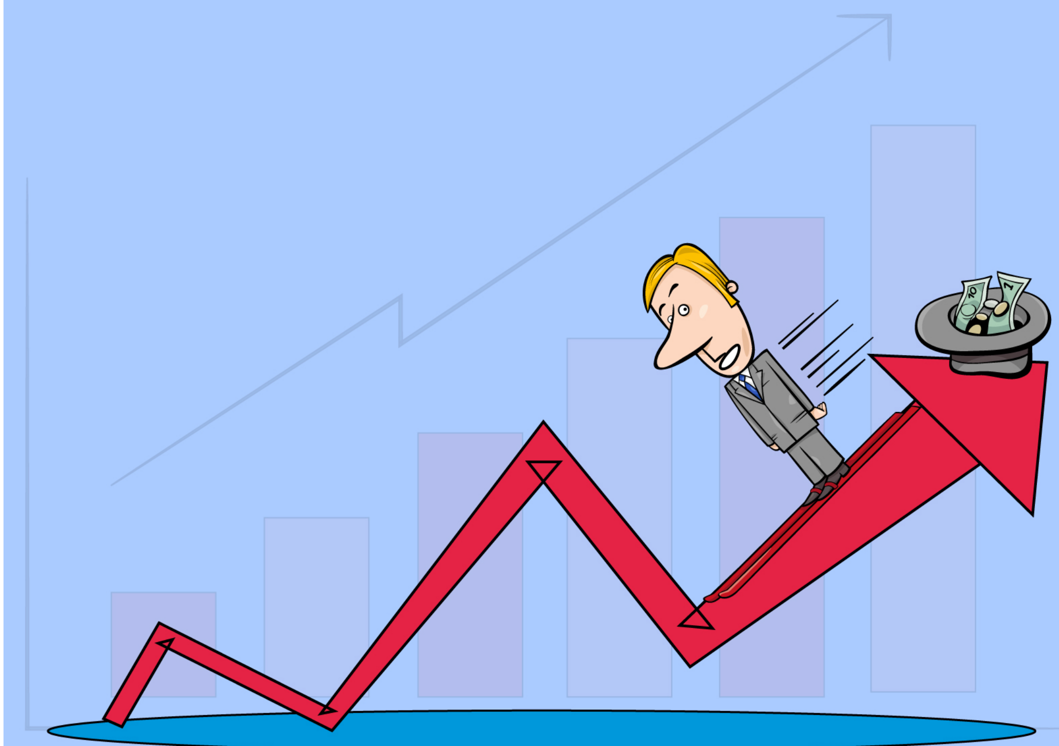
SUMMARY

HOW TO LIE WITH STATISTICS

DARRELL HUFF



QuickRead

FREE BOOK SUMMARIES

Summary of “How to Lie With Statistics” by Darrell Huff

Written by Lea Schullery

Learn to Identify How Companies Use Statistics to Deceive and Manipulate the Public

Introduction

When author Darrell Huff moved from Iowa to California, he remembers his father-in-law stating “There’s a mighty lot of crime around here.” And according to the newspaper his father-in-law was reading, there was. The problem with the statistics, however, was that it was based on a biased sample, and like many sophisticated statistics, “it was guilty of semi attachment: It assumed that newspaper space given to crime reporting is a measure of crime rate.” Unfortunately, averages, trends, and graphs are not always what they seem. There is a secret language of statistics that is meant to sensationalize, inflate, confuse, and oversimplify. Of course, statistical methods and statistical terms are necessary when reporting the mass data of social and economic trends, business conditions, and polls, but if writers aren’t using words with honesty and readers don’t know what they mean, the result can only turn into nonsense. Even worse, writers use statistics to deceive and dupe readers, they know all the tricks. Thankfully, you, an honest person, can learn these tricks for self-defense.

It is Near Impossible to Create a Truly Random Sample

Imagine you have a barrel of beans, some red and some white, and you want to find out exactly how many of each color you have. Well, there is only one way to do this: count 'em. However, this process is both time-consuming and difficult. No one wants to spill beans all over the floor and count them individually, right? You can, however, find out approximately how many are red in a much easier fashion. Simply pull out a handful of beans, count them, and figure that the proportion will be the same all through the barrel.

This method involves creating a sample, which is a carefully chosen data set used to represent the whole of the thing you wish to analyze - in this scenario, the color of beans. Sampling is the basis for concluding statistics; therefore, it is absolutely critical that you create a good sample. The criteria for a good sample is that it is large enough and selected properly to represent the whole well enough. If not, it may be inaccurate or biased. For example, *Time* magazine once stated that "the average Yaleman, Class of '24, makes $25,111 a year." But how did they get this number? First, they had to get a sample calculated from the amounts the Yale men *said* they earned. Did any of them exaggerate? Did others minimize the amount for tax purposes? We don't know.

Second, the sample came from those that graduated from yale in 1924, twenty-five years before the study was done. Did researchers contact each graduate? Surely, some addresses are unknown or questionnaires were sent out and not returned. And "Who are the little lost sheep down in the Yale" whose addresses are unknown? The big-time earners - Wall Street men? Corporation directors? Manufacturing executives? No. The addresses of the rich are not hard to come by. It's the clerks, writers, artists, etc. whose addresses are unknown. In other words, the sample wasn't good and failed to represent the whole.

For a sample to represent the whole, it must be random. Unfortunately, a purely random sample is both difficult and expensive to obtain. If you want a truly random sample, you must randomly select your sample group regardless of income, social class, age, race, etc. This seems easier said than done. For instance, if we go back to our barrel of red and white beans, let's assume that we pulled out a random handful for our sample. What if the barrel wasn't mixed up and you only grabbed a handful of red beans off the top? Your sample would conclude that the barrel is full of red beans. As a result, you would have experimented with sample bias, similar to that of the Yale graduates.

Non-Random Samples Lead to Sample Bias

So how can we avoid biased samples? We can spill the entire barrel of beans and count them individually, but we can also use a strategy called stratified random sampling. To get a stratified sample, you must first divide your universe into several groups in proportion to their known prevalence. It is here, pinpointing that proportion, where trouble can occur. Your information about the proportion may not be correct.

For instance, let's assume your universe is comprised of vegans, how do you know what proportion of them are a particular gender, race, or age group? Without a huge database of this information, it is incredibly difficult to know. Instead, you'll need a random sample within each subgroup and then randomly interview people from each subgroup, like black vegans, vegans under 18, low-income vegans, and so on. Even with all this in mind, it is difficult to keep your sample truly random. For instance, contacting each demographic proves difficult, after all, does everyone check their email? Have access to email? Own a phone? This is a running battle against bias, a battle that is often lost.

For example, *Literary Digest* once fell victim to sample bias when polling its readers about the upcoming 1936 election between Alf Landon and FDR. In the past, *Literary Digest* had correctly predicted the presidential election multiple times, this year however, was different. The magazine predicted Landon to win over FDR, but as we now know, FDR won by a landslide. So where did *Literary Digest* go wrong? Well, they conducted non-random sampling by relying on their reader's predictions via a telephone survey. This is critical information because people who could afford telephones and magazine subscriptions in 1936 were not a cross-section of voters. Economically, they were Republican voters who skewed the results of their polls.

To avoid sample bias, *Literary Digest* should have utilized stratified random sampling and divided their subgroups within their readership proportionally. From there, they should have produced a random sample within each subgroup and then reached them using diverse methods.

Beware of Averages

Let's say you are in the market to buy a house. You are looking for a property to buy and your real estate agent tells you that the average income in a particular neighborhood is $200,000 a year. This information entices you to buy the property; however, a year later you meet again with your agent who tells you the average income in the neighborhood is only $40,000 a year. Was your agent lying to you now, or was he lying last year?

Well, the answer is that the agent wasn't lying at all! "That is the essential beauty of doing your lying with statistics. Both these figures are legitimate averages, legally arrived at. Both represent the same data, the same people, the same incomes." The trick the agent used was to use a different kind of average each time, the word "average" having a loose meaning. That is because there are three common kinds of average: mean, median, and mode.

The $200,000 figure is the one the real estate agent used when calculating the mean, or the arithmetic average of the incomes of all the families in the neighborhood. To calculate the mean, you simply add up all the incomes and divide by the total number of variables. So where did the $40,000 figure come from? Well, the smaller figure is the median, which is the middle point of your sample. For example, the median of 3, 4, 7, 10, and 20 is 7 because half of the values are above 7 and half are below. So the median tells you that half the families in question have more than $40,000 a year and the other half have less.

The real estate agent could have also used the mode, which is the most frequent figure in a series. So if most of the families in the neighborhood make $44,000 then the modal income would be $44,000. In this case, the word "average" is virtually meaningless. So when you read an announcement by a corporate executive that the average salary is $100,000, you will want to question whether the executive used mean, median, or mode to arrive at that average figure. So when you see an average-pay figure, first ask yourself, “Average of what?” and “Who’s included?” Asking these questions can help you gain a better understanding of how that average was calculated.

Beware of Significance Bias

As you learned in the first chapter, a sample is only considered good if it is a good representation of the whole. And one of the ways to get a good sample is by making it large enough to become "statistically significant." For instance, we all know the probability of tossing a coin and it landing on heads, right? The probability is 50 percent. Well, let's check and see. When you toss a penny ten times, how often does it land on heads? Was it 50 percent? While you may get a 50-50 result, you probably won't. In fact, the author did this experiment and the penny landed on heads 8 out of 10 times, which gives a probability of 80 percent!

So why didn't it come up heads 50 percent of the time? Simply put, the experiment wasn't repeated enough. "If your patience holds out for a thousand tosses, you might be more likely to come out with a result very close to half heads... Only when there is a substantial number of trials involved is the law of averages a useful description or prediction." For this reason, reliable studies must use a statistically significant sample to avoid "significance-biased" results, like the coin toss. Unfortunately, some companies will take advantage of this significance bias when trying to astonish consumers with new products and services.

For example, a toothpaste company might advertise its product by putting "Users report 23% fewer cavities" on its packages. From these results, you might assume that this toothpaste is remarkable, right? Well, upon closer inspection you might read the small type which states the test group of users consisted of just twelve people. This sample size works well for the toothpaste company, and here's why. With a small sample size, the toothpaste company will keep count of each cavity for six months before switching to the company's toothpaste. Next, one of three things will happen: more cavities, fewer cavities, or about the same number. If the first or last of these possibilities occur, the toothpaste company will ignore findings and try again.

Sooner or later, a test group will show a big improvement worthy of a headline and an entire advertising campaign will be created. This is bound to happen regardless of whether the test group used the toothpaste or not. Ultimately, using a small group allows for dramatic results - a well-known fact that companies will take advantage of to sell their products or services.

Look Out for the Standard Error

Imagine there are two children, Peter and Linda, who take an IQ test at school. When the results come back, you learn that Peter's IQ is 90, and Linda's is 101. You also learn that the average IQ for a "normal" child is 100. From this information, you might assume that Linda is above average and Peter is below average. As it turns out, conclusions like these are sheer nonsense. Not only because IQ tests are not an accurate measure of intelligence but also because the numbers are inaccurate.

For the results of the IQ test to become more accurate, you would need to calculate the *standard error*. Let's imagine Peter and Linda take the IQ test a total of three times and each score is different, with such deviation, you'll need to find the standard error to better understand the IQ of Peter and Linda. To do this, you'll need to start with the average IQ - 100. Next, you'll need to add the deviation from each result to the average. Let's imagine Peter's scores were 90, 100, and 110. So the first deviation is 10 (90 to 100), the second deviation is 0 (100 to 100), and the third (100 to 110) is 10.

Next, you divide your total sum of deviations (10+0+10=20) by the number of results (3), and then you have 6.67 - your standard error! So what does this even mean? With this number, we now know that the IQs that range from 93 to 107 is "normal" for Peter; therefore, his IQ is 100+/-7. Of course, this number will become more precise the more you test your IQ. This becomes important as the "normal" range for IQ is 90-110, so when we originally believed Peter to be "below average" because of his score of 90, we were making an assumption based on incomplete statistical information.

Beware of Semiattached Figures

Let’s say you want to prove that a product you created works, but you can’t figure out how. What should you do? You know that statistics and numbers elicit trust among consumers; therefore, all you have to do is create a semiattached figure. Don’t worry, creating a semiattached figure is pretty easy. All you have to do is pick two or more things that seem alike - but aren’t- and draw a comparison between them. In a sense, you are simply ignoring a gap in your argument to achieve your desired conclusion.

For example, let’s say the product you created is a medicine to cure a common cold. While you can’t prove that your medicine cures colds, you can publish a laboratory report that half an ounce of it killed 31,108 germs in a test tube in eleven seconds. You can ensure that the laboratory is reputable or has an impressive name, reproduce the report, photograph a doctor-type model in a lab coat, and now you have your finished advertisement! What your study failed to do, however, is demonstrate whether or not the medicine works in the human body. After all, what works in a test tube may not perform in the human body, especially after being diluted with water, food, etc. To the average consumer, however, the statistics sound credible.

Unfortunately, advertisers aren’t the only people who will fool you with numbers. For instance, an article on driving safety published by *This Week* magazine once published that you would have four times a greater chance of staying alive driving at seven in the morning versus driving at seven at night. The evidence states, “Four times more fatalities occur on the highways at 7 p.m. than at 7 a.m.” While this may be true, it fails to mention that more people are killed in the evening simply because more people are on the highways at that time to be killed. In fact, more accidents occur in clear weather than foggy weather simply because clear weather is more common than foggy weather.

Additionally, more people were killed by airplanes in 1953 than in 1910. Does that mean that modern planes are more dangerous? Of course not! There are simply hundreds of times more people flying in 1953 than in 1910. Companies also use deception when using comparisons and percentages together. All they do is “forget” to state what they are comparing. For example, you might drink a glass of orange juice that contains “30 percent more juice,” and assume you’re drinking a healthier choice. But what is it 30 percent more of? The answer could be anything.

Beware the Post-Hoc Fallacy

An anti-smoking researcher once went through the trouble to prove if cigarette smokers make lower grades in college than non-smokers. As it turned out, they did! For anti-smoking groups, the news was promising for their argument that smoking makes dull minds. While this study was properly done, with a big enough sample that was honestly and carefully chosen, the conclusion that smoking makes dull minds follows a powerful fallacy that many often fall victim to.

The post-hoc fallacy states that we often assume causal relationships because we assume “if B follows A, then A has caused B.” Therefore, an assumption is being made that since smoking and low grades go together, then smoking causes those low grades. But the truth is the same conclusion could be made that people with low grades are driven to smoke tobacco and drink alcohol. It is only human nature that in statistics, we often look for correlations like these to explain the world around us. Unfortunately, not all correlations make sense!

This becomes even more complicated when we see that there are many different types of correlations. One correlation is produced by chance. You may be able to produce a particular set of figures to prove some unlikely thing, but if you try again, your next set may not prove it at all. We saw this with the toothpaste company who simply threw away results they didn’t want until they stumbled upon a test that worked in their favor. Another type of correlation fallacy is covariation, in which the relationship is real but it is impossible to be sure which variable is the cause and which is the effect. For instance, there is certainly a correlation between income and ownership of stocks. But do wealthy people buy more stocks? Or do stocks make you more wealthy? You cannot say that one has produced the other.

There are many types of correlation-causality fallacies, but we can be certain that while correlation is critical for causality, one does not always lead to the other.

Question Statistics and Be Aware of Common Tricks

“Misinforming people by the use of statistical material might be called statistical manipulation; in a word (though not a very good one), statisticulation.” And while some will certainly lie with statistics, not all lies come with malicious intent. Regardless of their meaning and intentions, how can you identify bad statistics and defend yourself against them? The first thing you can do is ask the right questions. Look at who conducted the study first and determine what their motives might be. For instance, studies that are sponsored by certain companies should be carefully examined since companies will want the study to produce results in their favor.

Next, you should be suspicious of small or poorly selected samples, as they are known to produce biased results. You should scrutinize the sample to ensure it is big enough to make it significant, to ensure it involves a variety of participants, and to ensure it didn’t omit important groups from the study. You should also make sure the authors provide the standard error and specify the type of average they are using. If this information is not provided, then it is best to assume that something is not right.

Lastly, you should watch out for a switch on the subject. If this happens, the author is likely to create a correlation between the raw figure and the conclusion. Therefore, you should ask yourself if the numbers actually lead to their conclusion, or determine if the person is falling victim to a post-hoc fallacy. Simply ask yourself, “Does it make sense?” After all, marketers, businesses, and others will always lie with statistics to produce the results they want; therefore, it is up to you to be aware of the tricks they use to stay one step ahead and keep yourself from falling victim to inflated or inaccurate statistics.

Final Summary

Advertisers, marketers, businesses, politicians, and others understand the power of numbers and statistics. Unfortunately, they also understand how to use these statistics to create biased results by easily misusing and misinterpreting them to make them in their favor. It is up to you to stay informed and understand how others will misuse statistics to deceive you. If you ask the right questions and keep an eye out for the tricks they use, you'll be able to fight back against statistics and never be duped again!