Statistical Literacy

Knowing What's Wheat and What's Chaff

Chase Brady

June 15, 2008

Introduction

The term "Statistical Literacy" has come to have two slightly different meanings. The first is simply familiarity with the terminology and methods of statistics. The second, more recent use of the term is indicated by the subtitle of this article. Here we are talking about the ability to discriminate between reliable statistical information and not-so-reliable information. It is this second meaning we are using here. Sadly, unreliable statistics abound, and being able to recognize them is becoming a necessary form of modern day self-defense. Fortunately, one does not need to know all that much about the mathematics of statistics in order to recognize the majority of the unreliable statistics. What one needs, instead, is a careful examination of the process by which information was obtained, and of the logic by which conclusions were drawn. If the process of gathering information is faulty, then you have bad data, and no mathematical magic is going to improve it. It's rather a "garbage in – garbage out" situation. What's worse, in most cases there is no mathematical magic that will even tell you that your data is bad. You have to examine the way it was collected and be alert for common mistakes. Another area where mistakes are commonly made is in drawing conclusions. Statistics has an air of legitimacy that makes conclusions seem obvious and inescapable, even when they are neither. We shall look at some of the common mistakes made in both processes – data gathering and conclusion drawing.

Consider the Source

Suppose you pick up a magazine and see an advertisement for tooth paste which claims that using the advertised brand reduced cavities by 20% in clinical studies. Most of us are naturally skeptical of statistics found in advertisements such as this. We are often advised to "Consider the source." There is a degree of wisdom in this, and yet such caution can be overdone. Certainly we would expect someone who is trying to sell us a product to tell us only positive things about that product, and we would not be so shocked to find that some of those positive things aren't altogether accurate. Those who have a vested interest in the outcome of a research project *may* not be giving us the most reliable information. This gives us reason to look closely at their figures and their methods. But we need to be a little cautious here. Having a vested interest in the outcome does not preclude doing valid research. The fact is, a large portion of the research that is done is in fact done by people who have a vested interest in the outcome. University professors and research associates nearly always have a strong interest in the outcome of their efforts. They often have pet theories they're trying to support or wish to question existing theories. If they have no other agenda, they at least want their research to be published. Nonetheless, we expect them to have the integrity to report their findings honestly and objectively, without allowing their personal bias to drive their conclusions. In cases where results are inconclusive and unpublishable, we expect them to accept the fact. The fact that the people funding or doing the research may have an agenda gives us cause to be skeptical, but not to dismiss their findings without examination.

Particular caution does need to be exercised when the source of information is an *advocate*. An advocate is "one that argues for a cause; a supporter or defender." [1] They range from professional spokesmen hired by companies to speak on their behalf to passionate volunteers seeking to alert society about social problems. On the one hand, society has a genuine need for many of these people. Social problems do not get addressed until someone brings them to public attention. On the other hand, advocates have an obvious agenda. A spokesman for a company is not paid to tell us negative things about that company. Someone urging action on a social problem must convince us of the severity of that problem. In the minds of some, the end justifies occasionally being slightly less than truthful.

No one is entirely innocent here, of course. Ultimately, the consumers of statistical information are you and I – the public. We have our own

prejudices and preformed opinions, and it often requires deliberate conscious effort to keep those from dictating our own conclusions. When we encounter research which tends to support our preconceptions, we tend to take it as scientific proof of what we knew all along, often without even looking at the details. When confronted with research which threatens our preconceptions, we want to go through the details with a fine tooth comb, looking for some reason to "Say it isn't so!" The first step towards statistical literacy may well be learning to examine those things we wish to believe and those we don't through the same lens.

Estimates

Some of the most dubious statistics we encounter arguably aren't statistics at all, or at least aren't very statistical. They are not the result of painstaking research or carefully conducted surveys, though they may sometimes utilize such results. They are called estimates, though they would perhaps be better labelled 'expert guesses', and we see them all the time. You read in a journal, "While one in five Americans suffers from mental illness, it often goes undiagnosed." [2] A question you should form the habit of asking is "How do they know that?". Think about it. How does one keep track of *undiagnosed* illnesses? It is possible to get a reasonably accurate estimate for this sort of thing, but it is difficult – and expensive. One would have to carefully select a random sample of people from all walks of life (including the homeless), and somehow test them for mental illness. Has this been done? For this particular quote, the author simply states without reference, so we don't know where she got this number. Sometimes the numbers we see quoted in this fashion represent little more than guesswork. We tend to assume that given any social problem, either government agencies or university professors with grant money are doing research and keeping track of the situation. This is not necessarily the case. When a social problem first comes to light, little or no research is likely to have been done. Even when attention is focused on them, some problems are nearly impossible to keep track of. How do you count the homeless, or illegal aliens, or illegal substance abusers, or feral cats? Attempts to count those who are hard to find or reluctant to self-identify are likely to be difficult, expensive, and ultimately, not altogether reliable.

Yet advocates need these numbers. When an advocate is trying to raise consciousness about a social problem, one of the first questions he or she

will be asked is "How many people are affected by this problem?". We as a society do not have the resources to effectively address every problem that we encounter. We have to prioritize. Knowing what percentage of society is affected or will be affected by a problem or program is indispensable if we wish to prioritize well. So the activists turn to experts for these numbers. The problem is, if no research has been done, the experts won't know. We seem to assume that within their field, experts know everything, or if they don't, then at least their guess is better than ours. This also, is not necessarily true. Ask a history professor whose specialty is American history who fired the first shot at the battle of the Little Bighorn. I doubt that he could help you much. He or she might know a great deal about the battle, but not who fired the first shot. There are some things that we just don't know, and even the experts can't guess with any reliability. So we have a critical need to know, but the knowledge just isn't there to be had, nor is there much to base a guess on. Often a sympathetic expert will venture a guess, saying something like, "Could be a hundred thousand, could be millions. That's very rough, but it's the best I can do. We just don't know." What we hear from the advocate is "Experts say there may be millions of people affected." The disclaimer has gotten lost. Experts are very careful about what they say when they are publishing their own work in a peer-reviewed journal. They are much more casual about comments made in an informal interview. Experts sometimes use formulas in their estimates, and the calculations make it seem very scientific, but if any of the numbers going into the calculation are themselves rough guesses, then so is the outcome. It's another of those "garbage in – garbage out" situations.

Another problem encountered in many estimates, whether they are based on careful statistical research or guesswork, is the problem of definitions. If you are going to count or estimate a number, you need a careful working definition of the thing being counted, one that distinguishes between what should be counted and what should not. A term which may seem clearly understood in general usage can become a problem when you begin to count. For example, take the color red. We all (except for those with certain visual difficulties) know what red means. We recognize it when we see it – or do we? Suppose you were given a bag containing 150 marbles, all solid colors, ranging over the entire spectrum, with no two precisely the same color. Now suppose you were asked to count the red ones. You would soon find yourself asking "Just what do we mean by red?" Is burgundy red? Rust? Crimson? Magenta? Where do we draw the line between red and not red?

The simple act of counting requires definitions of a precision most folks are not used to. We encounter this problem often when we are dealing with social statistics. The example of mental illness is a good one. When we hear the term "mental illness" we tend to think of behavior which is erratic and dysfunctional, possibly even dangerous. Often psychologists prefer a broader definition, including conditions such as bipolar disorder, anxiety and depression. Then there are phobias and compulsions, ADD, eating disorders, substance abuse, shyness. We could even include learning disabilities, test anxiety, and math anxiety. Are we interested only in chronic conditions, or should we also consider temporary conditions? Temporary anxiety and depression are very common among students, the unemployed and the recently divorced, or bereaved. Should we consider every one who has ever suffered from any of these conditions to be mentally ill? It becomes clear that we could broaden our definition far enough that the real oddballs would be the ones who are not being counted as "mentally ill". So just where should the line be drawn? Often, there is no widely accepted answer to a question such as this. The researchers must define the thing they are counting themselves, and the numbers they get will depend heavily on how they choose to do so. So we need to ask, "Just how is that being defined?"

Surveys and Sample Bias

There is one simple reality which everyone needs to understand about surveys. For any group larger than can be reasonably assembled in one place at one time, doing a survey which will give us reliable results is *a lot of work*. Often the hours spent collecting the data and the cost of hiring people to spend those hours far exceeds the available resources, rendering a proper survey impractical. "But wait!", you say. "This is the age of the Internet. If you can get a list of email addresses, it's no problem to send out a survey to thousands of people. You can even get software to track the results." Well, yes, that's true, but there was a qualifier in my statement. We want "reliable results." It's easy to do a survey. Doing one which will give us results we can trust is another matter altogether. I have on more than one occasion found myself sitting on a committee having to bite my tongue to keep from lecturing my colleagues about this. People have a hard time accepting the fact that email surveys are not likely to work. We don't have time or money to do door to door canvassing, or even phone surveys. While many will con-

cede that an email survey may not be ideal, most seem to feel that some information is better than none, even if it's not entirely reliable. I would like to convince you otherwise. Let's look at a couple of examples. (Actually, neither of these are email surveys, but the similarities will become clear.)

- One of the most accurate straw polls for a while was the Literary Digest poll. It gained a wide following for its straw polls of presidential elections in the early 20th Century. The Digest's straw polling methods were based on the then-current belief that the bigger the sample, the more accurate the results.

  During the last year (1936) the Digest polled, over 10 million questionnaires were mailed out, and over two million people responded. Compared to the roughly 1,500 or fewer people interviewed in a modern poll, two million respondents seem almost incalculable.

  It was, however, with the 1936 poll that the Literary Digest's techniques became a symbol of flawed polling, and the Digest thereafter would have a special place in polling research infamy. Based on a sample of two million, the Digest confidently predicted that Pennsylvania native Republican Alf Landon, now a Kansan, would overwhelm Democrat Franklin Roosevelt.

  Inconveniently both for the Digest and Alf Landon, Roosevelt easily won 46 of 48 states and 63 percent of the vote – handing Landon one of the worst defeats in presidential election history. Landon, it is reported, felt bad, but he survived. The Digest didn't. [3]

- Television news programs like to conduct call-in polls of public opinion. The program announces a question and asks viewers to call one number to respond 'yes' and another for 'no'. ... The ABC program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. More than 186,000 callers responded, and 67% said 'no'. (By comparison) ... a properly designed sample showed that 72% of adults wanted the UN to stay in the United States. [4]

The biggest single problem with both of these polls, as well as email surveys, is that they are *voluntary response surveys*. A large number of people are given the opportunity to participate, and data is collected from those who choose to do so. The sample, i.e. the people who choose to respond, is *self-selecting*. The problem is that those who choose to respond are often different from those who decline. This is called *nonresponse bias*. The *response rate*, i.e. the percentage of those given the opportunity to participate who actually do so, becomes critical. While there is no general agreement on what a good response rate should be, and in fact, it depends on the nature of the questions being asked, most estimates place acceptable response rates well above 50%. For example, the National Center for Education Statistics standards specify "Any survey stage of data collection with a unit or item response rate less than 85 percent must be evaluated for the potential magnitude of nonresponse bias before the data or any analysis using the data may be released." [5] Voluntary response surveys rarely achieve acceptable response rates.

The problem with a poor response rate is not just that you are not getting data from everyone. In fact, pollsters today routinely use relative small samples to assess the opinions of large populations, and do so with fairly reliable results. Pollster George Gallup made his reputation in the same 1936 election by correctly predicting Roosevelt as the winner based on a much smaller sample than the Literary Digest's. The trick is getting a *representative* sample, i.e. one that is not different from the larger population. To see why self-selecting samples are rarely representative, we consider who would tend to be the first to respond to a survey. In general, those with strong opinions, and particularly those who consider their opinions to be in the minority, are more likely to respond than those who don't really care that much, or who expect their opinion to be echoed by most everyone else. On controversial issues particularly, a vocal minority is likely to be vastly over-represented. For example, in the 1936 Literary Digest poll, Roosevelt was an incumbent president. It appears that many of those who did not like his policies went out of their way to make their opinions known. In the general elections, where the votes actually counted, it was a different story. Voluntary response surveys have their strongest appeal to those who have an axe to grind. The situation is getting worse, as many experts have noted that response rates in general are declining. [7] It seems that the proliferation of marketing surveys as well as others has left many Americans bored with the whole idea, and unwilling to participate.

So how do we get a good response rate from a survey? First we email them, then we send out letters to those who didn't respond to the email, then we proceed to phone calls, then we try to set up in person interviews, etc. That's why I said it's a lot of work. If we do not have the time or funding to survey everyone we'd like, then we chose the largest *simple random sample* of those people that we can expect to contact and we use them. For example, if you want to get information about all students at WVUP, but you can only realistically contact in person a hundred or so, then choose at random a hundred students and do everything short of harassment to get a response from them. 90% of 100 randomly chosen students is a much better sample than a self-selected 25% of 4000 – guaranteed.

Non-response bias is not the only type of sample bias we encounter. When we say the word "bias", we tend to think of a conscious predisposition for one outcome over another, but sample bias is much more subtle than that. As the previous examples illustrate, great care must be given to the process of selecting a sample to insure that no inadvertent bias occurs. *Sample bias* occurs anytime a segment of the population tends to be either favored or excluded by the selection process, whether it is deliberate or accidental. The Literary Digest's poll was also subject to another type of sample bias. They sent out over 10 million survey questionnaires. Where did they get a list of 10 million addresses? They used four sources:

- subscribers
- people in the phone book
- auto registry records
- voter registration records

Bear in mind, this was 1936, the midst of the great depression. The poorer folks would have been unlikely to subscribe to magazines, have phones (not everyone did in 1936), or own automobiles. FDR was the champion of the poor, and it is likely that many of his supporters were systematically excluded by the selection process. A few more examples of sample bias are:

- Polling customers at a shopping mall about their political opinions. There is evidence that people who dislike and avoid shopping malls tend to vote differently than those who frequent them.

- Comparing the grades of students who have participated in optional tutoring sessions to those who have not is an example of *self-selection bias*. Since choosing to participate in and of itself may show an unusual level of initiative, we would expect those students' grades to be better even if the actual tutoring is ineffective.

- Selecting a sample from the white pages of a phone book will systematically exclude those with unlisted numbers, those without phones, newcomers and those whose only phones are cell phones. The last categorically has grown significantly over the past few years, particularly among the young.

- Doing phone surveys between the hours of 5 and 7 PM is likely to exclude a lot of single people, since they are less likely to dine at home.

The best method of selecting a sample is taking a *simple random sample* of the entire population. Informally, simple random sampling is any selection process mathematically equivalent to drawing names from a hat. The problem is, you have to have a hat full of names to draw from, i.e. a list of all individuals in the population of interest. In cases where no such list can be had, another selection method must be used, and a great deal of care must be taken to avoid inadvertent sample bias.

Anecdotal Evidence

Everyone loves a good story. Advocates love to tell us stories about people who are affected by problems, or about those who have been helped by programs. Such stories can be very helpful in making those problems/programs seem more real to us. They give us the human side. What they do not give us is numbers. They tell us nothing about the number of people affected by the problem/program. When we are assessing the extent of a problem or the value of a program, there are two distinctly different questions that need to be considered:

1. What is the impact on the individual?

2. How many individuals are affected?

Stories can be quite helpful in regards to the first question, but are of little use in approaching the second. The problem is that humans have a tendency

to confuse the two questions, and often take the stories as an indication that more people are affected than actually are. This is particularly true for those who have some personal involvement in those stories. I've heard teachers who were involved in experimental programs comment, "I'm sure this program is effective. I've seen so many students turned around!" When confronted with test scores that indicate no significant change, I've heard them remark, "I don't care what the numbers say. I know what I've seen!" It is hard for us to accept that our personal experience might not be borne out by statistics. But the fact is, personal experience is not a reliable indicator of trends. The more personal the stories are, the less analytical we tend to be. We place undue weight on those things which seem to fulfill our hopes and discount those which do not. Statisticians refer to these personal stories/experience as *anecdotal evidence*.

According to Wikipedia,

> The expression *anecdotal evidence* has two quite distinct meanings.
>
> 1. Evidence in the form of an anecdote or hearsay is called anecdotal if there is doubt about its veracity: the evidence itself is considered untrustworthy or untrue.
> 2. Evidence which may itself be true and verifiable is used to deduce a conclusion which does not follow from it, usually by generalizing from an insufficient amount of evidence. For example 'my grandfather smoked like a chimney and died healthy in a car crash at the age of 99' does not disprove the proposition that 'smoking markedly increases the probability of cancer and heart disease at a relatively early age'. In this case the evidence may itself be true, but does not warrant the conclusion.
>
> In both cases the conclusion is unreliable; it might happen not to be untrue, but it doesn't follow from the 'evidence'." [6]

Part of the problem with anecdotal evidence is that we tend to hear only the stories that fit our agenda. In the example of the teacher, those students who are benefiting from the program tend to tell her so. She knows them, since they are going to be the better students. Those students for whom the

program has failed tend to fade away quietly. Their stories are much easier to overlook.

Designed Experiments and Observational Studies

One of the most common uses of statistics is to try to determine cause and effect relationships. Does smoking cause cancer? Does TV contribute to juvenile delinquency? Do high doses of vitamin C ward off colds? Questions like these can sometimes be answered by outlining a mechanism, that is, coming up with a convincing explanation of how cause creates effect. Sometimes no general agreement on a mechanism can be reached among experts. In either case, researchers often turn to statistically based studies to connect cause and effect.
Consider the following two hypothetical studies:

1. A medical researcher wants to know if a high cholesterol diet makes one more prone to type II diabetes. Based on surveys of diet, he identifies a group of 500 people having a high cholesterol diet and 500 people having a low cholesterol diet. He then surveys them about diabetes, and finds that 27 of the people in the high cholesterol group have been diagnosed with type II diabetes, while only 14 people in the low cholesterol group have such a diagnosis. He concludes that a high cholesterol diet increases the risk of diabetes.

2. An auto industry market analyst conducts surveys of automobile owners. Based on his surveys, he identifies 500 people who paid over $50,000 for their automobile, and 500 people who paid under $20,000 for their automobile. He then surveys both groups about their income. He finds that in the group which paid over $50,000 for their automobile, 207 have a six figure income. In the group that paid under $20,000 for their automobile, only 41 have a six figure income. He concludes that buying an expensive car tends to increase the buyers income.

Neither of these examples is real. I made them up. But few people would be surprised to see a study similar to the first one written up in the news and presented as valid. In fact, most of us wouldn't question it. On the other hand, the conclusion of the second study seems obviously invalid and

laughable. But look at the logical structure of both studies. In both cases, we start with a group which is characterized by some hypothetical cause, and another which is characterized by the lack of this cause. We then count the number in each group who exhibit the hypothetical effect and compare the results.

	Cholesterol and Diabetes			Car and Income	
cause	high	low		expensive	inexpensive
	cholesterol	cholesterol		car	car
effect	Diabetes	Diabetes		6 figures	6 figures
	27	14		207	41
risk factor	.054	.028		.414	.082
% increase	93			404	

The logical structure of both arguments is the same. In fact, the numbers seem more striking in the second study. The risk factor is the fraction of those in the cause group who exhibit the effect. For example, 27 of the 500 people in the high cholesterol group have diabetes, so the risk factor is

$$\frac{27}{500} = .054$$

The percent increase between the low cholesterol and high cholesterol groups is

$$\frac{.054 - .028}{.028} \approx .93 = 93\%$$

The increase in "risk" factor in the second study is 404%, much greater than in the first. Yet most folks would find the first study more credible than the second. What's going on here?

Both of these hypothetical studies are examples of *observational studies*. In an observational study, the researchers do not introduce the hypothetical cause, but rather select a sample which already exhibits that cause. The folks in the high cholesterol group were not asked by the researchers to eat a high cholesterol diet. The chose that diet themselves. Similarly, no one asked the expensive automobile group to pay that much for their car, they elected to do so on their own. The researchers selected those individuals because of the choices they had already made. This allows room for self-selection bias in the sample. It's easy to see that in the second study. When we select people

who have chosen to purchase a \$50,000+ car, we are automatically choosing a more affluent group. People on modest incomes can't afford that kind of car. Is there self-selection bias in the first study? Though it's nothing quite as obvious as in the second study, we can't really be sure there is none. It could be that people who eat a high cholesterol diet are generally less health conscious. Unfortunately, the fact that we are not able to identify any clear self- selection bias does not imply that none exists. Sample bias can be very subtle. All observational studies admit the possibility of self-selection bias in the sample.

So how can we do statistically based cause/effect research which avoids this kind of self-selection bias? Sometimes, we can do a *designed experiment*. In a designed experiment, the researchers do not select sample groups who already exhibit the hypothetical cause. Instead, they choose a group which is designed to be as representative as possible of the population, and then the researchers introduce the cause themselves. If we wanted to do a designed experiment to determine whether driving an expensive car results in an increase in income, we might choose two groups of people at random, give everyone in one group a \$50,000+ car, and give everyone in the other group a $\$20,000-$ car. Then we would keep track of any change in average income between the two groups. If we find after a few years, that the \$50,000+ group has seen a significantly higher increase in income, then we would have a pretty good argument. Since the groups were chosen at random, and the cars were given to them, there should be no self-selection bias muddying things up.

Can we do a designed experiment that would tell us whether high cholesterol leads to diabetes? We would need to select two groups of people at random, then ask one to adopt a high cholesterol diet, and the other to adopt a low cholesterol diet. We run into problems with this sort of thing. Do we have the right to ask people to adopt a high cholesterol diet, knowing that such a diet would adversely affect their health? How many years are we going to need to wait for results? There are many situations, particularly in the biomedical fields and the social sciences, where ethical and practical considerations prohibit doing proper designed experiments. In these situations, observational studies may be the best we can do. Observational studies are never as convincing as a properly designed experiment, but in situations where the need for answers is great, they are considered better than nothing. A single well designed experiment is often considered reasonably convincing. A single observational study, on the other hand, is seldom considered conclusive. Increases in risk factors in observational studies are also expected to

be high. The 93% increase in risk factor in the first example would not be considered high enough to warrant drawing any conclusions. Increases in risk factor for this type of study do not begin to become significant until we get to 200% or better. [9] (Note that a 200% increase would mean that 3 times as many people have the condition.) The cause/effect relationship between smoking and lung cancer is an example of one that has been well established by observational studies. While no one study can be considered absolutely conclusive, the sheer number of studies which come to the same conclusion, and the increases in risk factor that they report (2000% or better) leave little room for doubt.

The biggest problem with observational studies may be the way that they are presented in the popular media. The media seldom calls any attention to the distinction between designed experiments and observational studies, and tends to treat every study as if it were a "smoking gun," even when results are not statistically significant. Experts know that we should not draw conclusions from any single observational study, but rather look at *all* the research that has been done on a problem. Even then, sometimes the research is either inadequate or contradictory, and no conclusions can be drawn.

Suggestions For Further Reading

There is much more that can be said on the subject of statistical literacy. All I have attempted to do here is to get you thinking about the differences in good and bad statistics. Some excellent and quite readable books have been written on the subject. Dr. Joel Best's *Damned Lies and Statistics*[8] and *More Damned Lies and Statistics*[9] are highly recommended. Darell Huff's *How to Lie With Statistics*[10] is a classic. Although some of the examples are a bit dated, it is still quite entertaining and informative. *A Mathematician Reads the Newspaper*[11] by John Allen Paulos is another good one. All of these are relatively inexpensive and highly recommended.

References

[1] *The American Heritage Dictionary of the English Language*, 4th ed., Houghton Mifflin Company, 2004. <Dictionary.com http://dictionary.reference.com/browse/advocate>.

[2] *CRHRE's Weekly Mental Health Bulletin*: *Undiagnosed Mental Illness*, Center for Rural Health Research & Education, College of Health Sciences, University of Wyoming. http://www.uwyo.edu/health/MH_Bulletins/(5)Undiagnosed.htm.

[3] Dr. G. Terry Madonna and Dr. Michael Young, *Politically Uncorrected Column*: *The First Political Poll*, Center for Politics & Public Affairs, Franklin & Marshall College, June 18, 2002. http://www.fandm.edu/x3905.xml.

[4] David S. Moore et. al., *For All Practical Purposes*: *Mathematical Literacy in Today's World*, W. H. Freeman & Company, COMAP – Consortium for Mathematics and its Applications, 2003.

[5] *Statistical Standards*: *Processing and Editing of Data: Nonresponse Bias Analysis*, National Center for Education Statistics U. S. Department of Education. http://nces.ed.gov/StatProg/2002/std4. 4.asp.

[6] *Anecdotal Evidence*, Wikipedia, the Free Encyclopedia. http://en.wikipedia.org/wiki/Anecdotalevidence.

[7] Michael Greenland, *Special Report: Questionable Future*: *Declining Response Rate, Rising Costs*, National Science Foundation. http://www.nsf.gov/news/special reports/survey/index.jsp?id=question.

[8] Joel Best, *Damned Lies and Statistics*: *Untangling Numbers From the Media, Politicians, and Activists*, University of California Press, 2001.

[9] Joel Best, *More Damned Lies and Statistics*: *How Numbers Confuse Public Issues*, University of California Press, 2004.

[10] Darell Huff, *How to Lie With Statistics*, W. W. Norton & Company, 1954.

[11] John Allen Paulos, *A Mathematician Reads the Newspaper*, Anchor Books, 1997.