(Mis)reporting of Data

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 17, 2013

Beginning Quotations

I gather, young man, that you wish to be a Member of Parliament. The first lesson that you must learn is, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister. That is a political statistic.
– Winston Churchill

It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
– Sir Arthur Conan Doyle, *A Scandal in Bohemia* (1892)

A source's sincerity is no guarantee of a number's accuracy.
– Joel Best (*Stat-Spotting*)

I have here in my hand a list of 205 – a list of names that were made known to the Secretary of State as being members of the Communist party and who nevertheless are still working and shaping policy in the State Department.
– Joseph McCarthy

Be precise. A lack of precision is dangerous when the margin of error is small.
– Donald Rumsfeld

Week 9: (Mis)reporting of Data

Required Reading:
SGEP (283–305) —
10.1 The Social Construction of Statistics
10.2 Adjustments for Groups Not Comparable On a Variable, Such As Age

Suggested Reading:
10.0.3 Appendix: P. Lorillard Co. v. Federal Trade Commission (Court of Appeals Fourth Circuit; Decided: December 29, 1950)

Film: *The Manchurian Candidate* (127 minutes)

Introduction

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest*.

One recent issue was from Gerd Gigerenzer and colleagues, entitled "Helping Doctors and Patients Make Sense of Health Statistics".

It discusses aspects of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally.

Some parts of being statistically literate may be fairly obvious;

we know that just making up data, or suppressing information even of supposed outliers without comment, is unethical.

The topics touched upon by Gigerenzer et al. (2007), however, are more subtle.

If an overall admonition is needed, it is that context is always important, and the way data and information are presented is absolutely crucial to an ability to reason appropriately and act accordingly.

Rudy Guiliani Mistake

We begin with a quotation from Rudy Guiliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican presidential nomination:

I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it—in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine.

Not only did Guiliani not receive the Republican presidential nomination, he was just plain wrong on survival chances for prostate cancer.

The problem is a confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

Survival versus Mortality

To give a more detailed explanation, we define a five-year survival rate and an annual mortality rate:

five-year survival rate = (number of diagnosed patients alive after five years)/(number of diagnosed patients);

annual mortality rate = (number of people who die from a disease over one year)/(number in the group).

The inflation of a five-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed.

Moreover, such screening, particularly for cancers such as prostate, leads to an *overdiagnosis bias*, the detection of a pseudodisease that will never progress to cause symptoms in a patient's lifetime.

Besides inflating five-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (for example, incontinence, impotence, and other health-related problems).

Screening

Screening does not "prevent cancer," and early detection does not prevent the risk of getting cancer.

One can only hope that cancer is caught, either by screening or other symptoms, at an early enough stage to help.

It is also relevant to remember that more invasive treatments are not automatically more effective.

A recent and informative summary of the dismal state and circumstances surrounding cancer screening generally, appeared in the *New York Times* as a "page one and above the fold" article by Natasha Singer (July 16, 2009), "In Push for Cancer Screening, Limited Benefits."

Health Statistics Reporting

A major area of concern in the clarity of reporting health statistics is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter. We give examples that present the same information:

Relative risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced by about one third over the next ten years.

Absolute risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced from 3 in 1000 to 2 in 1000, over the next ten years.

A useful variant on absolute risk reduction is given by its reciprocal, the *number needed to treat*;

if 1000 people have this test every two years, one person will be saved from dying from the disease every ten years.

In addition to the use of relative and absolute risk, or the number needed to treat, a fourth way of presenting benefit would be as an increase in life expectancy.

For example, one might say that women who participate in screening from the ages of 50 to 69 increase their life expectancy by an average of 12 days.

This is misleading in terms of a benefit to any one particular individual;

it is much more of an all-or-nothing situation, like a lottery.

Nobody who plays a lottery gains the expected payout;

you either win it all or not.

Mismatched Framing

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm.

It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically.

The latter has been referred to as "mismatched framing" by Gigerenzer and colleagues.

Transparent Framing

An ethical presentation of information avoids nontransparent framing of information, whether intentional or unintentional.

Intentional efforts to manipulate or persuade people are particularly destructive, and unethical by definition.

As Tversky and Kahneman have noted many times, framing effects and context have major influences on a person's decision processes.

Whenever possible, give measures that have operational meanings with respect to the sample at hand (for example, the Goodman–Kruskal γ, the median or the mode, the interquartile range) and avoid measures that do not, such as the odds ratio.

Inflated Statistics

In a framework of misreporting data, we have the all-to-common occurrence of inflated and sensational statistics intended to have some type of dramatic effect.

As noted succinctly by Joel Best in his 2005 *Statistical Science* article, "Lies, Calculations and Constructions," "Ridiculous statistics live on, long after they've been thoroughly debunked; they are harder to kill than vampires".

We might see a three-stage process in the use of inflated statistics:

first, there is a tale of atrocity (think Roman Polanski's 1968 movie, *Rosemary's Baby*);

the problem is then given a name (for example, the presence of satanic cults in our midst);

and finally, an inflated and, most likely, incorrect statistic is given that is intended to alarm (for example, there are well over 150,000 active satanic cults throughout the United States and Canada).

Remember that when a statement seems counterintuitive and nonsensical, it probably is.

Context

Another issue in the reporting of data is when the context for a statement is important but is just not given (or is suppressed), resulting in a misinterpretation, or at least, an overinterpretation.

These examples are legion and many are summarized in the required reading. We give three exemplars here:

1) A recent article posted on the MSNBC website had the title, "1 in 5 US Moms Have Kids With Multiple Dads, Study Says," with the teaser line: "Poverty, lack of education and divorce perpetuate lack of opportunities".

The first short paragraph below leaves one with the question: can the child of a mother with only one child have multiple fathers?:

[O]ne in five of all American moms have kids who have different birth fathers, a new study shows. And when researchers look only at moms with two or more kids, that figure is even higher: 28 percent have kids with at least two different men.

2) In 2004, President Bush boasted that the average tax cut was $1,586; this was an arithmetical average so it was considerably inflated by many large numbers for the very wealthy.

A more transparent measure would have been the median of $470, where half of individuals and families got more of a cut than this, but half got less.

3) A "page one" article in the *New York Times* by Sam Roberts (January 16, 2007), had the eye-catching title "51% of Women Are Now Living Without Spouse."

Its first line reads: "For what experts say is probably the first time, more American women are living without a husband than with one, according to the *New York Times* analysis of census results."

A month later (February 11, 2007), the Public Editor of the *Times*, Byron Calame, had a column entitled "Can a 15-Year-Old Be a 'Woman Without a Spouse?'"

Part of the second paragraph reads: "But the new majority materialized only because *The Times* chose to use survey data that counted, as spouseless women, teenagers 15 through 17—almost 90 percent of whom were living with their parents."

Astroturfing and Push-Polling

Because of the importance of context and framing in presenting and interpreting data, it is important to avoid falling prey to the cliché that "the data speak for themselves."

This is never true;

data are constructed by a process requiring many decisions as to what is observed and presented.

This can be a benign procedure when the definitional boundaries are clear.

Sometimes, however, data collection is used with a more nefarious intent, with two of the most blatant named "astroturfing" and "push polling.

As defined in an OpEd by Ryan Sager in the *New York Times* (August 18, 2009), "Keep Off the Astroturf," "astroturfing" refers to the simulation of a (large) "grass-roots" movement by manufacturing a large number of responses.

Thus, whenever there are data given in the form of the number of, say, "concerned citizens," this can be highly inflated.

A "push poll" is a survey in which the data (that is, the responses) are not of any real interest. Questions are posed to "push" the respondent in a particular direction.

A good example of a push poll is given by Richard H. Davis in a *Boston Globe* article (March 21, 2004), entitled "The Anatomy of a Smear Campaign."

Davis was Senator John McCain's presidential campaign manager in 2000 when McCain was running against Bush in the Republican primaries.

During the contentious (and decisive) South Carolina contest, push polling was used to imply that McCain's adopted Bangladeshi-born daughter was his own illegitimate black child.

The question posed to McCain supporters was whether they would be more or less likely to vote for McCain if they knew he had fathered an illegitimate black child.

As we know, Bush went on to win the South Carolina primary, and ultimately the Presidency.

The Social Construction of Statistics

In a series of three entertaining books, Joel Best has made the point that most statistics presented to the public are socially constructed, and therefore, possibly not a perfect reflection of reality.

A critical consumer has to know why and by whom the numbers were generated and in what context they are being used.

So yet again, context and framing are crucial to how numbers should be interpreted.

The titles of these three books are given below along with their complete subtitles:

Damned Lies and Statistics: Untangling Numbers From the Media, Politicians, and Activists (2001)

More Damned Lies and Statistics: How Numbers Confuse Public Issues (2004)

Stat-Spotting: A Field Guide to Identifying Dubious Data (2008)

The Dark Figure

Best offers a variety of useful observations as to how we might go about being intelligent consumers of the numerical information that is continually provided by the media.

Criminologists, for example, have used a phrase, "the dark figure," to refer to the number of crimes that don't appear in official statistics.

More generally, every social problem has a dark figure of some sort defined by the difference between officially recorded incidents and the true number.

It may be big (as in reporting prostitution), or small (as in reporting homicides), but it exists for most social problems that come to mind.

Once a dark figure, such as declaring the number of homeless in the United States to be so many million, has been estimated and made public, that number goes through a "number laundering" process where the best guess source is now forgotten, and the number is treated as a fact.

People who repeat or create the "fact" may even come to have a stake in defending it.

Also, any estimate can be defended just by impugning the motives of someone disputing the figure.

Besides giving high-end guesses of a dark figure to promote attention to whatever social problems are being addressed, the use of especially broad definitions supports much larger estimates of a problem's size.

Depending on how instances of such things as "bullying," "hunger," or "mental illness" are defined, very different estimates of a dark figure could be generated.

Depending on how the boundaries of a vague term are delineated, the numbers then attached are open to differing interpretations as to severity.

Other instances of equivocation in definition are generally problematic as well.

For example, questions have been raised about a current epidemic of autism among children, and whether this might be attributable to environmental contaminants or vaccination.

It may be that the definition of what is considered autistic behavior has just been broadened to include an "Autism Spectrum," a term now used to describe pervasive developmental disorders, which include among others, Autistic Disorder, Asperger's Disorder, Childhood Disintegrative Disorder, Rett's Disorder, and the catch-all term, Pervasive Development Disorder Not Otherwise Specified.

What Population is Surveyed

It is important to know the population surveyed if appropriate interpretations are to be made of the responses.

The randomness of the chosen sample is a hoped-for ideal that is rarely achieved.

Even when we try to approximate randomness by, say, the random-digit dialing of landlines, the influence of high cell-phone usage in an area is unknown.

Asking about homelessness at bus stations or estimating one's electoral chances based only on the enthusiasm of the crowds at political rallies may suggest obvious biases.

But subtler things can happen that remain hidden.

Generally, the pseudo-randomness of a sample is more important than sample size, and the use of Internet-based surveys would seem dubious indeed.

Also, a large sample size may give a false impression of the accuracy of the data.

Who cares about a margin of error that is plus or minus three percentage points if you are surveying the wrong crowd?

In short, good statistics are based on more than guessing (for example, of the dark figure).

Clear and reasonable definitions that are not overly broad are necessary, as well as clear and reasonable measures with appropriate wording of questions that doesn't direct answers in a particular way. Finally, good representative samples are crucial to extrapolation and interpretation.

Joel Best's Names for Various Statistics

In Joel Best's discussion of statistical information of which one should be wary, he provides a number of clever names.

A *mutant* statistic is one mangled beyond recognition from its original form.

Best recalls a dissertation proposal reporting that "every year since 1950, the number of American children gunned down has doubled". Thus, if one child were gunned down in 1950, we would be up to 35 million in 1995 and one trillion by 2014—not very plausible given the earth's population.

Possibly the following statement might have been intended: "the number of American children killed each year by guns has doubled since 1950."

Another category of bad statistics are "missing numbers," referring to what has not been counted.

For example, there may be a declaration of an epidemic of school gun violence because of several salient occurrences (remember, the plural of "anecdote" is not "data");

or definitions that specify what is to be counted may be problematic to implement (for example, what characterizes a "missing child," or how does one count an occurrence of "child abuse").

We are reminded of two famous Donald Rumsfeld quotations:

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

Look for what's missing. Many advisors can tell a President how to improve what's proposed or what's gone amiss. Few are able to see what isn't there.

One particularly fraught name that Best uses is that of a "scary number".

Given the incredible needs of the cable and broadcast news media, numbers are continually sought that will command our attention, and thereby allow the industry to sell us more things we probably don't really need.

Risk numbers are given in bigger scary relative values rather than in smaller absolute terms; social problems are given big estimates (for example, inflated dark figures); trends always seem to be troubling; the various scenarios envisioned for whatever is being followed are invariably apocalyptic; and so on.

When we get frightened, the focus is on what scares us rather than on the actual risk. This type of reaction is named "probability neglect" in the judgment and decision-making literature.

Based on the social construction of the numbers we are bombarded with continually, the question to ask is not necessarily "is it true?", but rather, "how and by whom was it produced?"

This admonition applies to simple number misuse (for example, when colleges scam the numbers in the annual "Best Colleges" issue of *US News & World Report*), and to those that have serious national military implications (for example, the bodycount and pacification figures reported in the Vietnam war, or more recently, the counts of weapons of mass destruction in justifying the Iraq War).

Proofiness: The Dark Arts of Mathematical Deception

Charles Seife in his 2010 book, *Proofiness: The Dark Arts of Mathematical Deception*, provides several additional memorable names in pointing out various acts of numerical mendacity.

Seife defines "proofiness," a numerical version of Stephen Colbert's notion of "truthiness", as "the art of using bogus mathematical arguments to prove something that you know in your heart to be true—even when it's not".

One of the first terms Seife introduces is used for fabricated statistics—"Potemkin numbers"—the numerical equivalent of Potemkin villages.

To explain the historical allusion, we quote from Seife:

According to legend, Prince Grigory Potemkin didn't want the empress of Russia to know that a region in the Crimea was a barren wasteland. Potemkin felt he had to convince the empress that the area was thriving and full of life, so he constructed elaborate facades along her route—crudely panted wooden frameworks meant to look like villages and towns from afar. Even though these "Potemkin villages" were completely empty—a closer inspection would reveal them to be mere imitations of villages rather than real ones—they were good enough to fool the empress, who breezed by them without alighting from her carriage.

Another useful term introduced in *Proofiness* is *disestimation*, or the underestimating of the uncertainties associated with most numbers, and generally, taking a number too seriously. We mention three examples of disestimation:

the first, which immediately follows, is a humorous anecdote that Seife presents about a guide at a national history museum;

the second but not so humorous tale, is about the "smoothness" of Old Gold cigarettes;

the third concerns a belief that by continual recounts in contested elections, the "true winner" will eventually be identified.

There's an anecdote about an aging guide at a natural history museum. Every day, the guide gives tours of the exhibits, without fail ending with the most spectacular sight in the museum. It's a skeleton of a fearsome dinosaur—a *tyrannosaurus rex*—that towers high over the wide-eyed tour group. One day, a teenager gestures at the skeleton and asks the guide, "How old is it?"

"Sixty-five million and thirty-eight years old," the guide responds proudly.

"How could you possibly know that?" the teenager shoots back.

"Simple! On the very first day that I started working at the museum, I asked a scientist the very same question. He told me that the skeleton was sixty-five million years old. That was thirty-eight years ago."

The second example that exploits false precision for nefarious purposes occurred in the 1950s with the P. Lorillard Company and its marketing of Old Gold cigarettes.

Two excerpts are given in an appendix that relate to the history of this case.

The first item is a brief summary taken from a book by Susan Wagner, *Cigarette Country: Tobacco in American History and Politics* (1971);

the second is more lengthly and taken directly from the opinion in *P. Lorillard Co. v. Federal Trade Commission* (1950).

The third example of disestimation is the notion of *electile dysfunction* defined as a collective belief that a "true" election winner will finally emerge if only enough recounts are carried out.

Recent examples of this are the 2000 Florida vote in *Bush v. Gore* or the Minnesota Senate election between Al Franken and Norm Coleman.

Any true (but razor-thin) margin of winning will be swamped by errors in absentee ballots; write-ins; machine ((un)readable) failures; and so on.

In short, there is no way to determine a "true" winner in these closely contested elections; possibly, the fairest mechanism to both candidates would be a simple coin toss.

Adjustments for Groups Not Comparable On a Variable, Such As Age

Data may require preliminary adjustment when presented for groups of individuals on events related to a variable such as age (for example, the presence or absence of disease, death, injury, and suicide).

If the groups differ in their distributions on this variable, two types of adjustment are generally possible: direct and indirect.

Relying on age as the variable of interest, direct age adjustment adopts a standard population to eliminate the effects of age differences between the groups to be compared.

A well-known and extensively developed example of age adjustment is given in Mosteller and Tukey's 1977 textbook, *Data Analysis and Regression: A Second Course in Statistics* (pp. 225–229).

It concerns death rates in Maine and South Carolina in 1930 and is a good example where the comparison of overall crude death rates produces misleading results.

South Carolina had higher death rates than Maine in all but one age class (and even here they were close).

Nonetheless, the crude death rate of South Carolina (at 1288.8 per 100,000) made it look better than Maine (at 1390.8 per 100,000).

The reason for this misleading benefit of living in South Carolina (pun intended) is that Maine's population was generally much older than South Carolina's.

Average Age of Death

As another instance where context is crucial, it can be considered a fallacy to compare the average age at which, say, death occurs rather than comparing the risk of death between groups of the same age.

For example, suppose there are two groups of individuals: A: all elementary students in the United States; B: all members of special forces units in the United States military.

The average age of death is much less for group A than B, but that does not imply it is riskier to be an elementary student than a military commando.

Generally, the average age at death doesn't reflect the risk of death but only a characteristic of those who die.

It only looks at those who die and ignores all those who survive.

A similar reasoning anomaly happens when the age at onset of a particular kind of disease (for example, lung cancer in chemical factory workers) is compared to that for the general population.

An average age being lower only implies that factory workers may be younger to begin with.

The comparison needed is of the risk for lung cancer occurrence in groups having the same age.