May 14th, 9:00 AM - May 17th, 5:00 PM

The Rhetoric of Numbers: Statistical Inference as Argumentation

Mark Battersby
Capilano College

Title: The Rhetoric of Numbers: Statistical Inference as Argumentation
Author: Mark Battersby

No one doubts that numeric information can be used to provide good reasons for beliefs and judgments, and no one doubts that the same type of information can be used to mislead, intimidate, and illegitimately persuade. The study of how numeric information does and does not rationally persuade is a major research task that is already being undertaken by psychologists and statisticians (Kahneman, Gigerenzer). Interestingly what this research shows is that many of the ways that numeric information is presented fail to be adequately understood and appreciated by the audience. The rhetorical concern raised by this research is to find ways to communicate numeric information that can be readily understood and used by non-mathematicians. The more common concern of logicians with rhetoric has been the concern that persuasive techniques will lead people to accept beliefs without providing adequate reasons for these beliefs. Given the ubiquitous use of statistical information, in everything from informed medical consent to public policy decision-making, both problems can have significant consequences. As a teacher I am particularly concerned with finding ways to help students make sense of and evaluate statistical information. Such information presented in a credible and intelligible fashion can be of great value. One of the most central uses of statistical methods is inferential statistics. Inferential statistics provide the basis for polling and statistically based scientific research such as sociology, psychology, and epidemiology. While acknowledging the importance and value of such statistical methods, in this paper I argue that the presentation of research and polls based on statistical methodology is often misleading. I am not arguing that such research be ignored or dismissed, but rather that the claims emerging from such research be viewed as conclusions of informal (i.e. not statistical) arguments.

My basic assumption is that for a contemporary educated audience, numbers can speak louder than words. This means that the proper presentation of numeric information can often be more effective than arguments presented without numbers. It also means, that in those cases where the numeric information does not deserve a great deal of argumentative weight, appropriate caution and qualification needs to be exercised in its presentation. There are many such cases. In particular I will argue that the typical presentation of inferential statistics is flawed and misleading. The air of precision created by the use of concepts such as "margin of error" and "confidence level" is seldom warranted despite the respect that they invite.

The analogy I would like to draw is with the Ad Hominem fallacy. In the real world of argumentation it is almost always useful to know about the biases and motivation of the author of an argument. The problem is that many use the knowledge of an author's motivation or point of view to dismiss or ignore the actual arguments presented by the author. The problem could be characterized by pointing out that *Ad Hominem* remarks frequently have a persuasive (or dismissive) value significantly in

excess of their probative value or logical worth. I will argue that something similar happens with the presentation of numeric information, particularly inferences from samples. The language and conceptual framework of statistical inferences such as sampling, margin of error, statistical significance, confidence levels, and the like are frequently used without the logical (mathematical) pre-conditions for their use. Nonetheless, the conclusions are typically stated with a mathematical precision that usually carries a persuasive force in excess of their epistemological worth. Conclusions of most statistical inferences used in polling and research should be viewed with far less confidence than the numbers claim and suggest: the precision used in expressing confidence intervals and statistical significance is seriously misleading.

In the alternative I suggest that the proper way to view statistical inference is not as a mathematical inference but as part of an informal inductive argument. I will sketch the form of this argument and provide some illustrative examples.

While much of my paper is critical of the presentation of statistical information, I wish to make it clear that I am not critical of the use of statistics and experimental methodology as a means for coming to a well grounded understanding of our world My concern is with the undue rhetorical force that the presentation of such information typically carries.

Before proceeding I need to make a brief and simplified description of the logical basis of statistical inference. Sampling and the inferences made from samples are generally based on the assumption that the samples are random, meaning that every item or person in the population being sampled has an equal chance of being selected for the sample. If the sampling process does not guarantee equal chance of selection for all members of the population, then the process is biased and there is no *mathematical* basis for making the kind of inferences that are typically stated. Given a random sample and certain assumptions about the distribution of the population, probabilistic inferences can be made about the likelihood that a sample statistic is close to that of the actual value in the population. The results of such statistical inferences are expressed in terms of how likely (the so-called confidence level) the sample statistic is within a margin of error (±) of the population value.

Note that the prerequisite for this probabilistic reasoning is that the sample should be a *random sample* of the target population, not, as is often stated, that the sample should be "representative" of the population. The latter concept has a kind of intuitive attraction until you realize that it is impossible to say what a representative sample is unless it is a random sample. The concept of representativeness is based on the assumption that we can identify those properties a person in the sample possesses that count towards representativeness (e.g. gender, income, geographic location, eating habits). The claim of representativeness also assumes that we know the rate of people in the population who have these properties and can therefore check if the sample is representative, i.e. we can check if our sample has the approximately the same proportion of men and women, rich and poor etc. as in the population. Key problems with "representativeness" are that we don't know which properties are the relevant ones to use to determine "representativeness" and, in many cases, we don't know the actual proportions in the population. Not that we can't make reasonable claims about these issues, but however credible the claim for representativeness, a representative sample is not the random sample required as the basis for the statistical inference. A case can be made for the "representativeness" of a sample, and such cases are often made by pollsters and less frequently by researchers, but this case needs to form part of the argument for

any generalization based on the sample.

Unfortunately, pollsters and researchers typically treat the inference from a sample to the generalization about the population as a kind of mathematical deduction as follows:

1. Results of our sample of size X (typically around 1000 in national polls) is S (the so called statistic, e.g. "70% of the sample expressed support for Kyoto").

Therefore, (according to statistical theory) there is a 95% chance (we can be 95% confidant) that the population parameter, P, is S ± 3.1 percentage points. (P being the value that in theory would be obtained if all members of the population were surveyed.)

But this won't do. Samples are never truly random and this is well understood by pollsters. The qualifications regarding the sampling process should be part of the argument. Responsible pollsters often acknowledge (frequently in a footnote) the inappropriateness of such mathematical precision. For example, the Harris pollsters in the US append the following footnote to their polls:

> In theory, with a probability sample of this size, one can say with 95 percent certainty that the results have a statistical precision of plus or minus 3 percentage points of what they would be if the entire adult population had been polled with complete accuracy. Unfortunately, there are several other possible sources of error in all polls or surveys that are probably more serious than theoretical calculations of sampling error. They include refusals to be interviewed (non-response), question wording and question order, interviewer bias, weighting by demographic control data and screening (e.g., for likely voters). It is impossible to quantify the errors that may result from these" (http://www.harrisinteractive.com/harris_poll/index.asp?PID=309).

Note " impossible to quantify". True, but an informal, not quantified argument can be made that the sampling process produces a survey that is likely biased in certain way(s). For example, studies have been done that try to determine the biases introduced by non-responders (now there is a challenge!), and certainly studies can be made of people who don't have phone (Moore). There is also considerable information about the effects of question wording and question order, and of course some effort is made by pollsters to guard against easily dealt with sources of bias such as question order. Pollsters also make other adjustments that supposedly account for the non-randomness of their sample. But, as Harris admits (above), this is not statistics. To varying extents, pollsters take these issues and biases into account, but when they report their results they seldom include any arguments or even explain their efforts to adjust for "polling bias."

There is one well known situation in which pollsters make efforts to adjust their results in view of the difficulty they have in sampling their target population. National elections provide a kind of "gold test" of polling techniques. Pollsters make considerable effort to identify and poll only voters, and to adjust for other sources of bias in their polling. Despite these efforts, the results of presidential election polling published by Gallop (see appendix) suggest a much higher margin of error or much lower level of confidence than pollsters typically claim. About a third of Gallop's predictions were

outside the +/- 3% margin of error he claimed. These errors occur despite the fact that these polls are often of much larger samples and "adjusted" for representativeness by pollsters (Wheeler 142-143).

Since it is impossible to quantify the biases identified by Harris, the argument for the conclusion should make limited and cautious use of numeric information. The argument might look like the following:

1. Results of sample of size X (typically around 1000) is *S* (the so called statistic, e.g. 70% of the sample expressed support for Kyoto)
2. The polling techniques were as follows:……..
3. The reason to believe that this sample is close to what a genuine random sample of this size would have been (i.e. the reason to believe that this sample is more or less *representative* of the target population) is …

Therefore, there is a reasonable chance that the population parameter *P* is pretty close (though not better than ± 3.1 percentage points) the sample percentage *S*.

If such candour and transparency were common, pollsters might simply acknowledge that the target population of their polling is not all citizens or adults, but rather the group of people who have phones, answer their phones, speak the pollsters' language and are willing to answer their questions. It is unlikely that this target population is "representative" of the more general population, so there is a clear bias built into such a sample. Pollsters could acknowledge this problem, but argue that since the same polling techniques are used from survey to survey, polls do a good job of tracking over time the attitudes of this particular sub-population of the general populace. Such an argument is perhaps a bit cynical, but at least it is not deceptive.

While the confidence and or precision that pollsters claim for the conclusions/generalizations of their "arguments" are generally overstated, their generalizations are undoubtedly more trustworthy than either anecdotal evidence or those polls generated by self-selected samples (e.g. write in, phone in, or now "click in" surveys). In most of these cases there is not even a *prima facie* case for the claim that a sample of self-selected respondents is a random or "representative," sample and absolutely no basis for even *alluding to* the standard statistical methods and inference. I don't wish to overstate this standard dismissal of self-selected samples. Given the difficulties in getting random and unbiased samples using standard polling techniques, the sharp line usually drawn between polling techniques that preclude self-selection and those that allow for it is perhaps exaggerated. Take a personnel "climate survey" of a small company done by a mail out and request for response. Suppose that 120 of 180 employees respond. Their response is of course self-selected and almost sure to be biased in ways that are difficult to determine. Will the discontented respond disproportionately or will those who are happy respond in greater numbers? Hard to say and the use of the statistical concepts of margin of error and confidence level would clearly be inappropriate. But if efforts are made to ascertain whether the respondents are "representative" in terms, for example, of distribution throughout the company divisions, then non-statistical arguments could be made that the proportions in those replying were likely representative of the staff as a whole.

While pollsters present their "arguments" and generalizations with misleading precision they are still relatively clear about their target population. Such is seldom the

case with academic research.

The Problem Of The Uncertain Target Population

As most readers know, there are basically three ways to study humans: case studies, cohort studies, and experimental studies. In case studies, researchers isolate individuals to be studied initially on the basis of their having a symptom such as blood clots or lung cancer or violent behaviour. They then compare this group to another group (usually in the same hospital or institution but without the same symptoms). The comparison group is matched on the basis of a variety of factors depending on the nature of the study such as age, lifestyle, and economic background. The researchers then compare the two groups looking for differences in past behavior or conditions of the two groups that correlate with the current illness or behavior. For example, we might look for evidence that the lung cancer group smoked at a higher rate than a group without lung cancer, or, that the women with blood clots showed a higher frequency of birth control pill use, or, that violent criminals watched more violent television.

Results from such studies are fraught with uncertainty. Obviously, they do not involve random samples of any population. In fact the target population of these studies is often obscure. This is not to say that such studies have no value. The case study approach is often of great value, especially when trying to study a condition that is relatively rare, or recently emerging, such as blood clots in young women. But many researchers using the case study method also use mathematical techniques to justify the claim that there is or is not a *statistically significant* difference between e.g. the rate of blood clots among women who take birth control pills and those who don't. Certainly a *prima facie* case could be made for a correlation using this method, but the use of statistical inference which is based on the assumption of random sampling is misleading.

The best the researchers can tell us is that *had the groups been randomly selected from a population*, such differences that exist between the groups would have been statistically significant. Basically what researchers are looking for is a large enough difference between the two groups to provide evidence that a suspected cause such as the birth control pill should be further investigated. Used with this kind of candour and transparency, the arguments would have the appropriate non-formalness consistent with the nature of the case study method.

An interesting historical example of the kind of difficulties involved, and the use of statistics being misleading (in this case misleading to the researchers), was an early study on smoking (Stolley, 1995). In the early fifties, two studies of approximately 600-700 cases of lung cancer were done that compared the rate of smoking among lung cancer victims and a comparison population. How was the research done? By comparing the smoking history of hospital patients. While both studies found a slightly higher rate of smoking among the cancer victims than the comparison group of hospital patients, the differences were not great enough to be statistically significant, i.e. the difference in the rate of smoking between the group with lung cancer and the control group was not greater than that allowed for by the margin of error. In other words, the researchers could not be confidant the difference in rates was not due to chance.

While researchers still suspected there was a relationship between smoking and lung cancer, their study failed to demonstrate it. Why? With the advantage of hindsight we can see the problem. While none of the patients in the "control group" had lung cancer, many of them had illnesses to which we now know smoking contributes (e.g.

heart disease). As a result the control group was not "representative" of the non-lung cancer population—the control group was biased towards smokers. Its members smoked more than the healthy population obscuring the actually dramatic difference between the rates of smoking among people with and without lung cancer. These studies graphically illustrate the potential problems in using samples of convenience rather than truly random samples.

Texts on research methods usually acknowledge that case study results are only preliminary and suggestive, and this is usually noted in the studies themselves. Because such studies can legitimately provide a basis for applying for funding to support more reliable studies, there is a temptation to allow "misrepresentation" of the value of the results to enhance publicity facilitating the acquisition of more research funding.

Because of the limited possibility of studying humans in a randomized experimental controlled study, the most common approach to studying humans is the *cohort study* approach. The term "cohort study" is applied to different kinds of prospective studies. In one approach a group with the putative cause (e.g. smoking) and a "comparable" control group without the cause are followed over time and the incidence of an effect (e.g. lung cancer) is studied. The group without the putative cause is, of course, supposed to be like that with the cause except for the difference in exposure to the suspected cause. The difficulties in setting up such comparison groups are obvious. Without random assignment to control and experimental groups, the use of the usual statistical machinery is not really justified.

Another form of cohort study involves tracking a large group of people over a period of time as in the Harvard Health or Framingham studies. In this kind of study, the researchers follow a very large group of people keeping a record of what they hope are all the relevant details of their lives and then studying the data for correlations. Such studies avoid the problem of setting up comparison groups and because of the large samples involved and long time frame, appear to give credible results.

But such "data mining" is itself fraught with methodological problems. First we must assume, which is seldom argued for, (and often completely implausible) that these large groups are representative of "the" population in general. In what sense, for example, are Harvard graduates, likely be a "representative" sample of any population? Then there is the awkward fact that even if the studies were done on genuine random samples, at the 95% confidence level, it is likely that 1 out of every 20 apparent correlations is due to chance rather than an actual correlation.

While many statisticians warn against the problem of data mining, there is seldom mention of the far more egregious problem that the samples are not random samples of a target population. There is also the well recognized problem of the confounding factor. When the correlation between, for example, exercise and life expectancy is detected, the scientific challenge is to separate out from a constellation of lifestyle choices, the influence of one factor such as exercise -- healthy people tend to eat healthily and exercise. A variety of mathematical techniques have been developed to isolate individual associations. I do not pretend to understand the mathematics involved (and neither, I suspect do most researchers as the work is done by computers), but the lack of random sampling of the target population still means that these results cannot be statistically generalized. Which is the target population? Humans? North Americans? Americans? Harvard Graduates? Men?

It might be thought that most of the problems addressed above can be solved if it is possible to run a proper randomized experiment. Such studies involve the random

assignment of subjects into control and experimental groups with the experimental group receiving exposure to treatment or putative cause. As one commentator puts it:

> Experimental studies are less susceptible to confounding because the investigator determines who is exposed and who is unexposed. In particular, if exposure is allocated randomly and the number of groups or individuals randomized is large then even unrecognised confounding effects become statistically unlikely (Coggon, Ch. 9).

Obviously the commentator is thinking of the confounding effects of "self-selection" among subjects in a cohort study, but while randomization addresses this issue it does not address the more crucial issue of generalizability to a target population. This generalizability needs arguing for, and even if there is a "statistically significant difference" between the two groups studied, one cannot conclude that such a difference would also be true of the target population

Many researchers appear to believe that they have met the need for randomization when they randomly assign subjects to experimental and control groups as is typically done, for example, in the case testing for drug efficacy. But this is randomization of the group (often volunteers) that has agreed to be studied. This is not the same as starting with a random sample of a population, for example the population of the people with a certain illness. If the population the researchers were interested in studying was simply the group of people being actually studied in the experiment, then the margin of errors and confidence levels would be a justifiable indication of whether differences (between the two specific groups being studied) were likely due to chance or the result of the experimental factor such as the drug treatment. But of course, no one is only interested in the people being studied. Those being studied are supposed to be a sample of the greater target population. Without those selected for study being randomly selected there is *no statistical basis* for inferring from "statistically significant" results in the experiment to the same likelihood that these results are true of the (target) population. What the researchers owe the reader is not mathematics but a case, a non-mathematical case, that the groups studied are non-biased in important ways. A case that is often hard to make. As one writer on epidemiology puts it:

> Bias cannot usually be totally eliminated from epidemiological studies. The aim, therefore, must be to keep it to a minimum, to identify those biases that cannot be avoided, to assess their potential impact, and to take this into account when interpreting results. The motto of the epidemiologist could well be "dirty hands but a clean mind" (*manus sordidae, mens pura*) (Coggon Ch. 4).

Perhaps this is why texts on research often emphasize that experiments with statistically significant results still require replication. If the samples were genuinely random and of reasonable size and the results significant enough (not merely statistically significant), then the case for replication could only be based on the possibility of error or bias in areas such as measurement etc.. Without collecting random samples, the best method we have for controlling (by no means eliminating) biased sampling is through replication. This is one reason why carefully done meta-analysis is probably the most reliable method of evaluating claims. By melding together data from credible studies (credible, but still plagued by lack of true randomization) researchers doing meta-analysis

can make a case that the resultant data and inferences are less likely to be a product of biased sampling than any individual study. Less biased, but hardly free of bias. Much research is based on "samples" of convenience, which often means using people in peculiar institutions such as hospitals and universities – groups that are likely to be different than the general population in a variety of ways. Collecting such studies still raises the problem of selection bias. Making numerous studies of college youth all across North America, for example, which is a common modality of research and replication in psychology, should not give one confidence that these studies can claim to have plumbed the human mind.

A particularly controversial example of the problem statistical of inference from an experimental study being generalized to target population was in a study of gender and racial bias among doctors. The study is described in detail below. There were two rhetorical issues. The numeric information was presented in a manner that overstated the putative racial and gender biases of the doctors studied, and the doctors studied were assumed by the media to be a "representative" sample of doctor population. By interpreting the numbers as they did, the media were acting as if the doctors being studied were a random sample of the population of US doctors. While the researchers comments acknowledged the difficulty of generalizing from their data, the announced confidence levels and odds ratios "said" otherwise and the numbers (albeit misunderstood) spoke louder than the words. A kind of reverse of "poisoning the well"-- the qualifying remarks failed to temper the "message" contained in the numbers. [1]

In Feb 1999, a study published in, as they always say, the "highly respected" *New England Journal of Medicine*, alleged that race and sex of a patient influence how physicians manage symptoms of heart disease. While the study focused only on heart disease diagnosis and recommended treatment, papers such as the *New York Times* ran headlines such as "Doctors' bias may affect health care." The following is a brief summary of the study's methodology:

> In a randomized controlled study, Schulman et al. determined how often doctors recommended cardiac catheterization for hypothetical patients with chest pain. At two professional meetings, 720 primary care physicians were shown a videotaped interview with a patient (portrayed by an actor) and given other relevant data (cardiac risk factors and the result of a thallium stress test) and were then asked whether they would recommend catheterization. The investigators developed 18 hypothetical scenarios representing all possible combinations of the following factors: 3 descriptions of chest pain, 2 levels of cardiac risk, and 3 results of thallium stress tests. In order to isolate the influence of race, sex, and age on the physicians' decisions, each scenario was portrayed by eight actors (representing two races, both sexes, and two ages). The investigators then determined how often these "patients" with identical symptoms and medical histories were referred for cardiac catheterization.

The results, as presented by the authors in their abstract were:

> Logistic-regression analysis indicated that women (odds ratio, 0.60; 95 percent confidence interval, 0.4 to 0.9; P=0.02) and blacks (odds ratio, 0.60; 95 percent confidence interval, 0.4 to 0.9; P=0.02) were less likely to be referred for cardiac catheterization than men and whites, respectively. Analysis of race–sex

> interactions showed that black women were significantly less likely to be referred for catheterization than white men (odds ratio, 0.4; 95 percent confidence interval, 0.2 to 0.7; P=0.004).
>
> *Conclusions*
> Our findings suggest that the race and sex of a patient independently influence how physicians manage chest pain (Schulman et al).

The following table from their study summarizes the data:

TABLE 5. PREDICTORS OF REFERRAL FOR CARDIAC CATHETERIZATION.*

MODEL AND VARIABLE	ODDS RATIO (95% CI)†	P VALUE
Race and sex as separate factors		
Sex		
Male	1.0	
Female	0.6 (0.4–0.9)	0.02
Race		
White	1.0	
Black	0.6 (0.4–0.9)	0.02
Interaction of race and sex		
White male	1.0	
Black male	1.0 (0.5–2.1)	0.99
White female	1.0 (0.5–2.1)	>0.99
Black female	0.4 (0.2–0.7)	0.004

*Both models included all experimental factors as covariates, as well as the probability of coronary artery disease as estimated after the results of the stress tests were known. The first analysis included only the main effects. The second analysis explored a race–sex interaction.

†CI denotes confidence interval.

Ironically, the conclusion of the article in contrast to the abstract, is stated cautiously and moderately.

> Our finding that the race and sex of the patient influence the recommendations of physicians independently of other factors may suggest bias on the part of the physicians. (my emphasis)

And they admit that their sample while clearly not random might also not be representative:

> The recruitment of physicians at national meetings of major professional organizations may have resulted in non-representative samples. Physicians who attend professional meetings may be better informed than those who do not attend. Also, the physicians who volunteered for this project may have had a greater interest than others in coronary heart disease.
>
> An admission, but not perhaps a very thoughtful one. It is easy enough to think of

other ways in which this group might not be representative for example race, gender, and economic circumstance (going to conferences cost time and money).

Despite the qualifications in their article, the numbers they report in their tables (and the statements in the abstract – another rhetorical issue) spoke more powerfully. What the table seems to say is that if you are a black, and/or female you will likely be referred for catheterization at 60% the rate of white males (and if you are a black female, at 40% that rate) despite having the same presenting symptoms.

While fully untangling the message of the numbers is a bit complex because of the use (in this case, misleading use) of "odds ratio," even a cursory look at the table suggests that the differential reference to catheterization is primarily associated with being a "black female."

Commentators pointed out (Schwartz), and the authors in a subsequent response acknowledged (Schulman 1999b), that the use of "odds ratio" was rhetorically ill advised. Odds ratios are similar to risk ratios but only if the incidence of what is being studied is relatively small. In this case with most "patients" being referred for catheterization, the "risk" of being referred was extremely high and the odds ratio extremely misleading because it will almost always be read as a risk ratio. Every news reporter that wrote up the study treated the "odds ratio" as a "risk" ratio. In fairness to the reporters, Schulman admitted that the study's use of odds ratio was "potentially misleading." [2]

In a critical article, (Schwarz), a different table summarizing the data is presented. Looking closely at this table makes it even clearer that it is only black women who are referred at a significantly different rate than white men, and that even they are referred at only a slightly lower percentage.

TABLE 1. RATE OF REFERRAL FOR CARDIAC CATHETERIZATION, ODDS OF REFERRAL, ODDS RATIO, AND RISK RATIO ACCORDING TO SEX AND RACE.*

PATIENTS	MEAN REFERRAL RATE	ODDS OF REFERRAL	ODDS RATIO (95% CI)	RISK RATIO (95% CI)
	%			
Four strata				
White men†	90.6	9.6 to 1	1.0	
Black men	90.6	9.6 to 1	1.0 (0.5–2.1)	
White women	90.6	9.6 to 1	1.0 (0.5–2.1)	
Black women	78.8	3.7 to 1	0.4 (0.2–0.7)	0.87 (0.80–0.95)
Aggregate data				
White†	90.6	9.6 to 1	1.0	
Black	84.7	5.5 to 1	0.6 (0.4–0.9)	0.93 (0.89–0.99)
Men†	90.6	9.6 to 1	1.0	
Women	84.7	5.5 to 1	0.6 (0.4–0.9)	0.93 (0.89–0.99)
Overall	87.7	7.1 to 1		

*Referral rates for the four strata were inferred from aggregate rates and odds ratios reported by Schulman et al.[1] The odds of referral were calculated according to the following formula: referral rate÷(100%−referral rate). The risk ratio was calculated as the referral rate for the group in question divided by the referral rate for the reference group. CI denotes confidence interval.

†This was the reference group.

So there were at least three rhetorical difficulties with this report. 1. The choice

of "odds ratio" artificially inflated the appearance of difference in recommendations. 2. The aggregation of women on one hand and blacks on the other was completely misleading since all the differences were actually resulting from the difference in recommendations for black women. 3. There was no reason given for believing that this particular health issue and referral practice was "representative" of treatment approaches and differences in other domains.

Conclusion

Numbers often speak louder than words. The rhetoric of numbers requires a careful presentation of statistical information so that the audience will give numeric information its appropriate argumentative worth. The presentation of statistical inferences should be treated as informal argumentation with explicit acknowledgement of the issues surrounding population and sample selection that limit the applicability and appropriateness of statistical inference. There are of course reasonable inferences to be made from careful research and polls. These inferences are not simply the result of applying of statistical formulae. They require consideration and acknowledgement of the extent to which samples deviate from the mathematical assumption of random selection of a clearly defined population. The arguments that justify inference from a sample to a population should explicitly refer to the variety of non-mathematical considerations involved. Researchers and pollsters should explicitly address the greater uncertainty involved in inferences from non-random sampling methods. They should also provide a clear indication of what population is being studied and sampled. Carefully done polls and statistically based research are often the best means we have for making reasonable claims about the views of populations and causes of social and medical ills. They are useful tools for evaluating causal interventions and can provide a useful check on causal impressions. They are tools of judgment and informal argument, and should not be allowed to create illusory confidence.

Gallup Poll Accuracy Record

Year	Candidates	Final Gallup Survey	Election Result	Gallup Deviation
1996	Clinton	52.0	50.1	+1.9
	Dole	41.0	41.4	-0.4
	Perot	7.0	8.5	-1.5
1992	Clinton	49.0	43.3	**+5.7**
	Bush	37.0	37.7	-0.7
	Perot	14.0	19.0	-5.0
1988	Bush	56.0	53.0	+2.1
	Dukakis	44.0	46.1	-2.1
1984	Reagan	59.0	59.2	-0.2
	Mondale	41.0	40.8	+0.2
1980	Reagan	47.0	50.8	**-3.8**
	Carter	44.0	41.0	+3.0
	Anderson	8.0	6.6	+1.4
	Other	1.0	1.6	-0.6
1976	Carter	48.0	50.1	-2.1
	Ford	49.0	48.1	+0.9
	McCarthy	2.0	0.9	+1.1
	Other	1.0	0.9	+0.1
1972	Nixon	62.0	61.8	+0.2
	McGovern	38.0	38.2	-0.2
1968	Nixon	43.0	43.5	-0.5
	Humphrey	42.0	42.9	-0.9
	Wallace	15.0	13.6	+1.4
1964	Johnson	64.0	61.3	+2.7
	Goldwater	36.0	38.7	-2.7
1960	Kennedy	51.0	50.1	+0.9
	Nixon	49.0	49.9	-0.9
1956	Eisenhower	59.5	57.8	+1.7
	Stevenson	40.5	42.2	-1.7
1952	Eisenhower	51.0	55.4	**-4.4**
	Stevenson	49.0	44.6	+4.4
1948	Truman	44.5	49.5	**-5.0**
	Dewey	49.5	45.1	+4.4
	Wallace	4.0	2.4	+1.6
	Other	2.0	3.0	-1.0
1944	Roosevelt	51.5	53.8	-2.3
	Dewey	48.5	46.2	+2.3
1940	Roosevelt	52.0	55.0	-3.0
	Wilkie	48.0	45.0	+3.0
1936	Roosevelt	55.7	62.5	**-6.8**
	Landon	44.3	37.5	+6.8

Notes

[1] There was an excellent critique published in the NEJM, (Schwartz et al, 1999) and a more popular critique in the *Atlantic Monthly* (Satel, 2001) .

[2] Schulam offers the following odd defence of his report: "Our study hypotheses, as stated in the original grant application, were that blacks would be less likely to be referred for cardiac catheterization than whites and that women would be less likely to be referred than men. Our reporting of the sizes of the main effects of race and sex is therefore consistent with fundamental statistical principles (Schulman et al, 1999b).

References

Coggon, D., Rose, G., Barker, DJP. 1997 *Epidemiology for the Uninitiated*, 4th Ed, London: British Medical Journal Publishing Group at http://bmj.com/collections/epidem/epid.shtml.

Gigerenzer, Gerd. 2002 *Calculated Risks: How to Know When Numbers Deceive You.* Simon and Schuster.

Kahneman, Daniel, Paul Slovic and Amos Tversky (ed) 1982. *Judgment Under Uncertainty.* NY: Cambridge University Press.

Moore, David 1996. *Statistics: Concepts and Controversies.* NY: W.H. Freeman and Co.

Satel, Sally. 2001; "The Indoctrinologists Are Coming - *The Atlantic Monthly* Vol: 287, No. 1: 59-64.;

Schulman KA, Berlin JA, Harless W, et al. 1999a. "The effect of race and sex on physicians' recommendations for cardiac catheterization" *N Engl J Med*;Vol: 340:618-626.

Schulman KA, Berlin JA, Harless W, et al. 1999b. "Race, Sex, and Physicians' Referrals for Cardiac Catheterization" *N Engl J Med*;Vol 341:285-287.

Schwartz, Lisa M. Steven Woloshin, and H. Gilbert Welch. 1999. "Misunderstandings about the Effects of Race and Sex on Physicians' Referrals for Cardiac Catheterization" *N Engl J Med*. Vol 341:279-283.

Stolley, Paul and Tamer Lasky. 1995. *Investigating Disease Patterns: The Science of Epidemiology*. New York, New York: Scientific American Library.

Wheeler, Michael. 1976. *Lies, Damn Lies and Statistics*. NY:Dell Publishing.