Statistical literacy: Thinking critically about statistics

As published in the Inaugural issue of the Journal "Of Significance"
Produced by the Association of Public Data Users
www.apdu.org

Milo Schield

Associate Professor,
Department of Business, Accounting and MIS
Augsburg College, Minneapolis, MN 55454

schield@augsburg.edu
www.augsburg.edu/ppages/schield

STATISTICAL LITERACY: THINKING CRITICALLY ABOUT STATISTICS

Milo Schield, Augsburg College
Department of Business & MIS Minneapolis, MN

Abstract:
Statistical literacy is the ability to read and interpret data: the ability to use statistics as evidence in arguments. Statistical literacy is a competency: the ability to think critically about statistics. This introduction defines statistical literacy as a science of method, compares statistical literacy with traditional statistics and reviews some of the elements in reading and interpreting statistics. It gives more emphasis to observational studies than to experiments and thus to using associations to support claims about causation.

Keywords: Teaching, Epistemology, Critical Thinking, Strength of Belief, Observational Studies

Statistical literacy is a basic skill: the ability to think critically about arguments using statistics as evidence.

Consider the story of two hunters being chased by a bear. [Adapted from David Friedman, (1996)] The first says, "It's hopeless! This bear can run twice as fast as we can." The second, realizing the flaw in the argument says, "No, it's not hopeless! I don't have to outrun this bear. I just have to outrun you!" The truth of this statistic ("twice as fast") does not give strong support for this conclusion ("it's hopeless"). The second hunter was statistically literate; the first hunter wasn't.

A SCIENCE OF METHOD

Statistical literacy is a science of method. The sciences of method study how we think (Rand, 1966). The sciences of method can be classified by their focus (words versus numbers) and by their method (deductive versus inductive).

	METHOD OF REASONING	
FOCUS	Exclusively deductive	Primarily inductive Some deductive
WORDS	Logic	Critical Thinking
NUMBERS	Math, Probability, most Statistics	Statistical Literacy

In deductive reasoning an argument is either valid or invalid. When an argument is valid then the conclusion must be true if the premises are true. Deductive reasoning is called formal reasoning. Most courses in logic, mathematics and probability study deductive logic exclusively.

In inductive reasoning, an argument is judged on a continuum from weak and strong. The stronger the argument, the more reason we have to treat the conclusion as being true – assuming the premises are true. Inductive reasoning is often referred to as informal or practical reasoning. See Kelly's text (1994).

Note that statistical literacy is closely related to traditional statistics. Both cover the same topics: descriptive statistics, models, probability and statistical inference. Both focus on inferences: generalizations, predictions and explanations.

To understand the relation of statistical literacy to traditional statistics we need to see how the differences in method (inductive versus deductive) affect the choice of topics, the questions and the results.

RELATION TO TRADITIONAL STATISTICS

Traditional statistics focuses on deductive arguments using probability, independence and chance to deduce the associated variation. Chance, steady-state independence, is the premise – the cause. Variation, the resulting probability distribution, is the conclusion – the effect. The derivation of the binomial distribution and various sampling distributions are typically deductive. The question is "How likely is this sample statistic *if* due entirely to chance?"

In traditional statistics, predictions and tests are deductive – they involve 100% certainty. In regard to classical confidence intervals: there is 100% certainty that 95% of all 95% confidence intervals obtained from random samples will include the fixed population parameter. In regard to classical hypothesis tests, consider a sample mean located 2 standard errors from the mean of the null distribution. There is 100% certainty that this sample statistic (or one more extreme) will occur in less than 2.5% of all the samples drawn randomly from the null distribution.

Statistical literacy focuses more on inductive arguments. Statistical literacy questions whether chance is the indeterminate cause of an observed variation or whether there is some determinate cause. Here, chance is one of several explanations – chance is not the given as in theoretical statistics. The question is "How likely is this sample statistic *to be* due entirely to chance?"

Statistical literacy focuses more on inductive reasoning. If we obtain a 95% confidence interval from a single random sample, how should we act? How strongly is

one justified in treating the unknown parameter as though it is in this particular interval? Yes, in reality, the fixed population parameter is either inside this interval or it is not. But given our ignorance are we justified *in acting as though* there were a 95% chance that this particular confidence interval included the fixed population parameter?

In hypothesis testing, does a smaller p-value give us more reason to treat the alternate as true? Of course in reality the truth or falsehood of the null and alternate are fixed; there is no classical probability of their being true or false. But in making a decision, do we have greater reason for treating the null as false as the p-value decreases?

1. READING STATISTICS

Statistical literacy focuses on making decisions using statistics as evidence just as reading literacy focuses on using words as evidence. Statistical literacy is a competency just like reading, writing or speaking. Statistical literacy involves two reading skills: comprehension and interpretation. We will examine reading comprehension first and then turn to interpretation.

All too often, statistical illiteracy involves an inability to comprehend what is being read. Nuances of grammar and technical distinctions are simply overlooked. Consider three important distinctions: association versus causation, sample versus population, and the quality of a test versus the predictive power of a test.

ASSOCIATION VERSUS CAUSATION

To be statistically literate, one must be able to distinguish statements of association from statements of causation. All too often terms designating an association (e.g., factor, influence, related, followed, risk, link, etc.) are treated as asserting causation.

Consider these statements from a recent newspaper article. Major studies have found that 'TV violence is a contributing factor to increases in violent crime and antisocial behavior." The scholarly evidence "either demonstrates cumulative effects of violent entertainment or establishes it as a 'risk factor' that contributes to increasing a person's aggressiveness." Although some may assume this technical language proves causation, it simply describes associations. The real issue is how strongly does this evidence support the claim that TV violence is a causal factor.

To be statistically *literate, one must know whether a statement of comparison involves association or causation.* Consider three claims about the results of an observational study:

1. People who weigh more tend to be taller [than those people who weigh less.]
2. Weight is positively associated with height.
3. If you gain weight, you can expect to get taller.

The first statement is obviously an association. The second statement is often misinterpreted as asserting causation. The change in weight is mistakenly viewed as *a physical change within a given subject.* In fact, the change is a mental change: a shift in mental focus from below-average weight people to above-average weight people. The third statement obviously involves causality. From our experience, we recognize that for adults, number 3 is false. But some mistakenly conclude if number 3 is false then number 2 must be false.

Consider three claims about the results of another observational study.

1. Juveniles who watch more TV violence are more likely to exhibit antisocial behavior.
2. TV violence is positively associated with antisocial behavior.
3. If juveniles were to watch less TV violence, they would exhibit less antisocial behavior.

All too many readers mistakenly conclude if #2 is true, then #3 must be true. But the difference between #2 and #3 is the difference between association and causation. In an observational study, the truth of #2 is evidence for the truth of #3; the truth of #2 is not sufficient to prove the truth of #3.

To be statistically literate, one must be able to distinguish 'attributable to' from 'attributed to' or 'caused by'. 'Attributable' means 'could be attributed to' or 'could be caused by.' 'Attributable' does not mean 'attributed to' or 'caused by.' 'Attributable' is association; 'attributed' is causation.

- In 1995, 30% of the heroin usage among female arrestees in Los Angeles, California was *attributable to* their being female (as opposed to male). However, this association does not mean that the usage is caused by gender and is thus fixed.
- In 1995, 86% of the ratio of abortions to pregnancies among unmarried women were *attributable to* being unmarried.. However, this association does not mean that if all unmarried women were to get married, this rate would decrease.

SAMPLE VERSUS POPULATION

To be statistically literate, readers of statistics must be able to distinguish a sample statistic from a population parameter. All too often unwary readers presume that a statistic obtained from a sample is actually a property of the entire population. Consider the claim, "70% of adult Minnesotans oppose medically assisted suicide". We may interpret the term "adult Minnesotans" as meaning 'all adult Minnesotans.' But unless this survey was part of a census, the data was obtained from a much smaller sample. The term 'adult Minnesotans'

should be restated as 'adult Minnesotans who responded to this poll'.

To be statistically literate, the readers of statistics must be able to distinguish between the target population (the population of interest) and the sampled population (the population from which the sample was obtained). If the target population is difficult to locate (e.g., the homeless) then the data is often obtained from a related population (e.g., people in shelters). If some members of the target population refuse to participate, then the sampled population is only that subset of the population of interest who agree to participate in surveys.

QUALITY VERSUS POWER OF A TEST

To be statistically literate, readers of statistics must be able to distinguish the quality of a test from the predictive power of a test. The quality of a test is measured on subjects whose disease status is known prior to the test; the predictive power of a test is measured on subjects whose disease status is unknown prior to the test. A test may be of good quality (99% of the diseased subjects tested positive; 99% of the disease-free subjects tested negative). But when used to predict a rare condition or disease (a 1% prevalence), this same test may have poor predictive power (only 50% of the positive-test subjects had the disease.) All too often those who lack statistical literacy presume that the quality of a test (99% of the diseased tested positive) is the predictive power of the test (99% of the positives will have the disease).

The inability to distinguish quality from power reflects a deeper problem involving percentages: the inability to distinguish part from whole. The percentage of diseased who test positive is not the same as the percentage of test-positives who are diseased. Identifying part and whole can be quite elusive: the percentage of runners among females is not the same as the percentage of runners who are female. Reading and comparing percentages and rates is a part of statistical literacy that is too often ignored.

2. INTERPRETING STATISTICS

When those who are statistically illiterate misread a statistic, they tend to blame the statistic and not themselves. Most statistics are true; most statistics are not lies. When one misreads the meaning of a statistical claim, it is most tempting to say that statistics lie. But that judgment is often misplaced. All too often, it is the reader who misinterpreted what the statistic meant.

To be statistically literate, one must be able to interpret what a statistic means. Interpretation often involves asking good questions.

To be statistically literate, one must first ask" Is this statistic true?" In some cases, it is a simple error. The 1994 U.S. Statistical Abstract showed the birth rate among unmarried black women was 189.5 per 1,000 in 1991: more that twice the rate for 1990 and 1992. The 1991 rate was in error; it was subsequently shown as 89.5 per 1,000. In other cases, statistics have been 'manufactured' – all too often by those who believe their end justifies such means. Some statistics are presented in a misleading fashion. For more insight, see books by Campbell, Hooke, Huff, Jaffe and Paulos.

To be statistically literate, one must then ask" Is this statistic representative?" In some cases, a true statistic has been selected just because it supports a particular claim – not because is it representative. Thus, if someone wants to support the claim that raising highway speed limits will cause more deaths and select only those states in which this was the case (and ignore those states having the opposite experience) their statistic will be factual and true. But their sample is unrepresentative of the population, so the sample statistic may not be close to the population parameter. In other cases, the convenience sample that self-selected may be unrepresentative. Recall the Ann Landers column (23 January, 1976) that reported the results of a reader write-in survey: 70% of parents say "Kids not worth it" – if they could do it over again. The 70% is factual and true of those who chose to respond. But is it representative of all parents?

To be statistically literate, one must be able to distinguish whether a statistic is factual or inferential. A factual statistic may be false, but its truth-value is not very disputable (in a particular context). The truth-value of an inferential statistic is very disputable. Inferential statistics include predictions, generalizations and explanations. In the Ann Landers survey, the statistic (70%) was factual – for the self-selected sample. But for the entire population, that same statistic (70%) is inferential and in this case highly disputable.

THE QUALITY OF A STUDY

To be statistically literate, one must be able to distinguish an observational study from an experiment. In an experiment, the researcher has effective physical control over which subjects receive the treatment; in an observational study, the researcher has no physical control over who receives the treatment. Those who are statistically illiterate may mistakenly presume a study is an experiment if it involves any kind of treatment, if it involves a control group, or if it involves measurements that are objective. They may mistakenly presume a study is an observational study if it involves a survey, if it lacks a control group or if it involves measurements that are subjective (a self-report of things that are unobservable such as one's feelings or values).

To be statistically literate, one must be able to distinguish a good experiment from a bad one. When they are told the subjects were randomly assigned to the treatment and control groups (as in a clinical trial), readers may mistakenly conclude this study must be an experiment: a good experiment. But if the subjects in this study have knowledge of the treatment then their informed behavior may transform a good experiment into a bad one.

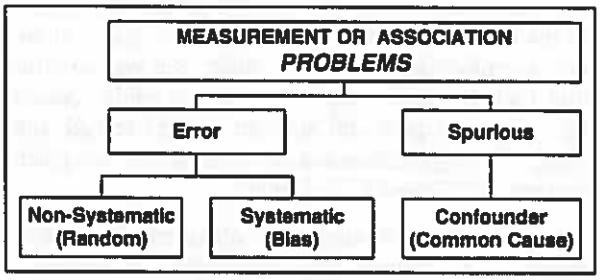
For example, consider an experiment that indicated magnets decrease pain. Fifty subjects having pain associated with post-polio syndrome were randomly assigned to two groups: the treatment group received concentric magnets; the controls received inert placebo 'magnets'. A major decrease in pain was reported by 75% of those in the treatment group -- 19% in the control group. [Natural Health, August, 1998, page 52.] How strongly does this result of this study support the claim that magnets decrease pain?

A statistically literate analyst would investigate the possibility of bias introduced by the Hawthorne effect: the change in behavior in those subjects who were aware of receiving the treatment. Could these subjects have detected whether they had a magnet or not? And if the researchers weren't double blinded about which subjects received the real magnets, could the researchers have inadvertently communicated their knowledge to the subjects? If the researchers weren't double-blinded, perhaps there was bias from the halo effect: seeing what the researcher wants to see. Perhaps the researchers inadvertently allowed their knowledge of whether or not the subject had a magnet to 'push' a subject's borderline response into the desired category.

Consider the quality of another experiment. A group of homeless adults were randomly assigned to either an in-patient program or an outpatient program. Obviously the subjects knew of the treatment. Their informed behavior may generate a Hawthorne effect: a modification of the subject's behavior owing to their awareness of some aspect of the treatment. In this case, those homeless who were assigned to the in-patient program were less likely to enter the program than those who were assigned to the outpatient program. A differential non-response in participation can create an associated bias in the results. And even if the same percentage failed to show up in each group, their informed knowledge of which group they were in may create a non-response bias in the observed results. This experiment may have been seriously compromised by the informed non-response.

PROBLEMS IN MEASUREMENT

To be statistically literate, one must know the various sources of problems in interpreting a measurement or an association. The first problem is error; the second problem is that of being spurious. A single measurement or association may involve both problems.



If there is error, it may be either systematic or random. Random error is often due to such a large number of small determinate causes that they are viewed collectively as being indeterminate. In flipping a fair coin, I may get 80% heads: 4 heads in 5 tries. The expected percentage is 50%; the difference (the random error) is due to a large number of small determinate causes (how the coin was flipped, height of the flip, etc.). We can minimize the influence of chance by taking the average of a larger number of measurements – by getting a larger sample.

Systematic error is due to bias: measurement bias and response bias. Examples of measurement bias include sampling from an unrepresentative subset of the target population and bad measuring instruments (e.g., bad questions in a survey). Examples of response bias include non-response bias (from those who chose not to respond) and non-truthful responses (from those do respond). Non-response is a growing problem in phone surveys as more people screen calls using caller-ID or answering machines. Evaluating any bias due to non-response is a critical element in presuming the statistics obtained from the sampled population are similar to the parameters of the target population.

A different kind of problem is when a measurement or association is spurious. It may be true – but it is not appropriate or relevant after taking something more important into account. Consider this example:

> A father and his young children were riding a New York subway. The children were definitely out of control. The father was slumped over with his head in his hands. When the father did nothing to control his children some of the passengers became irritated. They felt the father was failing in his responsibility. Finally one irritated passenger asked the father to control his children. The father lifted his head and explained that he and the children had left the hospital where his wife, their mother, had just died. The passengers immediately reversed their evaluation of the children and the father -- once they took into account the influence of this death on this family.

In this case, the initial evaluation of the passengers was not in error: the children were causing problems, the father was failing to control his children. The evaluation was correct given the context of the passengers. But the evaluation was spurious: it didn't take into account a confounding factor – a factor that was so influential that it would actually reverse the initial evaluation. What we take into account (or fail to take into account) strongly influences the conclusions we reach. See Rand (1965) and Kelly (1994).

So how does one minimize or eliminate these three problems? (1) Eliminating the problem of chance is often the easiest: just increase the size of the random sample. (2) Eliminating the problem of bias is more difficult. How does one know when a measuring device is faulty? Even when one knows that a large percentage of the subjects failed to respond (non-response), there is great difficulty in determining the influence of this non-response: the non-response bias. (3) Eliminating the influence of confounding factors can be extremely difficult if not impossible in an observational study. In the case of the passengers on the New York subway, how could they have known what had just happened to this family?

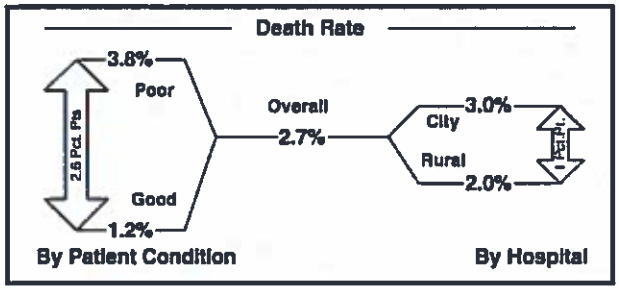
Interpreting statistics is an art – an art of making judgments under uncertainty. But as it becomes easier to obtain statistics on much larger samples, and as the providers of statistics become more professional, the problem of error is reduced and the problem of spurious associations remains. As the quality and quantity of data obtained in an observational study increases, the problem of confounding becomes the central problem.

SPURIOUS ASSOCIATIONS

To be statistically literate, one must ask of any result of an observational study, "Is this association spurious?" To understand a spurious association, one must understand Simpson's Paradox. A **spurious association** is both true and 'false' – but in different ways. It is true given what one has (or has not) taken into account (controlled for). It is 'false' or at least accidental because it does not persist after one takes into account (controls for) a more important confounding factor.

Simpson's Paradox is a *reversal* of an association between two variables after a third variable (a confounding factor) is taken into account. A **confounding factor** is a related factor: a factor that is *found* with (*con*) another. This reversal of an association is often used to support the claim that 'correlation is not causation' and that "observational studies should never be used to support claims about causation." A more precise claim is that 'correlation is not necessarily a sign of direct causation.' An observed correlation may be spurious: due entirely to a confounding factor – a common cause.

Simpson's paradox has been observed in a number of situations. At Berkeley, rejection for admission was more likely for women than for men at the college level. The confounding factor was the department. Rejection was more likely for men than for women at the departmental level. Thus, the association between sex and rejection was reversed after taking into account the department rejecting the admission. The department was a more important explanatory factor of being rejected than was the sex of the applicant. And the choice of department was significantly associated with sex: 90% of the women chose departments with high rejection rates whereas only 50% of the men chose such departments. (Freedman et al.)



Suppose the death rate in city hospitals was 3% while that in rural hospitals was 2%. How strongly does this support the conclusion that rural hospitals are safer than city hospitals? Not very strongly. This relation could be totally spurious if we have failed to take into account any confounding factors that are stronger.

For example, suppose the death rate by patient condition was 3.8% for poor condition and 1.2% for good condition. Suppose that the patients in poor condition are more likely to be found in the city hospitals (where such patients can get better care) than in the rural hospitals (where they can't). Then patient condition may explain the higher death rate in the city hospitals. Indeed, after taking into account the condition of the patients, we may find the city hospitals are actually safer than the rural hospitals.

ALTERNATE EXPLANATIONS

To be statistically literate, one must review three different kinds of explanations for any association obtained from an observational study. In interpreting an observational association between A and B, the three causal explanations are (1) A causes B, (2) B causes A, and (3) C (some confounding factor) causes both A and B. Once all three explanations are expressed, one can work at eliminating one and supporting another.

Suppose there is a fairly strong positive correlation between fire trucks and fire damage. (Source: www.autobox.com). The more fire trucks at a fire scene, the more damage done. Consider three explanations:

1. The fire trucks are causing the fire damage. Thus, the more fire trucks, the more damage they cause.
2. The fire damage is causing [the appearance of] the fire trucks. Thus, the more fire damage, the more fire trucks that are present.
3. A common factor (the size of the fire) is causing both [the appearance of] the fire trucks and the fire damage. Thus, the greater the fire, the more fire trucks present and the greater the fire damage.

In this case, #2 is implausible if the fire trucks arrive before the fire damage occurs. We can eliminate #1 by seeing if the fire damage was in fact caused by the presence of the fire trucks. If we can eliminate direct causality (#1 and #2), we are left with a #3 type explanation. However there can be more than one confounding factor (more than one common cause). But if we can find no other factor that better explains the association, then - *to the best of our knowledge* - the size of the fire is the cause of both. Our intellectual responsibility is to examine and eliminate plausible alternative explanations for the observed association.

There are no statistical tests for bias or for confounding factors. There is no confidence interval for the probability of being free of bias or confounding. This is what makes statistical literacy an art. Being able to evaluate the plausibility and consequences of bias and confounding is essential to being statistically literate.

SUMMARY

Statistical literacy focuses on understanding what is being asserted, asking good questions and evaluating evidence. To see this, reconsider the association between TV violence and antisocial behavior that was mentioned previously. The real issue is how strongly does this evidence support the claim that TV violence is a causal factor: that if TV violence were reduced, antisocial behavior will decrease. Without some measure of the strength of the association, without identifying what factors might confound this relation and without seeing what strength remains in the association after controlling for these confounding factors, a statistically literate reader would say the argument presented is very weak.

Could it be that those children who watch more TV violence receive less supervision by responsible adults? Could it be that children who watch more TV violence are more likely to be in homes that are unstable, disrupted or dysfunctional? If so, then we need to control for these factors. It may be that TV violence is causal. The question is how strong is the evidence.

Statistical literacy is more about questions than answers. It doesn't have many answers, but it should help one to ask better questions and thereby make better judgements and decisions. Statistical literacy is a liberal art – not a mathematical science. In this sense, statistical literacy is a most important skill in dealing with a growing torrent of statistical data. Statistical literacy helps one answer the question asked of most statistics: "What does this mean?"

This introduction is far from exhaustive. It does not discuss reading and interpreting tables and graphs. It does not discuss reading and interpreting statistical models, confidence intervals and hypothesis tests. But it does provide an introduction to viewing statistical literacy as an art – a most useful art for those who make decisions using statistics as evidence. For additional background, see Cohn, Freedman et al., Friedman, Meyer, Schield, Zeisel and Zeisel and Kaye.

REFERENCES

Campbell, Stephen K. (1974). *Flaws and Fallacies in Statistical Thinking*. *Prentice Hall*.

Cohn, Victor (1989). *News and Numbers* Iowa State University Press

Freedman, David, Robert Pisani, Roger Purves and Ani Adhikari. *Statistics* W.W. Norton & Co., 2nd ed.

Friedman, Gary (1994). *Primer Of Epidemiology*, McGraw Hill, 4th ed. p.214 4th ed.

Friedman, David (1996). *The Hidden Order*. Harper Business, 1st ed.

Hooke, Robert (1983). *How to Tell the Liars from the Statisticians*. Marcel Dekker Inc.

Huff, Darrell (1954). *How to Lie with Statistics*. W. W. Norton & Co.

Jaffe, A.J. and Herbert Spirer (1987). *MISUS D STATISTICS: Straight Talk for Twisted Numbers*. Marcel Dekker, Inc.

Kelley, David (1994). *The Art of Reasoning*. 2nd ed. W.W. Norton & Co.

Meyer, Phillip (1975). *Precision Journalism*. Indiana University Press

Paulos, John Allen (1995). *A Mathematician Reads the Newspaper*. Basic Books.

Phillips, John L, Jr. (1971). *How to Think About Statistics*. W.H. Freeman & Co.

Rand, Ayn (1966). *Introduction to Objectivist Epistemology*. Mentor Books, New American Library. P. 85

Schield, Milo (1998). *Statistical Literacy and Evidential Statistics*. ASA Proceedings of the Section on Statistical Education

Schield, Milo (1999). *Statistical Literacy and Simpson's Paradox*. ASA Proceedings of the Section on Statistical Education.

Zeisel, Hans (1947). *Say It With Figures*. Harper & Row.

Zeisel, Hans and David Kaye (1997). *Prove It With Figures: Empirical Methods in Law and Litigation*. Springer.

Contact: Dr. Schield can be reached at schield@augsburg.edu. For this paper and related materials, see Dr. Schield's homepage at www.augsburg.edu/ppages/schield.