EVIDENTIAL STATISTICS

Milo Schield, Augsburg College
Dept. of Business & MIS. 2211 Riverside Drive. Minneapolis, MN 55454

Abstract: The introductory course in statistics promises much more than it delivers. In moving from known sample statistics to unknown population parameters, it promises induction. But statistical inferences (confidence level, p-value, prediction and explanation) are entirely deductive. Although statisticians are well aware of the difference, students typically are not. Students have difficulty identifying whether a claim is deductive or inductive. They cannot distinguish "this outcome is highly unlikely if due to chance" (deductive) from "this outcome is highly unlikely to be due to chance" (inductive). To deliver what has been promised (to help students read and interpret data), introductory statistics should be broadened to include evidential statistics: the use of traditional statistics as evidence in inductive arguments with non-statistical conclusions. Such conclusions involve the strength of belief in inductive inferences: causal explanations, causal predictions and statistical generalizations. Evidential statistics views numerical statistics as premises -- not just as descriptive claims. Thus, evidential statistics deals with the ambiguities of language and the standards for professional communications in statistically based claims. Functionally, the relation between traditional statistics and evidential statistics is like that between micro and macroeconomics. Traditional statistics falls under logic and mathematics; evidential statistics falls under critical thinking. Students need both traditional and evidential statistics to adequately understand the nature, the power and the limitations of statistics.

Keywords: Teaching, Epistemology, Philosophy of Science

1. THE PROBLEM

Students who have completed introductory statistics often comment on how they found so little relation between the statistics they studied and the 'real world'. They see statistics as having very little value unless one is going to be a statistician. Teachers in Business and Communications don't see a clear correlation between taking statistics and being able to deal with arguments involving statistics. Teachers of statistics find this all very disconcerting since they see connections at many levels.

One part of the problem is that introductory statistics is "designed like a human anatomy course, and not a human physiology course. So much time is spent trying to get these students to understand where the basic organs are in the Statistical body, that they never get a chance to understand how the organs function together to maintain homeostasis." From "Testing basic statistical concepts." Posted to sci.stat.edu newsgroup on 2 June, 1997 by Robert Shilling, MPH at Loma Linda, CA. Email to rschill718@aol.com (RSchill718).

The bigger problem is that the introductory course promises much more than it delivers. It promises to teach students how to read and interpret data, how to obtain information about an unknown population parameter given a known sample statistic, and how to test the truth of a null hypothesis. It promises to teach students about prediction error and the strength of an explanation. Classical statistics promises to teach students statistical inference.

The specific problem is that classical statistics teaches only that part of statistical inference that is deductive. It consciously avoids anything inductive. The benefit of this choice is "objectivity" and rigor; the price is narrowness of focus or irrelevance.

Our wording is ambiguous. Our topics imply induction: "Estimating population parameters from sample statistics", "a test of hypotheses". These topics should read "Estimating sample statistics from population parameters" and "A test of significance of a sample statistic given the truth of a null hypothesis."

Students have difficulty distinguishing deduction and induction. They have difficulty seeing the difference between "this outcome is highly unlikely if due to chance" and "this outcome is highly unlikely to be due to chance."

Introductory statistics is limited to deductive statistics whereas students need more emphasis on inductive statistics. We promise both when we say, "we want to help students read and interpret data."

2. CLASSICAL STATISTICS

Classical statistics is mathematics; it is descriptive and deductive. Statistical inference is deductively based on mathematical probability using formulas and proofs. Data summaries are mathematically precise (cf., mean, standard deviation and correlation). The proofs are deductively valid; their conclusions are true whenever their premises are true (cf., the

central limit theorem, confidence levels and p-values). Classical statistics is applied mathematics. Data is studied as a source of numerical values, but most material aspects can be ignored. Data are just numbers that some find interesting.

Classical statistics gives the same results independent of whether the data came from an experiment or from an observational study. This is true of descriptive statistics (mean, correlation), confidence intervals, hypothesis tests, regression and ANOVA. The kind of study is really immaterial in classical statistics.

Classical statistics might be defined using H. O. Hartley's definition: "that branch of applied mathematics that is concerned predominately, though not exclusively, with stochastic phenomena." (Watts, 1968, p. 123).

Typically classical statistics begins with data on individual subjects. Various descriptive statistics and models are used to summarize this data. Classical statistics upholds the use of Bayes theorem provided the prior probabilities are classical probabilities such as relative frequencies.

Classically, the inferential question is how unlikely is this sample statistic (or population organization) *if due* entirely to chance. Classical statistics does not investigate how likely is it that a sample statistic *is due* to chance.

The most common result is a confidence interval or a p-value for a given factor or outcome. A great deal of classical statistics concerns the validity and reliability of various statistical constructs: confidence intervals, p-values, etc. The universal premise of statistical inference is that the variability is due strictly to chance. Chance is modeled by statistical independence.

The conclusions of classical statistics include:

- Correlation is not causation: an observational study can never establish causation.
- A random sample is sufficient: the classic statistical inferences apply -- regardless of whether the sample is representative or unrepresentative, and regardless of whether the study is an experiment or an observational study.
- Relations in tables can be misleading (Simpson's paradox). Correlations between variables can be misleading (Partial correlation).
- For a given set of data, a regression (or an ANOVA) is the same regardless of whether the levels were set in an experiment or observed in an observational study.
- Statistics studies relationships between variables: the nature of the entity or its properties is irrelevant to the truth of statistical inference.
- Statistical confidence is an objective mathematical concept. It is neither subjective nor psychological. It is not the probability that the parameter is in the interval. It is not a guide to action.
- Statistical significance (p-value) is an objective mathematical concept. The smaller the p-value, the more statistically significant is the sample statistic. When the null is rejected, the p-value is not the chance the null is true. The p-value has no epistemic significance concerning the truth of the alternate aside from 'smaller is better'.
- Classical statistics uses the terms 'reject' and 'fail to reject' to describe the decisions one makes, but micro-statistics gives no particular advice on when one should decide to reject (how to select a cutoff value of alpha or how to interpret p-values).
- Classical statistics examines 'explanations' and 'predictions', but these are merely a form of description or association. The explanations don't really explain in terms of causes; the predictions don't really predict what will happen in terms of the effects of causes. The explanations and predictions are entirely formal -- there is no real discussion of the entities involved: their nature and their behavior.

3. GOALS OF STATISTICS

Classical statistics (the study of statistics and statistical inference) can study different subjects:

- Study of chance. This anchors statistics as a branch of applied mathematics.
- Study of variation. Variation includes both non-systematic/indeterminate causes (chance) and systematic/determinate causes. This emphasizes the importance of modeling and process control.
- Studying data. Data links subjects and properties in reality to mathematical variables having variability and associations. Data can be explained as due to chance or as due to determinate causes.

In each case, the goal is to use the techniques of statistical inference to study a particular subject.

There is one goal that classical statistics cannot include:

- Studying *statistics as evidence.* This emphasizes how statistics are used (their function), subsumes

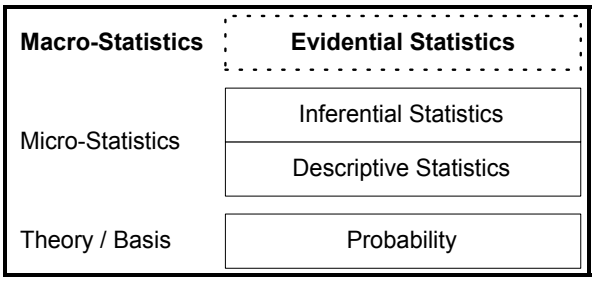
the previous goals as means to this end, and specifies a goal that is outside statistics.

4. EVIDENTIAL STATISTICS:

Evidential statistics studies statistics as a means rather than as an end in itself. Evidential statistics studies the use of classical statistics as evidence in upholding the truth of inductive inferences: explanatory inferences about the natures and causes of things and predictive inferences about the results of causing changes in things. Evidential statistics does not involve a new kind of numerical statistic -- but it does involve new kinds of statements about descriptive and inferential statements. Since identifying causality is typically discipline specific, evidential statistics focuses primarily on language (what that distinguishes classical and evidential statistics), on formal inconsistencies (what are common errors and abuses) and on formal consistencies (does this new data support or contradict).

Evidential statistics is based on classical statistics and epistemology (cf., critical thinking and philosophy of science)

The relation between evidential statistics (macro-statistics) and classical statistics (micro-statistics) can be illustrated as follows.



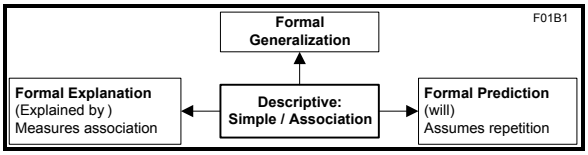
Another way to envision evidential statistics is in terms of the bases on a baseball diamond.

1st:	Classical	How likely if due to chance?
2nd:	Bayesian	How likely to be due to chance?
3rd:	Modeling	Reducing variability
Home	Experiment	Demonstrating causality

Classical Statistics gets one to first base and with deductive certainty. Evidential statistics gets one to 3rd base -- but does so with inductive arguments. With enough data, one can get to 2nd base without classical statistics. But the real work still remains!

5. MICRO AND MACRO STATISTICS

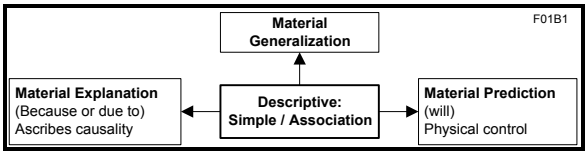
Micro-statistics (classical statistics) focuses on formal inferences -- inferences that are deductive.



Formal associations either describe a situation or are deductively derived therefrom.

- Formal prediction assumes repetition: "given a random sample from the same population, what would we predict."
- Formal explanation merely describes an association: "60% of the variance in the dependent variable is *explained by* modeling on this independent variable."
- Formal generalization is a mathematical induction -- a type of deduction. "If 68% of the means in the sampling distribution are within 2 standard deviations of the population mean, then in 68% of the random samples, the population mean will be within 2 standard deviations of the sample mean."

Macro-statistics focuses on material inferences -- inferences that are inductive.



Material associations involve inductive inferences:

- Material prediction infers the consequences of a given cause. (Suppose the model of house prices increases by $20,000 per bathroom. Can we expect to add $20,000 to the value of a random house by adding an extra bathroom?)
- Material explanation infers the causes of a given association. (Suppose the price of houses increases by $20,000 for each additional bathroom. Is the extra bathroom the sole, primary or important factor in causing this increase? How likely is it that adding a bathroom is the cause of the increase?
- Material generalization infers the causes (population) of a given sample: "How strongly does this sample support the claim that the population is normally distributed?"

6. HISTORY OF EVIDENTIAL STATISTICS

Evidential statistics has a long history. Florence Nightingale, the passionate statistician, used evidential statistics to support her claims that improved medical care would save lives. In 1859, she noted that for every soldier killed in battle in the Crimea, seven died after the battle. But she recognized this eye-catching statistic gave only weak support for her conclusion: that a cause of the high death rate was lack of medical care. She had no evidence to show that improved medical care would have made a difference.

Florence also presented a more mundane statistic: the death rate for young soldiers in peacetime was double that of the general population. By selection, this study controlled for battle-related deaths and for diseases not prevalent in Great Britain. Although less eye-catching, the two-fold statistic actually gave stronger support for her claim than did the seven-fold statistic. (Brown, 1988 *People Who Have Helped the World, Florence Nightingale*, p. 44)

Florence introduced many techniques designed to take into account (control for) confounding factors. She compared cases (soldiers in barracks) with civilian controls. She noted that mortality statistics should be age-specific and that crude death rates can be misleading. (Johnson & Kotz, 1997 *Leading Personalities in Statistical Sciences.* Wiley-Interscience)

Evidence-based statistics have been used ever since and in growing numbers. ☺ Evidence based statistics are discussed in *Statistics, the Principled Argument* by Abelson. Evidence based statistics are common used in law. See *Statistics and the Law* by DeGroot, Fienberg and Kadane, *The Evolving Role of Statistical Assessments as Evidence in the Courts* by Feinberg, and *A Probabilistic Analysis of the Sacco and Vanzetti Evidence* by Kadane and Schum. Evidence-based statistics are used in medicine. See *Evidence-Based General Practice* by Ridsdale. Evidence-based statistics are used in public policy. See *Statistics and Public Policy* by Spencer.

The most complex cases occur when evidence-based statistics involving one field (epidemiology, health, or education) are used to support legal claims. See *Phantom Risk, Scientific Inference and the Law* by Foster, Bernstein and Huber (1994, MIT Press).

7. CHIEF CLAIMS

Some of the chief claims of evidential statistics are:

- Correlation is a sign of causation. Correlation can provide evidence for asserting causation.
- As one controls for potentially confounding factors, the remaining correlation becomes a stronger indication that there is some determinate causality between the variables in the model.
- Bias is at least as important as random variation in interpreting the meaning of data. [cf., Bailar, John C. Amstat News, Nov., 1997. P.5]
- A representative random sample is definitely superior to a simple random sample. In the long run, random samples are representative, but not necessarily so in the short run.
- Evidential statistics studies the relationships between the properties of entities. What a thing is determines what it can do.
- The chance that a sample came from a normal population is related to (but not determined by) the chance of obtaining such a sample from a population that is normal.
- Statistical confidence is both objective and psychological. It does not prescribe an action, but it does calibrate an action. [Schield, 1997]
- Statistical significance (p-value) is epistemically related to the truth of the alternate hypothesis given the rejection of the null. [Schield, 1996]
- To maintain a fixed strength of belief in the truth of the alternate (given a rejection of the null) alpha must be decreased as the alternate seems more implausible. [Schield, 1996]

Consider common questions about classical tests. "I interpret statistical significance to mean that the test discerned differences in the two means that are a consequence of systematic error or bias, from the background precision error. … If the standard deviation is low, then the t-test finds systematic errors with greater confidence. … If a large number of replicates are used, again the t-test finds systematic errors with greater confidence. Is this true?" <Stanley110@aol.com> Stan Alekman [per 4/8/98 post on Sci.Stat.Edu].

8. STANDARDS FOR EVIDENCE:

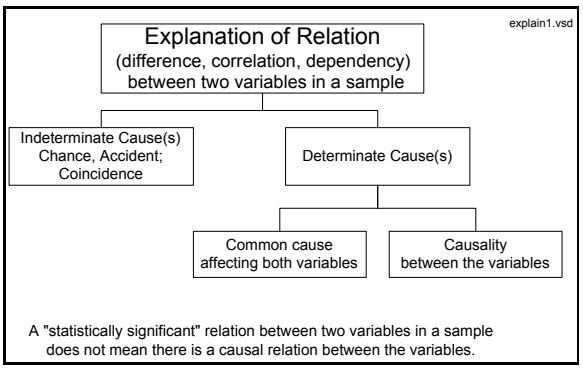
Evidence based statistics deals with both structural and materials matters. Structural matters include identifying the meaning of statistical claims (and noting any associated ambiguity). Material matters focus primarily on the context: the need for action. The context, in turn, sets the standard for what strength of evidence is adequate.

Examples of standards include:

- Emergency situations (immediate necessity). Something bad will happen if we don't act. Something bad may happen if we don't act.
- Political standards. Act whenever enough people, their political representatives, civil servants or tax-supported managers are convinced it is worth trying.
- Legal standards. Judge as guilty only if certain beyond reasonable doubt.
- Medical standards. Act only if the action will cause no harm beyond a reasonable doubt (Hippocratic Oath)
- Scientific standards. Unlimited time: Scientific induction. Virtual consensus by leaders in the discipline.

Although the standards for making a decision will vary with the context, the evaluation of the quality of the evidence is independent of that context. Truth and value are certainly related but they still maintain their separate identities.

For experiments, Mill's standards are generally accepted. For observational studies, there are two sets of standards in current use: the HKE standards and the xxx. {insert more here}



9. ETHICAL RESPONSIBILITIES

Evidential statistics will require much more explicit standards for ethical behavior. We should look at the AICPA and GAAP as examples of what we should do to promote the integrity of our discipline.

We should also learn from the AICPA and avoid any entanglement with any government agency in establishing rules. If there is controversy, then we might label different sets of rules and encourage those using a particular set to clearly identify the rules being used in their presentations.

Seeking government support to promote standards effectively transfers legitimacy to the government to change these standards in the future depending on the situation. Government is a political entity and has an interest in debasing our statistical currency. We would be better emulating the Actuaries or CMAs than in emulating the CPAs.

Our president has asked for papers on the theme of "Statistics -- A Guide to Policy". Evidential statistics can help in this effort. Sometimes the result will be negative -- one simply can't use certain statistics for a given purpose with any credibility, reliability or confidence. Sometimes the result will be highly tentative. If statistics is to be used as a guide to policy, the consumers of statistics must understand how to evaluate the strength of a statistic in supporting a non-statistical claim. If we are manufacturers of a product and our product is being continually misused, we have an obligation to teach others how to use our product lest we be guilty of professional negligence.

10. EDUCATIONAL RESPONSIBILITIES

The ASA, MAA and NCTM are working to establish standards for presenting statistics in high schools and colleges. To the extent that most of this effort involves micro-statistics, we are not meeting the needs of students. Evidential statistics must be introduced along with micro-statistics.

11. REASONS AGAINST

There are reasons for not teaching evidential statistics in an introductory statistics course.

Dilution of our strength. Mathematics is an extremely powerful discipline. It gains its power by sticking to conclusions that can be proven deductively. The so-called mathematical induction is really a form of deduction. To the extent that statistics is to be taught by mathematicians, evidential statistics should be excluded. It pollutes the deductive clarity

of mathematical reasoning. Real mathematics is not inductive; real mathematicians use induction for creation, but deduction is still the gold standard for validation. As mathematicians, we have no special knowledge of induction. We should stick to what we do best: deduction!

Context of knowledge. Doing induction about causality requires domain knowledge -- specific knowledge about the entities and their properties in that discipline. Statisticians, as such, have no expertise in any particular domain. Causality is always discipline specific.

Division of Labor. Even if statistics were to review the use of statistics in identifying causality in different disciplines, this should be done by those interested in the philosophy of science -- not by statisticians. Statistics is a branch of mathematics -- not a branch of philosophy.

Boundary Crossing. Statistics is fully capable of identifying causality in experimental studies. It is the use of observational studies that create the problem. Statistics should focus on the design of experiments and let those disciplines doing observational studies generate their own standards and protocols to argue for the existence of causality. There is nothing that statistics can or should do in this area.

Hubris. *It is intellectual arrogance to think that statistics should take on the role of teaching students how to think (logic), of how to argue (rhetoric), and of how to communicate (persuasive writing) about not just the scientific disciplines (Philosophy of science) but about all disciplines (Philosophy).* To think that statistics should deal with every argument that involves a numerical statistic is not justifiable. In many arguments, the statistic is only ancillary (not direct) and the statistic involved maybe so lacking in credibility as to be all but worthless.

Bad choice of wording. *Even if there is a place in the introductory course for an analysis of the role of statistics as evidence in arguments, we should not use the phrase 'macro-statistics' because of its obvious association with macroeconomics*. For some intellectuals, macroeconomics is the epitome of a science without principles. Both micro and macro economics have well-defined subject matters, have a number of refined concepts, and deal with socially important issues. But macroeconomics is sorely lacking in substantial principles on which there is a strong consensus within the discipline. While micro-economists tend to agree that raising the minimum wage hurts the poor, there is no substantial agreement among macro-economists on the relation between a change in the money supply and inflation. To link statistical evidence to macroeconomics is to place an intellectual millstone around those who seek to earn respect by extending their responsibility in teaching intro statistics.

12. REASONS FOR

College students need to be able to interpret statistics. Only a small fraction of those students who study statistics each year will ever conduct a statistical study, design a survey, conduct an experiment, generate confidence intervals or perform tests of significance. But students have a life-long need to be able to assess the truth of a statistical claim and to evaluate the support it provides in upholding the truth or falsity of a non-statistical claim.

Our colleagues expect us to help their students interpret statistics. Statistics is primarily a service course. Most students wouldn't study statistics unless it was required to obtain their major. Teachers in other disciplines expect that their majors will be better prepared to evaluate arguments containing statistics as evidence. Management majors are much more likely to make decisions in which statistics were given as reasons rather than as conclusions. Although students in some majors (e.g., psychology) may need to conduct surveys (market research), conduct statistical experiments (psychology), generate statistical models (economics and finance), and control production processes (industrial engineering), these same students must still be able to evaluate arguments involving statistics.

Unless the introductory course places more emphasis on analyzing arguments involving statistics, our colleagues may decide to drop statistics from 'required' to 'recommended'. Required courses should contribute directly and substantially to the ability of students to deal with their major. If statistics does not make an observable difference in the behavior of those who complete the course, then why should it be required? When teachers in other disciplines want to add a course to their major but cannot add something without cutting something else, statistics may be on their list of subjects that are under review for being cut.

13. BENEFITS OF

The focus on using statistics in arguments might help frequentists and Bayesians find a more common ground. Frequentists will need to use strength of belief in assessing the strength of inductive arguments; Bayesians will need to evaluate arguments based on frequency-based constructs (confidence intervals and p-values).

By focusing on statistics as evidence in arguments, students will obtain a better balance between deduction and induction. Too many students leave statistics thinking it is just mathematics: numbers, formulas, proofs and problems that have one right answer. Teachers of statistics cannot expect to convince students of the power and importance of statistics so long as they limit statistics to deductive arguments.

Perhaps the greatest benefit of including evidential statistics in the introductory course would be to elevate the intellectual status of our discipline outside our discipline. By helping students learn to evaluate the strength of inductive arguments, we are helping our students deal with inductive arguments that are completely non-statistical. Majors in history may occasionally deal with arguments involving statistics, but history is almost exclusively concerned with arguing about explanations. To the extent that taking a course in statistics helps students think more efficiently and effectively about explanations, statistics will be a critical part of every college program.

14. CONCLUSIONS

Reality is extremely complex. All too often statistics is used (misused) as a short cut to achieving real understanding. Statisticians must not stand by while their discipline is so vigorously misused and abused. If statisticians do not act with resolute clarity, statisticians may find their discipline to be as relevant to most people as Esperanto.

If we want our students to appreciate the power of statistics, we, as teachers and authors, must embrace evidential statistics -- macro statistics -- as a part of our discipline.

15. RECOMMENDATIONS

We must do the following:

Focus more on the verbal and less on the mathematical. Students need help in identifying what words and phrases are most critical in statistical arguments.

Focus more on induction and less on deduction. Focus more on strength of belief and less on validity.

Focus more on using statistics for decision making and less on using statistics to measure the variability due solely to chance.

Focus more on controlling for confounding factors in observational studies and less on the design of experiments.

Focus more on controlling for expected sources of bias (systematic error) and less on the minimum sample size necessary to reduce random non-systematic error to a certain amount.

Focus more on arguing whether an outcome is due to a determinate cause and if so, on which particular determinate cause.

Focus more on everyday arguments involving statistics and less on technical uses of statistical inference as contained in academic journals.

Focus more on how statistical confidence is related to psychological and epistemic confidence in taking an action.

Focus more on how statistical significance is related to psychological and epistemic significance concerning the truth of the alternate.

Focus more on statistics as a science of method and less on statistics as a blend of algebra and finite mathematics.

Statistics must return to its roots as the queen of the social sciences; statistics must come to see itself as a branch of philosophy rather than as a stepchild of mathematics. The teaching of introductory statistics must involve fewer topics and less probability in order to take on the role of statistics in arguments.

If choose to work toward these goals in teaching introductory statistics, we will have indeed chosen goals that are worthy of our best efforts. To the extent we achieve these goals, we may be able to take statistics from being "the worst course I ever had" to being "one of the most intellectually valuable courses I ever had".

REFERENCES

Abelson, Robert P., (1995). *Statistics as Principled Argument.* Lawerence Earlbaum Assoc. Pub.

Brown, Pam (1988). *People Who Have Helped the World, Florence Nightingale*, Exley Publications.

DeGroot, Morris H., Stephen E. Feinberg and Joseph B. Kadane (1994). *Statistics and the Law*, Wiley Interscience 2nd Ed.

Fienberg, Stephen E.(1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts* Springer-Verlag

Howson, Colin and Peter Urbach, (1993). *Scientific Reasoning* 2nd Ed. Open Court Publishing.

Kadane, Joseph B. and David A. Schum. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*, Springer-Verlag.

Kelley, David (1994). *The Art of Reasoning*. 2nd Ed.

Mayo, Deborah (1996). *Error and the Growth of Experimental Knowledge* University of Chicago Press.

Moore, David & George McCabe (1993). *Introduction to the Practice of Statistics*, 2nd ed., Freeman.

Ridsdale, L (1995). *Evidence-Based General Practice*, Saunders Company.

Rosenbaum,Paul R. 1995. *Observational Studies.* Springer-Verlag.

Shum, David A. (1994), *Evidential Foundations of Probabilistic Reasoning. Wiley Interscience.*

Spencer, Bruce B. (1997). *Statistics and Public Policy,* Clarendon Press.

Utts, Jessica (1991), *Seeing Through Statistics*, Duxbury Press.

Watts, Donald G. Editor (1968). *The Future of Statistics*, Academic Press.

Acknowledgments: Thomas V.V. Burnham, Gerald Kaminsksi, Linda Schield made many helpful suggestions. A draft of this paper was presented at WesCOTS in Colorado Springs. Dr. Schield can be reached at schield@augsburg.edu.