View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarWorks@UMass Amherst

University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Ethics in Science and Engineering National Clearinghouse

Science, Technology and Society Initiative

4-1-2000

Responsible Use of Statistical Methods

Larry Nelson
North Carolina State University at Raleigh

Charles Proctor
North Carolina State University at Raleigh

Cavell Brownie
North Carolina State University at Raleigh

Follow this and additional works at: https://scholarworks.umass.edu/esence


Part of the Engineering Commons, Life Sciences Commons, Medicine and Health Sciences Commons, Physical Sciences and Mathematics Commons, and the Social and Behavioral Sciences Commons

Recommended Citation
Nelson, Larry; Proctor, Charles; and Brownie, Cavell, "Responsible Use of Statistical Methods" (2000). *Ethics in Science and Engineering National Clearinghouse*. 301.
Retrieved from https://scholarworks.umass.edu/esence/301

This Teaching Module is brought to you for free and open access by the Science, Technology and Society Initiative at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Ethics in Science and Engineering National Clearinghouse by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Responsible Use of Statistical Methods focuses on good statistical practices. In the Introduction we distinguish between two types of activities; one, those involving the study design and protocol (a priori) and two, those actions taken with the results (post hoc.) We note that right practice is right ethics, the distinction between a mistake and misconduct and emphasize the importance of how the central hypothesis is stated. The Central Essay, I*dentification of Outliers in a Set of Precision Agriculture Experimental Data* by Larry A. Nelson, Charles H. Proctor and Cavell Brownie, is a good paper to study. The Applied Ethics section focuses on objectivity and trustworthiness; we note that the misuse of statistics may be more widespread than misconduct. We have two Central Theme sections; 1) on setting up statistically rigorous hypothesis, and 2) on statistics and data management. The Case Study is courtesy of Case Western Reserve University, from their NSPE based case collection. For our Study Question, we present an ongoing argument concerning the United States census and good statistical practices, asking if statisticians should be involved in deciding how the census should be done.

Our faculty guides for this module are Larry A. Nelson and Marcia Gumpertz, Department of Statistics. We would like to thank Cindy Levine of the NC State University Library for her article search assistance.

Table of Contents

Major Theme II:

Statistics and Data Management: Francis L. Macrina, Chapter 11, *Scientific Record Keeping,* Howard M. Kanare, *Organizing and Writing the Notebook,* Joan E. Sieber and Bruce E. Trumbo, *(Not) Giving Credit Where Credit Is Due: Citation of Data Sets,*

5) Case Study: from the Center for the Study of Ethics in the Professions (CSEP)

6) Study Question: Statistics and the United States Census. Resources: *U.S. Census Monitoring Board.* and Kenneth Prewitt and David W. Murray Letters: Politics of the Census.

7) Additional Resources: Articles, Books, Websites

1) Introduction

The Skagerrak Case

The phrase, "numbers don't lie," is one most of us have heard. On the face of it, this seems true. Numbers, if carefully collected and reported correctly, are objective facts. How is it possible then that two highly respected, experienced marine biologists, both spending years in the field collecting and analyzing data from Sweden's Skagerrak area, could be accusing each other of misusing statistics?

The two scientists, Dr. Alf Josefson and Dr. John Gray, conducted separate long-term studies of eutrophic levels, investigating changes in biomass in order to determine whether increased nutrient load from human activities were having an environmental impact. Josefson, collecting data from 14 different localities, found increases in 12 of them and extrapolated that, therefore, there was an overall increase in eutrophication in the Skagerrak area in general. Gray disagreed completely, saying that Josefson was not using his statistical data properly. Gray argued against Josefson's results on two fronts. First of all, he noted, it is improper statistics to pool data and extrapolate from the part to the whole without a high enough significance level and second, it is irresponsible to publicize results of this sort.

This argument, in all its details and with a discussion of ramifications for both public policy and ethics, was the subject of a provocative article in an October 1998 volume of Science and Engineering Ethics. The disagreement between the two researchers highlights two critical aspects of research: 1) the actual statistical analysis, and 2) how to disseminate the information from research. The "Precautionary Principle" is the principle of publicizing results from preliminary studies, when scientists feel that doing so will lessen the risks to either the environment or the public.

When Larry Nelson, of the Department of Statistics, North Carolina State University, talks about Good Statistical Practices, he describes two different types of activities. First there is the study design itself and the work that derives from following the protocol, the data collection and analysis: these are a priori actions. Working with the results are what Nelson labels post hoc activities. Thus, there are two levels of ethics to consider when doing statistics, the study design and the actions taken with the results.

"Of special interest here is that Gray accuses Josefson of rejecting accepted statistical norms and using the precautionary principle instead. He claims that 'surely the correct scientific approach is to rely on statistics…and that…the role of scientists is to produce the objective scientific evidence…' In a reply to Gray, Josefson poses the question, Should environmental scientists be silent until 95% confirmation has been obtained for a particular change, or are we allowed to warn on the basis of less security?"

Buhl-Mortensen, Lene and Stellan Welin. "The Ethics of Doing Policy Relevant Science: The Precautionary Principle and the Significance of Non-significant Results." Science and Engineering Ethics, October, 1998. 404-405.

Right Practice as Right Ethics

In the article quoted, authors Dr. Buhl-Mortensen and Dr. Welin discuss in detail the a priori issues of study design, hypothesis testing and the critical difference between Type I (false positive) and Type II (false negative) errors. Josefson and Gray were not accusing each other of misconduct—their disagreement was over correct statistical practice. What, for example, are the correct methods for extrapolation? Is it acceptable to pool data and if so, when and how? What is the correct level of significance to choose for a particular study? How can one lower the chance of error, protect against bias and make sure the hypothesis is stated in such a way as to get objective data?

Objectivity in research is the goal. By definition, acting ethically in research means being objective, free from bias and this objectivity needs to be part of the structure of the research study. If the questions are skewed, the data will be faulty and the analysis even further off the mark. Then, no matter how honest and complete the published results are, readers will be led astray; if the original protocol is faulty, the research will be less than useful. And, as we have seen in the disagreement over research results in the Skagerrak, a priori statistical decisions will set the stage for how best to deal with the results. Whether or not one believes that the Precautionary Principle—the idea of early warning—is a good idea is a post hoc question. But how to correctly gather data that may (or may not) bring up the need for an early warning is an a priori issue.

Poor Practice -- Making a Mistake or Misconduct?

There are instances that are not meant to be deceptive or unethical but begin out of an unintended mistake, carelessness or lack of rigor. This was Gray's criticism of his colleague, Josefson—lack of rigor. In an essay first published in Annals of Internal Medicine (1986), John C. Bailar III discusses good statistical practice. Bailar comments on the importance of careful inference as part of the slow, step by step nature of scientific discovery, warning that faulty understanding of statistical methods can lead, even when not intended, to deceptive practices. Bailar warns against calculating p-values at the post hoc stage, adding, "it is widely recognized that t-tests, chi-square tests, and other statistical tests provide a basis for probability statements only when the hypothesis is fully developed before the data are examined in any way." We quote from Bailar's essay, "Science, Statistics and Deception," in the box at the right and it is available electronically. The misuse of statistics, whether through

Some practices that distort scientific inferences

Failure to deal honestly with readers about non-random error (bias)

Post hoc hypothesis

Inappropriate statistical tests and other statistical procedures

Fragmentation of reports

Low statistical power

Suppressing, trimming, or "adjusting" data; or undisclosed repetition of "unsatisfactory" experiments

Selective reporting of findings

Bailar, John C. "Science, Statistics and Deception." Research Ethics: A Reader, Deni Elliott and Judy E. Stern Eds. Hanover:University Press of New England, 1997. 104.

sloppiness or lack of training is not misconduct in the strict sense of the narrow definition, fabrication, falsification, or plagiarism. As Nelson puts it, "It's just not good science." For example, in the extrapolation from a sample to a population, the correct size for n is important. If "n" is too small, estimates of parameters will lack precision and there will be a lack of power in tests of significance. If human or animal subjects are used for a study that is too small to be of value, this is an ethical problem. But if "n" is too large, the expense of the study and data analysis will be unnecessarily high and experimental material (e.g. animals, people) may be wasted.

"You need to frame the statistical questions so that they take in mind the ethical implications of the science. If there is a disconnect between the real question that needs to be answered and the results of your data set, you have not truly appreciated the problem and have an inherently flawed study."

Marcia Gumpertz,
NC State Department of Statistics

Marcia Gumpertz, Department of Statistics, North Carolina State University, emphasizes the critical relationship between asking the right questions and generating answers in a statistical study.

She notes:

"The relationship between the questions of interest and the hypotheses in a test of significance are of crucial importance. In the example of a hypothetical study examining levels of trace metals in drinking water to determine safe levels, we can imagine two different questions of interest."

"First, if we want to show that the levels of certain trace metals are dangerous, the null hypothesis would be that the levels of trace metals equal some values and the alternative hypothesis would be that the levels do not equal these values (or that they exceed these values). The researcher then controls the rate of type I errors at 5% or some other specified level, where a type I error would lead to a conclusion that the water is unsafe when in fact it is actually safe. A finding that the water is unsafe at the 5% level then gives a strong statement that there is a good deal of evidence pointing to the fact that the water is unsafe."

"If, on the other hand, the goal is to show that the water is safe, or substantially equal to some target level, the hypothesis should be set up the other way around. The null hypothesis would then be that the water is unsafe, and the alternative that the water is safe. Now the type I error, which is controlled at a rate specified by the researcher (often 5%), is the error of concluding that the water is safe when it is not. If the null hypothesis is rejected under this test then it gives high confidence that the water is really safe."

Caution: Picturesque Words Ahead

In his essay, Bailar talks of statistical practices that are not necessarily unethical but that can become problematic. One has to be very careful with these techniques since they are not misconduct and may even be standard statistical methods in working with data. But, if misused, one ends up in dangerous territory. And this is where the expertise of a statistician is so critical. It is an aspect of their job, their professional responsibility to know the correct techniques for working with data and what might be needed to correctly run different types of analysis. Again, we see Nelson's exhortation to "right practice" is to the point. He advises that these techniques can be acceptable if done properly.

It's interesting that these techniques are all given colorful terms and the rhetoric is a signal to proceed with caution, whether you are using or being asked to use such techniques, or hear about or read a paper using these techniques. For example, Bailar lists "**data dredging**" and "**trimming**" as "practices that distort scientific inferences." Sometimes you even hear these types of terms bandied about, as in "Oh, he was just **mining** for the data" or "what's wrong with a bit of **trimming**?"

Trimming refers to the practice of not reporting data points that are more than four standard deviations from the mean in a particular data set. The points being deleted are called **outliers**. (For a full discussion of outliers, see Dr. Larry Nelson et al. paper, "Identification of Outliers in a Set of Precision Agriculture Experimental Data." Some advise that outliers should be included in a footnote, along with the explanation for their absence from the final report. Data may be justifiably omitted if there is a clear reason why these data points are not representative of the entire data set; for example, points obtained while the equipment was malfunctioning or data obtained from plots in a flooded area of a field experiment. **Trimming outliers** is what Milliken did in his famous oil drop experiment to measure electron weight. (Whitbeck. "Responsibility for Research Integrity." Ethics in Engineering Practice and Research. New York: Cambridge University Press, 1998. 208.

It is important to be aware of differences between disciplines as they relate to statistical practices. In the social sciences, it is customary to report all data points. In the biological sciences on the other hand, in reporting data from designed experiments, means rather than individual observations are reported. In these cases, says Nelson of NC State University, it is accepted statistical practice, to "trim" non-representative outliers before computing the mean that will be reported.

Data **mining, dredging** or **fishing** refer to the practice (not uncommon in an era of computers) to look through previously collected data sets, looking for patterns. Some might say, just as with **trimming**, that this is a valid task when the scientist knows what he is looking for, has an intuition as to a possible pattern from previous experiments, or when in possession of large data sets. An inherent difficulty with this technique is that it is post hoc; good statistical practice demands that the hypothesis is articulated first and then the data collected, not vice versa.

And then there is the phrase **cooking the data**, which refers to the practice of creating a value, or group of values where you "know" or "expect" them to occur.

But, what if you are missing some data points that you need to run a particular analysis tool? **Imputation** is the term for estimating these missing points so that you have a complete set. This is not cooking data since the statisticians who do this are following statistical procedures to estimate the missing values. And yet, although there are established methods for deriving these data points, they are estimated values. David Resnik (2000) has argued for complete disclosure—in the case of imputation, is this enough? Perhaps statisticians will understand the basis of the imputed value, but will the general public? At what point will careful **imputation** in order to more fully utilize data move from responsible creativity to unintended bias into outright misrepresentation? And who should decide this?

(Larry Nelson, of NC State University, notes that with the current widespread computer programs for estimating values, the issue of imputation is not as critical.)

One term that is problematic is **noise**. David Resnik notes,

"One way of thinking about the role of statistics in research is to think of statistics as a device for amplifying and clarifying a signal. A signal is a real effect, phenomenon, or relationship that is represented by the data, while noise is an effect due to random fluctuations in the data." (Resnik, David. "Statistics, Ethics and Research." Accountability in Research, 8, 2000. 165)

But if the original research design is faulty, what might seem to be "random fluctuations" might indicate a skewing of data. One person's **noise** might be another person's need for further study.

"The selection and presentation of data are a professional responsibility and require the exercise of judgment. Discretion is required to recognize sources of 'noise' (that is, extraneous influences on observations of the phenomena under investigation) and to apply statistical methods to deal with noisy data, even where the source of the noise is unknown. Making the required judgments is therefore more complex than simply reliably recording data. Self-deception is also more of a risk when one must exercise discretion."

Whitbeck, Caroline. "Responsibility for Research Integrity," Part 1 and Part 2 Chapter 6. Ethics in Engineering Practice and Research. Cambridge: Cambridge University Press, 1998. 208.

Commentary on Central Essay

In the Introduction, we have said that there are two levels of ethical conduct. One is a priori and relates to study design. The other focuses on the application of a study, post hoc actions. In the central essay, "Identification of Outliers in a Set of Precision Agriculture Experimental Data," Nelson, Proctor, and Brownie show, by example, the exemplary methods, both for reporting data and for dealing with outliers. For the rest of this module we will focus on the first category, planning the study design.

Nelson indicated that you cannot emphasize good statistical planning enough: it is the critical first step. In his paper, we see how he has meshed the roles of the subject matter scientist and the statistician, right from the beginning. In his discussion of outliers, we see an example of how honest disclosure is done in a professional presentation.

Here is a Good Practices Checklist from Dr. Nelson:

- Planning in experimentation is important like it is in any other aspect of life. The data resulting from an experiment will be only as good as the planning and careful control that went into the experiment.

- The subject matter is where? scientist can initiate the planning process; however, statisticians can also be helpful in the planning process. They can help the researcher organize his ideas into a logical analytic framework. They may even question the researcher's basic hypothesis, causing a need for its revision.

- Planning does not ensure success but it does assure that experimental results won't be hampered with biases that result from bad design.

- Statistics cannot compensate for negative impacts of persisting in a faulty line of research.

What we can prevent with good planning

a. Costly waste of resources
b. Difficult statistical analysis
c. Data for which interpretation is controversial
d. An experiment which is precise but which answers the wrong questions

How to Measure the Success of an Experiment

- Were the original questions important?
- Were the assumptions from which the original questions emerged valid?
- Was there adequate precision and power?
- Was there the proper degree of generality?
- Was the experiment overambitious?
- Is there inappropriate use of a "pet" design that the researcher doesn't understand, but which is popular with colleagues?
- Has there been proper control checks and standardization in a series of experiments?
- Was there an extension of the purpose of an experiment after it was planned for another purpose?

Setting Up the Original Hypothesis Objectively

When Larry Nelson uses the phrase "good science" in setting up a protocol he is talking about creating an objective study that clearly articulates what needs to be studied and provided for specific analytic tools that fit both the study question and the data collection. He notes that "a clearly stated hypothesis will both aid and determine the design of your trial, and will help identify appropriate controls." Here are two of his rules for developing the research hypothesis:

1. The hypothesis must be formulated in a way that it is clearly related to the problem you wish to solve.

2. The hypothesis should be stated as simply as possible.

For example, here is a poor hypothesis: "Improved fallows will increase crop yields." This is too vague and does not point to any specific perimeters for the study. This is a better statement of the original: "Improved fallows using Sesbania or Tephrosia (specific cover crop species) will increase maize yields compared to continuous unfertilized maize or maize following a natural fallow." Says Nelson, "This is better. It is related to the problem, i.e. low maize yields, it suggests a solution and it identifies appropriate control treatments."

A key job for the statistician

An important task for the statistician is to increase the **statistical power** of the test associated with an experiment and this is done through proper study design. Nelson defines **statistical power** as "the probability of rejecting a false statistical null hypothesis," and comments that this is one of the most crucial tasks of the statistician. One may increase the power of a test by increasing the replication, by improving the precision of the experiment (through careful experimental control) and/or by increasing the significance level.

A good statistician will review the data by eye, a time consuming process, but one that insures that the correct analytic tool as well as the correct level of significance will be chosen. This idea of having a literal feel for the data reminds us of Evelyn Fox Keller's title to a book about Barbara McClintock's research: "A feeling for the Organism." This literal "feeling" is just as desirable for a data set as it is for a botanical sample.

"Planning does not assure experimental success but it does assure that the experiment is not put into jeopardy due to improper design, poor execution, or incorrect statistical procedure. Planning also leads us approximately to the correct size of our experiment. Experiments that are too small lack power. Experiments that are too large are costly. Planning leads to a more straightforward data analysis. Planning should be a joint creative effort between the researcher and the statistician."

Larry Nelson, Department of Statistics, NC State University.

3) Applied Ethics: Objectivity and Trustworthiness

Objectivity in Asking the Right Questions

One of the principles of scientific research is that it is value-free; if our goal is to understand the world, we need to gather information objectively. Thus the study designs must be objective. This is critical not only in order that the particular experiment have objective results, but given that knowledge is a collaborative endeavor, future work built on the study will continue in an objective fashion. If a design is faulty, not only will it miss the mark, but subsequent studies will be misguided. There is an ethical imperative to design studies that are inherently objective.

Is it possible, though, to be completely value free? Fred Leavitt (2001, p. 11) asks this provocative question, taking the position that complete objectivity is unrealistic, that the intellectual values that researchers hold, a priori, influence all stages of a study, from conception to publication. He notes how science is a process of hypothesis proving and disproving, saying that tentative preconceived notions influence theoretical models. Making a hypothesis is a tentative procedure. The researcher necessarily needs to assume a position for the sake of setting up the experimental test.

Does Leavitt go too far in questioning whether it is possible to be completely value free? For instance, is it possible to ask a question that is completely neutral, that in no way includes a preconceived idea?

Quantitative questions, queries that can be put in the form of a test of significance are one sort of research. Qualitative questions may not fit into the test of significance paradigm as easily and this is something to keep in mind when designing a good research study.

For example, a researcher might want to test his or her hypothesis that watching violent programs on television results in an increase in aggression in those who view these programs. This is not a simple question and to set up a research protocol that seeks to answer the question in quantitative terms is challenging. Nelson notes that here is the skill of setting up an appropriate study design. In the case of the above study, the co-variables become an integral part of the protocol.

"Then Koehler identified 195 scientists who believed in extrasensory perception and 131 skeptics. He mailed each a description of one of several versions of a fictitious parapsychological research report. The scientists judged studies that supported their beliefs as stronger methodologically than otherwise identical studies that opposed their beliefs. The scientists stated that prior beliefs did not, nor should not have, influenced their judgments of the quality of the new reports."

Leavitt, Fred. "What is Science?" Part 1 and Part 2, Chapter 1. Evaluating Scientific Research: Separating Fact From Fiction Upper Saddle River: Prentice Hall, 2001. 11.

Deciding how to interpret the data is a critical part of designing the study. David Pittenger, as did Fred Leavitt, asks if objectivity is truly possible. In the journal Ethics & Behavior, Pittenger goes so far as to say that making inferences from statistical data is a value judgment. He talks of Type I and Type II errors and the ethical responsibility inherent in assuming a risk of 5% or 1% probability of being wrong due to Type I error. Pittinger, Resnik, and Levitt are concerned that the misuse of statistics will lead to an ethical compromise. But more than that, he states that it is tremendously difficult for statistics to be completely objective, or value-free. Pittenger emphasizes the necessity of constant self examination and analysis to keep the study design and data interpretation neutral and questions if this is even possible.

"As I show, using inferential statistics requires the researcher to make value judgments regarding the importance of conclusions drawn from the data. Moreover, these value judgments have moral consequences that deserve careful consideration. Consequently, I believe that researchers should broaden their analysis of ethical principles to include the criteria they intend to use to evaluate the statistical and practical significance of their research."

Pittenger, David. "Hypothesis Testing as a Moral Choice." Ethics & Behavior, II. 2, 2001. 152.

Objectivity Relates to Trustworthiness

One of the characteristics that gives a researcher trustworthiness is the sense that they are objective, both in the a priori tasks of setting up the study and gathering the data and in the posteriori tasks of interpreting and publishing the results. As an example, a method to ensure objectivity is to select subjects randomly. Yet it is not uncommon to see advertisements in newspapers for "volunteers." Is this method of selection truly random? A protocol that uses human participants who have answered an advertisement is already selected for 1) readers of that particular publication, 2) individuals identifying with the advertised goals of the research study, and 3) if financial reward or other benefits are offered (for example, free health exams) then the sample population is not random at all.
It is difficult to know in this case to what population of individuals the results apply. This emphasizes the need to have a clear population in mind to which the results will apply.

In a useful overview article appearing in Science and Engineering Ethics (Volume I, Issue 4, 1995) Stephanie Bird and David Houseman comment on the difficulty of being clear of underlying assumptions and bias. For example, they note that not all animal studies can be extrapolated from species to species or that findings from a study of men can be extrapolated to women. If in the rush to publicize new findings, even with the intent to improve public health, if steps are missed and findings not verified, the long run affect is to undermine the faith the public will have in research.

"Trust is intimately linked to expectations and their fulfillment. A critical question is 'who is expecting what of whom and why?' As a corollary one can ask, 'Are those expectations appropriate? Are they justified?' "

Bird, Stephanie J. and David E. Houseman. "Trust and the Collection, Selection, Analysis and Interpretation of Data." Science and Engineering Ethics, 1. 4, 1995. 374.

Misuse of Statistics Relates to Trustworthiness and Objectivity

Caroline Whitbeck (1998) notes how negligent or careless behavior can range from cutting corners in haste to meet a publication deadline all the way to actually fabricating data. An honest mistake, if not openly admitted can end up with far reaching consequences.

Whitbeck tells the story of John Urban, a researcher at CalTech. Urban, pressed for time, submitted a paper to the journal Cell with data he fabricated, pending finishing the experiment in time for the actual publication at a later date. Since he "knew" how the experiment would come out, he felt this was not an unreasonable action. When investigated, however, he could not prove his actual experimental data sets were valid because his lab notebooks had been lost. Ultimately, although it was felt he had no intent to deceive, the paper was withdrawn. If time is not allowed for an experiment, what of the time needed for proper statistical analysis? And what of scientific objectivity—can you "know" how an experiment will turn out?

Whitbeck discusses this case, and others, distinguishing between what Nelson calls "poor science" and what she calls "recklessness" (p. 215). Since statistical analysis is a necessary part of research, negligence as to proper technique can have far reaching consequences. In many situations, once a paper is published, the results are not checked further and new research is built upon the report.

"Competence and care are elements of professional responsibility. Failure to give adequate attention and care more often than evil intent leads of a failure of professional responsibility…We saw in Chapter 2 that 'negligence' is a term of moral judgment and that some mistakes—negligent or reckless mistakes—are morally blameworthy. A careless act shows insufficient care and attention; a negligent one shows insufficient care in a matter where one is morally obliged to be careful."

Whitbeck, Caroline. "Responsibility for Research Integrity" Part 1 and Part 2. Chapter 6. Ethics in Engineering Practice and Research. New York: Cambridge University Press, 1998. 215-216.

Misuse May Be More Common Than Misconduct

In Resnik's article previously mentioned, he emphasizes this distinction between actual misconduct and lacking a correct understanding of statistics. He notes that actual intent to deceive (e.g. falsifying data) is one type of unethical behavior: he calls this "an act of commission." He compares this to an "act of omission" (e.g. not reporting all outliers). The former would be clearly an unethical act, while the latter would be a case of how sloppiness could lead to unethical ramifications. Since, says Resnik, proper statistical analysis is so crucial to reporting research, misuse is a great disservice to both the scientific community and the public at large. He notes

that scientists have the moral obligation to act with integrity at every level of research and when they do so, this advances the sense of trust and support by society at large.

Both Resnik and Whitbeck emphasize honest disclosure as a key to integrity in statistics. Depending on what is customary for your discipline, either all the data points or all the means need to be reported and if there are values not included, the reasons for so doing need to be noted. Since future research is built upon current studies, promising areas of study might be missed: one study might reject some values, labeling them "noise," or random results that are irrelevant to the current work. Another researcher might indeed see a pattern in this noise, something worth further examination. Thus, leaving out values without acknowledgment might become an inadvertent impediment to the further advancement of knowledge.

"This essay will argue that the research community needs to pay more attention to the appropriate use of statistical methods in discussion of research integrity, and it will propose some strategies for enhancing discussions of the ethical aspects of statistics in investigational, educational, and organizational settings. The essay will support its view by 1) explaining why statistics plays such a key role in research integrity, 2) describing how some common misuses of statistics in research violate ethical standards pertaining to honesty and error avoidance, and 3)reviewing evidence that suggests that the misuse of statistics is more prevalent (and perhaps more significant) than research misconduct (narrowly defined as FFP)."

Resnik, David B. "Statistics, Ethics, and Research: An Agenda for Education and Reform, Part 1 and Part 2." Accountability in Research, 8, 2000. 164.

Deleted: "

Choosing Between the Good and the Good

Thinking about the interface of honesty, trustworthiness and professional responsibility, there is another aspect to this that bears thinking about. In Module I, *Research Ethics: an Introduction,* Tom Regan noted that many ethical dilemmas involve decisions between two good alternatives.

Thinking about the Skagerrak Case again, one researcher might feel a pull to not publicize less than significant results for the good of the discipline, feeling that premature publication would not be professionally responsible. She might feel that although the studies do lead to some tentative conclusions that it is really not in the interests of the scientific method to give premature disclosure. For her, raising public health concerns that are not completely proven would be to act irresponsibly. At the same time she may be torn by a sense of professional responsibility to

protect the public's health, safety and well-being (See Module V, *Professional Responsibility and Codes of Conduct,* for a more thorough discussion of this) and want in some way to reconcile this with the difficulty of premature disclosure. This, simply put, is what the Precautionary Principle is about, a method to provide for premature disclosure when it is deemed in the public interest to do so, but results have not been scientifically, definitively proven.

Another researcher might say that his sense of professional responsibility for public safety outweighs his concern over premature disclosure. Another researcher might take the opposite view, commenting that in the long run, premature disclosure is not acting as a responsible professional. Here is the kind of question that statistics cannot answer. Even proper statistical practices cannot tell us what a correct ethical choice might be.

If you had to decide what to do about a case such as the one faced by researchers in the Skagerrak, what would you do?

4) Major Theme I: Setting up and Testing Hypotheses

A Good Overview Chapter in a Useful Basic Book About Statistics

We have spoken about the setting up of an hypothesis in terms of a null hypothesis that can either be proven or not—that is, proven or not within a certain level of significance. The best that can be done is to say that a specific hypothesis is proven up to a certain level of chance, usually 5% of being not due to chance. We can see why objectivity is so key: we are already admitting to a 5% probability of our results being due to chance. If we have set up a biased protocol, what does this do to the 5% probability value?

In a clearly written overview chapter "Significance: The Logic of Hypothesis Testing," authors R. P. Cuzzort and James S. Vrettos outline the basic concepts central to good statistics practices. Their chapter headings include: "Type I and Type II Errors," "Setting Alpha" (the probability level for an experiment,) "Sampling Errors," "Significance and Large Samples," "Chi-Square and Significance," and "ANOVA and Statistical Significance."

This overview chapter will be most useful to those who do not consider themselves already trained as statisticians; for statisticians it will probably be too elementary. However, we have included it as a reading selection here, and on electronic reserve, as a baseline, beginning read.

The authors make an interesting comment in the summary of this chapter on one aspect of publishing results. Generally, journals prefer to publicize research that is considered a "new finding" or "positive results." This can, and does lead to an inherent publication bias that can become problematic. Many years ago, the New York Times ran a story entitled, "Negative Data is Still Data." What should researchers do about this tendency on the part of professional journals to publish one type of report?

"One last word: a number of researchers in the social sciences have developed a kind of statistical bigotry with respect to statistical significance. If a relationship is significant, it is considered an important finding. Researchers are commonly disappointed by not finding a significant relationship and do not publish their results when nothing seems to be happening. The consequence has been a bias resulting from researchers publishing their findings when they are able to reject Ho, but not publishing their findings when they are unable to reject it.

Good research should find that support of Ho is as intellectually interesting as its rejection."

Cuzzort, R. P. and James S. Vrettos. " Significance: the Logic of Hypothesis Testing", Part 1 and Part 2. The Elementary Forms of Statistical Reason. New York: St. Martin's Press, 1996. 260.

A Provocative Article in Science and Engineering Ethics

Writing in the January 2001 issue of this journal, Harold Hillman makes a number of important points about both good statistical practice and the fine cusp between what he calls "fraud" and "Para fraud." Fraud is of course, the obvious and narrow definition of misconduct, what Hillman calls "sins of commission."

"Para fraud" would be what Nelson calls "poor practice," Whitbeck calls "sloppiness" and what Resnik labels as "misuse." In this article Harold Hillman gives us a list of "How Research Can Be Improved," emphasizing the importance of original thinking and encouraging a culture of "questioning the experts." Hillman also gives a useful list of what he calls "ground rules" in statistical practice. We reproduce this list in the box below. This essay is available electronically. Nelson notes that he doesn't necessarily agree with Hillman on all accounts. Do you?

"There is such a complicated range of difficulties in statistics, that it might be most useful to list some of the ground rules:

a) one cannot conclude from several series of similar experiments by different authors, each of which does not show a significant difference between two populations, that they altogether add up to a significant difference;

b) different statistical tests examining the same data cannot produce significantly different degrees of significance;

c) if one compares a hundred independent characteristics of two populations, 5% of them will be different by chance, with a probability of 0.05. Thus, if one goes on measuring many different characteristics of a population, or if one does not use all one's data in calculations, sooner or later, one will come across a run of results which will be apparently significantly different from the rest of the population. This may not be a truly biological difference, and can be tested by studying larger populations;

d) many tests of significance of differences between two populations are based on the assumption that the variable measured shows a normal distribution in both populations. Sometimes the populations are too small to permit one to know whether or not the characteristic is normally distributed. If it is not, that particular statistical test may well be invalid;

e) many statistical tests compare random populations. Of course, volunteers, observer-biased observations, and populations in which some values have been rejected on arbitrary grounds are not.

Hillman, Harold. "Research Practices in Need of Examination and Improvement." Science and Engineering Ethics, 7.1 (2001). 8

A Research Article Talks About Statistical Problems Encountered in Publication

We have seen the importance of avoiding both Type I (false positive,) Type II (false negative) errors, as well as the need for a protocol of high statistical power. "Statistical Power," as defined by Dr. Nelson, is "the probability of rejecting a false statistical null hypothesis." We have read several authors who caution against these types of poor practice, but how prevalent are these types of mistakes? In 1997, a team of researchers from both the School of Medicine at the University of South Dakota and the Department of Policy Studies published the results of a survey they had conducted. Looking at a range of journal articles about cardiovascular research, as published in Circulation Research, they focused on the statistical portion of each article to try and answer this question.

They chose two complete publication years to examine: Volume 246 (1984) and Volume 266 (1994.) They focused on the correct use of the t-test as well, with the finding that often the article authors did not clearly state whether a one tailed or two-tailed test was used. Further, they noted that 18% (1984) and 16% (1994) of the time, the t-test was used incorrectly for multiple comparisons. They took exception to the lack of clarity about which tests in general were used for particular purposes, saying "A common annoyance in the methods section was a statement that a particular test was used 'when appropriate.' " (Williams, et.al. p. 490)

One extremely interesting finding, in light of our discussion of problems of bias in statistics is that the assumptions upon which tests were based were often unclear or not stated. In fact, the highest number of what the authors call "common abuses of statistics" were in this category, as opposed to "uncorrected t-tests" or "vague usage," for example. In the box below we quote from the summary of this article.

"Despite much progress, casual inspection of many research articles should convince readers with a basic understanding of inferential statistics that all of the information that is needed for the evaluation of results frequently is not present. Descriptions of statistical methods often are vague, confusing and incomplete, and statistical assumptions usually are not addressed.

…as with any method used, the strengths and weaknesses of the statistical methods used should be discussed in the article. This discussion may include reasons for the selection of statistical tests, reasons for the selection of power levels and minimum acceptable difference, and ethical and financial considerations. Investigators may consider it important to achieve a high degree of power in tests for the primary variables measured."

Williams, John L., Christopher A. Hathaway, Kaia L. Kloster and Benjamin H. Layne. "Low Power, type II errors, and other statistical problems in recent cardiovascular research." American Journal of Physiology, 273.42. (1997). 493.

Major Theme II: Statistics and Data Management

Summary Chapter from a Well-Known Book on RCR

Up to now we have focused on the a priori aspect of good statistical practices. For the second part of the Central Theme section we will review a book chapter and two articles that discuss a critical post hoc aspect of ethical use of statistics—the correct management of data.

A good place to start is to read "Scientific Record Keeping," by Francis Macrina, chapter 11 of his book, Scientific Integrity: an Introductory Text With Cases. Macrina describes keeping laboratory notebooks, giving detailed instructions for data entry. Since the data set is the starting point for statistical analysis, the entries must be impeccable. He lists the components of the proper notebook: a bound book with numbered pages, preferably on acid-free paper since that is the most permanent. Use pen and ink; the consensus is that black colored ball-point is the most impervious to water and smudging. He also recommends setting out some sort of empty chart, or "matrix" as he calls it, to ready your book to receive your data in an organized fashion.

A good laboratory notebook, as we can see from the list "Data book Zen," is the basis for following through on a good research protocol. You can see that once you have developed your protocol, in tandem with your plans for statistical analysis, as Dr. Nelson advises, your notebook will become a valuable tool when you begin the statistical portion of your work.

Macrina describes what he calls a "laboratory central methods book," (Macrina, p.241) a reference manual for procedures and materials for the laboratory in general. This would be the place to record laboratory wide practices and details; for example, reagents and chemicals, supply sources, if specific strains of mice are used, that would be noted as well. He recommends repeating the details in the Materials and Methods section of your own lab notebook. Of course, he notes, each laboratory will have its own procedures and guidelines.

Data book zen

"*Useful data books explain:*
Why you did it
How you did it
Where materials are
What happened (and what did not)
Your interpretation
What's next"

"*Good data books*:
Are legible
Are well organized
Allow repetition of your experiments
Are the ultimate record of your scientific
contribution"

Macrina, Francis L. "Scientific Record Keeping." Scientific Integrity: an Introductory Text with Cases. Washington, DC: ASM Press, 2001. 232.

One of the more challenging aspects of data collection, especially in the changing world of new technologies, is the exact definition of data. There is the classic type of recorded laboratory observations. But field work will necessitate observations as well; in a field such as Sociology or Anthropology the data may be recorded conversations. [And the Institutional Review Board (IRB) consent form to go along with it.] There is a useful online document titled, Investigator's Handbook, published by the University of California.

One of the key points about keeping a meticulous notebook, whatever discipline you are working in, is to make sure that ownership issues are clear from the start. This is especially critical when working collaboratively, whether you are the subject matter scientist or the statistician on a project. And since the nature of the data can change as the work continues—from chemical compound, to a set of complex proteins, to a gel, to a photograph, for example—the notebook is the place where your ownership is clarified. If you are to do a statistical analysis of your gel results, you will need exact documentation of your results in photographs. In the box below we quote from Macrina's discussion of what constitutes data.

"What do we mean by data? Simply stated, data are any form of factual information used for reasoning. Data take many forms. Scientific data are not limited to the contents of data books. Much of what we would call data contained in data books is commonly classified as being intangible. That is, it contains handscript or affixed typescript that records and reports measurements, observations, calculations, interpretations, and conclusions. The term 'tangible data,' on the other hand, is used to describe materials such as cells, tissues or tissue sections, biological specimens, gels, photographs and micrographs, and other physical manifestations of research."

Macrina, Francis L. "Scientific Record Keeping." Scientific Integrity: an Introductory Text With Cases. Washington, DC: ASM Press, 2001. 233.

A classic book by Howard M. Kanare

Writing the Laboratory Notebook, (Washington, D.C., American Chemical Society, 1985) by Howard M. Kanare is a resource often quoted in discussions of the laboratory notebook. There is a copy of this book, available for three-day circulation, in the print reserve collection of the NC State University Library. Kanare outlines correct practices ranging from entering data and witnessing entries to suggestions for official notebook storage. Although the author is a chemist, his guidance is appropriate for research in any discipline.

"Simply put, write down whatever happens, when it happens. Don't wait until the end of the day to sit down and recollect your thoughts; you must plan for adequate time to write notes…Make note keeping, like safe working habits, an integral part of whatever you do."

Kanare, Howard M. "Organizing and Writing the Notebook." Writing the Laboratory Notebook, Washington, DC: American Chemical Society, 1985. 67.

An Article concerning the Correct Citation of Data Sets

Working collaboratively, the statistician and the subject matter scientist, have a mutual interest in a data set. How best to share the data, how to correctly and ethically give credit—these are questions that must be answered at the beginning of the project. And since science proceeds step by step, further research is often built on the shared data, meaning that the issue of correct citation needs to be resolved. One of the difficult issues is that a researcher may hesitate to share data, fearing that he or she will not receive appropriate credit. In this age of high pressure for results, this concern is reasonable.

As we noted in the Module VIII, *An Introduction to Intellectual Property – Copyright*, correct citation will enhance sharing and increase access. Dr. Marcia Gumpertz of NC State University notes that often statisticians are less hesitant than the subject matter scientist to share their data analysis results because they are researching statistical methods and are eager to share questions and problems in the interests of improving statistical tools. For the subject matter researcher, the raw data set and results are valuable in and of themselves.

In an informative article published in Science and Engineering Ethics, (Vol. I, Issue 1, 1995) authors Joan E. Sieber and Bruce E. Trumbo, both at California State University, the former in the Department of Psychology and the latter in the Statistics Department, report on their research into citation practices among scientists. They found there to be no universal standards as to correct practices. Citation practices varied—in some cases they noted that the Principal Investigator is not even named. Although some researchers resist sharing data, the authors found "a network of researchers who are donors as well as recipients of data." (Seiber and Trumbo, p. 19.) One of the authors' strong recommendations is for journals to set high standards and clearly inform their authors of their requirements.

"The meaning of openness, fairness and economy in research has changed with the universal forms of data. Openness in science now means not only openness of method and results but also of data…Today's universal use of computers in science education and research means that the data that would have been far too complex and cumbersome to document, archive and share can now be used even by appropriately instructed undergraduates. Thus, the technology to foster the three virtues—openness, fairness and economy –is now available to all scientists and science educators. What remains is to establish norms in science that foster these three virtues. Hence, our concern with norms in data citation."

Sieber, Joan E. and Bruce E. Trumbo. "(Not) Giving Credit Where Credit Is Due: Citation of Data Sets." Science and Engineering Ethics, (Volume 1.1, 1995. 18.

5) Case Study

The case, Engineer's Duty to Report Data Relating to Research, explores the issue of how much data to report when convinced of your hypothesis.

> Engineer A is performing graduate research at a major university. As part of the requirement for Engineer A to complete his graduate research and obtain his advanced degree, Engineer A is required to develop a research report. In line with developing the report, Engineer A compiles a vast amount of data pertaining to the subject of his report. The vast majority of the data strongly supports Engineer A's conclusion as well as prior conclusions developed by others. However, a few aspects of the data are at variance and not fully consistent with the conclusions contained in Engineer A's report. Convinced of the soundness of his report and concerned that inclusion of the ambiguous data will detract from and distort the essential thrust of the report, Engineer A decides to omit references to the ambiguous data in the report.
>
> Question:
>
> Was it unethical for Engineer A to fail to include reference to the unsubstantiative data in his report?

In the box above we quote from the Case Study website directly. The site contains reference material as well as a Discussion of the issues this case brings out.

You will find that with this case, as well as others, there are two levels of questions and/or concerns; firstly, there will be specific statistical issues, such as data reporting to consider and then, the deeper, more complex societal implications to ponder.

Access the original Case Study, Engineer's Duty to Report Data Relating to Research, read it thoroughly, including the Discussion. As we have done with previous modules, review *Tom Regan's Check List* from page 4 of Module 1. Doing this will enable you to see the inter-relationship of research ethics in general to the context specific concerns that occur when dealing with statistics.

For example, the *issue of reporting scientific results fully and honestly*– how does that link to Regan's point 8: *"Are any duties of justice involved? If so, who has what rights? Against whom?"* Would the general public be involved here in thinking about the issues of justice? Does the public have the right for complete disclosure or should that be left up to the experts to decide? On the other hand, do the experts have the right to present the data as they see fit, given that they have a deeper understanding of the data than that of the general public?

Think over this statement from the NSPE case study:

> "By misrepresenting his findings, Engineer A distorts a field of knowledge upon which others are bound to rely and also undermines the exercise of engineering research." (See Discussion)

Thinking back to the case presented in the Introduction, the Skagerrak Case, how does reporting each and every data point relate to environmental research and the Precautionary Principle? What problems do you see here as the research becomes more complex? How would you relate the crux of this Case Study, reporting all the data points, with the challenge of presenting complex material to the public? Cast a wide net in your thinking about "right balance" in terms of statistics when reviewing Regan's *Morally Relevant Questions.*

Again, as in the case studies for all the modules:

What seems to you to be *resolved* in your own mind?
What seems to you to be *unresolved* in your own mind?
What do you find challenging to *articulate*?

Dr. Larry Nelson asks,

"1. Was Elton guilty of falsification of research results in omitting the anomalous data? What type of additional information would be ethically relevant to this case?

2. What would have been a better approach for Elton to take?"

6) Study Question

At the annual meeting of the American Statistical Association, one topic that comes up for discussion is the national census. How best to account for every single person in the population? What is interesting is that the disagreement seems to follow political party lines, with those favoring actual literal head counts identifying with the Republican party and those favoring some form of statistical extrapolation aligning with the Democratic party. Why is this?

One of the problems is that actual head counts tend to undercount those in the population who move around a lot or live in group dwelling situations, often those in lower income brackets, or urban populations—groups which tend to vote Democratic. This is good for Republican sympathizers. Statistical adjustment of the actual head counts results in higher counts from these populations. This is good for the Democrats.

Does this seem to be a problem for statisticians, or is it really a political problem? If one takes the position that part of a professional's responsibility is for the health and welfare of the public, does that imply the need for expert statisticians to become involved?

In 1999, when Congress held hearings on this challenge, they noted, "The resulting net undercount of more than four million was comprised disproportionately of racial and ethnic minorities and children." U.S. Census Monitoring Board.

Previously, in the Case Study, we have considered the problem of extrapolation of data points in the abstract; here, we see the real life application in terms of funding government programs. This aspect of the census: to create data for program budget determination is not a political problem, it is a statistical one. Or is it? Can we separate out the political/societal from the statistical here? This brings us back to the classic theme: is science (in this case, research in statistical method) inherently objective?

"In the resultant morality play, adjustment advocates usually came off as earnest advocates for the poor, who could be aided by a simple application of statistical justice. Those who favored an enumerated count, on the other hand, were often cast as stubbornly refusing to use a readily available technical means to solve a social problem – 'correcting' the undercount by statistics. Lost in the fracas were genuine arguments about the feasibility and advisability of supplanting the standard enumeration with these technical means—a position ultimately validated not only by the Supreme Court decision of January, 1999, but as well on February 28, 2001 by the decision of the Census Bureau itself. The enumerated count prevailed for good technical, not political reasons."

Prewitt, Kenneth and David W. Murray. "Letters: Politics of the Census." Science, Feb 23, 2001.

7) Additional Resources

Articles

Best, Joel. " Telling the Truth About Damned Lies and Statistics." The Chronicle of Higher Education, May 4, 2001.

DeMets, David L. "Statistics and Ethics in Medical Research." Science and Engineering Ethics, 5, 1999. 97-117.

Gardenier, John, Roles for Statistician in Elections, Mathematics Awareness Month website, April 2008

Seltzer, William, Official Statistics and Statistical Ethics: Selected Issues, International Statistical Institute, 55th Session, 2005

Chance, a journal about statistics and society, is published by the American Statistical Association and Springer.

Vardeman, Stephen B., Morris, Max D. "Statistics and Ethics: Some Advice for Young Statisticians." The American Statistician, 57, 2003.

Books

De Laine, Marlene, Fieldwork. Participation and Practice: Ethics and Dilemmas in Qualitative Research. Sage Publications, 2000). See chapters such as "the moral career of the qualitative fieldworker and "field notes: ethics and the emotional self are of interest."

Smith, F.Gao and Smith, J.E. Key Topics in Clinical Research and Statistics. Taylor & Francis, 2003. Chapters covering a wide variety of concerns, from research design and principles of analysis to grant application and research ethics committees.

Websites

Ethical Guidelines for Statistical Practice, from The American Statistical Association. This is a good website to browse; it contains numerous hyperlinks and online resources.

Maria de los A. Medina, Ethics in Statistics is an online module from Connections: this module was funded by the National Science Foundation: "Collaborative

Development of Ethics Across the Curriculum Resources and Sharing of Best Practices," NSF-SES-0551779

The Internet Glossary of Statistical Terms, from The Animated Software Company.

The National Institute of Statistical Science focuses on cross disciplinary research involving statistics

RCR Data Acquisition and Management, a training module from Columbia University, sponsored by the Office of Research Integrity for its RCR education program.