Research Ethics in Use of Statistical Methods*

Udaya S Mishra

Disagreements and confrontations are common among social scientists regarding conclusions obtained by two researchers on a similar premise. Such disagreements highlight two critical aspects of research i.e. the actual statistical analysis and the manner in which research findings are disseminated. This leads to a consideration of two levels of ethics while doing statistics namely, the study design and the action taken with the results.

Statistics is a tool that facilitates the designing studies in social science research and guides interpretation of results emerging out of it. Application of statistics in social science research can also be misleading unless they qualify what is called good/honest statistical practice. If I put a checklist of what is called good statistical practice they will include the importance of planning; how to measure the success of an experiment; setting up the hypothesis objectively and above all increasing the statistical power of the test. On the other hand, what I refer to, as ethics is to distinguish between mistake and mis-conduct. Hence right practice is the right ethics. However, to make a distinction between a right practice and a wrong practice calls for laying down norms, violation of which should be named as misconduct or unethical. For instance, in statistical parlance, we often hear terms like trimming, dredging or mining of data, but one has to ensure: how much of that is desirable or tolerable in the sense that it does not have a bearing on the outcome/conclusions.

Disagreements and confrontations are common among social scientists regarding conclusions obtained by two researchers on a similar premise. Such disagreements highlight two critical aspects of research i.e. the actual statistical analysis and how to disseminate research findings. This leads to a consideration of two levels of ethics while doing statistics namely, the study design and the action taken with the results [Buhi-Mortensen and Welin 1998]. Apart from study design, setting hypothesis to be tested as well as the difference between types I (false positive) error and type II (false negative) error is of relevance. Disagreements are not often due to misconduct but differing views on the correct statistical practice. For instance, the contentious issues relate to (a) what are the correct methods of extrapolation? (b) Is it acceptable to pool data and if so when

* This article largely borrows from a module on ` Responsible use of Statistical Methods' taught at the North Carolina State University.

and how? (c) what is the correct level of significance to choose for a particular study? and (d) 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?.

There are instances, which are not meant to be deceptive or unethical, but results out of an unintended mistake carelessness or lack of rigor. Bailar (1997) mentions good statistical practice with an emphasis on careful inference. He warns that faulty understanding of statistical methods can lead, even when intended to deceptive practices. Fixing `p' values at a post-hoc stage is deceptive when 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. In fact, such instances are not uncommon in social science research interpretations when the inferences are over stretched beyond its deserving length leading to confronting conclusions and unscientific debates.

Testing of Hypothesis

The relationship between the questions of interest and the hypothesis is a test of significance is of crucial importance. Let us consider an example of a hypothetical study examining levels of trace metal in drinking water to determine safe levels. In this study we can have two different questions of interest:

First, if we want to show that the levels of certain trace metal are dangerous, the null hypothesis need to be that the levels of trace metals equal some values against an alternative hypothesis that the observed do not equal these stated levels (for they may exceed these values).Then, fixing a type I error level of 5% or some other specified level, the type I error will 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 lead s to a strong statement that there is a good deal of evidence pointing to the fact that water is unsafe.

On the other hand, if the goal is to show that water is safe or substantially equal to some target level, the hypothesis need to be stated differently. The null hypothesis would then be water is unsafe against an alternative of water is safe. Now the type I error which is controlled at a rate specified by the researcher (often 5 %), is the error concluding that the

water is safe when it is not. If the null hypothesis in this case is rejected then it gives high confidence that the water is really safe.

Treating the Data

As mentioned above regarding the common vocabulary of trimming and imputing of data, there could be reservations as well. 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 means (or averages) from designed experiments are reported. In such circumstance, it is an accepted statistical practice to "trim" non-representative outliers before computing the mean that will be reported. But in case of missing data points, the popular tool that is used is "imputation" which is an accepted statistical method , yet these are estimated values. [Resnik 2000] has argued for complete disclosure of the method of imputation but is this enough? Perhaps statistician will understand the basis of the imputed value but will the general audience? At what point will careful imputation in order to more fully utilize data move form responsible creativity to unintended bias into outright misrepresentation? And who should decide this?

There needs to be clear distinction between actual misconduct and lacking a correct understanding of statistics. Resnik (2000) notes that the actual intent to deceive (e.g. falsifying data) is one of the unethical behaviour: he calls this an act of commission, which he compared with an act of omission (e.g., not reporting all outliers). The former will be an unethical act, while the later will be a case of how sloppiness could lead to unethical ramifications. Hence proper statistical analysis is crucial to reporting research.

Objectivity and Trustworthiness

One of the characteristics that gives a researcher trustworthiness is the sense that they are objective both in apriori task of setting up the study and gathering the data and in the posteriori task of interpreting and publishing the results. As an example, a method to select subjects randomly. Yet it is not uncommon to find advertisements in newspapers for recruitment of volunteers. A protocol that uses human participants responding to such an advertisement is definitely not random in the sense that they are selected of being

readers of the newspaper, they identify with the goals of the research study and in case of any financial reward/ other benefit offered they are incentive selective.

Hillman (2001) lays down a set of ground rules for good statistical practice, which are as below:

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, and

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.

To sum up, there could be many more occasions when use of statistical principles are fraught with ethical compromises either due to not divulging the entire course of analysis or due to conveniently ignoring/overlooking mention of the inherent assumptions while interpreting the results.

References

Hillman, H (2001) `Research Practices in Need of Examination and Improvement' in *Science and Engineering Ethics*, Stephanie J. Bird and Raymond Spier, Eds., (Opragen Publications, Volume 7,No. 1, January 20001) p. 10.

Resnik, D (2000) `Statistics, Ethics and Research: An Agenda for Education and Reform' *Accountability in Research* Vol.8, pp.169.

Buhl-Mortensen, L and Welin S (1998) `The ethics of doing policy relevant science: the precautionary principle and the significance of non-significant results' in *Science and Engineering Ethics* (Surrey, Opragen Publications) p. 404-405.