Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 24

11-1-2005

Training Statisticians To Be Alert To The Dangers Of Misapplying Statistical Methods

Vance W. Berger
Biometry Research Group, National Cancer Institute, vb78c@nih.gov

DOI: 10.22237/jmasm/1130804580

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 587-590

1538 – 9472/02/$95.00

Training Statisticians To Be Alert To The Dangers Of Misapplying Statistical Methods

Vance W. Berger
Biometry Research Group
National Cancer Institute

Statisticians are faced with a variety of challenges. Their ability to cope successfully with these challenges depends, in large part, on the quality of their training. It is not the purpose of this article to present a comprehensive training plan that will overhaul the standard curriculum a statistician might follow under current training regimens (i.e., in a degree program). Rather, the objective is to point out important areas that appear to be under-represented in standard curricula and correspondingly overlooked too often in practice. The hope is that these areas might be better integrated into the training of the next generation of statisticians.

Key words: Assumptions; design-based analysis; exact conditional test; limitations; permutation test.

Introduction

The ability of statisticians to cope successfully with the wide variety of challenges they face depends, in large part, on the quality of their training. Key components of any training program for statisticians include mathematics, probability theory, statistical inference, and computing. Such classical statistics training would put the statistician in a position to offer solutions to a variety of problems, and defend these solutions. Yet "statistics can be used to form highly technical and even technically correct support for statements which are in fact not true" (Vardeman & Morris, 2003, p. 25). Kimball (1957) described a Type III error as the right answer to the wrong question; earlier Huff (1954) described this phenomenon as a semi-attached figure. It may be overly harsh to use so broad a brush to describe each right answer to a wrong question as an error. Optimal solutions for contrived problems that bear some resemblance to the true problems may also serve as appropriate, if not ideal, solutions for the true problem. On the other hand, an optimal solution to the surrogate problem may not be even a minimally acceptable solution to the true problem.

Vance W. Berger is Mathematical Statistician at the National Cancer Institute. E-mail: vb78c@nih.gov.

Few general rules exist to allow a statistician to be certain that the ideal solution to one problem is actually an appropriate solution to another related problem, so often subject matter knowledge must be used to evaluate a proposed solution to a given problem.

Unreasonable Assumptions

Many frequently applied statistical methods, including t-tests, linear regression, the analysis of variance (ANOVA), the analysis of covariance (ANCOVA), multivariate ANOVA (MANOVA), and the chi-square test, are based on random sampling and/or normality. In practice, these methods are often used even when neither of these conditions holds. It is also common for methods based on compound symmetry of the variance/covariance matrix, interval scaling of the data, proportional odds or hazards, common variances, or additivity to be used when these conditions do not hold. Statisticians must be concerned with such issues as 1) the evidence for or against each of these conditions holding in a given application and 2) the performance of specific analyses when some or all of these conditions fail to hold. Regarding the first issue, we note the impossibility of demonstrating that certain of these conditions hold in practice.

For example, although a statement such as 'the data are normally distributed' may appear innocuous, this statement simultaneously rules out

every distribution that is not Gaussian, including any distribution with finite support. Also, given the mean and variance, this statement specifies a fixed positive probability of a data point falling in any interval, no matter how far from the largest or smallest observations. As such, this seemingly simple statement actually represents an uncountable number of sub-statements, many of which could not possibly be true. The question is not so much whether the statement is true as it is how well would a procedure derived with the assumption perform without it. This raises the question of what exactly is the true question, when all the assumptions have been stripped away.

If a p-value is required for a between-group comparison, then the true question is 'How likely would it be, if there were no treatment effect, to obtain results as extreme as or more extreme than those which were found'? The answer to this question is a probability, and the relevant probability space is defined based on the observed outcome and all other outcomes that could have occurred given the study design. With random sampling from a normal distribution, the probability space would be based on repeated sampling from a normal distribution. Perhaps a t-test would be used, because it is the optimal solution to the problem of comparing the means of normal populations with equal but unknown variances. But, how well does the t-test perform as an answer for the original question?

To answer this question, the correct answer to the original question must be defined. If there is random allocation but not random sampling, then the platinum standard is an exact design-based permutation test (Tukey, 1993). The frequent assurances that standard statistical methods are robust to violations of their assumptions tend to be based on studies of performance when one assumption at a time is violated. In reality, if an analysis requires assumptions to be valid, then it is vulnerable to the possibility that two of its assumptions may be violated simultaneously. In this case, robustness may be lost (Hunter & May, 1993).

In some cases it may not be possible or feasible to compute an exact p-value. But if the exact p-value is available, as it often is, then the numerical difference between it and the approximate p-value is a better measure of robustness than the usual checks that are made of assumptions. Using this metric, Berger (2000) presented a real data set (specifically, sotalol for reinfarctions) whose assumptions appeared to have been met, yet the exact Smirnov test p-values were 0.0485 (two-sided) and 0.0258 (one-sided), and the approximate p-values were 0.9910 and 0.6823, respectively. This discrepancy can be attributed to the poor approximation of the approximate Smirnov reference distribution to the exact one. That is, the value of the test statistic remains the same whether the exact or approximate test is being used, but the p-value it produces fluctuates wildly as the reference distribution to which it is compared varies.

This is hardly an isolated example, nor is the phenomenon specific to the Smirnov test. Little (1989) presented another real data set, specifically a 2×2 table with cell counts {(170,2);(162,9)}. Each expected cell count is at least 5, so the usual check of the chi-square assumption would be passed, and the chi-square test would tend to be used in practice. Yet at the one-sided 0.025 alpha level the chi-square test would find significance (p=0.0162). and would not even be close to the border, although Fisher's exact test would not reach statistical significance (p=0.0299). Three more examples follow. Using the exact Wilcoxon test, Williams, et al. (2000) demonstrated that compared to routine appointments, open access reduces secondary care costs for inflammatory bowel disease.

Barber and Thompson (2000) unwittingly demonstrated that for this data set, either the normality assumption was sufficiently flawed or the difference in means was sufficiently accompanied by shifts in shape and/or scale that the t-test failed to detect this true difference. Likewise, in a study of the effect of neuromuscular training, Hewett, et al., (1999) used the chi-square test to analyze knee injuries in female athletes. Clancy (2000) commented:

> Because the observed and expected number of knee injuries was less than five in at least one cell, an approximate method is inappropriate. An appropriate method in this instance would have been a Fisher's exact test. Incidentally, use of this exact method demonstrated no statistical significance …, suggesting that the extreme variability present in the

> small sample resulted in an incorrect finding when an approximate method was used. This provides all sports medicine researchers with a potent example of why appropriate statistical analysis is extremely important. (p. 615)

Chaudry, et al. (2002) found p-values of 0.004, 0.016, 0.006, 0.001, and <0.001, using t-tests, for five measures (interest, importance, relevance, validity, believability) of readers' perceptions of papers with and without declaration of competing interests. Jacobs (2003) pointed out that the t-test was applied inappropriately, and, using an exact test, found three of these p-values to be non-significant (interest, p=0.054; importance, p=0.21; relevance, p=0.054). Clearly, assumption-based tests are at times used when they should not be. Bross (1990) stated,

> [T]he user of a statistical method has the responsibility for dealing with the *scientific* question: Are the assumptions valid? In particular, when human health and safety might be jeopardized ..., a statistician has a direct responsibility to protect the public health and safety by following fail-safe principles in dealing with any assumptions. (p. 1216)

Some assumptions are more realistic than others, but if they were known to be true, then they would not be assumptions. As such, one could argue that all things being equal, it is best not to rely on assumptions unless there is a good reason to.

In some cases, there are good statistical methods that require no assumptions at all. For example, design-based between-group permutation tests of the null hypothesis of no difference require no assumptions in randomized clinical trials (Berger, 2000). In other cases, progress can be measured by a reduction, but not elimination, of assumptions. Weerahandi and Berger (1999), for example, derived analyses of growth curves that retain the normality assumption but dropped other assumptions. The use of assumption-minimizing methods, along with the proper respect for uncertainty regarding any assumptions that are made, might be regarded as part and parcel of good statistical practice.

Biased Sampling

Without a reason to suspect systematic bias in the sampling procedure, information about the sample would be used, without adjustment, to draw inferences about the population. This would be optimal in the case of unbiased (perhaps random) sampling. Although it is uncommon for a clinical trial to employ random sampling from the target population, this approach is still used in practice, because the sample is still thought to represent the target population from which it was drawn. Whether or not this is true varies with the situation, but there are cases in which the sampling is biased in a known way. Many randomized clinical trials utilize what is called an open-label run-in phase prior to randomization.

Such a run-in phase is characterized by each patient being exposed to the same treatment. On the basis of their response during this run-in phase, patients are selected for or excluded from the subsequent randomization. Generally, good or bad responders are excluded as the run-in phase used placebo or the active treatment, respectively. But, the treatment used in the run-in phase is then used again as one of the treatments to which patients may be randomized. The effect is over-representation of either active responders or of control non-responders (or, sometimes, both). The advantage for the active treatment group can greatly exaggerate the estimated magnitude of treatment effect (Berger, Rezvani, & Makarewicz, 2003). An optimal analysis should provide a good answer to the question of whether or not treatment A is more effective than treatment B in the sample. But with run-in selection, this optimal answer represents an intentionally distorted answer to the question of whether or not treatment A is more effective than treatment B in the target population.

Conclusion

It is hoped that the next generation of statistical researchers will work towards deriving better solutions to the important practical questions that need answering. Often, this will involve deriving more powerful assumption-minimizing analyses. We also hope that the next generation of statistical practitioners will appreciate and use these maximally robust procedures more comprehensively. A good step for aspiring statisticians to take now, to help become part of

the solution later, would be to take classes in non-parametric analyses and robust methods, and to develop an interest in the nature of experiments (including limitations) and the way that data sets are generated. It is also useful for one to recognize what it is that (s)he does not know. All too often it is heard that data are used to prove or conclusively demonstrate a hypothesis, when in fact the inference from data analysis is inductive, and not deductive, so proof is not attainable. If, e.g., assumptions were used in an analysis, then the appearance of a treatment effect could be 1) a real treatment effect; 2) a Type I error; or 3) an artifact due to the assumption not being true. A low p-value allows one to probabilistically rule out the second of these explanations, but not the third. Even if the analysis did not explicitly rely on any assumptions, there is still the implicit assumption that an apparent treatment effect cannot be attributed exclusively to a bias. Selection bias, e.g., can create the appearance of a treatment effect where in fact none exists (Berger, 2005).

Even if every known bias can be ruled out, it is still possible that some other bias exists but is yet to be discovered. Hence, there may be any number of explanations for a given observation (such as a data pattern apparently indicative of a treatment effect), and introspection may help anticipate problems not yet identified, and may allow statisticians to perform analyses and design studies that not only gain acceptance in the present, but also stand the test of time in the future (Berger & Matthews, 2005).

References

Barber, J. A. & Thompson, S. G. (2000). Would have been better to use t-test than Mann-Whitney U-test. *British Medical Journal, 320,* 7251, 1730.

Berger, V. W. (2000). Pros and cons of permutation tests. *Statistics in Medicine, 19,* 1319-1328.

Berger, V. W. (2005). Selection bias and covariate imbalances in randomized clinical trials. Chichester: John Wiley & Sons,

Berger, V. W. & Matthews, J. R. (2005). Conducting today's trials by tomorrow's standards. *Pharmaceutical Statistics, 4*, 155-159.

Berger, V. W., Rezvani, A., & Makarewicz, V. (2003). Direct effect on validity of response run-in selection in clinical trials. *Controlled Clinical Trials, 24*, 2, 156-166.

Bross, I. D. (1990). How to eradicate fraudulent statistical methods: Statisticians must do science, *Biometrics, 46*, 1213-1225.

Chaudhry, S., Schroter, S., Smith, R., & Morris, J. (2002). Does declaration of competing interests affect readers' perceptions? A randomized trial, *British Medical Journal*, *325*, 1391-1392.

Clancy W. G. (2000). Letter to the editor. *American Journal of Sports Medicine*, *28*, 4, 615.

Hewett, T. E., Lindenfeld, T. N., Riccobene, J. V., & Noyes, F. R. (1999). The Effect of Neuromuscular Training on the incidence of knee injury in female athletes. *The American Journal of Sports Medicine, 27*, 6, 699-706.

Huff, D (1954). *How to lie with statistics.* New York: W. W. Norton & Company.

Hunter, M. A. & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology*, *34*, 384-389.

Jacobs, A. (2003). Clarification needed about possible bias and statistical testing. *British Medical Journal USA*, *3*, 93.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association,52,* 133-142.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician, 43*, 4, 283-288.

Tukey J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials, 14*, 266-285.

Vardeman, S. B. & Morris, M. D. (2003). Statistics and ethics: Some advice for young statisticians. *The American Statistician, 57*(1), 21-26.

Weerahandi, S. & Berger, V. W. (1999). Exact inference for growth curves with intraclass correlation structure. *Biometrics, 55*(3), 921-924.

Williams, J. G., Cheung, W. Y., Russell, I. T., Cohen, D. R., Longo, M., and Lervy, B. (2000). Open access follow-up for inflammatory bowel disease: Pragmatic randomized trial and cost effectiveness study. *British Medical Journal, 320*, 544-548.