OUR TEACHING MISTAKES LEAD TO MISTAKES IN RESEARCH

Leroy Wolins, Iowa State University

I teach statistics courses to social scientists and, of course, I have at my finger tips many textbooks in statistics. I also read the social and behavioral science literature and find that much of what we teach is used and usually correctly. Rarely do I find incorrect applications--I usually have to scan 20-30 journal articles before I find one that applies statistical methods incorrectly. Despite this, over the years I have accrued a wide assortment of published research where mistakes are involved and I have sorted through these and attempted to classify them into broad, meaningful classes which are mutually exclusive and exhaustive.

I have not been successful at this classification task. Although I like a few of my categories, most of them are "fuzzy" and many of these publications fit several broad classes. Today I want to talk about one narrow "fuzzy" class which I call "Mistakes in Non-parametric Analyses Based on Mistakes in Textbooks". I chose this narrow class because I can present it in fifteen minutes and it seemed appropriate to a statistical audience.

The Mistakes

Although the class is narrow, relatively, I am going to talk about what appears to be several different ideas. To get this started consider three data sets A, B, and C, each based on a sample size of four and each of these observational units are measured twice; once under treatment one (T_1) and again under treatment two (T_2). Further, consider that the actual measurement procedure resulted in scores that did not behave "nicely" so that only the rank of the eight scores are reported. For example, these eight scores could come from four autistic children measured with (T_1) and without (T_2) a medicine according to the amount of time the children attend to external sources of stimulation.

Table 1. Three Data Sets; A, B, and C; Involving Two Treatments and Four Observational Units

	A		B		C	
	T_1	T_2	T_1	T_2	T_1	T_2
1)	1	2	1	5	1	8
2)	3	4	2	6	2	7
3)	5	6	3	7	3	6
4)	7	8	4	8	4	5

How do we analyze such data? Well, I use Otts (1977) book for a course I teach so what does he say? He starts Chapter 11, "Nonparametric Methods", with, "Some studies yield data identified by rank only--", so I'm in the right chapter and the first thing I find is "The Sign Test" which is for "paired data" and I read it and find I can use it. But if I do, I get the same answer for all three of the data sets which doesn't seem right so I read on and find "Wilcoxon's Signed-Rank Test" ... "which makes use of the sign and the magnitude of the rank of the differences between pairs of measurements ...". Make use of the magnitude? What happened here? I thought we were talking about "... data identified by rank only ...". Oh well, I am glad I read it carefully because I only want to use the ordinal information and not magnitudes of differences occurring at different places on my "messy scale" so I read on and find "Wilcoxon's Rank Sum Test" which some people call the Mann-Whitney U test but I find I can't use that because that test is for comparing two populations and I have a sample from only one population and I read on and find nothing relevant so I go to Snedecor and Cochran (1967), Ostle and Mensing (1975), and others (Ferguson (1971); Steel and Torrie (1960); Bradley (1968)) and find much the same discussion except these other books do not mention that the signed-rank test ranks magnitudes. Maybe that is not important. I look further and I find Blalock (1979) and Conover (1967) and indeed the assumptions stated in both books include the interval measurement one so I am left with the sign test. Right? Wrong!

Why can't I use the rank sum test for these data. It is a permutation test that tells me out of all 8! permutations only $(4!)^2$ of them would give me results as extreme as data sets B and C and the rank sum test indicates the results are "significant" whereas the sign test indicates $p > .05$. There is no reason for not using this rank sum test. It uses information in the data without using "information" that is not in the data.

But what about the differences between data set B and C? Certainly one should be more comfortable with the B results than the C results but the rank sum test gives the same result for both of these data sets. I looked further and found one cannot incorporate this feeling into a non-parametric statistic requiring only ordinal information. What one should do, I suggest, is to teach our students that there is a concept of non-additivity that pertains to ordinal data.

But what about situation A? The probability derived from the sign test is 1/16 and the rank sum test gives an even larger probability. Yet the fact that $T_1 > T_2$ for all four observations and the rank ordering of the four observations is the same for both treatments seems to provide even more compelling evidence against H_o than situation C where the rank ordering resulting from the two treatments are opposite to what one should expect.

I also note, temerariously, that t-test results in a very small p-value for situations A and B but not for situation C.

Apropos the autistic children it seems the three results should invite three different inferences. Situation A suggests the medicine is effective but more data is needed before implementation. Situation B indicates the medicine is effective and one should start implementation on some limited scale. Situation C suggests that the medicine helps those who are not very bad off in the first place but is of doubtful use for the severly introverted.

These inferences are not based exclusively on any of these statistical tests. Although a probability basis for these inferences are available from these statistics they are vague: they either use "information" not in the data or ignore relevant information. There is no good way to come up with a single number to indicate our faith in the results. In making these three inferences I have resorted to the other "vagueness", intuition, which I consider to be the lesser evil, but that is not my main point. My main point is that authors of textbooks owe it to their audiences to point out this particular vagueness which is: there is no right way to analyze statistically paired observations using only ordinal information.

You should ask, "Why have I made this point in this context since it is almost always true that statistical summarizations do not use all the information in data?" I answer, in other situations the information loss is due to failure to meet assumptions but in this case it is not possible to design a statistical test that uses the block information when the treatments are compared. That is, in situations A and B, the concordance is high but the sign test gives the same answer and in situation C the concordance is low but again we get the same answer. If one used the t-test, situation C is identified as different from the other two and situations A and B are distinguished because the treatments are estimable in context of the assumptions. Thus the sign test and the rank sum test may loose some information because they <u>cannot</u> use it, whereas in other situations such information is useable but there is usually some loss because of failure to meet assumptions.

These five articles from the current literature, though not necessarily wrong, illustrate the confusion about this topic.

Foa, E. G., G. Stekettee and J. B. Milby. Differential Effects of Exposure and Response Prevention in Obsessive-Compulsive Washers. <u>Journal of Consulting and Clinical Psychology</u>. 1980, 48, 71-79. (Used sign test when the scores were completely ordered.)

Dunleary, R. A. and L. E. Baade. Neuropsychological correlates of severe asthma in children 9-14 years old. <u>Journal of Consulting and Clinical Psychology</u>. 1980, 48, 214-219. (Used "Wilcoxon matched-pairs signed ranks test" because "many of the Halstead battery tests provided only nominal measurement, the number of subjects was small and the test data were not normally distributed".)

Durlach, P. J. and R. A. Rescorla. Potentiation rather than overshadowing in flavor-aversion learning: an analysis in terms of within-compound associations. <u>Journal of Experimental Psychology</u>. Animal Behavior Processes, 1980, 6, 175-187. (Used Wilcoxon on least signed rank test - no reason given.)

Knudson, R. M., A. A. Sommers and S. L. Golding. Interpersonal perception and mode of resolution in marital conflict. <u>Journal of Personality and Social Psychology</u>, 38, 751-763. (Used "... Wilcoxon matched-pairs, signed ranks test ..." because "... the rating procedures described above do not yield either interval or ratio scale measurement".)

Lewis, T. L. and D. Maurer. Central vision in the newborn. <u>Journal of Experimental Child Psychology</u>, 29, 475-480. (Used "Wilcoxon test of matched-pairs" apparently because the measurement errors were heterogeneous. However N = 46. Means were reported with no variability statistic.)

REFERENCES

(1) Blalock, H. M., Jr. (1979). <u>Social Statistics</u> (Revised Second Edition), McGraw-Hill, New York.

(2) Bradley, James V. (1968). <u>Distribution-free Statistical Tests</u>, Prentice-Hall, Englewood Cliffs, New Jersey.

(3) Conover, W. J. (1971). <u>Practical Nonparametric Statistics</u>, John Wiley, New York.

(4) Ferguson, G. A. (1971). <u>Statistical Analysis in Psychology and Education</u> (Third Edition), McGraw-Hill, New York.

(5) Ostle, B. and R. W. Mensing (1975). <u>Statistics in Research</u> (Third Edition), Iowa State University, Ames, Iowa.

(6) Ott, Lyman (1977). <u>An Introduction to Statistical Methods and Data Analysis</u>. Duxbury Press, North Scituate, Massachusetts.

(7) Snedecor, G. W. and W. G. Cochran (1967). <u>Statistical Methods</u> (Sixth Edition), Iowa State University, Ames, Iowa.

(8) Steel, R. G. and J. H. Torrie (1960). <u>Principles and Procedures of Statistics</u>, McGraw-Hill, New York.