A PLEA FOR A <u>STATISTICAL FORMULA REDUCTION ACT</u>

by

Dennis M. Roberts

Penn State University

May 1992

It should come as no revelation that statistics is one subject that raises anxiety goosebumps on psychology students' bodies. In many cases, the mere thought of taking a statistics course is sufficient to make one feel ill. It is easy to blame students for these problems, citing such reasons as inadequate math preparation and poor study habits as the major culprits. Certainly, there is substantial informal evidence to back up such claims. However, students are not the only culprits in this common scenario. The quality of statistical instruction, and the quality of statistical books, also share some blame in this regard.

Recently, Roberts and Kunst (1990) attempted to make a case for the discontinuance of the use of the Spearman rank-order correlation formula and, other than historical mentions, references to the Spearman rank-order correlation. It was argued that continued use of the Spearman correlation not only encourages claiming that a Spearman correlation is being calculated when in fact it actually is a Pearson coefficient, but also fosters the belief that the Spearman correlation tells you something that the Pearson correlation does not. How often has it been said, if the data are at the ordinal level, then use the Spearman rank-order correlation (as if when you use the Pearson formula on the ranks it does not represent the same thing as does the Spearman)?

In this same context, other correlation procedures are given space in articles and textbooks that goes far beyond the realm of reasonableness. For example, we know that the point biserial correlation is simply an alternative Pearson formula when the X and

Y data have one dichotomous variable. Also, it is common knowledge that the phi coefficient is another alternative Pearson formula when both the X and Y variables are dichotomies. However, the context in which these algebraically equivalent formulas were developed was the desire to be able to calculate Pearson correlations more efficiently and without as much pain. In its raw form, and with realistic sets of data, the Pearson formula must be one of the most dreaded statistics to calculate that ever lived!

It should be obvious that the rationale that these alternative formulas provide a more efficient vehicle for calculation became invalid long ago. The advent of the hand-held calculator in the 1970's with a few statistical features made the need for using shortcut formulas for Pearson correlations obsolete. Why then, has there continued to be such emphasis on these formulas in both the statistical and textbook literature?

For example, a more or less (probably less) random selection of general purpose social science statistics books from my shelf (Christensen & Stoup, 1991; Glass & Hopkins, 1984; Heiman, 1992; and Howell, 1987) showed approximate means for page coverage for the point biserial and phi coefficients to be 2.2 and 2.3 respectively. The ranges went from .5 to 3.25 for point biserial and from 1.25 to 3 for phi. All of these sources provide and discuss the computational formulas for both the point biserial and phi. These are, of course, in addition to their previous presentations of the Pearson formulas. Also, most give some context in which these formulas are used, and some interpretation of such values.

Given that the point biserial and phi coefficient formulas are exactly the same as the originally presented Pearson formula, and given that they yield the same identical result, would it not make sense to simply indicate that fact? It is perfectly legitimate to mention in a section on Pearson correlation that, at various times, shortcut formulas have been developed when one or both of the X and Y variables are of the dichotomous variety. It would also be correct to point out that there was a time when such calculation simplification was helpful, but with calculators and computers, these formulas no longer serve any pedagogical usefulness. In addition, the more space is devoted to such unnecessary elaborations, the more one is likely to see some misinformation presented. For example, in one case (Christensen and Stoup, 1991), the authors state "Although a point biserial correlation has essentially the same meaning as a Pearson r and, indeed, a Pearson r could have been computed on the same data and given the same result ... (p. 124)". The implication from this statement is that although the point biserial is similar to the Pearson, it still is somehow different. Such statements require students to make discriminations among correlation names and formulas that lead to unnecessary confusions. No wonder it is hard for students to keep things straight! Again, how often do you hear students and faculty say: if I have one continuous variable and one dichotomous variable, do I use the Pearson or the point biserial?

The examples above are taken from correlation analysis. There are many other examples from other areas of statistics and quantitative analysis that represent a similar type of problem. Some procedures that were developed at one point in our statistical history have long outlived their usefulness. We should have the courage to say that they have outlived their functionality and remove them from our verbiage and scholarly communications. My plea for a statistical formula reduction act requests that we make a proactive effort to reduce the number of formulas we use so as to reduce confusion in both our instruction and in our analysis reporting activities.

References

Christensen, L. B. & Stoup, C. M. (1991). Introduction to statistics for the social and behavioral sciences (2nd ed.). Brooks/Cole Publishing Co.: Pacific Grove.

Glass, G. V. & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd. ed.). Prentice-Hall, Inc.: Englewood Cliffs.

Heiman, G. W. (1992). Basic statistics for the behavioral sciences. Houghton Mifflin Co.: Boston.

Howell, D. C. (1987). Statistical methods for psychology (2nd. ed.). Duxbury Press: Boston.

Roberts, D. M. & Kunst, R. E. (1990). A case against continuing use of the Spearman formula for rank-order correlation. Psychological reports, V66, 339-349.