An Examination of Some Commonly Held Attitudes Regarding the Nature and Usefulness of Nonparametric Tests

R. Clifford Blair
University of South Florida

Introduction

Although nonparametric statistical tests enjoyed a certain degree of popularity among educational and psychological researchers during the 1950's (Glass, Peckham and Sanders, 1972), attitudes concerning the usefulness of such procedures have changed markedly since that time. This change in attitudes is reflected by Glass et al (1972) who characterize the 1950's movement to nonparametrics as "unnecessary" and "unproductive." These authors go on to imply that researchers who use such procedures are not doing so on the basis of an informed decision, but rather, are simply caught up in a "herd" psychology.

One of the most popular statistical texts used in the training of psychologists is that of Guilford and Fruchter (1978). Although these authors admit that nonparametric tests may be of some very limited usefulness in small sample situations, they go on to admonish the reader "Where there is any choice...we should prefer a parametric test, except where a quick, rough test will do."

The purpose of this paper is to (1) identify the reasoning that underlies the common belief in education and psychology that nonparametric tests are of little or even no use in analyzing research data and (2) to compare and/or contrast this reasoning with relevant information available in the literature. Because of its common use and because it has been widely discussed in the literature, primary focus will be on the two independent means *t*-test.

Perceptions Regarding the Robustness of the *t*-test

Although it is oftentimes admitted that research data have "...an exasperating tendency to manifest themselves in a form which violates one or more of the assumptions underlying the usual tests of significance..." (Boneau 1960), this tendency is usually dismissed as being of little importance in so far as the *t*-test is concerned. This lack of concern is based on the belief that the *t*-test is robust to certain violations of its underlying assumptions. This is particularly true when the assumption is that of population normal distribution. The authors of one of the more popular statistical texts designed for use by educational and psychological researchers state, "Violation of the assumption of normality in the *t*-test of $H_0 : \mu_1 - \mu_2 = 0$ has been shown to have only *trivial* effects on the level of significance and power of the test and hence should be no cause for concern" [italics added] (Glass and Stanley, 1970). It is important to note that the authors do not qualify this statement as to amount of skew in the population, sample sizes employed, use of equal sample sizes, use of one or two tailed tests, or level of significance chosen for the test.

In what has probably become the most widely cited study of the *t*-test's robustness, Boneau (1960) staes "The purpose of this paper is to expound further the invulnerability of the *t*-test and its next of kin the F test to ordinary onslaughts stemming from violation of the assumptions of normality and homogeniety." Boneau (1960) goes on to state at a later point "Thus is would appear that the *t*-test is *functionally a distribution-free test*, providing the sample sizes are sufficiently large (say, 30, for extreme violations) and equal."[Italics added.]

In a recent text, Hopkins and Glass (1978) state "Fortunately, much research has revealed (see Glass, Peckham, and Sanders, 1972) that violation of the assumption of normality has almost no practical consequences in using the *t*-test." At a later point these same authors state "...the condition of normality can be largely disregarded as a prerequisite for using the *t*-test."

The unqualified dogmatism inherent in the Glass and Stanley (1970) statement combined with the exuberance of Boneau's (1960) statements as well as those of Hopkins and Glass (1978) might lead the reader to believe that a modicum of temperance is called for. In fact, temperance in statements that reflect on the robustness of the *t*-test may be strongly condemned. Consider for example the following statement by Hawkridge (1970).

> The question of using *t* and F tests with such skewed distributions was brought to our notice during one particular study...One of our staff at AIR...was statistical consultant for the study, and he went to some trouble to investigate the claims about the robustness of *t* and F tests. He showed that although Norton (in Lindquist, 1953) and Boneau (1960) had defended the robustness of *t* and F tests, the more recent work of Bradley...had raised new doubts about the violation of certain assumptions. This is not the place to go into detail about this debate, but Bradley's view... is that nonparametric statistics should be used when parametric assumptions are violated, rather than their normally more efficient parametric counterparts. [This quotation from Hawkridge (1970) was taken from Glass et all (1972).]

Reacting to the above quotation, Glass et al, (1972) state, "Incautious statements concerning the robustness of the ANOVA to non-normality could send applied statistics off on a rerun of the unproductive 1950s stampede to nonparametric methods. Hawkridge (1970, p. 36) threatened the safety of the herd with this warning...." Thus, even the mildest suggestions that nonparametric methods might be profitably substituted for parametric methods in situations where data are taken from non-normal, e.g., highly skewed, distributions can draw heavy fire in one of the most respected educational research journals.

The statistical consultant mentioned by Hawkridge (1970) must have done his research carefully for in spite of the fact that Bradley has done extensive studies of the *t*-test's robustness (or lack thereof), he has encountered great difficulties in having his results published. His difficulties have been particularly acute in so far as American, statistically oriented psychology journals are concerned. Commenting on this state of affairs, Bradley (1978) states "...my robustness studies, conducted over a decade ago, have appeared as government technical reports and have been briefly abstracted in my own book (Bradley, 1968), but have never been published in detail in any readily accessible source in the open literature...." Bradley (1978) goes on to present evidence to support his contention that his lack of success in publishing his findings is due to a strong referee bias that precludes the publishing of any evidence that shows a lack of robustness on the part of the *t*-test. It is perhaps noteworthy that these comments by Bradley (1978) appeared in the *British Journal of Mathematical and Statistical Psychology* and that his article on an easily encountered class of non-robustness-conducive populations (Bradley, 1977) appeared in the *American Statistician*.

Even the "briefly abstracted" material mentioned in the above statement was not published without serious challenge. In regard to this, Bradley (1978) states, "One of the publisher's consultants for a book I had written (Bradley, 1968) concentrated all his outraged fire upon the few pages that abstracted my robustness studies and discussed the virtues and disadvantages of parametric tests."

Further insight regarding the attitudes of some authors toward the robustness issue and the usefulness of nonparametric tests can be gained by considering the following statement by Glass et al (1972). "There must surely be some breaking point at which a distribution is so pathologically skewed that nominal levels of significance and power are seriously misleading.... Thus, while holding the general conclusions...in mind, the prudent data analyst would nonetheless attempt to estimate the skewness, kurtosis, and variances of the populations he has sampled and reference the tables of data presented above in the event that any of these values is extreme." By using the term "pathologically" skewed, these authors seem to imply that any distribution under which the *t*-test is non-robust may have come from extremely skewed distribution. It should also be noted that these authors do *not* recommend the use of nonparametric tests with such "pathological" distributions and, as was pointed out earlier, strongly recommend against the use of such tests. It appears therefore, that these authors believe that any population under which the *t*-test is non-robust may be so flawed as to preclude data analysis, or at least, call the interpretation of results into question.

Summarizing the discussion in this section, there is a tendency in education and psychology to make very assertive statements regarding the robustness of the *t*-test to population non-normality. Further, there is a tendency to resist any statements or evidence that would imply a lack of robustness in any but "pathological" situations. Some authors even resist recommending nonparametric tests in the so called pathological situation, implying that this data is some how not legitimate.

Evidence Regarding the Robustness of the *t*-Test to Population Non-Normality

The first difficulty encountered in trying to assess the robustness of the *t*-test to population non-normality arises from the lack of a generally accepted definition of robustness. Given a nominal significance level of .01 and an actual Type I error rate of .02, does one conclude that the error rate has been increased by a trivial .01 or does one conclude that the error rate has been substantially increased since it is twice the intended value? Clearly, personal point of view plays a major role in determining whether results obtained from a study show the *t*-test to be robust or non-robust in a given set of circumstances. Despite this ambiguity, sufficient evidence exists to allow for the conclusion that the *t*-test is quite robust in many non-normal population situations. The question still remains as to whether or not a researcher may reasonably expect to sample populations that, because of their non-normal shapes, are conducive to non-robustness in the *t*-test.

Bradley (1977) has given both a rationale and empirical evidence to support his claim that radically non-normal data occur with some frequency in the social and behavioral sciences. One such population had skew of 3.42 and kurtosis of 17.29. Bradley (1977) goes to some trouble to point out that these distributions arise for perfectly legitimate reasons that are unrelated to outliers.

Table 1 shows a portion of the results obtained by Bradley (1964) from his Monte Carlo study of the robustness of the *t*-test. The first column of the table gives the sizes of the two samples (n_1,n_2) while the second column gives the population from which the respective samples were taken. In this column, "L" designates an L shaped population that was generated during the course of a routine psychological experiment and "N" designates an approximately normal distribution. Thus, the two column designation "32,16 L,N" represents the situation in which the first sample, of size 32, is drawn from the L shaped population and the second sample, of size 16, is drawn from the approximately normal population. The remainder of the tables gives left and right tail cumulative probabilities under nominal values.

As was noted earlier, robustness determinations, at least to some degree, must rest in the eye of the beholder. Nevertheless, Table 1 contains many examples of situations that most reasonable observers would see as highly condusive to non-robustness. In addition, many observers wouls find questionable the assertion by Boneau (1960) that the *t*-test becomes "functionally a distribution-free test" when $n_1 = n_2 \geq 30$.

One further point regarding the robustness of the *t*-test should be made. Bradley (1977) states "...the strength of the evidence for robustness appears to derive partly from selectivity in investigating only the more familiar population shapes–which may be far less prevalent than their familiarity would suggest."

In summary, there is sufficident evidence in the literature to allow for the conclusion that the *t*-test is remarkably robust to many types of departures from population normality. However, there are rational as well as empirical reasons to believe that researchers in the social sciences may encounter population shapes that are condusive to non-robustness in the *t*-test.

Perceptions Regarding the Relative Power of Parametric and Nonparametric Tests

Bradley (1972) points out that nonparametric tests were often perceived as being "...quick and dirty substitutes for their parametric counterparts...." and that they were "...widely regarded by practitioners to be distinctly inferior in efficiency..." to their parametric counterparts. The Guilford and Fruchter (1978) quotation given earlier would indicate that this attitude still prevails among some authors of statistical texts designed for use by social scientists.

Boneau (1960) strongly defends the use of the *t*-test rather than some nonparametric procedure in the non-normal situation. In the course of his discussion, Boneau (1960) states "...tests which make no assumptions about the distribution from which one is sampling will tend not to reject the null hypothesis when it is actually false as often as will those tests which do make assumptions. This lack of power of the nonparametric tests is a decided handicap...."

In a recent and popular text entitled "Basic Statistics for the Behavioral Sciences," (Stanley and Glass 1978) only one reference to nonparametric statistics is found in the subject index. Upon finding this reference one discovers that it is part of a footnote in an introductory chapter. In regards to nonparametric statistics these authors state simply "...methods which make fewer assumptions but are also less efficient."

Table 1

Empirical Cumulative Probabilities of *t* Statistics Which Have Normal Theory Cumulative Probabilities Given By Column Headings. Results Are From A Monte Carlo Study by Bradley (1964).

		Left-Tail Cum. Prob.				Right-Tail Cum. Prob.			
n_1,n_2	Pop.	.0100	.0250	.0500	.1000	.0100	.0250	.0500	.1000
4,2	L,L	.0126	.0208	.0364	.1477	.0054	.0148	.0276	.0585
6,2	L,L	.0113	.0260	.1010	.1550	.0045	.0119	.0213	.0465
4,4	L,L	.0026	.0118	.0228	.0723	.0043	.0117	.0259	.0739
4,2	L,N	.1081	.1514	.1942	.2505	.0296	.0440	.0623	.1029
6,2	L,N	.1294	.1644	.1989	.2372	.0379	.0530	.0729	.1108
4,4	L,N	.0390	.0742	.1127	.1727	.0052	.0089	.0215	.0582
4,2	N,L	.0069	.0121	.0189	.0506	.0112	.0263	.0515	.1028
6,2	N,L	.0088	.0119	.0272	.0744	.0054	.0123	.0277	.0693
4,4	N,L	.0051	.0091	.0210	.0577	.0388	.0692	.1065	.1710
8,4	L,L	.0064	.0263	.0451	.1439	.0022	.0062	.0142	.0414
12,4	L,L	.0161	.0336	.0872	.1401	.0010	.0031	.0083	.0248
8,8	L,L	.0016	.0072	.0250	.0941	.0021	.0081	.0264	.0983
8,4	L,N	.0739	.1086	.1450	.1938	.0089	.0171	.0332	.0780
12,4	L,N	.0742	.1021	.1301	.1768	.0105	.0215	.0398	.0865
8,8	L,N	.0357	.0652	.1009	.1611	.0017	.0077	.0236	.0681
8,4	N,L	.0018	.0081	.0230	.0601	.0117	.0303	.0602	.1207
12,4	N,L	.0039	.0138	.0326	.0738	.0031	.0114	.0344	.0898
8,8	N,L	.0018	.0096	.0243	.0653	.0368	.0650	.1045	.1662
16,8	L,L	.0100	.0339	.0663	.1314	.0006	.0020	.0097	.0600
24,8	L,L	.0223	.0415	.0755	.1269	.0003	.0006	.0030	.0385
16,16	L,L	.0028	.0145	.0452	.1118	.0025	.0143	.0418	.1057
16,8	L,N	.0473	.0728	.1038	.1532	.0043	.0119	.0288	.0742
24,8	L,N	.0447	.0676	.0971	.1464	.0055	.0151	.0347	.0832
16,16	L,N	.0274	.0516	.0830	.1369	.0018	.0099	.0251	.0701
16,8	N,L	.0042	.0134	.0290	.0713	.0119	.0284	.0581	.1226
24,8	N,L	.0073	.0175	.0381	.0821	.0037	.0147	.0399	.1052
16,16	N,L	.0027	.0090	.0272	.0710	.0263	.0514	.0823	.1346
32,16	L,L	.0160	.0364	.0670	.1203	.0002	.0046	.0260	.0971
48,16	L,L	.0200	.0400	.0658	.1135	.0002	.0027	.0187	.0947
32,32	L,L	.0072	.0221	.0531	.1116	.0067	.0216	.0481	.1047
32,16	L,N	.0350	.0613	.0896	.1375	.0030	.0099	.0305	.0791
48,16	L,N	.0304	.0537	.0834	.1338	.0042	.0135	.0340	.0843
32,32	L,N	.0237	.0480	.0765	.1300	.0035	.0129	.0336	.0773
32,16	N,L	.0049	.0144	.0371	.0834	.0105	.0317	.0630	.1205
48,16	N,L	.0086	.0207	.0421	.0881	.0060	.0213	.0508	.1127
32,32	N,L	.0025	.0110	.0318	.0788	.0208	.0433	.0709	.1210

Hence, it appears that some authors regard nonparametric tests as "quick and rough" procedures, inherently less powerful than their parametric counterparts.

Evidence Regarding the Relative Power of the *t*-Test and Wilcoxon's Rank-Sum Test

Writers such as those quoted above do not recognize, or at least do not make clear to the reader, the fact that the *t*-test's optimal power properties are attained *under normal theory*. Thus, there is no mathematical or statistical basis for the claim that the *t*-test is more powerful than its nonparametric counterparts when population shapes are unspecified. Even when this fact is recognized, authors sometimes seem reluctant to point out that nonparametric tests may be *more* powerful than their parametric counterparts in the nonnormal situation. For example, Runyon and Haber (1971) point out in a footnote, "It must be reiterated that the parametric tests are more powerful only when the assumptions underlying their use are valid. When the assumptions are not met, a nonparametric treatment may be *as powerful as* the parametric." [Italics added]

In point of fact, certain nonparametric tests may be much more powerful than their parametric counterparts when sampling is from a variety of non-normal distributions. For example, Blair and Higgins (1980a) used Monte Carlo techniques to study the relative power of Wilcoxon's rank-sum test and the independent means *t*-test under uniform, Laplace, half normal, exponential, mixed normal and mixed uniform distributions. Their study showed that the nonparametric procedure was frequently the more powerful test, and that the magnitude of the Wilcoxon's power superiority was often *very* large. (The difference in the proportion of false null hypotheses rejected by the two statistics was as large as .94. This advantage was in favor of the Wilcoxon test and occurred under the mixed normal distribution.) On the other hand, the *t*-test was seldom the more powerful test, and in those situations where it held the advantage, the magnitude of the power advantage was usually quite modest.

The empirical results obtained in the Blair and Higgins (1980a) study are in full accord with the asymptotic results obtained by Hodges and Lehman (1956). These latter authors showed that while the asymptotic relative efficiency of the Wilcoxon to the *t* approaches infinity, it can never be lower than .864. Asymptotic studies by Blair and Higgins (1980b) also show large advantages for the Wilcoxon test in situations where sampling is from mixtures of two normal populations. (A table showing the asymptotic relative efficiencies of various nonparametric tests to their parametric counterparts can be found in Bradley (1972).)

It appears, that there is no mathematical or statistical basis for the claim that parametric rather than non-parametric tests should be used in non-normal population situations because of efficiency advantages maintained by the parametric tests. Indeed, when the evidence concerning the relative power of Wilcoxon's rank-sum test and the *t*-test is examined, one must conclude that truly large power advantages can be gained by using the nonparametric procedure.

Conclusions

The anti-nonparametrics sentiments expressed above are not held by all authors of statistical texts designed for the social and behavioral sciences. However, these sentiments are held by many influential writers of texts and articles. Such writers often exaggerate or fail to adequately qualify their statements concerning the robustness of certain parametric tests and/or fail to recognize, or at least make clear to their readers, the fact that their claims of power superiority for parametric tests are based on normal theory assumptions. In point of fact, certain nonparametric tests, such as Wilcoxon's rank-sum procedure, not only maintain stable Type I error rates in the non-normal population situation but also show large power advantages over parametric counterparts in many circumstances.

It is hoped that this article will stimulate those who have come to think of nonparametrics as "quick and dirty" procedures to carefully reassess the bases for their attitudes. A reassessment of this type might well lead to the conclusion that nonparametric tests, like parametric tests, are valuable research tools and as such should not be dismissed so lightly.

Footnotes

[1] Box and Anderson (1955) also use the term "pathological" in connection with non-normal populations that may produce non-robust results for the Z test.

[2] Bradley (1978) has proposed a definition for robustness that would eliminate this ambiguity if generally accepted.

[3] See Glass et al (1972) for a review of this evidence.

References

Blair, R. C., and Higgins, J. J. (1980a), "A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's *t* Statistic Under Various Non-Normal Distributions," Journal of Educational Statistics, (in press).

Blair, R. C., and Higgins, J. J. (1980b), "A Note on the Asymptotic Relative Efficiency of the Wilcoxon Rank-Sum Test Relative to the Independent Means *t*-Test Under Mixtures of Two Normal Distributions," British Journal of Mathematical and Statistical Psychology, (in press).

Boneau, C. A. (1960), "The Effects of Violations of Assumptions Underlying the *t*-Test," Psychological Bulletin, 57, 49-64.

Box, G. E. P., and Andersen, S. L. (1955), "Permutation Theory In the Derivation of Robust Criteria and the Study of Departures From Assumptions," Journal of the Royal Statistical Society, 17, 1-34.

Bradley, J. V. (1964), "Studies in Research Methodology VI. The Central Limit Effect for a Variety of Populations and the Robustness of Z, *t*, and F," Aerospace Medical Research Laboratories Technical Report AMRL-TR-64-123.

Bradley, J. V. (1968), Distribution-free Statistical Tests, pp. 24-44. Englewood Cliffs, N.J.: Prentice-Hall.

Bradley, J. V. (1972), "Nonparametric Statistics," in Statistical Issues, A Reader for the Behavioral Sciences, ed. R. E. Kirk, Monterey, California: Brooks/Cole, 329-338.

Bradley, J. V. (1977), "A Common Situation Conducive to Bizarre Distribution Shapes," The American Statistician, 31, 147-150.

Bradley, J. V. (1978), "Robustness?," British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Glass, G., Peckham, P. D., and Sanders, J. R. (1972), "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance," Review of Educational Research, 42, 237-288.

Glass, G., and Stanley, J. C. (1970), Statistical Methods in Education and Psychology, Englewood Cliffs, N. J.: Prentice-Hall.

Guilford, J. P., and Fruchter, B. (1978), Fundamental Statistics in Psychology and Education (6th ed.), New York: McGraw-Hill.

Hawkridge, D. G. (1970), "Designs for Evaluative Studies," in American Institutes for Research, Evaluative Research, Palo Alto, California: AIR, 24-47.

Hodges, J. L., and Lehman, E. L. (1956), "The Efficiency of Some Nonparametric Competitors of the t-Test," Annals of Mathematical Statistics, 27, 324-335.

Hopkins, K. D., and Glass, G. (1978), *Basic Statistics for the Behavioral Sciences*, Englewood Cliffs, N. J.: Prentice-Hall.

Lindquist, E. F. (1953), *Design and Analysis of Experiments in Education and Psychology*, Boston: Houghton Mifflin.

Runyon, R. P., and Haber, A. (1971), *Fundamentals of Behavioral Statistics* (2nd ed.), Reading, Massachusetts: Addison-Wesley.