A Power Comparison Study of Parametric and Non-Parametric Tests Under Severe Violations of the Parametric Assumptions of Normality and Homogeneity

Richard Edward Ryan
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses
Part of the Educational Psychology Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/3286

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
A POWER COMPARISON STUDY OF PARAMETRIC AND NON-PARAMETRIC TESTS UNDER SEVERE VIOLATIONS OF THE PARAMETRIC ASSUMPTIONS OF NORMALITY AND HOMOGENEITY

by

Richard Edward Ryan

A Thesis
Submitted to the
Faculty of the School of Graduate Studies in partial fulfillment of the Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
May 1967
RYAN, Richard Edward
A POWER COMPARISON STUDY OF PARAMETRIC AND NON-PARAMETRIC TESTS UNDER SEVERE VIOLATIONS OF THE PARAMETRIC ASSUMPTIONS OF NORMALITY AND HOMOGENEITY.

Western Michigan University, M.A., 1967
Education, psychology

University Microfilms, Inc., Ann Arbor, Michigan
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>II</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td>III</td>
<td>19</td>
</tr>
<tr>
<td>Case I:</td>
<td>19</td>
</tr>
<tr>
<td>Case II:</td>
<td>21</td>
</tr>
<tr>
<td>Case III:</td>
<td>25</td>
</tr>
<tr>
<td>Case IV:</td>
<td>25</td>
</tr>
<tr>
<td>IV</td>
<td>34</td>
</tr>
<tr>
<td>V</td>
<td>39</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

My thanks for aid in writing this thesis go to the faculty and staff of the Department of Psychology of Western Michigan University for their invaluable inspiration, guidance, and training. In particular, my thanks go to Professor Eston J. Asher, my Major Thesis Advisor, and Professors Frank Fatzinger and John E. Nangle who served on my thesis committee. Another vote of special gratitude goes to Jack R. Meagher, Director of the Computer Center, and his entire staff for their contribution of time and effort towards those sections of the study which required the use of the department's IBM 1620 computer.

Richard Edward Ryan
A POWER COMPARISON STUDY OF PARAMETRIC AND NON-
PARAMETRIC TESTS UNDER SEVERE VIOLATIONS OF THE
PARAMETRIC ASSUMPTIONS OF NORMALITY AND HOMOGENEITY

One of the most important aspects of modern psychological research
has been the evaluation of obtained data through some accepted statis-
tical method. Statistics have quite often dictated the form and even
the outcome of research by limiting the investigator to a stereotyped
experimental paradigm. Thus, an experiment reveals results which are
applicable to a certain statistical test, and may be evaluated by that
test, but are not commensurate with reality or the purpose and interest
of the investigator. The researcher is forced to alter his experimen-
tal paradigm, to transform his original data in order to fit the sta-
tistical prerequisites, or to shop around for a test that will eval-
uate his data so that the results are not hopelessly compromised and
useless for generalization and comparison.

Two tests most commonly utilized by psychologists are the par-
ametric t and F tests. Both the t and analysis of variance have
sharply limiting prerequisites. Two assumptions are of statistical
import: (a) the variance must be the same for each treatment popula-
tion, (b) the sampling must come from normal populations. The pre-
requisites of normality and homogeneity of variance are often difficult
to assume since they are based on the chance shape of the sampling
distribution and the very real possibility that many populations under
psychological investigation do not show a normal distribution. This
problem has long disturbed psychological researchers and demanded a
practical solution.
As one solution, special $t$ tests have been constructed to counter the deleterious effect of non-normality and heterogeneity. Welch's (1947) approximation of Student's $t$ distribution, the $t'$ statistic, is recommended by Winer (p. 38, 1962) as the soundest method for testing the hypothesis about the difference between two means when the population variances are assumed to be unequal.

More recently, a subfield of statistics was introduced which claimed to avoid the pitfalls of assuming normality and homogeneity as well as the time and trouble of transforming data. Non-parametric (distribution free) statistics are usually not based on the original data obtained in the sample, but utilize ranks or algebraic signs. The meaningfulness of the result of a parametric test depends on the validity of the assumptions of normality and homogeneity. Since non-parametric tests made no such assumptions they were considered to be more useful and valid for research in the behavioral sciences.

Objections to non-parametric statistics have usually taken two major forms. One objection is the assumption that parametric tests, especially the $t$ test, are so robust that even glaring discrepancies in the assumptions of normality and homogeneity do not appreciably affect its statistical power (decision to reject the null hypothesis when it is actually false). The second major attack on the use of non-parametric statistics claims that since the distribution free methods disregard data and are wasteful of facts which the more refined $t$ and $F$ tests consider, they can be demonstrated to be substantially less powerful.

Pearson (1931) and Cochran (1947) both agreed that the $t$ and $F$ tests were robust enough to withstand minor deviations in normality.
and homogeneity. Cochran stated that, "No serious error is introduced by non-normality in the significance levels of the $F$-test or of the two-tailed $t$-test" (p. 24).

The first comprehensive compilation of non-parametric techniques was initiated by Lincoln Moses (1952) and presented an overall view of distribution free methods in research. He cited the following as advantages of the non-parametric methods: (a) whatever may be the form of the distribution from which the sample has been drawn, a non-parametric test of a specified significance level actually has that significance level; (b) if samples are small; e.g., six cases, there is in effect no alternative to a non-parametric test (unless the parent distribution really is known); (c) the methods are usually easier to apply than the classical techniques.

Non-parametric procedures were also recommended by Blum (1954) for use in behavioral research. Blum, after reviewing the literature on statistical treatments, called for widespread use of distribution free statistics in educational research since the basic assumptions underlying the parametric techniques could not be met.

Box (1954) investigated the effect of unequal variances upon the $t$ test. Box, using a variance differer . e ratio of 1 to 3 found that, "Moderate departures from assumptions do not seriously affect the accuracy of decisions by Means of the $t$ test". Box, like Pearson and Cochran, assures the researcher that the $t$ test is robust with respect to the parametric assumption of normality of distributions.

Considerations of parametric vs. non-parametric testing were profoundly influenced by two texts which appeared in the mid-1950's. In direct opposition to the positive recommendations of Moses and
Blum, Lindquist (1953) published the results of the Norton (1951) study. Norton had attempted to show that the $F$ test was robust enough to withstand violations of its basic assumptions. He randomly sampled from distributions having the same means but differing in variance and homogeneity in a predetermined fashion. Norton found that the $F$ test was very robust when both samples came from the same population, regardless of population shape. Furthermore, for sampling from populations having the same shape, but different variances; or having different shapes, but the same variance, there was little serious distortion of the outcome of the $F$ test. However, serious discrepancies resulted when sampling was from populations with both heterogeneity of variance and differing population shapes. Lindquist concluded that the $F$ test and the $t$ test were definitely robust enough to use, even when non-parametric methods would seem more applicable.

On the other hand, the use of distribution free methods was strongly endorsed by Siegle (1956) in his text, *Non-parametric statistics*. Siegle gathered all the non-parametric techniques into one text and made distribution free statistical methods easily available for the first time. Although he enthusiastically urged the use of non-parametrics, he rather ruefully bowed to the generally held opinion that distribution free methods were less powerful than their parametric counterparts. Siegle expounded on the necessity for larger $N$'s to offset this power loss.

Gaito (1959), in his refutation of Siegle's work, agreed with Lindquist that there should be only a limited use of non-parametric statistics due in part to the wastefulness of non-parametric methods.
and in part to the robustness of the $t$ and $F$ tests. In Gaito's work the two major objections to non-parametric statistics were combined for the first time. Gaito mentioned Siegle's own admission that non-parametric tests are less powerful than parametric ones. However, in discussing the robustness of $t$ and $F$ tests relative to violations of normality and homogeneity, Gaito cautioned that when sample sizes differ appreciably, this robustness may not manifest itself. "If the numbers within the groups differ greatly (which is usually not the case) deviations from normality and homogeneity of errors will have a greater effect" (p. 118). In light of this problem, Gaito urged that when normality or homogeneity are in doubt, extremely large and equal sample sizes should be obtained so as to bring into play the mollifying effect of the Central Limits Theorem (means of large enough samples from any population are normally distributed). Gaito, in effect, condemned small sizes and non-parametric statistics while urging the use of the traditional $t$ and $F$ tests, bolstered by large sample sizes.

Both Edwards (1960) and Anderson (1961) agree with Gaito's evaluation of the robustness of the $t$ and $F$ tests. However, they differ in their assessment of non-parametric power. Edwards followed the traditional view that non-parametric statistics are less powerful, not only when the normality and homogeneity assumptions have been established, but even when they are not established. Anderson, on the other hand, quoting the work of Dixon and Massey (1957) finds that at least for rank order methods, non-parametric tests are nearly as powerful as parametric tests even under equinormality. He does maintain, however, that the loss of power involved in dichotomizing data for a median-type test is considerable. Anderson concludes that
parametric tests are more versatile and meet the everyday needs of psychology in a superior fashion than the non-parametric tests. Until non-parametric tests are produced which match this versatility, the author says, "non-parametric tests are to be considered only as useful minor techniques in the analysis of numerical data" (p. 315).

In answer to Gaito and Edwards' charge of the superiority of parametric statistics in power considerations, Hodges and Lehmann (1961) published further data on the power question. The authors reported to the Berkeley Symposium on Mathematical Statistics and Probability that two non-parametric tests, the Normal Scores Test and the Wilcoxon Test, had been experimentally shown to have more power than the t test when the assumptions of the t test were violated.

Furthermore, Boersma, DeJonge and Stellwagen (1964), upon comparing the power of the omnibus F test and the non-parametric L test, under conditions of strict normality and homogeneity of variance, found that the non-parametric test was more powerful when there was a monotonic order among treatment means.

Definitive studies were carried out by Boneau (1960, 1962), in an attempt to settle the power controversy. In 1960, Boneau attempted to evaluate the effect of heterogeneity and non-normality on the parametric t test. On the results of this preliminary study, Boneau reaffirmed the robustness of the t test. Boneau concluded that normality and heterogeneity may be readily violated in the situation where: (a) the two sample sizes are equal, and (b) the assumed underlying population distributions are of the same shape. "If these conditions are met, then no matter what the variance differences may be, samples of as small as five cases will produce results for which the true probability
of rejecting the null hypothesis at the .05 level will more than likely be within .03 of that level."

In the 1962 study, Boneau modifies his evaluation of the robustness of the parametric \( t \) test by comparing it with the non-parametric \( U \) test in situations where heterogeneity and normality have been specifically introduced. Attention was brought to bear on the fact that in specific instances the Wilcoxon-Mann-Whitney \( U \) test is as powerful as the normal \( t \) test.

The relative power displayed by the simple \( t \) test when presented with grave violations of its basic assumptions left the usefulness of parametric procedures in grave doubt. Boneau (1962) had tested the \( t \) and \( U \) tests under most of the likely distortions which could occur in sampling from two unknown distributions. The \( t \) and \( U \) tests were examined under: (a) equal sample sizes, heterogeneity, non-normality, (b) unequal sample sizes, homogeneity, non-normality, and (c) unequal sample sizes, heterogeneity, normality. Boneau found that in all such cases the parametric \( t \) test had slight, but non-significant, power advantages over the non-parametric \( U \) test.

The present paper was intended to explore further the relationship between parametric and non-parametric tests when the assumptions of the parametric tests have been systematically distorted. There were two general questions to be answered: (1) how does the power of the \( t \) and \( U \) tests compare when certain severe violations are made in parametric assumptions?; and, (2) do the special \( t' \) tests, devised by Welch (1947) and reported by Winer (1962), give any more powerful results than the "regular" \( t \) test?
Although non-normality and heterogeneity are serious violations of the assumptions underlying parametric testing they do not seriously affect the outcome of the robust \( t \) test. However, there are further considerations which, combined with non-normality and heterogeneity, do great damage to the confidence levels expressed by statistical testing. Boneau (1962) concludes that when sample sizes are as large as 25 or 30, the \( t \) test becomes absolutely non-parametric due to the tremendous influence of the Central Limit Theorem. Therefore, the larger the sample size, the easier it is for the researcher to ignore any violation of parametric assumptions and use the \( t \) test for any data. However, in a large proportion of present-day research, large sample sizes are not available. Paradoxically, this unavailability of subjects is usually linked to populations whose normality and homogeneity of variance are under severe question. How does the power of a parametric test compare to a non-parametric test under a set of circumstances where sample sizes are small?

Another severe restriction to the use of "robust" parametric tests under conditions of non-normality and heterogeneity is the equality of sample sizes. In both of their respective endorsements of the robustness of the normal \( t \) test, Gaito (1959) and Boneau (1962) took special pains to point out the dangers in using a parametric test if a combination of non-normality, heterogeneity and unequal sample size were encountered. In such a situation (where small random samples of unequal sizes are drawn from two populations differing in shape and variance) the power of the parametric test would be severely evaluated in comparison to the power of a non-parametric test.
A second objective of this paper is to evaluate the special \( t \) tests which have been constructed to overcome some of the limitations imposed by normality and homogeneity of variance. It is of special interest to compare the power of the normal \( t \) with the power of the \( t \) tests in the situation described above, where the power of a parametric test meets its most difficult challenge.
METHOD

In order to compare the ability of the various test to withstand violations of heterogeneity and non-normality, it was necessary to construct data which would display the necessary aberrations, and yet be truly random and consistent with proper sampling techniques. The procedure, initiated by Boneau in his 1962 study, utilizes the storage, speed and versatility of a computer to translate theoretical considerations into empirical data.

To properly evaluate the relative powers of the $t$, $t'$, and $U$ tests it was necessary to compute a large number of $t$, $t'$, and $U$ values, each based upon samples drawn at random from distributions having specified characteristics. The present study was performed on the IBM 1620 computer programmed to perform the following operations: (a) the generation of a random number, (b) the transformation of this random number into a random deviate from the appropriate distribution (rectangular or normal), (c) the accumulation of random deviates until the proper sample size is reached for each of two distributions (normal and rectangular), (d) the computation of a $t$, $t'$, and $U$ statistic for each set of data, and (e) the construction of a frequency diagram of resultant $t$, $t'$, and $U$ scores for use in the comparison of the relative power of parametric and non-parametric tests. This operation was performed internally and the results were punched out on IBM cards.

Amplification of the above operations is necessary and will be presented in respective order.
(a) It was necessary to place in the computer a 10 digit number taken from a table of random numbers. This number is then multiplied by one of a sequency of permutations of the 10 digits (0, 1, 2, 3, ..., 8, 9) randomly selected by the machine. The random number to be utilized consists of the middle 10 digits of the product of the previously generated random number and the selected permutation. The randomness of numbers generated in this manner was tested in 1962 by Boneau by sorting 5000 of the numbers "into 50 categories on the basis of the first 2 digits. A Chi Square test was then performed to determine the fit of the obtained distribution to a theoretical one consisting of 100 scores in each of the 50 categories" (p. 52). Boneau reports that the obtained Chi Square of 47.83 is extremely close to the theoretical median of the Chi Square distribution with 50 degrees of freedom; 49.332.

(b) The next step was to obtain the individual random deviates (the scores from the appropriate population). The random numbers obtained in the fashion described above were considered to be numbers between 0 and 1 and interpreted as "the cumulative probability for a particular score from the prescribed population" (Boneau, 1960, p. 52). Individual random scores for the normal and the rectangular distributions having that probability were selected from tables inserted in the machine. This is the same procedure used when one enters the ordinary z table to obtain the z score associated with a proportion of the area under the normal probability curve, such as .42220 is associated with the score 1.42. In this case the random deviate 1.42 corresponds to the cumulative percentage .42220.
The populations chosen for this study were the normal and the rectangular. These present a comparison in flatness of curve, especially as the variance differs for the rectangular distribution. The tables of deviates corresponding to the two populations were constructed so that the mean of each population was 0 and the variance 1 (as is the case with the z distribution for the normal curve). In order to change the size of the variance (when necessary) all random deviates were multiplied by 2 (the standard deviation of the variance) to obtain a variance of 4. Only variances of 1 and 4 were used; this contrast being enough to emphasize the distorting results of heterogeneity.

(c) The sample sizes chosen were 4 and 8. Since small sample size is harmful to the parametric \( t \) and \( t' \) tests and combines with non-normality and heterogeneity to disrupt these significance tests, such sample sizes were utilized. The difference in sample size should also act to further disrupt the parametric test, since the larger sampling might possibly come from a distinctly non-normal and heterogeneous sample. When the computer had provided the necessary random deviates from each population to make up the correct sample size, it then performed three statistical tests on this data.

(d) The first test, the normal \( t \) test, was taken from Boneau's (1960) work and is the regular textbook test for computing the \( t \) value.

\[
t = \frac{M_1 - M_2}{\sqrt{\frac{\sum X_1^2 - N_1 M_1^2}{N_1} + \frac{\sum X_2^2 - N_2 M_2^2}{N_2} - \frac{1}{N_1} + \frac{1}{N_2}}}
\]
The second test, the $t$ test, for testing the hypothesis about the difference between two means assuring that population variances are not equal, was taken from Winer (1962, p. 37) and is as follows:

$$t^* = \frac{(\overline{x}_a - \overline{x}_b) - (a - b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

where the degree's of freedom must be calculated by Welch's (1949) formula:

$$f = \frac{f_a f_b}{f_b c^2 + f_a (1-c)^2}$$

where $f_a = n_a - 1$

$$f_b = n_b - 1$$

and where $c = \frac{s_a^2}{n_a} \div \frac{s_b^2}{n_b}$

This test should throw an interesting light on whether an improved, but still parametric, $t$ test which was constructed to withstand the violation of heterogeneity has power comparable to the normal $t$ and to a non-parametric test in an extreme violation of the assumptions underlying the parametric test.

The third test is the non-parametric $U$ test of Mann and Whitney. This test is described by Dubois (1965, p. 476) and the procedure involves a ranking technique. Both sets of data, from the two distributions, are ranked in a single series from low to high. Two
sums of ranks are then obtained: $T_n$ for the sample of $n$ cases and $T_m$ for the sample of $m$ cases. $U$ is the smaller of the following quantities: 

(a) 

$$U = nm + \frac{n(m+1)}{2} - T_n$$

(b) 

$$U' = mm + \frac{m(m+1)}{2} - T_m$$

The $U$ test is one of the best non-parametric examples. Not only is it based on ranking techniques which are almost as powerful as parametric techniques (see Anderson, 1961), but Siegle (1956) has published complete tables for the significance level instituted by Welch. Boneau (1962) calls the $U$ test "a worthy protagonist". Hodges and Lehmann (1956) mention that even under strict normality the $U$ test is 95\% as powerful as the $t$ test.

(e) After the computer had taken the samples (from the appropriate distributions) 1000 times and had calculated 1000 $t'$s, 1000 $t''$s, and 1000 $U'$s, each based on the same data, it delivered a frequency distribution of the $t'$s, $t''$s, and $U'$s for inspection. This study investigated four sampling conditions. It was necessary to run each condition four separate times (to obtain power data — see below). Therefore, each sampling condition resulted in four sets of 1000 $t'$s, 1000 $t''$s, and 1000 $U'$s. Since it took 12,000 random deviates (sample sizes 8 and 4) to produce one set of 1000 $t'$s, $t''$s, and $U'$s, each sampling condition required the generation of 48,000 random numbers, the translating of these numbers to 1000 statistical test procedures (times three since there were three tests used). The time required
to run such a sampling condition, even when optimally programmed on
the IBM 1620, was 18 hours. The total operation for all four conditions
lasted 72 hours, required the generation of 192,000 random numbers and
random deviates, the working of some 12,000 statistical tests, and
utilized 48,000 IBM cards.

At this point, it is advisable to discuss the manner in which the
comparative power of the three tests was evaluated.

In order to facilitate the explanation of this technique, it is
necessary to introduce a nomenclature system devised by Boneau and
used in his 1962 study. The conditions of sampling are symbolically
represented in the following manner. For example, \( N(0,1)4 - R(0,1)8 \)
indicates that the first sample is from a normal population with a
mean of zero and a variance of 1, the sample size being 4. The second
sample (of the two upon which the \( t, t^*, \) and \( \Pi \) tests are run) came
from a rectangular distribution with a mean of 0, a variance of 1, the
sample size being 8. Since the original steps of the program assure
that the samples chosen will be either normal with a mean of 0 and a
variance of 1, (as in the \( z \) distribution) or from a rectangular dis-
tribution (equi-probability of selecting any score) with a mean of 0
and a variance of 1, it is quite simple to change either the variance
or the mean of any sample through the usual statistical rules of
multiplication and addition. This is the method by which the various
variances were assigned to the appropriate samples. For example,
\( N(0,1)4 - R(0,1)8 \) is easily translated to the distribution \( N(0,4)4 - R(0,1)8 \)
(normal population, 0 mean, variance of 4, with sample size 4 —
rectangular population, 0 mean, variance of 1, with sample size 8) by
multiplying all the individual scores of the normal sample by 2 (the
standard deviation of the variance) to assign a variance level of 4 to that sample. In like manner, the means of a sample may be changed through addition and in this manner the power of certain statistical tests may be measured. This method, introduced by Boneau (1960), utilizes the ability of the investigator to alter the means of certain samples at will. Since power may be defined as the "decision to reject the null hypothesis when it is indeed false", the experimenter has only to create a situation where the null hypothesis is indeed false by a predetermined amount, run several statistical tests on these samples and compare the number of times a certain test rejects the null hypothesis for the same data. The results will be an empirical comparison of the actual power expressed in any given test. This empirical comparison may be accomplished by altering the mean for any given sampling condition. In the following text and figures, the letter x will be inserted in any symbolic reference to a sampling condition to indicate that it takes on the necessary values; $N(x,1)4-R(0,1)8$. Four mean difference levels were used for each sampling consideration — 0, 1, 2, 3. Figure 1 pictures the frequency distribution received from the t test for the sampling condition $N(x,1)4-R(0,4)8$. As can be seen when the null hypothesis (the equality of means of the two samples) is true, most of the cases fall in the region of acceptance. The proportion falling in the region of rejection when the null hypothesis is true is .032. This is the empirical "Alpha" level and can be compared to the theoretical level of the two-tailed t test which is .050 at this level by definition. As the difference in means grows larger, the power of the tests also grows in comparison until at the mean difference level of 3, the t test is very powerful indeed.
FIGURE 1. Empirical t distribution for mean difference of (a) 0.00, (b) 1.00, (c) 2.00, (d) 3.00 for sampling consideration $N(x_1)4-R(0,4)8$. The area to the right of the vertical line is the region of rejection for the .05 level of significance.
To recapitulate, the data for this investigation must be created. A 10 digit random number was generated by a multiplication process and converted to random deviates from specific populations. These random numbers were further altered, at necessity, through multiplication and addition to conform to the specific sampling conditions required. These "created" random deviates were injected into the computing formulas for for \( t \) test, the \( t' \) tests, and the Mann-Whitney \( U \) test. The powers of the three tests (ability to reject a null hypothesis when it is indeed false) were compared under sampling conditions of heterogeneity, non-normality and heterogeneous sample sizes. Four sets of 1000 \( t' \)'s, \( t'' \)'s, and \( U \)'s were collected for each sampling condition. These results were tabulated by the computer and arranged in frequency distributions. The empirical power levels were obtained by running four sets of 1000 \( t' \)'s, \( t'' \)'s, and \( U \)'s for each sampling condition; while altering the actual mean difference from 0 through 4. The empirical power values considered as functions of the actual difference between population means are the data of this study.
RESULTS

In all, four sampling considerations were used to test the powers of the $t$, $t'$, and $U$ statistics. The samples were taken from the normal and rectangular populations. Variance and sample sizes were interchanged to make up the four considerations.

Case I: $N(x, 1) \sim R(0, 4) 8$

The first sampling consideration tested was the case $N(x, 1) \sim R(0, 4) 8$, where four samples were taken from the normal population (mean $x$, variance 1) and eight samples from the rectangular population (mean 0, variance 4). Figure 1 pictures the frequency diagrams produced for the $t$ scores with actual mean differences - 0, 1, 2, 3. Figure 2 depicts the power functions that resulted from the empirical comparison of the $t$, $t'$, and $U$ tests. Actual values used in Figure 2 (as for all power comparison graphs) are listed in Appendices A through D. As can be noted in Figure 2, the empirical Alpha levels (the proportion of cases which fall in the region of rejection when the null hypothesis is true, Mean\(^1\) = Mean\(^2\)) which are diagrammed at the 0 level on the abcissa, all fall below the theoretical .050 power level for the three tests. The $t$ test has an empirical power of .032, while the $t'$ has an empirical power level of .030, and the $U$ test has an empirical level of .029. As can be noted in Figure 2, the relative levels of power remain the same throughout the diagram at the mean difference levels of 1, 2, and 3. Boneau, in his 1962 study, comments on this phenomenon, "The violation of this assumption (homogeneity of
FIGURE 2. Empirical power functions for $t$, $t'$, and $U$ tests for sampling consideration $N(x,1)4-R(0,4)8$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
variance) coupled with heterogeneous sample sizes changes the Alpha level of both the \( t \) and the \( U \) tests, and produces power functions which seemingly are roughly appropriate for the true Alpha level rather than the nominal one" (p. 253). Since the Alpha level powers are not equal, any comparison of power is meaningless unless the data is corrected.

Based on the hypothesis of Boneau and on the evaluation of the empirical data, Figure 3 depicts the power comparison levels for \( N(x,1)4-R(0,4)8 \), corrected at all levels in the same proportion as that of the empirical Alpha level in its relation to the theoretical .05 level. As can be noted in Figure 3, the power of the corrected data shows that the tests now reach maximum power at a mean difference of two rather than three. However, there is little or no power difference among the competing tests.

Case II: \( N(x,4)4-R(0,1)8 \)

The second sampling consideration tested was the case \( N(x,4)4-R(0,1)8 \), where four samples were taken from the normal population (mean \( x \), variance 4) and eight samples were taken from the rectangular population (mean 0, variance 1). Figure 4 depicts the empirical frequency diagrams received for the \( t \) scores with actual mean differences of 0, 1, 2, 3. Figure 4 shows the moderate flattening effect which the violations of assumptions have had on the distribution of the \( t \) scores, an effect which is depicted in the heightened Alpha level of the power functions. Figure 5 demonstrates the empirical power functions for this sampling consideration. As can be noted, the \( t \) test seems much more powerful than either the \( t' \) or the \( U \) tests. However, when the Alpha level power of \( t \) is considered (.171) in comparison
FIGURE 3. CORRECTED power functions for $\bar{x}$, $\bar{x}'$, and $\bar{u}$ tests for sampling consideration $N(x, 1) - R(0, 4)$.
FIGURE 4. Empirical \( t \) distribution for mean difference of (a) 0.00, (b) 1.00, (c) 2.00, (d) 3.00 for sampling consideration \( H(s,4) \sim N(0,1) \). The area to the right of the vertical line is the region of rejection for the .05 level of significance.
FIGURE 5. Empirical power functions for $t$, $t^*$, and $U$ tests for sampling consideration $N(x, \sigma^2) \sim \mathcal{N}(0,1)^8$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
to the Alpha level power of $t^*$ (.057) or $\bar{U}$ (.095), it can be seen that $t$ is a much less conservative test than $t^*$ for this empirical data.

The $t$ test, as established by the empirical Alpha level, is not operating at its assigned .05 power level. When the proper corrections are established, Figure 6 depicts the actual power comparison. In Figure 6, the $t^*$-test is shown to be more powerful than the $\bar{U}$ test and both are more powerful than the $t$, when all are operating at the .05 level.

Case III: $N(x,1)\&-R(0,4)$

The third sampling consideration, $N(x,1)\&-R(0,4)$, has much in common with the second consideration. The frequency distributions for the $t$ test exhibit a somewhat flattened image which heightens the Alpha level of the tests as depicted in Figure 7. The power functions for this sampling consideration (Figure 8), show that both the $t$ and $\bar{U}$ tests display a superiority over the $t^*$ test at the mean-difference level of two. As in the second consideration, the power of $t^*$ seemingly falls behind. However, when the corrected levels are examined, Figure 9 depicts a definite rise in the power of the $t^*$ and the three tests seem to show almost equal power.

Case IV: $N(x,4)\&-R(0,1)$

The fourth and final sampling consideration tested was the case $N(x,4)\&-R(0,1)$, where eight samples were taken from the normal population (mean $x$, variance 4) and four samples from the rectangular population (mean 0, variance 1). Figure 10 depicts the empirical frequency distributions for the $t$ statistic. Figure 11 shows the
FIGURE 6. CORRECTED power functions for $t$, $t'$, and $U$ tests for sampling consideration $N(x, \sigma) \cup R(0,1)$. Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 7. Empirical $t$ distribution for mean difference of (a) 0.00, (b) 1.00, (c) 2.00, (d) 3.00 for sampling consideration $N(x,\sigma)\sim\sigma(0,4)$. The area to the right of the vertical line is the region of rejection for the .05 level of significance.
FIGURE 8. Empirical power functions for $t$, $t'$, and $U$ tests for sampling consideration $N(x,1)$-$R(0,4)$.
Figure 9. CORRECTED power functions for $t$, $t'$, and $U$ tests for sampling consideration $\mathcal{N}(x,1)\& R(0,4)$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 10. Empirical distribution for mean difference of (a) 0.00, (b) 1.00, (c) 2.00, (d) 3.00 for sampling consideration $N(x, \sigma^2) - R(0,1)$4. The area to the right of the vertical line is the region of rejection for the .05 level of significance.
FIGURE 11. Empirical power functions for $t$, $t'$, and $U$ tests for sampling consideration $N(x,4)\& R(0,1)\&$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
power functions resulting from the empirical comparison of the three previous considerations in that there are differing Alpha levels, and the test which holds the highest Alpha level maintains its superiority throughout. In this case, the $t^*$ test displays a higher Alpha level (although all three tests are actually lower in power than the theoretical value of .05) and the power functions for the mean difference levels of 1, 2, and 3 seem to uphold Boneau's conclusion that the actual Alpha level is maintained throughout. Figure 12 shows the corrected data and the $t$ test shows superior power, especially at the important lower mean-difference levels.
FIGURE 12. CORRECTED power functions for $t$, $t'$, and $U$ tests for sampling consideration $H(x,4)\in\mathcal{N}(0,1)$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
DISCUSSION

Figures 13 and 14 present composite pictures of the power frequencies for all four sampling conditions; both original data (Figure 13) and data corrected to the .05 level (Figure 14).

The corrected data (Figure 14) shows that the statistical tests appear more powerful when the deviant variance (4) is associated with the larger sample size. Sampling considerations 1 and 4 both seem to attain excellent power curves at the mean-difference level of two. On the other hand, in situations such as sampling considerations 2 and 3, where deviant variance (4) is associated with the smaller sample size, the power of all the tests lag behind. It is interesting to note, when comparing individual tests, that the only situation where the \( t \) and \( U \) tests (the two tests designed specifically for violations of normality and homogeneity) are seen to exceed the power of the \( t \) test, is in the situation where the extreme variance is associated with the smaller sample size (conditions 2 and 3). Considering what is known of the Central Limits Theorem, we might suspect that in conditions 1 and 4, the larger sample sizes reduced the violating effect of the deviant variance, producing a more nearly normal situation than in conditions 2 and 3, where the deviant variance was associated with the smaller sample size. The data would seem to bear out this conjecture since the \( t \) test, even after correction, (Figure 14, conditions 1 and 4) holds a positive power advantage, although not be much, over the other two tests.
FIGURE 13. Composite of Empirical power function graphs for 4 specified
sampling conditions: Cases I, II, III, IV.
FIGURE 14. Composite of CORRECTED power function graphs for 4 specified sampling conditions: Cases I, II, III, IV.
However, the consideration of corrected power functions is actually meaningless to the practical investigator. He has only empirical data and no "Alpha" level results to correct his findings. Figure 13 shows vastly different results than Figure 14; and Figure 13 is the actual empirical data received in this study. It seems apparent that the violations of non-normality, heterogeneity, and differing sample sizes radically change the theoretical significance level of the tests. It also seems apparent that the relative effects of this significance alteration remain constant throughout all mean-difference levels. In this consideration lies the practical problem for the psychological investigator. In sampling consideration 3, for example, (Figure 13) the t test appears the most powerful and if utilized, in this situation at least, would indeed show highly significant results. Yet these results would be a prime example of a Type I error, described by Dubois (1965) and Weinberg and Schumaker (1962), as more damaging to psychological research than the Type II error. The empirical level of significance for t in situation 3 is not .05, but .171, over three times the proper theoretical level. For this same situation, U shows an empirical Alpha level of .095. An examination of the charts and the data in Appendices A, B, C, and D reveals that the t test is highly unstable when presented with serious violations. Both the U test and the t' test are also affected, but to a lesser degree. The t' seems to be the most conservative of all three tests, and, upon comparison, deviates the least from the theoretical .05 level. (t deviates, in total, for all power sampling considerations .176 units from .05; U deviates, in total, .088; t' deviates, in total, .056 - See Appendix E).
Practical considerations, therefore, seem to dictate the conclusion that although $t$, $t'$, and $U$ have little difference in power when extreme violation of assumptions (small samples of unequal sizes drawn from populations differing in shape and variance) is introduced into a sampling situation; power is not the final criterion. Most texts on statistics urge investigators to use conservative tests, to err in the Type II manner rather than in the Type I. The $t'$ and $U$ tests provide a more conservative, less wandering significance test with no appreciable lack of power when extreme violations to the assumptions of normalcy and homogeneity are introduced.
SUMMARY

Recent investigations have cast doubt on the usefulness of the non-parametric statistics, claiming that the $t$ test is robust enough to compete with the power of non-parametric tests even under severe violations of normality and homogeneity.

It has been noted, however, that both (a) small sample sizes and (b) unequal sample sizes have a more severe effect on the parametric tests when combined with non-normality and heterogeneity.

This study attempted to evaluate three tests, the $t$ test, the $t'$ test (parametric), and the Wilcoxon $U$ test (non-parametric). Using a computer technique, the three tests were evaluated for their power strength under severe conditions of (a) non-normality, (b) heterogeneity, (c) small sample size, and (d) unequal sample size.

The results showed that, under severe violation of assumptions, all three tests fluctuated in their empirical significance levels and maintained this variance throughout the entire sampling consideration. In effect, the tests acted as if they were operating at the significance level they displayed for the mean-difference level of 0.00 (when $M^1 = M^2$). When the power functions were corrected, it was seen that for situations where extreme variance was coupled with small sample size the $U$ test (non-parametric) and the $t'$ test (modified parametric) showed greater power than the $t$ test (classical parametric). However, when the extreme variance was connected with the larger sample size, the $t$ test showed a superiority over the $t'$ and $U$ tests. All differences in power, however, were slight.

39

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In conclusion, it was shown that any power comparison from empirical data under the conditions described above was meaningless. Any test might show more power merely because it was operating at a higher significance level for that particular set of data. The $t_1$ and $U$ tests were recommended for use when data is suspected of severe violations since these two tests were generally more conservative and showed considerably less variation in significance levels than did the $t$ test.
REFERENCES


41


Welch, B. L., The generalization of Student's problem when several different population variances are involved. *Biometrika*, 1947, 34, 28-35.


## APPENDIX A

<table>
<thead>
<tr>
<th>MEAN-DIFFERENCE LEVEL</th>
<th>EMPIRICAL</th>
<th>CORRECTED TO .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>.021</td>
<td>.050</td>
</tr>
<tr>
<td>t*</td>
<td>.039</td>
<td>.050</td>
</tr>
<tr>
<td>ALPHA U</td>
<td>.035</td>
<td>.050</td>
</tr>
</tbody>
</table>

| 1.00                   | t         | .091            | 216 |
| 1.00                   | t*        | .144            | 194 |
|                        | U         | .136            | 144 |

| 2.00                   | t         | .387            | 921 |
| 2.00                   | t*        | .523            | 644 |
|                        | U         | .451            | 670 |

| 3.00                   | t         | .778            | 1.000 |
| 3.00                   | t*        | .883            | 1.000 |
|                        | U         | .816            | 1.000 |

Data for power functions for sampling consideration $N(x, 4) \& R(0, 1)$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>MEAN-DIFFERENCE</th>
<th>EMPIRICAL</th>
<th>CORRECTED TO .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEVEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>.058</td>
<td>.050</td>
</tr>
<tr>
<td>t*</td>
<td>.032</td>
<td>.050</td>
</tr>
<tr>
<td>ALPHA</td>
<td>.057</td>
<td>.050</td>
</tr>
<tr>
<td>t</td>
<td>.290</td>
<td>.250</td>
</tr>
<tr>
<td>t*</td>
<td>.118</td>
<td>.195</td>
</tr>
<tr>
<td>1.00</td>
<td>.223</td>
<td>.184</td>
</tr>
<tr>
<td>t</td>
<td>.796</td>
<td>.686</td>
</tr>
<tr>
<td>t*</td>
<td>.408</td>
<td>.633</td>
</tr>
<tr>
<td>2.00</td>
<td>.722</td>
<td>.637</td>
</tr>
<tr>
<td>t</td>
<td>.993</td>
<td>.856</td>
</tr>
<tr>
<td>t*</td>
<td>.904</td>
<td>.870</td>
</tr>
<tr>
<td>3.00</td>
<td>.992</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Data for power functions for sampling consideration $N(x,1)$ & $R(0,4)$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Data for power functions for sampling consideration N(x,4)4-R(0,1)8.
### APPENDIX D

<table>
<thead>
<tr>
<th>MEAN-DIFFERENCE LEVEL</th>
<th>EMPIRICAL</th>
<th>CORRECTED TO .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>.032</td>
<td>.050</td>
</tr>
<tr>
<td>( t^* )</td>
<td>.030</td>
<td>.050</td>
</tr>
<tr>
<td>( \text{ALPHA} )</td>
<td>.029</td>
<td>.050</td>
</tr>
</tbody>
</table>

| 1.00                    |           |                 |
| \( t \)                | .215      | .335            |
| \( t^* \)              | .194      | .334            |
| \( \text{ALPHA} \)     | .194      | .303            |

| 2.00                    |           |                 |
| \( t \)                | .746      | 1.000           |
| \( t^* \)              | .672      | 1.000           |
| \( \text{ALPHA} \)     | .708      | 1.000           |

| 3.00                    |           |                 |
| \( t \)                | .989      | 1.000           |
| \( t^* \)              | .972      | 1.000           |
| \( \text{ALPHA} \)     | .957      | 1.000           |

Data for power functions for sampling consideration \( N(x,1) \sim \mathcal{N}(0,4) \).
### APPENDIX E

<table>
<thead>
<tr>
<th>SAMPLING CONSIDERATION</th>
<th>EMPIRICAL ALPHA LEVEL</th>
<th>UNITS FROM .05 THEORETICAL LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $N(x,4.8-R(0,1.4)$</td>
<td>$t$</td>
<td>.021</td>
</tr>
<tr>
<td></td>
<td>$t'$</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>$U$</td>
<td>.035</td>
</tr>
<tr>
<td>2. $N(x,1.8-R(0,4.4)$</td>
<td>$t$</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td>$t'$</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>$U$</td>
<td>.057</td>
</tr>
<tr>
<td>3. $N(x,4.4-R(0,1.8)$</td>
<td>$t$</td>
<td>.171</td>
</tr>
<tr>
<td></td>
<td>$t'$</td>
<td>.057</td>
</tr>
<tr>
<td></td>
<td>$U$</td>
<td>.095</td>
</tr>
<tr>
<td>4. $N(x,1.4-R(0,4.8)$</td>
<td>$t$</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>$t'$</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>$U$</td>
<td>.029</td>
</tr>
</tbody>
</table>

**TOTALS:** $t$ test | .176 | $t'$ test | .088 | $U$ test | .056

Deviation of statistical tests from theoretical Alpha level of .05 in sampling considerations I - IV.