Florida Journal of Educational Research
Fall 1985, Vol. 27, No. 1, Pp. 83-97

A Comparison of Random Normal Scores Test Under the F and Chi-Square Distributions to the 2x2x2 ANOVA Test

Shlomo Sawilowsky
University of South Florida

ABSTRACT. The use of the parametric ANOVA Test when the underlying population is non-normally distributed is a violation of the test's assumption of normality. Alternatives include the use of nonparametric procedures. However, to date, useful procedures have not been developed to detect interactions, especially those of the higher order. Based on the rank transform, the Random Normal Scores Test has been suggested as a powerful alternative to the ANOVA Test. The major support rests upon asymptotic theory. This study is an empirical analysis of the Random Normal Scores Test performed under the F and Chi-square distributions. In the balanced 2x2x2 layout, for various population distributions, sample sizes, and nominal alpha levels selected, the Random Normal Scores Tests were shown to be non-robust and not powerful alternatives to the ANOVA Test.

Researchers in education and related disciplines have long been concerned by the existence of non-normally distributed variables (Bradley, 1968, 1977; Blair, 1981; Bloom, 1984; and Walberg, Strykowski, Rovai, and Hurg, 1984). This concern is with using parametric procedures that have been derived under the assumption of population normality on variables that are non-normally distributed. Instead, when testing hypotheses of shift in location parameter, researchers have advocated the use of nonparametrics. The advantage of nonparametric tests is that they do not make the assumption that the data

were sampled from a normal population.

In the analysis of variance layout, Iman (1974) and Iman and Conover (1976) proposed the use of the rank transformation as a solution strategy. Some theoretical evidence has been advanced to support the efficacy of their technique (Iman, Hora, and Conover, 1984). In fact, the Asymptotic Relative Efficiency (ARE) of the Rank Transformation Test, in comparison to the ANOVA Test, has the potential to be as high as infinity under certain non-normal distributions. A drawback, at least in theory, is that in performing the transformation, the assumption of independence of observations is violated. The effects of this violation are not known. For an empirical study of the rank transformation in higher-order analysis, see Sawilowsky, 1985.

Although the use of the rank transform has some promise for solving the analysis problem, there is the theoretical possibility of a more powerful procedure. Ranks are uniform, evenly spaced values that do not reflect the nature of the normal distribution. The impact of this condition, under the normal distribution, is that the ARE falls below 1.0 (Puri, 1964). Fisher and Yates (1949) noted that the integration of the rank transform with the substitution of "normal scores" would raise the ARE, under the normal distribution, to 1.0. Of the various types of normal deviates, Bell and Doksum (1965) suggested the use of random normal scores. The rationale, then, for using the random normal scores approach is as follows: a researcher does not know the nature of the parent population. If the underlying distribution happens to be normal, there will only be little, if any, power loss by using the Random Normal Scores Test. However, if the distribution is non-normal, there is the potential for unlimited power advantages.

To perform the Random Normal Scores Test, the original data are pooled together from their respective cells and ranked from lowest to highest. The ranks are in turn replaced by randomly selected deviates from a normal distribution. This is accomplished by replacing the lowest rank with the

lowest random normal deviate, the second lowest rank is replaced with the second lowest deviate, and so forth. The resulting values are returned to their original cells. Then, the usual parametric ANOVA Test, which is based on the F distribution, is applied on the random normal deviates.

Since the deviates are drawn from a normal distribution, the subsequent ANOVA Test's assumption of population normality is more completely satisfied than by the substitution of uniform ranks. The Random Normal Scores Test, as with the Rank Transformation Test, tests the hypothesis of identical populations, and in doing so, it is sensitive to location parameters. Therefore, it may be compared to the hypothesis of difference in means tested by the ANOVA Test.

A refinement of the random normal scores approach was suggested by Bradley (1968) and Blair (1980). The Random Normal Scores Test employs the usual parametric ANOVA Test on the deviates. In the formula for the F ratio, the denominator, the Mean Square Within (MSW), estimates the population variance. Suppose the deviates were purposefully drawn from a random normal distribution with a mean of zero and standard deviation of 1.0. In that case, the MSW is known, and the denominator may be directly replaced with the known population variance of 1.0. The resulting ratio reduces to the numerator. In doing so, the F distribution is transformed into a Chi-square distribution divided by the degrees of freedom (Puri, 1964; Hajik and Sidak, 1967). In the 2x2x2 layout, the degrees of freedom for each main effect and interaction is 1.0, yielding the Chi-square distribution.

There are certain implications of using the Chi-square instead of the F distribution. For example, consider the two sample independent means t-test. The denominator of the t ratio is also an estimate of the population variance. If Sigma (the population variance) is known, the t-test becomes a z-test, which is clearly a more powerful procedure. Since Sigma is not an estimation of the population variance, but rather, the actual population variance, the z-test is based on more precise information. The

implication of using random normal scores, under the Chi-square distribution, is that by substituting the known population variance of 1.0 in the denominator and obtaining critical values from the Chi-square table, a more powerful test will result.

The purpose of this study is to compare the robustness and relative power properties of the balanced 2x2x2 ANOVA Test to the Random Normal Scores Test under the F and the Chi-Square distributions, for a variety of distributions, sample sizes, and alpha levels. For the purposes of robustness, Bradley's (1978) liberal definition will be used. With nominal alpha at .05 and .01, the acceptable range of rejection is .025 - .075 and .005 - .015, respectively. The standard to which the power properties will be compared is the level established by the ANOVA Test.

Methodology

A descriptive exploratory design was used in this study. The major tool used to both generate the data and describe its characteristics is the Monte Carlo.

Data Generating Procedure

To investigate the robustness characteristics, random variates were generated and F ratios calculated for the main effects and interactions in the balanced 2x2x2 layout. This process was done for 5000 repetitions for each treatment condition studied. With nominal alpha at .05, for example, there should have been approximately 4750 non-significant F ratios for each main effect and interaction. Then, certain effects were made non-null. A shift in means was introduced by adding a constant to the appropriate observations. To complete the robustness aspect of the study, the main effects and interactions which did not receive a treatment were checked to ensure that they remained null.

Next, the relative power properties were examined. When a constant is added to the observations of the appropriate cells, a true difference in population means occurs. Depending on the size of this constant,

the resultant difference in means has an increased probability of being statistically significant. To simulate an A main effect, a constant was added to the observations of all cells labeled with an A1, e.g., A1B1C1, A1B1C2, A1B2C1, and A1B2C2. Similarly, a B and C main effect was achieved by adding constants to the observations of all cells labeled with a B1 and a C1, respectively.

Ordinal interactions were generated by adding constants to the observations of cells labeled A1B1, A1C1, B1C1, and A1B1C1, for the AxB, AxC, BxC, and AxBxC interactions, respectively. Disordinal interactions were generated by adding the constants in the following manner: for the AxB interaction, the constant was added to the observations of the A1B1 and A2B2 cells; and that same constant was subtracted from the A1B2 and A2B1 cells. That is, the constant was added to the observations of the A1B1C1, A1B1C2, A2B2C1, and A2B2C2 cells; and subtracted from the observations of the A1B2C1, A1B2C2, A2B1C1, and A2B1C2 cells. The remaining lower order disordinal interactions were generated in the same fashion. The ordinal higher order interaction was generated by adding a constant to the observations of the A1B1C1 cell. The disordinal higher order interaction was generated by adding a constant to the observations of the A1B1C1, A1B2C2, A2B1C1, and A2B2C2 cells; and subtracting that same constant from the A1B1C2, A1B2C1, A2B1C2, and A2B2C1 cells.

The properties of the parametric ANOVA Test for small, medium, and large treatment combinations were set as the standards to which the alternative tests were compared. A small treatment was considered as the size of the constant added that would bring the power level of the ANOVA Test to approximately .25. A medium and large treatment would result in power levels of approximately .5 and .75, respectively.

Two sample sizes were investigated. The smaller sample size was n = 2 observations per cell, and the larger sample size was n = 20 observations per cell. For the smaller sample size, forty-two treatment combinations of main effects and interactions were investigated, beginning with no treatment. For the study of main effects, the A, B, and C main effects

were set to low; the A main effect was set to high and B and C remained low; the A and B main effects were set to high and the C remained low; and all three main effects were set to high. For the study of the interactions, these combinations were repeated, adding a medium interaction to each treatment. First, the AxB interaction was made non-null, followed by the AxC, BxC, and AxBxC. This process was done for both ordinal and disordinal interactions.

Twenty-six treatment combinations were studied for the larger sample size. Due to the cost and efficiency of computer time, only ordinal interactions were generated for the larger sample size. The treatment situations are representative of plausible effects of a treatment in practice. As well, they represent a thorough and systematic sample of the theoretical permutations of conditions that may occur.

To maximize the generalizability of the study for various populations, a variety of distributions were investigated. The distributions were the normal, uniform, t with three degrees of freedom, exponential, and a mixed normal. The first three distributions are symmetrical. The normal and uniform are light-tailed, and the t with three degrees of freedom is heavy-tailed. The exponential is highly skewed. The mixed normal is a combination of two populations, which occurs frequently in education and related disciplines (Bradley, 1977; Blair and Higgins, 1980; Blair, 1981).

The computer used in this study was the IBM System/370 with the 3081/D24 Processor Complex. The source code was written in VS Fortran IV, Level 77, Release 1.2. To generate the various distributions, the IMSL GGNML, GGEXN, and GGUBS subroutines, Release 9.2, were accessed.

Results

Robustness

The results of the Monte Carlo have been charted in 390 tables. Due to the large number of tables, only a representative sampling are presented here. Tables 1 and 2 contain the rate of rejections for the tests under the normal distribution, for samples of size n =

Table 1

Normal Distribution; n = 2; No Treatment

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	.052	.055	.053	.054	.053	.054	.046
	.01	.011	.012	.010	.010	.012	.011	.011
RNS_F	.05	.049	.055	.052	.050	.053	.056	.046
	.01	.008	.010	.012	.011	.010	.010	.010
RNS_x2	.05	.039	.036	.040	.042	.043	.043	.043
	.01	.007	.008	.007	.006	.007	.008	.008

Table 2

Normal Distribution; n = 20; No Treatment

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	.044	.050	.042	.045	.055	.050	.048
	.01	.008	.011	.009	.007	.011	.011	.010
RNS_F	.05	.046	.049	.050	.046	.056	.047	.049
	.01	.009	.011	.010	.008	.012	.010	.012
RNS_x2	.05	.047	.050	.050	.046	.057	.050	.048
	.01	.009	.011	.008	.008	.012	.009	.010

Table 3

Normal Distribution; n = 20; High A, B, C Main Effects, Medium Ordinal AxB, AxC, BxC, AxBxC Interactions

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	1.00	1.00	1.00	.967	.969	.971	.479
	.01	1.00	1.00	1.00	.900	.903	.901	.249
RNS_F	.05	1.00	1.00	1.00	.643	.643	.654	.067
	.01	1.00	1.00	1.00	.400	.403	.398	.017
RNS_x2	.05	1.00	1.00	1.00	.311	.323	.312	.009
	.01	.999	1.00	.999	.089	.089	.090	.000

Table 4

Uniform Distribution; n = 20; High A, B, C Main Effects, Medium Ordinal AxB, AxC, BxC, AxBxC Interactions

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	1.00	1.00	1.00	.976	.972	.975	.508
	.01	1.00	1.00	1.00	.908	.907	.912	.278
RNS_F	.05	1.00	1.00	1.00	.655	.647	.641	.119
	.01	1.00	1.00	1.00	.407	.398	.383	.031
RNS_x2	.05	1.00	1.00	1.00	.314	.312	.304	.018
	.01	1.00	1.00	1.00	.081	.086	.086	.001

Table 5

t with 3 df Distribution; n = 20, High A, B, C Main Effects, Medium Ordinal AxB, AxC, BxC, AxBxC Interactions

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	.998	.998	.998	.947	.942	.948	.488
	.01	.997	.997	.997	.866	.873	.873	.272
RNS_F	.05	1.00	1.00	1.00	.545	.546	.548	.044
	.01	1.00	1.00	1.00	.309	.310	.310	.008
RNS_x2	.05	1.00	1.00	1.00	.202	.207	.211	.003
	.01	1.00	1.00	1.00	.043	.042	.040	.000

Table 6

Exponential Distribution; n = 20; High A, B, C Main Effects, Medium Ordinal AxB, AxC, BxC, AxBxC Interactions

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	1.00	1.00	1.00	.968	.966	.971	.508
	.01	1.00	1.00	1.00	.896	.894	.897	.280
RNS_F	.05	1.00	1.00	1.00	.345	.332	.334	.062
	.01	1.00	1.00	1.00	.146	.144	.143	.015
RNS_x2	.05	1.00	1.00	1.00	.089	.085	.083	.058
	.01	1.00	1.00	1.00	.013	.012	.011	.000

Table 7

Mixed Normal Distributions; n = 20; High A, B, C Main Effect, Medium Ordinal AxB, AxC, BxC, AxBxC Interaction

Test	a	Effect A	B	C	AxB	AxC	BxC	AxBxC
AN	.05	1.00	1.00	1.00	.944	.941	.941	.506
	.01	1.00	1.00	1.00	.854	.850	.851	.303
RNS_F	.05	1.00	1.00	1.00	.062	.061	.059	.052
	.05	1.00	1.00	1.00	.014	.013	.012	.012
RNS_{χ^2}	.05	1.00	1.00	1.00	.001	.001	.001	.001
	.01	1.00	1.00	1.00	.000	.000	.000	.000

2 and n = 20, respectively, with no treatment added. Tables 3 - 7 contain the rates of rejection for the tests under the five distributions, with samples of size n = 20, and high A, B, and C main effects, the AxB, AxC, and BxC lower order ordinal interactions, and the AxBxC higher order ordinal interactions present.

With no treatment present, all three tests had Type I error rates well within the range of Bradley's (1978) liberal definition of robustness.

For all distributions, alpha levels, sample sizes, and treatment combinations, the ANOVA Test preserved nominal alpha for null effects.

The Random Normal Scores Test under the F distribution had the tendency to become liberal. In the presence of multiple interactions, the rejection rate for null effects was often above the upper limit of the robustness range. This test had the most difficulty in preserving nominal alpha for the higher order interaction. These results occurred across the distributions and alpha levels. However, when the sample size was increased to n = 20, the test came much closer to the robustness range, although it remained liberal in some situations.

The Random Normal Scores Test under the Chi-square distribution had the tendency to become ultra-conservative. Regardless of the distribution or alpha level, in the presence of multiple main effects and interactions, the rejection rate for null effects approached .000. When n was increased to 20 observations per cell, the tendency to become ultra-conservative was reduced considerably. Also, the test fared much better with disordinal interactions than it did with ordinal interactions. However, the test was often unable to reject null effects above the lower limit of the robustness range.

Power

The Random Normal Scores Test under the F distribution was generally only slightly less powerful than the ANOVA Test in detecting main effects for the first three distributions. The maximum advantage of the ANOVA Test was .1. The Random Normal Scores Test

was much more powerful under the exponential and mixed normal distributions, with a maximum power advantage of .75. The increase in sample size slightly increased the advantage. However, the alpha levels did not appear to have much influence on the results.

Under the first three distributions, with interactions present, this test was less powerful than the ANOVA Test. The maximum advantages of the ANOVA Test ranged from .1 to .3 for the normal, uniform, and t distributions, increasing with the number of effects that were made non-null.

For the exponential and mixed normal distributions, neither test was consistently the more powerful with lower order, disordinal interactions. However, with ordinal, lower order interactions and the higher order interaction, the ANOVA Test was always more powerful. Once again, the increase in the size of the sample made the Random Normal Scores Test under the F distribution more competitive. However, the different alpha levels had little effect on the results.

The Random Normal Scores Test fared much worse under the Chi-square distribution. At the smaller sample size, for the normal, uniform, and t with 3 df distribution, the test was always less powerful than the ANOVA Test. Specifically, with main effects present, the test was usually slightly less powerful. However, with interactions present, the ANOVA Test had maximum advantages as high as .9, .5, and .5 for the main effects, lower order, and higher order interactions, respectively.

For the exponential and mixed normal distribution, with interactions present, the same pattern emerged. Although the test was much less powerful with interactions present, with one main effect present, the Random Normal Scores Test was often more powerful than the ANOVA Test, with a maximum power advantage of .75.

The test fared slightly better at the larger sample size, and for disordinal interactions. The alpha level did not appear to have much of an effect on these results.

Conclusions

The Random Normal Scores Test has the tendency to become liberal or conservative under the F and Chi-square distributions, respectively. That is, with no treatment present, or in the presence of a low main effect, the rate of rejection for null effects was usually robust. However, as the main effect increased from low to high, and from one to three main effects present, the tests began to lose their robustness. With the introduction of interactions, the maximum inflation or deflation occurred.

The nonparametric tests were less robust in the presence of ordinal interactions than they were in the presence of disordinal interactions. This may underscore a difficulty with Iman (1974) and Iman and Conover's (1976) studies, as they limited their investigation to disordinal interactions.

As the sample size increased, the alternative tests improved. Nevertheless, in many cases, they failed to remain within Bradley's (1978) liberal definition of robustness.

Since these tests failed to demonstrate robust characteristics, the question of their relative power compared to the ANOVA Test becomes problematic. However, the results do tend to substantiate some of the theoretical underpinnings based on the ARE. For example, in detecting main effects, the alternative procedures were often more powerful.

Under the normal, uniform, and t distributions, the power loss for main effects was usually slight. Under the exponential and mixed normal distributions, the Random Normal Scores Tests were much more powerful than the ANOVA Test in detecting main effects. This suggests the possibility of using these procedures in the t-test or one-way ANOVA layout. However, for interactions, and even more so, for the higher order interaction, the advantages were quickly lost. It is possible that the tests simply lack the sensitivity to detect interactions.

The transforming of the Random Normal Scores Test under the F distribution to the Chi-square distribution, as suggested by asymptotic theory, worked in reverse. It made the procedure ultra-

conservative and much less powerful.

In conclusion, the use of the Random Normal Scores Test in the 2x2x2 layout, for the distributions and sample sizes studied, is generally unwarrented. These procedures are neither robust nor powerful in comparison to the ANOVA Test. However, it must be recalled that the ANOVA Test, although robust and superior in terms of its relative power compared to these procedures, is nevertheless affected by the violation of the assumption of population normality. A suitable alternative is still required.

Bibliography

Bell, C. B. & Doksum, K. A. (1965). Some new distribution-free statistics. The Annals of Mathematical Statistics, 36, (1), 203-214.

Blair, R. C. (1980). A comparison of the power of the two independent means t test to that of the Wilcoxon's Rank-Sum Test for samples of various populations. (Unpublished doctoral dissertation, University of South Florida, Tampa, FL.)

Blair, R. C. (1981). A reaction to 'Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance.' Review of Educational Literature, 51, (4), 499-507.

Blair, R. C. & Higgins, J. J. (1980). 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, 5, 309-335.

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, May, 41, (8), 4-17.

Bradley, J. V. (1968). Distribution-free statistical tests. Englewood-Cliffs, NJ: Prentice-Hall.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. The American Statistitian, 31, 147-150.

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

Fisher, R. A. & Yates, F. (1949). Statistical tables for biological, agricultural, and medical research. (3rd ed.). NY: Hafner Publishing Co.

Hajek, J. & Sidak, Z. (1967). Theory of rank tests. NY: Academic Press.

Iman, R. L. (1974). A power study of a rank tranform for the two-way classification model when interactions may be present. The Canadian Journal of Statistics, Section C: Applications, 2, 227-239.

Iman, R. L. & Conover, W. J. (1976). A comparison of several rank tests for the two-way layout. Alburquerque, NM: Sandia Laboratories, SAND76-0631, Dec.

Iman, R. L., Hora, S. C., & Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. Journal of the American Statistical Association, 79, 674-685.

Puri, M. L. (1964). Asymptotic efficiency of a class of c-sample tests. The Annals of Mathematical Statistics, 35, 102-121.

Sawilowsky, S. N. (1985). Robust and power analysis of the 2x2x2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. (Unpublished doctoral dissertation, University of South Florida, Tampa, FL.)

Walberg, H. J., Strykowski, B. F., Rovai, E., & Hurg, S. S. (1984). Exceptional performance. Review of Educational Literature, 54, 87-112.

AUTHOR

SHLOMO SAWILOWSKY, Visting Assistant Professor, Institute for Instructional Research and Practice (Content Area Teacher Test), University of South Florida, Tampa, Florida 33620