DOCUMENT RESUME

ED 091 417 TM 003 631

AUTHOR Killian, C. Rodney; Hoover, H. D.
TITLE An Investigation of Selected Two-Sample Hypothesis Testing Procedures When Sampling From Empirically Based Test Score Models.
PUB DATE [Apr 74]
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)
AVAILABLE FROM C. Rodney Killian, Drake University, 26th Street and University Avenue, Des Moines, Iowa 50311 (FREE)

EDRS PRICE MF-$0.75 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS *Hypothesis Testing; Investigations; *Models; *Nonparametric Statistics; Population Distribution; Sampling; *Statistical Analysis
IDENTIFIERS *Mann Whitney U Test; Student T Statistics

ABSTRACT

The power of the t, expected normal scores, Mann-Whitney U, Tukey, a modified Mann-Whitney U, and an adaptive procedure were investigated when sampling from population models empirically developed from test score distributions. The models used were selected members of the beta family. This investigation was unique in that not only did the means of the alternative distributions increase under change in location parameter, but the shape of the distribution changed as well. In general, the t statistic displayed superior power over the other procedures. Closely behind t were the expected normal scores and Mann-Whitney U procedures with the others following. (Author)

Session 21.18

AN INVESTIGATION OF SELECTED TWO-SAMPLE HYPOTHESIS TESTING PROCEDURES WHEN SAMPLING FROM EMPIRICALLY BASED TEST SCORE MODELS

C. Rodney Killian
Drake University

and

H. D. Hoover
University of Iowa

A paper presented at the American Educational Research Association convention in Chicago, April 16-19, 1974.

INTRODUCTION

Frequently, educational and psychological researchers are confronted with the problem of determining whether two independent samples come from the same population. The major purpose of this study was to investigate the small sample power and the control of Type I error of selected two-sample statistical procedures when sampling from populations encountered in educational and psychological research. For such research, test scores are often used as the criterion measure reflecting the performance of an individual belonging to a target population or an experimentally accessible population.

Numerous statistical procedures have been proposed to detect differences in central tendency between two populations when independent, random samples are drawn from each. The two-sample statistical procedures investigated were: the Student's t-test (t), the Mann-Whitney U test (U), the Terry-Hoeffding normal scores test (S), a Tukey quick test (T) as developed by Tukey (1959), a modified Mann-Whitney U test (W) was developed by Randles and Hogg (1972), and an adaptive two-sample nonparametric procedure (A) also developed by Randles and Hogg (1972).

The modified Mann-Whitney U test was a Mann-Whitney U statistic based upon the [(N+1)/4]* largest observations and the [(N+1)/4]*

---

*[p] denotes the greatest integer contained in p.

smallest observations in the combined sample. The adaptive procedure used in this research was a modification of a procedure described by Randles and Hogg (1972). The adaptive scheme classified the underlying distribution as having either light or non-light tails and a modified Mann-Whitney U or Mann-Whitney U statistic was used accordingly. The criterion for classification was the range of the sample divided by the mean deviation from the sample median.

The t-test and its primary nonparametric competitor, the Mann-Whitney U (or the two-sample Wilcoxon), have been extensively researched with regards to power and control of Type I error. Most of this research has concentrated on sampling from populations whose forms are similar to well known theoretical distributions such as normal, uniform, exponential, logistic, double exponential, etc., and has considered that the two populations differ in location parameter and/or scale parameter. However, in educational and psychological research two crucial questions to raise are:

1. How often are the underlying population distributions really normal, uniform, exponential, logistic, double exponential, etc.
2. How well do the various two-sample statistical procedures detect differences between two populations when sampling from population distribution types that exist in the field of educational and psychological research?

CONSTRUCTION OF MODELS

In this investigation the following three considerations were taken into account in the construction of population distribution models.

1. The population models should exemplify test score distributions that frequently occur in educational and psychological research.
2. The population models should be bounded because test score distributions are bounded at both the upper and lower ends.
3. Since test score distributions are bounded, any change in central tendency from one distribution to another is likely to change other characteristics as well, such as skewness and kurtosis.

Descriptive statistics for raw score distributions of the Iowa Tests of Basic Skills and the Iowa Tests of Educational Development based upon National and Iowa norms were provided by Brandenburg (1972). While these data indicate that distributions of raw scores from standardization populations tend to be nearly symmetrical and very light-tailed, most educational and psychological research is not conducted by a sampling from such populations. Instead, a more common practice is to use students from an intact classroom, building, or school system, and randomly divide them into experimental and control groups. Data from the files of the Iowa Testing Programs including 122 classes ($N \leq 30$), 41 buildings ($30 < N \leq 90$) and 16 systems

re the basis for the construction of the models. A few in-
ng generalizations drawn from the data were:

1. The sample means and standard deviations were curvilinearly related.
2. The sample means and measures of skewness were linearly related.
3. The sample means and measures of kurtosis were curvilinearly related.
4. The measures of skewness and kurtosis were curvilinearly related.
5. The above relationships held for classes, buildings, and systems.
6. The majority of test score distributions were light-tailed and positively skewed.

Five beta distribution models were selected to represent these population distributions. These beta models, along with summary descriptive statistics including skewness ($\sqrt{\beta_1}$) and kurtosis (β_2), are illustrated in Figure 1.

PROCEDURE

To generate samples from the five beta distributions rapidly and efficiently an algorithm based on the inverse of the generalized lamda distribution, as developed by Schmeiser (1971), was used. Samples of size (5,5), (5,10), (10,5), (10,10), and (20,20) were investigated. For each combination of sample sizes, two empirical power functions

were obtained. The first using a beta distribution with $\mu = .275$ (distribution I of Figure 1) as the X-distribution sampling model with successive Y-distribution sampling models being those of distributions I through V of Figure 1. The second using a beta distribution with $\mu = .3875$ (distribution II of Figure 1) as the X-distribution sampling model with successive Y-distribution sampling models of II through V of Figure 1.

RESULTS

In the evaluation of any hypothesis testing procedure, the control over Type I error must be given careful consideration. The empirical Type I error rates are given in columns one and six of Table I. For the exact tests (all except t), the majority of empirical significance values were within expected binomial variation levels ($\sigma_p = .007$ for $p = .05$ and $N = 1000$). Only the empirical values obtained at the .05 level of significance are shown in the table. Similar results were obtained at the other significance levels of .01 and .10.

Even though the t-test is not an exact procedure for these distributions, close agreement between empirical and nominal levels was observed. An exception was its being somewhat erratic in the instances where the sample sizes were unequal. Similar results for the t statistic have been reported by other investigators and along with those given here confirm the apparent Type I error robustness of the t statistic for such non-normal population models.

Overall, the t-test exhibited the greatest empirical power in the situations investigated.[1] For small, equal samples of size 5, the

---

[1]Empirical results for the modified Mann-Whitney U and the adaptive procedure were not obtained for sample size m=n=5 as tabled critical values were not available for this case.

t-statistic was slightly superior to the other statistics investigated with its superiority being more pronounced at the .01 level of significance. In the situations where the sample sizes were unequal, the t-test, with one exception, was always the most powerful. The exception occurred when the smaller sized sample came from the X-distribution and the change in location parameter was small. Here the Mann-Whitney U and the normal scores procedure were slightly more powerful than the t.

The normal scores and Mann-Whitney U tests were very competitive with the t. In fact, with a small change in location parameter and equal sample sizes of 10 and 20, the Mann-Whitney U and the normal scores procedures displayed greater empirical power than did the t. Little difference was observed in the performance of the normal scores test and the Mann-Whitney U test. Close agreement in empirical power values of the normal scores and Mann-Whitney U was also observed in other investigations (Gibbons, 1964; van der Laan and Oosterhoff, 1967; and Neave and Granger, 1968).

Based upon the results of this investigation the t, normal scores, and Mann-Whitney U statistics would be recommended for use in detecting a difference between two population means when sampling from score distributions similar to the models used. Moreover, the two rank tests, namely normal scores and Mann-Whitney U, involve simpler arithmetic and may be preferred over the t if tables of their critical values are readily accessible.

As a quick procedure to test equality of two population means, Tukey's test performed well for the distributions sampled. This procedure compared favorably to the others when sample sizes were small as well as for small changes in location parameter.

The adaptive procedure usually had an empirical power value between those values obtained for the modified Mann-Whitney U and Mann-Whitney U and closer to the modified Mann-Whitney U. This outcome was accredited to the criterion value of 2.25 (suggested in a personal communication with Randles and Hogg) used to choose between the Mann-Whitney U and the modified Mann-Whitney U. Since the adaptive procedure consistently resulted in empirical power values below those of t, S, or U; it would not be recommended for use in detecting differences in two population means when sampling from distributions similar to the models of this study.

The modified Mann-Whitney U did not exhibit high empirical power values and thus we do not recommend its use when sampling from distributions similar to our models. Randles and Hogg (1972) have shown that this statistic has high relative power values under a shift alternative when sampling from uniform distributions. Even though the uniform distribution is light-tailed, it is quite different from the models of this study.

Very little difference in the obtained power functions was observed when the null distribution was markedly skewed (0.75) as compared to when this distribution was more moderately skewed (0.34) the noted difference being that S and U had more similar empirical power functions when the null distribution was more moderately skewed.

BIBLIOGRAPHY

Brandenburg, D. C. "The Use of Multiple Matrix Sampling in Approximating an Entire Empirical Norms Distribution." Unpublished Ph.D. Dissertation, University of Iowa, 1972.

Gibbons, J. D. "On the Power of Two-Sample Rank Tests on the Equality of Two Distribution Functions." Journal of the Royal Statistical Society, B, 26(1964), pp. 293-304.

Neave, H. R. and Granger, C. W. J. "A Monte Carlo Study Comparing Various Two-Sample Tests for Differences in Mean." Technometrics, 10(1968), pp. 509-522.

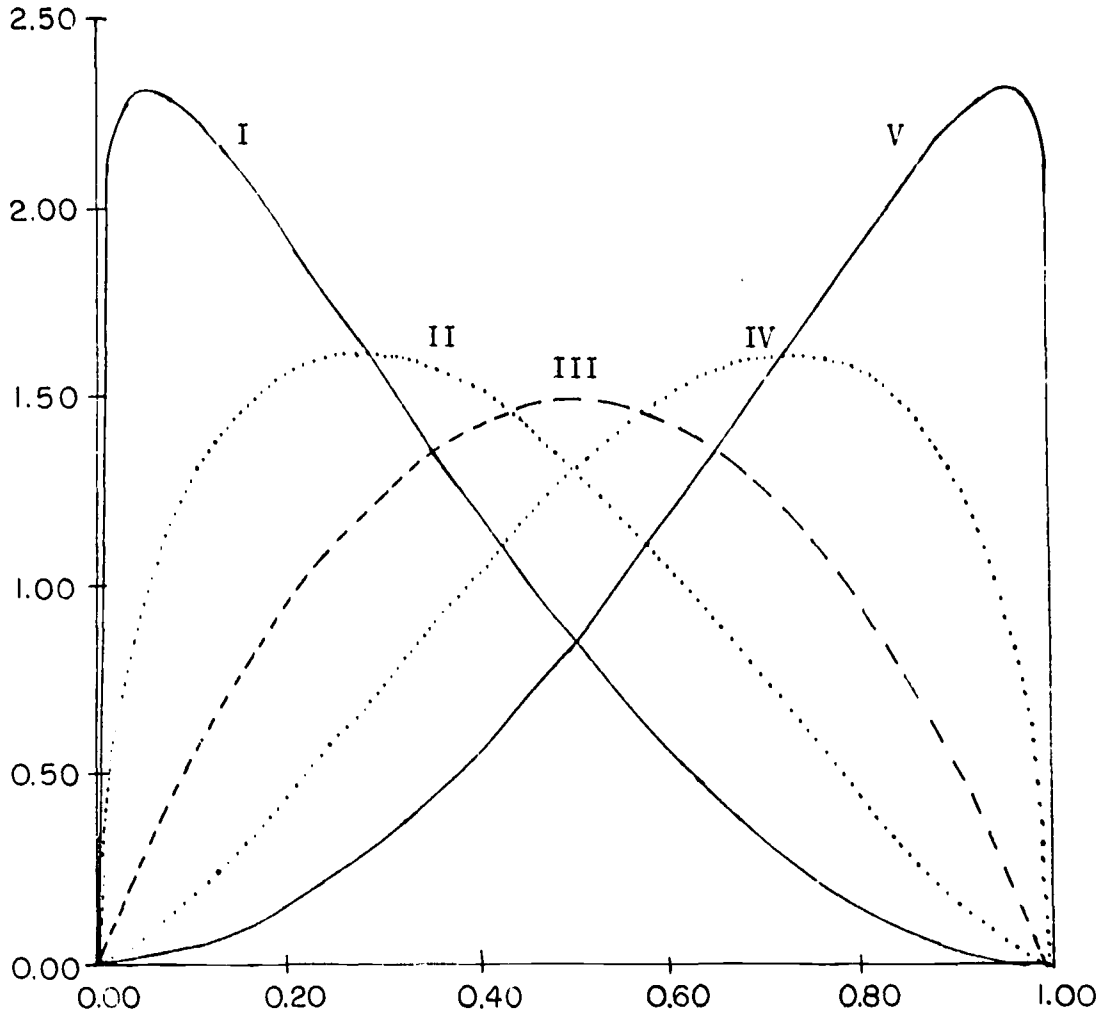
Randles, R. H. and Hogg, R. V. "Adaptive Distribution-Free Tests." Technical Report No. 17, University of Iowa, 1972.

Schmeiser, B. W. "A General Algorithm for Generating Random Variables." Unpublished Master's thesis, University of Iowa, 1971.

Tukey, J. W. "A Quick, Compact, Two-Sample Test to Duckworth's Specifications." Technometrics, 1(1959), pp. 31-38.

van der Laan, P. and Oosterhoff, J. "Experimental Determination of the Power Functions of the Two-Sample Rank Tests of Wilcoxen, van der Waerden and Terry by Monte Carlo Techniques --- I. Normal Parent Distributions." Statistica Neerlandica, 21(1967), pp. 55-68.

Figure 1
Distributions Used as Population Models



Distributions	μ	σ^2	$\sqrt{\beta_1}$	β_2
I	.275	.0399	.75	2.87
II	.3875	.0475	.34	2.30
III	.500	.0500	.00	2.14
IV	.6125	.0475	-.34	2.30
V	.7250	.0399	-.75	2.87

TABLE 1

Empirical Power Functions at $\alpha = .05$ for Each of Six Two-Sample Test Statistics (C) When Sampling from Selected Beta Distributions of the Family a + b = 4.0

		$\mu_X = .2750$					$\mu_X = .3875$			
N_X, N_Y	C \ μ_Y	.2750	.3875	.5000	.6125	.7250	.3875	.5000	.6125	.7250
5,5	t	.055	.206	.499	.727	.904	.044	.205	.433	.731
	S	.055	.201	.481	.694	.880	.043	.196	.412	.694
	U	.051	.199	.476	.694	.884	.045	.190	.410	.698
	T	.057	.196	.455	.664	.871	.045	.189	.401	.678
10,10	t	.047	.313	.748	.957	.999	.050	.306	.724	.954
	S	.043	.333	.737	.952	.998	.048	.300	.703	.951
	U	.046	.319	.739	.946	.999	.057	.297	.694	.948
	T	.055	.251	.621	.855	.958	.043	.249	.575	.861
	W	.053	.294	.677	.905	.986	.048	.281	.633	.911
	A	.041	.294	.673	.900	.987	.053	.271	.646	.913
20,20	t	.049	.533	.945	.999	1.000	.048	.493	.927	.999
	S	.055	.548	.949	.999	1.000	.054	.477	.923	1.000
	U	.053	.546	.944	.999	1.000	.050	.487	.920	.999
	T	.047	.343	.715	.914	.976	.054	.265	.674	.915
	W	.059	.510	.922	.997	1.000	.047	.435	.886	1.000
	A	.057	.518	.927	.997	1.000	.049	.442	.895	.999
5,10	t	.043	.226	.546	.864	.971	.048	.224	.563	.840
	S	.053	.245	.539	.839	.964	.052	.224	.542	.833
	U	.065	.237	.542	.833	.957	.054	.225	.534	.827
	T	.050	.232	.497	.774	.911	.057	.189	.476	.761
	W	.056	.257	.508	.774	.917	.052	.193	.496	.769
	A	.060	.244	.514	.775	.914	.055	.199	.492	.772
10,5	t	.051	.240	.579	.864	.964	.055	.263	.532	.867
	S	.049	.229	.560	.835	.946	.049	.249	.515	.863
	U	.052	.225	.576	.837	.951	.047	.247	.519	.852
	T	.048	.210	.510	.761	.897	.054	.233	.441	.788
	W	.047	.206	.510	.766	.899	.051	.224	.461	.789
	A	.050	.205	.513	.759	.895	.054	.224	.473	.792