Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 5

5-1-2002

Power Analyses When Comparing Trimmed Means

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

H. J. Keselman
University of Manitoba, kesel@ms.umanitoba.ca

Cover Page Footnote
The research reported in this article was supported, in part, by a grant from the Natural Sciences and Engineering Research Council of Canada.

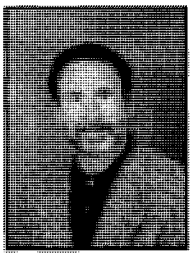
Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 24-31

1538 - 9472/02/$30.00

Power Analyses When Comparing Trimmed Means

Rand R. Wilcox
Department of Psychology
University of Southern California

H. J. Keselman
Department of Psychology
University of Manitoba



Given a random sample from each of two independent groups, this article takes up the problem of estimating power, as well as a power curve, when comparing 20% trimmed means with a percentile bootstrap method. Many methods were considered, but only one was found to be satisfactory in terms of obtaining both a point estimate of power as well as a (one-sided) confidence interval. The method is illustrated with data from a reading study where theory suggests two groups should differ but nonsignificant results were obtained.

Keywords: Bootstrap, Robust methods

Introduction

Power is a fundamental concern when comparing measures of location corresponding to two independent groups. Of course, when we fail to detect a difference, this might be because there is little or no difference between the measures of location, or perhaps the sample size was inadequate for detecting a difference that is substantively important. Surely the best-known and most commonly used method when addressing power is to assume both groups have normal distributions with a common variance, specify a (standardized) difference between means, choose $1 - \beta$, the desired probability of rejecting the hypothesis of equal means for this specified difference, and then determine the required sample size to achieve this goal. Cohen (1977) provided an excellent summary of this strategy.

Another but less commonly used method for dealing with power is to use a post hoc analysis. That is, collect data, and based on the observed values estimate power or determine what sample size is required to achieve a desired amount of power. As is evident, power is not an issue if the null hypothesis is rejected, but otherwise it is. A classic illustration of this approach is the two-stage strategy derived by Stein (1945). Extensions of the method to two or more groups have been proposed by various researchers during the ensuing years, a summary of which can be found in Wilcox (1996). Included are exact heterosecedastic methods when sampling from normal distributions. That is, under normality, both the probability of a Type I error and power can be simultaneously controlled. Other approaches to controlling power are reviewed by Hewett and Spurrier (1983).

Our goal in this paper is to consider how power analyses might be made when comparing 20% trimmed means rather than means. But unlike Stein-type procedures, our goal is to obtain both a point estimate of power plus a one-sided confidence interval. That is, if we fail to reject, we want to estimate power, based on the observed data. In particular, we want to estimate the power curve, the probability of rejecting as a function of the difference between the population trimmed means.

Our interest in 20% trimmed means stems from both its theoretical advantages summarized by Staudte and Sheather (1990) and Huber (1981), among others, plus its practical advantages when trying to deal with nonnormality. In particular, methods based on 20% trimmed means provide good control over Type I errors for a broader range of situations versus methods based on means, they maintain relatively high power under arbitrarily small departures from normality that destroy power when using means, and they provide accurate confidence intervals over a much

Rand R. Wilcox is a Professor of Psychology at the University of Southern California. He is a fellow of the Royal Statistical Society and the American Psychological Society. He has published over 170 journal articles, and has recently written his fifth book on statistics. Harvey Keselman is a fellow of the American Psychological Association and the American Psychological Society. He is a Professor of Psychology at The University of Manitoba, Winnipeg, Manitoba, Canada. His areas of interest include the analysis of repeated measurements, multiple comparison procedures, and robust estimation and testing.

broader range of situations versus conventional methods for means. Theory and simulations also indicate that trimmed means do a better job of reducing bias when testing hypotheses. Student's two-sample t, for example, is biased, meaning that the probability of rejecting is not minimized when the null hypothesis is true.

That is, power can decrease as the difference between the means increases. Comparing 20% trimmed means with a percentile bootstrap method virtually eliminates this problem among situations considered in extant publications. Moreover, the percentile bootstrap, used in conjunction with 20% trimmed means, performs remarkably well when the goal is to use a test that is reasonably equal-tailed. For a nontechnical summary of the many problems associated with means in particular and least squares in general, and how modern robust methods address these problems, see Wilcox (2001a). For a recent review of problems associated with conventional methods, see Keselman, Huberty, Lix, Olejnik, Cribbie, Donohue, Kowalchuk, Lowman, Petosky, Keselman, and Levin (1998). Readers interested in the practical details of how to apply robust methods can refer to Wilcox (1997).

To help motivate this paper we begin with the usual planning strategy of determining an appropriate sample size under normality when using Student's t. The left panel of Figure 1 shows two normal curves having means 0 and 1 and a common variance of one. Let

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma},$$

where μ_j and σ_j^2 are the mean and variance associated with the j^{th} group (j=1,2), and by assumption $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Cohen argued that for normal distributions Δ=.8 is a large effect, so from this view surely we want power to be reasonably high for the situation depicted in Figure 1. If we sample twenty-five observations from each group and test for equal means at the .05 level, power is .96. So based on this power analysis under normality, sample sizes of 25 would seem to suffice. But suppose we sample from the two distributions shown in the right panel instead. As is evident, they appear to be very similar to the distributions shown in the left panel, only the curves in the right panel do not extend as far up the y-axis. Now Student's t has power .28 so the sample sizes are inadequate if a difference of $\mu_1 - \mu_2 = 1$ is judged to be important.

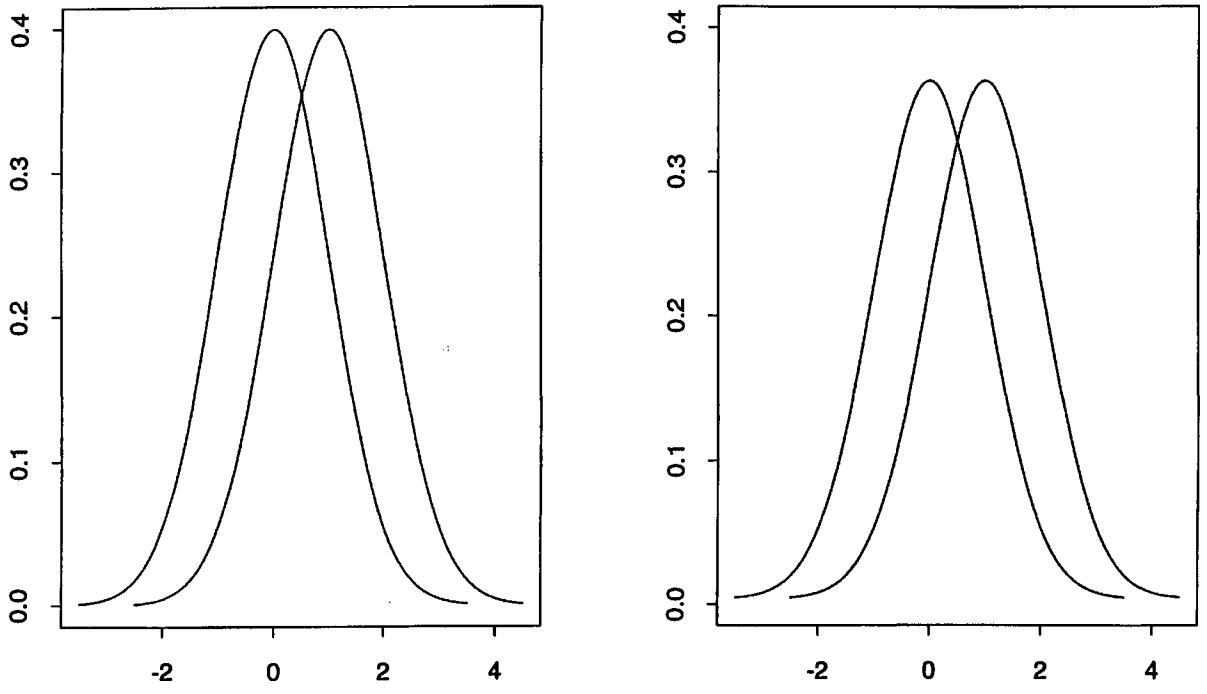


Figure 1

The distributions in the right panel of Figure 1 are from the family of mixed (or contaminated) normals. For the particular mixed normal considered here, sampling is from a standard normal distribution with probability .9, and with probability .1 an observation is sampled from a normal distribution with mean zero and standard deviation ten. The resulting distribution is very similar to a standard normal (in the Kolmogorov sense), but there is a crucial difference. The mixed normals in the right panel of Figure 1 have variances 10.9 and this is why power is relatively low. That is, a very slight departure from normality can result in very low power rendering our choice for sample sizes inadequate when conventional power analyses are made. This result follows immediately from Tukey (1960) as well as from general results on robustness summarized in Huber (1981), Staudte and Sheather (1990) and Wilcox (1997).

When attention is restricted to means, a Stein-type method of power analysis will tend to catch the error just described once observations are available. A member of this class of methods that can be used in the situation at hand was derived by Bishop and Dudewicz (1978). (Related methods derived by Hochberg, 1975, and Tamhane, 1977, can be used to control the length of a confidence interval.) Once data are available, the Bishop-Dudewicz method indicates how many additional observations are required to achieve some specified level of power given a difference between the means that is deemed important. If few or no additional observations are required, this indicates that the original sample size was adequate. Briefly, the required sample size depends on the sample variances. Not surprisingly, the larger the sample variances, the larger the required number of observations in order to achieve high power. Because the sample variances tend to be large when sampling from the mixed normal distribution considered here, versus sampling from a normal, the Bishop-Dudewicz method will tend to detect the fact that the original sample sizes were inadequate for the situation depicted in the right panel of Figure 1.

An obvious concern is that obtaining additional observations can be difficult. What would be nice is a method that achieves high power in both of the situations depicted in Figure 1. Methods based on a 20% trimmed mean accomplish this goal. For the normal distributions, if we apply Yuen's (1974) method for trimmed means, power is .89 (based on a simulation with 10,000 replications), and for the contaminated normals it is .78. That is, relatively little power is lost under normality versus using means, and power is not destroyed under a small departure from normality. This is one of several reasons 20% trimmed means have appeal. But if we fail to reject when comparing 20% trimmed means, again we have the issue of assessing why. That is, an estimate of power becomes important.

A natural strategy for assessing power, when using trimmed means, is to use some analog of the Bishop-Dudewicz method. Theoretical results leading to Yuen's (1974) method (e.g., Staudte and Sheather, 1990; Wilcox, 1997) suggest an obvious analog, but we found that in simulations, control over power was unsatisfactory. Various modifications were tried, but all of them gave unsatisfactory results. There are some rather obvious bootstrap methods for estimating power (e.g., Efron & Tibshirani, 1993). Unfortunately, the estimate can be rather inaccurate with small or even moderately large sample sizes. (Some of the many variations that were considered and found to be unsatisfactory are briefly described below.) Yet, another concern is that Yuen's method can be less satisfactory than two basic bootstrap methods for comparing trimmed means (e.g., Wilcox, 1997).

One of these is the percentile t bootstrap and another is the percentile method. It is known that when comparing means, the percentile t bootstrap outperforms the percentile method (e.g., Westfall & Young, 1993). However, for trimmed means, there is little separating the two methods when comparing two groups. But when there are more than two groups, the percentile method begins to perform better, in terms of probability coverage and Type I errors, than the percentile t (Wilcox, 2001b). Moreover, results in Singh (1998) suggest how the power of both bootstrap methods might be improved. Wilcox (2001b) found that Singh's approach, when applied to the percentile method, gives reasonable control over the probability of a Type I error if the smallest sample size is at least 15. But when using the percentile t, Singh's method performed rather poorly. For these reasons we focus on the percentile bootstrap method.

Our goal, therefore, is to find a reasonable point estimate of power and to assess the accuracy of this estimate by computing a .95 one-sided confidence interval, the idea being that we want a conservative estimate of power. For example, if we estimate power to be .8, and our one-sided confidence interval for the actual amount of power is (.7, 1), then we can be reasonably certain that power is at least .7 and this can be used to judge the sample sizes under consideration. (Of course we could compute a two-sided confidence interval for the estimated power, but the upper end of such a confidence interval seems less interesting than the lower end.)

Methodology

For two independent groups let X_{ij} be a randomly sampled observation for the j^{th} group, (j = 1,2; i = 1, ..., n_j). The corresponding population 20% trimmed means are labeled μ_{t1} and μ_{t2} and the goal is to test

$$H_0 : \mu_{t1} = \mu_{t2}$$

We begin by describing the method for testing H_0 after which we turn to the problem of estimating power.

For the j^{th} group let $X^*_{1j},\ldots,X^*_{n_j j}$ be a bootstrap sample. That is, for each j, the values $X^*_{1j},\ldots,X^*_{n_j j}$ are obtained by randomly sampling, with replacement, n_j values from $X_{1j},\ldots,X_{n_j j}$. Let $p^* = P\left(\bar{X}^*_{t1} > \bar{X}^*_{t2}\right)$, where for the j^{th} group $\bar{X}^*_{tj}$ is the 20% trimmed mean based on the bootstrap sample. (See Wilcox (1997, p. 32) for details on how to compute a trimmed mean.) That is, p^* is the probability that, when resampling from the empirical distribution associated with the first and second groups, a bootstrap sample trimmed mean from group 1 is greater than the bootstrap sample trimmed mean from group 2. Notice that p^* reflects the degree to which the empirical distributions differ. If the empirical distributions are identical, then $p^* = .5$. Moreover, if the null hypothesis of equal trimmed means is true, then results in Hall (1986a), in combination with the influence function of the trimmed mean, imply that p^* should have, approximately, a uniform distribution. The reason is that

$$\bar{X}_t = \mu_t + \frac{1}{n}\sum IF(X_i)$$

plus a remainder term that goes to zero as n gets large, where $IF(X_i)$ is the influence function (e.g., Staudte & Sheather, 1990). That is, the sample trimmed mean can be written as an average of independent identically distributed random variables. That p^* converges to a uniform distribution under the null hypothesis also follows from general results in Hall (1986b). Consequently, the closer p^* is to 0 or 1, the more evidence there is that the null hypothesis should be rejected. Reject $H_0 : \mu_{t1} = \mu_{t2}$ at the α level if $p^* \leq \alpha/2$ or if $p^* \geq 1 - \alpha/2$.

The value of p^* can be estimated in a simple manner. For the j^{th} group, obtain B bootstrap trimmed means $\bar{X}^*_{btj}$, $b=1,\ldots,B$. Let $I_b=1$ if $\bar{X}^*_{bt1} > \bar{X}^*_{bt2}$, otherwise $I_b=0$, where $\bar{X}^*_{btj}$ is the trimmed mean based on the b^{th} bootstrap sample. Then an estimate of p^* is

$$\hat{p}^* = \sum_{b=1}^{B} I_b/B$$

The hypothesis of equal trimmed means is rejected if $\hat{p}^* \leq \alpha/2$ or $\hat{p}^* \leq 1-\alpha/2$. It is readily verified that the hypothesis testing procedure just described is the percentile bootstrap method. Furthermore, the view of the percentile bootstrap just given provides a useful way of addressing power.

For convenience, let $\hat{p}^*_m = \min\left(\hat{p}^*,\ 1-\hat{p}^*\right)$ in which case H_0 is rejected if $\hat{p}^*_m \leq \alpha/2$, where as usual α is the desired probability of a Type I error. Then power is $1-\beta = P\left(\hat{p}^*_m \leq \alpha/2\right)$ when H_0 is false. One of our failed attempts at estimating power was to approximate the bootstrap sampling distribution of the bootstrap trimmed means with a normal distribution. The mean and variance of this distribution are easily estimated using well-known properties of the trimmed mean. But the resulting estimate of power was found to be unsatisfactory. Next we tried a Cornish-Fisher approximation of the bootstrap sampling distributions using results in Wilcox (1994). Again estimated power was unsatisfactory.

A method that was partially successful was a nested bootstrap estimate of power. This approach provided a reasonably unbiased estimate of the true power level under a shift model, but the standard error of the estimate was such that the assessed power might be inaccurate to the point of being misleading. That is, in many situations the actual amount of power is over-estimated, giving a false sense that the sample sizes used are adequate. What is needed is some way of computing confidence intervals for the actual power level, but a reasonable method for accomplishing this goal, when using the nested bootstrap, has not been found.

Now we describe the one method we have found so far that gives good results, in simulations, under a shift model. (Handling situations where distributions have unequal variances is discussed below.) As is evident, power is related to the standard errors of the trimmed means. Roughly, the strategy is to devise a function for estimating power when distributions are normal, where the estimate is based on some specified difference between the population trimmed means, say $\delta = \mu_{t1} - \mu_{t2}$, and the standard error of $\bar{X}_{t1} - \bar{X}_{t2}$. Then, given data, an estimate of power is obtained simply by estimating the standard error and plugging it into the function just described. That is, we estimate $1-\beta = P\left(\hat{p}^*_m \leq \alpha/2\right)$. To get a one-sided .95 confidence interval for the actual power level, we compute a one-sided .95 confidence interval using a percentile bootstrap in conjunction with our power function. To elaborate, temporarily consider a single random sample, $X_1,\ldots, X_n$. Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the order statistics and let

$$W_i = \begin{cases} X_{(g+1)}, & \text{if } X_i \leq X_{(g+1)} \\ X_i, & \text{if } X_{(g+1)} < X_i < X_{(n-g)} \\ X_{(n-g)}, & \text{if } X_i \geq X_{(n-g)}. \end{cases}$$

Set $g = [.2n]$, where $[.2n]$ is the greatest integer $\leq .2n$. The (20%) Winsorized sample mean is

$$\bar{X}_w = \frac{1}{n}\sum W_i$$

and the sample Winsorized sum of the squared deviations is

$$SSD_w = (g+1)(X_{(g+1)} - \bar{X}_w)^2 + (X_{(g+2)} - \bar{X}_w)^2 + \ldots + (X_{(n-g-1)} - \bar{X}_w)^2 + (g+1)(X_{(n-g)} - \bar{X}_w)^2. \quad (1)$$

The (20%) sample Winsorized variance is

$$s_w^2 = \frac{1}{n-1}\sum (W_i - \bar{X}_w)^2.$$

Letting σ_ω^2 be the population Winsorized variance, theoretical results, based on the influence function of the 20% trimmed mean, indicate that the squared standard error of the sample trimmed mean is

$$\frac{\sigma_w^2}{.36n}$$

(e.g., Staudte & Sheather, 1990). Following Yuen (1974), we estimate this squared standard error with

$$d = \frac{SSD_w}{h(h-1)},$$

where $h = n - 2g$ is the "effective" sample size. Of course one could use instead $s_w^2/(.36n)$, but in small sample sizes d has been found to perform better when testing hypotheses (Wilcox, 1997). There is no indication that the alternate estimate of the standard error provides added value for the problem at hand, so d is used henceforth. Returning to the two-sample case, let

$$\tau^2 = \frac{\sigma_{w1}^2}{.36n_1} + \frac{\sigma_{w2}^2}{.36n_2}$$

be the population squared standard error of $\bar{X}_{t1} - \bar{X}_{t2}$, where σ_{wj}^2 is the population Winsorized variance for the j^{th} group. Our immediate goal is to find a function that determines power under normality given δ and τ. We have been unsuccessful at finding an analytic function that has practical value, so we determined our required function by setting $n_1 = n_2 = 100$ and via simulations based on 10,000 replications, we then determined the function γ that approximates power for a wide range of γ values corresponding to $\delta = 0$ up to a δ for which power is close to one. Details about how to compute γ are in an appendix. We then checked the accuracy of this function when sample sizes are small and equal ($n_1 = n_2 = 20$) and when sample sizes are unequal ($n_1 = 20$ and $n_2 = 40$).

So given a γ, now we have a function for estimating power under normality and when the standard error is known. When the standard error τ is not known we simply estimate it with

$$S = \sqrt{d_1 + d_2}, \quad (2)$$

where d_j is the value of d for the j^{th} group, and given a δ we now estimate power with $\gamma(\delta, S)$. But this estimate will not be exact, so we need a one-sided confidence interval to get a conservative estimate of power. To do this we use a percentile bootstrap method. We generate bootstrap samples from each group in the manner already described and for the b^{th} bootstrap sample we let S_b^* be the bootstrap estimate of the standard error which can be used to obtain a bootstrap estimate of power. This estimate is labeled $G_b^* = \gamma_b^*(\delta, S_b^*)$, $b = 1, \ldots, B$. Letting $G_{(1)}^* \le \ldots \le G_{(B)}^*$ be these B values written in ascending order, and setting $L = [.05n]$, we take $G_{(L)}^*$ as the lower end of our .95 confidence interval for the actual power, γ. We have tried both $B = 800$ and $B = 2{,}000$. Based on the simulations below, all indications are that $B = 2{,}000$ offers no practical value over $B = 800$, so $B = 800$ is assumed henceforth.

Results

A two-step simulation study was used to check our power estimation method for both normal and non-normal distributions. The first step was to use simulations, based on 10,000 replications, to estimate the actual power for four types of distributions: Normal, symmetric with heavy tails, asymmetric with relatively light tails, and asymmetric with relatively heavy tails. That is, we chose a set of δ values so that the true power, γ, would have a reasonable range of values between 0 and 1. (The actual values for γ will be described momentarily.)

Given n_1 and n_2, we generated observations for both groups and increased the values in the second group by δ. Then for each replication we rejected the hypothesis of equal trimmed means if $\hat{p}_m^* \le \alpha/2$, and γ was estimated with the proportion of times H_0 was rejected.

In the second step, we ran another simulation where power is estimated with our proposed method. That is, for each replication we computed S which then is used to obtain a point estimate of γ, then we used our bootstrap method to compute a .95 confidence interval for γ, and the actual probability coverage was estimated with the proportion of confidence intervals containing the value of γ determined in step 1. The nominal probability coverage was set at .95, so the intended probability of not containing the true power is $\alpha = .05$. That is, we estimate α with $\hat{\alpha}$, the proportion of intervals not containing the true power.

In our simulations, observations were generated from g-and-h distributions which includes normal distributions as a special case. If Z is a standard normal random variable, then an observation, X, from the g-and-h distribution is given by

$$X = \left(\frac{e^{gZ} - 1}{g}\right) e^{hZ^2/2}.$$

When $g = 0$, this last expression is taken to be $X = Ze^{hZ^2/2}$. The case g=h=0 corresponds to a standard normal random variable. With g= 0, X has a symmetric distribution with increasingly heavier tails as h gets large. As g increases from 0, the distribution becomes more skewed. Hoaglin (1985) gave a detailed description of the g-and-h distribution. (For some additional properties, see Wilcox, 1997.) Table 1 lists the skewness (κ_1) and kurtosis (κ_2) for the four distributions considered here. When $h > 1/k$ and $g > 0$, $E(X - \mu)^k$ is not defined and the corresponding entry in Table 1 is left blank. It might be argued that g=h=1 is an unrealistic departure from normality, but one of our goals is to determine how our method performs under seemingly extreme conditions.

Table 1: Some properties of the g-and-h distribution.

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	1.0	0.00	—
1.0	0.0	6.18	113.9
1.0	1.0	—	—

Table 2 contains $\hat{\alpha}$ values (estimated one-sided probability coverage for γ) for $n_1 = n_2 = 20$. Simulations were conducted with $n_1 = 20$ and $n_2 = 40$; similar results were obtained with other sample sizes and are not reported. Next, we ran simulations with $n_1 = 40$ and $n_2 = 20$, but the second group has a standard deviation four times as large as the first group. For normal distributions the results were:

δ:	1.0	1.4	2.0	2.4	3.0
$\hat{\alpha}$:	.021	.030	.038	.045	.045
γ:	.189	.314	.539	.685	.859

Similar results were obtained when sampling from a symmetric heavy-tailed distribution, but unsatisfactory results were obtained when sampling from the two skewed distributions considered here. More precisely, the $\hat{\alpha}$ values now exceed .1. Setting $n_1 = n_2 = 40$ does not correct this problem. That is, our proposed method now overestimates power. So in practical terms, if our technique indicates that power is low for given values of δ, n_1 and n_2, all indications are that this is indeed the case. If the estimated power is judged to be sufficiently high, our simulations indicate that this will be the case for a shift model (where distributions differ in location only), or situations where distributions are symmetric. So we have some perspective on whether the sample sizes are sufficiently large. But for skewed distributions having unequal variances, the actual power might be less than what is indicated. So progress has been made for some important special cases, but more needs to be done.

Table 2: Estimates of α, $n_1 = n_2 = 20$.

		$g = h = 0$			
δ:	0.2	0.4	0.6	0.8	1.0
$\hat{\alpha}$:	.022	.015	.011	.018	.015
γ:	.101	.257	.506	.741	.904
		$g = 1, h = 0$			
δ:	0.4	0.6	0.8	1.0	1.2
$\hat{\alpha}$:	.022	.027	.039	.051	.071
γ:	.208	.382	.571	.734	.842
		$g = 0, h = 1$			
δ:	0.6	1.0	1.4	1.8	2.2
$\hat{\alpha}$:	.014	.018	.031	.043	.077
γ:	.199	.433	.652	.796	.884
		$g = h = 1$			
δ:	0.8	1.2	1.6	2.0	2.8
$\hat{\alpha}$:	.014	.025	.039	.065	.120
γ:	.265	.445	.601	.703	.827

Note: $\delta = \mu_{t1} - \mu_{t2}$, $\hat{\alpha}$: estimate of α, γ: actual power being estimated.

An Illustration

We illustrate our method with data from a reading study. (The data were generously supplied by Frank Manis, Department of Psychology, University of Southern California.) For one of the measures studied, theoretical arguments suggest that two particular groups

should differ, but no significant difference was found using the percentile bootstrap method described previously. (Non-significant results were obtained with Student's t as well.) Figure 2 shows an estimate of the power curve for these data. (The *S-PLUS* functions used to create this plot are available from the first author upon request.) The upper solid line is the estimated power and the lower dashed line marks the one-sided confidence interval. The estimate is that a difference between the trimmed means of $\mu_{t1} - \mu_{t2} = 600$ corresponds to power equal to .8, approximately, and the confidence interval indicates that power could be as low as .6. So in this particular case all indications are that power is inadequate except for a very large difference between the trimmed means. That is, the empirical results do not provide a compelling argument that the theory is wrong because if the groups differ by a substantial amount, there is a low probability of detecting this based on the sample sizes used.

Conclusion

The main result in this paper is a method for detecting situations where power is too low. Our method of estimating power provides perspective regarding a shift model regardless of whether the distributions differ in scale. Moreover, all indications are that a reasonably accurate estimate of power can be had when distributions differ in scale provided they are symmetric. But more needs to be done. In particular, an accurate confidence interval for power is needed when distributions are skewed and have unequal variances. Another goal of possible interest is a .95 confidence band for the estimated power curve. That is, rather than compute a .95 confidence interval for each δ of interest, compute a confidence band where for all δ values, the simultaneous probability coverage is .95.

References

Bishop, T., & Dudewicz, E. (1978). Exact analysis of variance with unequal variances. Test procedures and tables. Technometrics, 20, 419-420.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Hall, P. (1986a). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics, 14*, 1453–1462.

Hall, P. (1986b). On the bootstrap and confidence intervals. *Annals of Statistics, 14*, 1431–1452.

Hewett, J. E., & Spurrier, J. D. (1983). A survey of two stage tests of hypotheses: Theory and application. *Communications in Statistics–Theory and Methods, 12*, 2307–2425.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.), *Exploring data tables, trends, and shapes*. 461–515. New York: Wiley.

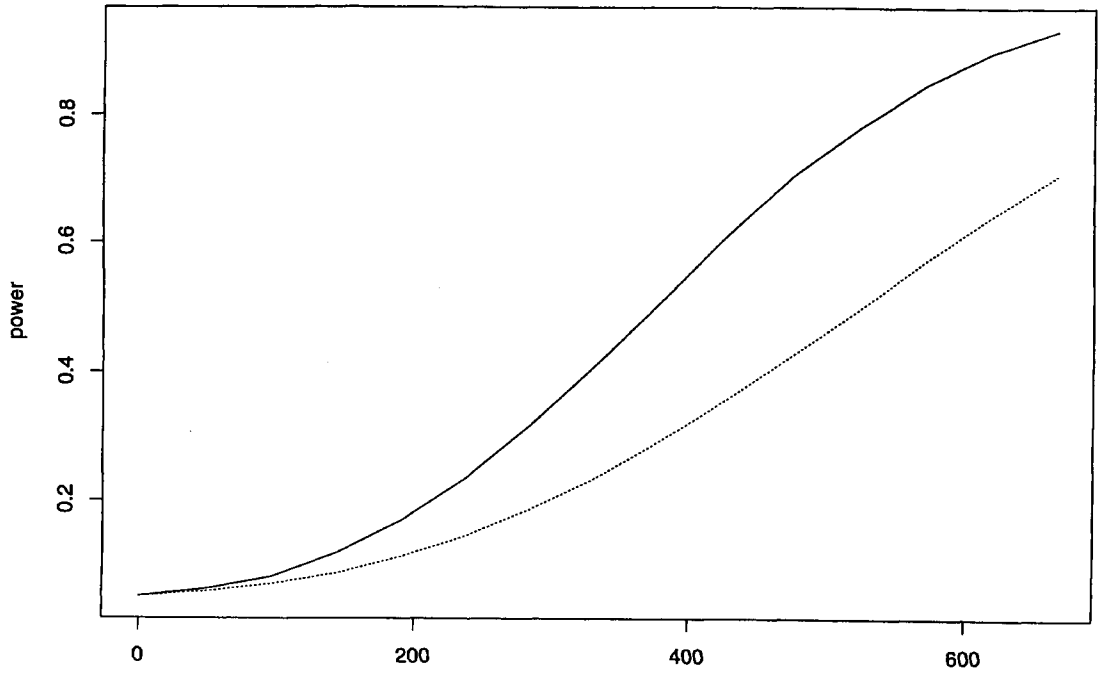


Figure 2

Hochberg, Y. (1975). Simultaneous inference under Behrens-Fisher conditions: *A two sample approach. Communications in Statistics, 4*, 1109–1119.

Huber, P. (1981). *Robust statistics*. New York: Wiley.

Hutson, A. D., & Ernst, M. D. (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society, 62*, 89–94.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donohue, B., Kowalchuk, R. K., Lowman, L. L., Petosky, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Annals of Statistics, 26*, 1719–1732.

Staudte, R. G., & Sheather, S. J. (1990) *Robust estimation and testing*. New York: Wiley.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Statistics, 16*, 243–258.

Tamhane, A. (1977). Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics, A6*, 15–32.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: Wiley.

Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal, 36*, 259–273.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego: Academic Press.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (2001a). *Fundamentals of modern statistical methods*. New York: Springer.

Wilcox, R. R. (2001b). Pairwise comparisons of trimmed means for two or more groups. *Psychometrika, 66*, 421-444.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika, 61*, 165–170.

Appendix

The construction of our function for estimating power began by determining power corresponding to multiples of the standard error of the difference between the trimmed means. We considered 36 multiples of the standard error, beginning with zero (in which case $\mu_{t1} - \mu_{t2} = 0$) and ending with 4.375 (meaning that the difference between the trimmed means is 4.375 standard errors). These 36 multiples of the standard error are given by $y_i = (i - 1)/8$, $i = 1, \ldots, 36$. So, for example, $y_{36} = 4.375$ is the largest difference between the trimmed means we considered in standard errors. Let h_i ($i = 1, \ldots, 36$) be given by 500, 540, 607, 706, 804, 981, 1176, 1402, 1681, 2008, 2353, 2769, 3191, 3646, 4124, 4617, 5101, 5630, 6117, 6602, 7058, 7459, 7812, 8150, 8479, 8743, 8984, 9168, 9332, 9490, 9607, 9700, 9782, 9839, 9868.

For example, suppose $h_1 = 500$ and $h_3 = 706$. The value h_i, divided by 10,000, is the power corresponding to $\mu_{t1} - \mu_{t2} = y_i\sqrt{2 * .01155}$ when $n_1 = n_2 = 100$ and sampling is from a standard normal distribution. The constant .01155 is approximately equal to the squared standard error of the 20% trimmed mean. That is, $h_i/10000$ gives power as a multiple of the standard error of $\bar{X}_{t1} - \bar{X}_{t2}$. Thus, for any two distributions, given δ and S, an estimate of the standard error of $\bar{X}_{t1} - \bar{X}_{t2}$ given by equation 2, power is estimated as follows: $v = [8 * \delta/S] + 1$, where [.] is the greatest integer function, and let $d = 8 * \left(\frac{\delta}{S} - \frac{v-1}{8}\right)$. Then, the estimated power is taken to be $\hat{\gamma} = \frac{h_v}{10000} + d * \left(\frac{h_{v+1}}{10000} - \frac{h_v}{10000}\right)$. In the event $v = 36$, h_{v+1} is taken to be 10000 in the previous equation. If $v > 36$, $\hat{\gamma} = 1$.

Note: The research reported in this article was supported, in part, by a grant from the Natural Sciences and Engineering Research Council of Canada.