Psychological Methods
2003, Vol. 8, No. 3, 254–274

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.3.254

Modern Robust Data Analysis Methods: Measures of Central Tendency

Rand R. Wilcox
University of Southern California

H. J. Keselman
University of Manitoba

Various statistical methods, developed after 1970, offer the opportunity to substantially improve upon the power and accuracy of the conventional *t* test and analysis of variance methods for a wide range of commonly occurring situations. The authors briefly review some of the more fundamental problems with conventional methods based on means; provide some indication of why recent advances, based on robust measures of location (or central tendency), have practical value; and describe why modern investigations dealing with nonnormality find practical problems when comparing means, in contrast to earlier studies. Some suggestions are made about how to proceed when using modern methods.

The advances and insights achieved during the last half century in statistics and quantitative psychology provide an opportunity for substantially improving psychological research. Recently developed methods can provide substantially more power when the standard assumptions of normality and homoscedasticity are violated. They also help deepen our understanding of how groups differ. The theoretical and practical advantages of modern technology have been documented in several books (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Hoaglin, Mostelier, & Tukey, 1983, 1985; Huber, 1981; Rousseeuw & Leroy, 1987; Wilcox, 1997a, 2001, 2003) and journal articles. Yet, most applied researchers continue to believe that conventional methods for making inferences about means perform well in terms of both controlling the Type I error rate and maintaining a relatively high level of statistical power. Although several classic articles describe situations in which this view is correct, more recent publications provide a decidedly different picture of the robustness of conventional techniques. In terms of avoiding actual Type I error probabilities larger than the nominal level (e.g., $\alpha = .05$), modern methods and conventional techniques produce similar results when groups have identical distributions. However, when distributions differ in skewness or have unequal variances, modern methods can have substantially more power, they can have more accurate confidence intervals, and they can provide better control over the probability of a Type I error. We also indicate why some commonly used strategies for correcting problems with methods based on means fail, and we summarize some recent strategies that appear to have considerable practical value.

Articles summarizing some of the basic problems with conventional methods and how they might be addressed have appeared in technical journals (e.g., Wilcox, 1998a) and basic psychology journals (e.g., Wilcox, 1998b). Wilcox (2001) also provided a nontechnical description of practical problems with conventional methods and how modern technology addresses these issues. Our goal here is to expand on this previous work by summarizing some recent advances.[1] But first, we review basic principles motivating modern methods.

Rand R. Wilcox, Department of Psychology, University of Southern California; H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada.

Work on this article by H. J. Keselman was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We are grateful to M. Earleywine and David Schwartz for comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles, California 90089-1061. E-mail: rwilcox@usc.edu

[1] This article does not discuss recent advances related to rank-based methods, but this is not to suggest that they have no practical value. For a summary of important and useful developments related to rank-based techniques, see Brunner, Domhof, and Langer (2002); Cliff (1996); and Wilcox (2003).

We do not promote a single approach to data analysis but rather argue that modern technology as a whole has much to offer psychological research. No single statistical method is ideal in all situations encountered in applied work. In terms of maximizing power, for example, modern methods often have a considerable advantage, but the optimal method depends in part on how the groups differ in the population, which will be unknown to the researcher. Modern methods can also provide useful new perspectives that help us develop a better understanding of how groups differ. In the Appendix to this article, we provide an overview of statistical software that can implement the modern methods to be described.

Some Basic Problems

We begin with the one-sample case in which the goal is either (a) to test

$$H_0: \mu = \mu_0,$$

the hypothesis that the population mean μ is equal to some specified constant μ_0, or (b) to compute a confidence interval for μ. When data in a single sample are analyzed, two departures from normality cause problems: skewness and outliers.

Skewness

First, we illustrate the effects of skewness. Imagine we want to assess an electroencephalographic (EEG) measure at a particular site in the brain for individuals convicted of murder. If we randomly sample n participants, the most commonly used approach is to estimate the population mean μ with the sample mean, M. The conventional $1 - \alpha$ confidence interval for μ is

$$M \pm t_{1-\alpha/2}\left(\frac{SD}{\sqrt{n}}\right), \tag{1}$$

where SD is the sample standard deviation and $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of Student's t distribution with $n - 1$ degrees of freedom. This confidence interval is based on the assumption that

$$T = \frac{M - \mu}{SD/\sqrt{n}} \tag{2}$$

has a Student's T distribution. If this assumption is reasonably correct, control over the probability of a Type I error, when hypotheses are tested, will be achieved.[2]

Following an example by Westfall and Young (1993), suppose that unknown to us, observations are sampled from the skewed (lognormal) distribution shown in the left panel of Figure 1. The dotted line in the right panel of Figure 1 shows the actual distribution of T when $n = 20$. The smooth symmetrical curve shows the distribution of T when a normal dis-

[2] We are following the convention that uppercase letters represent random variables and lowercase letters represent specific values (e.g., Hogg & Craig, 1970). So T represents Student's T random variable, and t represents a specific value, such as the .975 quantile.

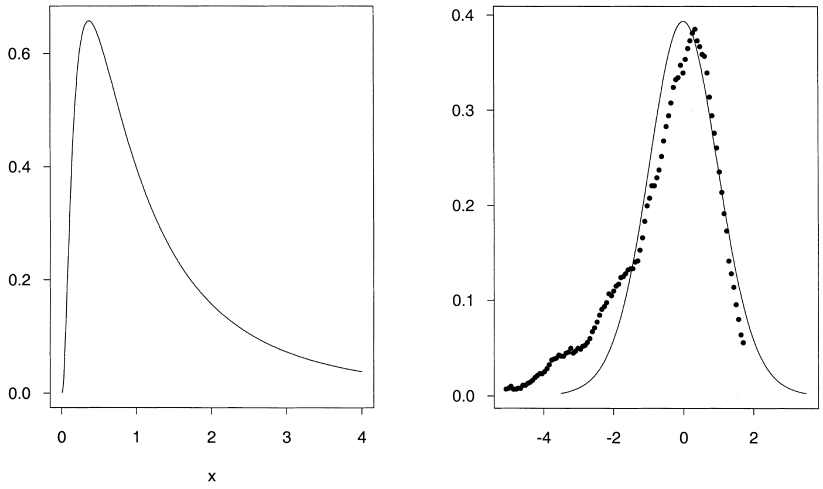

Figure 1. The left panel shows a lognormal distribution. The right panel shows an approximation of the distribution of Student's T when 20 observations are sampled from a lognormal distribution. The solid line represents the distribution under normality.

tribution is sampled. As is evident, there is a serious discrepancy between the actual distribution of T versus the distribution under normality, and this results in very poor control over the probability of a Type I error when $n = 20$. Westfall and Young (p. 40) noted that problems with controlling the probability of a Type I error persist even when $n = 160$. When the stated alpha is equal to .10, and when one is performing a two-tailed test, the actual probability of a Type I error for the lower tail is .11 (vs. $\alpha/2 = .05$). For the upper tail, the actual probability of a Type I error is .02. Thus, when one is performing the usual two-tailed test at the .10 level, the actual probability of a Type I error is .11 + .02 = .13.

Any hypothesis-testing method is said to be *unbiased* if the probability of a Type I error is minimized when the null hypothesis is true; otherwise, it is said to be biased. In cases such as the example above, where for the lower tail the actual probability of a Type I error is less than $\alpha/2$, power can actually decrease as the true mean increases (Wilcox, 2003). This means that for a two-tailed test, Student's T is biased. That is, there are situations in which we are more likely to reject when the null hypothesis is true versus when it is false.

Extending the Westfall and Young (1993) results, we find that for the same skewed distribution, but with $n = 200$, the actual Type I error probability when a lower tailed test is used is .07, and for $n = 250$ it is .06. That is, if we happen to be sampling from a lognormal distribution, and if an actual Type I error probability less than or equal to .08 is deemed adequate, nonnormality is not an issue in terms of Type I error probabilities and accurate confidence intervals with a sample size of at least 200.

Now, to expand on the problems with Student's T summarized by Westfall and Young (1993), suppose observations are sampled from the skewed distribution[3] shown in Figure 2. Among a sample of observations, outliers are more common versus the lognormal in Figure 1. The dotted line in the left panel of Figure 3 shows 5,000 T values with $n = 20$. The smooth symmetrical curve is the distribution of T assuming normality. Under normality there is a .95 probability that T will be between −2.09 and 2.09; these are the .025 and .975 quantiles of the distribution of T. However, when one samples from the dis-

[3] This distribution arises by sampling from a chi-square distribution having 4 degrees of freedom, and with probability .1 multiplying an observation by 10. (That is, generate X from a chi-square distribution, generate u from a uniform distribution, and if $u < .1$, multiply X by 10.) When one is sampling from this distribution with $n = 20$, the median number of outliers is 2; the mean is about 1.9; and with probability about .95, the number of outliers is less than or equal to 3. Based on $n = 10{,}000$ values, the skewness and kurtosis of this distribution are 4.98 and 34.89, respectively. The lognormal distribution has skewness and kurtosis equal to 6.2 and 113.9, respectively.

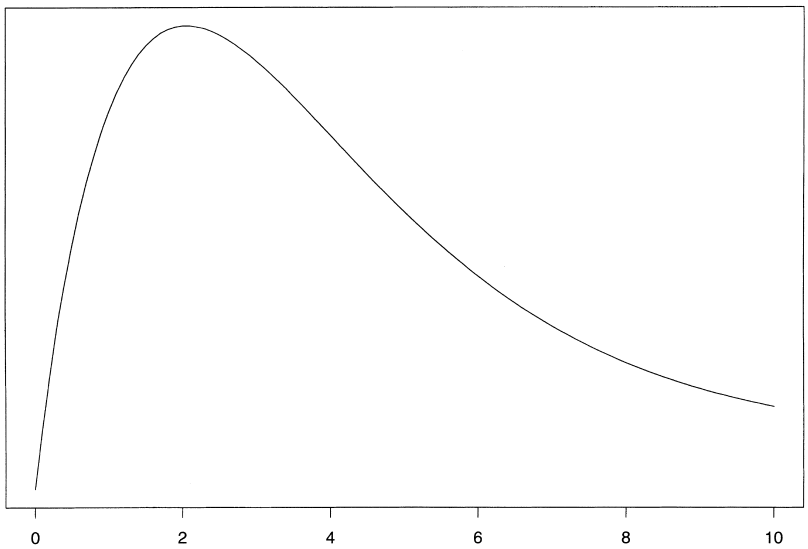

Figure 2. A skewed, heavy-tailed distribution that is formed in a manner similar to the contaminated normal distribution. More precisely, sample observations from a chi-square distribution with 4 degrees of freedom, and with probability .1, multiply an observation by 10.

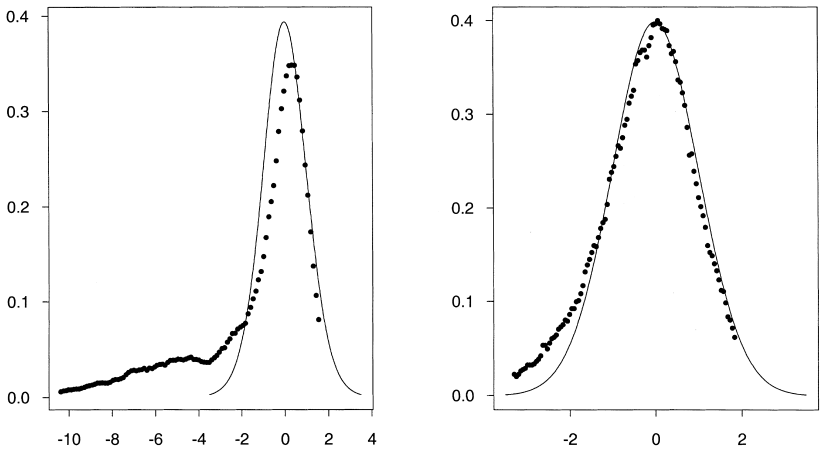

Figure 3. Probability density function of Student's T when sampling from the skewed, heavy-tailed distribution in Figure 2. The left panel is with $n = 20$, and the right is with $n = 300$. Even with $n = 300$, control over the probability of a Type I error is unsatisfactory. The solid line represents the distribution of T under normality. The dotted line represents the actual distribution of T.

tribution in Figure 2, these quantiles are approximately −8.50 and 1.29, respectively. The right panel of Figure 3 shows a plot of 5,000 T values when $n = 300$. There is closer agreement with the assumed distribution. Under normality, the .025 and .975 quantiles are approximately −1.96 and 1.96, respectively. However, the .025 and .975 quantiles of the actual distribution of T are approximately −2.50 and 1.70. The mean of the distribution shown in Figure 2 is 7.60. If we conduct a one-tailed test of H_0: $\mu \geq 7.60$ at the .05 level, the actual probability of a Type I error is approximately .1, twice the nominal level. The correct critical value under normality is −1.96, but when sampling from the distribution in Figure 2, it is −2.50. In practical terms, the results of these examples illustrate that when the sample size is small ($n = 20$), the Type I error rate and the confidence intervals can be highly inaccurate when the data are skewed. These problems diminish as the sample size increases, but substantial problems may arise even with $n = 300$ when outliers are likely to occur.

Realistic Departures From Normality

One could argue that in theory problems might arise but that in applied work problems are never as serious as suggested by Figures 1 and 3. The bumpy curve in Figure 4 shows a bootstrap approximation of the sampling distribution of Student's T based on data (with $n = 20$) taken from a study comparing hangover symptoms of sons of individuals with alcoholism with a control group.[4] (These data had skew and kurtosis equal to 2.30 and 7.60, respectively.) Figure 4 was created by resampling with replacement 20 observations from the first group, computing T, repeating this 5,000 times, and plotting the results. Again, there is a serious discrepancy between the estimated distribution and the distribution of T under normality.

Figure 5 shows a second example using data from a study about the sexual attitudes of young adults.[5] Undergraduate males were asked how many sexual partners they desired over the next 30 years. The bumpy curve in Figure 5 shows an approximation of the distribution of Student's T. (The skew and kurtosis are 5.68 and 37, respectively.) Despite a larger sample size ($n = 104$), there is still a substantial difference between the bootstrap estimate of the actual distribution of T versus the distribution under normality. In fact, there were actually 105 observations, one of which was an extreme outlier that we excluded. If this outlier is included, it makes the bootstrap estimate of the distribution of T even more asymmetrical than shown in Figure 4. The hypothetical distribution in Figure 2, as well as the illustration in Figure 3, clearly does not underestimate problems that can occur.

In Figures 4 and 5, the sampling distribution of T

[4] These data were provided by M. Earelywine (personal communication, 1998).

[5] These data were provided by W. Pedersen (personal communication, 1999).

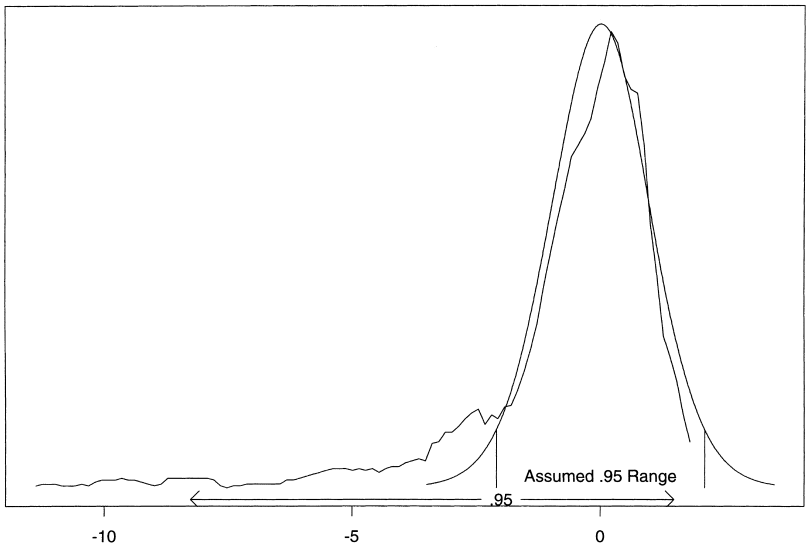

Figure 4. An approximation of the distribution of *T* based on data from a study dealing with sons of individuals with alcoholism. The approximation is obtained by resampling data with replacement, computing *T,* repeating this process 5,000 times, and plotting the resulting *T* values. The *x*-axis represents the possible values of *T.* The arrows indicate the location of the central 95% of the *T* values. When one is sampling from a normal distribution, these central *T* values fall between the two vertical lines of the smooth symmetrical curve.

was approximated by resampling with replacement observations from the original data, computing *T,* and repeating this process 5,000 times. A reasonable concern is that perhaps this does not give us a true indication of how Student's *T* performs in applied work. That is, the original data might provide a poor approximation of the true (population) distribution, and if we could sample repeatedly from the true distribution, perhaps Student's *T* would perform in a more satisfactory manner. However, other work, summa-

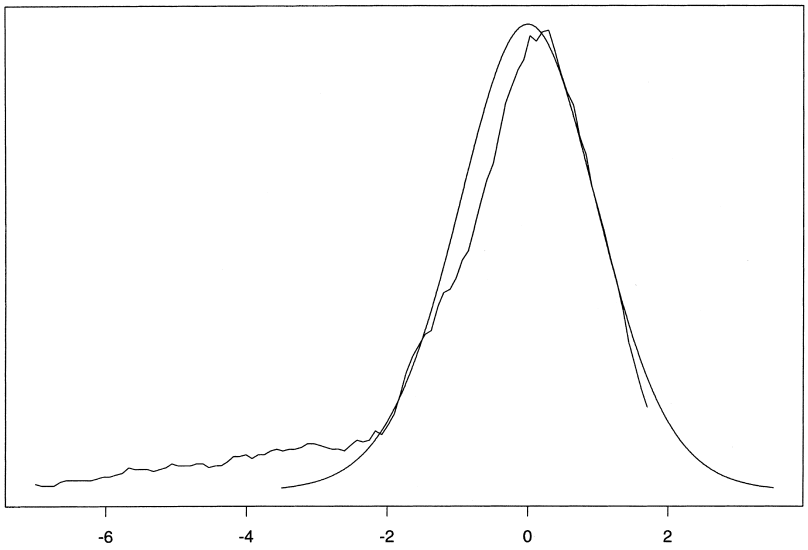

Figure 5. An approximation of the distribution of *T* based on data from a study dealing with the sexual attitudes of young adults. The approximation is based on the same strategy used to create Figure 4.

rized in Wilcox (1997a), suggests that the figures presented here do not overestimate the practical problems with Student's *T*.

Outliers and Power

Even when distributions are symmetrical in the population, practical problems can occur. These situations arise when one is sampling from heavy-tailed distributions, in which case outliers are common. Indeed, small departures from normality, which can be very difficult to detect, can substantially reduce power when using conventional methods (e.g., Staudte & Sheather, 1990).[6] The basic problem is that small changes in any distribution, including normal distributions as a special case, can greatly increase the variance, which in turn lowers power.

As an example, imagine that the data include two groups, say individuals who have or do not have schizophrenia. Assume that 90% of these individuals do not have schizophrenia and that these individuals have a standard normal distribution ($\mu = 0$ and $\sigma = 1$, where σ is the population standard deviation) on the dependent measure. Further assume that the individuals with schizophrenia also have a normal distribution with the same mean but with a much larger standard deviation of 10 ($\mu = 0, \sigma = 10$). If we pool these two groups and randomly sample an individual, we are randomly sampling from what is called a *contaminated normal distribution.* Figure 6 shows this particular contaminated normal distribution and a standard normal distribution. What is important here is that there is little visible difference between these two distributions, yet their standard deviations differ substantially—for the standard normal $\sigma = 1$, but for the contaminated normal distribution $\sigma = 3.30$. This example illustrates the fundamental principle that the standard deviation is highly sensitive to the tails of a distribution.

The sensitivity of the variance to slight changes in the tails of a distribution has many important implications, one being the effect on power. To illustrate this effect in the two-sample case, consider the left panel of Figure 7, which shows two normal distributions ($\mu_1 = 0$, $\sigma_1 = 1$; $\mu_2 = 1$, $\sigma_2 = 1$). If we sample 25 observations from each and test the hypothesis of equal means with Student's *T* at the .05 level, power is .96. Now look at the right panel of

[6] The best-known metric is the Kolmogorov metric, which is used by the Kolmogorov–Smirnov test of fit. Others are the Lipshitz metric, the Lévy metric, and the Prohorov metric, but the details are too invovled to give here (see Huber, 1981).

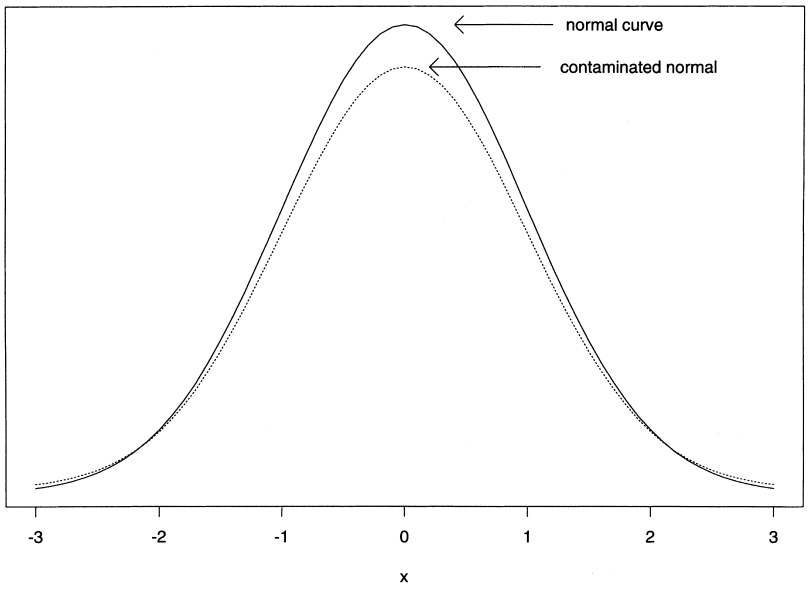

Figure 6. Normal and contaminated normal distributions. The solid line represents a standard normal distribution, and the dashed line represents a contaminated normal distribution. The normal distribution has variance 1, and the contaminated normal distribution has variance 10.9, illustrating that variance is highly sensitive to the tails of a distribution.

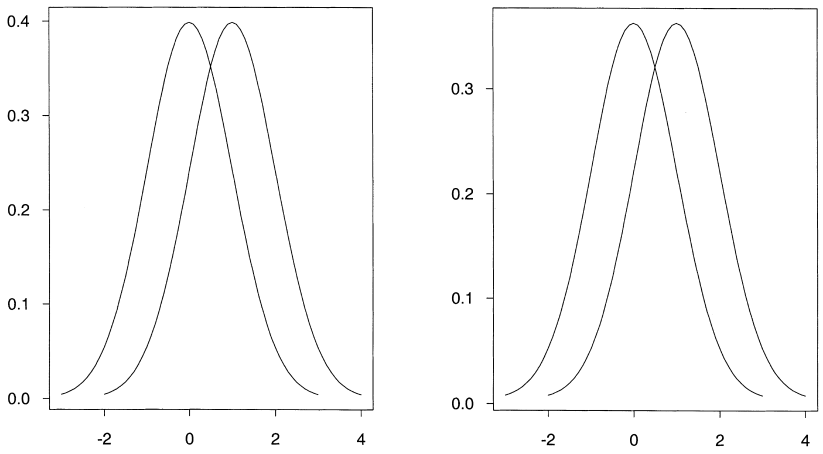

Figure 7. Power and nonnormality. In the left panel, Student's *T* has power .96, but in the right panel power is only .28.

Figure 7, which contains two contaminated normal distributions. (The same amount of contamination is used as in Figure 6: $\mu_1 = 0$, $\sigma_1 = 3.30$; $\mu_2 = 1$, $\sigma_2 = 3.30$). Despite the right panel's obvious visual similarity to the left panel, if we sample from the two contaminated normal distributions, now power is only .28. Power decreases mainly because outliers are more common when one is sampling from the contaminated normal distribution. One effect of the outliers is that they inflate the sample variance, SD^2, which in turn lowers the value of Student's *T*. A second effect is that the probability of a Type I error is less than the nominal alpha level, and this contributes somewhat to lower power.

Of course, as the sample size increases, the power of Student's *T* will increase when one is sampling from the contaminated normal distribution. However, for the contaminated normal distribution, the power of the conventional *t* test will always be less than the power achieved with modern techniques. Modern methods are specifically designed to avoid potential power problems associated with methods based on means due to sampling from heavy-tailed distributions.

The low power associated with Student's *T* might appear to contradict earlier publications claiming that Student's *T* maintains high power under nonnormality. The reason for the apparent discrepancy is that these researchers studied power based on a standardized difference between the means:

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

where $\sigma_p^2 = \sigma_1^2 = \sigma_2^2$ is the assumed common variance. Unfortunately, Δ is not robust; very slight changes in a distribution in the population can substantially alter its value. In the left panel of Figure 7, for example, $\Delta = 1$, but in the right panel it is only .3. This result illustrates that relying on Δ can grossly underestimate the degree to which two distributions differ. Considering only Δ, power should be less for the situation in the right panel versus the left. However, from a practical point of view, surely it is desirable to have about as much power when sampling from the contaminated distributions in the right panel versus the two normal distributions in the left panel. Methods that achieve this goal are now available, some of which are described below.

In summary, there are three general problems when using Student's *T*. First, skewness can cause problems when one is trying to control the probability of a Type I error. Problems are serious when one is sampling from relatively light-tailed skewed distributions[7] and they get worse as one moves toward situations in which outliers are common. Second, Student's *T* can

[7] The lognormal distribution has been characterized by some as being *heavy tailed,* a conclusion that seems reasonable based on its kurtosis, but Gleason (1993) argued that it is more accurately characterized as being *light tailed*. We have checked the expected number of outliers when sampling from a lognormal distribution, and our results support Gleason's view. As illustrated by Figure 6, the variance is highly sensitive to the tails of a distribution. Conventional measures of skewness and kurtosis are even more sensitive.

be biased when one is sampling from a skewed distribution, which in turn can result in low power relative to alternative modern methods (described later). Many attempts have been made to correct problems due to skewness, but no satisfactory technique has been found when attention is restricted to means. Third, outliers can result in very low power when one is using Student's T compared with modern methods even when one is sampling from a perfectly symmetrical distribution.

Simple Transformations

A common recommendation when dealing with nonnormality is to use simple transformations of the data. That is, each value is transformed in the same manner by taking logarithms, or square roots, or using more complex approaches such as the class of Box–Cox transformations (e.g., Cook & Weisberg, 1999). Unfortunately, simple transformations do not guard against low statistical power when one is sampling from heavy-tailed distributions (e.g., Rasmussen, 1989; also see Doksum & Wong, 1983). Simple transformations can alter skewed distributions so that they are more symmetrical, but they do not deal directly with outliers. Sometimes the number of outliers is decreased, but typically outliers remain, and in some cases the number of outliers actually increases. For example, Wilcox (2003, Table 8.4) reported data from a study of self-awareness. For the second group in that example, the data are skewed to the right and a single outlier is found. When logarithms are taken, the data have a more symmetrical distribution, but the number of outliers increases from one to three. What is needed are transformations that deal directly with outliers and eliminate their deleterious effects. One such transformation is trimming, which is discussed later in this article.

The Two-Sample Case: New Versus Old Findings

The properties of the one-sample t test just summarized provide a framework for explaining the discrepancy between conclusions of earlier and more recent investigations regarding the robustness of conventional techniques such as the two-sample t test and analysis of variance (ANOVA). Consider again the example of the EEG measure of murderers, and imagine the researcher's goal is to compare these participants with a control group. Let $F_1(x)$ be the probability that a randomly sampled murderer has an EEG measure less than or equal to x. Thus, $F_1(3)$ is the probability of a reading less than or equal to 3, and $F_1(1)$ is the probability of a reading less than or equal to 1. Similarly, let $F_2(x)$ be the probability that a randomly sampled participant from the control group has an EEG measure less than or equal to x. These two distributions are identical if for any x we might pick, $F_1(x) = F_2(x)$. Of course, in this particular case, both groups have identical variances as well as the same amount of skewness. If equal sample sizes are used, it can be shown that when groups have the same skewness, the difference between the sample means will have a perfectly symmetrical distribution. This suggests that, generally, the actual probability of a Type I error will not exceed the nominal level of alpha when one is sampling from identical, nonnormal distributions. Empirical evidence for this view was reported by Sawilowsky and Blair (1992), and this is exactly what was found in numerous other studies summarized in Wilcox (1997a).

However, unlike the better known earlier robustness studies, more recent investigators have considered situations in which the two distributions differ in shape. Theory suggests that such situations are more likely to cause practical problems when one is using conventional methods or indeed any method aimed at comparing means. These more recent studies show that when distributions differ in skewness, conventional methods might provide poor control over the probability of a Type I error as well as inaccurate confidence intervals (e.g., Wilcox, 1996). That is, nonnormality becomes a practical issue when distributions differ in shape.

Yet another problem with conventional methods occurs when there is *heteroscedasticity* (unequal variances in the groups). Even when groups have a normal distribution, heteroscedasticity can cause conventional tests to be biased. Moreover, a basic requirement of any statistical method is consistency—it should converge to the correct answer as the sample size increases. Thus, if alpha is set to .05, the actual probability of a Type I error should approach .05 as the number of observations gets large. Cressie and Whitford (1986) described general conditions in which Student's T does not enjoy this property because it is using the wrong standard error.

When sample sizes are equal and each group has a normal distribution in the population, comparing two groups with Student's T controls the probability of a Type I error fairly well except possibly with very small sample sizes, regardless of how unequal the variances might be (Ramsey, 1980). However, when

we look at a broader range of situations, this is the exception rather than the rule. In fact, a rough rule is that the more groups we compare, the more conventional methods become sensitive to violations of assumptions.

Practical problems with heteroscedasticity might appear to contradict classic studies. Box (1954) showed good control over the Type I error rate under conditions of normality when the group variances were not equal. However, Box's numerical results were based on situations in which, among the groups being compared, the largest standard deviation divided by the smallest did not exceed $\sqrt{3}$. As this ratio increases, or when group distributions are nonnormal, practical problems arise (e.g., Wilcox, 1987). Wilcox (1987) found in a survey of studies that this ratio frequently exceeds 4. A more recent review of the education and psychology literature by Keselman et al. (1998) found ratios as large as 23.8. Grissom (2000) reviewed just one issue of the *Journal of Consulting and Clinical Psychology,* finding that the ratios of the variances exceeded 3 in many cases and went as high as 282.

Another classic study by Ramsey (1980) demonstrated that if two groups are being compared via Student's *T,* sampling is from normal distributions, and equal sample sizes are used, reasonable control over the probability of a Type I error is achieved. The only exception Ramsey identified was for very small sample sizes. More recent investigations (e.g., Wilcox, Charlin, & Thompson, 1986) extended the conditions studied by Ramsey and showed that problems controlling the probability of a Type I error can appear (a) under normality with equal sample sizes when comparing four or more groups; (b) under normality with unequal sample sizes when there are two or more groups; or (c) under nonnormality when comparing two or more groups, even when sample sizes are equal.

A reasonable suggestion for trying to salvage homoscedastic methods is to test the assumption of equal variances. However, in practice this strategy can fail even under normality (Markowski & Markowski, 1990; Moser, Stevens, & Watts, 1989; Wilcox et al., 1986). Tests for equal variances rarely have enough power to detect differences in variances of a magnitude that can adversely affect conventional methods. In some situations, even testing for unequal variances at the $\alpha = .25$ level does not increase power sufficiently. Additional concerns and issues related to popular methods for testing the hypothesis of equal variances are described in DeCarlo (1997), Keyes and Levy (1997), and Wilcox (1990, 2002a).

Alternative methods based in part on weighted least squares estimation can be used to address heteroscedasticity. Improvement over conventional methods is achieved, but problems persist, even under normality (Wilcox et al., 1986). Large sample sizes do correct problems with Type I errors and power when using these heteroscedastic techniques, but methods for judging the adequacy of sample sizes require more research before one can be recommended.

In summary, Student's *T* provides adequate control over the Type I error rate when one is comparing groups with identical distributions. However, in many situations the distributions will not be identical and it is difficult for researchers to discern this in practice. Further, when the distributions associated with groups differ in skewness, or have unequal variances, or when outliers are likely to occur, the power of Student's *T* can be relatively poor. Poor power when there are outliers is a problem that plagues any method based on sample means.

Comparing Skewed Distributions

A brief consideration of three strategies for comparing two skewed distributions might be helpful before continuing. One strategy would be to compare all of the quantiles simultaneously. For example, we could compare the .10 quantiles of the two groups, the .25 quantiles (first quartile), the .50 quantiles (medians), the .75 quantiles (third quartiles), and the .90 quantiles. This strategy provides information about how both the central portions and the tails of the distributions differ. Even when one is comparing symmetrical distributions, this strategy, known as the *shift function,* provides a more detailed sense of how groups differ beyond any comparison based on a single measure of location or dispersion. The probability of at least one Type I error can be controlled exactly using an extension of the Kolmogorov–Smirnov test (Doksum & Sievers, 1976). Moreover, Doksum and Sievers proposed an interesting and useful graphical method for getting a global sense of how groups compare. Power appears to be relatively good when there are no tied values among the pooled data (Wilcox, 1997b). However, to the extent that tied values do occur, power can be low relative to other approaches to be described. (When using S-PLUS, or the free software R, which can be obtained as indicated in the Appendix, the computations are per-

formed by the function sband in Wilcox, 1997a, 2003, which also plots the shift function.)

Another strategy for comparing skewed distributions is to compare groups based on a single measure of location (or central tendency) that is near the central portion of the data. One advantage of using a single measure of location is that it can be applied to virtually all of the commonly used experimental designs. Comparing medians is perhaps the most obvious strategy, but if the goal is to maintain high power under normality or when sampling from relatively light-tailed distributions, alternatives to the median are preferable, two of which are described later in this article.

A third strategy is to use a specific measure of location that reflects the tails of a distribution and has some particular theoretical or practical interest. For example, if only the top 10% of applicants are hired, researchers may wish to compare the .90 quantiles. However, even when there is interest in comparing the tails of skewed distributions, alternatives to means can have practical value. Wilcox (2003, section 8.3) presented an R (or S-PLUS) program pb2gen for comparing specific quantiles.

Discarding Outliers

One strategy for dealing with outliers is to discard them and apply methods for means to the data that remain. There are two fundamental problems with this strategy. The first is that outlier detection methods based on means and variances can easily fail to identify outliers. A second and perhaps more serious problem is that when extreme values are discarded, the remaining observations are no longer independent under random sampling (e.g., Hogg & Craig, 1970). The dependence induced by eliminating outliers invalidates the derivation of the standard error of the sample mean. Wilcox (2001) illustrated how, from a practical point of view, ignoring this latter issue can be highly unsatisfactory.

Here we illustrate the problem with the data ($n = 20$) used to create Figure 4. Imagine that the two smallest and two largest observations are discarded. If we compute the usual standard error based on the remaining data, we get 1.63. This is a theoretically unsound estimate because it does not take into account the correlation among the remaining data. A theoretically correct estimate yields 2.59 and can be calculated using the R or S-PLUS function trimse. The standard error of a trimmed mean is related to what is called the *Winsorized variance* (Staudte & Sheather, 1990). Computing the Winsorized variance is easily done as described in Wilcox (1997a, 2003). Here, we used the S-PLUS (or R) function trimse in Wilcox (2003). (See the library of functions described in the Appendix.) In this illustration, a fixed proportion of observations was trimmed. When discarding only those points that are declared to be outliers, similar problems occur, but now a different method for getting a theoretically correct estimate of the standard error is needed. For example, if we use a rule based on the median (given by Equation 3 below) for detecting outliers and we compute the usual standard error based on the remaining data, we get .13. However, using a theoretically correct method for estimating the standard error, implemented in the R or S-PLUS function bootse (Wilcox, 2003), yields an estimate of .57, more than 4 times larger. Moreover, given that the standard errors computed using the usual formula are too small when outliers are discarded, poor control over the probability of a Type I error can result.

Situations may arise with a large sample size in which using a theoretically correct estimate of the standard error makes little practical difference. Currently, however, the only way to determine whether this is the case is to use a theoretically correct estimate. That is, there are no guidelines on how large the sample size must be or how many outliers can be discarded before the usual formula for the standard error yields problematic results. Our recommendation is that the theoretically correct standard errors be computed routinely using one of the software packages discussed in the Appendix.

Detecting Outliers

Outlier detection rules based on the mean and sample standard deviation suffer from a problem called *masking*. Suppose the value X is declared an outlier if it is more than 2 standard deviations from the mean. In symbols, declare X an outlier if

$$\frac{|X - M|}{SD} > 2.$$

Masking means that the very presence of outliers can destroy the ability of this method to detect unusually large or small values. Consider, for instance, the values 2, 2, 3, 3, 3, 4, 4, 4, 100,000, 100,000. Surely 100,000 is unusual versus the other values, but 100,000 is not declared an outlier using the method just described. Outliers inflate both the sample mean

and the standard deviation, but in a certain sense they have more of an effect on the standard deviations, which causes outliers to be missed.

There are various outlier detection methods for dealing with masking (Barnett & Lewis, 1994). Perhaps the best known is the standard box plot with improvements recently provided by Carling (2000). Among robust methods, one popular procedure for dealing with masking is to replace the sample mean with the median and to replace the standard deviation with a measure of dispersion called the *median absolute deviation* (MAD) statistic (see Wilcox, 1997a, p. 24). Based on the observations, $X_1, \ldots X_n$, MAD is the median of the set of absolute values of the differences between each score and the median. That is, MAD is the median of $|X_1 - Mdn|, \ldots, |X_n - Mdn|$. Then the ith value, X_i, is declared an outlier if

$$\frac{|X_i - Mdn|}{\text{MAD}/.6745} > 2.24. \quad (3)$$

(The constant .6745 rescales MAD so that the denominator estimates σ when one is sampling from a normal distribution.) Thus, in the case of a normal distribution, the left side of Equation 3 estimates $|X - \mu|/\sigma$. Applying this rule in the previous illustration, $Mdn = 3.50$ and MAD/.6745 = .74, so for 100,000, the value for the outlier statistic is 134,893, which is clearly greater than 2.24.

The outlier detection method just described, based on MAD and the median, is a commonly used robust method. However, our use of this rule is not intended to suggest that looking at data, and scrutinizing it for unusual values, is to be avoided. Exploratory methods play an important role when one is analyzing and understanding data. However, for certain purposes, particularly when testing hypotheses, specifying an explicit rule for detecting outliers is essential.

Consider a location estimator $\hat{\theta}$, such as the sample mean or median. When one is testing hypotheses, the typical strategy is to derive an expression for the standard error of $\hat{\theta}$. However, when one is using estimators aimed at dealing with outliers, theoretically sound estimates of standard errors are not immediately obvious. For example, if we use a trimmed mean (described later in article), the standard error is estimated based in part on a Winsorized variance. If we compute a one-step M estimator that uses MAD and the median to determine whether a value is unusually large or small, the resulting expression for the standard error has a complex form that differs substantially from the standard error of a trimmed mean (Huber, 1981). If we use a subjective method for eliminating outliers when testing hypotheses, it is unclear how to get a theoretically correct estimate of the standard error.

Another reason for specifying a specific outlier detection rule is so that we can study its performance when attention is focused on estimating a measure of location. If the goal is to get good power under a wide range of situations, we search for an estimator that has a relatively low standard error, regardless of the distribution in the population. Such studies cannot be undertaken without specifying a specific outlier detection method. Again, this is not to argue against exploratory methods. However, when attention turns to inferential methods, being precise about how outliers are detected is essential.

Robust Measures of Location

There are two general strategies for comparing two or more groups that address the problem of low power due to nonnormality: robust measures of location and rank-based methods. Space limitations preclude a consideration of rank-based methods (see Wilcox, 2003). Robust measures of location can be further subdivided into two types. The first type simply removes or trims a fixed proportion of the smallest and largest observations and averages the data that remain. The amount trimmed is not determined by first checking for outliers. The optimal amount of trimming in any given application is unknown, but 20% appears to be a reasonable default value based on the criteria of achieving a small standard error and controlling the probability of a Type I error (e.g., Wilcox, 2003). The term *20% trimming* means that if the data are put in ascending order, the largest 20%, as well as the smallest 20%, are trimmed. If we have 10 observations, 20% trimming consists of removing the two largest and smallest values and averaging the rest.

The second general approach to robust estimation of location is to check the data for outliers, remove any that are found, and average the values that remain. One estimator related to this type that has received considerable attention is the one-step M estimator (see Huber, 1981; Staudte & Sheather, 1990; Wilcox, 2001, 2003). Let MADN = MAD/.6745, which estimates σ under normality. Count the number of observations, i_1, for which $(X_i - M)/\text{MADN} < -1.28$, and count the number of observations, i_2, for which $(X_i - M)/\text{MADN} > 1.28$. A one-step M estimator is

$$\hat{\theta}_{os} = \frac{1.28(\text{MADN})(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2},$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the observations written in ascending order. As previously indicated, an expression for the standard error of this estimator has been derived and a method for estimating it is available (Wilcox, 1997a). However, the more effective methods for testing hypotheses based on a one-step M estimator, described below, do not use an estimate of the standard error.

The term $1.28(\text{MADN})(i_2 - i_1)$ in $\hat{\theta}_{os}$ arises for technical reasons (e.g., Staudte & Sheather, 1990; Wilcox, 1997a). Ignoring it yields the so-called modified one-step M estimator (MOM):

$$\frac{1}{n - i_1 - i_2} \sum_{i=i_1+1}^{n-i_2} X_{(i)}.$$

Now, however, to achieve a reasonably small standard error under normality, i_1 is the number of observations for which $(X_i - M)/\text{MADN} < -2.24$, and i_2 is the number for which $(X_i - \text{M})/\text{MADN} > 2.24$. The one-step M estimator is a bit more satisfactory in terms of achieving a relatively small standard error, but MOM has advantages when one is testing hypotheses and sample sizes are small.

Comments on Trimmed Means

Fixing the proportion of observations to be trimmed, without looking at the data, avoids certain theoretical problems when one is testing hypotheses. In particular, expressions for standard errors are much easier to derive (see Wilcox, 2003, for details). Also, generalizations of heteroscedastic methods for means to trimmed means are available that improve control over the probability of a Type I error, reduce bias, and provide substantial increases in power in situations in which all methods based on means perform poorly. To maintain high power under normality, yet achieve high power when outliers are common, a good compromise between the mean and median is a 20% trimmed mean.

Figure 8 graphically illustrates one of the practical advantages of a 20% trimmed mean. Twenty observations were randomly sampled from a standard normal distribution, the mean and 20% trimmed mean were computed, and this process was repeated 5,000 times. The means are plotted as a solid line and the trimmed means as a dashed line in Figure 8. Under these optimal conditions, the sampling distribution of the sample mean has a smaller standard error than the sampling distribution of the 20% trimmed mean. However, the advantage of the mean over the 20% trimmed mean is not very striking. Now we repeat this computer experiment, only we sample from the con-

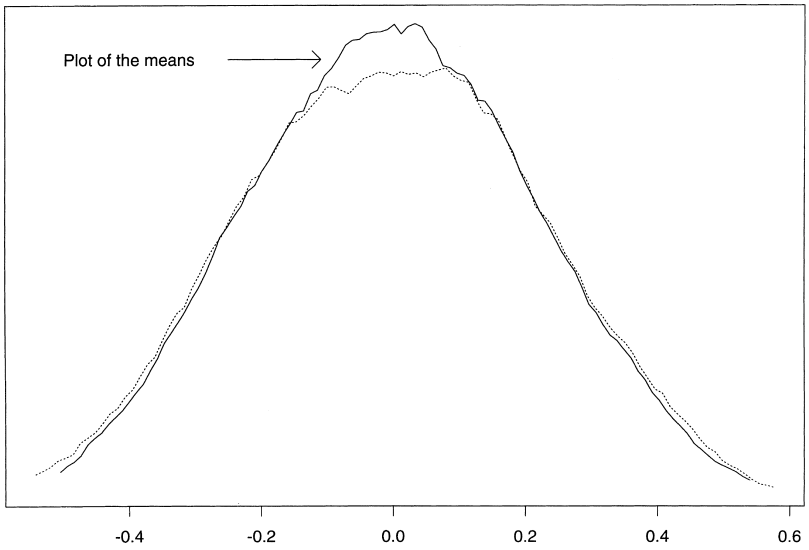

Figure 8. A plot of 5,000 means and 20% trimmed means (dashed line). Each mean and 20% trimmed mean is based on 20 observations randomly sampled from a normal distribution. Despite the fact that there was 20% trimming from both tails, the 20% trimmed mean has a standard error nearly as small as as the standard error of the mean.

taminated normal distribution shown in Figure 6. Figure 9 shows the results. Now the sampling distribution of the 20% trimmed mean is more tightly centered around zero, the value being estimated. That is, the standard error of the 20% trimmed mean is substantially smaller than the standard error of the mean, even though 8 of 20 observations are being trimmed in each sample.

One way of explaining the large standard error of the sample mean in Figure 9 is as follows. The standard error of the sample mean is $\sigma/\sqrt{n}$, where σ^2 is the variance of the distribution from which observations were randomly sampled. The contaminated normal distribution has a larger variance than the standard normal distribution. The basic problem is that σ^2 is extremely sensitive to the tails of any distribution. In contrast, the standard error of the sample trimmed mean is less affected by changes in the tails of a distribution. As a result, its standard error can be substantially smaller.

Because the fact that trimming data can result in a substantially lower standard error may be difficult to grasp, we provide an alternative perspective. Imagine that 20 observations are randomly sampled from a standard normal distribution, and consider the smallest value. The probability that the smallest value is within 1 standard deviation of the population mean is only .03. Similarly, the probability that the largest observation is within 1 standard deviation of the mean is .03. That is, there is a high probability that the smallest and largest values will be relatively far from the population mean.

Now suppose we put the observations in ascending order. Now the probability that the two middle values are within a half standard deviation of the population mean is .95. This would seem to suggest that the middle values should be given more weight when one is estimating μ, because they are more likely to be close to μ than the smallest and largest values. Moreover, it might seem that the median would be a more accurate estimate of μ, on average, than the sample mean. Yet when sampling from a normal distribution, we know that the sample mean is optimal. This result occurs because when we put observations in ascending order, they are no longer independent. That is, if $X_1, \ldots, X_n$ is a random sample, and if we put these values in ascending order yielding $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, it can be seen that for any $i < j$, $X_{(i)}$ and $X_{(j)}$ are dependent (e.g., Hogg & Craig, 1970; Wilcox, 2001) and have a nonzero correlation. Under normality and random sampling, these correlations are such that the sample mean provides a more accurate estimate of μ (based on mean square error) versus the median or any trimmed mean. However, when there are small

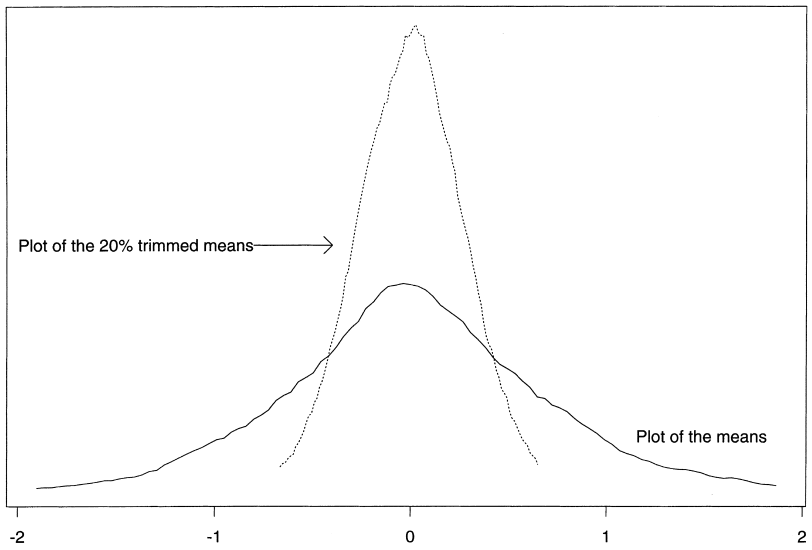

Figure 9. The observed distribution of 5,000 means and 20% trimmed means, illustrating that the standard error of a 20% trimmed mean can be substantially smaller than the standard error of the mean. Each mean and 20% trimmed mean is based on 20 observations randomly sampled from the contaminated normal distribution in Figure 6. The range of values is larger versus Figure 8 because when one is sampling from a contaminated normal distribution, there is more variability among the sample means.

departures from normality, the correlations no longer salvage the mean from the deleterious effects of the smallest and largest values among a batch of numbers. Extreme values increase the standard error of the mean in general, and trimming them reduces this problem. When working with means, problems with power were noted nearly a half a century ago by Bahadur and Savage (1956), and results published by Laplace (see Hald, 1998, p. 449) foreshadowed the difficulties we know today. Indeed, Laplace described general conditions under which the median has a smaller standard error than the mean.

How Much Trimming Should Be Used?

How many observations should be trimmed when using a trimmed mean? We do not want to trim too much in the event that sampling is from a normal distribution. However, we do not want to trim too little in case we are sampling from a heavy-tailed distribution such as the contaminated normal distribution in Figure 6. Of course, in practice, we will not know the distribution in the population. On the basis of the criterion of achieving a relatively small standard error, Rosenberger and Gasko (1983) recommended 20% in general, but for small sample sizes they suggested using 25% instead. More trimming might be beneficial in situations in which a large number of outliers frequently occur. In terms of Type I errors, theory and simulations suggest that problems due to nonnormality diminish as the amount of trimming increases (Wilcox, 1996). This provides an argument for using medians, but it ignores power. As a means of avoiding low power under normality and substantially improving on our ability to control the probability of a Type I error, again, 20% trimming is an excellent choice (Wilcox, 2003).

An Illustration

In theory, the sample mean can have a large standard error relative to some other estimators, but can it really make a difference which estimator is used when working with real data? Consider the data given in Wilcox (2001, p. 83). The general goal was to investigate issues related to self-awareness and self-evaluation and to understand the processes involved in reducing negative affect when people compare themselves with some standard of performance or correctness. One phase of the study consisted of comparing groups in terms of their ability to keep a portion of an apparatus in contact with a specified target. For the first group, the estimated standard error of the sample mean is 136 versus 56.10 for the 20% trimmed mean. For the second group the estimated standard errors for the mean and 20% trimmed mean were 157.90 and 69.40, respectively. Thus, using trimmed means can result in much higher power. The usual Student's T test has a significance level of .47. Yuen's (1974) test for trimmed means, which addresses both outliers and heteroscedasticity, has a significance level of .05. Yuen's test was calculated with the S-PLUS function `yuen in` (Wilcox, 1997a, 2003).

What Is Being Tested When Robust Estimators Are Used?

Robust estimators such as MOM, M estimators, and trimmed means are aimed at estimating the typical response given by participants. When sampling from a symmetrical distribution, their population values are identical to the population mean. For this special case only, trimmed means, M estimators, and MOM provide alternative methods for estimating and testing hypotheses about μ. However, when distributions are skewed, they estimate not μ but rather some value that is typically closer to the bulk of the observations.

To avoid misconceptions about trimmed means in particular, and robust estimators in general, it might help for us to elaborate somewhat on how a population trimmed mean is defined. (For a more technical description, see Staudte & Sheather, 1990.) Consider again the lognormal distribution in the left panel of Figure 1. The .10 and .90 quantiles of this distribution are 0.28 and 3.60, respectively. The population 10% trimmed mean is the average of all possible observations that fall between the .10 and .90 quantiles, namely 0.28 and 3.60. Similarly, the .20 and .80 quantiles are 0.43 and 2.32, respectively. The population 20% trimmed mean is the average for all possible participants, if they could be measured, provided their observed outcome is between 0.43 and 2.32. In contrast, the population mean is the average of all possible outcomes regardless of how small or large they might be. Put another way, a population trimmed mean is the population mean of a transformed distribution in which the tails of the distribution are eliminated.

A sample trimmed mean does not represent an attempt to estimate the population mean of the whole distribution. Rather, the goal is to estimate the population trimmed mean. When distributions are skewed, the population trimmed mean is typically closer to the bulk of the observations than the population mean.

For the lognormal distribution in Figure 1, the population mean is 1.65, and the population 20% trimmed mean is 1.11.

The left panel of Figure 10 shows the location of the population mean, median, and 20% trimmed mean for a lognormal distribution. When computing a 20% trimmed mean based on a sample from this distribution, we are attempting to estimate 1.11, the population trimmed mean shown in Figure 10. The right panel shows the population mean, median, and 20% trimmed mean for the skewed, heavy-tailed distribution in Figure 2. Now a sample 20% trimmed mean represents an attempt to estimate the value 3.90. Note that in both cases, the population mean lies relatively far from the most typical values. When one is comparing two groups, methods based on trimmed means are used for testing the hypothesis that the population trimmed means are identical.

Basic Bootstrap Methods

In terms of getting accurate confidence intervals and controlling the probability of a Type I error, extant investigations indicate that when one is comparing groups based on trimmed means or M estimators, some type of bootstrap technique has practical value. Roughly, as the amount of trimming decreases, the benefit of some type of bootstrap method increases. From a technical point of view, analytic solutions exist for both the 20% trimmed mean and one-step M estimator (e.g., Luh & Guo, 1999; Wilcox, 1993, 1997a), but the bootstrap offers a practical advantage in some situations, and it appears to be the best method for general use. Consequently, we quickly review two basic bootstrap methods here.

Bootstrap-t Method

We first illustrate the bootstrap t using sample means and then show how this method can be extended to 20% trimmed means. For a sample of observations, $X_1, \ldots, X_n$, a bootstrap sample is obtained by resampling with replacement n observations from $X_1, \ldots, X_n$, which we label $X_1^*, \ldots, X_n^*$. For example, if our original sample has observations 1, 2, 4, 5, 3, one possible bootstrap sample is 4, 4, 3, 2, 3. The basic idea behind the bootstrap-t method is to estimate the null distribution of some appropriate analog of Student's T. In the one sample case, for example, when the goal is to make inferences about the population mean, μ, we approximate the distribution of

$$T = \frac{M - \mu}{SD/\sqrt{n}} \tag{4}$$

as follows: (a) Generated a bootstrap sample; (b) computed the mean and standard deviation based on this bootstrap sample, which are labeled M^* and SD^*, respectively; and (c) compute T^* as follows:

$$T^* = \frac{M^* - M}{SD/\sqrt{n}}.$$

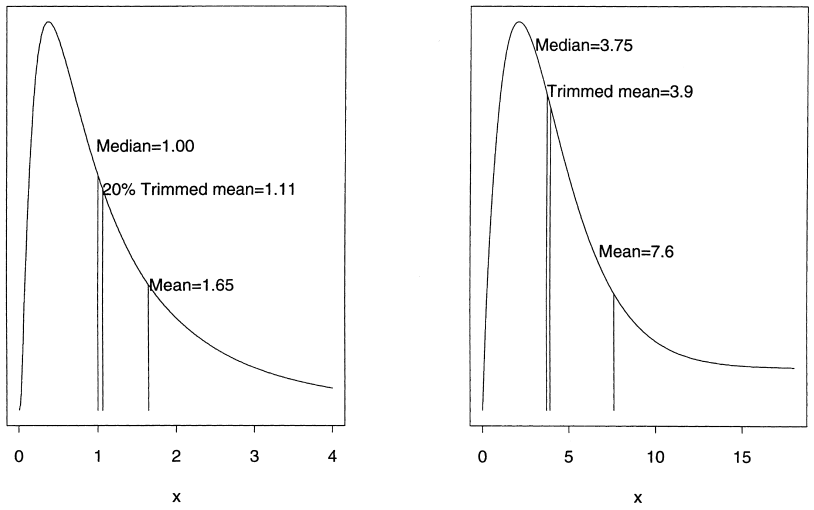

Figure 10. The left panel shows the location of the population mean, median, and 20% trimmed mean for a lognormal distribution. The right panel shows their location for the (contaminated chi-square) distribution in Figure 2.

An approximation of the distribution of T is obtained by repeating Steps (a)–(c) B times, yielding $T_1^*, \ldots, T_B^*$. That is, we can test hypotheses about μ if we know the distribution of T, but rather than assume normality to determine its distribution, we empirically determine its distribution based on the data available to us.

The bootstrap t is extended to trimmed means by computing T^* as just described but replacing M with the sample 20% trimmed mean, replacing M^* with the 20% trimmed mean based on a bootstrap sample, and replacing $SD/\sqrt{n}$ with an appropriate estimate of the standard error of a 20% trimmed mean (see Wilcox, 1997a, 2003).

A basic issue is how many bootstrap samples should be used. One approach is to determine how large B must be to get reasonably good control over the probability of a Type I error. With 20% trimmed means, $B = 600$ seems to be sufficient (cf. Hall, 1986). Booth and Sarkar (1998) argued that B should be large enough so that if a different collection of B bootstrap samples were used, this would have at most a very small effect on the estimated significance level. They derived an approximation of how large B must be so that the variance associated with a bootstrap sampling distribution is approximately equal to the usual squared standard error. For their example, they concluded that $B = 800$ should be used.

Percentile Bootstrap Method

Another basic strategy is the percentile bootstrap, which, in contrast to the bootstrap t method, does not use estimates of standard errors. To illustrate this strategy, consider again testing some hypothesis about a single population mean (μ). The percentile bootstrap uses the central $1 - \alpha$ portion of the bootstrap sample means as a $1 - \alpha$ confidence interval for μ. For example, if, among 800 bootstrap sample means, the central 95% have values between 1.40 and 3.90, then the .95 confidence interval for μ is (1.40, 3.90). The theoretical motivation for this method arises as follows. Suppose the goal is to test H_0: $\mu = \mu_0$, where μ_0 is a specified value (e.g., IQ = 100). For a random sample of observations, let p^* be the probability that a bootstrap sample mean exceeds μ_0. If the null hypothesis is true and the sample size is sufficiently large, then p^* will have, approximately, a uniform distribution (e.g., Hall, 1988; Liu & Singh, 1997). Thus, to test H_0, compute the proportion of bootstrap sample means greater than μ_0 and label it $\hat{p}^*$. Let

$$\hat{p}_m^* = \min(\hat{p}^*, 1 - \hat{p}^*).$$

Then, for a two-tailed test, $2\hat{p}_m^*$ is the estimated significance level, so reject H_0 if $2\hat{p}_m^* \leq \alpha$.

The percentile bootstrap method is readily extended to two groups and any measure of location, θ. Let θ_1 and θ_2 be the value of θ (e.g., population trimmed means) for Groups 1 and 2, respectively, and consider testing

$$H_0: \theta_1 = \theta_2.$$

A set of B bootstrap estimates are computed for each group, where the two groups are denoted by $g = 1$ or $g = 2$. Let $\hat{\theta}_{gb}^*$ be the bth bootstrap estimate of θ for the gth group ($b = 1, \ldots, B$). Let $\hat{p}^*$ be the proportion of times $\hat{\theta}_{1b}^*$ is less than $\hat{\theta}_{2b}^*$ among the B bootstrap estimates. (Otherwise stated, if there are A instances where $\hat{\theta}_{1b}^* < \hat{\theta}_{2b}^*$, and there are B bootstrap samples, then $\hat{p}^* = A/B$.) Set

$$\hat{p}_m^* = \min(\hat{p}^*, 1 - \hat{p}^*).$$

Then $2\hat{p}_m^*$ is the estimated significance level and, as in the one-sample case, H_0 is rejected when $2\hat{p}_m^* \leq \alpha$. The computations are performed with the R or S-PLUS function pb2gen in Wilcox (2003) and can be used with any measure of location. The function trimpb2 is designed specifically for trimmed means. Extensions for comparing multiple groups, including dependent groups, have been derived, but no details are given here.

When comparing means (no trimming), it is well-known that the percentile bootstrap method performs rather poorly relative to the bootstrap t (e.g., Westfall & Young, 1993). However, when one is using measures of location that are relatively insensitive to outliers, it currently seems that some version of the percentile bootstrap method generally has a practical advantage. This suggests using some type of percentile bootstrap method when comparing groups based on MOM or trimmed means. Findings in Wilcox (2002b) and Keselman, Wilcox, Othman, and Fradette (2002) support this approach so far.

ANOVA

Complete details about recent ANOVA methods are beyond the scope of this article, but some comments might be informative. We note that all of the more common designs can be handled with the techniques and software in Wilcox (2003) when means are replaced by robust measures of location, and SAS software is available as noted in the Appendix. When

performing an omnibus test or dealing with multiple comparisons, based on estimators that empirically check for outliers and discard any that are found (such as M estimators), special modifications and variations of percentile bootstrap methods are the only techniques to date that have performed well in simulations.

Imagine the goal is to test

$$H_0: \theta_1 = \cdots = \theta_G,$$

the hypothesis that the typical responses among G independent groups are identical. Note that if H_0 is true, then, by implication, all pairwise differences are zero. That is, for any two groups j and k, $j < k$, H_0 implies that $\theta_j - \theta_k = 0$. We define the difference between the typical responses in groups j and k as $\delta_{jk} = \theta_j - \theta_k$. Let $\hat{\delta} = \hat{\theta}_j - \hat{\theta}_k$ be an estimate of δ_{jk} based on data. Note that there are $L = (G^2 - G)/2$ pairwise differences. If the null hypothesis is true, then the vector corresponding to all pairwise differences should not be too far from the vector $(0, \ldots, 0)$, which has L elements. Here we merely note that a bootstrap method for implementing this approach is available. The theoretical justification for this method stems from work by Liu and Singh (1997), but certain modifications are required to control the probability of a Type I error when sample sizes are small. Keselman et al. (2002) examined this approach and Wilcox and Keselman (2003) extended this procedure to provide a test based on MOM with dependent groups.

As for trimmed means, all indications are that when sample sizes are small, bootstrap methods are generally the best choice. Even with large sample sizes and a large number of bootstrap replications (B), typically analyses can be performed quickly on modern computers, often requiring less than a minute.

Choosing a Method and a Measure of Central Tendency

On the surface, the MOM estimator might seem more appealing versus a one-step M estimator, or trimmed means in general, and the 20% trimmed mean in particular. MOM is flexible in terms of how many observations are discarded as outliers, it reduces to using the usual mean when no outliers are found, it can handle a relatively large number of outliers, and it allows different amounts of trimming in the left tail of a distribution versus the right. And situations do arise in which it has a small standard error compared with other estimators that might be used. However, from a broader perspective, choosing an estimator is a more complicated issue.

First, consider the goal of choosing an estimator based solely on the criterion that it has a relatively small standard error. From a theoretical point of view, the one-step M estimator has excellent properties. It was designed to compete well with the mean when sampling from a normal or light-tailed distribution, and it competes well with the median when instead sampling is from a heavy-tailed distribution. Like MOM, it contains the possibility of no trimming, and for various situations it offers a bit of an advantage over MOM. An argument for using MOM or the one-step M estimator is that they outperform a 20% trimmed mean in situations in which outliers are very common. That is, if more than 20% of the largest values are outliers, a 20% trimmed mean might have a relatively large standard error compared with a one-step M estimator or MOM.

However, consider the more general goal of choosing a single measure of location for routine use when testing hypotheses based on multiple criteria. If simultaneously we want to achieve high power, accurate probability coverage (i.e., confidence intervals are accurate), relatively low standard errors, a negligible amount of bias, and good control over the probability of a Type I error, it currently seems that a percentile bootstrap method with a 20% trimmed mean is a good choice. Another approach, not described here, consists of heteroscedastic methods based on 20% trimmed means that are not based on some bootstrap technique (see Wu, 2002, for results when using actual data). These methods are based on extensions of heteroscedastic techniques for means, such as the methods in Welch (1938, 1951) or Johansen (1980; see Keselman et al., 2002).

We hasten to add that we suggest flexibility over rigid adherence to one approach or even one measure of location. Different methods (including rank-based techniques and the shift function) are sensitive to different features of the data. Depending on the specific research question, these methods might add perspectives that help us develop a more informed understanding of how the groups differ and by how much. In some situations, multiple methods might be needed to get a sufficiently detailed understanding of how groups differ. Moreover, the reality is that situations can arise in which something other than a 20% trimmed mean is preferable. There is, of course, the issue of controlling the probability of a Type I error when multiple methods are applied. If many methods

are used to compare groups, and there is some indication that groups differ in a particular manner, it might be necessary to repeat a study with the goal of determining whether the apparent difference can be substantiated. As is evident, repeating a study can be time-consuming and expensive, but this seems preferable to routinely missing an important difference because of the statistical method used.

One potential approach to the choice of method would be to identify diagnostic criteria that can inform the choice of a measure of location, or some hypothesis-testing method, prior to testing any hypotheses. That is, some type of preliminary analysis is performed with the goal of deciding which method should be used to compare groups. For example, one could estimate skewness and kurtosis in an attempt to justify using means over trimmed means or MOM. However, the estimates of these quantities tend to have much larger standard errors than the mean and variance, they are highly sensitive to slight changes in a distribution, and they are particularly affected by outliers. Consequently, the expectation is that they perform poorly as a diagnostic tool, and currently it seems that this approach does not have practical merit. We have considered many other diagnostic strategies, all of which have proven to be rather unsatisfactory.

Conclusion

Many issues and techniques have not been described, but hopefully we have conveyed the idea that we now have a vast array of tools for substantially increasing power in a wide range of commonly occurring situations. Moreover, modern methods help improve our understanding of how groups compare, as well as how much they differ, and the precision of our estimates can be assessed much more accurately than ever before.

Modern methods are not designed to test hypotheses about means except in situations in which perfectly symmetrical distributions are being compared. We outlined how population trimmed means are defined, a more technical discussion of this issue can be found in Huber (1981) or Staudte and Sheather (1990), and formal definitions of population values for MOM and one-step M estimators are available as well. This is one reason why different methods can lead to different results.

We close with seven summary observations that we believe are of value to researchers.

1. We encourage researchers to be cautious when interpreting nonsignificant results in cases in which group differences are expected. Failing to reject the null hypothesis might be because the null hypothesis is true, but it may occur because of poor experimental design or lack of statistical power. Traditional statistical methods may have low statistical power because, in many cases, they are relatively insensitive to group differences.

2. Conventional methods generally offer at most a small advantage in statistical power over modern methods when standard assumptions are approximately true. This is because modern methods are designed to perform nearly as well under these circumstances.

3. As we move toward situations in which groups differ in terms of variances, or skewness, or sampling is from heavy-tailed distributions with many outliers, at some point conventional methods for means break down whereas modern methods continue to perform well in terms of probability coverage and power. The only known way of determining whether the choice between modern and conventional methods makes a difference is to try both.

4. If heteroscedastic methods for means are used, problems with low power, poor control over the probability of a Type I error, and bias will become negligible if the sample size is sufficiently large. However, it is unknown how large the sample size must be. When dealing with skewed distributions, regardless of how large the sample size happens to be, a criticism is that the mean can lie in the tails of a distribution and be rather atypical (Staudte & Sheather, 1990).

5. Even if a plot of the data suggests that it has a nearly symmetrical distribution, low power can result when one is using tests of means due to outliers or heteroscedasticity.

6. The 20% trimmed mean is likely to perform well relative to other choices. However, there are cases in which exceptions will occur (see Keselman et al., 2002). For example, a one-step M estimator or MOM can handle situations in which the proportion of outliers in one of the tails of the data happens to exceed 20%.

7. At some level, no single method can be expected to compete well in all possible situations that might be encountered simply because different methods are sensitive to different features of the data. However, modern methods have much to offer because they perform well in a larger set of situations than conventional techniques.

References

Bahadur, R., & Savage, L. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Statistics, 27,* 1115–1122.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Booth, J. G., & Sarkar, S. (1998). Monte Carlo approximation of bootstrap variances. *American Statistician, 52,* 354–357.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way model. *Annals of Mathematical Statistics, 25,* 290–302.

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments.* New York: Wiley.

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis, 33,* 249–258.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Erlbaum.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics.* New York: Wiley.

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal, 28,* 131–148.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2,* 292–307.

Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika, 63,* 421–434.

Doksum, K. A., & Wong, C.-W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association, 78,* 411–417.

Gleason, J. R. (1993). Understanding elongation: The scale contaminated normal family. *Journal of the American Statistical Association, 88,* 327–337.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155–165.

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930.* New York: Wiley.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics, 14,* 1431–1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics, 16,* 927–953.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data tables, trends, and shapes.* New York: Wiley.

Hogg, R. V., & Craig, A. T. (1970). *Introduction to mathematical statistics.* New York: Macmillan.

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

Johansen, S. (1980). The Welch–James approximation of the distribution of the residual sum of squares in weighted linear regression. *Biometrika, 67,* 85–92.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (in press). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology.*

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods, 1,* 288–309.

Keyes, T. K., & Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics, 22,* 227–236.

Liu, R. Y., & Singh, K. (1997). Notions of limiting *P* values based on data depth and bootstrap. *Journal of the American Statistical Association, 92,* 266–277.

Luh, W. M., & Guo, J. H. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variance. *British Journal of Mathematical and Statistical Psychology, 52,* 303–320.

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *American Statistician, 44,* 322–326.

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample *t*-test versus Satterthwaite's approximate *F* test. *Communications in Statistics—Theory and Methods, 18,* 3963–3975.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's *t* test with unequal variances. *Journal of Educational Statistics, 5,* 337–349.

Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology, 42,* 203–211.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297–336). New York: Wiley.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from normality. *Psychological Bulletin, 111,* 353–360.

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330–336.

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

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 61,* 165–170.

Wilcox, R. R. (1990). Comparing means and variances when distributions have non-identical shapes. *Communications in Statistics—Simulation and Computation, 19,* 945–971.

Wilcox, R. R. (1993). 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). A note on testing hypotheses about trimmed means. *Biometrical Journal, 38,* 173–180.

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

Wilcox, R. R. (1997b). Some practical reasons for reconsidering the Kolmogorov–Smirnov test. *British Journal of Mathematical and Statistical Psychology, 50,* 9–20.

Wilcox, R. R. (1998a). The goals and strategies of robust methods (with discussion). *British Journal of Mathematical and Statistical Psychology, 51,* 1–39.

Wilcox, R. R. (1998b). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York: Springer.

Wilcox, R. R. (2002a). Comparing the variances of independent groups. *British Journal of Mathematical and Statistical Psychology, 55,* 169–176.

Wilcox, R. R. (2002b). Multiple comparisons among dependent groups based on a modified one-step *M*-estimator. *Biometrical Journal, 44,* 466–477.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* San Diego, CA. Academic Press.

Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA *F, W,* and *F** statistics. *Communications in Statistics—Simulation and Computation, 15,* 933–944.

Wilcox, R. R., Keselman, H. J. (2003). Repeated measures one-way ANOVA based on a modified one-step *M*-estimator. *British Journal of Mathematical and Statistical Psychology, 56,* 15–26.

Wu, P. C. (2002). *Central limit theorem and comparing means, trimmed means one-step M-estimators and modified one-step M-estimators under non-normality.* Unpublished doctoral disseration, University of Southern California, Los Angeles.

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

Appendix

Analysis of Variance Software

Three software packages can be used to apply modern analysis of variance (ANOVA) methods: R, S-PLUS, and SAS. The software R and S-PLUS are used in nearly the same manner, but R is a freeware variant of S-PLUS. Information about S-PLUS is available at http://www.insightful.com. The R freeware can be downloaded from http://www.R-project.org.

When one is using S-PLUS, the methods outlined in this

article can be applied by downloading the files allfunvl and allfun2 from www-rcf.usc.edu/~rwilcox/ using the `Save As` command. Complete details about these functions can be found in Wilcox (2003). On some systems, when the file allfunv1, for example, is downloaded, it is stored in a file called allfunv1.txt. On other systems, the file might be stored in allfunv1.html. Store these files in the directory being used by S-PLUS. The directory being used should be indicated at the top of the screen when S-PLUS is running on a PC. On some machines this directory is

Programs Files\sp2000\users\default.

Next, run the `source` command on both files. Thus, for the first file, use

```
source("allfunvl").
```

Then all of the S-PLUS functions in this file are added to your version of S-PLUS until they are removed. Descriptions of how to use these functions can be found in Wilcox (2003).

The free software R comes with a manual that can be accessed by clicking on help once R is running. To apply the modern methods described here and in Wilcox (2003), download the files Rallfunv1 and Rallfunv2 from http://www-rcf.usc.edu/~rwilcox/ and store them in the directory being used by R. On Rand R. Wilcox's PC, this directory is Programs Files/R/rw1041. To incorporate the functions in these files into R, again use the `source` command. Thus, the command `source("Rallfunv1")` adds all of the functions in the file Rallfunv1, and they remain in R until they are removed with the `rm` command. S-PLUS and R commands are nearly identical, but because of slight differences, separate files or the R functions were created. Nearly all of the S-PLUS functions in Wilcox (2003) are used in exactly the same manner as when using R. The only known exceptions occur when one is creating certain three-dimensional plots or when using certain multivariate methods. In these cases, R will produce an error saying that some function (such as cov.mve or interp) does not exist. To correct this, click on packages, and then click on lqs as well as akima. Then use the commands

```
library(lqs)
library(akima).
```

Both R and S-PLUS are relatively easy to use. For example, if the data for four groups are stored in the variable dat (in either list mode or in a matrix), the command `lincon(dat)` will perform all pairwise comparisons among the groups based on 20% trimmed means and control the familywise error rate. The command `lincon(dat, tr=0)` performs comparisons of means instead, so the optional argument `tr` indicates how much trimming is to be used and defaults to 20% if not specified.

SAS/IML (Version 8) software for both completely randomized and correlated groups design using both conventional and robust estimators is available from H. J. Keselman at http://www.umanitoba.ca/faculties/arts/psychology/. For detailed illustrations on how to use this SAS software, see Keselman, Wilcox, and Lix (in press). Both omnibus hypotheses and linear contrasts can be tested, the amount of trimming can be adjusted, and bootstrap techniques are available as well. (Currently, some of the R and S-PLUS functions can perform analyses that have not yet been written in SAS.)

Received October 31, 2001
Revision received February 12, 2003
Accepted March 16, 2003 ■