Review of Educational Research
Spring 1995, Vol. 65, No. 1, pp. 51–77

ANOVA: A Paradigm for Low Power and Misleading Measures of Effect Size?

Rand R. Wilcox
University of Southern California

Over 30 years ago, Tukey made it evident that slight departures from normality can substantially lower power when means are compared, and that a popular measure of effect size can be highly misleading. At the time there were no methods for dealing with the problem raised in Tukey's paper, and some of the more obvious and seemingly intuitive solutions have since been found to be highly unsatisfactory. Today there are practical methods for not only dealing with the problem raised by Tukey, but also achieving more accurate confidence intervals and control over the probability of a Type I error. More generally, there are many robust and exploratory ways of comparing groups that can reveal important differences that are missed by conventional methods based on means, and even modern methods based solely on robust measures of location. This article reviews these new techniques.

Suppose two methods of teaching high school algebra are being investigated. One group learns algebra using Method A, and an independent group uses Method B. Once data are collected, how should these two groups be compared? Of course, the most common approach in education and psychology is to compare means, though some would use the Mann-Whitney-Wilcoxon test instead. Today there are many alternative solutions that can reveal important and interesting differences that are missed by these more conventional techniques. Some of these newer methods are based on replacing the mean with some other measure of location. The term *measure of location* is formally defined below, but for now it suffices to think of it as a measure intended to represent the typical individual under study. The mean and median are the two best-known examples of measure of location, but other important measures of location are available which are reviewed in this article. Much of this article deals with comparing measures of location, but it is stressed that when attention is restricted to comparing measures of location, interesting differences between groups can be missed. One of the main goals in this introductory section is to describe methods that can supplement techniques based solely on measures of location and why they can be useful. These methods are generally called robust and exploratory techniques, and are described in more detail in the two books by Hoaglin, Mosteller, and Tukey (1983, 1985), which are written at a relatively nontechnical level. Here the goal is to supplement these two books by describing some recent graphical tools and related techniques for comparing groups. A related goal is to point out that popular measures of effect size can be highly misleading for reasons that are explained below.

Before continuing, one point should be stressed. While this article is critical of standard methods for comparing means and related measures of effect size, it is not being implied that all published research is inaccurate. On the positive side, when a researcher reports a significant result when testing some hypothesis, a truly significant result has probably been found. The concern is that many nonsignificant results might have been highly significant if some alternative method had been used, and there is the concern that significance levels and reported probability coverage of confidence intervals might be highly inaccurate. A more general and perhaps more important issue is that flexibility and alternative ways of viewing data can be very important.

To begin to appreciate why global comparisons of distributions can be useful, as opposed to comparing measures of location, first consider the problem of comparing two independent and normally distributed random variables. More realistic situations are considered below. To be concrete, suppose Y corresponds to an experimental method for teaching algebra having mean μ_1, and X represents a control or standard method with mean μ_2. Suppose $\mu_1 = \mu_2 = 9$, but Y has variance $\sigma^2 = .2$, while X has variance $\sigma^2 = 1$. Assume that high Y values indicate that the experimental method is more effective than the control. Of course, any method for comparing means should not reject H_0: $\mu_1 = \mu_2$, and for the situation at hand, none of the methods considered in this article for comparing alternative measures of location should reject, either. Yet there is a potentially important difference: Students who do poorly under the standard method benefit from the experimental method, but the experimental method is detrimental for students who do well using the standard technique.

One way of dealing with this problem is to compare the quantiles of the two groups. For the situation at hand, one might also compare variances, but comparing quantiles can be more revealing for reasons that will become evident. Let x_p be the pth quantile corresponding to the random variable X. Thus, $x_{.5}$ is the median. Consider graphing x_p versus $\delta(x_p) = y_p - x_p$, the difference between the quantiles. If $\delta(x_p) > 0$, students who are at the pth quantile of the standard method would typically perform better under the experimental method; and if $\delta(x_p) < 0$, the experimental method is detrimental. Figure 1 shows a graph of δ versus x_p for the two normal distributions in the illustration for the deciles, $p = .1, .2, .3, .4, .5, .6, .7, .8$ and $.9$.

The quantity $\delta(x_p)$ is known as a shift function and has been studied by Doksum (1974, 1977) as well as Doksum and Sievers (1976). One concern is that groups might differ in more complicated ways than is the case in the illustration just given, and Doksum (1977) reanalyzes some data on the effect of radiation on mice to demonstrate this point. Here, data analyzed by Salk (1973) is used to provide yet another illustration. A portion of Salk's study dealt with weight gain in newborns weighing at least 3,500 grams at birth. These newborns were separated into two groups. The first group was continuously exposed to the recorded sound of a human heartbeat, while the other group acted as a control. The equality of the means of the two groups was tested with Student's t test and rejected at the .05 level of significance. Figure 2 shows an estimate of the shift function where the quantiles are estimated using the method derived by Harrell and Davis (1982). The Minitab macro dshift.mtb in Wilcox (in press-c) performs the necessary computations. The point here is that the graph descends, levels off,

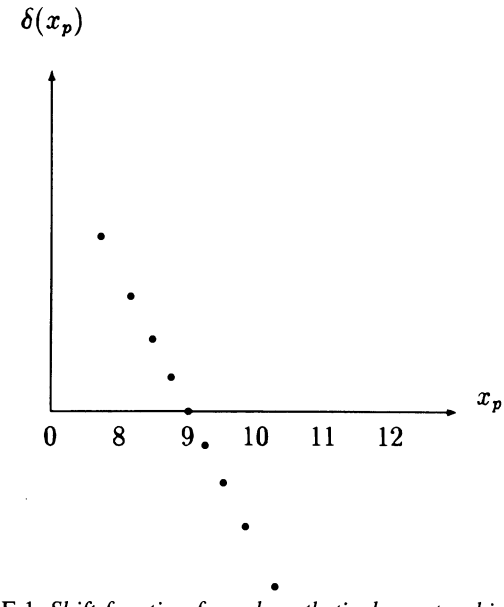

FIGURE 1. *Shift function for a hypothetical new teaching method*

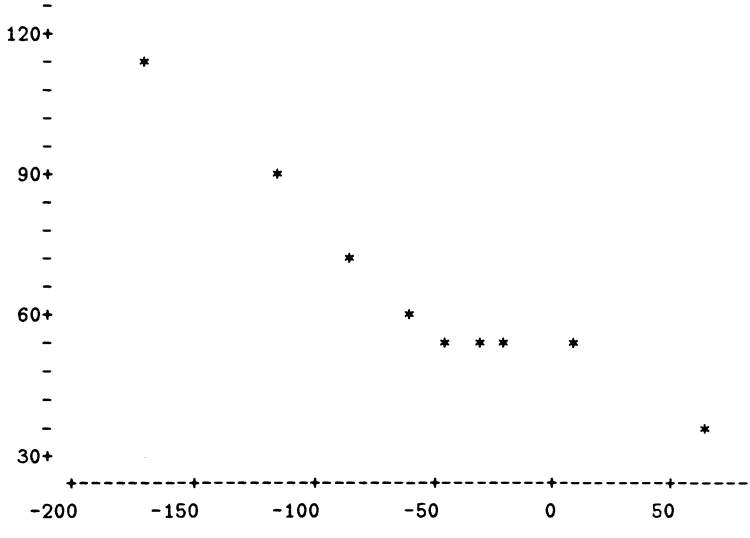

FIGURE 2. *Estimated shift function for data on weight gain in newborns*
Note. Data are from Salk (1973).

and then descends again. This indicates that the largest effect occurs for babies who have a relatively large loss of weight without the treatment. Clearly $\delta(x_p)$ can be important because it provides a more refined method of determining who might benefit from a particular treatment. Of particular importance is that $\delta(x_p)$ can reflect differences not indicated by $\mu_1 - \mu_2$ or by the difference in any other measure of location. A method for obtaining simultaneous confidence intervals for $x_p - y_p$ when working with deciles is outlined in Wilcox (in press-c); the necessary computations are done with his Minitab macro mq.mtb, and details of how this method performs can be found in Wilcox (in press-b).

Several other methods have been proposed which also capture the spirit of the shift function. The best known is the Kolmogorov-Smirnov test, but it has relatively low power, and its power can increase slowly as the sample sizes get large (Randles & Wolfe, 1979, p. 384). Doksum and Sievers (1976) derived a confidence band for all quantiles using a Kolmogorov-Smirnov statistic, and a related method was derived by Switzer (1976). Earlier related work is the quantile-quantile plot of Gnanadesikan and Wilk (1968), and a method based on ranks was proposed by O'Brien (1988). O'Brien provides additional illustrations of why these techniques can be important. Yet another graphical approach is to draw a boxplot of two groups, one on top of the other and on the same scale. An illustration is given in Wilcox (in press-c, chapter 8). For results on the properties of the boxplot, see Brant (1990). For graphical methods aimed at comparing multiple groups, see Tukey (1993). Finally, global measures of how groups differ can be obtained by estimating the distributions corresponding to each group using methods such as those in Silverman (1986).

Some Practical Reasons for Using Robust Measures of Location and Effect Size

Now consider the problem of comparing groups based on some measure of location. There are two related concerns about power when sample means are used: outliers and heavy-tailed distributions. To illustrate the first, consider the data in Wilcox (1992d) which deals with a study on self-awareness. The sample means are 448 and 599. Using Welch's test of H_0: $\mu_1 > \mu_2$, the significance level is .24. However, comparing the groups in terms of 20% trimmed means or one-step M-estimators, which are robust measures of location described below, yields a significant difference at the .05 level. Moreover, various papers described below indicate that these tests generally provide better control over Type I error probabilities, especially when distributions are skewed. The reason for the discrepant results is that the first group has two extreme outliers (unusually large values relative to the bulk of the available data) that are revealed by a boxplot, and the second group has an extreme outlier, as well. Outliers inflate the standard error of the sample mean, which, in turn, lowers power. Many robust estimators of location have standard errors that are relatively insensitive to outliers, and in the illustration this is why it was possible to reject. However, even when there are no outliers, different measures of location can give different results for reasons explained below.

Ever since Tukey (1960) published his work on the contaminated normal distribution, it has been clear that slight departures from normality can have devastating effects on power when means are compared, and that popular mea-

sures of effect size can be misleading, as well. The contaminated normal is an example of a distribution with a heavy or thick tail relative to the normal distribution. Tukey's paper had no immediate impact on applied research because it was unclear at the time how a researcher might address the problem he discussed. During the ten years following Tukey's paper, a few mathematical statisticians began laying a foundation for getting better results. Of particular interest is the theory of robustness developed by Huber (1964) and Hampel (1968). By the year 1974, there were some practical solutions for the one- and two-sample cases, but general techniques for more complicated designs were not available even a few years ago. Today, any of the common experimental designs can be improved.

Figure 3 shows a standard normal and a contaminated normal distribution. A contaminated normal is formed by sampling from a standard normal distribution (having mean 0 and variance 1) with probability $1 - \epsilon$; alternatively, sampling is from a normal distribution having mean 0 and standard deviation K. (Readers interested in further technical details can refer to Hoaglin et al., 1983). In Figure 1, $\epsilon = .1$ and $K = 10$. There is an obvious similarity between the two distributions. The tails of the contaminated normal actually lie above the tails of the normal, but this is difficult to discern without extending the range of x values and drawing the figure on a much larger scale. The main point is that while the standard normal has variance 1, the contaminated normal has variance $1 - \epsilon + \epsilon K^2 = 10.9$. Put another way, if sampling is from the contaminated normal, even when the variance is known, the length of the confidence interval for the population mean μ, will be over 3 times longer than what it is when sampling from the normal distribution instead, and this has obvious implications for power. The problem is that σ, the standard deviation, is highly sensitive to the tails of a distribution, so seemingly small departures from normality can inflate the standard error of the sample mean, $\sigma/\sqrt{n}$, where n is the sample size. One might try to salvage the situation by arguing that the contaminated normal distribution is actually a large departure from normality. The extreme quantiles differ substantially, which results in large differences between the standard deviations. But in terms of common metrics for comparing distributions (the Kolmogorov, Lévy, and Prohorov, the latter two being discussed by Huber, 1981), there is little difference between the normal and contaminated normal. Gleason (1993) compares them in terms of what he calls elongation and concludes that there is little difference.

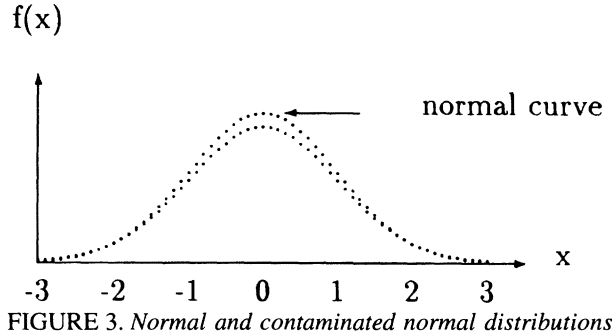

FIGURE 3. *Normal and contaminated normal distributions*

To add some perspective, look at Figure 4, which shows two normal distributions with means 0 and 0.8, both having variance 1. Consider the common measure of effect size

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

where μ_j is the mean of the jth group, and the groups have a common standard deviation $\sigma_1 = \sigma_2 = \sigma$. Cohen (1977, p. 26) defines a medium measure of effect size as one large enough to be visible to the naked eye, and for normal distributions he concludes that Δ values equal to .2, .5, and .8, correspond to small, medium, and large effect sizes. Thus, Figure 4 has $\Delta = .8$, which is large. Now look at Figure 5, which shows two contaminated normals, again with means 0 and 0.8. Then $\Delta = .24$, which is supposedly small, yet by appearances the effect size is large.

It has already been illustrated that outliers can substantially reduce power when one is working with sample means, and in a similar manner, heavy-tailed distributions inflate the standard error of the sample mean. As a result, power might be low relative to methods based on other measures of location. For example, suppose sampling is from two normal distributions with $\mu_1 - \mu_2 = 1$ and variances equal to 1. Then with $\alpha = .05$, and $n_1 = n_2 = 25$, Student's t test has power .957, while Welch's (1951) test has power approximately equal to .93. If, instead, sampling is from contaminated normal distributions, again with

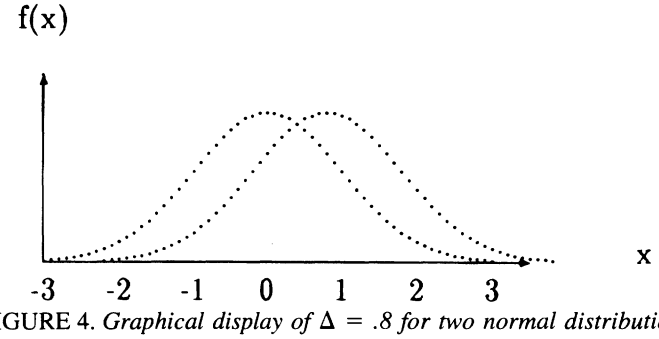

FIGURE 4. *Graphical display of* $\Delta = .8$ *for two normal distributions*

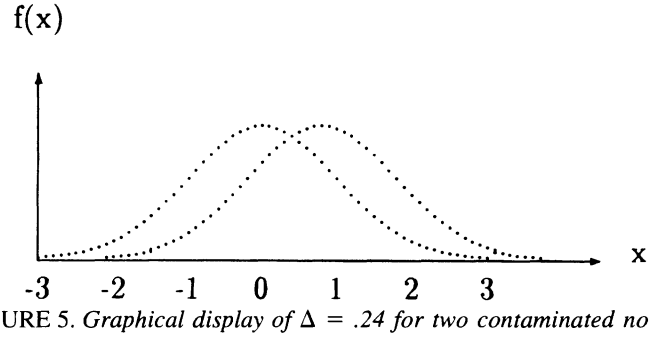

FIGURE 5. *Graphical display of* $\Delta = .24$ *for two contaminated normals*

$\mu_1 - \mu_2 = 1$, then these two procedures have power approximately equal to .28, based on simulations with 10,000 replications, and this agrees with the usual approximation for power derived under the assumption of normality. However, powers for three other procedures compared by Wilcox (1992d) range between .78 and .80, with little loss in power when distributions are normal. One of these is Yuen's (1974) method for trimmed means, based on 20% trimming, while the other two are recently developed methods for the median and a measure of location called a one-step *M*-estimator. (More details are given later in the article. Computational details, plus Minitab macros for applying these techniques, can be found in Wilcox, in press-c). Thus, to the extent that one wants to reject when groups differ, these newer procedures are of interest. There are a variety of other issues that one might consider before choosing a procedure; these are discussed below. Note that when dealing with power under nonnormality, Glass, Peckham, and Sanders (1972) focus exclusively on power as a function of Δ. Hence, the results just reported are consistent with their paper, but the point here is that other procedures can have substantially more power.

To describe yet another problem related to outliers, let $\overline{X}_1$ and $\overline{X}_2$ be the sample means corresponding to two groups, and suppose $\overline{X}_2 > \overline{X}_1$, and that H_0: $\mu_1 = \mu_2$ is rejected with $\alpha = .05$ using either Welch's method or Student's t test. Now suppose the largest observation in the second group is made even larger, so $\overline{X}_2 - \overline{X}_1$ increases as well. It would seem that there is stronger evidence for concluding that $\mu_2 > \mu_1$, yet eventually both tests for means will not reject because of the corresponding increase in the standard error. Illustrations are given in Staudte and Sheather (1990) as well as Wilcox (in press-c).

It should be stressed that outliers are not the only reason for considering robust methods. Small shifts in a distribution can have a large impact on the mean which might render it a potentially misleading measure of the typical individual. Consider, for example, a chi-square distribution with 4 degrees of freedom which has mean 4. If the distribution is contaminated by multiplying an observation by 10 with probability .1, it does not change very much based on typical metrics for comparing distributions. However, the mean is affected considerably as depicted in Figure 6, which also shows the location of the median, $\theta = 3.75$, and the 20% trimmed mean, $\mu_t = 4.2$, which is discussed in more detail in

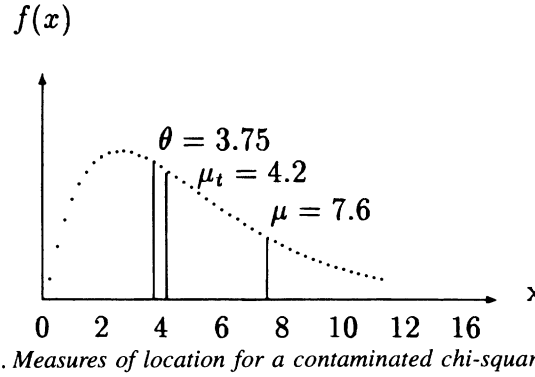

FIGURE 6. *Measures of location for a contaminated chi-square distribution*

subsequent sections. The mean is increased to 7.6, and its value is so far in the right tail, one might argue that it is not a very good reflection of the typical person.

During the early years of modern mathematical statistics, there were reasons to suspect that strict adherence to means, when describing and comparing groups, might be a problem (Pearson, 1931; Student, 1927). These results were overshadowed by theoretical results, summarized by Tan (1982), indicating that when two or more groups have identical distributions, conventional hypothesis testing procedures provide good control over the probability of a Type I error when distributions are nonnormal. In terms of power, however, there are several serious problems. Tukey's (1960) results imply that departures from normality that are difficult to detect can inflate the variance by a large amount, and this implies that power will be lowered. Moreover, Tukey argued that heavy-tailed distributions and outliers are common in applied research. If this is true, low power due to nonnormality could be common. Surveys of data occurring in applied work indicate that severe departures from normality can arise in practice (Micceri, 1989; Wilcox, 1990a).

The remainder of this article provides a brief and relatively nontechnical introduction to robust measures of location and scale, and how they can be used to compare groups. Some of these issues are summarized by Wilcox (1992d, 1993d). This article takes a much broader view of the problems that arise and the solutions that have been proposed. Elementary descriptions of many of the recently developed robust methods examined here are given in the textbook by Wilcox (in press-c). With Wilcox's book comes a floppy containing over 150 Minitab macros for applying techniques not found in standard statistical packages. Consequently, numerical descriptions of these methods are kept to a minimum. Here, attention is focused on general technical issues and practical details that are relevant to experienced researchers who want a relatively nontechnical understanding of how and why modern robust methods are relevant to their research.

Building a Mathematical Foundation for More Robust Methods

Applied researchers who do not religiously follow developments in mathematical statistics might still have the impression that robust methods are based on ad hoc procedures. The purpose of this section is to indicate that this is not at all the case. While technical details are kept to a minimum, some mathematical comments are needed in order to convey how the foundations of modern robust procedures were developed.

There are two fundamental and related problems that must be addressed. The first is finding analogs of the *population* mean and variance, μ and σ^2, that are relatively insensitive to slight changes in a distribution. When dealing with effect size, for example, it is desirable to have a measure that is consistent with Δ when distributions are normal, but which does not give a distorted view of the magnitude of the effect size under slight departures from normality. The second problem is finding estimators of location and scale with standard errors that are also relatively insensitive to slight changes in a distribution. In particular, small departures from normality should not drastically color our perceptions about how groups differ, nor should they substantially lower power. Today there is a well established mathematical foundation for addressing these problems which is summarized in the books by Huber (1981); Hampel, Ronchetti, Rousseeuw, and Stahel (1986);

and Staudte and Sheather (1990). This section briefly reviews some aspects of these results that are particularly relevant to applied researchers. One of the most striking results is that intuitions about how to deal with outliers and heavy-tailed distributions, based in part on notions related to normal distributions, turn out to be completely unsatisfactory. For example, discarding outliers and applying Student's t test to the data that remain results in a procedure based on estimated standard errors that are not even asymptotically correct. That is, the estimated standard errors do not converge to the correct value as the sample sizes increase. Another common misconception is that outliers can be detected using rules based on s, the sample standard deviation. Because s can become highly inflated due to outliers, methods for detecting outliers that are based on s are subject to "masking." That is, the very presence of outliers hinders the ability of methods based on s to find them. A more satisfactory approach is to use a measure of dispersion that is insensitive to outliers. (See Brant, 1990; Davies & Gather, 1993; Hadi, 1994; Barnett & Lewis, 1984; Goldberg & Iglewicz, 1992; Rousseeuw & Leroy, 1987; and Rousseeuw & van Zomeren, 1990 for methods dealing with the detection of outliers.)

Three different ways of assessing an estimator play a prominent role in modern statistical methods: (a) the breakdown point, (b) the influence function, and (c) the continuity of a descriptive measure. These concepts are described by Staudte and Sheather (1990, p. 44) as quantitative robustness, infinitesimal robustness, and qualitative robustness. But before delving into greater detail, we should perhaps be more precise about what is meant by a *measure of location.* It is intended to represent the typical subject, but what basic properties should it have? There are four properties that are usually imposed (e.g., Staudte & Sheather, 1990, p. 101), and a fifth condition (described below) that is sometimes added.

Suppose X has distribution F, and let $\theta(X)$ be some descriptive measure of the distribution. Then $\theta(X)$ is a measure of location if for any constants a and b,

$$\theta(X + b) = \theta(X) + b, \qquad \text{(Condition 1)}$$
$$\theta(-X) = -\theta(X), \qquad \text{(Condition 2)}$$
$$X > 0 \text{ implies } \theta(X) > 0, \text{ and} \qquad \text{(Condition 3)}$$
$$\theta(aX) = a\theta(X). \qquad \text{(Condition 4)}$$

Condition 1 is called location equivariance and simply requires that if b is added to every observation, then a measure of location should be increased by the same amount. Conditions 1–3 imply that any measure of location should have a value within the range of values of X. Condition 4 is called scale equivariance and guarantees that estimators of measures of location, such as the sample mean, give results independent of the unit of measurement. In the context of testing hypotheses, it should not be possible to multiply all the observations by the constant a and come to different conclusions about whether two groups differ. There are many measures of location in addition to the mean and median. To make judgments about how measures of location compare, additional desiderata must be imposed.

The discussion of the contaminated normal and chi-square distributions suggests one approach to characterizing robust and resistant measures of location, and these distributions are examples of what is called a contamination neighbor-

hood. That is, they represent a family of distributions that are in some sense close to a distribution of interest, F. Other neighborhoods have been defined, but they can be conceptually inadequate as argued by Huber (1981, pp. 11–12). A more satisfactory point of view has been developed, but involves technical details that go beyond the scope of this article. (In terms of determining whether a distribution is close to F, the Lévy and Prohorov distances play an important role, but the technical details are not covered in this review. See Huber, 1981, for further information.) It is possible, however, to gain some sense of how to proceed by first considering a few comments about mathematics.

Consider any function, say $f(x)$, not necessarily a probability function, and suppose we want to impose restrictions on $f(x)$ so that it does not change substantially with very small changes in x. Two restrictions that might be imposed are that $f(x)$ is differentiable, and that the derivative be bounded. For example, if $f(x) = x^2$, and x is positive, then the rate at which $f(x)$ increases gets larger as x goes to infinity. This rate is $2x$, the derivative of $f(x)$. A key element of modern robust statistical methods is an analog of these two restrictions for working with measures of location and scale.

The analog is obtained by viewing descriptive measures as functionals, which are just mappings that assign real numbers to distributions. For example, the population mean μ can be written as

$$T(F) = \int x dF(x),$$

where F is any distribution function. That is, $T(F)$ is a rule that maps any F into a real number. An advantage of this approach is that there are analogs of derivatives which indicate how sensitive $T(F)$ happens to be in terms of slight changes in F, and it also leads to asymptotically correct estimators of standard errors. Derivatives of functionals are called Gâteaux derivatives, and in the robustness literature they are generally known as an *influence function.* Thus, a basic requirement of any measure of location and scale is that it have a bounded influence function. The population mean has the influence function $IF(x) = x - \mu$, which is unbounded. That is, it can be made arbitrarily large by increasing x. Put in more practical terms, for a skewed distribution, very small changes in a distribution, F, can have an arbitrarily large effect on μ. The variance has an unbounded influence function as well, and this is why power can be poor when one is working with means, and Δ can be a misleading measure of effect size even with infinitely large sample sizes.

The influence function approximates the relative influence on $T(F)$ of small departures from F. Roughly, it reflects the influence of adding an observation with value x to an already large sample. More formally, suppose δ_x is a probability function where the value x occurs with probability 1. Then $T((1 - \epsilon)F + \epsilon\delta_x)$ is a functional, such as some measure of location, evaluated for a distribution where with probability $1 - \epsilon$ an observation is sampled from F, and with probability ϵ the observed value is x. Put another way, there is probability $1 - \epsilon$ of getting a "good" observation, and probability ϵ of getting a "bad" or contaminated value, x. Then $T((1 - \epsilon)F + \epsilon\delta_x) - T(F)$ measures the change due to including

x with probability ϵ. Dividing this difference by ϵ and taking the limit as ϵ goes to zero provides the formal definition of the influence function.

One of the more important roles played by the influence function is that when one is dealing with a measure of location other than the mean, it indicates how the standard error should be estimated. Let $X_1, \ldots, X_n$ be a random sample, and let $\hat{\theta}$ be an estimator of θ, some measure of location that is of interest. For the more popular measures of location appearing in the robustness literature,

$$\hat{\theta} = \theta + \Sigma\, IF(X_i), \tag{1}$$

plus a remainder term that goes to zero in probability as the sample size increases. Thus, $\hat{\theta}$ can be written as the sum of independent and identically distributed random variables. Consequently, once the influence function has been determined, an estimator of the standard error is available as well from basic principles in mathematical statistics.

As an illustration, consider the sample trimmed mean given by

$$\bar{X}_t = (X_{(g+1)} + \cdots + X_{(n-g)})/(n - 2g),$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the observations written in ascending order, and $g = [kn]$, where k is some predetermined constant between 0 and .5, and the notation $[kn]$ indicates that kn is rounded down to the nearest integer. (For example, $[9.9] = 9$.) In words, the sample trimmed mean is computed by removing the g largest and smallest observations and averaging the values that remain. (Choosing k, the proportion of observations to be trimmed, is discussed in the following section.) The influence function of the population trimmed mean can be used to show that the standard error of the sample trimmed mean, $\bar{X}_t$, should be estimated as follows. Let

$$\bar{X}_w = \frac{1}{n}((g + 1)X_{(g+1)} + X_{(g+2)} + \cdots + X_{(n-g-1)} + (g + 1)X_{(n-g)})$$

be the Winsorized mean, and let

$$SSD = (g + 1)(X_{(g+1)} - \bar{X}_w)^2 + (X_{(g+2)} - \bar{X}_w)^2 + \cdots + (X_{(n-g-1)} - \bar{X}_w)^2 + (g + 1)(X_{(n-g)} - \bar{X}_w)^2.$$

The Winsorized sample variance is

$$s_w^2 = \frac{SSD}{n - 1}.$$

Then an asymptotically correct estimator of the standard error of the sample trimmed mean is $s_w/\{(1 - 2k)\sqrt{n}\}$. Note that if you trim and then compute the sample variance with the data that remain, you do not get the Winsorized sample variance, s_w^2. The estimate of the standard error follows from the result that

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

plus a remainder term that goes to zero in probability as n gets large, where μ_t is the population trimmed mean and $IF(X_i)$ is the influence function of the trimmed mean. That is, $\bar{X}_t$ can be written as the average of independent and identically distributed random variables, plus a term that can be ignored asymptotically, even though the order statistics used to compute $\bar{X}_t$ are dependent. When x is between x_k and x_{1-k}, the k and $1 - k$ quantiles of F, $IF(x) = (x - \mu_w)/(1 - 2k)$, where μ_w is the population Winsorized mean estimated by $\bar{X}_w$. When $x > x_{1-k}$, $IF(x) = (x_{1-k} - \mu_w)/(1 - 2k)$, while for $x < x_k$, $IF(x) = (x_k - \mu_w)/(1 - 2k)$. Note that the influence function is bounded below by $(x_k - \mu_w)/(1 - 2k)$, and it is bounded above by $(x_{1-k} - \mu_w)/(1 - 2k)$. (Readers interested in more technical details can refer to Huber, 1981, or Staudte & Sheather, 1990.)

Another common criterion is the so-called *finite sample breakdown point* of an estimator. This refers to the proportion of observations which, when altered, can make an estimator arbitrarily large. Put another way, the finite sample breakdown point reflects the amount of contamination that can be tolerated. For example, the finite sample breakdown point of the sample mean is only $1/n$ because if a single observation is increased indefinitely, then the sample mean goes to infinity. The sample variance has a finite sample breakdown point of only $1/n$, as well. The limiting value of the finite sample breakdown point, as n gets large, provides a measure of the global stability of an estimator. For the sample mean this limit is zero, while for the trimmed mean it is k, the proportion of observations trimmed from both tails. The best that can be achieved is a finite sample breakdown point approximately equal to .5. The sample median has this property, and its standard error is relatively small when one samples from heavy-tailed distributions, but its standard error is relatively large for normal distributions, resulting in potentially low power compared to other methods for comparing groups.

M-Estimators

There are many measures of location, including *R*-estimators, which are related to tests of hypotheses based on ranks. No attempt is made to describe all of these estimators here. A class of estimators of location that should be discussed, and which has received considerable attention in the statistics literature, is *M*-estimators, which contain maximum likelihood estimators as a special case. In their simplest form, *M*-estimators are estimates of measures of location, θ, where θ satisfies

$$E\{\Psi(x - \theta)\} = 0, \tag{2}$$

and Ψ is some function chosen to have desirable properties. (For a more extensive yet relatively nontechnical discussion of *M*-estimators, see Hoaglin et al., 1983.) Note that for $\Psi(x) = x$, θ becomes the population mean, μ. *M*-estimators are mathematically appealing because they have a simple influence function, which aids in the choice of Ψ. The two most popular choices for Ψ are Tukey's biweight

and Huber's Ψ, the latter given by $\Psi(x) = \max\{-k, \min(x, k)\}$. The constant k is a tuning constant chosen so that the estimator has good properties under normality. Typically $k = 1.28$, the .9 quantile of the standard normal distribution, and this value is used henceforth. There are serious estimation problems associated with Tukey's biweight (Freedman & Diaconis, 1982), so this approach is not discussed further. Using Huber's Ψ, estimating θ is easily accomplished and the estimator has a relatively high finite sample breakdown point (see Huber, 1981, p. 144), but other difficulties remain. Recall from Condition 4 for measures of location that if X has measure of location θ, then bX should have measure of location $b\theta$, whenever $b > 0$. In general, M-estimators satisfy all but the last criterion. This is easily corrected by incorporating a measure of scale into Equation 2. The measure of scale typically used is the median absolute deviation statistic given by

$$MAD = \text{med}|X_i - M|,$$

where M is the usual sample median. That is, subtract the median from each observation and then compute the absolute value of the resulting n observations, in which case MAD is the median of the absolute values just computed. MAD has a finite sample breakdown point of approximately .5, which is one reason it is commonly used. Iterative methods must be used to estimate θ, but typically one iteration via the Newton-Raphson method suffices. This yields

$$\hat{\theta} = \frac{1.28(MAD)(i_2 - i_1) + S}{n - i_1 - i_2},$$

where i_1 is the number of observations X_i for which $(X_i - M)/MAD < -k$, i_2 is the number for which $(X_i - M)/MAD > k$, and

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

In other words, $\hat{\theta}$ empirically determines whether an observation is unusually large or small by comparing $(X_i - M)/MAD$ to k; it then trims these observations and averages those that remain, but it also makes an adjustment based on a measure of scale, MAD. Note that when equal amounts of trimming are done ($i_1 = i_2$), MAD no longer plays a role in the estimate of θ. The estimator $\hat{\theta}$ is generally known as a one-step M-estimator of location. Although MAD has a finite sample breakdown point of .5, a criticism is that it is relatively inefficient. Rousseeuw and Croux (1993) consider alternative measures of deviation, but the implications of these measures, in terms of testing hypotheses, have not been explored.

Bickel and Lehmann (1975) add a fifth criterion to the four that define a measure of location. Roughly, if the quantiles of a random variable X are greater than or equal to the quantiles of the random variable Y, then any measure of location for X should be larger than the corresponding value for Y. In general,

robust M-estimators do not satisfy this last criterion. Trimmed means satisfy all of the criteria described in this section, so some authorities believe they should be preferred in practice. Others argue that the one-step M-estimator is still reasonable because its value is close to the "bulk" of a distribution, and because it empirically determines how much trimming is done, while the trimmed mean does not. In particular, if a distribution is skewed to the right, say, it might seem preferable to trim more observations from the right tail compared to the left. Also, the one-step M-estimator includes the possibility of not trimming any observations at all.

M-estimators in general, and the one-step M-estimator in particular, are asymptotically normal. For symmetric distributions the one-step M-estimator has a fairly simple influence function, but for asymmetric distributions it takes on a rather complicated form (Huber, 1981, p. 140). Despite this, there are practical methods for making inferences about a one-step M-estimator, and estimates of its standard error can be obtained, as well. Details are given in some of the remaining sections of this article.

For completeness, improvements on M-estimators appear to be possible (Morgenthaler & Tukey, 1991), but many complications remain to be resolved, so they are not discussed. Still another possibility is the minimum volume ellipsoid estimator used by Rousseeuw and Leroy (1987). Also, other methods of empirically determining how much trimming to do have been suggested (e.g., Hogg, 1974), but technical difficulties typically render them of questionable value. (See the discussion following Hogg's article, particularly the comments by P. Huber.)

Testing Hypotheses: The One-Sample Case

This section takes up the problem of making inferences about a measure of location in the one-sample case. A general point of considerable practical importance is that the measure of location chosen can be extremely important in terms of Type I errors, accuracy of a confidence interval, the interpretation of a measure of effect size, and power. In particular, rigid adherence to means can result in missing important differences even when there are large sample sizes. The ideal hypothesis testing procedure would have as much or more power than any other technique, but such a procedure has not been derived. In fact, despite the negative features related to means which are described in this section and the next, there are situations in which comparing means can have higher power than all other methods, even under nonnormality. Details and illustrations are given below.

Attention is focused on trimmed means and one-step M-estimators of location because currently they seem to be two of the more important estimators for consideration in applied work. Issues that must be addressed are whether good control over Type I errors can be achieved, whether confidence intervals have accurate probability coverage, whether power compares well to methods based on means when distributions are normal, and whether there is a large difference in power when distributions have heavy tails.

First consider trimmed means. Tukey and McLaughlin (1963) devised a method for testing hypotheses which uses Student's t distribution to approximate the null distribution with degrees of freedom $\nu = n - 2g - 1$. Results in Patel, Mudholkar, and Fernando (1988) support this approximation of the null distribution. More specifically, letting μ_t be the population trimmed mean,

$$T_t = \frac{\sqrt{n}\,(1 - 2k)(\bar{X}_t - \mu_t)}{s_w}$$

has, approximately, a Student's t distribution with $n - 2g - 1$ degrees of freedom. Using T_t to test hypotheses can yield substantially more power than Student's t test because the Winsorized standard deviation, s_w, is less affected by heavy-tailed distributions and outliers. Note that s_w effectively "pulls in" extreme values.

To have any practical utility, a value for k, the amount of trimming, must be determined. Results in Rosenberger and Gasko (1983) indicate that $k = .2$ is a good choice, although with very small sample sizes they recommend $k = .25$ or higher. Wilcox (1994c) derived analytic results showing that for skewed distributions, the more trimming is done, the better the control over the probability of a Type I error. However, if too much trimming is done, power might be poor for normal distributions. Wilcox found that a good compromise is $k = .2$, which gives good power for a shift model under normality, and which can yield substantially more power when there are outliers or distributions have heavy tails. Another point of view is that k be chosen according to how many outliers one expects based on experience with similar types of data. Yet another approach is to empirically determine the amount of trimming according to some criterion of interest, such as a relatively small standard error. These so-called adaptive trimmed means have been studied recently by Léger and Romano (1990a, 1990b) and Léger, Politis, and Romano (1992).

While Student's t test is generally thought to be robust under nonnormality, control over the probability of a Type I error can be poor, especially when one is dealing with one-sided tests. For example, when one is testing H_0: $\mu > 0$, and sampling is from a lognormal distribution, the actual probability of a Type I error is .124 with $n = 80$ and $\alpha = .05$ (Westfall & Young, 1993, p. 40; Noreen, 1989, p. 74). Increasing n to 160, it is .109. The problem is that the null distribution is skewed, and the expected value of the test statistic is not zero as it is assumed to be. Sutton (1993) compared several methods and found that a method suggested by Johnson (1978), used in conjunction with a bootstrap procedure, gives good results depending on whether a distribution is skewed to the right or left (cf. Kleijnen, Kloppenburg, & Meeuwsen, 1986). (For a relatively nontechnical description of the bootstrap method, see Noreen, 1989; Wilcox, in press-c. For a review of the bootstrap written at a more technical level, see DiCiccio & Romano, 1988.) The Tukey-McLaughlin test for trimmed means gives better control over the probability of a Type I error than does Student's t test, but as with all methods for comparing means, problems remain.

Westfall and Young (1993) suggest that when one is working with the mean, a particular form of Efron's (1982) bootstrap method be used. Their method improves control over the probability of a Type I error, but even with $n = 80$, the actual probability of a Type I error can be as high as .08. Just how serious this is will depend on the situation and judgments made about the importance of a Type I error. (Bradley, 1978, argues that when one is testing at the .05 level, the actual probability of a Type I error should not exceed .075.) The main point here is that combining the bootstrap with trimmed means gives even better results. For the lognormal distribution considered here, the probability of a Type I error

is .06 with only $n = 12$. Complete computational details are given in Wilcox (1994a), and thorough discussion of the general method is provided by Westfall and Young (1993). Briefly, set $C_i = X_i - \bar{X}_t$, randomly sample n observations, with replacement, from the C_i values, and compute T_t based on the n values just sampled yielding T_t^*. Repeat this process B times yielding $T_{t1}^* \ldots, T_{tB}^*$. Let $\hat{p}$ be the proportion of these B values less than T_t, the test statistic based on the observed data, $X_1, \ldots, X_n$. Then $\hat{p}$ estimates the significance level when one is testing H_0: $\mu_t \geq 0$. If $\hat{p} < \alpha$, reject. Westfall and Young recommend $B = 10{,}000$ although $B = 1{,}000$ seems to give good results. The Minitab macro trim1b.mtb, which is included on the floppy accompanying Wilcox's (in press-c) book, performs the calculations.

When one is working with one-step M-estimators, and when distributions are skewed, it appears there is no simple method for approximating the null distribution. However, results in Wilcox (1992a) suggest that fairly accurate confidence intervals can be obtained using a percentile bootstrap method. That is, obtain a bootstrap sample by randomly sampling n observations with replacement from $X_1, \ldots, X_n$. Let $\hat{\theta}^*$ be the resulting value of the one-step M-estimator. Repeat this process B times yielding $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, which yields an estimate of the sampling distribution of $\hat{\theta}$, the one-step M-estimator based on the observed values $X_1, \ldots, X_n$. The percentiles of the estimated sampling distribution can be used to compute a confidence interval. $B = 399$ appears to give good results when one is computing a .95 confidence interval. $B = 399$ rather than $B = 400$ is recommended because of results in Hall (1986). The B bootstrap values can also be used to obtain an estimate of the standard error. Other forms of the bootstrap might give better results in some sense, but this remains to be seen.

Comparing Two or More Groups

Let θ_1 and θ_2 be any measures of location corresponding to two independent groups. This section takes up the problem of testing H_0: $\theta_1 = \theta_2$ or computing a $1 - \alpha$ confidence interval for $\theta_1 - \theta_2$. As before, there are the general concerns about whether a measure of location is robust, about the possibility that a nonrobust measure can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power.

For working with two or more groups, this article considers only heteroscedastic methods—that is, methods that allow unequal variances. There is now a vast literature indicating that the usual t test can be unsatisfactory in terms of Type I errors when distributions have unequal variances, and perhaps more importantly, that heteroscedastic methods can have substantially more power and more accurate probability coverage when computing confidence intervals. This is consistent with the review by Glass et al. (1972), and today there is even more evidence that homoscedastic methods are unsatisfactory, even under normality, as indicated in the review by Wilcox (1993d). This might appear to contradict well-known results in Box (1954), but Box limited his numerical results to situations where the ratio of the largest standard deviation to the smallest is less than or equal to the square root of 3. Surveys of published studies indicate that larger ratios are common, in which case problems arise even when sampling from normal

distributions. In fact, when distributions have different shapes, Cressie and Whitford (1986) describe general circumstances where Student's t test is not even asymptotically correct. That is, the variance of Student's t test does not approach 1 as the sample sizes get large, so in particular it does not approach a standard normal distribution as is typically assumed. One of the negative aspects of Student's t test is poor control over Type I error probabilities when groups have unequal variances. It might be argued that it is unrealistic to have equal means but unequal variances, in which case Type I errors are not a concern; but when two groups are being compared, the problem of getting accurate confidence intervals remains, and homoscedastic methods have peculiar power properties, as well (Wilcox, in press-a). A natural approach is to test for equal variances and use conventional methods if a nonsignificant result is obtained; but even under normality, tests for equal variances do not have enough power to detect unequal variances in situations where violating the assumption causes problems (Markowski & Markowski, 1990; Moser, Stevens, & Watts, 1989; Wilcox, Charlin, & Thompson, 1986), and there is the additional problem that most methods for comparing variances do not control Type I error probabilities (Wilcox, 1992c). This is true of the methods compared by Conover, Johnson, and Johnson (1981). The one method found by Wilcox (1992c) to control Type I errors also has relatively low power.

One of the best-known methods for comparing means and handling unequal variances is Welch's (1951) adjusted degrees of freedom procedure. (See also Brown & Forsythe, 1974a, 1974b; Nanayakkara & Cressie, 1991; James, 1951; Krutchkoff, 1988; and Matuszewski & Sotres, 1986.) Algina, Oshima, and Lin (1994) describe situations where even with equal sample sizes, Welch's method can be unsatisfactory. For example, with $n_1 = n_2 = 50$, the actual probability of a Type I error can be .08 when one is sampling from a lognormal distribution with $\alpha = .05$. With $n_1 = 33$ and $n_2 = 67$, the Type I error probability is .11. Similar problems are reported by Wilcox (1990b). As in the one-sample case, problems arise when one is working with skewed distributions. Cressie and Whitford (1986) describe correction terms based on estimated skewness, but their approach can actually make matters worse (Wilcox, 1990a). Tiku's (1982) method was also found to be unsatisfactory. Oshima and Algina (1992) compared two heteroscedastic methods for means and again found problems with controlling Type I errors. In general, heteroscedastic methods improve upon homoscedastic techniques, but problems remain.

Yuen (1974) extended the Tukey-McLaughlin test for trimmed means to the two-sample case; this reduces to Welch's (1951) test for means when there is no trimming. An omnibus test for more than two groups is described by Wilcox (in press-a). For multiple comparison procedures, see Wilcox (in press-c). As in the one-sample case, analytic results indicate that confidence intervals for the difference between two trimmed means are generally more accurate than confidence intervals for means, especially when distributions are skewed (Wilcox, 1994c). However, problems with controlling Type I error probabilities still remain, particularly when a directional test is performed. Better results can be obtained by using a straightforward analog of the bootstrap method for trimmed means described in the previous section. Computational details are given in Wilcox

(1994a), and the Minitab macro trim2b.mtb, which comes with Wilcox's (in press-c) book, performs the calculations.

An appealing feature of trimmed means is that any of the common experimental designs can be analyzed, including repeated measures (Wilcox, 1993a) and random effects designs (Wilcox, 1994b). For linear contrasts, multiple comparisons, and two-way designs, see Wilcox (in press-c). More complicated designs are easily handled using similar techniques, but there are no published simulation results on how they perform. Results for the random effects model are striking not only in terms of power, but in improved control over the probability of a Type I error (Wilcox, 1994b). For example, the conventional F test can have Type I error probabilities exceeding .3, and for nonnormal distributions the heteroscedastic method suggested by Jeyaratnam and Othman (1985) can have a Type I error probability as high as .4. In contrast, using trimmed means with 20% trimming yields Type I error probabilities that never exceed .075 for the same situations.

Comparing one-step M-estimators for two or more independent groups can be accomplished using a bootstrap method (Wilcox, 1993b). (For a homoscedastic method of comparing groups using M-estimators, see Schrader & Hettmansperger, 1980.) Currently, control over the probability of a Type I error appears to be reasonably good with $\alpha = .05$ and sample sizes of at least 20. In general, trimmed means with 20% trimming seem to give good control over Type I error probabilities for a larger range of situations. In principle, dependent groups can be compared using one-step M-estimators and a method similar to the one used for independent groups, but there are no results on how well this approach controls Type I errors.

Though comparing groups with the usual sample median can mean relatively low power when distributions are normal, several improved methods for estimating the population median which can have substantially smaller standard errors have appeared in the statistical literature. From Parrish (1990), an improved estimator that deserves consideration is one derived by Harrell and Davis (1982). Yoshizawa, Sen, and Davis (1985) show that the Harrell-Davis estimator is asymptotically normal. Their method consists of using a weighted linear combination of all the ordered observations. Given n, the weights can be chosen so that the population median is estimated. Weighted sums of order statistics are generally called L-estimators, which include trimmed means as a special case. Comparing two or more independent groups can be accomplished with a bootstrap method as described in Wilcox (1991b), and dependent groups can be compared using a similar technique (Wilcox, 1992b). (See also Hettmansperger, 1984; and Lunneborg, 1986.)

To illustrate the extent to which different methods can give different results, Table 1 shows the power of six methods for three distributions when two independent groups with $n_1 = n_2 = 25$ are compared. The mean of the first is 0, and the other measures of location considered here are equal to 0, as well. The second group has a mean of 1. The first distribution is normal, the second (CN1) is contaminated normal with $\epsilon = .1$ and $K = 10$, and the third (CN2) is also contaminated normal but with $K = 20$. Method M uses the usual sample median as described by Wilcox and Charlin (1986). Method C compares medians using the Harrell-Davis estimator, while Method H compares groups using one-step

TABLE 1
Power of six methods for comparing two groups

Distribution	Welch	M	Yuen (10%)	Yuen (20%)	C	H
Normal	.931	.758	.914	.890	.865	.894
CN1	.278	.673	.705	.784	.778	.804
CN2	.162	.666	.383	.602	.639	.614

Note. Welch = Welch's (1951) adjusted degrees of freedom procedure; M = method using the usual sampling median as described by Wilcox & Charlin (1986); Yuen (10%) = Yuen's (1974) extension of the Tukey-McLaughlin test with 10% trimming; Yuen (20%) = Yuen's (1974) extension of the Tukey-McLaughlin test with 20% trimming; C = method comparing medians using the Harrell-Davis estimator; H = method comparing groups using one-step M-estimators.

M-estimators. As is evident, Welch's method is the most affected by nonnormality, with power dropping from .931 to .162.

Though the latter three methods in Table 1 compare well to Welch's procedure, there are situations where Welch's method can have more power than any of the other methods considered here. The problem is that the mean is not equal to the trimmed mean, for example, when distributions are skewed, so it is possible to have $\mu_1 - \mu_2 > \mu_{t1} - \mu_{t2}$. That is, the difference between the means is larger than the difference between the trimmed means. To see how this might happen, look at Figure 6, which shows a skewed distribution with mean 7.6 and a 20% trimmed mean of 4.2. Now imagine a second distribution that is symmetric with mean 4.2. Because the distribution is symmetric, the trimmed mean is also 4.2, so $\mu_1 - \mu_2 = 7.6 - 4.2 = 3.4$, while $\mu_{t1} - \mu_{t2} = 0$. Put another way, detecting outliers does not necessarily imply that using means will result in less power relative to other methods based on robust measures of location.

An anonymous referee summarized a general strategy and point of view for approaching the two-sample problem which is roughly as follows. Though the means of two distributions might be of some interest, a more general issue is where the distributions differ and by how much. For example, the shift function of the self-awareness data indicates that the upper deciles, starting with the median, differ substantially compared to the lower deciles, and other global differences between the distributions can be investigated in a variety of ways. The issue of outliers can be addressed by considering, in addition to "automatic" outlier detection methods such as the boxplot, how data were collected and the many delicate ways in which outliers could disguise themselves. If of interest, the distributions could be summarized by some measure of location, and comparisons of robust measures of location could be compared to less robust (unbounded influence and low breakdown) measures such as the sample mean.

Transformations

Another common and natural approach to nonnormality is to transform the observations and apply methods for means. For example, it is common to replace each X_i with its logarithm. This often makes data look more normal, but outliers are not necessarily eliminated, and in some cases power remains poor. Wilcox

(in press-c) reanalyzed data from an actual study by replacing observations with their logarithms, but outliers remained and highly nonsignificant results remained nonsignificant; when trimmed means or one-step M-estimators are used, however, significant results are obtained. Rasmussen (1989) studied the Box-Cox transformation; he found it to be useful in some cases but concluded that it can be unsatisfactory when dealing with low power due to outliers.

Methods Based on Ranks

Still another approach to nonnormality is to compare groups based on ranks, For example, a simple strategy is to pool all the observations, assign ranks, and then apply Student's t test to the ranks corresponding to each group. It turns out that this is tantamount to applying the Mann-Whitney test (Conover & Iman, 1981). The Mann-Whitney test is satisfactory when there is no difference between groups and the distributions are, in fact, identical; when distributions differ, however, problems arise. For example, the Mann-Whitney test is both biased and inconsistent (Kendall & Stuart, 1973). That is, when the null hypothesis is false, there are situations where the probability of rejecting is less than the nominal α value, and there are also situations where power does not approach 1 as the sample sizes get large. One difficulty is getting a consistent estimate of the standard error of the average ranks when the null hypothesis is false. In particular, the estimated standard error used by the Mann-Whitney test can converge to the wrong value as the sample sizes increase.

Zaremba (1962) derived an improvement on the Mann-Whitney test, but it can have relatively low power. Fligner and Policello (1981) proposed a method for comparing medians based on ranks. Their procedure is easy to use, it is closely related to a procedure where Welch's test is applied to the ranks, and it can have relatively high power when distributions have heavy tails. The Fligner-Policello procedure is unbiased and consistent if both distributions are symmetric. It is noted that for asymmetric distributions, the Fligner-Policello procedure is actually testing $H_0 : p = .5$, where p is the probability that a randomly sampled observation from the first group is larger than a randomly sampled observation from the second. Inferences about p are of interest in their own right, as argued by Cliff (1993). For results on computing a confidence interval for p, see Mee (1990). The Minitab macro mee.mtb in Wilcox (in press-c) does the necessary computations. For more recent results on using ranks, see Thompson (1991) and Akritas (1991).

Comments About Multiple Comparison Procedures

This section briefly comments on performing all pairwise comparisons of two or more groups. For working with means, Dunnett's (1980) T3 and C procedures stand out as providing relatively accurate confidence intervals (Hochberg & Tamhane, 1987). An important point is that Dunnett's method is designed to control the experimentwise Type I error probability (the probability of at least one Type I error among all tests to be performed) without any dependence on any other preliminary test. In particular, it does not require that an omnibus test for equal means be performed first, and in fact, if it is made contingent on an omnibus test being significant, results in Bernhardson (1975) imply that power might actually be lowered. (Analogs of Dunnett's method for trimmed means

are described in Wilcox, in press-c). This does not mean that omnibus tests have no value, but they should be used with caution.

When distributions are normal, so called step-down procedures offer advantages in terms of all-pairs power, the probability of detecting all true differences among several groups. These techniques start with an omnibus test for all J groups. If significant, one then tests all subsets consisting of $J - 1$ groups, and the process continues until one tests all pairs of groups. Complete details are not important here; interested readers can refer to Hochberg and Tamhane (1987) or Wilcox (1991a; in press-c). The point here is best made with a simple illustration. Suppose five groups are being compared, and all five groups have unequal means. Also suppose the first four are normal, but the last group is nonnormal with heavy tails. Then the last group can cause any omnibus test for equal means to have low power, in which case the differences among the groups are unlikely to be detected, even though Dunnett's T3 has high power when comparing the first four groups. A numerical illustration is given in Wilcox (in press-c). Of course, if two or more groups have heavy-tailed distributions, power can again be relatively poor. However, replacing means with trimmed means can correct this difficulty, so step-down techniques might be of interest provided one is willing to sacrifice confidence intervals for a potential increase in all-pairs power. For more complete details relevant to trimmed means, see Wilcox (in press-c).

Measures of Scale

As a final note, there are many robust measures of scale in addition to the nonrobust sample standard deviation, s. One of these is *MAD*, already mentioned, which has a finite sample breakdown point of approximately .5, but which is relatively inefficient. Lax (1985) compared about 150 measures of location and found the so-called biweight midvariance to have good properties based on the criteria he used. In particular, it has high efficiency. It is related to M-estimators of location and discussed at some length, in conjunction with many other measures, by Iglewicz (1983). From Goldberg and Iglewicz (1992), it appears to have a finite sample breakdown point of approximately .5, but a formal proof has not been found. For normal distributions, the biweight midvariance has a value very similar to the standard deviation (Shoemaker & Hettmansperger, 1982), but unlike s, it is not overly affected by heavy tails. A method for comparing the biweight midvariances of two independent groups was investigated by Wilcox (1993c), and it appears to perform well over a wide range of distributions; software for performing the test accompanies the textbook by Wilcox (in press-c). For general results on measures of scale based on trimming and Winsorization, which include a derivation of their influence function, see Welsh and Morrison (1990).

One other issue concerning choosing a measure of scale should be mentioned. Let τ_x be a measure of scale associated with the random variable X. Bickel and Lehmann (1976) define τ_x to be a measure of dispersion if $\tau_x > \tau_y$ whenever the quantiles of the distribution of $|X|$ are larger than the corresponding quantiles of the distribution of $|Y|$, and they argue that measures of scale should be measures of dispersion. The biweight midvariance is not a measure of dispersion, while the Winsorized variance is. Another measure of scale that is also a measure of

dispersion is the percentage bend variance discussed by Shoemaker and Hettmansperger (1982).

References

Akritas, M. G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part I. *Journal of the American Statistical Association, 86,* 457–460.

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19,* 91–101.

Algina, J., Oshima, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19,* 275–291.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data.* New York: Wiley.

Bernhardson, C. (1975). Type I error rates when multiple comparison procedures follow a significant *F* test of ANOVA. *Biometrics, 31,* 719–724.

Bickel, P. J., & Lehmann, E. L. (1975). Descriptive statistics for nonparametric models. II. Location. *Annals of Statistics, 3,* 1045–1069.

Bickel, P. J., & Lehmann, E. L. (1976). Descriptive statistics for nonparametric models III. Dispersion. *Annals of Statistics, 4,* 1139–1158.

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

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

Brant, R. (1990). Comparing classical and resistant outlier rules. *Journal of the American Statistical Association, 85,* 1083–1090.

Brown, M. B., & Forsythe, A. B. (1974a). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics, 30,* 719–724.

Brown, M. B., & Forsythe, A. B. (1974b). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16,* 129–132.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 494–509.

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

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances with applications to the outer continental shelf bidding data. *Technometrics, 23,* 351–361.

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

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association, 88,* 782–792.

DiCiccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society, B50,* 338–370.

Doksum, K. A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two sample case. *Annals of Statistics, 2,* 267–277.

Doksum, K. A. (1977). Some graphical methods in statistics. A review and some extensions. *Statistica Nederlandica, 31,* 53–68.

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

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association, 75,* 796–800.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Fligner, M. A., & Policello, G. E., II. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76,* 162–168.

Freedman, D. A., & Diaconis, P. (1982). On inconsistent *M*-estimators. *Annals of Statistics, 10,* 454–461.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42,* 237–288.

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

Gnanadesikan, R., & Wilk, M. B. (1968). Probability plotting methods for nonlinear models in the two-sample case. *Biometrika, 55,* 1–18.

Goldberg, K. M., & Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics, 34,* 307–320.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, B56,* 393–396.

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

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* Unpublished doctoral dissertation, University of California, Berkeley.

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

Harrell, F. E., & Davis, D. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69,* 635–640.

Hettmansperger, T. P. (1984). Two-sample inference based on one-sample sign statistics. *Applied Statistics, 33,* 45–51.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69,* 909–923.

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.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35,* 73–101.

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

Iglewicz, B. (1983). Robust scale estimators and confidence intervals for location. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 404–431). New York: Wiley.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38,* 324–329.

Jeyaratnam, S., & Othman, A. B. (1985). Test of hypothesis in one-way random effects model with unequal error variances. *Journal of Statistical Computation and Simulation, 21,* 51–57.

Johnson, N. J. (1978). Modified *t* tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association, 73,* 536–544.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kleijnen, J. P. C., Kloppenburg, G. L. J., & Meeuwsen, F. L. (1986). Testing the mean of an asymmetric population: Johnson's *t* test revisited. *Communications in Statistics—Simulation Computation, 15,* 715–732.

Krutchkoff, R. G. (1988). One-way fixed effects analysis of variance when the error variances may be unequal. *Journal of Statistical Computation and Simulation, 30,* 259–271.

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performances in long-tailed symmetric distributions. *Journal of the American Statistical Association, 80,* 736–741.

Léger, C., & Romano, J. P. (1990a). Bootstrap adaptive estimation: The trimmed mean example. *The Canadian Journal of Statistics, 18,* 297–314.

Léger, C., & Romano, J. P. (1990b). Bootstrap choice of tuning parameters. *Annals of the Institute of Mathematical Statistics, 42,* 709–735.

Léger, C., Politis, D. N., & Romano, J. P. (1992). Bootstrap technology and applications. *Technometrics, 34,* 378–398.

Lunneborg, C. E. (1986). Confidence intervals for a quantile contrast: Application of the bootstrap. *Journal of Applied Psychology, 71,* 451–456.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60.

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

Matuszewski, A., & Sotres, D. (1986). A simple test for the Behrens-Fisher problem. *Computational Statistics and Data Analysis, 3,* 241–249.

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association, 85,* 793–800.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Morgenthaler, S., & Tukey, J. W. (1991). *Configural polysampling.* New York: Wiley.

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.

Nanayakkara, N., & Cressie, N. (1991). Robustness to unequal scale and other departures from the classical linear model. In W. Stahel & S. Weisberg (Eds.), *Directions in robust statistics and diagnostics* (Part II). New York: Springer-Verlag.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses.* New York: Wiley.

O'Brien, P. C. (1988). Comparing two samples: Extensions of the *t,* rank-sum, and log-rank tests. *Journal of the American Statistical Association, 83,* 52–61.

Oshima, T. C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcox's H_m test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology, 45,* 255–264.

Pagurova, V. Z. (1978). A test for comparison of mean values in two normal samples. *Selected Translations in Mathematical Statistics and Probability, 14,* 149–158.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics, 46,* 247–257.

Patel, K. P., Mudholkar, G. S., & Fernando, J. L. (1988). Student's *t* approximation for three simple robust estimators. *Journal of the American Statistical Association, 83,* 1203–1210.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika, 23,* 114–133.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* New York: Wiley.

Rasmussen, J. L. (1989). Data transformations, 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 & outlier detection.* New York: Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85,* 633–639.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88,* 1273–1283.

Salk, L. (1973). The role of the heartbeat in the relations between mother and infant. *Scientific American, 235,* 26–29.

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance. *Biometrika, 67,* 93–101.

Shoemaker, L. H., & Hettmansperger, T. P. (1982). Robust estimates and tests for the one- and two-sample scale models. *Biometrika, 69,* 47–54.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* New York: Chapman and Hall.

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

Student. (1927). Errors of routine analysis. *Biometrika, 19,* 151–164.

Sutton, C. D. (1993). Computer intensive methods for tests about the mean of an asymmetrical distribution. *Journal of the American Statistical Association, 88,* 802–810.

Switzer, P. (1976). Confidence procedures for the two-sample problems. *Biometrika, 63,* 13–25.

Tan, W. Y. (1982). Sampling distributions and robustness of *t, F* and variance-ratio in two samples and ANOVA models with respect to departures from normality. *Communications in Statistics—Theory and Methods, A11,* 2485–2511.

Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association, 86,* 410–419.

Tiku, M. L. (1982). Robust statistics for testing equality of means or variances. *Communications in Statistics—Theory and Methods, 11,* 2543–2558.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, W. Hoeffding, S. Ghurye, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448–485). Stanford, CA: Stanford University Press.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorizing 1. *Sankhya, A25,* 331–352.

Tukey, J. W. (1993). Where should multiple comparisons go next? In F. Hoppe (Ed.), *Multiple comparisons, selection, and applications in biometry* (pp. 187–207). New York: Marcel Dekker.

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

Welsh, A. H., & Morrison, H. L. (1990). Robust *L* estimators of scale with an application in astronomy. *Journal of the American Statistical Association, 85,* 729–743.

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

Wilcox, R. R. (1990a). Comparing the means of two independent groups. *Biometrical Journal, 32,* 771–780.

Wilcox, R. R. (1990b). Comparing variances and means when distributions have non-identical shapes. *Communications in Statistics—Simulation and Computation, 19,* 155–173.

Wilcox, R. R. (1991a). A step-down heteroscedastic multiple comparison procedure. *Communications in Statistics—Theory and Methods, 20,* 1087–1098.

Wilcox, R. R. (1991b). Testing whether independent groups have equal medians. *Psychometrika, 56,* 381–395.

Wilcox, R. R. (1992a). Comparing one-step *M*-estimators of location corresponding to two independent groups. *Psychometrika, 57,* 141–154.

Wilcox, R. R. (1992b). Comparing the medians of dependent groups. *British Journal of Mathematical and Statistical Psychology, 45,* 151–162.

Wilcox, R. R. (1992c). An improved method for comparing variances when distributions have non-identical shapes. *Computational Statistics & Data Analysis, 13,* 163–172.

Wilcox, R. R. (1992d). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science, 1,* 101–105.

Wilcox, R. R. (1993a). Analysing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology, 46,* 63–76.

Wilcox, R. R. (1993b). Comparing one-step *M*-estimators of location when there are more than two groups. *Psychometrika, 58,* 71–78.

Wilcox, R. R. (1993c). Comparing the biweight midvariances of two independent groups. *The Statistician, 42,* 29–35.

Wilcox, R. R. (1993d). Robustness in ANOVA. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 345–374). New York: Marcel Dekker.

Wilcox, R. R. (1994a). On testing hypotheses about measures of location. Unpublished technical report, Department of Psychology, University of Southern California.

Wilcox, R. R. (1994b). A one-way random effects model for trimmed means. *Psychometrika, 59,* 289–306.

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

Wilcox, R. R. (in press-a). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology.*

Wilcox, R. R. (in press-b). Comparing two independent groups via multiple quantiles. *The Statistician.*

Wilcox, R. R. (in press-c). *Statistics for the social sciences.* New York: Academic Press.

Wilcox, R. R. (in press-d). Three multiple comparison procedures for trimmed means. *Biometrical Journal.*

Wilcox, R. R., & Charlin, V. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics, 11,* 263–274.

Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the ANOVA *F, W* and *F** statistics. *Communications in Statistics—Simulation and Computation, B15,* 933–944.

Yoshizawa, C. N., Sen, P. K., & Davis, C. E. (1985). Asymptotic equivalence of the Harrell-Davis median estimator and the sample median. *Communications in Statistics—Theory and Methods, 14,* 2129–2136.

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

Zaremba, S. K. (1962). A generalization of Wilcoxon's test. *Monatshefte Math., 66,* 359–370.

Author

RAND R. WILCOX is Professor, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061. He specializes in quantitative methods.

Received February 2, 1994
Revision received July 29, 1994
Accepted October 17, 1994