Journal of Clinical Child and Adolescent Psychology
2002, Vol. 31, No. 3, 399–412

METHODOLOGICAL ARTICLE

Understanding the Practical Advantages of Modern ANOVA Methods

Rand R. Wilcox
Department of Psychology, University of Southern California

Examined the fundamental problems associated with standard hypothesis testing techniques. This article explains why many articles have failed to detect problems due to nonnormality and discusses the basics of modern methods aimed at correcting these problems.

Based on hundreds of published articles, it is now known that when groups differ, the analysis of variance (ANOVA) F test and related techniques that assume normality and homoscedasticity (equal variances) can perform poorly. In practical terms, if researchers interested in clinical child and adolescent psychology want to detect important differences among groups and accurately assess how the groups differ, and by how much, modern technology has much to offer. In fact, even highly nonsignificant results based on an F test can become significant. Moreover, modern methods offer improved control over the probability of a Type I error and more accurate confidence intervals.

One of the more commonly used research methods in studies of children and adolescents in clinical psychology is to compare groups based on measures intended to reflect the typical response given by the participants. Of course, the sample mean associated with each group is the most common way of representing the typical response, and there are well-known methods for comparing groups based on means such as the analysis of variance (ANOVA) *F* test, plus inferential methods based on linear contrasts, which include, as a special case, popular multiple comparison procedures (such as Tukey's honestly significant difference method). These routinely used methods are based on two crucial assumptions: normality and equal variances. A half century ago, there was some evidence suggesting that violating these assumptions posed no serious threat in applied work, and these limited results greatly influence how data are analyzed today. But during the past 40 years, hundreds of articles published in statistics and quantitative psychology journals have shown that violating these assumptions can cause serious problems in very realistic situations (see Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 1997, 2001, in press). Simultaneously, it seems fair to say that the lines of communication between statisticians and applied psychological researchers have nearly broken down because even the insights that began to appear during the 1960s—which have very serious implications for any applied researcher—are unknown by most psychologists (e.g., Tukey, 1960).

Today, improvements on the ANOVA *F* test can be roughly classified into one of three major categories with a myriad of details in each. One goal here is to outline these three categories, explain their practical importance, and describe how modern technology deals with relevant issues. Software for applying modern methods is available and is briefly described.

Before continuing, perhaps it should be stressed that this article is not an indictment against all published research dealing with child and adolescent psychology. Rather, the theme of this paper is that we have the technology for substantially improving our ability to detect true differences among groups and to describe how the groups differ. Roughly, when a significant result is reported based on an *F* test or some appropriate linear contrast, modern insights indicate that groups probably differ, but great care must be taken when interpreting what this difference is like. Said another way, standard methods perform well in terms of Type I errors when groups do not differ in any way, which means that in addition to having equal means, they also have equal variances and

Requests for reprints should be sent to Rand Wilcox, Department of Psychology, University of Southern California, University Park Campus, Los Angeles, CA 90089–1061. E-mail: rwilcox@usc.edu

the same amount of skewness. (More precise details are given later.) Where conventional methods break down is when groups differ in some manner; extremely poor properties can result in completely missing important differences among groups, and standard measures of effect size can indicate a small effect size when in fact a large effect exists. Moreover, characterizing differences and assessing the precision of estimated differences via confidence intervals is fraught with peril. Generally, the ANOVA F is designed to be sensitive to differences among the means of the groups under study, but in reality it is sensitive to a wide range of other features, which include unequal variances as a special case. The good news is that modern technology can deal with these problems in a very effective manner.

Understanding the Effects of Nonnormality

Many applied statistics books still claim that when working with means, nonnormality is not a serious concern except possibly when sample sizes are very small. These claims are not based on wild speculations, but it is now known that this view is incorrect, and we understand why earlier studies missed serious problems with conventional techniques. In some realistic situations, nonnormality is a concern even with a sample size of 300, and some problems persist no matter how large the sample size might be! A nontechnical explanation of why it was once thought that nonnormality could be ignored, and why we now know better, can be found in Wilcox (2001). Presumably some readers are not familiar with modern insights regarding problems due to nonnormality, so a brief review is provided here. There are, in fact, two types of nonnormality that play a crucial role. The first has to do with what are called heavy-tailed distributions, roughly meaning that unusually large or small values (called outliers) tend to appear. The other has to do with skewness. Situations in which outliers are rare but skewness exists can be a very serious concern. If distributions are perfectly symmetric but outliers tend to appear, any method based on means can have poor power. When distributions are both skewed and outliers tend to appear, all standard hypothesis testing methods based on means can have very poor properties.

Skewed Distributions

To elaborate in a reasonably concrete manner, imagine a study dealing with cognitive functioning among children as measured by a test of memory function that ranges from 0 to 100 with higher scores indicating better memory. Further imagine that 20 children are randomly sampled and that the investigator wants to test if the population mean is above or below the test norm that defines a clinically meaningful level of cognitive impairment. If the mean is 30, then the investigator might conduct a traditional one sample, Student's t test based on the sample mean for the 20 children, which will be labeled $\overline{X}$. That is, we compute $T = \sqrt{n}(\overline{X} - 30)/s$, where n is the sample size and s is the usual standard deviation.

Now imagine that we repeat the study many times, each time randomly sampling 20 participants. We can depict this as follows.

Study:	1	2	3	…
Means:	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	…
T:	T_1	T_2	T_3	…

Using Student's t is based on the assumption that a plot of the values $T_1, T_2, \ldots$ follows a t distribution with $n - 1$ *df*. The central limit theorem roughly says that if the sample size used to compute each sample mean is not too small, and if we were to plot the sample means just shown, we will get a curve that is approximately normal. A variation of this theorem tells us that the same is true about the t values. A practical problem is determining how large the sample size must be to assume normality, but there is no mathematical result that provides a precise answer. Early investigations focused on very light-tailed distributions where outliers are extremely rare. It was found that even with a sample size of 20 to 25 observations, a plot of the sample means is indeed well approximated by a normal curve. Some books written by statisticians more than 40 years ago implicitly assumed that, as a result, plots of the t values would be approximately normal as well, but more modern investigations reveal that this is not true (e.g., Westfall & Young, 1993).

Imagine, for example, that test scores follow the (lognormal) distribution shown in Figure 1, which is known to be relatively light tailed (e.g., Gleason, 1993). The general shape of this distribution is not uncommon for many variables studied by clinical child psychologists. Figure 2 shows a plot of 5,000 t values generated on a computer with each t based on 20 randomly sampled observations. The smooth symmetric curve about zero is the plot we obtain under normality. Notice that the tails of the two curves differ substantially. Westfall and Young (1993) showed that, as a result, when using Student's t, we get poor control over the probability of a Type I error and inaccurate confidence intervals. Moreover, an undesirable power property can occur: We are more likely to reject when the null hypothesis is true versus when it is false. (For more details and illustrations, see Wilcox, 2001, pp. 80–85.) With a sample size of $n = 160$, and when sampling from the distribution in Figure 1, problems with controlling the probability of a Type I error are not completely corrected (Westfall & Young, 1993), but with $n = 200$, good control over the probability of a Type I error can be had in this particular situation when testing at the .05 level. (For theoretical results in-

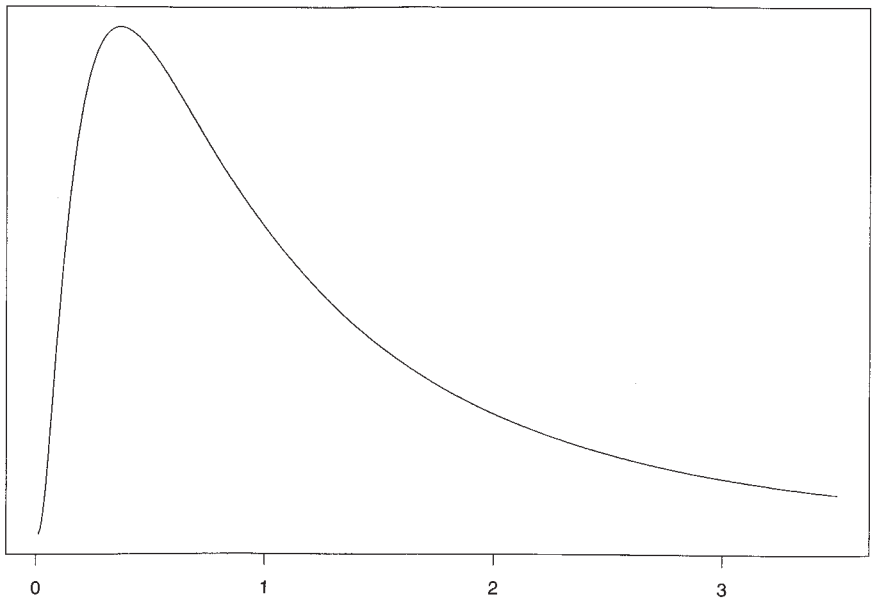

Figure 1. An example of a skewed, relatively light-tailed distribution where the expected number of outliers is small versus the sample size.

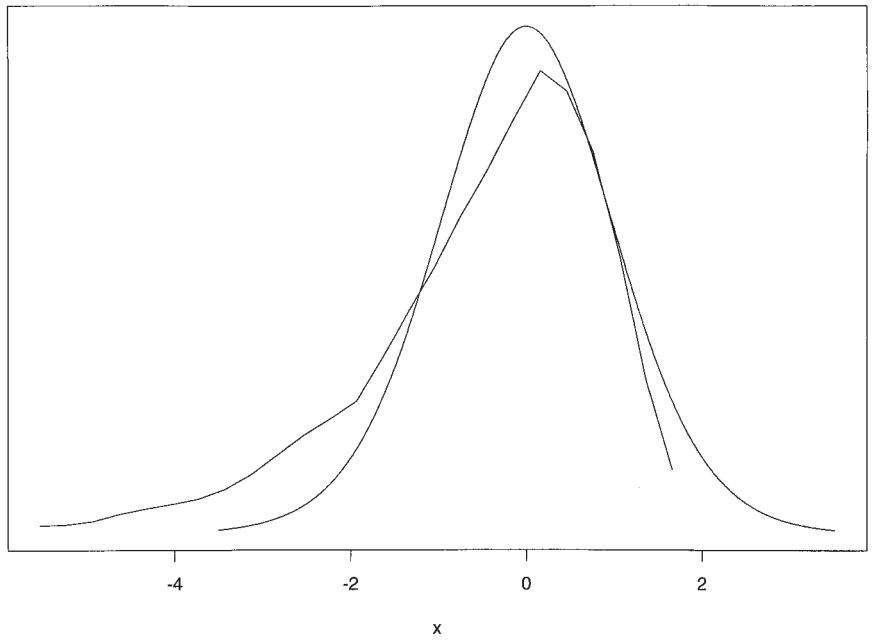

Figure 2. The approximate distribution of *t* when sampling from the distribution in Figure 1, *n* = 20. The smooth symmetric curve is the distribution of *t* under normality.

dicating that Student's *t* can have extremely poor properties, even with a much larger sample size; see Basu & DasGupta, 1995.)

One way of trying to defend Student's *t* is to suggest that the illustration in Figure 2 represents an extreme case that never occurs in practice, but empirical investigations indicate that the problems illustrated by Figure 2 are underestimated. For example, based on $n = 104$ participants where the goal was to investigate the sexual attitudes of young adults, Wilcox (in press) found that Student's *t* is approximately shaped as shown in Figure 3. So when using *t*, the applied researcher is assuming that with probability .95, *t* will have a value between –1.98 and 1.98, but the data suggest that, in reality, *t* will have a value between –5.57 and 1.49 instead. In practical terms, poor control over the probability of a Type I error will be obtained, and inaccurate confidence intervals will result. In some situations, *t* breaks down even with $n = 300$ (e.g., Wilcox, in press). Moreover, Figure 3 underestimates problems because in reality the sample size was $n = 105$ but an extreme outlier was removed. If the extreme outlier is included,

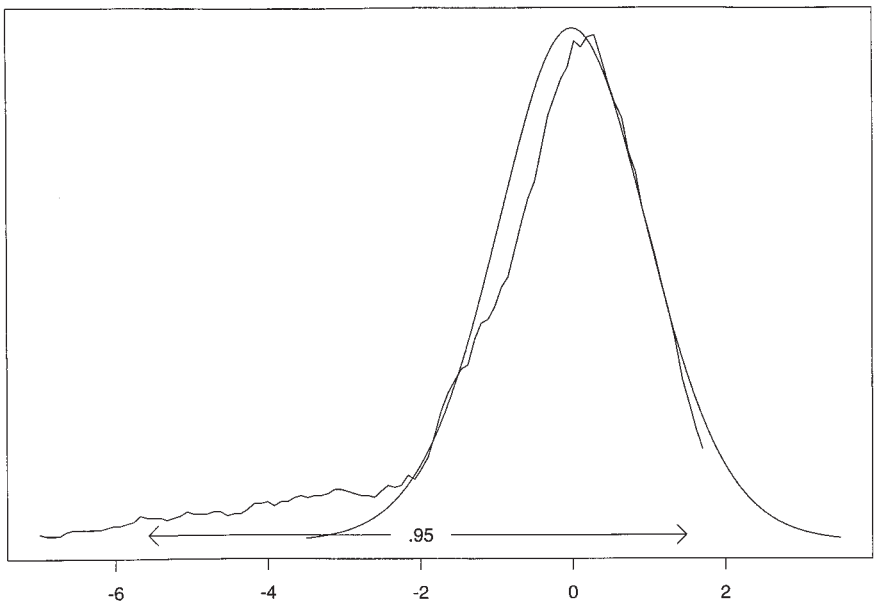

Figure 3. An approximation of the distribution of *t* based on data from a study dealing with the sexual attitudes of young adults.

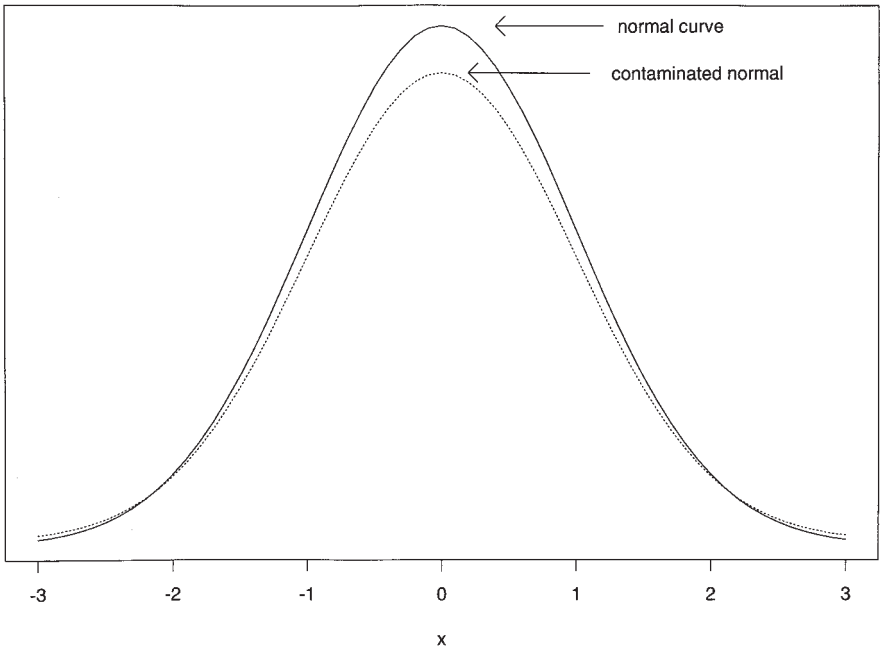

Figure 4. Normal and contaminated normal distributions. The solid line is a standard normal distribution and the dashed line is a contaminated normal. The normal curve has variance 1 but the contaminated normal has variance 10.9.

the sample mean is 65. If we resample observations with replacement and use Student's *t* at the.05 level to test the hypothesis that the mean is greater than 65, the actual probability of a Type I error is .22.

Symmetric Distributions

The problems just described diminish as we move toward symmetric distributions, but even when sampling from a perfectly symmetric distribution, another very serious problem occurs, even under arbitrarily small departures from normality (e.g., Staudte & Sheather, 1990). The problem stems from the fact that very slight changes in the tails of any distribution, including normal distributions as a special case, can greatly inflate the variance, which results in poor power when using any method based on means relative to alternative methods that have been developed in recent years. The classic example stems from Tukey (1960) and is based on a so-called contaminated normal distribution. Figure 4 shows a standard normal distribution with mean zero and variance one and a particular con-

taminated normal, the construction of which is not particularly important here. The important point is that although the contaminated normal appears to be very similar to the normal curve (and it is very similar using standard mathematical methods for measuring the difference between two distributions), its variance is 10.9, nearly 11 times larger than the variance of the standard normal curve!

To illustrate an important implication, imagine that 25 observations are sampled from each of the two groups depicted in the left panel of Figure 5. The distributions are normal, and comparing means with Student's *t* at the.05 level, power (the probability of rejecting) is .96. But if we sample from the two distributions shown in the right panel, power is only .28!

Figure 5 illustrates yet another important point. There has been an increasing awareness and demand for reporting effect sizes, but what has become a standard approach to this problem is potentially very misleading, even under very small departures from normality. Cohen (1977) defined a large effect size as something visible to the naked eye. Suppose we measure effect size with a standardized difference, Δ. That is, Δ is the difference between the means divided by the assumed common standard deviation. Then Cohen concludes that for normal distributions, .2, .5, and .8 are small, medium, and large effect sizes, respectively. In the left panel of Figure 5, $\Delta = 1$. But for the right panel Δ is only .3, yet by Cohen's definition the difference between the groups is large. That is, Δ can be a very misleading measure of effect size and can seriously underestimate the degree to which groups differ.

Figure 5 also illustrates why many studies failed to discover that nonnormality can substantially reduce power. Early investigations studied power based on Δ without taking into account the extreme sensitivity of the variance to the tails of a distribution. Based on Δ, the difference between the two distributions in the left panel of Figure 5 is substantially larger versus the right panel, and from this perspective power should be lower for the situation in the right panel when comparing means. But clearly the two situations are similar in some sense, and surely it is desirable to have approximately the same amount of power when sampling from the distributions in the right panel of Figure 5 versus the left. Modern methods achieve this goal.

Why Not Transform the Data?

Often transforming data is recommended when dealing skewed distributions. Two common suggestions are to take logarithms or the square root of every observation, but there are some issues that should be kept in mind. First, transforming data means that an attempt at making inferences about the mean of the original scores has been abandoned. Second, the more obvious transformations do not eliminate the effects of outliers. That is, outliers remain and can still lower power by a substantial amount. Third, if each observation is transformed in the same manner, situations arise where the distribution of the observed scores remains skewed (even when using what is called a Box-Cox transformation).

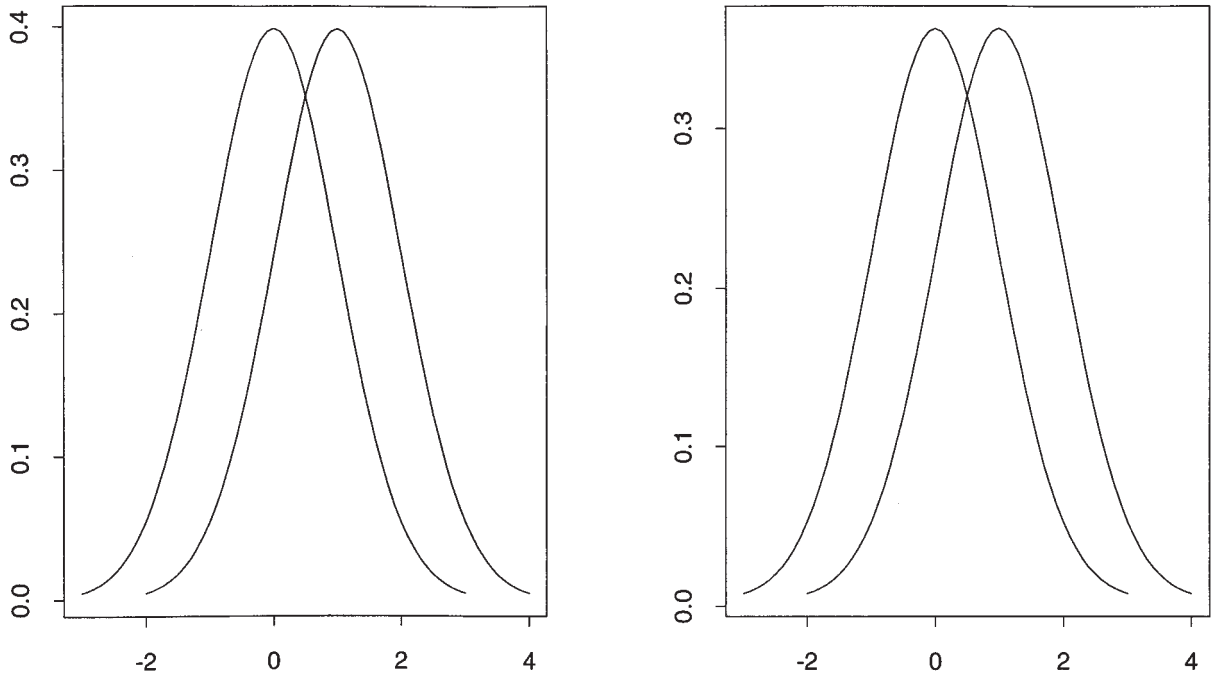

Figure 5. Power and nonnormality when using Student's *t*. In the left panel, power is .96 but in the right panel, where the distributions are contaminated normals, it is only .28.

Heteroscedasticity (Unequal Variances)

The problems with nonnormality just described can be addressed in a very effective manner, but before continuing, there is another issue that should be mentioned that has to do with an important difference between modern methods and conventional techniques. Conventional methods use the correct standard error when groups do not differ (meaning they have identical distributions), and in particular the variances are equal. But under general circumstances conventional methods use the wrong standard error, which can result in poor power, undesirable power properties (meaning power can actually decrease as the difference among the means increases), and inaccurate confidence intervals, even under normality. In contrast, modern methods use the correct standard error regardless of whether groups differ, and this can increase accuracy (when computing confidence intervals) and provide better power.

Most readers are probably familiar with a classic paper by Box (1954), which seemed to suggest that the ANOVA F is insensitive to violations of the equal variance assumption. When comparing two groups only, both having normal distributions, and when the sample sizes are equal, it can be shown that good control over the probability of a Type I error is achieved accept with very small sample sizes (Ramsey, 1980). But with unequal sample sizes, problems can occur under very realistic situations, and as the number of groups being compared increases, problems arise even when using equal sample sizes. When sampling from nonnormal distributions, problems get much worse, and even when comparing two groups with equal sample sizes, control over the probability of a Type I error can be unsatisfactory.

The statements just made do not contradict results reported by Box (1954). Box's numerical results were limited to normal distributions in which, among J groups, the largest standard deviation divided by the smallest is less than or equal to $\sqrt{3}$. If this ratio is a bit larger, the ANOVA F no longer performs well under normality (except when comparing two groups with equal sample sizes), and nonnormality results in even more serious problems as pointed out in numerous articles summarized in Wilcox (1987, 1997, in press). Glass, Peckham, and Sanders (1972) appear to be the first to realize that there are practical problems, and many articles have since confirmed that unequal variances can result in poor control over the probability of a Type I error.

A basic requirement of any method is that, as the sample sizes increase, the method should converge to the correct answer under random sampling. For instance, if the claim is that with probability .95, a confidence interval will contain the true difference between the means ($\mu_1 - \mu_2$), or that the probability of a Type I error is.05, then the accuracy of these claims should increase as we increase the number of observations. Modern methods that do not assume equal variances achieve this goal, but Cressie and Whitford (1986) described general circumstances under which Student's t does not.

A natural suggestion for salvaging methods that assume equal variances is to test the assumption of equal variances, and, if we fail to reject, proceed with a standard technique. Popular commercial software contains methods for easily implementing this strategy. However, this approach is known to fail (e.g., Markowski & Markowski, 1990; Moser, Stevens, & Watts, 1989; Wilcox, Charlin, & Thompson, 1986). The reason is that methods that test the hypothesis of equal variances do not have enough power to detect situations in which the variances differ enough to cause practical problems—even under normality.

Yet another way of trying to address problems with unequal variances is to use error bars, a method that is employed with increasing frequency in some areas of psychological research. The strategy is to estimate the mean and standard deviation of the first group and determine the interval corresponding to one standard error from the mean. The same process is applied to the second group, and if the resulting intervals do not overlap, the temptation is to declare that the population means differ. Graphs of these intervals would seem to provide a convenient summary of the data, but, unfortunately, this simple approach violates basic principles and has terrible properties—even under normality. It can be shown that when comparing the two groups, the wrong standard error is being used and the method does not control the probability of a Type I error (e.g., Wilcox, in press, section 8.4). A more detailed analysis is given by Schenker and Gentleman (2001). And even if this problem could be ignored, problems with nonnormality, already discussed, remain a serious concern.

In sum, there is little separating conventional homoscedastic methods versus heteroscedastic methods when comparing groups having identical distributions, so in particular the assumption of equal variances is true. But when groups differ, contemporary heteroscedastic techniques that satisfy basic theoretical requirements (meaning they use a correct expression for the standard error) have considerable practical value.

Interpreting the ANOVA F and Tests of Linear Contrasts

Despite the practical problems with the ANOVA F test and related techniques, these methods do tell us something about our data when comparing two or more groups. Consider two groups and for the first group let $F_1(x)$ be the probability that a randomly sampled participant has an observed value that is less than

or equal to x. So, for example, $F_1(2)$ is the probability that a randomly sampled observation is less than or equal to 2, and $F_1(6)$ is the probability that an observed value is less than or equal to 6. In a similar manner, for Group 2, $F_2(x)$ is the probability that an observed value is less than or equal to x. The two distributions are said to be identical if for any x we might pick, $F_1(x) = F_2(x)$. That is, if we were to graph the distributions, they would be indistinguishable, so in particular they would have equal means, equal variances, the same amount of skewness, and so on.

Now consider comparing two groups with Student's t when the groups have identical distributions. The hypothesis of equal means is true, and if equal sample sizes are used, it can be shown that the distribution of the difference between the sample means is symmetric about zero. Translation: Theory, empirical studies, and simulations indicate that as far as avoiding Type I error probabilities well above the nominal level, Student's t should work reasonably well when sampling from nonnormal distributions having identical distributions. So, for example, if we test at the .05 level, the actual probability of a Type I error will be approximately.05 or less. In fact, from a modern perspective, the research conducted during the 1950s on the effects of nonnormality on the ANOVA F was concerned with this special case, and even with unequal sample sizes it seems that the probability of a Type I error can be controlled. The empirical study by Sawilowsky and Blair (1992) is sometimes cited as justification that Student's t performs well under nonnormality, but they focused on situations where groups have identical distributions and did not consider situations where the distributions differ. It is when distributions differ that practical problems with Student's t emerge. Said another way, it appears that in terms of controlling the probability of a Type I error, Student's t provides a satisfactory test of

$$H_0 : F_1(x) = F_2(x) \tag{1}$$

for any x. That is, it tests the hypothesis that two groups have identical distributions. It is designed to be sensitive to differences between the means, but in reality it is sensitive to many features associated with the distributions, including variances and skewness. Moreover, as soon as we conclude that distributions differ, assessing the magnitude of the difference between the means becomes fraught with peril.

If we agree that when distributions differ, in general the means differ as well, then it is perfectly legitimate to conclude that the means differ when a significant result is obtained with Student's t. In theory, distributions can have different shapes with equal means, but some would argue that the likelihood of this happening in practice is so small that this possibility can be ignored. What modern investigations tell us is that, although arguments for concluding that the means differ can be made, the main reason for obtaining a significant result is unclear. That is, the primary reason might not be due to differences between the means but rather unequal variances, differences in skewness, or some other difference between the distributions that we have not considered.

Perhaps one more point should be stressed. Failure to reject with any conventional method for means is not remotely convincing evidence that the null hypothesis of equal means should be accepted. When groups differ, all conventional methods can have poor power, making it virtually impossible to detect a substantively important difference. Of course, exceptions arise, but determining whether it is safe to trust inferences about means is a nontrivial problem, and the more obvious methods have proven to be unsatisfactory.

A Summary of Modern Insights Regarding Conventional Techniques

The main points regarding modern insights into conventional ANOVA methods can be summarized as follows:

- A significant result indicates that distributions differ, and arguments can be made that, by implication, the means differ as well.
- Once we conclude that distributions differ, an accurate assessment of the magnitude of the difference between the population means can be difficult and may be even impossible. In addition, the main reason for rejecting might not be due to differences among means, but rather some other difference among the distributions such as unequal variances or differences in skewness. Modern methods (to be described) help address these concerns.
- Arbitrarily small departures from normality can result in very poor power when using any method based on means, relative to modern methods to be described, and the power of conventional ANOVA methods can be reduced substantially when there is skewness or heteroscedasticity. So when using conventional ANOVA methods, failure to reject does not imply that the null hypothesis should be accepted.
- Recent studies finding practical problems with nonnormality do not contradict earlier studies where no problems were found. Theory indicates that for groups with nonnormal but identical distributions, practical problems should be minimal, and this view is supported by various journal articles. Theory also indicates that distributions with nonnormal and nonidentical distributions cause problems, and more recent articles illustrate that very serious concerns do indeed occur. Early studies that greatly influence current prac-

tice did not consider the effects of nonidentical distributions.

- A rough rule is that as we move toward more complicated designs, the more sensitive conventional methods become to violations of standard assumptions.
- The practical problems with conventional ANOVA methods have been known for many years, and contemporary techniques can make a substantial difference.

During the past 40-plus years, three major advances have come together that make it possible to effectively address the problems with conventional methods just outlined: improved inferential methods, the theory of robustness (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990), and fast computers. The improved inferential methods available today consist of two key components. The first is the ability to handle unequal variances, and the second is the ability to reduce other serious problems associated with nonnormality. This latter component consists of using some form of what is called a bootstrap method. Bootstrap methods that allow unequal variances certainly offer a practical advantage, but certain problems cannot be eliminated when working with means. However, combining modern inferential techniques with robust methods results in vastly more accurate inferences, and in commonly occurring situations the increase in power can be very substantial. In practical terms, highly nonsignificant results can become significant when using a more accurate method developed during the past 20 years. Said another way, if groups do not differ in any manner (they have identical distributions), modern methods offer no practical advantage, but if groups differ, standard techniques can break down. The practical problem is that we do not know whether groups have identical distributions, so the argument for contemporary methods is that they are designed to perform well in both situations.

Comments on Outliers

As already noted, situations where outliers occur can substantially reduce power when using any method based on means. Based on traditional training, it might seem that outliers are rare, but Tukey (1960) predicted that the exact opposite is true, and modern outlier detection methods support his view.

Detecting and Dealing with Outliers

A popular and seemingly natural method for detecting outliers is to declare the value X an outlier if it is more than two standard deviations from the mean. For example, the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 50 have a mean of $\bar{X} = 10.4$, a standard deviation of $s = 14.15$, and because $|50 - 10.4| > 2 \times 14.14$, declare 50 to be an outlier. However, this outlier detection method has long been known to be unsatisfactory because it suffers from what is called masking. To illustrate the problem, consider 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 standard deviation, but they have more of an effect on the standard deviation, causing outliers to be missed. What is needed is an outlier detection method that is not itself affected by outliers. Such methods have been developed, the best known being the boxplot. (See Barnett & Lewis, 1994, for a broad treatment of how to detect outliers. Recent improvements on the boxplot, plus some other methods that have considerable practical value, are described in Wilcox, in press.) The important point here is that when these methods are applied to data, outliers are the rule rather than the exception. But even if no outliers are found, this does not mean that standard methods based on means will have relatively high power because skewness and heteroscedasticity also affect power, as already noted.

Because outliers can lower power, a seemingly natural strategy for dealing with outliers is to discard them and apply methods for means to the data that remain, but from a technical point of view, this strategy fails. The problem is that by discarding extreme values, the remaining observations are no longer independent under random sampling (e.g., Hogg & Craig, 1970; Wilcox, 2001), and this dependence invalidates the derivation of the standard error of the sample mean. From a practical point of view, ignoring this issue can be unsatisfactory. Consider, for example, the data in Wilcox (in press, Table 3.2). If we discard outliers (using a method described later in this article and outlined in the appendix) and then compute the standard error of the mean as if the outliers never existed, we get .134. However, using a (bootstrap) estimate that is theoretically sound when discarding extreme values, we get .505. Generally, discarding outliers and applying standard methods for means to the data that remain eliminates our ability to control the probability of a Type I error. However, there are methods for discarding outliers and getting excellent control over the probability of a Type I error, some of which are outlined in the following.

Robust Measures of Location

Robust measures of location are values intended to reflect the typical response of participants without the negative features associated with means. There are two general types. The first empirically checks for outliers and discards any that are found, but now special methods for testing hypotheses must be used, the most effective of which are based on some type of bootstrap

method. The second is based on so-called trimmed means. Each approach has both advantages and disadvantages. Complete details are impossible here, but some of the main issues are discussed.

Trimmed Means

A trimmed mean is computed by removing a certain percentage of the largest and smallest observations and averaging the values that remain. The proportion of observations trimmed is fixed in advance, which is convenient because it provides a relatively simple method for analyzing data. (A theoretically sound estimate of the standard error that is fairly easy to use can be derived.) The median is a trimmed mean where the maximum possible amount of trimming is used, and the sample mean is a trimmed mean where no trimming is done at all. The term *10% trimming* indicates that 10% of the largest observations, as well as 10% of the smallest observations, are trimmed. If we have 10 observations, the largest value is 35, and the smallest value is 6; 10% trimming consists of removing these two values and averaging the rest.

Methods for comparing multiple groups based on medians are available (Wilcox, in press), but it has long been known that when sampling from a normal distribution, on average the mean beats the median in power by a fair amount. But, as early as 1775, Laplace was aware of conditions where the median beats the mean, and by 1818 he was able to characterize general conditions under which this is true. How might we maintain relatively high accuracy when sampling from light-tailed distributions, including normal distributions as a special case, yet achieve reasonably high accuracy when outliers are common? One approach is to use a compromise amount of trimming, something between the two extremes of no trimming and the maximum amount. What modern insights reveal is that alternatives to the mean and median perform relatively well under normality, but even for very slight departures from normality they can offer a substantial increase in accuracy.

To illustrate this point, 20 observations were sampled from the contaminated normal shown in Figure 4, the mean and 20% trimmed mean were computed, and this process was repeated 5,000 times. Figure 6 shows a plot of the results. As is evident, the 20% trimmed means are more tightly centered around zero. That is, on average, they provide a more accurate estimate of the center of the distribution versus the mean. (For additional details, written at a nontechnical level regarding why trimming increases accuracy, see Wilcox, 2001, pp. 144–146.) Moreover, theory and simulations indicate that 20% trimming deals effectively with problems due to skewness.

As for asymmetric distributions, again it is very common for the trimmed mean to have a smaller standard error than the mean, but now it no longer estimates μ; typically it estimates a quantity closer to the central portion of a distribution. Put another way, the sample mean can provide a poor reflection of what is typical, even with large sample sizes. For example, Wilcox (2001, p. 16) described data from a study in which (with a sample size of 105) 97% of the observations were less than the sample mean. Even with infinitely many observations, the sample mean can be far from the bulk of the observations (e.g., Staudte & Sheather, 1990). In contrast, a 20% trimmed mean will

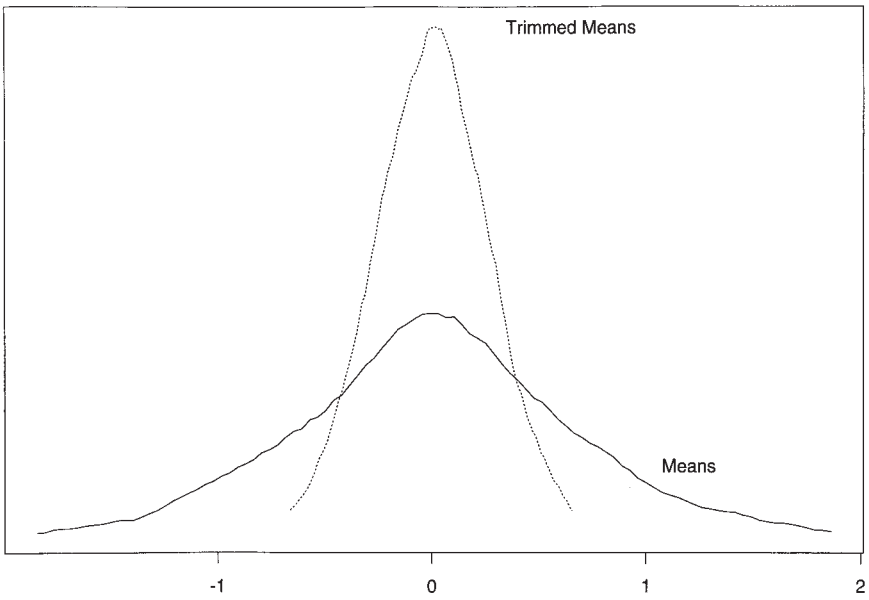

Figure 6. A plot of 5,000 trimmed means and 5,000 means, each based on 20 observations randomly sampled from the contaminated normal distribution shown in Figure 4. This illustrates that small departures from normality can substantially reduce the accuracy of the mean.

tend to be closer to the bulk of the observations. Insofar as we want a measure of location to reflect what is typical, this is desirable.

Hypothesis Testing Based on Trimmed Means

Given the goal of comparing the typical response among one group versus another, all methods based on means are readily extended to trimmed means. The first hypothesis-testing method based on a trimmed mean, which was limited to a single group, was derived by Tukey and McLaughlin (1963). A heteroscedastic extension to two groups was studied by Yuen (1974), and Yuen's method is readily extended to any situation where the goal is to compare more than two groups (Wilcox, 1997, in press). Methods for comparing dependent groups are available as well. All of these methods use a theoretically correct estimate of the standard errors. The computations are easy to do, but they are not obvious based on traditional training. (See Wilcox, 1997, in press, for details.)

In theory, the sample mean can be highly inaccurate relative to a 20% trimmed mean, but can it really make a difference which estimator is used when working with real data? Consider data from a self-awareness study (viz., Wilcox, 2001, p. 83). For the first group, the estimated standard error of the sample mean was 136 versus 56.1 for the 20% trimmed mean. This means that a confidence interval based on a 20% trimmed means will have less than half the length of the confidence interval based on Student's *t*. So using trimmed means can result in much higher power. In this particular study, there were two groups, and the usual Student's *t* test had a significance level of.47, but Yuen's (1974) test for trimmed means had significance level .052. Moreover, theory and simulations indicate that Yuen's test provides good control over the probability of a Type I error over a broader range of situations versus any method based on means. Experience suggests that in some cases the standard error of the sample mean will be a bit smaller than the standard error of a 20% trimmed mean, but the sample mean seems to rarely if ever offer a substantial advantage, and it is fairly common for trimmed means to have substantially smaller standard errors.

Estimators That Discard Outliers

Both theory and simulation studies indicate that trimming considerably reduces problems due to skewness and outliers. But despite the advantages of 20% trimmed means, it suffers from at least two practical concerns. First, situations arise where the proportion of outliers exceeds 20%, meaning that, in terms of power, more trimming or some other measure of location—that is relatively unaffected by a large proportion of outliers—is needed. Second, if a distribution is highly skewed to the right, say, then at least in some situations it seems reasonable to trim more observations from the right tail versus the left.

An alternative to fixing the amount of trimming is to remove values that are flagged as outliers, but, as already noted, we need an outlier detection rule that is not subject to masking. Among robust methods, one popular outlier detection technique is based on the median and a measure of variation called the median absolute deviation (MAD) statistic. Under normality, MAD/.6745 estimates the standard deviation, σ, which suggests a method for deciding whether X is an outlier. (The computation of MAD and an illustration of how it is used to detect outliers is relegated to the appendix.) If we discard outliers and average the values that remain, we get what is called a *modified one-step M-estimator* (MOM). MOM is very similar to what are called *skipped estimators*, which were proposed by Tukey.

Some advantages of MOM are that it can handle a large number of outliers, it performs well under normality, and it contains the possibility of using the mean. Moreover, excellent control over the probability of a Type I error can be had in situations where no method based on means has been found to perform in a satisfactory manner. An inconvenience is that when testing hypotheses, a computer-intensive bootstrap method is required. But in the age of the computer, this would seem to be a minor concern. As with trimmed means, MOM can be used in all of the commonly encountered experimental designs (Wilcox, in press).

Outliers Can Be Interesting

Can arguments be made that extreme values are informative and that we want to know something about the mean even if it lies in the tails of a distribution? Yes, in some cases such an argument can be made. The thing to keep in mind is that making accurate inferences about the mean can be very difficult unless sample sizes are extremely large. Special methods for making inferences about the tails of distributions are available (e.g., Wilcox, 1997, in press) and might be considered.

Basic Bootstrap Methods

Both homoscedastic and heteroscedastic hypothesis-testing methods are based on basic principles developed by Laplace about two centuries ago. About a century later, small sample size adjustments began to appear, which are routinely used today. With the realization that these methods can be inaccurate, a fundamental issue is whether alternative inferential methods

can be found that have both theoretical and practical advantages, particularly when the sample sizes are small. Based on hundreds of published articles, the answer is that such methods exist and consist of some variation of a so-called bootstrap method. Bootstrap methods contain two basic forms that are particularly important in applied work. There are many variations of these methods, but they go beyond the scope of this article. Readers interested in these variations and book-length descriptions of bootstrap techniques are referred to Chernick (1999), Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall and Hall (1995), Lunneborg (2000), Mooney and Duval (1993), and Shao and Tu (1995).

All bootstrap methods are based on what are called bootstrap samples. To be concrete, imagine that the goal is to test the hypothesis H_0: $\mu = 20$, and that we get the following observations:

2, 4, 6, 6, 7, 11, 13, 13, 14, 15, 19,
23, 24, 27, 28, 28, 28, 30, 31, 43.

The sample size is 20 and the sample mean is $\overline{X} = 18.6$. A bootstrap sample is obtained by randomly resampling *with replacement* 20 observations from the observed scores. More generally, with n observations, we resample n observations with replacement. For the data at hand, a bootstrap sample might be

14, 31, 28, 19, 43, 27, 2, 30, 7, 27,
11, 13, 7, 14, 4, 28, 6, 4, 28, 19

(which was generated on a computer from the original observations). The mean of this bootstrap sample is typically labeled $\overline{X}^*$ to distinguish it from $\overline{X}$, the mean based on the original scores. Here, $\overline{X}^* = 18.1$.

Percentile Bootstrap

The first of the two basic bootstrap methods is applied as follows. Notice that we can repeatedly generate bootstrap samples yielding a collection of bootstrap sample means. For illustrative purposes, assume 1,000 bootstrap sample means have been generated, and notice that a certain proportion of them will have a value greater than the hypothesized value, 20. Label this proportion $\hat{p}^*$ and let $\hat{p}^*_m$ be equal to $\hat{p}^*$ or, $1 - \hat{p}^*$, whichever is smaller. Then $2\hat{p}^*_m$ is the estimated significance level. So, for example, if the proportion of bootstrap sample means greater than 20 is $\hat{p}^* = .9$, the estimated significance level is $2(1 - .9) = .2$. The percentile bootstrap method can be extended to comparing multiple groups as described in Wilcox (in press), and it can be used with any of the robust estimators described here.

The Bootstrap-*t* Method

The other basic bootstrap technique is called a bootstrap-*t* (and sometimes a percentile-*t*) method. In the simplest case, when dealing with a single group, the goal is to use the data at hand to estimate the actual distribution of Student's t versus approximating the distribution based on the assumption that observations are randomly sampled from a normal distribution. Consider the problem of testing H_0: $\mu = \mu_0$, where μ_0 equals some specified constant, and suppose the goal is to have a Type I error probability equal to .05. Given some data, ordinarily we compute

$$T = \sqrt{n}(\overline{X} - \mu_0)/s,$$

and if T is sufficiently small or large, based on critical values read from a table of Student's t distribution, we reject. But this method can be highly inaccurate under nonnormality—the wrong critical values are being used—and the bootstrap-*t* attempts to deal with this problem by estimating a more accurate critical value based on the data. In particular, we generate a bootstrap sample as in the percentile method and compute the mean and standard deviation, which are labeled $\overline{X}^*$ and s^*. Let

$$T^* = \sqrt{n}(\overline{X}^* - \overline{X})/s^*.$$

For illustrative purposes, assume this process is repeated 1,000 times yielding 1,000 T^* values. Put these 1,000 values in ascending order. If T falls outside the middle 95% of the 1,000 T^* values, reject H_0.

The Percentile Bootstrap Versus the Bootstrap-*t*

Of course, a practical issue is determining when the bootstrap-*t* should be preferred over the percentile bootstrap. When comparing groups based on means, all indications are that the bootstrap-*t* is preferable (e.g., Westfall & Young, 1993). But when using measures of location relatively insensitive to outliers, the percentile bootstrap has distinct advantages (Wilcox, in press). For example, when using trimmed means with at least 20% trimming, generally a percentile bootstrap method beats the bootstrap-*t*. And when using measures that empirically check for outliers and discard any that are found (such as MOM), currently some variation of the percentile bootstrap is a must.

The Accuracy of Bootstrap Methods

When first encountered, there is no particular reason to suspect that in terms of accuracy the bootstrap-*t* offers a practical advantage over the conventional Student's t test, but all indications are that this is the case. And in the event sampling is from a normal distribution, the bootstrap-*t* performs nearly as well as Stu-

dent's t. For example, with $n = 20$, and when sampling from a normal distribution, the actual probability of Type I error when testing at the .05 level is approximately .054. When sampling from the (lognormal) distribution shown in Figure 1, the actual probability of Type I error when using Student's t is .14, but switching to the bootstrap-t, it is only .078. So some would say that even the bootstrap-t is not completely satisfactory in this case, but it is substantially better than Student's t. Where problems get especially serious when using any method based on means is when sampling from a skewed, heavy-tailed distribution. For example, Wilcox (in press, section 7.4) describes a situation in which with $n = 20$, when using Student's t at the .05 level, the actual probability of a Type I error is .202. Switching to the bootstrap-t, it is.198, not much better. Increasing the sample size to 100, these values become .198 and .168, so the situation improves as the sample size increases. The bootstrap-t offers an advantage, but both methods remain unsatisfactory. For the situations considered here, the percentile bootstrap method performs poorly.

Now consider what happens when using a 20% trimmed mean instead. With $n = 20$, and using the Tukey and McLaughlin (1963) method, the actual probability of a Type I error when sampling from the same heavy-tailed distribution is .020, in contrast to .202 when using Student's t. The bootstrap-t can be extended to 20% trimmed means, and for the situation at hand, the probability of a Type I error is .014. Using the percentile bootstrap with a 20% trimmed mean, it is .066. This illustrates the more general finding that when using a percentile bootstrap with 20% trimmed means, the actual probability of a Type I error tends to be closer to the nominal level compared to many competing methods, including all methods based on means. Moreover, when using an extension of the percentile bootstrap method to compare multiple groups with 20% trimmed means, good control over the probability of a Type I error can be achieved over a very broad range of situations where all methods based on means have been found to be unsatisfactory (e.g., Wilcox, in press).

Software

Easy-to-use software for both modern robust methods and bootstrap techniques is available. The free software described in Wilcox (in press) includes most of the software in Wilcox (1997), plus software for many new methods that have not appeared in any other books; it can be downloaded at www-rcf.usc.edu/~rwilcox/. Software that links these S-PLUS functions to SPSS is being developed by James Jaccard at the State University of New York, Albany. (See his web page at www.zumastat.com.) For SAS software aimed at trimmed means, see Keselman, Wilcox, and Lix (2001); their software can be obtained from www.umanitoba.ca/faculties/arts/psychology. Virtually any experimental design based on means can now be extended to robust estimators, including a between by within design and analysis of covariance.

Rank-Based Methods

A common recommendation for dealing with nonnormality is to switch to a rank-based method. There are two general types that are important here. To describe the first, consider two independent groups and let p be the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. The Wilcoxon–Mann–Whitney test is based on a direct estimate of p (but often this is not made clear) and is aimed at testing H_0: $p = .5$. The other type is based on testing the hypothesis of identical distributions, and typically this is done using the average ranks among groups. Heteroscedastic analogs of standard methods are now available (e.g., Brunner, Domhof, & Langer, 2002; Cliff, 1996; Wilcox, in press) and are recommended over more conventional techniques. When groups differ, typically there are general conditions where conventional methods use the wrong standard error that can affect power.

As for choosing between rank-based methods versus robust measures of location, each provides a different perspective on how groups differ. Generally, rank-based methods tell us nothing about how participants compare in terms of a typical response reflected by some robust measure of location. Simultaneously, methods based on robust measures of location generally do not reflect the information conveyed by rank-based techniques. That is, at some level, comparisons are meaningless. As for maximizing power, it seems that often robust measures of location are best, but the only certainty is that exceptions occur, one reason being that by design, they are sensitive to different features regarding how groups compare. A positive feature of rank-based methods is that they guard against low power due to outliers. To see why, consider the values 2, 6, 7, 9, and 10. Then 10 has a rank of 5, and if the value 10 is increased to 1 million, its rank remains 5. That is, the variation among the ranks is unchanged. Some quantitative authorities argue that rank-based methods should be used to the exclusion of all other approaches, including methods based on means, but experts are not in agreement about this issue. (See Cliff, 1996, chapter 1, for a discussion of why some quantitative experts prefer rank-based methods.) Even if measures of location are used as the main tool for comparing groups, modern rank-based techniques would seem

to be very useful when characterizing how groups differ and by how much.

A Summary of Conventional Versus Contemporary Methods

Now we consider three overall characterizations of conventional versus contemporary methods. First, significant results based on conventional methods tend to remain significant when using modern robust methods, although exceptions are occasionally encountered by the author. Second, when groups do not differ in any manner, meaning they have identical distributions, modern techniques offer little or no advantage over standard methods. The problem is, of course, that we do not know whether groups have identical distributions. All indications are that if groups differ, modern technology can be very useful. Third, currently there is only one diagnostic tool for determining whether contemporary technology makes a difference: applying modern methods. Some years ago the author tried a variety of diagnostic tools aimed at avoiding contemporary robust methods and relying on methods for means only. None of the methods performed well and therefore were never published.

Concluding Remarks

Many issues and techniques have not been described, but hopefully the idea has been conveyed that modern ANOVA methods have great practical value. Not only do contemporary methods offer more power, they provide new perspectives on how groups of participants differ and by how much, and the precision of our estimates can be assessed much more accurately than ever before. Generally, we have the technology for improving our understanding of important issues related to clinical child and adolescent psychology.

References

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

Basu, S., & DasGupta, A. (1995). Robustness of standard confidence intervals for location parameters under departure from normality. *Annals of Statistics, 23,* 1433–1442.

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.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide.* New York: Wiley.

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

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

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge, England: Cambridge University Press.

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

Glass, G., Peckham, P., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis 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.

Hall, P. G., & Hall, D. (1995). *The bootstrap and edgeworth expansion.* New York: Springer-Verlag.

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

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2001). A generally robust approach to hypothesis testing in independent and correlated groups designs. Unpublished technical report, Dept. of Psychology, University of Manitoba.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Duxbury.

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

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

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.

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

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.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55,* 182–186.

Shao, J., & Tu, D. (1995). *The jackknife and the bootstrap.* New York: Springer-Verlag.

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

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448–503). 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/winsorization 1. *Sankhya A, 25,* 331–352.

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

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

Wilcox, R. R. (in press). *Applying contemporary statistical methods.* 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.

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

Received January 28, 2002
Accepted March 28, 2002

APPENDIX

Detecting Outliers Using the Median and MAD

First compute the median, M, of the randomly sampled observations, $X,\ldots,X_n$. Next, compute the median of the values $|X_1 - M|,\ldots,|X_n - M|$. The result is a measure of scale called the MAD statistic, which plays a prominent role in modern robust methods. Then label the value X_i an outlier if

$$\frac{|X_i - M|}{MAD/.6745} > 2.24,$$

where 2.24 was chosen to get good results under normality. (This outlier detection rule is a special case of a general multivariate method studied by Rousseeuw & van Zomeren, 1990.)

To illustrate the computations, consider again the values 2; 2; 3; 3; 3; 4; 4; 4; 100,000; 100,000. As previously shown, the outlier detection rule based on the mean and standard deviation fails to detect any outliers. The median is $M = 3.5$. Subtracting the median from each value and taking the absolute value of this difference yields 1.5; 1.5; 0.5; 0.5; 0.5; 0.5; 0.5; 0.5; 99,996.5; 99,996.5. The median of these 10 values is MAD and is equal to 0.5. So two outliers are detected: 100,000 and 100,000

Computing MOM

The MOM estimate of location consists of checking for outliers using M and MAD as just illustrated, discarding any that are found and then averaging the values that remain. The constant 2.24, used when checking for outliers, is motivated in part by the goal of having a reasonably small standard error when sampling from a normal distribution. In the previous illustration, MOM = 3.125.