How Many Discoveries Have Been Lost by Ignoring Modern Statistical Methods?

Rand R. Wilcox
University of Southern California

Hundreds of articles in statistical journals have pointed out that standard analysis of variance, Pearson product–moment correlations, and least squares regression can be highly misleading and can have relatively low power even under very small departures from normality. In practical terms, psychology journals are littered with nonsignificant results that would have been significant if a more modern method had been used. Modern robust techniques, developed during the past 30 years, provide very effective methods for dealing with nonnormality, and they compete very well with conventional procedures when standard assumptions are met. In addition, modern methods provide accurate confidence intervals for a much broader range of situations, they provide more effective methods for detecting and studying outliers, and they can be used to get a deeper understanding of how variables are related. This article outlines and illustrates these results.

All psychologists are taught the standard *t* test, the analysis of variance (ANOVA) *F* test, Pearson product–moment correlation, and least squares regression. When I was in graduate school, I was assured that these methods were robust when distributions were not normal or when groups had unequal variances. In essence, I was led to believe that by the year 1955, all practical problems had been addressed. Looking at various psychology journals, this view still reflects conventional wisdom. However, a more accurate description of standard hypothesis-testing methods is that they are robust when there are no differences.

As hundreds of articles in statistical journals have pointed out and for reasons summarized in several books (e.g., Birkes & Dodge, 1993; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Hoaglin, Mosteller, & Tukey, 1983, 1985; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 1996, 1997a), standard methods are not robust when differences exist or when there is an association between random variables. In particular, arbitrarily small departures from normality result in low power; even when distributions are normal, heteroscedasticity can seriously lower the power of standard ANOVA and regression methods. The practical result is that in applied work, many nonsignificant results would have been significant if a more modern method, developed after the year 1960, had been used. A related and perhaps more serious concern is that standard confidence intervals and measures of effect size can be extremely misleading under small departures from normality. Another advantage of modern techniques is that they provide more effective methods for identifying and studying outliers—outliers being any unusually large or small values among a batch of numbers.

Of course, some statistics books, aimed at psychologists, admit that problems might arise when one is using standard techniques, and they offer advice about how to proceed. Typically, a nonparametric method is suggested. However, most of the standard nonparametric methods are now obsolete. That is, modern rank-based methods have something to offer, but they are not known by the typical researcher trying to keep current in his or her own area of expertise. This problem reflects the more general concern that there is an ever widening gap between modern statistical methods and techniques used by psychologists.

More than 30 years ago, theoretical methods were developed for dealing with problems that arise under small departures from normality. Of particular note is the theory of robustness developed by Huber and Hampel and summarized in Huber (1981) and Hampel et al. (1986). These methods simultaneously deal with problems that are due to small changes in observed values. That is, typical statistics such as the sample mean, the sample variance, the Pearson product–moment correlation, and the least squares estimate of regression parameters can be drastically affected by a single unusual value, and modern robust methods are designed to deal with this problem. Initially, it was unclear how to test hypotheses and to compute confidence intervals by using modern estimators, but today this task is easily done. It turns out that even when there are no outliers, but distributions are skewed, modern methods offer a substantial advantage over standard techniques. Despite this advantage, most

Editor's note. J. Bruce Overmier served as action editor for this article.

Author's note. Correspondence concerning this article should be addressed to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061. Electronic mail may be sent to rwilcox@wilcox.usc.edu.

Vol. 53, No. 3, 300–314

Rand R. Wilcox

applied researchers remain unaware of modern techniques and the practical advantages they offer. Another problem is that most quantitative articles tend to be too technical for applied researchers who do not routinely work with theoretical statistics. My goal in this article is to give a very nontechnical description of the problems that arise and to illustrate that modern methods can make a huge difference in applied work.

In terms of power and accurate probability coverage, standard ANOVA and regression methods are affected by three characteristics of data that are commonly seen in applied work: skewness, heteroscedasticity (unequal variances among groups), and outliers. Each of these characteristics, irrespective of the other two, can substantially diminish the chances of (a) detecting true differences between groups, (b) detecting true associations among random variables, and (c) obtaining accurate confidence intervals for the parameters of interest. Taken together, these three features become a very serious concern. Problems with common measures of effect size arise as well.

Problems With Student's *t* Test

I begin with a seemingly simple problem: Choose a measure intended to reflect the typical person under study and compute a confidence interval for it. Of course, the usual measure is the population mean (μ), the average of all the individuals if only they could be measured. One cannot measure all persons that are of interest, so one estimates the population mean with the sample mean (M). The standard $1 - \alpha$ confidence interval for the population mean is

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

where SD is the sample standard deviation based on a random sample of n participants, and $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of Student's t distribution with $n - 1$ degrees of freedom. From basic principles, if the goal is to test the null hypothesis that $\mu = 10$, say, then the null hypothesis would be rejected if the confidence interval given by Equation 1 does not contain the hypothesized value of 10. Two serious problems arise when this method is applied to data, namely, heavy-tailed distributions and skewness, and new problems are introduced when attention is turned to comparing two or more groups. A particularly serious problem is having unequal variances that can lower power.

First, consider how heavy-tailed distributions and outliers affect power—the probability of rejecting when in fact the null hypothesis is false. Power is related to the variance of the sample mean, which is σ^2/n, where σ^2 is the population variance. Of course, one estimates the population variance with the sample variance (s^2). From basic principles, if the population variance is known, then the confidence interval for the population mean becomes

$$M \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \qquad (2)$$

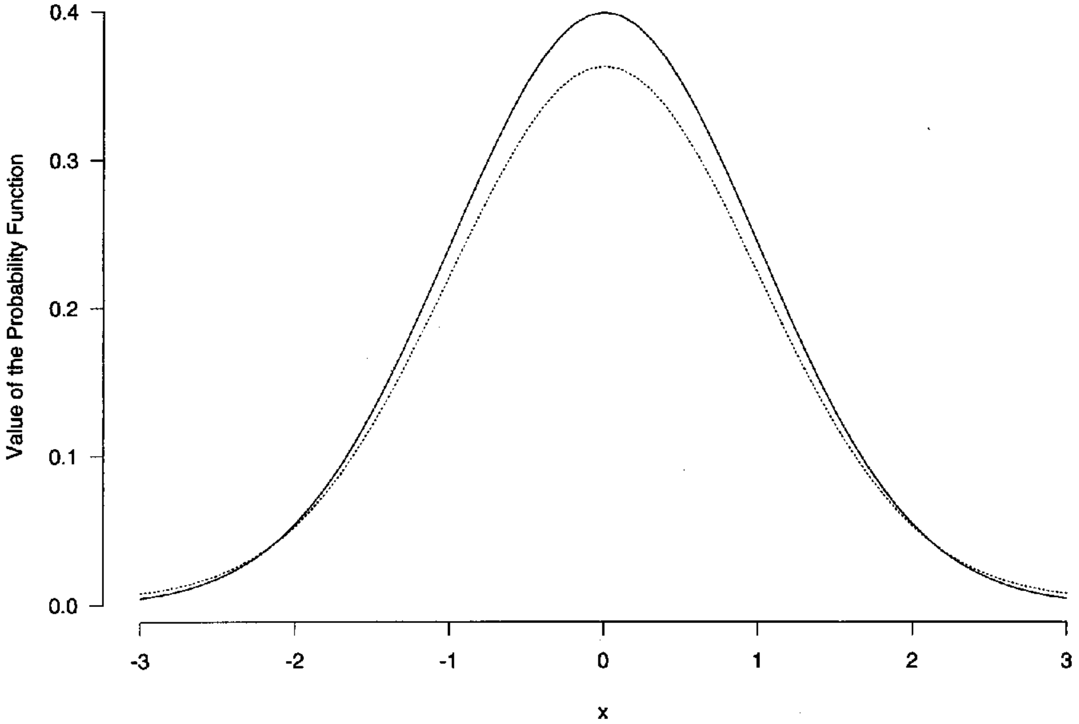
where z is the $1 - \alpha/2$ quantile of a standard normal distribution. As most introductory books on statistics point out, as the population variance goes up, power goes down.

Temporarily assume that observations are randomly sampled from a symmetric distribution (of course, asymmetric distributions are important, but I focus on one problem at a time). When one is sampling from a normal distribution, power is best when one uses the sample mean in conjunction with Equation 1 or Equation 2. The reason is that among all of the unbiased estimators of the population mean, none has a smaller variance than the sample mean.

I now illustrate how small departures from normality can substantially lower power. The classic illustration is based on the mixed or contaminated normal distribution. Let X be any random variable. Suppose that for a randomly sampled participant, there is a .9 probability that an observation comes from a standard normal distribution and a .1 probability of sampling from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 10$. This is called a mixed normal distribution because its distribution is a mixture or weighted sum of two normal distributions, namely, $.9\Phi(x) + .1\Phi(x/10)$, where Φ is the standard normal distribution. Figure 1 shows the standard normal distribution and the mixed normal distribution just described. This is an example of a heavy-tailed distribution, meaning that the tails are "thicker" than a normal distribution, which in turn means that outliers are more likely than situations where a distribution is normal.

Note that the normal and mixed normal distributions appear to be very similar. On the basis of any of several

Figure 1
Normal and Contaminated Normal Distributions

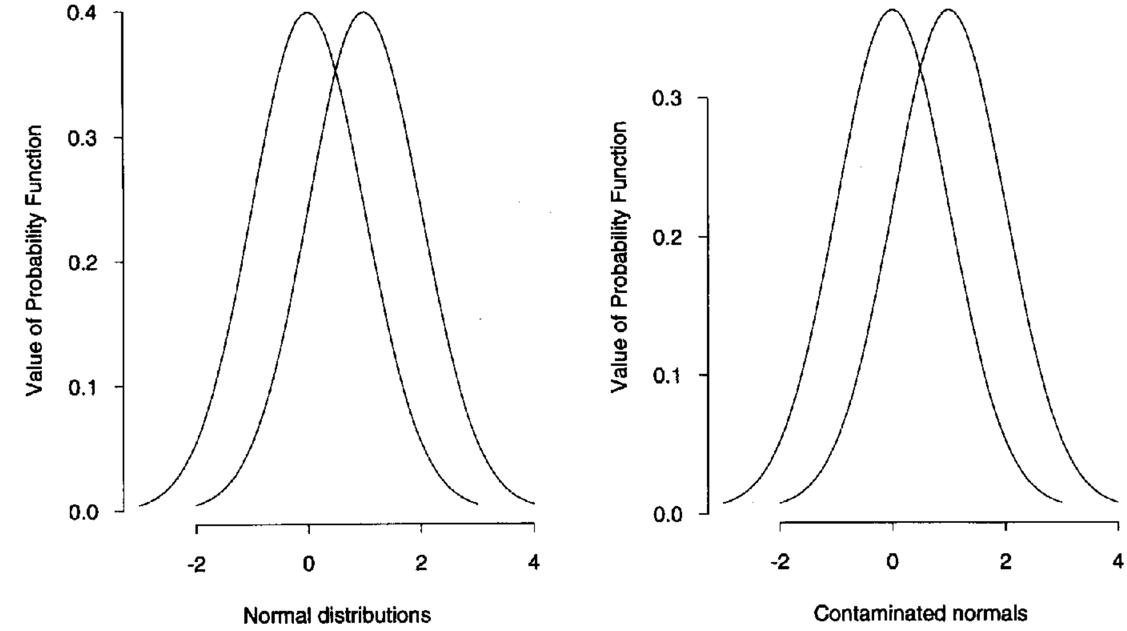


Note. The solid line is a standard normal distribution, and the dashed line is a contaminated normal distribution.

measures of the global difference between distributions, this is indeed the case. The most common measure of the global difference between two distributions is the Kolmogorov distance, which is the maximum value of the absolute difference between the two cumulative distributions. If distributions are identical, then this distance is zero, and the maximum distance is one. For the standard normal and mixed normal distributions considered here, the distance is very small (less than .04), indicating that the distributions are similar. Because the Kolmogorov distance is small, the Kolmogorov test for normality can have low power, making it unlikely that an applied researcher would detect any departure from normality. However, for the standard normal distribution, $\sigma^2 = 1.0$, and for the mixed normal distribution, $\sigma^2 = 10.9$! Because the distribution of the mixed normal is the weighted sum of two normal distributions, the variance turns out to be $.9 + .1(10^2) = 10.9$ (see Wilcox, 1997a, for more details). The contaminated normal distribution illustrates that the population variance is not robust, meaning that a small change in the tails of a distribution can drastically alter the value of the population variance. Put another way, if distributions are normal, one has some sense of how two distributions will appear if one is given their variances, but if one allows the possibility that distributions are not normal, knowing the variance alone, one cannot know how much the distributions differ as measured by the Kolmogorov distance. Also, the population variance is not robust when distributions are skewed, meaning that small changes in the tail of a skewed distribution can result in large changes in the population variance.

An important point is that modern methods do not assume or require that distributions are mixed normals. Rather, mixed normals illustrate the very general concern that very small departures from normality can inflate the population standard deviation. In consequence, the squared standard error of the sample mean (σ^2/n) can become inflated with small departures from normality, and power can be drastically lowered. For example, suppose that for two independent groups with normal distributions, the difference between the means is $\mu_1 - \mu_2 = 1$ and both groups have variance equal to one. The left panel of Figure 2 shows the two distributions. Then, with $n = 25$ and $\alpha = .05$, power is .96 when using Student's test for means. The right panel of Figure 2 shows two contaminated normals; again with $\mu_1 - \mu_2 = 1$, there is little visible difference from the left panel, but now power is .28. Under slight departures from normality, potential discoveries will be lost! What is needed is an estimator

Figure 2
Power and Nonnormality



that performs about as well as the sample mean when distributions are normal but continues to perform well (have a relatively low standard error and high power) under slight departures from normality toward any heavy-tailed distribution.

Table 1 shows the variance of several alternative estimators when sampling from one of four distributions, where M_t indicates a trimmed mean with 10% or 20% trimming, $\hat{\mu}_m$ is what is called an M estimator with Huber's Ψ (Huber, 1981), and $\hat{\theta}_{.5}$ is an estimator of the median derived by Harrell and Davis (1982). A one-wild distribution means that sampling is from a normal distribution and one of the observations is multiplied by 10. A slash distribution starts with a standard normal and divides by an independent, uniform random variable. Both the one-wild and slash distributions are symmetric with heavier than normal tails.

Table 1
Variances of Selected Estimators (n = 10)

	Distribution			
Estimator	Normal	Lognormal	One-wild	Slash
M	.1000	.4658	1.0900	∞
M_t (10%)	.1053	.2238	0.1432	∞
M_t (20%)	.1133	.1775	0.1433	0.9649
Mdn	.1383	.1727	0.1679	0.7048
$\hat{\mu}_m$ (Huber)	.1085	.1976	0.1463	0.9544
$\hat{\theta}_{.5}$	.1176	.1729	0.1482	1.4731

Ideally, the variance of an estimator would be as small or smaller than any other estimator that might be used because this would mean relatively short confidence intervals and high power. None of the estimators in Table 1 achieve this goal primarily because nothing can beat the sample mean when sampling from a normal distribution. However, if the goal is to avoid complete disaster, meaning that the standard error should not be large relative to some other estimator that might be used, the sample mean is the least satisfactory. The two estimators that perform relatively well are the 20% trimmed mean and the M estimator ($\hat{\mu}_m$).

Both trimmed means and M estimators use a type of trimming. That is, extreme values are removed when a measure of location is being estimated. 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. The term *10% trimming* means that 10% of the largest observations, as well as 10% of the smallest observations, are trimmed. If one has 10 observations and the largest value is 35 and the smallest value is 6, 10% trimming consists of removing these two values and averaging the rest. (Some articles call this 20% trimming instead.) In contrast, M estimators empirically determine whether an observation is an outlier, and if it is, adjustments for it are made. M estimators include the possibility of no trimming, in which case $\hat{\mu}_m = M$. When one is using trimmed means, symmetric trimming is best when the goal is to achieve accurate confidence intervals. M estimators allow the possibility of asymmetric trimming (one tail might be trimmed whereas the other tail is not), but this leads to technical complications that must be handled with special techniques. I will not burden the reader with the details of how to compute an M estimator, but it is easy to compute and is easily applied to data by using the software in the books by Wilcox (1996, 1997a). It should be noted that for symmetric distributions, all of the estimators in Table 1 estimate the mean (μ). There are some obvious concerns about trimming, and each is discussed in due course.

Of course, the distributions in Table 1 are artificial. Can it really make a difference which estimator one uses when working with real data? Consider data from a self-awareness study (viz., Wilcox, 1996, Table 8.11). For the first group, the estimated standard error of the sample mean was 136 versus 56 for the 20% trimmed mean and 54 for $\hat{\mu}_m$. This means that a confidence interval, based on trimmed means for example, will have less than half the length of the confidence interval based on means, which means that using trimmed means can result in much higher power. For these data, the usual t test had a significance level of .47, but Yuen's (1974) test for trimmed means had a significance level of .052. If a more modern method for comparing trimmed means were used (Wilcox, 1997a, section 5.3.2), one would reject at the .05 level. Experience suggests that in some cases, the standard error of the sample mean will be a bit smaller than the standard error of these other estimators, but the sample mean seems to rarely, if ever, offer a substantial advantage, and it is fairly common for trimmed means and M estimators to have substantially smaller standard errors. Ignoring the potential problems with means is not in the interest of applied researchers or psychology as a science.

It is stressed that there are formal methods and criteria for deriving location estimators with standard errors that are relatively insensitive to slight changes in a distribution. As a special case, their standard errors are not overly affected by slight departures from normality. In light of these results, the preceding illustration is not surprising and is much more common than psychologists are trained to expect. One reason is that outliers occur in situations where they might seem unlikely. Experience indicates that more often than not, outliers will be found. A boxplot (e.g., Hoaglin et al., 1983; Wilcox, 1996) can be used to check for outliers, but even if no outliers are detected, modern methods can be important, as I explain later.

How do the trimmed mean and the M estimator achieve such low standard errors? As previously noted, both use a type of trimming, but this raises some obvious concerns that must be addressed. The first point that needs to be stressed is that in samples of observations, outliers inflate the estimated standard error of the sample mean. That is, outliers can inflate the sample variance, which in turn can mean relatively long confidence intervals and relatively poor power. Only one outlier can destroy power. By incorporating some mechanism into an estimator that reduces or eliminates the effects of outliers, relatively low standard errors can be achieved.

The more one trims, the more outliers one can have among n randomly sampled observations without getting relatively high standard errors. (This is not completely obvious on the basis of the information given here, but the standard errors of the trimmed mean and the M estimator have been derived, and examining these expressions verifies the statement just made.) For example, if $n = 50$ and 10% trimming is used, there can be as many as 5 outliers without getting an inflated standard error, but 6 outliers might cause practical problems. Similarly, the 20% trimmed mean can handle up to 10 outliers, or 20% of the sample size. The M estimators and the median can handle situations in which up to half of the observations are outliers, but if sampling is from a normal distribution, the median performs poorly, relative to no trimming, in terms of its standard error.

But Outliers Are Important, Interesting, and Informative

One concern about trimming is that outliers can be interesting and informative. Modern robust methods do not deny this, this is assumed to be evident, and a great deal of effort has been made in finding effective methods for detecting outliers. Moreover, it turns out that in terms of identifying and studying outliers, complete reliance on the mean and the variance is relatively ineffective. But more modern methods, based on robust measures of location and scale, have much to offer. That is, what is needed when one is trying to identify and study unusual observations are measures of location and scale that are not themselves affected by outliers. When dealing with regression, again, modern methods that are insensitive to unusual points play a major role in studying outliers, as I later illustrate.

Detecting outliers turns out to be especially difficult when one is dealing with multivariate data, but reasonably effective methods have been derived (e.g., Barnett & Lewis, 1994; Rousseeuw & van Zomeren, 1990; Wilcox, 1997a). A natural strategy is to simply apply a boxplot to each of the variables, but this can be unsatisfactory, and in fact it can miss outliers because it does not take into account the overall structure of the points. For example, a boxplot was applied to five variables taken from a study of reading by Wilcox and Doi (1996). A total of 10 outliers was detected, but when the more modern method by Rousseeuw and van Zomeren is applied, 20 outliers are detected instead. When dealing with multivariate data, a particularly difficult aspect of the problem is finding measures of location and scale that (a) are insensitive to outliers and (b) simultaneously satisfy additional properties that allow one to take into account the overall structure of the points. From a practical point of view, solutions have been derived and can be easily applied with existing software (Wilcox, 1997a).

To elaborate a little, suppose one observes the values 2, 3, 4, 5, 6, 7, 8, 9, 10, and 50 and decides that an observation is an outlier if it is more than two standard deviations from the mean. For these data, $|M - 50| = 2.8\ SD$, where *SD* is the sample standard deviation. Thus, the value 50 would be declared an outlier. But suppose another outlier was added by changing the value from 10 to 50. Then, $|M - 50| = 1.88\ SD$, so 50 would not be declared an outlier, yet surely 50 is unusual versus the other values. If the two largest values in this last example were increased from 50 to 100, then $|M - 100| = 1.89\ SD$; the value 100 still would not be declared an outlier. If the two largest values were increased to 1,000, even 1,000 would not be flagged as an outlier! The problem is that both the sample mean and the sample standard deviation are being inflated by the outliers, which in turn masks their presence.

Although outliers are interesting, if the goal is to achieve high power under nonnormality or to find a measure of location that reflects the typical participant when distributions are skewed, outliers wreak havoc. A natural reaction is that trimming could not possibly work because information is being lost somehow, but a more precise description is that outliers make it difficult to get a relatively short confidence interval for the population mean. The reason trimming works is related to the erroneous strategy of throwing away outliers and applying standard methods to the data that remain. As advocated here, trimming is not simply the throwing away of extreme values and then applying standard methods.

Why Not Discard Outliers and Apply Standard Methods to the Remaining Data?

A common and seemingly natural strategy is to search for outliers, to remove any that are found, and then to apply standard hypothesis-testing methods to the data that remain. This approach fails because it results in using the wrong standard error. Briefly, if extreme values are thrown out, the remaining observations are no longer independent, so conventional methods for deriving expressions for standard errors no longer apply. This result is well-known in mathematical statistics (e.g., Hogg & Craig, 1970), and a relatively nontechnical explanation is given in Wilcox (1997a), but perhaps it is too technical to give here. Instead, I give a different and informal explanation that also sheds light on why trimming can reduce the standard error of an estimator.

Suppose one randomly samples five observations, say X_1, X_2, X_3, X_4, and X_5, and then one puts these five observations in order, yielding $X_{(1)} \leq X_{(2)} \cdots \leq X_{(5)}$. Suppose this process is repeated many times. If, for example, the fourth and fifth largest values among five randomly sampled observations are independent, then the correlation between these two numbers should be zero, and a scatterplot of the points (the sampling distribution of the pairs of points $X_{(4)}$ and $X_{(5)}$) should show no visible association, only a random pattern. To illustrate that they are dependent, five observations were generated from a standard normal distribution, and the two largest observations were recorded. This process was repeated 500 times, resulting in 500 pairs of numbers, each pair representing the fourth and fifth largest values among $n = 5$ randomly sampled values. Figure 3 shows a scatterplot of the points. Clearly, there is an association, and the correlation is .60. Note that whatever the value is for the largest observation, $X_{(5)}$, $X_{(4)}$ must have a smaller value, which is why they are dependent.

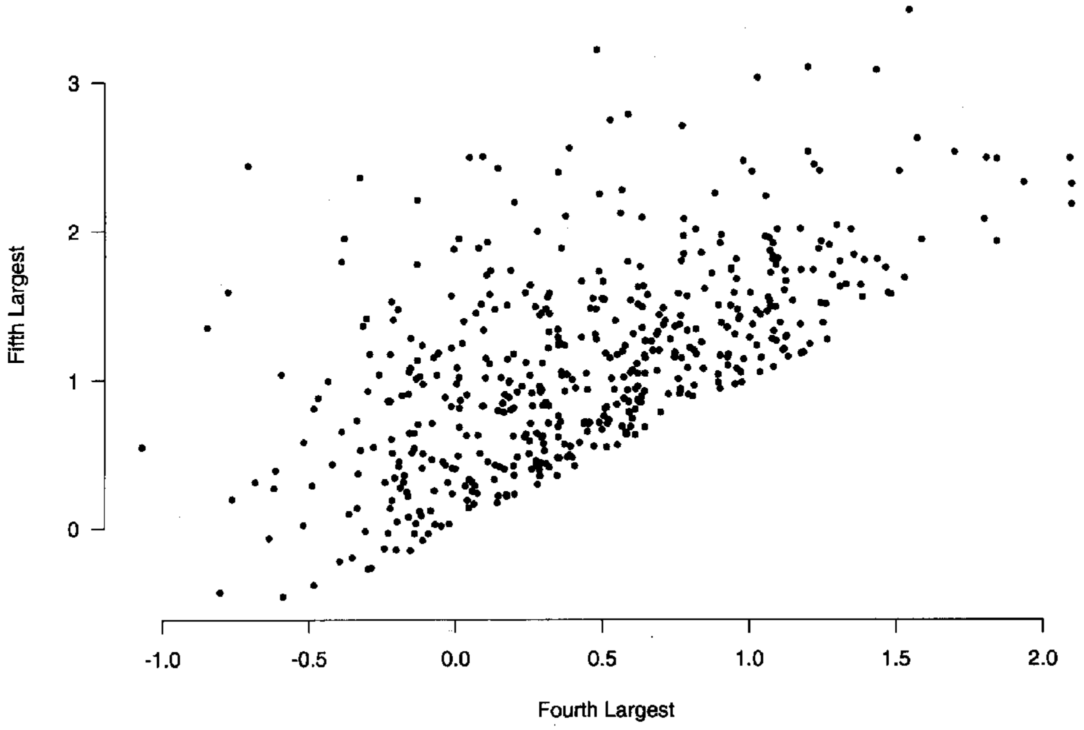
The *i*th largest observation, $X_{(i)}$, is called the *i*th order statistic. Another important point is that the variance of the *i*th order statistic is not equal to the variance of the distribution from which it was sampled. $X_{(1)}$, for example, does not have the same variance as X_1; it is smaller. The reason is that large values for $X_{(1)}$ are less likely than they are for X_1 simply because $X_{(1)}$ is the smallest of the five values. In the illustration, X_1 has a variance of 1.0, but $X_{(1)}$ has a variance of approximately 0.4. More generally, if $X_{(i)}$ is the *i*th largest observation, its variance is not 1.0, and it is correlated with the *j*th largest observation, $X_{(j)}$, for any $i \neq j$.

Now the variance of the sample mean is derived from the result that if $X_1, \cdots, X_n$ are independent,

$$\text{VAR}(\Sigma X_i) = \Sigma \text{VAR}(X_i).$$

That is, the variance of the sum is equal to the sum of the variances, and this plays a role in justifying s^2/n as an estimator of the variance of the sample mean. But if the observations are correlated, this last expression no longer holds; the covariances among the observations must be taken into account. In the illustration, if one discards the smallest and largest observations and sums those that remain, one gets $X_{(2)} + X_{(3)} + X_{(4)}$, and the average of these three numbers is an example of a trimmed mean. But the variance of this sum is not easily determined because the individual terms do not have the same variance as the distribution from which they were sampled, and the covariances among these three random variables must be taken into account. Consequently, it is not readily apparent how to estimate the standard error of the trimmed mean, and when M estimators are used, where the amount of trimming is empirically determined, additional complications are introduced. There are effective methods for dealing with this problem of determining the standard error of an estimator when extreme values are discarded, but the details are too involved to be given here (see Staudte & Sheather, 1990; Wilcox, 1997a). The important point is that these methods are easily applied with existing software and they have great practical importance when one is dealing with low power due to outliers. Even undergraduates taking an introductory statistics course can learn how to do this by hand when working with trimmed means, but when using M estimators, a computer must be used.

Figure 3
Scatterplot of the Two Largest Observations (n = 5)



What If Distributions Are Skewed or There Are No Outliers?

A common misconception is that robust methods are appropriate only when distributions are symmetric, so by implication, standard methods should be used if distributions are skewed. Another misconception is that if there are no outliers, modern robust methods offer no practical advantages. On the basis of both theoretical and simulation results, modern methods give better results when distributions are skewed, whereas standard methods can fail miserably. In terms of power, this is especially true when distributions have relatively light tails, meaning that outliers are relatively rare. In particular, standard methods can have peculiar power properties (power can go down as one moves away from the null hypothesis), confidence intervals can have probability coverage substantially different from the nominal level, and the sample mean can poorly reflect the typical participant under study.

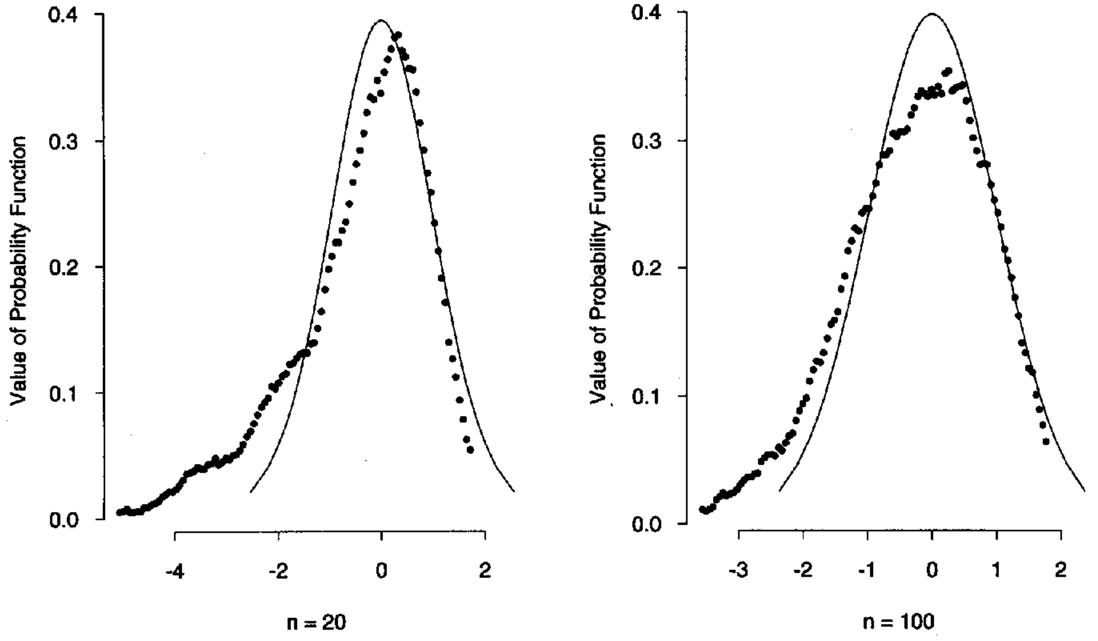
To gain some insight as to why skewness can cause problems, suppose observations are randomly sampled from a lognormal distribution that is skewed. The t test is based on the assumption that

$$T = \frac{\sqrt{n}(M - \mu)}{SD}$$

has a Student's t distribution with $n - 1$ degrees of freedom. In particular, T is assumed to have a symmetric distribution around zero. The left panel of Figure 4 shows the actual distribution of T when $n = 20$. The solid line, symmetric around zero, shows the assumed Student's t distribution, and as is evident, it differs substantially from the actual distribution of T. In fact, the mean of T is approximately -0.5, not zero, as is commonly assumed. (For nonnormal distributions, M and SD are dependent, and this is why the mean of T can differ from zero.) The result is that there are situations where power goes down as one moves away from the null hypothesis, the standard confidence interval for the population mean can have poor probability coverage, and control over the probability of a Type I error can be poor. Of course, increasing n improves matters, but how large must n be before T can be used? The right panel of Figure 4 shows the distribution of T when $n = 100$, where again sampling is from a lognormal distribution. Poor control over the probability of a Type I error and unsatisfactory probability coverage are still problems. According to Westfall and Young (1993), even $n = 160$ is not large enough to eliminate problems, and it is unknown how large n must be so as to ensure good results with T.

For example, suppose one wants the probability of a Type I error to be $\alpha = .05$; one randomly samples n

Figure 4
Probability Density Function of Student's t When Sampling From a Lognormal Distribution



Note. The solid line is the assumed distribution.

= 20 observations from a lognormal distribution having a mean of 1.649 and tests the null hypothesis that $\mu >$ 1.649. The actual probability of a Type I error is .153. Increasing n to 160, the actual probability of a Type I error drops to only .109.

Theory states that as the probability of sampling outliers goes down, or when the tails of a distribution get thinner, the problems illustrated by Figure 4 get worse! Outliers are common when one is sampling from a lognormal distribution, so the illustrations just given are not based on an extreme case. The t test can be made even worse by sampling from a lighter tailed distribution. For example, Wilcox (1997a) reported situations where one tests at the .05 level but the actual probability of a Type I error is .42.

Yet another concern with skewed distributions is that the population mean or the sample mean might not provide a good reflection of the typical participant. Consider again the self-awareness data (Wilcox, 1996). The sample mean was 448, yet about 80% of the observations had values that were less than 448. That is, the data suggest that the mean was close to the .8 quantile. By trimming, one gets a measure closer to the bulk of the observations. For the data at hand, the median is 262, and the 20% trimmed mean is 283.

How Much Trimming Should Be Used?

When using trimmed means, one must first decide how much trimming should be done. This issue has been examined from various perspectives, and a good choice for general use is 20%. Of course, this choice is not always optimal in terms of minimizing the standard error, but no trimming or 10% trimming can result in very poor power and highly inaccurate confidence intervals, whereas 20% trimming competes well with no trimming when distributions are normal. If the criterion is accurate probability coverage, it is known that the more skewed a distribution happens to be, the more unsatisfactory the standard confidence interval for the mean becomes, particularly in situations where outliers are unlikely to appear. There are also theoretical and simulation results showing that the more one trims, the more accurate the probability coverage is, but one does not want to trim too much if one wants to avoid low power when distributions are normal.

An analogue of Student's t test has been derived for situations where a trimmed mean is used instead. It involves computing what is called the *Winsorized variance,* which in turn is multiplied by a constant, the value of which depends on the amount of trimming used (for details, see, e.g., Wilcox, 1996, p. 118). Although trimming improves matters, practical problems remain. However, both theory and simulations indicate that if one combines trimmed means with what is called the percentile-t bootstrap method, even better probability coverage can be obtained. Briefly, an analogue of the t test has been derived for trimmed means; the null distribution has approximately a Student's t distribution, but even better approximations of the null distribution can be obtained with a computer (for computational details, see Wilcox, 1997a, section 4.3.1.).

For example, it was previously pointed out that there are situations where one tests at the .05 level with Student's *t*, but the actual probability of a Type I error is .42. Switching to the analogue of Student's *t* for trimmed means, the actual probability of a Type I error is .063. If one uses a percentile-*t* bootstrap method with the 20% trimmed mean, the Type I error probability is .053.

When two or more independent groups are compared, all of the problems discussed so far remain, and there is the additional problem that unequal variances contribute to poor power and inaccurate confidence intervals. In fact, if there are unequal sample sizes and groups differ, there are general conditions under which the usual confidence interval for $\mu_1 - \mu_2$ (the difference between the means) has probability coverage that does not even converge to the nominal level as the sample sizes get larger (Cressie & Whitford, 1986). More precisely, Student's *t* in the two-sample case is assumed to approach a standard normal distribution as the sample sizes get large, so, in particular, its variance approaches 1.0. But if the variances corresponding to the two groups are unequal, there are general conditions where the variance of Student's *t* statistic does not approach 1.0, contrary to what is commonly assumed.

As another illustration, again consider the self-awareness data (Wilcox, 1996). If the means of the two groups are compared, the significance level is .47. When comparing 20% trimmed means (with Yuen's [1974] test; see Wilcox, 1996, p. 138), the significance level drops to .0525, and if trimmed means are compared with a percentile-*t* bootstrap method, one rejects at the .05 level. Of course, 20% trimmed means do not offer the most power in all situations—nothing does—but this illustrates that modern methods can result in a substantially different perspective about how two groups compare.

What about M estimators? They compare fairly well with trimmed means in terms of controlling the probability of a Type I error or achieving high power, but trimmed means perform well over a broader range of situations. However, when attention is turned to correlation and regression, M estimators have advantages over trimming. Interested readers can refer to Wilcox (1996, 1997a). Some illustrations of the practical advantages of M estimators are given below.

Note that for skewed distributions, trimmed means and M estimators are not estimating the same quantity as the sample mean, and perhaps there are situations where this offers an advantage to means, in terms of power. Although it is possible, this is relatively unlikely because trimmed means and M estimators usually have standard errors as small as or much smaller than the mean. Situations arise where comparing means results in a slightly lower significance level, but it is difficult to find situations where means are rejected at the .05 level and trimmed means are not. Presumably such situations arise—again no single method is perfect—but it seems relatively easy to find situations where trimmed means have substantially lower significance levels. If groups do not differ, it makes little difference which measure of location is used, and it is not surprising to find situations where methods based on means have lower significance levels. Of course, one does not know whether they differ, but if they do, the choice of an estimator might make a substantial difference.

Other Issues

A natural strategy is to test assumptions (normality and equal variances) and if they are not significant, to use a standard method. This strategy fails because conventional tests of assumptions do not have enough power to detect situations where standard assumptions are unsatisfactory (e.g., Wilcox, 1996). The only known way of determining whether modern robust methods make a difference is to apply them to the problem at hand.

Another reasonable suggestion is to transform the data, but simple transformations fail. For example, the common strategy of using the logarithm of all observations does not always remove the effect of outliers, and more sophisticated transformations have been shown to be unsatisfactory as well (Wilcox, 1996).

Modern methods have been extended to one-way and higher designs, including repeated measures designs. It is hoped that as the number of groups increases, some of the problems associated with Student's *t* become negligible, but the exact opposite is true.

Although it is a seemingly rare event, a few researchers have the incorrect notion that modern robust methods are designed to find better estimates of the population mean, the population variance, and the population correlation. Generally, this is not of interest from a modern point of view because these parameters are not robust. That is, even if all participants could be measured, the resulting values could be misleading under arbitrarily small departures from normality. (An exception is when distributions are symmetric, in which case the goal is to find a good estimator of the population mean, meaning that its standard error should not be affected by small changes in the distribution.) In the next section of this article, I elaborate on this important issue.

What Is a Robust Parameter?

For most applied researchers, it seems that the term *robust* means that a particular hypothesis-testing procedure controls the probability of a Type I error. Among modern statistical methods, it has a more general meaning that applies to both parameters and estimators. In terms of hypothesis testing, small changes in a distribution should not result in large changes in power or probability coverage. As a special case, when one is sampling from normal distributions, small shifts away from normality should not drastically affect the value of the population mean and variance. As previously illustrated, the population variance is not robust, and similar problems plague the population mean. There are three formal criteria for judging the robustness of a parameter, but the details go beyond the scope of this article. For a description written at an intermediate level of difficulty, see Staudte and Sheather (1990). For a less technical description, see Wilcox (1997a). It turns out that the population mean, the population variance, and the population correlation do not satisfy any of these criteria.

Under random sampling, it is known that as the sample size gets large, the sample mean approaches the population mean. Despite this result, it is possible for the empirical distribution to be arbitrarily close to the distribution generating the observations (in the Kolmogorov sense), but the difference between the sample mean and the population mean can be arbitrarily large! This problem does not arise when trimmed means or M estimators are used.

In light of all the negative consequences of using means, should they be discarded? Presumably the answer is no, but using means to the exclusion of modern measures of location is not remotely satisfactory and it is a relatively uninformative way to proceed.

Correlation

When attention is turned to correlation and especially regression, the problems that plague methods for means remain, and in some ways, they become worse. There are formal methods for deriving robust analogues of the population correlation and least squares regression, but no details are given here. In fact, there are dozens of robust correlations. Five robust correlations, beyond Spearman's ρ and Kendall's τ, are described in Wilcox (1997a) and have value depending on the goals of the applied researcher. Here, the goal is to illustrate how a modern measure of correlation versus the usual test of independence based on Pearson product–moment correlation can yield substantially different conclusions. As will become evident, modern methods also provide insight into how outliers affect the overall assessment of how variables are related. A basic problem with correlation is that it is not resistant. That is, a single unusual value, or a small change in many values, can affect a Pearson product–moment correlation to the point that one fails to detect associations that are revealed when more modern methods are used.

Percentage-Bend Correlation

The percentage-bend correlation (r_{pb}) is computed as described in Table 7.1 of Wilcox (1997a). (An S-PLUS function can be downloaded from the web site maintained by Academic Press, as described in section 1.7 of Wilcox, 1997a. A Minitab macro for computing the percentage-bend correlation comes with the textbook by Wilcox, 1996.) One reason this measure of association is presented here is that its population value is equal to zero under independence. Not all robust measures of association have this property. Another reason is that the resulting test of independence has been found to provide good control over the probability of a Type I error for a broader range of distributions versus the standard test based on Pearson product–moment correlation, especially when one is dealing with multiple pairs of random variables (Wilcox, 1997c). The test statistic for independence, using the percentage-bend correlation, is

$$t_{\text{pb}} = r_{\text{pb}}\sqrt{\frac{n-2}{1-r_{\text{pb}}^2}},$$

and the (two-sided) null hypothesis is rejected if $|t_{\text{pb}}| > t$, where t is the $1 - \alpha/2$ quantile of Student's t distribution with $n - 2$ degrees of freedom. That is, one should use the usual test with correlation replaced by the percentage-bend correlation. The simple strategy of replacing standard estimators with some robust analogue in a conventional test statistic usually fails, but it happens to perform well for the problem at hand.

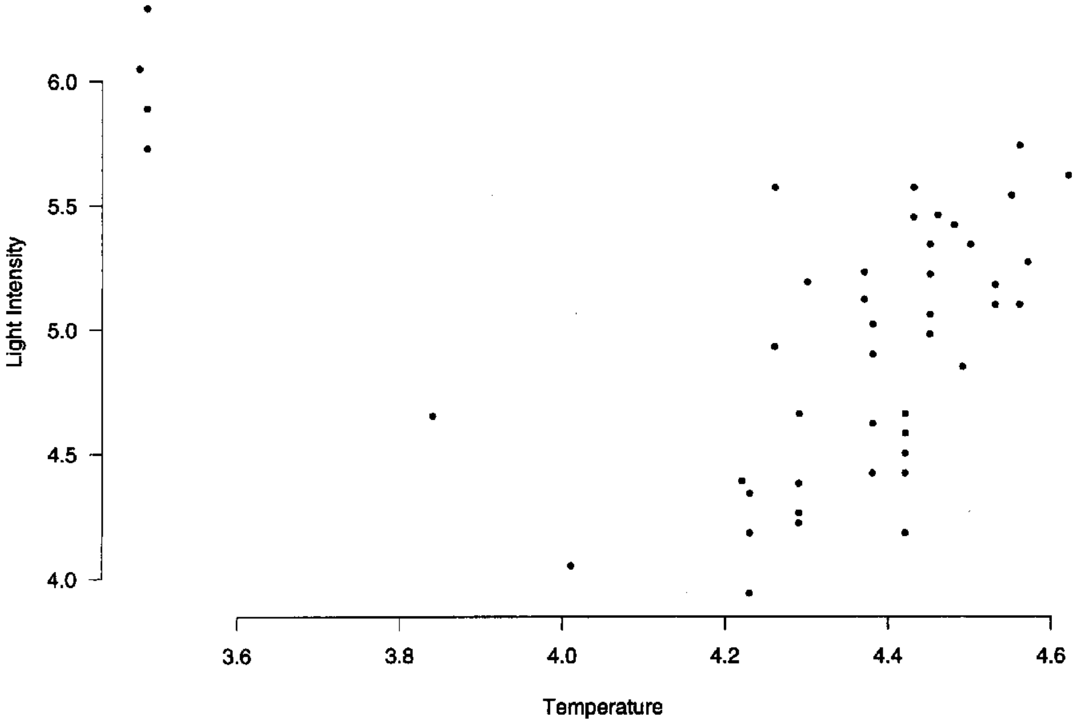
When using the percentage-bend correlation, one must choose the value of a parameter that determines how many outliers can be handled. (This is the parameter β in Table 7.1 of Wilcox, 1997a.) The situation is similar to choosing how much trimming to do when using trimmed means. If $\beta = 0$ is used, $r_{\text{pb}} = r$ and a single outlier can be a problem. If $\beta = .2$, about 20% of the observations can be outliers, which appears to be a good choice for general use. In particular, even if distributions are normal, there is little advantage to using r over r_{pb}, but if $\beta = .5$, this is not necessarily the case. However, there are situations where $\beta > .2$ might be needed. Two versions are used here: one that is slightly resistant to outliers ($\beta = .1$) and one that is moderately resistant ($\beta = .2$). The point that is illustrated here is that for the seemingly simple problem of testing for independence, modern robust methods can make a difference.

Some Illustrations

The first illustration is based on data collected by M. Earlywine and reproduced in Wilcox (1997a). Each participant consumed a specified amount of alcohol, and a measure of hangover symptoms was recorded. This was done on three different occasions. The Pearson product–moment correlation (r) between Time 1 and Time 2 had a significance level of .108, and for Time 2 and Time 3, the significance level was .075. When the slightly resistant percentage-bend correlation coefficient was used, the significance levels were .024 and .006, respectively. If one were to use the more resistant form of the percentage-bend correlation, then the significance levels would drop to .007 and .002, respectively, a rather dramatic decrease versus the Pearson product–moment correlation. This suggests that the variables are dependent, a result that would have been missed if correlation was used at the .05 level.

A more dramatic illustration is provided by the star data in Rousseeuw and Leroy (1987, p. 27), which consists of the logarithm of the effective temperature at the surface of 47 stars versus its light intensity. First, look

Figure 5
Star Data



at Figure 5, which shows a scatter plot of the data. Note that for the bulk of the data, $X > 4$, and the scatter plot suggests that there is a positive association between light intensity and temperature. However, $r = -.21$ with a significance level of .15. This implies that as light intensity increases, temperature goes down, but surely this is an inadequate summary of the data because, in general, the reverse is true. The reason the correlation is negative is that the four isolated points in the left portion of Figure 5 are outliers that dominate its value. As previously noted, even one outlier can cause the correlation to be negative when the remaining points have a positive association. The slightly resistant percentage-bend correlation coefficient is equal to .065 with a significance level of .665. The outliers are having less of an influence, but for these data, more resistance is needed. The moderately resistant percentage-bend correlation is equal to .265 with a significance level of .072. (If β is increased from .20 to .26, the correlation is now .29 with a significance level of .048.)

Once again, I am not suggesting that the outliers that affect the correlation are uninteresting. Quite to the contrary, there is the obvious issue of trying to determine why these points seem to differ from the bulk of the observations. If one simply considers correlation, one is missing the fact that a few points dominate its value, and there is no hint that interesting outliers might exist. For the star data, it seems that there is a positive association if $X > 4$, but something interesting seems to occur for smaller values of X. If one simply fits a least squares regression line to the data, one gets a negative slope, and again one misses the positive association when $X > 4$, and clearly this positive association is interesting too. Also, if one tests the null hypothesis that $\rho = 0$ and gets a nonsignificant result, why does this occur? It might be because there is indeed no association, or perhaps there is an association, but an outlier prevented it from being detected.

It might appear that the significance level decreases as the resistance of an estimator increases, but exceptions occur. Using the methods in Goldberg and Iglewicz (1992) as well as Rousseeuw and van Zomeren (1990), it can be seen that the data analyzed here have outliers that lower power. These outlier detection methods are based on robust multivariate measures of location and scatter, the robust measure of scatter being a robust analogue of the usual covariances between random variables. Goldberg and Iglewicz's method is limited to the bivariate case, but Rousseeuw and van Zomeren's method can be used in the multivariate case.

Regression

There are many robust analogues of the ordinary least squares (OLS) regression estimator. Several of them offer substantial improvements in terms of power and resistance to outliers while sacrificing very little when the error term is normal and homoscedastic. Robust regression estimators eliminate or reduce the effects of outliers, but again, it is stressed that when one is using a resistant estimator, it is not being suggested that unusual points have no interest. Robust methods add valuable insight into how variables are related, and modern regression methods play a role in identifying which points are indeed unusual as compared with the bulk of the points being studied. That is, one of the tools that is useful when searching for unusual points is a regression method that is not itself sensitive to outliers. By finding a regression line that gives a good fit to the bulk of the points, one is better able to determine which points are unusual. Rousseeuw and van Zomeren (1990) described one such method, a summary of which is in Wilcox (1997a).

Which regression estimator is best? Several estimators can handle a large number of outliers; many of these have poor efficiency in certain situations, meaning that they have relatively large standard errors, but several offer substantial advantages. In terms of achieving high power, the methods in Wilcox (1997a, chap. 8) currently stand out. The adjusted M estimator (see section 8.5.4 of Wilcox, 1997a) seems particularly attractive in terms of efficiency, as compared with the OLS estimator, but situations arise where some other robust method has higher power. That is, no single method is perfect. In terms of computing accurate confidence intervals, modern methods give reasonably good results for a much broader range of situations than does the standard technique.

The main message is that the OLS estimator is one of the poorest choices researchers could make. In some cases, its standard error is more than 100 times larger than certain modern methods! Even when the error term is normal but heteroscedastic, the OLS estimator and the standard confidence interval can be unsatisfactory. For example, when the sample size equals 20, the actual probability coverage can be less than .5 when a .95 confidence interval for the slope is being computed (Wilcox, 1997a, p. 209). Put another way, when one is testing at the .05 level, the actual probability of a Type I error can exceed .5! In contrast, modern methods have probability coverage close to the nominal level for these same situations. Currently, what works best when confidence intervals are being computed is a bootstrap percentile method in conjunction with any one of several robust estimators. The S-PLUS function regci in Wilcox (1997a) can be used with any estimator of interest, and tests that all p predictors are equal to zero can be made as well with the function regtest.

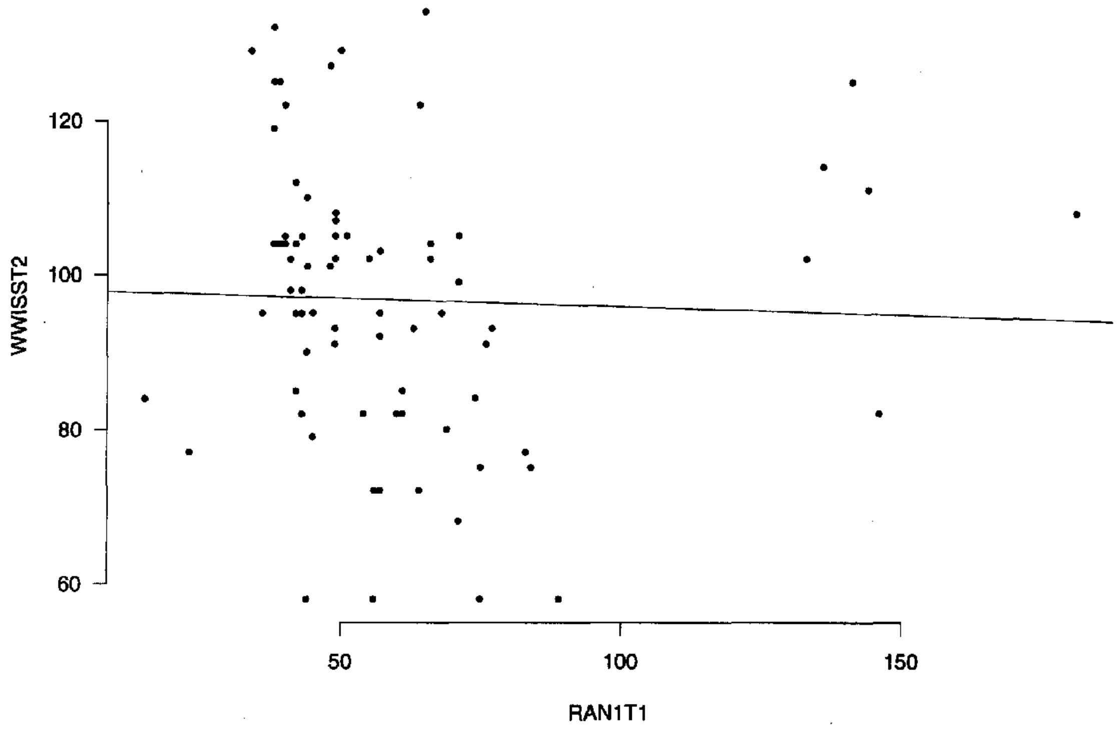
A criticism of robust regression estimators is that the usual strategy of checking for curvature, by checking the residuals, can fail (e.g., Cook, Hawkins, & Weisberg, 1992; McKean, Sheather, & Hettmansperger, 1993). One approach for dealing with this problem is to use some type of smoother, which is a method of examining the shape of a regression line without assuming any particular parametric form, such as a straight line. An illustration is given below. As will become evident, even when the OLS estimator is used, scatter plots of the points might be deceptive.

Some Illustrations

Rather than describe the details of how robust regression methods are computed, I simply provide some illustrations of their importance. The first illustration is based on data collected by L. Doi and studied by Wilcox and Doi (1996). The general goal is to examine predictors of reading ability. One specific goal was to study the ability of a measure of speeded naming for digits (RAN1T1) to predict a measure of the ability to identify words (WWISST2). Figure 6 shows a scatter plot of the points plus the OLS regression line, which is nearly horizontal. The estimated slope is −0.02, and the correlation is −.04 with a significance level of .76. Thus, it might seem that there is little or no association between these two variables. Now look at Figure 7, which shows a relplot of the same data plus a running interval smoother. A relplot is a bivariate analogue of the boxplot derived by Goldberg and Iglewicz (1992). The inner ellipse of a relplot contains half the points, and points outside the outer ellipse are labeled *outliers*. A running interval smoother was created with the S-PLUS function runmean in Wilcox (1997a). It estimates the regression line when the goal is to predict the 20% trimmed mean of WWISST2 with RAN1T1. Note that for the bulk of the points, there seems to be a negative association. Also, the outliers are raised high enough so as to mask this association when one is using OLS or correlation. If one uses the biweight midregression or the Winsorized regression methods in Wilcox (1997a), one rejects at the .05 level. If the range of RAN1T1 values is restricted so that outliers are eliminated, the slope is found to be significantly different from zero at the .05 level using the adjusted M estimator in Wilcox (1997a) in conjunction with a percentile bootstrap. Without restricting the range, the M estimator is not significant. (If points with outlying residuals are eliminated and standard methods are applied, one gets the wrong standard error.)

The Theil-Sen estimator has been found to compete well with the OLS estimator when there is only one regressor, its efficiency compares well with the M estimator used here, and accurate confidence intervals can be computed even when the error term is heteroscedastic and highly nonnormal and the sample size is as small as 20 (Wilcox, 1997b). The estimator is easily applied when there is only one predictor (e.g., Conover, 1980). When there is more than one predictor, complications arise; perhaps M estimators are better for general use, but this needs further study. (When M estimators are used, the software in Wilcox, 1997a, can handle as many predictors as desired.) When a percentile bootstrap method is used, there seems to be little or no difference between the Theil-Sen estimator and the M estimator in terms of probability coverage. The Theil-Sen estimator is reasonably insensi-

Figure 6
Scatterplot and Least Squares Fit of RAN1T1 Versus WWISST2



tive to outliers, and for the data at hand, it estimates the slope to be −0.282. The .95 confidence interval (using the percentile bootstrap method) is (−0.63, −0.01), so the null hypothesis that $\beta_1 = 0$ would be rejected although there is some possibility that the slope is fairly close to zero. That is, without restricting the range of the RAN1T1 values, once more a significant result is obtained. A Winsorized correlation, with 20% Winsorization, has a significance level of .02, again using all of the data, and the percentage-bend correlation with $\beta = .2$ has a significance level of .015.

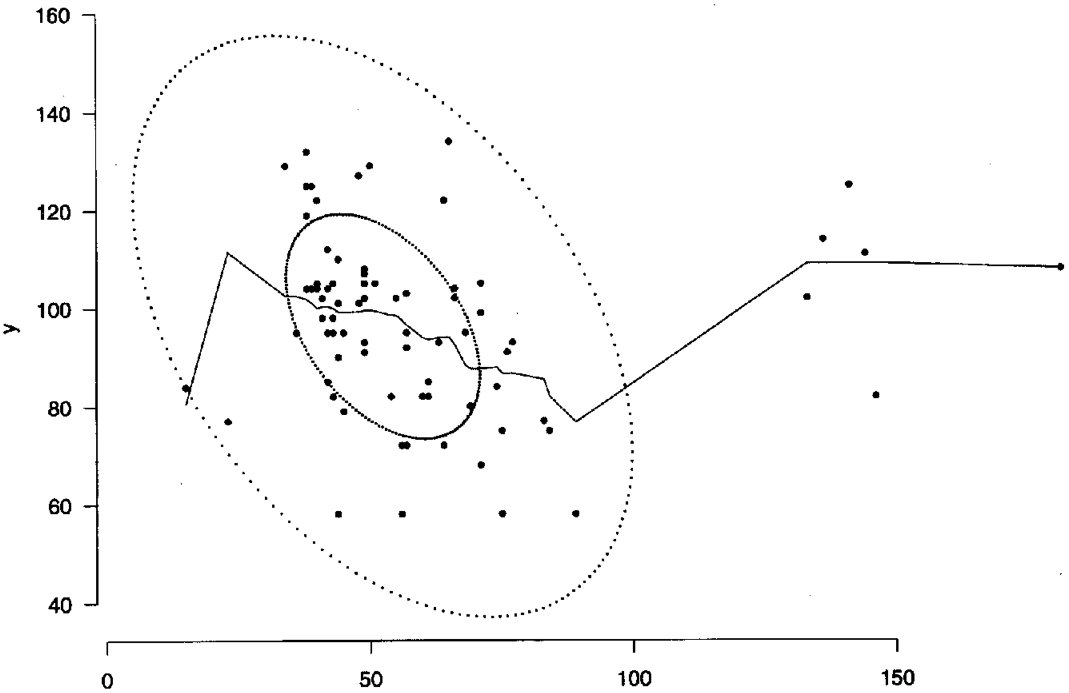
Again, it is stressed that unusual points are of interest, and it is not being suggested that they have no value. For the problem at hand, there is interest in knowing how WWISST2 is related to RAN1T1 when RAN1T1 has a value greater than 100, but with only six such points, it is difficult to know for sure. The only message here is that these six points seem to mask an association when RAN1T1 has a value less than 100.

As another illustration, consider the Pygmalion data recently discussed by Snow (1995) and originally collected by R. Rosenthal. The study compared children in an experimental condition, for whom positive expectancies had been suggested to teachers, with children in a control condition, for whom no expectancies had been suggested. Both pretest and posttest reasoning IQ scores were obtained. For the experimental group, the usual correlation had a significance level of .052, whereas the two robust versions considered here have significance levels of .021 ($\beta = .1$) and .007 ($\beta = .2$). That is, the percentage-bend correlation provides more convincing evidence that the variables are dependent. The OLS estimator of the slope is 0.57 versus 0.647 using the Theil-Sen estimator. The standard .95 confidence interval for the slope is (−0.006, 1.138).

Even when attention is restricted to the OLS estimator, modern methods have something to offer when one is computing confidence intervals or testing hypotheses. In particular, the modified percentile bootstrap method described in Table 13.5 of Wilcox (1996), which was designed specifically for the OLS estimator, performs well in simulations. Applying this method to the problem at hand, the .95 confidence interval for the slope is (0.059, 1.777). Note that this interval is longer than the standard interval, the ratio of the lengths being (1.777 − .059)/(1.138 + .006) = 1.5, a rather substantial difference. Despite this difference, the more accurate confidence interval rejects the hypothesis of a zero slope, but the standard method does not. (The modified percentile method yields a confidence interval that is not necessarily symmetric about the estimate of the slope, which is why the confidence interval can be longer yet reject when the standard method does not.)

Again consider the Pygmalion data (Snow, 1995). The posttest scores of the two groups differed signifi-

Figure 7
Relplot and Smooth of RAN1T1 Versus WWISST2



cantly, based on both the usual *t* test and robust methods. Generally, significant results with a standard method remain significant when a robust technique is used. However, if robust methods are used to compare the slopes and the intercepts, when comparing the regression lines for predicting posttest scores, given pretest scores, no differences are found. If a robust analogue of analysis of covariance is used, which makes no parametric assumptions about the form of the regression line, again highly nonsignificant results are obtained (Wilcox, 1997a). That is, if pretest scores are taken into account, there is no compelling reason to believe that expectancies influence test scores. Of course, this one reanalysis does not resolve the controversy surrounding studies about Pygmalion in the classroom, but it does illustrate that new and improved methods might be used to study this issue.

Concluding Remarks

Experience suggests that it is easy to find situations where modern methods reject and standard methods for means do not. In contrast, it is difficult, but presumably possible, to find real data where methods based on means reject but trimmed means do not. Modern rank-based methods, developed after the year 1965, deserve serious consideration as well. Situations arise where rank-based methods reject but methods based on trimmed means do not, and it is common for the reverse to happen too. (Boxplots often provide an explanation for why this happens.) If the goal is to avoid low power, the worst method is the ANOVA *F* test.

Regression is very difficult, and I do not mean to suggest that modern methods solve all practical problems. Unfortunately, I cannot point to a single method that always gives the best results, but many estimators can be eliminated, and recommendations can be made on how to proceed. The adjusted M estimator and the biweight midregression estimator appear to perform relatively well over a broad range of situations. If there is one predictor, the Theil-Sen estimator deserves consideration as well. Which estimator is best depends on the situation at hand and cannot be determined prior to looking at the data. Although no single method is perfect, the least satisfactory approach can be identified: Apply OLS regression, or simply report correlations and assume all is well. Despite this negative result, the more general message is very positive: Researchers have the technology to vastly improve on standard ANOVA and regression techniques. Thus, although at first the message in this article might seem discouraging, it really reflects a great opportunity for improving psychological research.

REFERENCES

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

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression.* New York: Wiley.

Conover, W. J. (1980). *Practical nonparametric statistics.* New York: Wiley.

Cook, R. D., Hawkins, D. M., & Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fit. *Journal of the American Statistical Association, 87,* 419–424.

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

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

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

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

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

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

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

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

McKean, J. W., Sheather, S. J., & Hettmansperger, T. P. (1993). The use and interpretation of residuals based on robust estimation. *Journal of the American Statistical Association, 88,* 1254–1263.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and 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.

Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science, 4,* 169–171.

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

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

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

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

Wilcox, R. R. (1997b). *A note on the Theil-Sen regression estimator when the regressor is random and error term is heteroscedastic.* Unpublished manuscript, University of Southern California.

Wilcox, R. R. (1997c). Tests of independence and zero correlation among P random variables. *Biometrical Journal, 39,* 183–193.

Wilcox, R. R., & Doi, L. M. (1996). *Predictors of reading ability: An application of modern robust regression methods plus some new exploratory techniques.* Unpublished manuscript.

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