MedCrave
Step into the World of Research

Journal of Psychology & Clinical Psychiatry

Review Article

Open Access

CrossMark

Gaining a deeper and more accurate understanding of data via modern robust statistical techniques

Abstract

Many new and improved statistical techniques have been derived for comparing groups and studying associations. Under general conditions, compared to conventional methods based on means and least squares regression, these new methods provide more power, more accurate confidence intervals, and new perspectives that help deepen our understanding of data. Although it is well known in the statistics literature that these new methods can make a practical difference when analyzing data, it is evident that their practical advantages are not well known outside the group of statisticians who specialize in robust techniques. The paper reviews the basics of these new methods, including some recent advances, and it illustrates their practical advantages using data from various studies.

Keywords: non-normality, outliers, heteroscedasticity, regression, anova, curvature, sexual attitudes, depressive symptoms, biomarkers, reading ability, life satisfaction, alcoholism

Volume 1 Issue 2 - 2014

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

Correspondence: Rand R Wilcox, Department of Psychology, University of Southern California, USA, Email rwilcox@usc.edu

Received: May 20, 2014 | **Published:** June 21, 2014

Introduction

Consider the classic, routinely taught and used statistical methods for comparing groups based on means. A fundamental issue is whether these methods continue to perform well when dealing with non-normal distributions. Contrary to what was once thought, it is now known that for a broad range of situations this is not the case.[1–7] A positive feature of standard methods for comparing means is that they control the Type I error probability reasonably well when groups have identical distributions. So in addition to equal means, the variances and the amount of skewness are the same among the groups being compared. But a practical concern is that if groups differ, classic method scan yield inaccurate confidence intervals and can have relatively poor power, where power refers to the probability of detecting true differences among the groups. Yet another concern is that they can result in a highly misleading summary of the data associated with the bulk of the participants.

One strategy for dealing with non-normal distributions is to use classic rank-based methods, but it is now known that under general conditions these techniques have practical concerns as well. Like methods for comparing means, they perform well in terms of controlling the Type I error probability when comparing groups that have identical distributions. But otherwise they perform poorly compared to more modern rank-based techniques.[4,5,8,9] Also, rank-based methods are not generally designed to make inferences about some measure of central tendency in contrast to the robust methods reviewed here. Section 3 elaborates on this point.

When dealing with regression, Pearson's correlation and the usual (least squares) regression methods perform well in terms of controlling the Type I error probability when there is no association. But otherwise, they inherit the practical concerns associated with methods based on means and new problems are introduced.[1–7] In practical terms, a few outliers can mask a strong association among the bulk of the participants as will be illustrated in section 7.

During the last fifty years, many new and improved statistical techniques have been derived for comparing groups and studying associations.[1–7] A rough characterization of inferential methods based on means is that even small changes in the distributions under study, such as slight departures from a normal distribution, can have an inordinately large impact in terms of power and more generally any characterization of how groups compare and how variables are related. Modern robust methods are designed to avoid this problem. Although the practical utility of these methods is well known in the statistics literature, they are relatively unknown among most researchers trying to understand data. In an attempt to close this gap, the paper reviews the basics of modern robust methods, it summarizes some recent advances, and it illustrates some of their practical advantages.

Concerns about methods based on means and variances

Appreciating the practical importance of modern robust methods requires in part some understanding of when and why more conventional methods, based on means and least squares regression, are unsatisfactory. But before addressing this issue, it helps to first comments about methods aimed at detecting outliers.

Detecting outliers

Outliers can destroy power (the probability of detecting true differences among groups and true associations among variables) and as previously indicated they can result in a highly misleading summary of the data. Some awareness of these concerns is apparent among most psychologists because researchers frequently indicate that checks for outliers were made. A common strategy for detecting outliers is to declare a value an outlier if it is more than two or three standard deviations from the mean. More formally, consider a random sample of n observations: $X_1, \ldots, X_n$. Then the two standard deviation rule is to declare the ith observation, X_i, an outlier if

$$\frac{|X_i - \hat{X}|}{s} \geq 2$$

Where X the usual sample mean and s is is the standard deviation. But this method is well known to be unsatisfactory, roughly because outliers can inflate the standard deviation s, which results in outliers being masked.[2,4] The general strategy for dealing with this problem is to replace the mean and standard deviation with measures of

central tendency and scale (measures of variation) that are relatively insensitive to outliers. Two such methods are the box plot and the so-called MAD-median rule. From basic principles, the box plot is based on the interquartile range, meaning that more than 25% of the values would need to be outliers for it to break down. As for the MAD-median rule let M be the usual sample median and let MAD indicate the median absolute deviation statistic, which is the median based on |X1−M|, . . . , |Xn− M|. Then the MAD median rule declares Xi and outlier if

$$\frac{|X_i - M|}{MADN} > 2.24,$$

Where MADN is MAD/.6745 (Under normality, MADN estimates the standard deviation.) The MAD-median rule can accommodate more outliers than the box plot without breaking down, but the choice between the two methods is not always straightforward. No details are given here, however, because this goes well beyond the scope of this review and the details are not directly relevant to the main points made here (For a recent summary of outlier detection methods designed for multivariate data).[5]

Practical concerns with student's t test

Consider the one-sample Student's t test. At one time it was thought to perform reasonably well when dealing with non-normal distributions. Indeed, many introductory statistics book still claim that with a sample size of about 30, normality can be assumed. If sampling is from a symmetric distribution, then generally the actual Type I error probability will be at or below the nominal level. But two things were missed. The first has to do with the central limit theorem and skewed distributions. Early studies found that with a relatively small sample size, say 30, the sample mean has, to a good approximation, a normal distribution under fairly weak conditions. But this does not necessarily mean that Student's t will provide good control over the Type I error probability or reasonably accurate confidence intervals.

As an illustration, consider a lognormal distribution, which is skewed and relatively light-tailed, meaning that for a random sample, the proportion of points declared an outlier will be relatively small based on a box plot or the MAD-median rule. Suppose that for a nominal .05 Type I error probability, Student's t is judged to be reasonably accurate if the actual Type I probability is between .025 and .075. Then approximately 200 observations are required. When dealing with a skewed distribution where outliers are relatively common, now 300 observations can be required.[5]

This has implications regarding the two-sample t test. Again suppose a researcher wants the Type I error probability to be .05. If the distributions being compared have the same amount of skewness, then Student's t is satisfactory in the sense that the actual Type I error probability will not exceed .05. But otherwise, the actual Type I error probability can be substantially larger than .05 as illustrated in.[4]

The second thing that was missed has to do with the high sensitivity of the population variance to the tails of a distribution, which has serious implications in terms of power. Even arbitrarily small changes in a distribution, in a sense made precise, for example, in,[1,5] can substantially impact the population variance, which in turn can substantially alter power when testing hypotheses based on the sample mean. In practical terms, important differences among the groups being compared can be missed.

A classic example is shown in Figure 1. The left panel shows two normal distributions, both of which have variances equal to one. With sample sizes n1=n2=25, power is .96 when testing at the .05 level with Student's t. Now look at the right panel. Now power is only .28 despite the apparent similarity with the normal distributions in the left panel. The reason is that these distributions are not normal; they are mixed normal distributions that have variance 10.9. This illustrates the general principle that inferential methods based on means are highly sensitive to the tails of the distributions, roughly because the tails of a distribution can have an inordinate influence on the variance.

Some unsatisfactory strategies for dealing with skewed distributions and outliers.

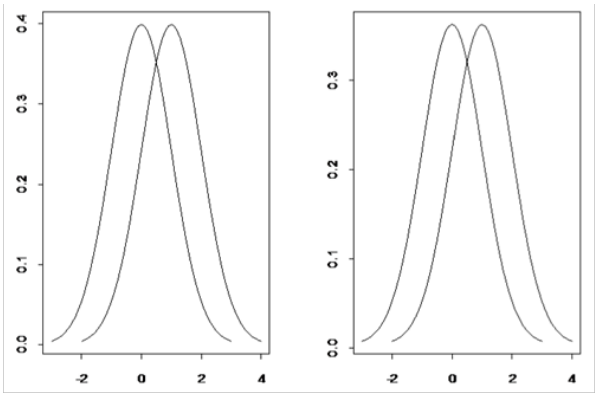

Figure 1 In the left panel, power is .96 based on student's t, α = .05 and sample sizes n1= n2= 25. But in the right panel, the distributions are not normal and power is only .28.

Simple transformations are sometimes suggested for salvaging methods based on means, such as taking logs. But by modern standards this approach is unsatisfactory: typically distributions remained skewed and the deleterious impact of outliers remains.[10–12]

A seemingly natural way of dealing with outliers is to simply remove them and apply some method for means to the remaining data. This is reasonable provided a compelling argument can be made that the outliers are invalid. But otherwise, this can result in highly inaccurate conclusions regardless of the sample sizes.[4,5,13] The fundamental reason is when extreme values are eliminated from a random sample; the derivation of Student's t test is no longer valid. There are technically sound techniques for dealing with outliers, but they are not obvious based on standard training. The important point here is that these more modern methods can make a substantial difference when analyzing data.

Comments on rank-based methods

Although not the main focus of this review, some brief comments about rank-based methods are warranted. It is noted that nearly all of the more obvious rank-based methods for comparing groups have been improved substantially.[4,5,8,9] The main point here is that they are sensitive to different features of the distributions under study compared to methods based on robust measures of central tendency such as the median.

Consider, for example, the Wilcoxon-Mann-Whitney (WMW) test. The WMW test is sometimes suggested as a method for comparing medians, but under general conditions it does not address this goal.[14,15] Let P be the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. It is known that the WMW test is based on an estimate of P.

But it does not provide a satisfactory test of H_0: P=.5 or an accurate confidence interval for P, roughly because the derivation of the WMW test assumes groups have identical distributions. So in effect, like Student's t, when the WMW method rejects, it is reasonable to conclude that the distributions differ, but the details regarding how they differ and by how much is unclear. Methods have been derived that provide accurate inferences about P even when distributions differ, two of which have been found to perform relatively well in simulations.[4] The main point here is that surely there are situations where information about P has practical value. Simultaneously, information about measures of central tendency is useful too.

Inferences based on robust measures of central tendency

In terms of measures of central tendency, there are two general ways of dealing with outliers. The first is to trim some specified proportion of the data, the best-known example being the usual sample median. The usual sample median guards against outliers simply because it trims all but one or two of the values. For situations where outliers are likely to make up a relatively large proportion of the data, the median can result in much higher power compared to other measures of central tendency that might be used. But when outliers are relatively rare, such as when sampling from a normal distribution, the sample median can result into relatively low power, roughly because it trims nearly all of the data. A way of dealing with this issue is to trim fewer observations with the goal of guarding against the negative impact of outliers but still having a reasonably high power when dealing with distributions for which outliers are relatively rare.

A γ trimmed mean is

$$\hat{X}_t = \frac{1}{n-2g}\left(X_{(g+1)} + \cdots + X_{(n-2g)}\right)$$

where $0 \le \gamma < .5$, and $X(1) \le X(2) \le \ldots \le X(n)$ are the observations written in ascending order and $g = \lfloor \gamma n \rfloor$, where $\lfloor \gamma n \rfloor$ is the value of γn rounded down to the nearest integer. This estimator guards against outliers because the g smallest and g largest values are trimmed.

An argument for using γ = .2 is that under normality, it performs about as well as the sample mean in terms of power, while simultaneously providing a reasonably good protection against outliers. Also, both theoretical results and simulation studies indicate that as the amount of trimming increases, concerns about controlling the Type I error probability diminish. However, for extreme amounts of trimming, including the median, the Tukey—McLaughlin breaks down. A method that does perform well[4,5] is called a percentile bootstrap technique and is described momentarily.

Testing hypotheses based on a trimmed mean might seem straightforward: Apply Student's t to the data not trimmed. But this approach performs poorly regardless of how large the sample size might be. There are technically sound methods for testing hypotheses based on a trimmed mean, the most basic method being one derived by Tukey and McLaughlin,[16] but the computational details are not provided. Rather, the focus is on the relative merits of a trimmed mean.

The second strategy for dealing with outliers is to empirically determine which values are unusually large or small and then down weight or eliminate them. For example, one could eliminate any outliers using the MAD-median rule. There are several ways of proceeding and their relative merits are described in.[4,5] (The computational details are not important for present purposes and therefore not provided.) Currently, the best hypothesis testing technique when using this strategy is a percentile bootstrap method.

At some level, this second strategy would seem to be preferable to using a trimmed mean. But this issue is not straightforward.[4,5] Indeed, Wu[17] ran simulations based on data from several dissertations. Trimmed means did not dominate, but in terms of power, generally 20% trimmed mean performed better than methods based on the strategy of removing only values that are flagged as outliers.

To describe the basic percentile bootstrap method, suppose that θ_1 and θ_2 are the population medians associated with two independent group sand consider the goal of testing H_0:θ_1=θ_2. The method begins by randomly sampling observations from the first group (with replacement), doing the same for the second group, and observing whether the median for the first groups is less than the median for the second. This process is repeated many times and the proportion times the median from the first group is less than the median from the second is noted as denoted by P*. Then a p-value is 2P* or 1−2P*, whichever is smaller. (When there are tied values, a slight modification of this process is used.) A confidence interval for the difference between the population medians is readily computed as well. In the process just described, at each step compute the difference between the medians. Put these differences in ascending order in which case the middle 95% from a .95 confidence interval. This method continues to work very well when using a 20% trimmed mean instead, but when using the mean, it performs poorly.

It is noted that all of the usual ANOVA designs can be analyzed using robust measures of central tendency[4,5] Moreover, they can be easily applied using the software described in section 6.

Comments on comparing quantiles

When distributions are skewed, robust estimators can better reflect the typical response compared to the mean. However, situations are encountered where comparing the tails of distributions, via some quantile, is important and useful as illustrated in section 7. (A quantile is just a percentile divided by 100. The population median is the .5 quantile). A seemingly natural strategy is to use a simple generalization of percentile bootstrap method just described. However, when using the better-known estimates of quantiles, this approach performs poorly in terms of controlling the Type I error probability. A method that does perform well in simulations is available.[18,19] It uses a percentile bootstrap method in conjunction with a quantile estimator that is based on a weighted average of all of the data.

Robust regression estimators

Least squares regression inherits all of the concerns associated with the mean, and new concerns are introduced. For example, outliers among the dependent variable can destroy power and outliers among the independent variable can result in a poor fit to the bulk of the data. From a technical point of view, one can remove outliers among the independent variable, typically called leverage points, and still test hypotheses in a technically sound manner when using least squares regression. However, this is not the case when dealing with outliers among the dependent variable. Regarding outliers among the independent variable, it is noted there are two kinds: good and bad, which are illustrated in Figure 2. Roughly, a good leverage point is one that is consistent with the regression line associated with the bulk of the data. One positive appeal of a good leverage point is that it results in a smaller standard error compared to situations where there are no

leverage points. A bad leverage point has the potential to substantially affect the least squares regression line in a manner that completely distorts the association among the majority of the data.

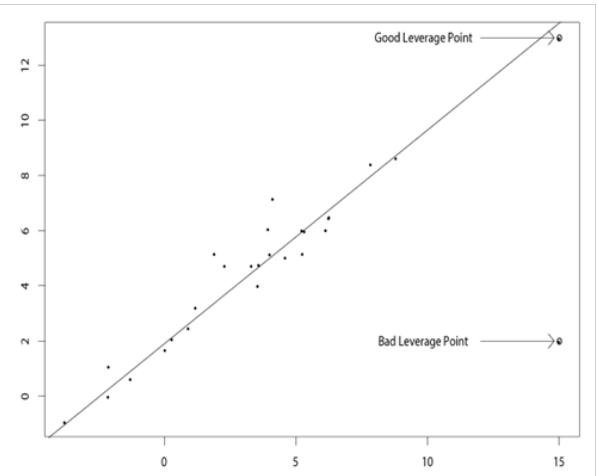

Figure 2 An illustration of good and bad leverage points.

Numerous robust regression estimators have been proposed [5]. Here, to illustrate the extent they can make a practical difference compared to using the least squares regression estimator, the focus is on the Theil–Sen estimator.[20,21] This is not to suggest that it dominates all other robust regression estimators that might be used. It does not dominate and indeed there are good arguments for seriously considering other regression estimators.

For a random sample $(X_1, Y1), \ldots, (X_n, Yn)$, the Theil–Sen estimator is computed as follows. For any i<i', for $X_i \neq X_i$

$$S_{ii'} = \frac{Y_i - Y_{i'}}{X_i - X_{i'}}.$$

The Theil–Sen estimate of the slope is b1ts, the median of all the slopes represented by Sii'. The intercept is estimated with My − b1Mx, where My and Mx are the usual sample medians of the y and x values, respectively. Several methods for extending this estimator to more than one independent variable have been proposed,[5] but no details are given here. (For some theoretical properties of the Theil—Sen estimator[22] In terms of testing hypotheses based on a robust regression estimator, currently the best way of handling heteroscedasticity, as well as robust estimators that might not be asymptotically normal, is to use a percentile bootstrap method. This is easily done with extant software mentioned in section 7.

A criticism of the Theil–Sen estimator is that when testing the hypothesis that the slope is zero, power might be relatively poor when there are tied (duplicated) values among the dependent variable. A modification of the Theil–Sen estimator is available that deals with this problem.[23]

Another general approach is to determine the slopes and intercept that minimize some robust measure of variation applied to the residuals. A somewhat related approach is least trimmed squares where the goal is to determine the slopes and intercepts so as to minimize the sum of the squared residuals, ignoring some specified proportion of the largest residuals. (This strategy has some similarities to a trimmed mean.) Yet another general approach is based on what are called M-estimators. A crude description is that M-estimators empirically down weight or eliminates unusually large or small residuals. (The formal definition of an M-estimator is more involved.) Two that perform relatively well are the Coakley and Hettmansperger[24] estimator and the MM-estimator derived by Yohai.[25] There are conditions where all of these estimators offer a substantial advantage over least squares regression in terms of power, control over the probability of a Type I error, and the ability to avoid misleading summaries of data due to outliers. When standard assumptions are true, least squares does not offer much of an advantage. Although the Theil–Sen estimator and the MM-estimator seem to perform relatively well, there are practical reasons for considering other estimators not listed here.[5]

Two things are stressed. First, even among the better robust regression estimators, the choice of estimator can make a practical difference as illustrated in. Second, although the better robust regression estimators are designed to deal with outliers among both the independent variables and the dependent variable, removing outliers among the independent variables can make an important practical difference.[5]

Comparing regression lines

When using robust regression estimators, it is briefly noted that methods for comparing the slopes and intercepts associated with two or more independent groups are available. For recent results, see.[26] There are also robust analogs of ANCVOA (analysis of covariance) that do not assume parallel regression lines.[4,5,27] For results on comparing the slopes and intercepts associated with dependent groups.[28] These methods are readily applied with the software illustrated in section 7.

Curvature

There are many nonparametric methods for estimating a regression line, generally called smoothers, the bulk of which are focused on estimating the conditional mean of Y. Smoothers provide a flexible approach to approximating a regression surface without specifying a particular parametric regression model. Experience makes it clear that the more obvious parametric models for dealing with curvature can be highly unsatisfactory, particularly when dealing with more than one predictor. Some illustrations supporting this statement are given in section 7. Modern methods for dealing with curvature can reveal strong associations that otherwise would be missed using the more traditional techniques, as will be seen.

Imagine that the typical outcome of some dependent variable is given by some unknown function $m(X_1, \ldots, X_p)$ of the p predictors $(X_1, \ldots, X_p)$. A crude description of smoothing techniques is that they identify which points are close to $(X_1, \ldots, X_p)$, and then some measure of location, such as a trimmed mean or the median, is computed based on the corresponding Y values. The result is $\hat{m}(X_1, \ldots, X_p)$, an estimate of the measure of location associated with Y at the point $(X_1,...,X_p)$. The estimator is called a smoother, and the outcome of a smoothing procedure is called a smooth.[29]

When dealing with a robust measure of central tendency such as a trimmed mean, a relatively simple but effective method is the running interval smoother. For more precise details, see section 11.5.4 in.[5] For the special case where the goal is to estimate quartiles or some other percentile of interest, see.[30–32] The main point here is that the plots of the regression line resulting from a smoother can be invaluable when trying to detect and describe an association. Moreover, there are inferential methods based on smoothers. For example, one can test the hypothesis that a regression line is straight or that it has a particular parametric form. Smoothers provide a much more flexible approach to comparing groups when there is a covariate. These modern ANCOVA

methods effectively deal with both curvature (regression lines that are not straight) and non-normality.

Measures of association

Even with a large sample size, Pearson's correlation is not robust: outliers can have an inordinate impact on its value. Moreover, the population Pearson correlation ρ is not robust in the sense that arbitrarily small changes in the tails of the (marginal) distributions can have a substantial impact on its value.

The left panel of Figure 3 shows a bivariate normal distribution having correlation .8 and the right panel shows a bivariate normal distribution having correlation .2. Now look at Figure 4. This bivariate distribution indicates a strong association from a graphical perspective: it has an obvious similarity with the left panel of Figure 3, yet Pearson's correlation is only .2. The reason is that one of the marginal distributions has a mixed normal distribution. Chapter 9 in[5] summarizes various strategies for dealing with this problem. Some of these methods deal with outliers among the marginal distributions without taking into account the overall structure of the data. Two well-known examples are Spearman's rho and Kendall's tau. Others take into account the overall structure of the data when checking for outliers. As a simple example of what this means, it is not unusual for someone to be young and it is not unusual for someone to have heart disease, but it is unusual for someone to be both young and have heart disease. One approach to measuring the strength of the association among the bulk of the points is to use a skipped correlation, meaning that outliers are removed using a method that takes into account the overall structure of the data and then Pearson's correlation is applied to the remaining data. In,[4] p. 705, for an illustration that this approach can better deal with outliers than Spearman's rho and Kendall's tau see,[4] p. 705.

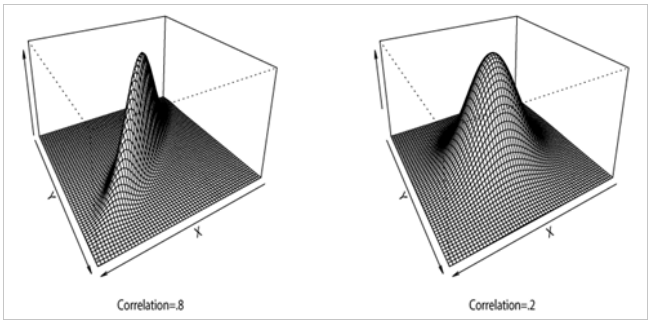

Figure 3 When both X and Y has a normal distribution, increasing ρ from .2 to .8 has a noticeable effect on the bivariate distribution of X and Y.

Figure 4: Two bivariate distributions can appear to be very similar yet have substantially different correlations. Shown is a bivariate distribution with ρ = .2, but the graph is very similar to the left panel of Figure 2 where ρ = .8

Another approach is to measure the strength of an association based on a fit to the data. This can be done using what is called explanatory.[5]

Software

Of course, modern robust methods have no practical value without access to appropriate software. In terms of being able to apply a wide range of robust methods, the software R is easily the best choice. For an extensive summary of R functions for applying robust methods, with illustrations.[4,5] The illustrations in section 7 are based on these R functions. (The R functions that were used here are stored in the file Rallfun-v24, which can be downloaded from http://college.usc.edu/labs/rwilcox/home.) In contrast, SPSS is a poor choice in terms of gaining access to modern methods.

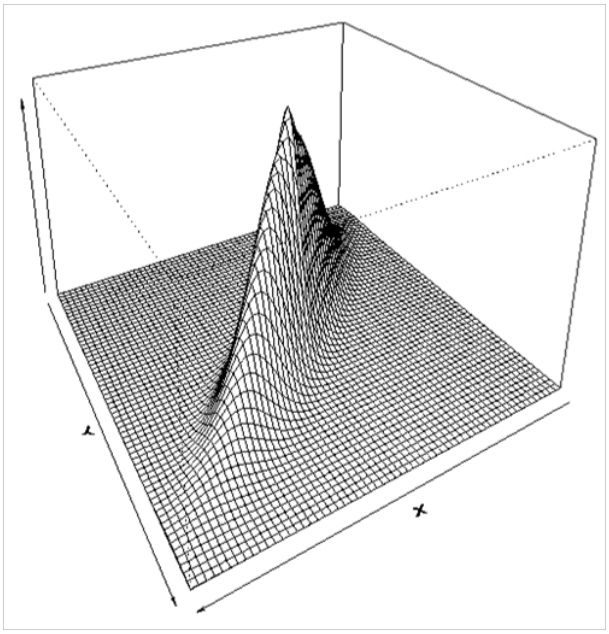

Figure 4 Two bivariate distributions can appear to be very similar yet have substantially different correlations. Shown is a bivariate distribution with ρ = .2, but the graph is very similar to the left panel of Figure 2 where □ = .8.

Some illustrations

The first example is based on data from a sexual attitude study.[33] One portion of the study asked participants how many sexual partners they wanted during the next thirty years. There were 105 males and 156 females and an issue is whether the responses from males differ in any way from females. The sample means for the males and females are 64.9 and 2.8, respectively. (There was one extreme response among the males: 6000. The next largest response was 150.) Comparing the means using Welch's test (via the built-in R function t. test), the p-value is p = .28. Moreover, the median response for both groups is 1, indicating that the typical male response does not differ from the typical response among females. But despite these results, there is an indication that the distributions differ in the upper tails. Comparing the upper quartiles (using the R function qcomhd with the argument q=.75), a significant difference is found (p < .001); the estimated difference between the .75 quantiles is 6.77-2.81=3.96. It is noted that a significant difference is also found using a 1% trimmed mean, which effectively eliminates the one extreme outlier among the males.

The next example stems from a study dealing with the effects of consuming alcohol and was supplied by M. Earleywine. Group 1 was a control group and measures reflect hangover symptoms after consuming a specific amount of alcohol in a laboratory setting. Group 2 consisted of sons of alcoholic fathers. The sample size for both groups is 20. Comparing means, the estimated difference is 4.5, p = .14. The .95 confidence interval is (−1.63, 10.73). Box plots indicate that the data are skewed with outliers, casting doubt on the accuracy of the confidence interval based on means. Using 20% trimmed means (R function yuenv2) yields an estimated difference of 3.7, p = .076 and a .95 confidence interval (−0.456, 7.788). Note that the length of the confidence intervals differ substantially; the ratio of the lengths is .67. Using instead a percentile bootstrap method, again using a 20% trimmed mean (via the R function trimpb2), p = .0475 and now the .95

confidence interval is (0.083, 8.333), the point being that the choice of method can make a practical difference. Comparing medians (with the R function pb2gen and the argument est=hd) gives similar results: p = .038 and a .95 confidence interval (0.0955, 8.498).

The impact of removing leverage points, when using least squares regression, is illustrated in Figure 5. The left panel deals with data on the average influent nitrogen concentration (NIN) in 29 lakes and the mean annual total nitrogen (TN) concentration. Note the obvious leverage points in the lower right portion of the plot. The nearly horizontal (solid) line is the least squares regression line using all of the data. Based on the usual Student's t test, the p-value associated with the slope is .72. The other regression (dotted) line is the least squares regression line when the leverage points are removed.

The right panel in Figure 5 is from a reading study where a measure of digit naming speed (RAN1T) was used to predict the ability to identify words (WWISST2). The nearly horizontal (solid) line is the least squares regression line using all of the data and the other (dotted) line is the regression line when the obvious leverage points are removed.

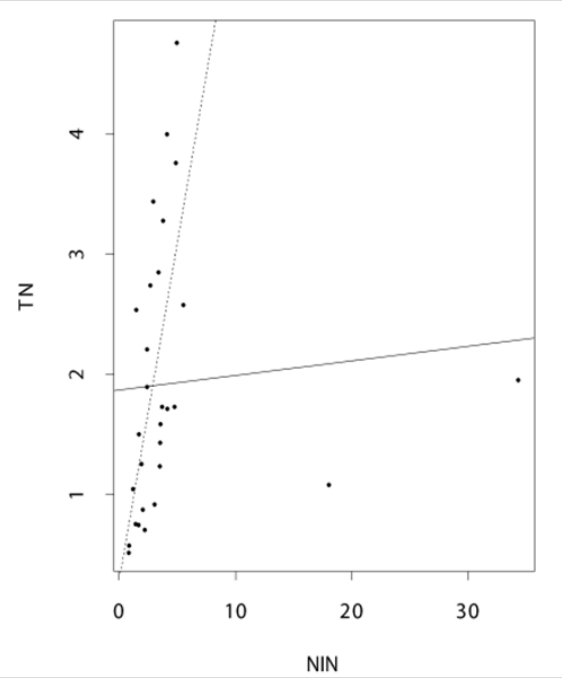

Figure 5 When using least squares regression, eliminating leverage points based on the MAD-median rule can make a substantial difference as illustrated with the lake data in the left panel and the reading data in the right panel. (NIN is on the average influent nitrogen concentration in 29 lakes and TN is the mean annual total nitrogen concentration. The nearly horizontal lines are the least squares regression lines with leverage points included. In the right panel RAN1T is a measure of digit naming speed and WWISST2 is a measure of the ability to identify words.

Robust regression estimators have the advantage of being able to deal with outliers among the dependent variable. For the situation just considered, switching to a robust regression estimator makes little difference when testing the hypothesis that the slope is zero. But to illustrate that a robust regression estimator can make a substantial difference, consider again the reading data, only now WWISST2 is taken to be the independent variable and RAN1T is the dependent variable. Using least squares and the usual Student's t test of the hypothesis of a zero slope, p = .76; the estimate of the slope is −0.056. Using the modified Theil–Sen estimator in,[23] via the R function regci (with the argument regfun=tshdreg), now the estimate of the slope is −.23 and p = .007. Using the MM-estimator (by setting regfun=MMreg), p < .001.

The next set of examples stem from the Well Elderly 2 study.[34,35] One portion of the study dealt with the association between the cortisol awakening response (CAR) and various measures of stress and well-being. (The CAR is the change in an individual's cortisol level upon awakening and measured again 30-60 minutes later). One of the measures, labeled SF36, reflected self-reports of physical health after intervention. Least squares regression finds no significant association based on Student's t, p = .99 and the estimate of the slope was .01. Again, outliers among the independent variable can mask a true association. Removing them (with the R function ols and outliers among the independent variable removed by setting the argument xout=TRUE), p = .082 and the slope is now estimated to be −7.1. Again using least squares and the R function olshc4 with the goal of dealing with heteroscedasticity, p = .03 (with the argument xout=TRUE), and the slope is estimated to be −9.47. (With xout=FALSE, meaning that outliers among the independent variable are retained, p = .997.) Testing the hypothesis that the slope is zero via the R function regci (with the arguments xout=TRUE and regfun=tshdreg), p = .007 and the slope is estimated to be −6.2. Note that the p-values range between .007 and .99 depending on which method is used, and the slope estimates range between .01 and −6.2.

However, look at the right panel of Figure 6, which shows a nonparametric estimate of the regression line based on a running interval smoother. (The R function rplot was used with argument xout=TRUE and the argument est=tmean, which means that the conditional 20% trimmed of SF36, given CAR, is being estimated.) A test of the hypothesis that the regression line is straight (using the R function lintest) was significant (p < .001). The plot suggests a distinct bend close to where CAR is equal to −.1. Using only the data for which the CAR is greater than −.1, with outliers among the CAR values removed, the slope differs significantly from zero (p = .003), with the strength of the association estimated to be .11. For CAR less than −.1, no association is indicated, the point being that close attention to curvature can be important from a substantive point view.

Next consider the association between the CAR and a measure of depressive symptoms, which is labeled CES-D. The left panel of Figure 6 shows a nonparametric estimate of the regression line. (Again the R function rplot was used with xout=TRUE.) Now there appears to be a distinct bend close to where CAR is equal to zero. Testing the hypothesis that the regression line is straight (via the R function lintest), the p-value is less than .001. Focusing only on the data for which CAR is positive, the modified Theil–Sen estimator yields a significant result (p = .017), the slope is estimated to be 33.5 and a measure of the strength of the association (using a generalization of Pearson's correlation) is .27 (obtained with the R function tshdreg and the argument xout=TRUE). For CAR negative, no association is found. So the results suggest that when CAR is positive (cortisol decreases shortly after awakening), typical CES-D scores tend to increase. But when CAR is negative, there appears to be little or no association with CES-D.

Cortisol and dehydroepiandrosterone (DHEA) are considered to be valuable markers of the hypothalamus-pituitary-adrenal (HPA) axis, while salivary alpha-amylase (sAA) reflects the autonomic nervous system. The final example deals with the association between DHEA and cortisol, taken as the independent variables, and sAA, measured upon awakening, again using the Well Elderly data. Figure 7 shows an estimate of the regression surface. (The R function lplot was used

with the argument xout=TRUE, which eliminates outliers among the independent variables.) Figure 7 indicates that typical sAA values tend to be highest when simultaneously cortisol is low and DHEA is relatively high.

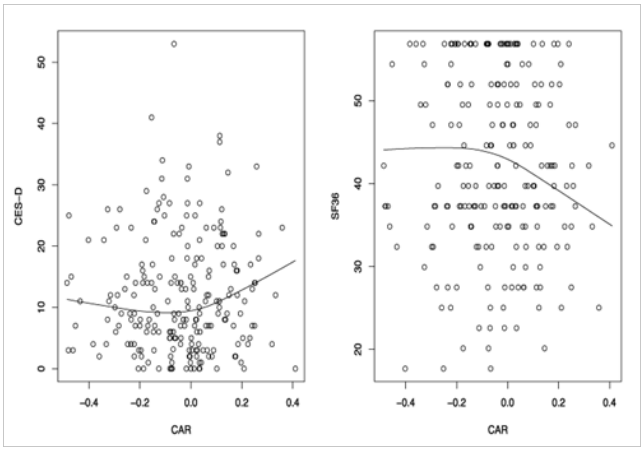

Figure 6 The left panel shows the estimated regression line for predicting CES-D using the cortisol awakening response (CAR) based on a non-parametric regression estimator called a running interval smoother. The right panel is the running interval smoother for predicting SF-36.

However, Figure 7 suggests that curvature might be an issue with the nature of the association changing as DHEA increases. To check on this possibility, the data were split into two groups according to whether DHEA is less than 100 pg/mL. For DHEA less than 100 pg/mL, no association is found. But for DHEA greater than 100 pg/mL, the slopes for both cortisol and DHEA are significant ($p = .037$ and .007, respectively). Now the slope associated with cortisol is negative and it differs significantly from the slope when DHEA is less than 100 pg/mL ($p = .027$).

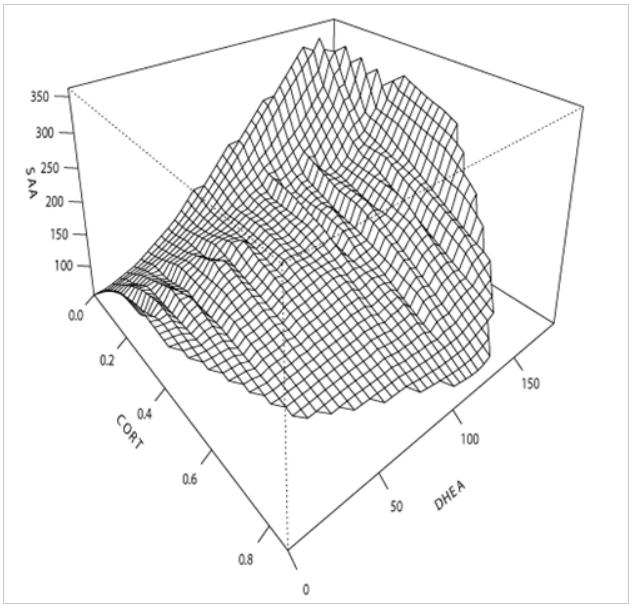

Figure 7 The estimated regression surface for predicting sAA using the cortisol awakening response (CAR) and DHEA. Fitting a regression plane, both slopes are positive with the slope for cortisol non-significant ($p = .81$) and the slope for DHEA significant ($p = .01$).

The first example in this section (based on the sexual attitude data) illustrated that the measure of location that is used can make a substantial difference. No significant differences were found using means or medians, but differences were found in terms of a 1% trimmed mean and the .75 quantiles. This final example is similar in the sense that significant differences in the tails of the distribution are indicated, but no differences are found using medians, 20% trimmed means and a modified one-step M-estimator. The data are again from the Well Elderly study and the goal was to compare a control group to an experimental group that received intervention, based on CES-D. Figure 8 shows a plot of the distributions. (The solid line corresponds to the control group.) Note that there appears to be little difference between the distributions near the modes, but the tails appear to differ. Comparing the .8, .85 and .9 quantiles, the corresponding p-values are .061, .006 and .012, respectively. From a substantive point of view, most participants had relatively low depressive measures. So intervention would seem unlikely to lower CES-D scores. But among individuals with relatively high CES-D scores, participants in the intervention group were found to have significantly lower measures of depressive symptoms.

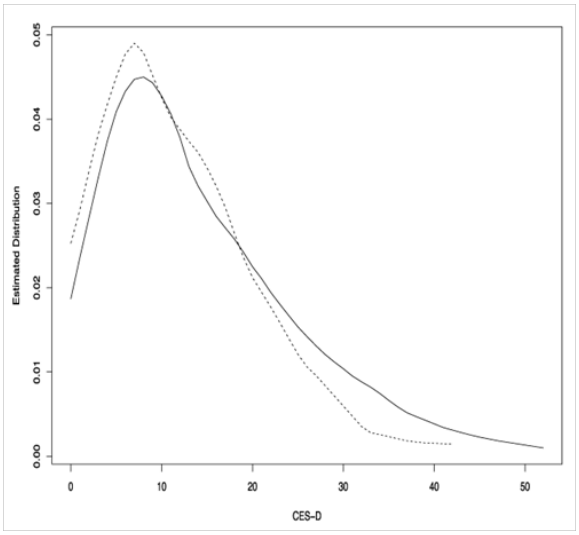

Figure 8 Distribution of CES-D for a control group (solid line) and a group that received intervention.

Conclusion

There are many issues and techniques beyond those outlined and illustrated in this review. Taken as whole, these new and improved methods raise a seemingly natural question: Which method is best? It is suggested, however, that there is a better question: How many methods are required for understanding how groups compare and the association among two or more variables? As was illustrated in section 7, different conclusions and perspectives can result depending on which method is used. The relative merits of many methods have been studied extensively. But in the end, no single method dominates and as illustrated here, the choice of method can make a practical difference. Of course, testing multiple hypotheses raises the issue of controlling the probability of one more Type I error probabilities. But without multiple perspectives, researchers are severely limited in their quest to understand data. A possible way of dealing with this issue is to put more emphasis on exploratory studies. The main point is that we have the technology for learning more from data. All that remains is taking advantage of what this technology has to offer.

Acknowledgments

None.

Citation: Wilcox RR. Gaining a deeper and more accurate understanding of data via modern robust statistical techniques. *J Psychol Clin Psychiatry.* 2014;1(2):48–55. DOI: 10.15406/jpcpy.2014.01.00012

Conflicts of interest

Author declares there are no conflicts of interest.

Funding

None.

References

1. Staudte RG, Sheather SJ. Robust Estimation and Testing. *J*ohn Wiley & Sons, Inc., USA. 1990.
2. Rousseeuw PJ, Leroy AM. Robust Regression & Outlier Detection, John Wiley & Sons, Inc., USA. 1987.
3. Heritier S, Cantoni E, Copt S, et al. Robust Methods in Biostatistics. John Wiley & Sons, Inc., USA. 2009.
4. Wilcox RR. Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction. Chapman & Hall/CRC press. 2011.
5. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. (3rd edn), Elsevier. 2012.
6. Huber PJ, Ronchetti E. Robust Statistics. (2nd edn), John Wiley & Sons, Inc., USA. 2009.
7. Maronna RA, Martin DR, Yohai VJ. Robust Statistics: Theory and Methods. *John Wiley & Sons,* Inc., USA. 2006.
8. Brunner E, Domhof S, Langer F. Nonparametric Analysis of Longitudinal Data in Factorial Experiments. Wiley, New York, USA. 2002.
9. Cliff N. Ordinal Methods for Behavioral Data Analysis. Lawrence Erlbaum Associates, Mahwah, *New Jersey*, USA. 1996.
10. Doksum KA, Wong CW. Statistical tests based on transformed data. *Journal of the American Statistical Association*. 1983;78(382):411–417.
11. Rasmussen JL. Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*. 1989;42(2):203–213.
12. Wilcox RR, Keselman HJ. Modern robust data analysis methods: Measures of central tendency. *Psychol Methods*. 2003;8(3):254–274.
13. Bakker M, Wicherts JM. Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in t Tests. *Psychol Methods (in press)*. 2014.
14. Hettmansperger TP. Statistical Inference Based on Ranks. Wiley. 1984.
15. Fung KY. Small sample behaviour of some nonparametric multi-sample location tests in the presence of dispersion differences. *Statistica Neerlandica.* 1980;34(4):189–196.
16. Tukey JW, McLaughlin DH. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/ Winsorization 1. *Sankhya A*. 1963;25(3): 331–352.
17. Wu P-C Central limit theorem and comparing means, trimmed means one-step M-estimators and modified one-step M-estimators under non-normality. Unpublished doctoral dissertation, Dept. of Education, University of Southern California. 2002.
18. Wilcox RR, Erceg-Hurn D. Comparing two dependent groups via quantiles. *Journal of Applied Statistics*. 2012;39(12):2655–2664.
19. Wilcox RR, Erceg-Hurn D, Clark F, et al. Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*. 2013;84(7):1543–1551.
20. Theil H. A rank-invariant method of linear and polynomial regression analysis Indaga- tiones Mathematicae. 1950;12(5):85–91.
21. Sen PK. Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*. 1968;63(324):1379–1389.
22. Peng H, Wang S, Wang X. Consistency and asymptotic distribution of the Theil– Sen estimator. *Journal of Statistical Planning and Inference*. 2008;138(6):1836–1850.
23. Wilcox RR, Clark F. Robust regression estimators when there are tied values. *Journal of Modern and Applied Statistical Methods.* 2013;12(2).
24. Coakley CW, Hettmansperger TP. A bounded influence, high breakdown, ef- ficient regression estimator. *Journal of the American Statistical Association*. 1993;88(423):872–880.
25. Yohai VJ. High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*. 1987;15(2):642–656.
26. Wilcox RR, Clark F (in press) Heteroscedastic global tests that the regression parame- ters for two or more independent groups are identical. Communications in Statistics– Simulation and Computation.
27. Wilcox RR. A heteroscedastic method for comparing regression lines at specified design points when using a robust regression estimator. *J Data Sci*. 2013;11(2): 281–291.
28. Wilcox RR, Clark F (in press) Comparing robust regression lines associated with two dependent groups when there is heteroscedasticity. Computational Statistics.
29. Tukey JW. Exploratory Data Analysis. Pearson Education Inc., USA.1977.
30. Harrell FE, Davis CE. A new distribution-free quantile estimator. *Biometrika*. 1982;69(3):635–640.
31. Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika.* 1944;81(4):673–680.
32. Wilcox RR, Clark F. Comparisons of two quantile regression smoothers. Unpublished technical report. 2004.
33. Pedersen WC, Miller LC, Putcha-Bhagavatula AD, et al. Evolved sex differences in sexual strategies: The long and the short of it. *Psychological Science*. 2002;13(2):157–161.
34. Clark F, Azen SP, Zemke R, et al. Occupational therapy for independent-living older adults. A randomized controlled trial. *JAMA.* 1997;278(16):1321–1326.
35. Jackson J, Mandel D, Blanchard J, et al. Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials* . 2009;6(1):90–101.