

[Page One](#)[Campus
Computing News](#)[Simeon Support
to be Phased
Out](#)[The Force is just
around the
corner ... really!](#)[Faculty/Staff
NDS Directory
Update](#)[Today's Cartoon](#)[RSS Matters](#)[SAS Corner](#)[The Network
Connection](#)[List of the Month](#)[WWW@UNT.EDU](#)[Short Courses](#)[IRC News](#)[Staff Activities](#)[Subscribe to
Benchmarks
Online](#)

Research and Statistical Support

University of North Texas

RSS Matters

An Introduction to Robust Measures of Location Using GNU S-Plus

By [Dr. Rich Herrington](#), Research and Statistical Support Consultant

This month we demonstrate the use of robust estimators of location using the GNU S-Plus language, "R". R is a statistical programming environment that is a clone of the S and S-Plus language developed at Lucent Technologies. In the following document we will illustrate the use of a GNU Web interface to the R engine on the "rss" server, <http://rss.acs.unt.edu/cgi-bin/R/Rprog>. This GNU Web interface is a derivative of the "Rcgi" Perl scripts available for download from the CRAN Website, <http://www.cran.r-project.org> (the main "R" Website). Scripts can be submitted interactively, edited, and re-submitted with changed parameters by selecting the hypertext link buttons that appear below the figures. For example, clicking the button below:

opens a new window with a "browser entry form" where the program code that has been submitted is displayed. The script can be edited and re-submitted to produce a new program output. Scrolling down the browser window displays text from the program execution. Selecting the "Display Graphic" link will open another browser window where graphics will be displayed. Readers are encouraged to change program parameters to see what the effect will be on results.

Introduction to Robust Estimation

Conventional wisdom has often promoted the view that standard ANOVA techniques are robust to non-normality. However, this view is with respect to type I error (Wilcox, 1998). When it is assumed that there are no differences between groups in a group difference testing setting, then the probability level corresponding to the critical cut-off score, used to reject the null hypothesis, is found to be close to the nominal level of .05. However, many statistical journals have pointed out that standard methods are not robust when differences exist (Hampele, 1973; Tukey, 1960). As early as 1960, it was known that slight deviations away from normality could have a large negative impact on power whenever means were being compared, and that popular measures of effect size could be misleading (Tukey, 1960). Later, a theory of robustness was developed by Huber (1964) and Hampel (1968). Today, there is a well established mathematical foundation for dealing with these issues (Huber, 1981; Rousseeuw & Leroy, 1987). Moreover, basic coverage of the theory and the use of computer software in performing robust analyses can be found in introductory textbooks (Rand Wilcox, 1997, 2001).

Dealing with Outliers

It is often assumed in the social sciences that data conform to a normal distribution. Numerous studies have examined real world data sets for conformity to normality, and have strongly questioned this assumption (Hampel, 1973; Tukey, 1960; Micceri, 1989; Stigler, 1977). Sometimes we may believe that a normal distribution is a good approximation to the data, and at other times we may believe this to be only a rough approximation. Two approaches have been taken to incorporate this reality. One approach is a two-stage process whereby influential observations are identified and removed from the data. So-called outlier analysis involves the calculation of leverage and influence statistics to help identify influential observations (Rousseeuw & Leroy, 1987). The other approach, robust estimation, involves calculating estimators that are relatively insensitive to the tails of a data distribution, but which conform to normal theory approximation at the center of the data distribution. These robust estimators are somewhere between a nonparametric or distribution free approach, and a parametric approach. Consequently, a robust approach distinguishes between plausible distributions the data may come from, unlike a nonparametric approach, which treats all possible distributions as equal. The positive aspect of this is that robust estimators are very nearly as efficient (very nearly optimal estimators) as the best possible estimators (Huber, 1981). It is possible to get a sense of how much you can violate the normality assumption before inferences are compromised.

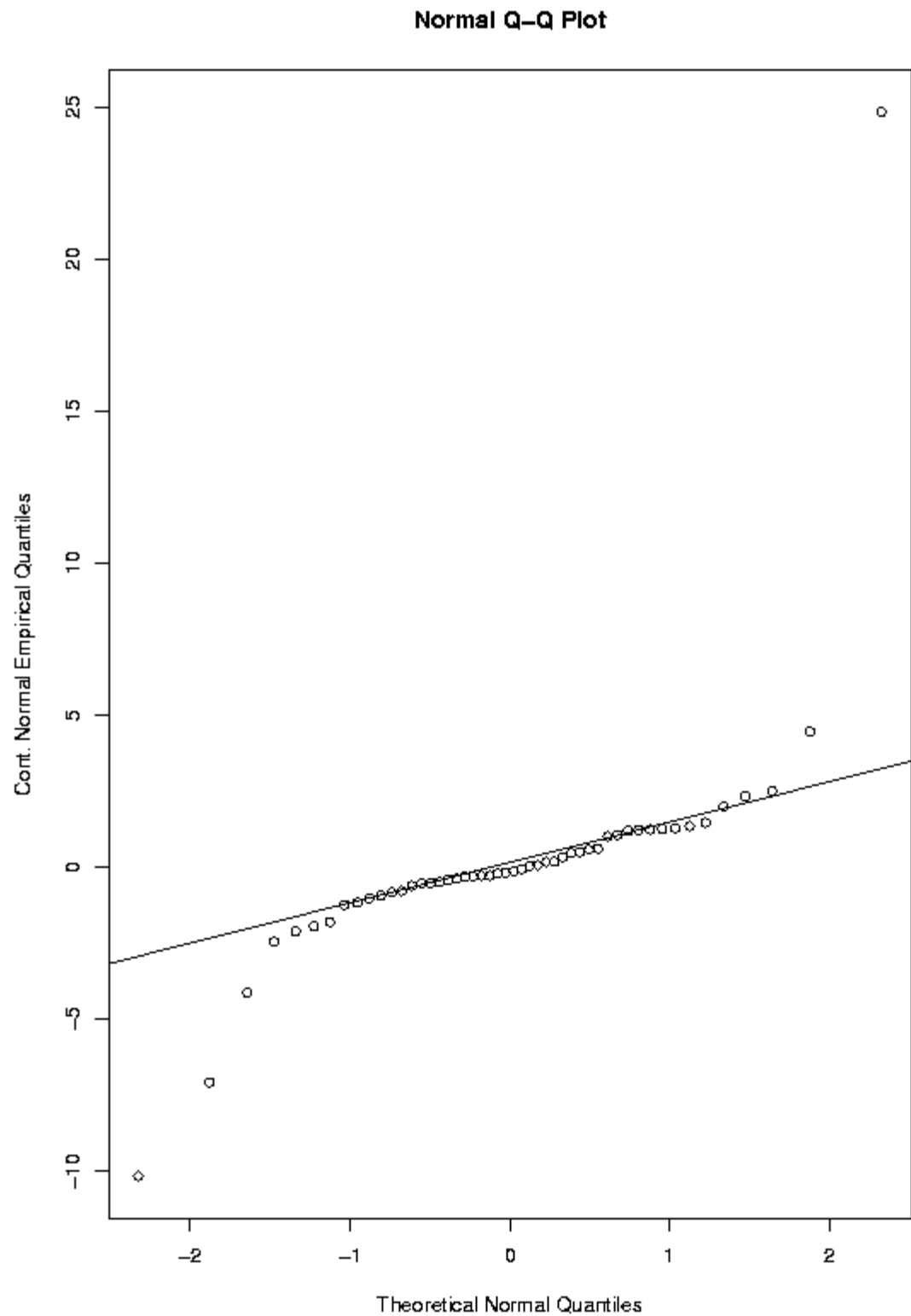
Symmetric and Asymmetric Distributions

Historically, statisticians have focused on estimators that assume symmetry in the population. The reason for this is that estimators of location are best understood when a distribution's natural candidates for location all nearly coincide (e.g. mean, median, mode). Additionally, when a distribution is treated in a symmetric way so that no bias arises, a trade off is not needed between bias and variability (e.g. M-estimators with odd influence functions are unbiased estimators whenever the distribution is symmetric). Moreover, whenever a distribution is admitted as skew, there is some question as to what measure of location we are trying to estimate. That is, asymmetric distributions do not have a natural location parameter as the center of symmetry, of a symmetric distribution (Hoaglin, Mosteller, & Tukey, 1983). It is a common practice to re-express the data, such as in a functional transformation (e.g. log-transformation), so that the data more nearly resembles a symmetric distribution. Often, if the departure from symmetry is not too large, it is found that estimators that rely on symmetry are still satisfactory (Hoaglin, Mosteller, & Tukey, 1983). The use of quantile-quantile plots can aid in the assessment of skewness (see below). In the case of M-estimators for location, we would like the M-estimate to be an unbiased, robust estimate of the population mean. This goal can be realized in the case of a symmetric distribution.

The Contaminated Normal

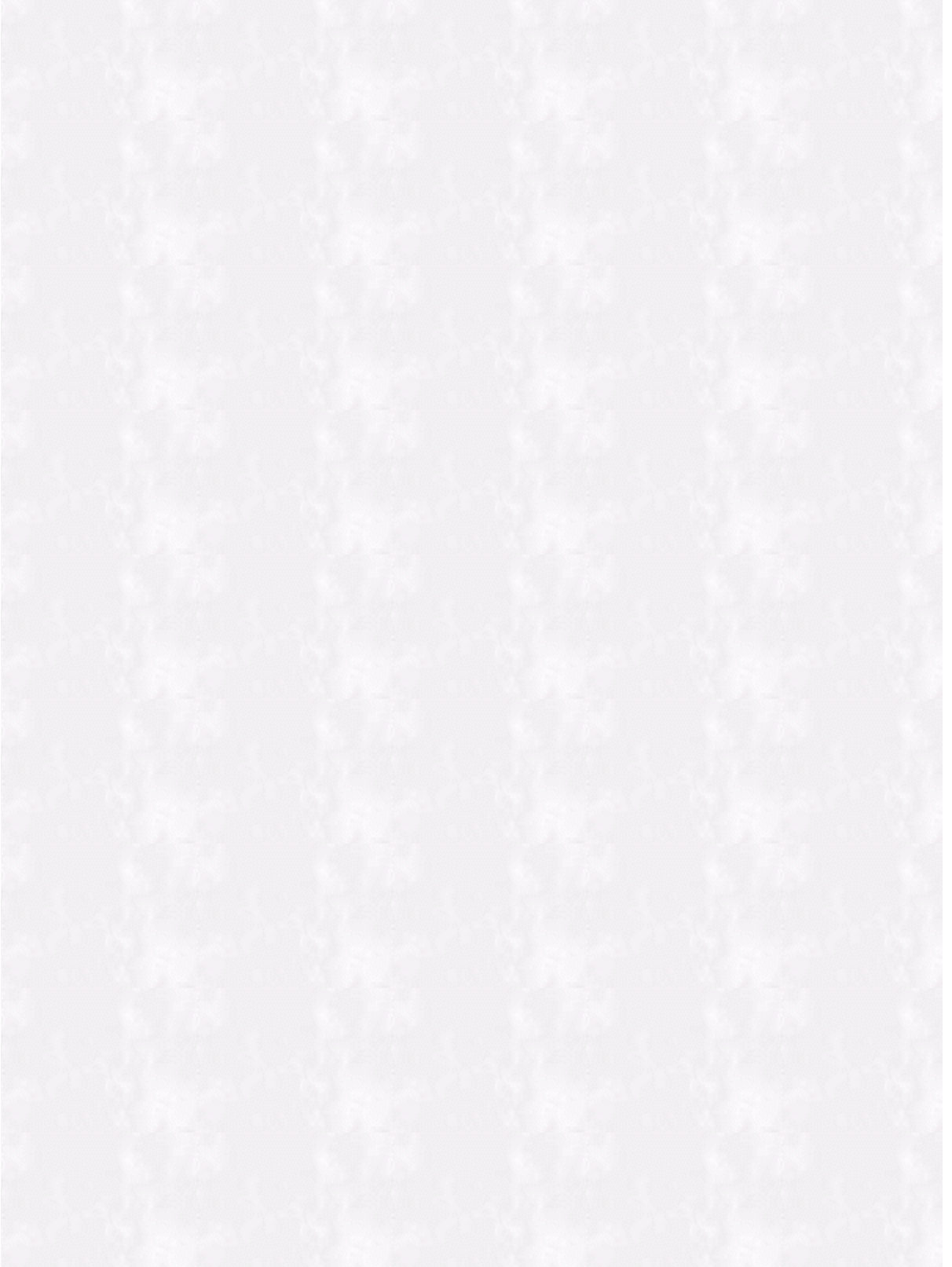
For example, Tukey (1960) showed that classical estimators are quite sensitive to distributions which have heavy tails. The approach Tukey took was to sample from a continuous distribution called the contaminated normal (CN). The contaminated normal is a mixture of two normal distributions, one of which has a large variance; the other distribution is standard normal. The contaminated normal has tails which are heavier, or thicker, than the normal distribution. This can be illustrated by the use of quantile-quantile plots. The empirical quantiles of a data set are graphed against the theoretical quantiles of a reference distribution (i.e. normal distribution). Deviations away from the straight line

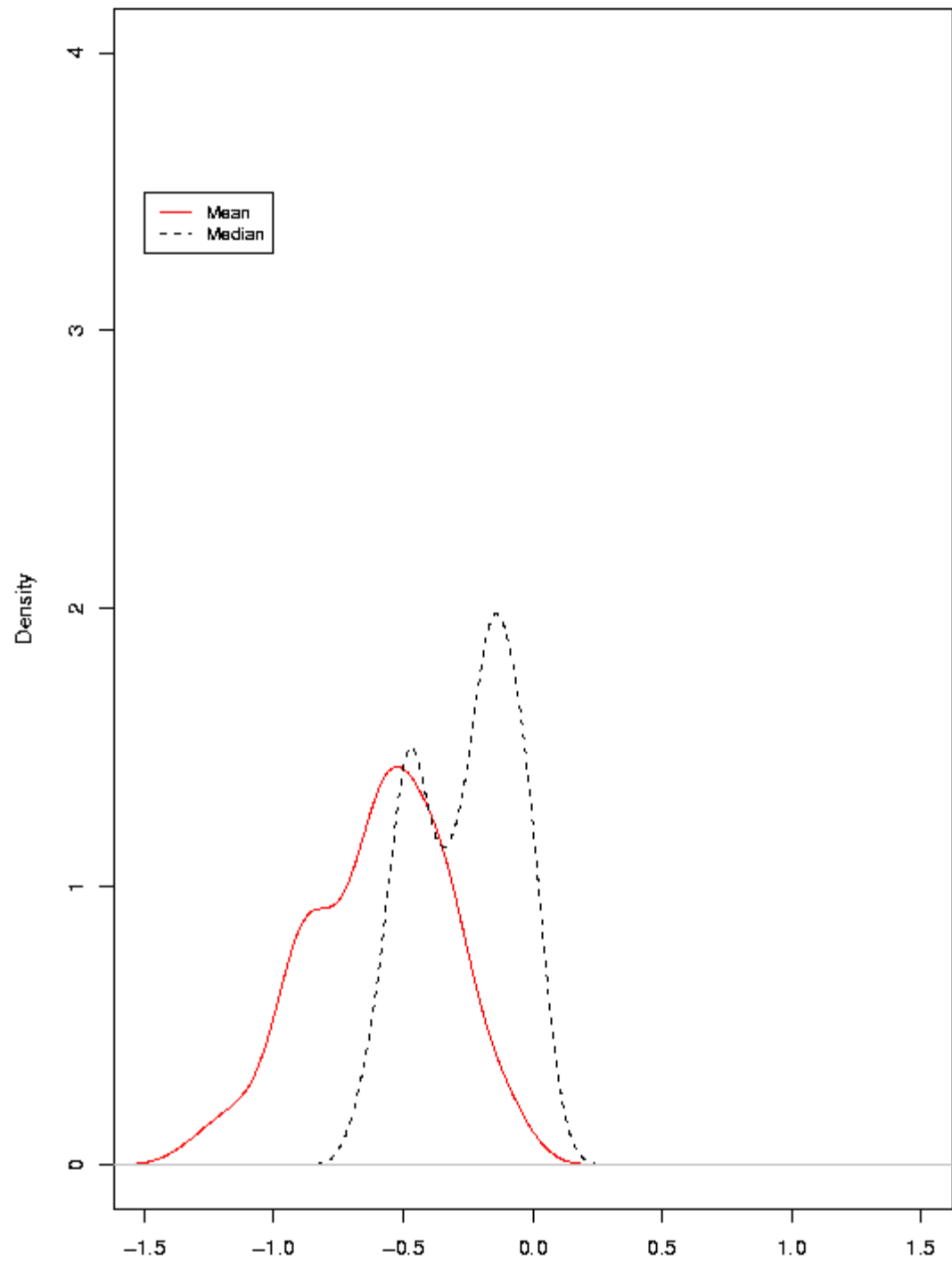
indicate deviations away from the reference distribution. In the figure below, the quantile-quantile plot illustrates a heavy-tailed distribution.



Robust estimators are considered resistant if small changes in many of the observations or large changes in only a few data points have a small effect on its value. For example, the

median is considered an example of a resistant measure of location, while the mean is not. In the figure below, the sampling distributions of the mean and median are plotted when sampling from the contaminated normal distribution (CN). Sampling occurred from the CN distribution where there is a 90% probability of sampling from $N(0, 1)$ and 10% probability of sampling from $N(0, 10)$ and the population mean for the CN is zero. Notice that there is substantially more variability in our estimate of the population mean when using the sample mean to estimate the population mean, than when using the sample median to estimate the population mean. Also, the sample median is closer to the population mean of zero, than is the sample mean.





The Trimmed Mean

One problem with the median however is that its value is determined by only 1 or 2 values

in the data set information is lost. The trimmed mean represents a compromise between the mean and the median (Huber, 1981). The trimmed mean is computed by putting the observations in order. Next, trim the numbers by removing the d largest and d smallest observations, and then compute the average of the remaining numbers. d can be between 0 and $n/2$. Trimming enough data gives the sample median. Rules of thumb are that 20%-25% ($d=.2*n$) trimming works well in a wide range of settings (Wilcox, 1997). Another approach to selecting the trimming amount is to calculate the mean for 0, .10, .20 and then use the trimming value that corresponds to the smallest standard error (Leger and Romano, 1990).

M-Estimators

The trimmed mean is based on a preset amount of trimming. A different approach is to determine empirically the amount of trimming necessary. If the data come from a normal distribution, then light or no trimming is necessary. If the data come from a heavy tailed distribution, then a heavier amount of trimming is desired in both tails. If the distribution has a heavy right tail, then more trimming might be desired from the right tail; or if the distribution has a heavy left tail, more trimming from the left tail might be appropriate. Essentially, M-estimators accomplish this appropriate amount of trimming by meeting certain statistical criterion for what is considered a good estimator (i.e. maximum-likelihood principle). For the M-estimator, the degree of trimming is determined by a trimming constant, k .

Desirable Properties of a Robust Estimator

A good robust estimator is asymptotically consistent and unbiased (the estimator converges on the true population value as sample size increases). Additionally, a good robust estimator should be efficient when the underlying distribution is normal, but still be relatively efficient when the tails of the distribution deviate from normality. That is, the variance of the sampling distribution for the estimator should be small whether we are sampling from a normal or non-normal distribution. When sampling data from a normal distribution, the mean is a minimum variance estimator. That is, the mean is considered an optimal estimator because the variance of its sampling distribution is as small as possible assuming an underlying normal distribution. While the mean is an optimal estimator, it does not possess other characteristics which are associated with a good estimator. Whenever sampling from a non-normal distribution, the mean can lose many of the properties which make it an optimal estimator. Efficient estimators exist for situations where non-normality is present. These estimators are referred to as robust estimators.

Comparing Estimators - Asymptotic Relative Efficiency

Efficiency refers to the variance of the sampling distribution for the estimator. High efficiency estimators have small variance in the sampling distribution for the estimator. Efficiency will affect the power of a test procedure in that less variance in the sampling distribution for the estimator being tested, will lead to higher power for the statistical test. Here are two ways of viewing efficiency. Finite sample efficiency refers to the variance of the sampling distribution for the estimator as it is applied in small sample settings. Asymptotic efficiency refers to the way an estimator performs as the sample size gets larger. It is a common practice to compare estimators to one another using Asymptotic Relative Efficiency (ARE). For a fixed underlying distribution, we define the Relative Efficiency (RE) of one estimator to another estimator as the ratio of the two variances of the estimators, and ARE is the asymptotic value of RE as the sample size goes to infinity. For

example, to compare the efficiencies of the mean and median, one would sample from a fixed underlying distribution and fixed sample size (i.e. normal distribution), then divide the variance of the median into the variance of the mean. As the sample size increases, this ratio will converge to the ARE of the two estimators. In this way, estimators can be compared with respect to the different types of non-normality that is found in data analysis settings.

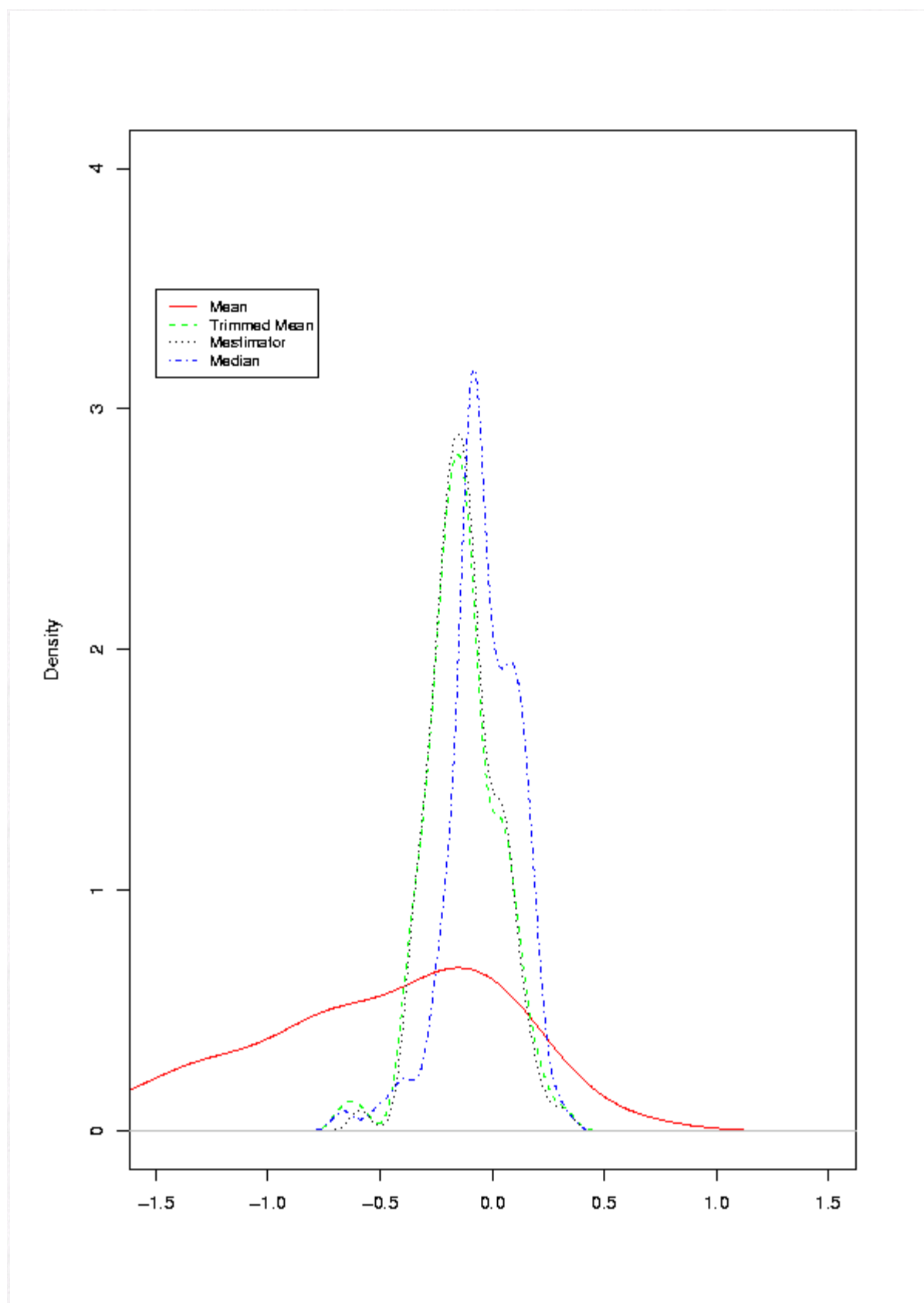
Robustness Properties: High Breakdown and Resistance

High breakdown is the largest percentage of data points that can be arbitrarily changed and not unduly influence the estimator (e.g. location parameter). For example, the median has 50% breakdown. That is, for 100 rank ordered data points, the first 49 points can be changed arbitrarily such that the values are still less than the median, and the median will not change. The mean is not considered a robust estimator because changing one observation arbitrarily can greatly influence the mean. This implies that the mean has a breakdown of $(1/n) \times 100$. As n increases, the breakdown of the mean linearly decreases in an unbounded fashion. In comparison, the median has a much higher breakdown than the mean, and as such, is considered a more robust estimate of location.

A Comparison of Four Robust Estimators of Location

The median has a breakdown of 50%. The trimmed mean has a breakdown that corresponds to the degree of trimming that is utilized. For example, a 20% trimmed mean has a breakdown of 20%. The mean has a breakdown of $(1/n) \times 100$, where n is the sample size. For Huber type estimators, the breakdown will depend on the trimming constant k . In the figure below, the sampling distributions of the sample mean, sample trimmed mean, sample M-estimator, and the sample median are plotted. Sampling occurred from the CN distribution where there is a 90% probability of sampling from $N(0, 1)$ and 10% probability of sampling from $N(0, 10)$ with a mean of zero. We see that the sample median, sample m-estimate and sample trimmed mean are all considerably closer to the population mean of zero. Additionally, there is less variability in these estimates, than the sample mean.





References

Hampel, F.R. (1973). Robust estimation: A condensed partial survey. *Z.*

Wahrscheinlichkeitstheorie and Verw. Gebiete, 27, 87-172.

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

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

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

Leger, C. & Romano, J.P. (1990). *Bootstrap adaptive estimation: the trimmed-mean example*. *The Canadian Journal of Statistics*, 18(4), 297-314.

Micceri, T. (1989). *The unicorn, the normal curve, and other improbable creatures*. *Psychological Bulletin*, 105(1), 156-166.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley, New York.

Sawilowsky, S. S. & Blair, R. C. (1992). *A more realistic look at the robustness of Type II error properties of the t-test to departures from population normality*. *Psychological Bulletin*, 111, 352-360.

Stigler, S.M. (1977). *Do robust estimators work with real data?* *Annals of Statistics*, 5, 1055-1098.

Tukey, J. (1960). A Survey of Sampling From Contaminated Distributions. In: *Contributions to Probability and Statistics*. Olkin, I. (Ed). Stanford University Press: Stanford, California.

Wilcox, Rand R. (2001). *Fundamentals of Modern Statistical Methods*. Springer-Verlag, New York.

Wilcox, Rand R. (1998). *How Many Discoveries have Been Lost by ignoring Modern Statistical Methods?* *American Psychologist*, Vol 53, No. 3, 300-314.