Paper 128-2010

SAS JMP® simulation-based empirical determination of robust scale estimator and comparison of outlier discrimination performance

Tarun Chandra and Gennadiy Gorelik, EmpiriQA LLC, Long Grove, IL

Abstract

A robust scale estimator based on an empirically-derived correction factor (ECF) is proposed when the sample median is more efficient than the mean. ECF values for simulated samples of varying size (n=2 to 80) drawn from the standard normal distribution are obtained in SAS JMP® using the distribution of sample standard deviation (s) to sample median absolute deviation (MAD) ratios. ECF central tendency estimates based on median of s/MAD ratio distributions are compared to corresponding MAD-based theoretical correction factor (TCF) and σ-based asymptotic correction factor (ACF=1.4826) estimates. Results indicate TCF > ECF > ACF and ECF (and TCF) asymptotically approach ACF as sample size increases. %Bias in estimating σ using ECF, TCF and ACF were compared using mean, median and root mean square (RMS) of normalized (relative to "σ") s-distribution. Outlier discrimination performance of non-robust scale estimator, "s", and robust estimators, "MAD_n", based on ECF and TCF was also compared using simulations for single outlier-contaminated samples of varying size, n, drawn from a standard normal distribution. Different outlier scenarios were considered by combining (n-1) samples from N[μ = 0, σ =1] with one outlier drawn from a statistically distinct normal distribution N[μ >> 0, σ =1]. For sample sizes n≤ 40, robust scale estimator MAD_n = ECF*MAD yielded better outlier discrimination when compared to MAD_n =TCF*MAD with "s" being the worst. In practical applications involving small sample sizes and occasional outlier contamination with no assignable cause, an ECF-based scale estimator could yield robust and reliable estimates of scale.

Introduction

In manufacturing and quality control applications it is common practice to take a small representative sample from a batch of production material, measure a physical, chemical or biological characteristic of the sample, express the measurement in terms of classical location (e.g. mean) and/or a scale (e.g. SD) estimators and compare these outputs to established specifications. The presence of outliers in the measured values can play havoc with these classical estimators especially for small sample sizes. Robust statistics provide an alternative approach to classical statistical estimators of location and scale. These robust estimators are more resistant to the statistical influences of outliers in real data. Like the Mean, the Median is an unbiased measure of the central tendency of a distribution. For a symmetric distribution, it is therefore equal to the mean. Just like the SD, the Median Absolute Deviation (MAD_n) is an unbiased measure of scale. This paper attempts at evaluating the advantages and limitations of applying robust statistical estimators like the median and MAD_n to measurements involving small sample sizes (n ≤ 5).

The robust equivalent of the mean statistic is the median [4]. A median is defined as the numeric value separating the upper half of a sample, a population, or a probability distribution, from the lower half

$$Median=med(X)$$

Similarly, the robust equivalent of the standard deviation is the median absolute deviation (MAD). MAD is based on the deviation of the individual data points from the Median, given by

$$MAD = med|X-med(X)|$$

To make the MAD estimator consistent with the standard deviation, σ, a Correction Factor b_n is used

$$MAD_n = b_n\ med|X-med(X)| = b_n\ MAD$$

The sample median and the MAD are simple and easy to compute. Their robustness makes them very useful for screening data for outliers [2], by computing the extreme value:

$$extreme = med|Xi-med(X)| / MADn$$

for each Xi and flagging those Xi for which this statistic exceeds a certain cutoff (for example, error rate for Six Sigma at 4.5 sigma level).

The breakdown point [6, 9] of an estimator is the proportion of incorrect observations an estimator can handle before giving an arbitrarily large result [8]. The median estimator has the finite sample breakdown point of (n-1)/2n. This means that for four (4) replicates the sample median estimator can handle one outlier. The mean estimator has the finite sample breakdown point of 1/n, what means that even one outlier ruins the sample mean.

Relative Efficiency of Median

The relative efficiency [3, 5, 7] of the median, measured as the ratio of the variance of the mean to the variance of the median, depends on the underlying distribution and the sample size. Through simulations performed in SAS JMP the effect of different distribution shapes and sample sizes on median efficiency was evaluated.

Median efficiency of t-distribution

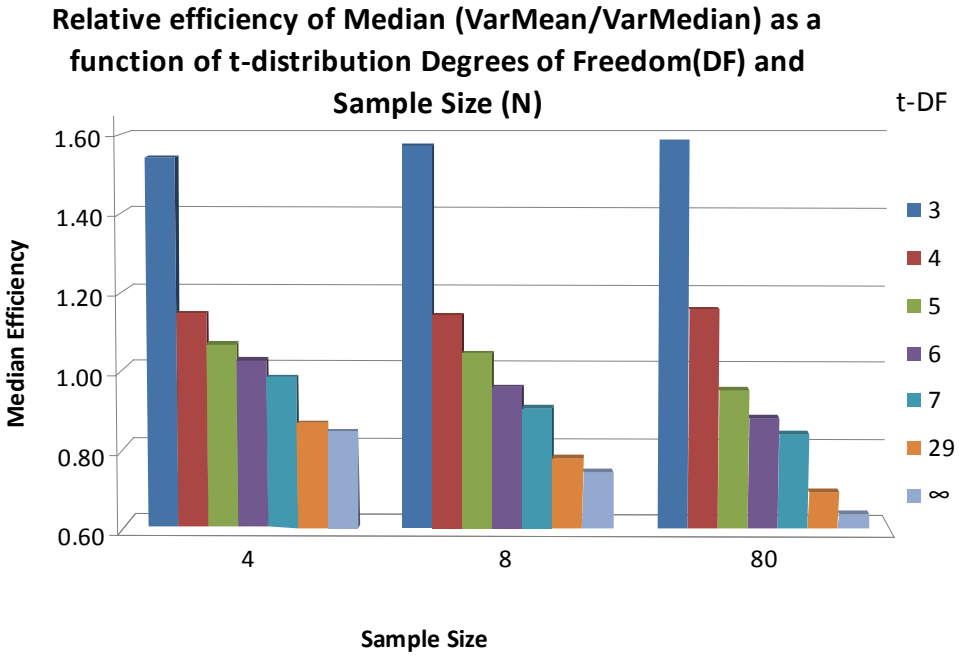
To define the operating space where Median is more efficient than Mean (efficiency >=1) for symmetrical t-distribution

1. Simulation of 80 Replicates from Student's t-distribution with Degrees of Freedom DF=3, 4, 5, 6, 7, 29, and ∞ (Normal Distribution) using JMP formula
   X=t Quantile(Random Uniform(), :DF) + mean (mean=0) was done.
1. Mean and Median of each sample (size 2, 3, 4, 5, 6, 7, 8, 20, 40, 80) were calculated.
2. 10,000 runs were simulated (10,000 rows in JMP data file).
3. Variance of mean and median (across the 10,000 runs) and the relative efficiency of the median were calculated.
4. To calculate simulation error of the median efficiency, steps 1 through 4 were repeated 10 times and standard deviation of the 10 efficiency values was calculated.

As results of simulation for Student's t distribution-based samples, presented in the table and figure below, the efficiency of the Median depends on the Degrees of Freedom (DF) and Sample Size (N).

Table 1: Relative efficiency of Median as a function of t-distribution Degrees of Freedom (DF) and Sample Size (N)

DF	N		
	4 reps	8 reps	80 reps
3	1.54	1.57	1.59
4	1.15	1.14	1.16
5	1.07	1.05	0.95
6	1.03	0.96	0.88
7	0.99	0.91	0.84
29	0.87	0.78	0.69
∞	0.85	0.74	0.64
Error(DF=3)	0.08	0.07	0.03

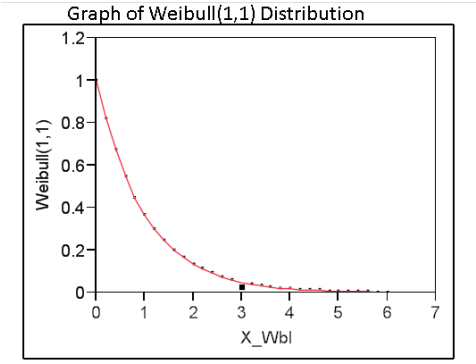


The median efficiency tends to increase with decreasing DF and N. So, the median is more efficient than the mean (efficiency>=1) for all samples from the t-distributions with DF = 3 and 4. At DF of 5 and 6, employing the Median instead of the Mean is reasonable for small sample sizes N<= 8 where median is at least as efficient as the Mean. In other cases (DF>6) including normal distribution (DF =∞) the Median is less efficient than the Mean, independent of Sample Size.

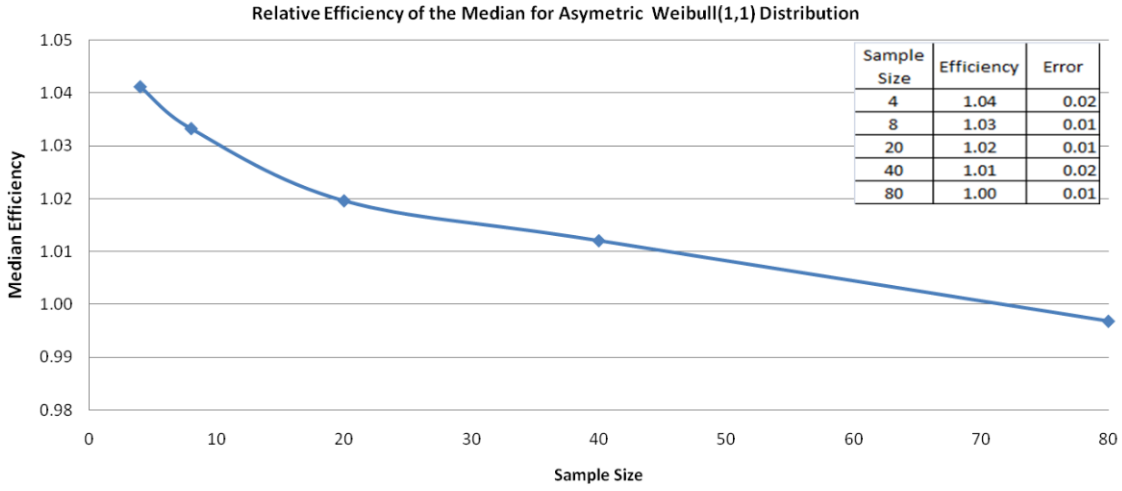
Efficiency of Weibull distribution

To estimate the impact of distribution shape asymmetry on the median efficiency, the Weibull distribution was used for simulations. Simulations were done similar to those for the symmetrical t-distribution using JMP formula:

X=Random Weibull(1, 1)



Based on the JMP simulations, the Weibull(1,1) sample Median is at least as efficient as the Mean, independent of Sample Size (efficiency>~1).



Sample Size	Efficiency	Error
4	1.04	0.02
8	1.03	0.01
20	1.02	0.01
40	1.01	0.02
80	1.00	0.01

Efficiency of Bimodal Distribution

Four scenarios of bimodal distributions were considered by evaluating a combination of samples drawn from two distinct t-distributions:

- Three replicates from standard t-distribution with mean=0 in combination with one replicate from t-distribution with a shifted mean
  - Sample Size N=4
  - Mean shift for second t-distribution: 0.0, 0.2 and 0.4
  - Efficiency of Median vs. DF and magnitude of mean shift was evaluated
- Two replicates from standard t-distribution with mean=0 in combination with two replicates from t-distribution with shifted mean
  - Sample Size N=4, 8 and 80
  - Mean shift for second t-distribution: 0.4
  - Efficiency of Median vs. DF was evaluated
- All four replicates from t-distribution with shifted mean
  - Sample Size N=4, 8 and 80
  - Efficiency of Median vs. shift was evaluated
- Four replicates from Weibull(1,1) distribution in combination with replicates from t-distribution
  - Sample Size N=4, 8 and 80

–Mean shift for t-distribution: 0.4

–To check effect of multimodality, the replicates picked from standard t-distribution had DF=7, because value of Efficiency for this DF is close to one

–Efficiency of Median vs. Sample Size was evaluated

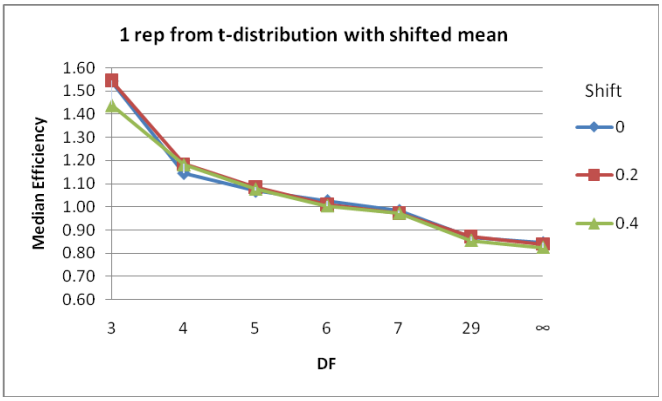
Results of Bimodal distribution Simulation

For Bimodal t-distribution with one and two replicates from distribution with shifted mean, range of DF and Sample Sizes with efficient Median (efficiency>1) doesn't change significantly in comparison to single t-distribution

1 rep from t-distribution with shifted mean

DF ↓	Shift		
	0	0.2	0.4
3	1.54	1.54	1.44
4	1.15	1.18	1.18
5	1.07	1.09	1.08
6	1.03	1.01	1.00
7	0.99	0.97	0.97
29	0.87	0.87	0.86
∞	0.85	0.84	0.82
Error(DF=3)	0.08	0.05	0.07

.

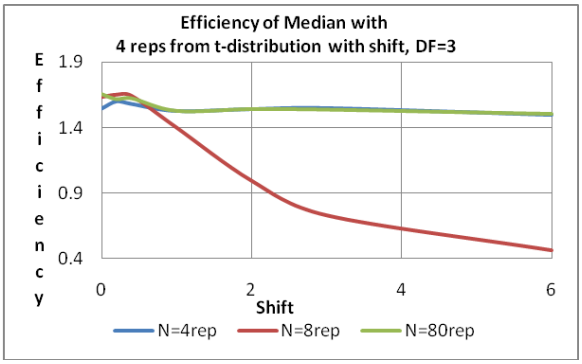


Sample Size	4	8	80	4	8	80
DF ↓	t-distribution			Bimodal t-distr with shift=0.4*		
3	1.54	1.57	1.59	1.57	1.66	1.66
4	1.15	1.14	1.16	1.20	1.17	1.14
5	1.07	1.05	0.95	1.07	1.04	0.98
6	1.03	0.96	0.88	0.98	0.93	0.85
7	0.99	0.91	0.84	0.97	0.91	0.83
29	0.87	0.78	0.69	0.86	0.78	0.68
∞	0.85	0.74	0.64	0.83	0.73	0.64
Error(DF=3)	0.08	0.07	0.03	0.09	0.08	0.09

*2 reps from t-distribution with mean 0.4

For Bimodal t-distribution, with constant number of shifted replicates (4) and DF=3, efficiency of Median is not different from Student's t-distribution for large sample size (N=80 replicates}, when the effect of shifted replicates on mean and median is not significant. When number of standard and shifted replicates is comparable (N=8 replicates) efficiency of median decreases as magnitude of mean shift increases.

Mean Shift	Sample Size (N)		
	N=4rep	N=8rep	N=80rep
0	1.54	1.64	1.66
0.2	1.60	1.65	1.62
0.4	1.58	1.64	1.62
1	1.52	1.40	1.53
2	1.54	0.99	1.54
3	1.55	0.73	1.54
6	1.49	0.46	1.50



N=4rep (second column) corresponds to unimodal t-distribution

For Bimodal t-distribution with four replicates from the Weibull(1,1) distribution, efficiency of median when N<=40 is higher in comparison to single t-distribution with DF=7 when efficiency is close to one only when Sample Size N<=4.

Sample Size	Efficiency	Error
8	1.43	0.03
20	1.12	0.01
40	1.01	0.01
80	0.91	0.01

Correction Factor for Sigma (σ) estimate

MAD = med|X-med(X)|

To make the MAD estimator consistent with the standard deviation, σ, the Correction Factor b_n is used.

$$MAD_n = b_n \text{ med}|X-\text{med}(X)| = b_n MAD$$

The factor b_n (Theoretical Correction Factor, TCF) depends on sample size n and the underlying distribution. In the case of the Normal distribution, the asymptotic value of b_n is 1.4826 [2]. In other words, the expectation of 1.4826 times the MAD for large samples of normally distributed X_i is approximately equal to the population standard deviation. Other distributions behave differently: for example for large samples from a uniform continuous distribution, this factor is about 1.1547 (the square root of 4/3). This necessitates the empirical estimation of the correction factor (Empirical Correction Factor, ECF) specific for each type of sample distribution with unknown parameters using actual data.

To compare the Empirical Correction Factor (ECF) with Theoretical Correction Factor (TCF) for the normal Distribution, the following JMP Simulation procedure was performed: (correction factor estimate does not depend on parameters of normal distribution)

- Simulation of 80 Replicates from normal distribution using JMP formula
  *X=Random Normal() * :sigma + :mean (mean=0)*
- Mean, Median, SD (STD in JMP) and MAD of each sample (size of 2, 3, 4, 5, 6, 7, 8, 20, 40, 80) were calculated.
- 10000 runs were simulated (10000 rows).
- Proportion of STD/MAD for each run was calculated
- Empirical Correction factor ECF was calculated as the Median of simulated (STD/RSTD) distribution
  ECF=Col Quantile(:Name("std_N/MAD_N"), 0.5), N-Sample Size
- To calculate Simulation Error of the Correction factor, steps above were repeated 10 times and standard deviation of 10 Correction factor values was calculated.

Effect of parameters of simulated distribution (μ, σ) on Empirical Correction factor value (ECF)
ECF was calculated using Median of simulated (STD/RSTD) distribution

Effect of Population Mean and SD variability

Sample size	μ=0, σ=1	μ=0, σ=10	μ=10, σ=10	Theoretical TCF	Asymptotic ACF	% difference= (TCF-ECF)/TCF
	Empirical Correction factor (ECF)					
3	1.94	1.94	1.93	2.22	1.48	12.7
4	1.78	1.77	1.78	2.02	1.48	12.1
5	1.68	1.69	1.67	1.79	1.48	6.0
6	1.65	1.65	1.66	1.78	1.48	7.2
8	1.59	1.59	1.59	1.67	1.48	5.0
20	1.52	1.51	1.52	1.54	1.48	1.6
40	1.50	1.50	1.50	1.51	1.48	0.8
80	1.50	1.50	1.50	1.50	1.48	-0.3

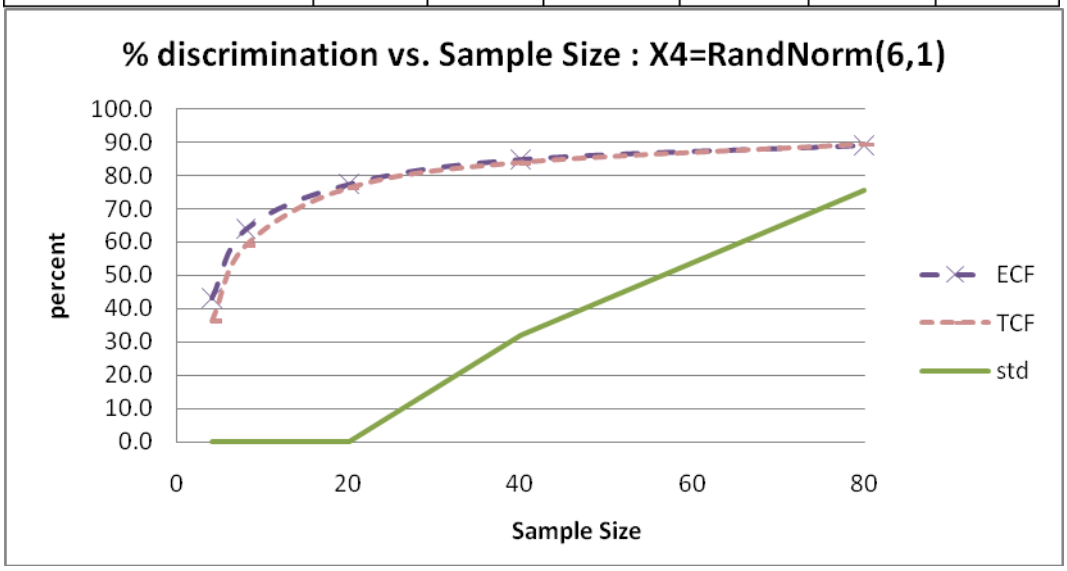
Error of Simulation for ECF ~ 1%

Outlier discrimination: Description of Simulation

- All Simulations were done using JMP
- 80 replicates were simulated
  - 79 Replicates from Normal Distribution using standard normal distribution
    *X=Random Normal() * :sigma + :mean (mean=0, sigma=1)*
  - one rep (outlier) was simulated from second normal distribution with mean of 6, 10 or 20 and sigma=1
- 10000 runs with 80 replicates were simulated (10000 rows).
- Mean and Median and STD, MAD= med|X-med(X)| and STD/MAD of each sample were calculated.
- Empirical Correction factor ECF was calculated as Median of simulated (STD/MAD) distribution
  - *ECF=Col Quantile(:Name("std_N/MAD_N"), 0.5), N-Sample Size*
- Robust extreme values were calculated using
  - *extreme = med|X-med(X)| / MAD_n*
    with MAD_n = ECF*MAD and MADn = TCF*MAD
  - for comparison, extreme values using mean and standard deviation, instead of median and MAD were also calculated
    extreme = |X-mean(X)| / STD
- If Extreme value is greater than Cut Off of 4.5 then outlier is defined (Success);
- If Extreme value is lower than 4.5 then outlier is not defined (Failure);
- % discrimination of outliers (% Success) vs. sample size for standard and robust methods were calculated and compared.

% Discrimination of Outliers (Extremes)
All replicates except for one are from RandNorm(0, 1), the outlier replicate is from RandNorm(shift, 1)

extreme value was calculated using →		%failure			% discrimination		
shift	Sample size	std	ECF	TCF	std	ECF	TCF
6	4	100.00	56.85	63.89	0.00	43.15	36.11
	8	100.00	36.09	41.00	0.00	63.91	59.00
	20	100.00	22.63	23.85	0.00	77.37	76.15
	40	67.82	15.18	16.27	32.18	84.82	83.73
	80	24.31	10.92	10.79	75.69	89.08	89.21
10	4	100.00	21.71	29.90	0.00	78.29	70.10
	8	100.00	3.53	4.86	0.00	96.47	95.14
	20	100.00	0.17	0.20	0.00	99.83	99.80
	40	0.44	0.00	0.00	99.56	100.00	100.00
	80	0.00	0.00	0.00	100.00	100.00	100.00
20	4	100.00	0.30	0.94	0.00	99.70	99.06
	8	100.00	0.00	0.00	0.00	100.00	100.00
	20	99.99	0.00	0.00	0.01	100.00	100.00
	40	0.00	0.00	0.00	100.00	100.00	100.00
	80	0.00	0.00	0.00	100.00	100.00	100.00



%discrimination increases with sample size and with magnitude of outlier (value of shift). The %discrimination is always higher for robust methods (ECF and TCF) compared to the classical SD estimator with ECF demonstrating equivalent or better (especially for small sample sizes) discrimination than TCF.

Conclusion

- The data space where the median is more efficient than the mean was determined. Using JMP simulations, the Empirical Correction Factor (ECF) was estimated for different types of distributions (normal, asymmetric and bimodal). The ECF was applied to the estimation of sigma and for outlier discrimination. Simulation results indicate ECF performance as equivalent to or better than that of the theoretical correction factor (TCF).

- Application of ECF to calculation of robust scale estimator and percent outlier discrimination for normal samples, gives results more sensitive to outliers than with the theoretical correction factor TCF and the classical SD scale estimator.

References

1. Wackerly D., Mendenhall W., Scheaffer R., Mathematical Statistics with Application, p.370, Duxbury Press,
2. Rousseeuw, P., Croux C., “Alternatives to the Median Absolute Deviation”, Journal of the American Statistical Association, 1993, Vol.88, No. 424
3. The distribution efficiency of the median. Statistical methods, by George Waddel Snedecor, William Gemmell Cochran, p.136.
4. Statistical Median, from Wolfram Math World (http://mathworld.wolfram.com/StatisticalMedian.html)
5. Example of efficiency for mean vs. median, the blog of John D. Cook - research statistician at M. D. Anderson Cancer Center and associate faculty for the UT Graduate School of Biomedical Sciences, (http://www.johndcook.com/blog/2009/03/06/student-t-distribution-mean-median).
6. Stat 5601 (Charles J. Geyer) Efficiency and Breakdown Point, Charles J. Geyer - Associate Professor School of Statistics University of Minnesota (http://www.stat.umn.edu/geyer/old02/5601/examp/eff.html).
7. http://en.wikipedia.org/wiki/Efficiency_(statistics)
8. http://en.wikipedia.org/wiki/Robust_statistics
9. Breakdown Point Theory Notes. Charles J. Geyer. February 2, 2006. 1 Introduction. These are class notes for Stat 5601.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Name: Tarun Chandra
Enterprise: EmpiriQA LLC
Address: 1580 RFD
City, State ZIP: Long Grove, IL 60047
Phone: (262)412-5407
Fax:
E-mail: tchandra@empiriqa.com
Web: www.empiriqa.com