Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Symmetric quantiles and their applications

Yuang-Chin Chiang [a] , Lin-An Chen [b] & Hsien-Chueh Peter Yang [c]

[a] Institute of Statistics , National Tsing Hua University , Hsinchu, Taiwan

[b] Institute of Statistics , National Chiao Tung University , Hsinchu, Taiwan

[c] Department of Risk Management and Insurance , National Kaohsiung First University of Science and Technology , Taiwan

Published online: 22 Jan 2007.

Journal of Applied Statistics
Vol. 33, No. 8, 807–817, September 2006

Symmetric Quantiles and their Applications

YUANG-CHIN CHIANG,* LIN-AN CHEN** & HSIEN-CHUEH PETER YANG†

**Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan, **Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan, †Department of Risk Management and Insurance, National Kaohsiung First University of Science and Technology, Taiwan*

ABSTRACT *To develop estimators with stronger efficiencies than the trimmed means which use the empirical quantile, Kim (1992) and Chen & Chiang (1996), implicitly or explicitly used the symmetric quantile, and thus introduced new trimmed means for location and linear regression models, respectively. This study further investigates the properties of the symmetric quantile and extends its application in several aspects. (a) The symmetric quantile is more efficient than the empirical quantiles in asymptotic variances when quantile percentage α is either small or large. This reveals that for any proposal involving the α th quantile of small or large α s, the symmetric quantile is the right choice; (b) a trimmed mean based on it has asymptotic variance achieving a Cramer-Rao lower bound in one heavy tail distribution; (c) an improvement of the quantiles-based control chart by Grimshaw & Alt (1997) is discussed; (d) Monte Carlo simulations of two new scale estimators based on symmetric quantiles also support this new quantile.*

KEY WORDS: Regression quantile, scale estimator, trimmed mean

Introduction

The empirical quantile has long been very popular in constructing location and scale estimators and this quantile has been successfully generalized to the regression case by Koenker & Bassett (1978). In order to improve the efficiency of a location estimator, and the trimmed mean, Kim (1992) developed the metrically trimmed mean for a location model which, through comparison of asymptotic variances, was shown to be more efficient than the ordinary trimmed mean. Later, Chen & Chiang (1996) defined the symmetric quantile and used it to propose the symmetric trimmed mean as an extension of Kim's trimmed mean to the linear regression model. They observed that this symmetric trimmed mean of small trimming percentages can have asymptotic variances very close to the Crammer-Rao lower bounds when regression errors obey heavy tail distributions.

Correspondence Address: Lin-An Chen, Institute of Statistics, national Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan. Email: lachen@stat.nctu.edu.tw

0266-4763 Print/1360-0532 Online/06/080807–11 © 2006 Taylor & Francis
DOI: 10.1080/02664760600743464

Two questions arise from the fact that efficiency is gained by the symmetric trimmed mean. (1) For a trimmed mean, the role of a quantile is to classify the data into a set of good observations and a set of suspected outliers. Clearly, the efficiency of a quantile in the estimation of a population quantile must affect the efficiency of the resulting data classification. Therefore, could efficiency be gained by using the symmetric quantile rather than the empirical quantile as the estimator of the population quantile? (2) Can the efficiency of a symmetric trimmed mean carry over to other quantile-based proposals? We deal with the first question in three ways. (a) We compare the asymptotic variances of the symmetric quantile and the empirical quantile to discover their efficiencies in the role of estimating the population quantile. (b) We analyze a real data set by computing the confidence regions of observations through the symmetric quantile and empirical quantile separately to determine their ability for data classification. We will find that the former is better to catch the main trend shown by the data. (c) For studying the efficiency of the symmetric trimmed mean in advance, we show that its asymptotic variance may achieve the Cramer-Rao lower bound when the errors follow an extreme contaminated distribution. This result has not been shown to hold by other robust estimators. To answer the second question above, we propose a quantile control chart, using the symmetric quantiles, that gains efficiencies and is different from the one of Grimshaw & Alt (1997) and two symmetric-quantile-based scale estimators.

This work displays the symmetric quantile in a more general form and, in the next section, studies its large sample distribution. A comparison of asymptotic variances for the empirical quantile and the symmetric quantile is presented in the third section. The fourth section discusses the benefits of constructing a quantile control chart as in Grimshaw & Alt (1997) with the symmetric quantiles. In the fifth section, a data analysis is displayed and, under an extreme contaminated normal distribution, an optimal result performed by the trimmed mean based on symmetric quantiles is introduced. The sixth section introduces two new scale estimators.

Symmetric Quantile Class

Unlike the way in which the empirical quantile is constructed based on the cumulative distribution function, the so-called symmetric quantile of Chen & Chiang (1996) is formulated based on a folded distribution function. However, these two quantiles are identical when a symmetric error distribution is assumed, which makes them comparable. Here we consider this quantile concept of Chen & Chiang (1996) in a more general setting.

Definition 1. For random variable y with cumulative distribution function F_y, consider the folded cumulative function about constant c, known or unknown, as

$$F_s(a) = P(|y - c| \leq a), \quad a \geq 0$$

The symmetric λ th quantile of F_y about c is the pair $\{c_s^-(\lambda), c_s^+(\lambda)\}$, where $c_s^-(\lambda) = c - F_s^{-1}(\lambda)$ and $c_s^+(\lambda) = c + F_s^{-1}(\lambda)$ and where the function $F_s^{-1}(\lambda)$ is the λ th quantile of cumulative function F_s.

If F_y is continuous, the symmetric quantile satisfies $\lambda = P(c_s^-(\lambda) \leq y \leq c_s^+(\lambda))$. If we further assume that F_y is symmetric at μ, it can be seen that $\mu_s^-(1 - 2\alpha) = F_y^{-1}(\alpha)$ and $\mu_s^+(1 - 2\alpha) = F_y^{-1}(1 - \alpha)$ for $0 < \alpha < 0.5$. According to Ferguson (1967), function $F_s^{-1}(\lambda)$ may be formulated as a solution of a minimization problem.

Theorem 1. If $0 < \lambda < 1$ and if c is any constant, we have

$$F_s^{-1}(\lambda) = \arg\min_{a>0} E_{F_y}(|y-c|-a)(\lambda - I(|y-c| \le a))$$

which provides an identification of the symmetric quantile.

By letting $\varepsilon_c = y - c$, then random variable y obeys the location model $y = c + \varepsilon_c$. Suppose that we now have a random sample $y_1, \ldots, y_n$ from this location model. Let $\hat{c}$ be a statistic satisfying $n^{1/2}(\hat{c} - c) = O_p(1)$. We define next the sample type symmetric quantile for the location model.

Definition 2. The sample symmetric quantile for the location model is defined as the pair $\{\hat{c}_s^-(\lambda), \hat{c}_s^+(\lambda)\}$ with $\hat{c}_s^-(\lambda) = \hat{c} - \hat{F}_s^{-1}(\lambda)$ and $\hat{c}_s^+(\lambda) = \hat{c} + \hat{F}_s^{-1}(\lambda)$ where $\hat{F}_s^{-1}(\lambda)$ is for estimating $F_s^{-1}(\lambda)$ as

$$\hat{F}_s^{-1}(\lambda) = \arg\min_{a>0} \sum_{i=1}^{n} (|y_i - \hat{c}| - a)(\lambda - I(|y_i - \hat{c}| \le a))$$

We now introduce the concept of the symmetric quantile for the linear regression model $y = x'\beta + \varepsilon$ where x is a constant vector with value 1 in its first element. We assume that regression error ε has a cumulative distribution function F and we denote the vector $\beta = (\beta_0, \ldots, \beta_p)'$. Koenker & Bassett (1978) defined the α th regression quantile as $\beta(\alpha) = \beta + (F^{-1}(\alpha), 0, \ldots, 0)'$ where they defined a sample type α th regression quantile as

$$\widehat{\beta}(\alpha) = \arg\min_{b \in R^{p+1}} \sum_{i=1}^{n} (y_i - x_i'b)(\alpha - I(y_i \le x_i'b))$$

We consider the conditional regression symmetric quantile based on a distribution at $x'\beta_c$ with vector $\beta_c = (c, \beta_1, \ldots, \beta_p)'$.

Definition 3. The symmetric type conditional quantile centered at $x'\beta_c$ is defined as the pair $\{x'\beta_c - F_s^{-1}(\lambda), x'\beta_c + F_s^{-1}(\lambda)\}$ where

$$F_s^{-1}(\lambda) = \arg\inf_{a>0} E_F(|y - x'\beta_c| - a)(\lambda - I(|y - x'\beta_c| \le a))$$

We refer to the pair $(\beta_s^-(\lambda), \beta_s^+(\lambda))$ with $\beta_s^+(\lambda) = \beta_c + \begin{pmatrix} F_s^{-1}(\lambda) \\ 0_p \end{pmatrix}$ and $\beta_s^-(\lambda) = \beta_c - \begin{pmatrix} F_s^{-1}(\lambda) \\ 0_p \end{pmatrix}$ as the symmetric regression quantiles.

It can be seen that the symmetric conditional quantile pair centered at $x'\beta_c$ is $\{x'\beta_s^-(\lambda), x'\beta_s^+(\lambda)\}$. When F is continuous, the symmetric conditional quantile satisfies $\lambda = P(x'\beta_s^-(\lambda) \le y \le x'\beta_s^+(\lambda))$. Moreover, if we further assume that F is symmetric at 0 and we let $c = \beta_0$ then, for $0 < \alpha < .5$, we have $F_s^{-1}(1-2\alpha) = F^{-1}(1-\alpha)$, which also implies that $\beta_s^-(1-2\alpha) = \beta(\alpha)$ and $\beta_s^+(1-2\alpha) = \beta(1-\alpha)$. This makes estimators of these two regression quantiles comparable in either their large sample or small sample properties.

Again, suppose that we have drawn a set of regression observations $\begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ x_n \end{pmatrix}$ from the linear regression model and an estimator $\hat{\beta}_c$, computed based on these observations, satisfying $n^{1/2}(\hat{\beta}_c - \beta_c) = O_p(1)$. Then to estimate the symmetric regression quantile we need only to estimate quantity $F_s^{-1}(\lambda)$.

Definition 4. The sample type symmetric λ th regression quantiles are $\hat{\beta}_s^{+}(\lambda) = \hat{\beta}_c + \begin{pmatrix} \hat{F}_s^{-1}(\lambda) \\ 0_p \end{pmatrix}$ and $\hat{\beta}_s^{-}(\lambda) = \hat{\beta}_c + \begin{pmatrix} -\hat{F}_s^{-1}(\lambda) \\ 0_p \end{pmatrix}$. Here $\hat{F}_s^{-1}(\lambda)$ is defined by

$$\hat{F}_s^{-1}(\lambda) = \arg\min_{a>0} \sum_{i=1}^{n} (|y_i - x_i'\hat{\beta}_c| - a)(\lambda - I(|y_i - x_i'\hat{\beta}_c| \le a)) \tag{1}$$

for estimating $F_s^{-1}(\lambda)$.

The role of the symmetric regression quantile in classifying observations may be interpreted in the following theorem.

Theorem 2. Let U and Z^{+} denote the numbers, respectively, of positive and zero elements in the set $\{r_i = y_i - x_i'\hat{\beta}_s^{+}(\lambda)\}$; L and Z^{-} denote the numbers, respectively, of negative and zero elements in the set $\{r_i = y_i - x_i'\hat{\beta}_s^{-}(\lambda)\}$; and B denotes the number of elements in the set $\{y_i : x_i'\hat{\beta}_s^{-}(\lambda) < y_i < x_i'\hat{\beta}_s^{+}(\lambda)\}$. Then the λ th sample symmetric regression quantiles for the linear regression model satisfy

(a) $n\lambda - (Z^{+} + Z^{-}) \le B \le n\lambda$ and
(b) $U + L \le n(1 - \lambda) \le U + L + Z^{+} + Z^{-}$

This theorem may be proved analogously to Theorem 3.4 of Koenker & Bassett (1978) and thus it is skipped here. This specifies the numbers of observations falling in and out of the λ th quantile, denoted by $\varepsilon_{ci} = y_i - x_i'\beta_c$ and $\hat{\beta}_c = (\hat{\beta}_{c0}, \ldots, \hat{\beta}_{cp})'$. The main result, which provides a representation of a symmetric regression quantile, is stated in the following theorem.

Theorem 3. For either γ denoted $-$ or $+$, the symmetric regression quantile has the following representation

$$n^{1/2}(\hat{\beta}_s^{\gamma}(\lambda) - \beta_s^{\gamma}(\lambda)) = \begin{pmatrix} \gamma(g^{+}(\lambda))^{-1} n^{-1/2} \sum_{i=1}^{n} [\lambda - I(|\varepsilon_{ci}| \le F_s^{-1}(\lambda))] \\ 0_p \end{pmatrix}$$
$$+ n^{1/2} \begin{pmatrix} (1 + \gamma(g^{+}(\lambda))^{-1} g^{-}(\lambda))(\hat{\beta}_{c0} - c) \\ \hat{\beta}_{c1} - \beta_1 \\ \cdots \\ \hat{\beta}_{cp} - \beta_p \end{pmatrix} + o_p(1)$$

where $g^{+}(\lambda) = f_c(F_s^{-1}(\lambda)) + f_c(-F_s^{-1}(\lambda))$ and $g^{-}(\lambda) = f_c(F_s^{-1}(\lambda)) - f_c(-F_s^{-1}(\lambda))$ and where f_c is the p.d.f. of ε_{ci}.

The proof of this theorem is analogous to Theorem 3.1 of Chen & Chiang (1996) and thus is also skipped.

Corollary 1. If F is symmetric at zero and we let $\beta_c = \beta$, then

$$n^{1/2}(\hat{\beta}_s^{\gamma}(1-2\alpha) - \beta_s^{\gamma}(1-2\alpha)) = n^{1/2}(\hat{\beta} - \beta) + \begin{pmatrix} \gamma(2f(F^{-1}(1-\alpha)))^{-1} n^{-1/2} \sum_{i=1}^{n} (1 - 2\alpha - I(|\varepsilon_i| \leq F^{-1}(1-\alpha))) \\ 0_p \end{pmatrix} + o_p(1).$$

The benefits of the symmetric quantile compared with the empirical quantile, through the point of asymptotic variances computed from the above representation, are described in next two sections.

Comparison of Asymptotic Variances for Location Model

For asymmetric error distribution, the symmetric regression quantile and the regression quantile of Koenker & Bassett (1978), actually represent, respectively, estimators of two different parameter vectors, which makes it difficult to compare their asymptotic variances. We here restrict the errors with symmetric distribution to compare asymptotic variances of these quantile estimators. For simplicity, we also consider only the location model where the asymptotic variances of the α th empirical quantile and symmetric quantile are $\alpha(1-\alpha)f^{-2}(F^{-1}(\alpha))$ and, from Corollary 1 $(1-\alpha)(2\alpha-1)f^{-2}(F^{-1}(1-\alpha)) + 0.25f^{-2}(0)$, respectively. With these formulas, we evaluate the efficiency of the symmetric quantile defined the following

$$\frac{\text{Asymptotic variance of empirical quantile}}{\text{Asymptotic variance of symmetric quantile}}$$

where the quantiles are to estimate $F_y^{-1}(\alpha)$ where the error variable is considered to have the contaminated normal distribution

$$(1-\delta)N(0,1) + \delta N(0,\sigma^2) \tag{2}$$

Table 1 lists the efficiency defined above for the cases where $\delta = 0.1$ and 0.2, $\sigma = 1,3,5,10$ and $\alpha = 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95$ and 0.98, where we note that results for α and $1-\alpha$ are exactly the same.

Table 1. Efficiency of symmetric quantile

α	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.98
$\delta = 0.1$									
$\sigma = 1$	0.87	0.84	0.84	0.84	0.87	0.92	1.01	1.21	1.47
3	0.91	0.89	0.90	0.92	.98	1.09	1.30	1.84	1.90
5	0.91	0.90	0.91	0.94	1.01	1.15	1.44	2.39	2.02
10	0.90	0.89	0.90	0.93	1.01	1.16	1.49	2.70	2.03
$\delta = 0.2$									
$\sigma = 3$	0.88	0.87	0.88	0.92	1.00	1.14	1.40	1.78	1.98
5	0.89	0.88	0.91	0.97	1.08	1.28	1.68	2.04	2.00
10	0.89	0.89	0.93	1.01	1.16	1.45	1.98	2.10	2.03

Based on Table 1, we have the following comments concerning the estimation of population quantile $F_y^{-1}(\alpha)$.

(a) It is relatively efficient to use the empirical quantile when the quantile percentage α is close to 0.5 in either direction. However, it is efficient to use the symmetric quantile when the α is either small or large.
(b) It implies that the symmetric quantile is more efficient than the empirical quantile in detecting outliers since outliers usually lie below the lower population quantile and above the upper population quantile. This confirms the results in Kim (1992) and Chen & Chiang (1996) that trimmed means based on symmetric type quantiles may be more efficient since this quantile is more capable in dividing observations into a good subclass and a bad subclass of observations.
(c) Any proposal using quantiles of small or large percentages α will be efficient if it is constructed by the symmetric quantiles.

What are the situations where practical statistical procedures involve quantiles of large or small percentages α? As we have seen, the trimmed means proposed by Kim (1992) and Chen & Chiang (1996) are just these procedures. One statistical technique that is very powerful in improving a product's quality is the so-called process capability index, which is defined as

$$\frac{UCL - LCL}{F_y^{-1}(0.99865) - F_y^{-1}(0.00135)}$$

where *USL* and *LSL* are values representing, respectively, the upper and lower specification limits. Higher values of the index indicate that the manufacturing process is more capable. Traditionally, $F_y^{-1}(0.00135)$ and $F_y^{-1}(0.99865)$ are estimated by their corresponding empirical quantiles. We now have an estimator of this index with smaller asymptotic variance than the usual one by using symmetric quantiles as estimators of $F_y^{-1}(0.00135)$ and $F_y^{-1}(0.99865)$. In addition to this application, in subsequent sections we introduce a control chart and two scale estimators that all involve symmetric quantiles.

Control Chart for Quantile Function Values

The traditional $\bar{X}$ and R charts in quality control are efficient in detecting changes in mean and variation when the ideal assumption of normal distribution is valid. However, their efficiencies can be remarkably reduced due to departures from normality and in the presence of outliers. Grimshaw & Alt (1997) proposed using a quantile control chart where the control limit of the chart is estimated by a confidence band for the quantile vector $(F_y^{-1}(\alpha_1), \ldots, F_y^{-1}(\alpha_p))'$, for some α_i, $i = 1, \ldots, p$. They showed that these charts are quite effective in detecting changes in the distributional shape that are undetected in $\bar{X}$ and R charts. Moreover, they also pointed out that for effective use of a quantile control chart we should select quantile percentages α_i, $i = 1, \ldots, p$ so that their corresponding differences $F_{yO}^{-1}(\alpha_i) - F_{yI}^{-1}(\alpha_i)$, with F_{yI} and F_{yO} respectively representing the distribution functions of in-control and likely out-of-control processes, are large.

In the following, Table 2 gives the differences of α th population quantiles of standard normal distribution $N(0,1)$ and the contaminated normal distribution of 0.8 $N(0,1) + 0.2$ $N(0,\sigma 2)$, for several values of α, where the contaminated one represents the out-of-control statistical process and the normal one is in control. Smaller and larger quantile

Table 2. Differences of population quantiles for standard normal and contaminated normal distributions of $\delta = 0.2$

α	0.55	0.60	0.70	0.75	0.80	0.85	0.90	0.95
$\sigma = 3$	0.024	0.041	0.089	0.119	0.163	0.182	0.348	0.729
5	0.024	0.051	0.114	0.154	0.163	0.312	0.528	1.749
10	0.029	0.061	0.129	0.184	0.258	0.392	0.768	5.134
20	0.034	0.061	0.139	0.199	0.283	0.442	0.998	11.854

differences between the in-control and out-of-control distributions means, respectively, insensitivity and sensitivity in detection of distributional change by the quantile control chart.

From this table and with the symmetric property of the distribution, it is clear that the largest difference values occurred only on the larger (close to 1) and smaller (close to 0) values of α. This observation indicates the following things that we may be concerned with in constructing a quantile control chart: (a) the quantile control chart should choose α_i, $i = 1, \ldots, p$, small or large values; (b) these quantiles, $F^{-1}(\alpha_i)$, of small and large values of α, should be estimated by symmetric quantiles for efficiency in estimation of control limits.

An Example of Quantile Confidence Region and Asymptotic Efficiency for a Trimmed Mean based on Symmetric Quantiles

In addition to being an estimator of its population version, one important function of the quantile function is to identify a region that covers a subset of observations with a predetermined proportion γ. One application of this coverage region includes constructing robust-type location and scale estimators based on observations in this region, for example, the trimmed mean, Winsorized mean and trimmed variance. Another application is that some estimators are constructed based on the width (area, etc) of the region, for example, the interquartile range and the process capability index presented in the third section of this paper.

Among the available coverage regions constructed by the existing quantile functions, how can we determine one to use? Criteria may be set by comparing the asymptotic variances or mean square errors of the estimators constructed by these coverage regions. We do this later, after comparing them through the sizes of the regions' areas or volumes.

First we consider an example of real data with outliers and asymmetric errors in order to compare the coverage regions constructed by, respectively, Koenker & Bassett's and symmetric regression quantiles. The example we now consider is a data set of international phone calls that appeared in the Belgian Statistical Survey, as presented in Rousseeuw & Leroy (1987). The plot of the phone calls (in tens of millions) in Rousseeuw & Leroy shows an upward trend over years. However, the tendency contains heavy contamination from year 64 to 69 (1964–1969). We let $A(KB)$ and $A(SQ)$ denote the two areas for the coverage regions: one covered by two Koenker & Bassett's regression quantiles $\hat{\beta}(\alpha)$ and $\hat{\beta}(1-\alpha)$ and the other covered by two symmetric quantiles $\hat{\beta}_s^-(1-2\alpha)$ and $\hat{\beta}_s^+(1-2\alpha)$. The efficiency of coverage region by quantile Q, $Q = KB$ or SQ, is defined as

$$E_Q = \frac{\min\{A(C_{SQ}), A(C_{KB})\}}{A(C_Q)}$$

In Table 3, we display the computed efficiencies E_{KB} and E_{SQ} for the international phone calls data.

In this example, the efficiencies of the coverage regions based on Koenker & Bassett's regression quantiles at confidence coefficients, $\gamma = 0.8$ and 0.9, are relatively larger than those based on symmetric quantiles; however, the discrepancies are not significant. On the other hand, the efficiencies of the coverage regions based on symmetric quantiles for confidence coefficients less than or equal to 0.7 are all significantly larger than those based on Koenker & Bassett's quantiles.

In estimating regression parameters, Chen & Chiang (1996) and Chen (1997) applied the symmetric quantile in constructing a trimmed mean. They also showed that this trimmed mean has asymptotic variance closer to the C-R bounds than the usual robust estimators when there is a heavy tail error distribution such as a contaminated normal. We will further prove here a theory for the attainment of C-R bound by this trimmed mean when the contaminated variance goes to infinity.

The λ th symmetric trimmed mean in Chen & Chiang (1996) is

$$\hat{\beta}_s(\lambda) = \left(\sum_{i=1}^{n} x_i x_i' \psi(y_i)\right)^{-1} \sum_{i=1}^{n} x_i y_i \psi(y_i)$$

where $\psi(y_i) = I(x_i'\hat{\beta}_s^-(\lambda) \leq y_i \leq x_i'\hat{\beta}_s^+(\lambda))$. This leads to the following, which is the main result in this section.

Theorem 4. Suppose that error variable ε has a contaminated normal distribution as

$$(1 - \delta)N(0, \sigma^2) + \delta\ N(0, \gamma\sigma^2) \tag{3}$$

where $0 < \delta < l$, $\gamma > 0$. Also, we assume that $\beta_c = \beta$, and $\hat{\beta}_c$ has a representation with bounded influence function. Then, as $\gamma \to \infty$, $\hat{\beta}_s(1 - \delta)$ has an asymptotic covariance matrix achieving the C-R lower bound as

$$n^{-1}(1 - \delta)^{-1}\sigma^2 Q^{-1} \tag{4}$$

Table 3. Efficiencies of coverage regions

1-2α	SQ	KB
0.90	0.714	1.000
0.80	0.983	1.000
0.70	1.000	0.063
0.60	1.000	0.071
0.50	1.000	0.056
0.40	1.000	0.045
0.30	1.000	0.052
0.20	1.000	0.214

Proof

Denote by $\tilde{g}_\gamma$ the contaminated distribution (3). The C-R bound for β is $n^{-1}(E_{\tilde{g}_\gamma}(\partial \ln \tilde{g}_\gamma(\varepsilon)/\partial\varepsilon)^2)^{-1}Q^{-1}$ which can be seen is convergent to the quantity of equation (4) as $\gamma \to \infty$. On the other hand, the contaminated normal distribution of equation (3) satisfies $\varepsilon\, f(\varepsilon) \to 0$ as $\varepsilon \to \infty$. Since, $\hat{\beta}_c$ has a bounded influence function, the asymptotic covariance matrix of $\hat{\beta}_s(1-\delta)$ is $n^{-1}Q^{-1}(1-\delta)^{-2}E_{g_\sigma}\varepsilon^2 I(|\varepsilon| \le F_s^{-1}(1-\delta/2))$ where g_σ is the distribution of $N(0, \sigma^2)$. The above result is induced from a representation of the symmetric trimmed mean in Chen & Chiang (1996) as

$$n^{1/2}(\hat{\beta}_s(1-2\alpha)-\beta) = \lambda^{-1}Q^{-1}[2F_s^{-1}(1-\alpha)f_c(F_s^{-1}(1-\alpha))Qn^{1/2}(\hat{\beta}_c-\beta) + n^{-1/2}\sum_{i=1}^{n} x_i\varepsilon_i I(|\varepsilon_i| \le F_s^{-1}(1-\alpha))] + o_p(1).$$

However, as $\gamma \to \infty$, then $F_s^{-1}(1-\delta/2) \to \infty$. Thus, the above variance is also the quantity of equation (4), which proves the theorem.

This result improves the theory in Chen & Chiang (1996) in two aspects. (1) A theory where a trimmed mean under a heavy tail distribution attains the C-R bound is developed. This property has not been seen for usual robust estimators. (2) The best trimming percentage is specified in this extreme distribution.

Two Scale Estimators Based on Symmetric Quantile

Developing robust-type scale estimators is also an interesting topic in the statistical literature. Welsh (1986) studied the Bahadur representations for median deviation and interquartile range. Welsh & Morrison (1990) introduced an interesting class of scale L-estimators with trimmed variance as a special case. Moreover, Staudte & Sheather (1990) provided a comprehensive review of scale estimators and Monte Carlo simulation. Here we introduce two easily computed alternative scale estimators based on a symmetric quantile.

One simple robust scale estimator of dispersion, popular in the literature, is the ‘quantile range’ $\hat{\tau}(1-2\alpha) = F_n^{-1}(1-\alpha) - F_n^{-1}(\alpha)$, where F_n is the empirical distribution, that measures the width, denoted by $\tau(1-2\alpha)$, of $100(1-2\alpha)\%$ center interval $(F^{-1}(1-\alpha) - F^{-1}(\alpha))$. As a special case, the interquartile range $\tau(0.5)$ is purely used as a robust-type scale parameter. Another example using the quantile range is the process capability index. An alternative approach measuring the distance of a sample subspace with probability λ is

$$\tau_s(\lambda) = c_s^+(\lambda) - c_s^-(\lambda) = 2F_s^{-1}(\lambda).$$

For convenience, we call this the symmetric quantile range. It is clear that $\tau_s(\lambda) = \tau(\lambda)$ when the distribution F is symmetric and c is the central point. The following theorem is a representation of $\hat{\tau}_s = \hat{c}_s^+(\lambda) - \hat{c}_s^-(\lambda)$ that is induced from Theorem 3.

Theorem 5. If $0 < \lambda < 1$, then

$$n^{1/2}(\hat{\tau}_s(\lambda) - \tau_s(\lambda)) = 2(f_c(F_s^{-1}(\lambda)) + f_c(-F_s^{-1}(\lambda)))^{-1}\ddot{E}n^{-1/2}\sum_{i=1}^{n}[\lambda - I(|\varepsilon_i| \le F_s^{-1}(\lambda))]$$

$$+ (f_c(F_s^{-1}(\lambda)) - f_c(-F_s^{-1}(\lambda)))n^{-1/2}(\hat{c} - c)\ddot{E} + o_p(1).$$

We now consider the second alternative choice of robust scale estimator. Trimming variance has been introduced by Staudte & Sheather (1990, p. 124) as

$$d_t = (\alpha_2 - \alpha_1)^{-1}\int_{F_y^{-1}(\alpha_1)}^{F_y^{-1}(\alpha_2)}\left(y - (\alpha_2 - \alpha_1)^{-1}\int_{F_y^{-1}(\alpha_1)}^{F_y^{-1}(\alpha_2)} y \mathrm{d}F_y\right)^2 \mathrm{d}F_y$$

where its estimator, called the sample trimmed variance, is simply replacing the population quantile $F_y^{-1}(\alpha_i)$ by the empirical quantile $F_n^{-1}(\alpha_i)$ for $i = 1$ and 2. An analogue of trimming variance for interpreting the dispersion is denoted as

$$d_{st} = \lambda^{-1}\int_{c^-(\lambda)}^{c^+(\lambda)}\left(y - \lambda^{-1}\int_{c^-(\lambda)}^{c^+(\lambda)} y \mathrm{d}F_y\right)^2 \mathrm{d}F_y$$

We call this version the symmetric trimmed variance. When F_y is symmetric at c, and if we let $\lambda = 1 - 2\alpha$ and $\alpha = \alpha_1 = 1 - \alpha_2$, then we have $d_t = d_{st}$. Thus, the sample symmetric type trimmed variance is

$$\hat{d}_{st} = (n\lambda)^{-1}\sum_{\hat{c}^-(\lambda)\le y_i \le \hat{c}^+(\lambda)}\left(y_i - (n\lambda)^{-1}\sum_{\hat{c}^-(\lambda)\le y_i \le \hat{c}^+(\lambda)} y_i\right)^2$$

This provides an alternative version of the trimmed scale estimator.

We do not further study their large sample properties, although a Monte Carlo study for these two robust scale estimators is performed. We consider a simulation study with sample size $n = 40$ and replication 1000 where the location model

$$y = \theta + \varepsilon$$

with error ε being assumed to be the contaminated normal of equation (2) with $\delta = 0.1$, 0.2. In estimation, we compute the quantile range and symmetric quantile range of $\lambda = 0.7$, 0.8, 0.9 and in Table 4, we display the efficiency of the symmetric type quantile

Table 4. Efficiency of symmetric quantile range

λ	$\delta = 0.1$ $\sigma = 3$	0.1 5	0.1 10	0.1 25	0.2 3	0.2 5	0.2 10	0.2 25
0.7	0.116	0.119	0.124	0.105	0.133	0.131	0.122	0.104
0.8	0.604	0.576	0.581	0.627	0.593	0.577	0.646	0.957
0.9	0.327	0.650	2.997	9.945	0.410	1.306	4.540	28.52

Table 5. Efficiency of symmetric trimmed variance

λ	$\delta = 0.1$ $\sigma = 3$	0.1 5	0.1 10	0.1 25	0.2 3	0.2 5	0.2 10	0.2 25
0.7	1.986	1.745	1.846	1.787	1.871	1.704	1.699	1.815
0.8	1.673	1.669	1.649	1.499	1.643	1.955	2.161	16.72
0.9	1.620	2.217	8.499	90.65	2.128	3.373	8.234	12.95
0.95	2.319	5.078	7.368	12.27	3.016	3.977	4.069	4.522

range as

$$\frac{\text{Mean squares error of quantile range}}{\text{Mean squares error of symmetric quantile range}}$$

Larger values of this ratio indicate that better efficiency is obtained by the symmetric quantile range.

Basically, this table reveals that symmetric quantile ranges are relatively more efficient when contaminated variances are large and quantile percentages are also large. For comparison, we display the simulation result of truncated variance estimators in Table 5, where the efficiency is defined as

$$\frac{\text{Mean squares error of sample trimmed variance}}{\text{Mean squares error of symmetric sample trimmed variance}}$$

As shown in this table, it is surprising that the symmetric-type sample truncated variance estimators are uniformly better than the sample truncated variance estimators.

References

Chen, L.A. & Chiang, Y.C. (1996) Symmetric type quantile and trimmed means for location and linear regression model, *Journal of Nonparametric Statistics*, 7, pp. 171–185.

Chen, L.-A. (1997) An efficient class of weighted trimmed means for linear regression models. *Statistica Sinica*, 7, pp. 669–686.

Ferguson, T.S. (1967) *Mathematical Statistics: A Decision Approach* (New York: Academic Press).

Grimshaw, S.D. & Alt, F.B. (1997) Control charts for quantile function values, *Journal of Quality Technology*, 29, pp. 1–7.

Kim, S.J. (1992) The metrically trimmed means as a robust estimator of location, *Annals of Statistics*, 20, pp. 1534–1547.

Koenker, R. & Bassett, G.J. (1978) Regression quantiles, *Econometrica*, 46, pp. 33–50.

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

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

Welsh, A.H. (1986) Bahadur representations for robust scale estimators based on regression residuals, *The Annals of Statistics*, 14, pp. 1246–1251.

Welsh, A.H. & Morrison, H.L. (1990) Robust L estimation of scale with an application in astronomy, *Journal of the American Statistical Association*, 85, pp. 729–743.