國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

分位數全距

Interpercentile Distance

研究生：樊揚波

指導教授：陳鄰安博士

中華民國 九十四 年 六 月

分位數全距

Interpercentile Distance

研究生：樊揚波　　Student：Yang-Bo Fan

指導教授：陳鄰安博士　　Advisor：Dr. Lin-An Chen

國立交通大學理學院

統計研究所

碩士論文

A thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Statistics
June 2005
Hsinchu, Taiwan, Republic of China

中華民國 九十四 年 六 月

分位數全距

研究生：樊揚波

指導教授：陳鄰安博士

國立交通大學統計研究所

摘　　要

此論文中，我們介紹了一個新的全距，叫做「眾數型態分位數全距」。考慮到測量穩健性估計量時，breakdown point 是一個重要的準則，且早期在建構的穩健性估計量中，breakdown points 都在 0.5 以下(參見 Hampel (1986))，我們將會說明這個新的分位數全距 breakdown point 可以提高到接近 1。我們也利用模擬資料來比較此眾數型態及傳統分位數全距的 MSE。更進一步地，我們會把此分位數全距延伸到建構製程能力指標上。

Interpercentile Distance

Student: Yang-Bo Fan
Adviser: Dr. Lin-An Chen

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

We introduce a new type of range, called the mode type interpercentile distance. With the fact that the breakdown point is one important criterion for measuring the robust type estimators and the fact that the proposed robust estimators are all with breakdown points less than or equal to 0.5)(see this point in Hampel et al. (1986)), we will show that this new interpercentile distance may have breakdown point as large as close to 1. Simulation for comparing this interpercentile distance and the traditional one will also be conducted through the mean square error (MSE). Moreover, an extension of this interpercentile distance to construct a process capability index will also be introduced.

致　　謝

非常感謝陳鄰安老師的教導與協助，不厭其煩地指導我順利完成此論文，在此獻上我誠心的感謝與祝福。並且謝謝彭南夫老師、江永進老師及徐南蓉老師，在口試時給我建議，提供許多寶貴的意見。

謝謝所上老師們的授課，老師們的專業知識及親切隨和的上課氣氛使我這兩年受益匪淺，交大統研所像是個大家庭，很幸運我能在這樣的環境裡學習。

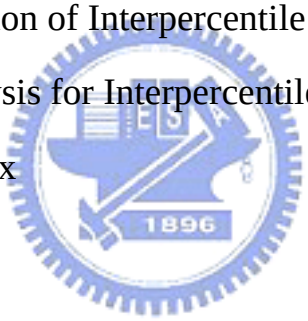
謝謝我研究室的各位成員，婉菁、怡娟、淑儀、如美、秀仁、翊琳及雅靜，在我遇到挫折時熱心地協助我，一同討論一同成長，妳們是很棒的夥伴喔！

另外，要謝謝我的家人們，在我感到沮喪時不斷地支持我，給我繼續往前的力量，陪伴我度過學習路上的低潮期，謝謝你們！

樊揚波

2005.6

Contents



1. Introduction

Measuring the center and variability of a random variable (r.v.) with a distribution function (d.f.) F, unknown or partially unknown, are two most important topics in statistical inference. Basically the variability tries to tell us about the variable X how close together or spread out. In statistics, it is as important to study the variability as the center of a r.v. For example, when products that fit together (such as pipes) are manufactured, it is important to keep the variations of the diameters of the products as small as possible; otherwise, they will not fit together properly.

Although there are varieties of measures for variability in the literature, however, very few of them are designed with clear explanation for their role in measuring spreadness and closeness. One exceptional case is that the variance is explained as a measure of maximum dispersion from the mean. There may not be able to set a line of measure for variability such that spreadness and closeness stay on its two ends. For example for interpretation, suppose that we have a class of students taken an examination of some course and obtain two interval estimates $(10, 85)$ and $(65, 93)$ estimating population intervals covering the student's score with the same probability 0.9. Furthermore, assume that the two population intervals are the longest and the shortest ones under the same confidence coefficient. Then two ranges, 75 and 28, are both estimates of variability, one measuring the interval spreading most widely and the other one measuring the closest interval, both with the same coverage probability. Classifying a measure of variability in its role for spreadness or closeness does make sense for user. Our interest will be interval concerning towards the side of closeness.

Among many choices of variability formulation, the interpercentile distance $\tau(\alpha, \beta) = F^{-1}(\beta) - F^{-1}(\alpha), 0 < \alpha < \beta < 1$, provides variety of versions that are widely applied in practice. The reason is that the symmetric range $F^{-1}(1-\alpha) - F^{-1}(\alpha)$, one in the class, has been shown very useful in application and computationally easy. For examples of application, an alternative formulation of the normal standard deviation is $d\tau_{med}(1-2\alpha)$ with constant d satisfying $d\tau_{med}(1-2\alpha) = \sigma$. The other one application is the 0.5 median range $\tau_{med}(0.5)$, also being called the interquantile range. Both formulations are aiming for providing robust versions of variability measure (see these in Staudte and Sheather (1990)). Very important application of the 0.9973 median range $\tau_{med}(0.9973)$ is done on quality improvement in industry. The process capability index is for judging if a manufacturing process in industry is in control. There are

many types of process capability indices constructed by 99.73% median range, the distribution F been considered symmetric or asymmetric (see Vannaman (1995) and Pearn, Chen (1997) and Kotz and Lovelace (1998)) whereas the simplest version is

$$C_p = \frac{USL - LSL}{\tau_{med}(0.9973)} \tag{1.1}$$

with LSL and USL the lower and upper specification limits determined by engineers.

We consider two criterions for the selection of a interpercentile distance. First, we consider the robustness of breakdown point for an estimator as the largest fraction of the data that can be moved arbitrary without perturbing the estimator to the boundary of the parameter space. Thus the higher the breakdown point, the more robust the estimator against extreme outliers. Among the widths of the quantile interval class $\{(F^{-1}(\alpha), F^{-1}(\gamma + \alpha)) : 0 < \alpha < 1 - \gamma\}$, we define the minimum one as a new measure of variability. Obviously it measures the width of the closest interval in a given coverage probability. As a scale parameter, we will show that it meets several desirable properties of a scale point. This measure of variability is strongly dependent on the shape of a distribution so that it may not blindly be the range of a central interval. Application of this width to build an interpercentile distance, like quantity, alternative representation of standard deviation for any distribution. We will consider all nonparametric estimations for this new interpercentile distance and we will compare it with the traditional symmetric type interpercentile distance through simulations in terms of mean squares error (MSE) and breakdown point. We will also introduce a new process capability index. Illustration of these new procedures are given based on several distributions including the normal, gamma and exponential distributions.

2. Breakdown Point for Some Robust Location Estimators

First we given one example for iid random variable case. Let $y_1, ..., y_n$ be iid random variables.

Sample Mean and Sample Median

The sample mean $\bar{y}$ has breakdown point is $\frac{1}{n}$ which converges to zero and the sample median has breakdown point approximated 0.5.

Hodges-Lehmann Estimator

Consider the location model. The Hodges-Lehmann (HL) estimator (see Hodges

and Lehmann (1963)) is defined as

$$\hat{\theta} = \text{med}_{1 \le i < j \le n} \frac{y_i + y_j}{2}.$$

Then the breakdown point of the Hodges-Lehmann estimator is $1 - (\frac{1}{2})^{1/2}$ which is approximately 0.293.

Suppose that we have a linear regression model

$$y_i = x_i'\beta + \epsilon_i, i = 1, ..., n$$

where x_i is a p-vector of independent variables.

Least Squares Estimator

The least squares (LS) estimator is defined as

$$\hat{\beta}_{LS} = \arg\min_b \sum_{i=1}^{n} (y_i - x_i'b)^2$$

which has breakdown point $\frac{1}{n}$ that converges to zero as n goes to infinity.

Least Median of Squares Estimator

The least median of squares (LMS) estimator, proposed by Rousseeuw (1984), $\hat{\beta}_{LMS}$ solving

$$\arg\min_b \text{med}_i (y_i - x_i'b)^2.$$

The breakdown point of the LMS estimator is $\frac{(n+1)/2}{n}$ which converges to 0.5 as $n \to \infty$.

Least Trimmed Squares Estimator

The least trimmed squares (LTS) estimator proposed by Rousseeuw (1983). Let b be any p vector in R^p. By letting $r_i = y_i - x_i'b, i = 1, ..., n$ and $(r^2)_{i:n}, i = 1, ..., n$ be the order statistics of $r_i^2, i = 1, ..., n$, the LTS estimators is defined as

$$\hat{\beta}_{LTS} = \min_b \sum_{i=1}^{h} (r^2)_{i:n}$$

where $h = [\frac{n}{2}] + 1$. Then the breakdown point of the LTS estimator is $\frac{[(n+1)/2]}{n}$ which converges to 0.5 as $n \to \infty$.

Least Winsorized Squares Estimator

For the location estimation problem, the least Winsorized squares (LWS) estimator (see Rousseeuw (1987)) is defined as

$$\hat{\beta}_{LWS} = \min_b \sum_{i=1}^{h} (r^2)_{i:n} + (n-h)(r^2)_{h:n}$$

where $h = [n/2] + 1$. Then the breakdown point of the LWS estimator is $\frac{[(n+1)/2]}{n}$ which converges to 0.5 as $n \to \infty$.

Least Absolute Values Regression Estimator

The least absolute values regression estimator is defined as

$$\hat{\beta}_L = \arg\min_b \sum_{i=1}^{n} |y_i - x_i' b|$$

which has breakdown point $\frac{1}{n} \to 0$ as $n \to \infty$.

Huber's M-Estimator

The Huber's M-estimator (M) is defined as

$$\arg_b \sum_{i=1}^{n} x_i \psi(y_i - x_i' b) = 0$$

where

$$\psi(z) = \begin{cases} z, & \text{if } |z| < k \\ k \text{ sgn } (z), & \text{if } |z| > k \end{cases}$$

where k is positive constant, usually taking 1.5. The breakdown point of the Huber's M-estimator is $\leq \frac{1}{p}$.

3. Interpercentile Distance

We say that τ_0, a nonnegative function of r.v. X and percentage $\gamma, 0 < \gamma < 1$, is a measure of dispersion if it satisfies

(a). $\tau_0(X + b, \gamma) = \tau_0(X, \gamma)$ for $b \in R$.

(b). $\tau_0(aX, \gamma) = |a|\tau_0(X, \gamma)$ for $a \in R$.

Intuitively, the members in the following family of quantile differences

$$\{F^{-1}(\gamma + \alpha) - F^{-1}(\alpha) : 0 < \alpha < 1 - \gamma\}$$

may serves a γ-range for the distribution. However, not every one in the family satisfies the requirements for a measure of dispersion. It is well known that the median range

$\tau_{med}(1-2\alpha) = F^{-1}(1-\alpha) - F^{-1}(\alpha), 0 < \alpha < 0.5$, is a measure of dispersion (see the proof in Staudte and Sheather (1990)). We are interested in a measure of dispersion that is a quantile combination of the following form

$$\tau_0(X,\gamma) = \inf_{0<\alpha<1-\gamma}\{cF^{-1}(\alpha) + dF^{-1}(\gamma+\alpha)\} \tag{3.1}$$

with $d > 0, c \in R$.

The following theorem provides the condition that the minimization quantile combination in (3.1) is a measure of dispersion.

Theorem 3.1. For given $c, d \in R$, τ_0 of (2.1) is a measure of dispersion if $c = -d$.

Proof. We know that the population quantile F^{-1} satisfies $F^{-1}(X+b,\alpha) = F^{-1}(X,\alpha)+ b$ for $b \in R$ and $F^{-1}(aX,\alpha) = aF^{-1}(X,\alpha)$ if $a > 0$ and $aF^{-1}(X,1-\alpha)$ if $a \le 0$. Now,

$$\begin{aligned}
&\tau_0(X+b,\gamma) = \inf_{0<\alpha<1-\gamma}\{cF^{-1}(X+b,\alpha) + dF^{-1}(X+b,\gamma+\alpha)\} \\
&= \inf_{0<\alpha<1-\gamma}\{cF^{-1}(X,\alpha) + dF^{-1}(X,\gamma+\alpha) + (c+d)b\} \\
&= \tau_0(X,\gamma) + (c+d)b
\end{aligned}$$

which is equal, for satisfying condition (a), to $\tau_0(X,\gamma)$ only if $c+d=0$. Then $d=-c$. It is obvious that (b) holds for $a > 0$. To prove (b) for $a \le 0$, we let $d = 1$.

$$\begin{aligned}
&\tau_0(aX,\gamma) = \inf_{0<\alpha<1-\gamma}\{F^{-1}(aX,\gamma+\alpha) - F^{-1}(aX,\alpha)\} \\
&= \inf_{0<\alpha<1-\gamma}\{aF^{-1}(X,1-(\gamma+\alpha)) - aF^{-1}(X,1-\alpha)\} \\
&= \inf_{\gamma<\beta<1}\{aF^{-1}(X,\beta-\gamma) - aF^{-1}(X,\beta)\} \\
&= -a\ \inf_{\gamma<\beta<1}\{F^{-1}(X,\beta) - F^{-1}(X,\beta-\gamma)\} \\
&= |a|\tau(X,\gamma)
\end{aligned}$$

whch finishes (b). □

Theorem 3.2. Suppose that F has a symmetric continuous density f that is unimodal (meaning that $f(x)$ is trictly decreasing about its center of symmetry). Then $\tau_{mod} = F^{-1}(\frac{1+\gamma}{2}) - F^{-1}(\frac{1-\gamma}{2})$.

Proof.

$$\frac{\partial}{\partial\alpha}(F^{-1}(\gamma+\alpha)-F^{-1}(\alpha))=0$$
$$\frac{1}{f(F^{-1}(\gamma+\alpha))}-\frac{1}{f(F^{-1}(\alpha))}=0$$
$$F^{-1}(\gamma+\alpha)=F^{-1}(\alpha)$$
$$\gamma+\alpha=1-\alpha$$
$$\alpha^{*}=\frac{1-\gamma}{2}$$

4. Nonparametric Estimation of Interpercentile Distance

We consider a nonparametric estimation technique for estimating the unknown interpercentile distance. Parametric methods of data analysis rely on distributional assumptions on the underlying data. Nonparametric methods however, are fully data-driven and hence are particularly suited for the less understood random experiments of highly complexity.

Let $X_{(1)},...,X_{(n)}$ be the order statistics of a random sample of sample size n drawn from a distribution F. By letting $h=[n\gamma]+1$, we define the estimator of γ interpercentile distance as the shortest width of h consecutive sample as

$$\hat{\tau}_{mod}=\arg_{h,h+1,...,n}\min\{X_{(h)}-X_{(1)},X_{(h+1)}-X_{(2)},...,X_{(n)}-X_{(n-h+1)}\}.$$

Having introduced the mode type interpercentile distance as an alternative for the traditional interpercentile distance defined through the ordinary quantile function, we now examine two finite sample numerical aspects. First, it is the fact that the aim for using an interpercentile distance is essentially for robustness consideration. It is the interesting to see if the mode type interpercentile distance is more efficient than the traditional interpercentile distance when the sample is drawn from distributions with outliers. Second, most traditional statistical methods are efficient when the underlying distribution is symmetric. It is then also interesting to see the results of these two interpercentile distances for sample with asymmetric distributions.

The first two questions are answered through a Monte Carlo study using the least squares estimator as the predetermined estimator. How these estimators perform in

the presence of outliers is of particular concern. We consider the following power function model

$$X_i = \mu + \epsilon_i, i = 1, ..., n.$$

With sample size $n = 50$, the distribution of error variable ϵ is the contaminated normal distribution

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

with $\delta = 0.1, 0.2, 0.3$ and $\sigma = 5, 10, 25$. The replication number is 1000. For number ith replication, we compute two interpercentile distance $\hat{\tau}^i_{med}$ and $\hat{\tau}^i_{mod}$. Then we define the mean squares errors,

$$MSE_{med} = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\tau}^i_{med} - \tau_{med})^2 \text{ and } MSE_{mod} = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\tau}^i_{mod} - \tau_{mod})^2.$$

In Table 1, we display the MSE's under the contaminated normal distributions. The purpose of the Monte Carlo study is to evaluate the small-sample behavior of these two interpercentile distances.

Table 1. MSE's for the median and mode type interpercentile distance under contaminated normal distributions

γ	MSE_{mod}	MSE_{med}	MSE_{mod}	MSE_{med}	MSE_{mod}	MSE_{med}
	$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.3$	
	$\sigma = 5$					
$\gamma = 0.5$	0.0999	0.0757	0.1194	0.1097	0.1396	0.1623
$\gamma = 0.6$	0.1188	0.1153	0.1477	0.2052	0.1925	0.3475
$\gamma = 0.7$	0.2006	0.1271	0.2678	0.2634	0.4361	0.5531
$\gamma = 0.8$	0.2291	0.4217	0.4869	1.4770	1.3679	2.4403
$\gamma = 0.9$	0.9840	1.9936	5.8309	3.6026	15.397	5.8729
	$\sigma = 10$					
$\gamma = 0.5$	0.1044	0.0801	0.1299	0.1245	0.1694	0.2635
$\gamma = 0.6$	0.1256	0.1276	0.1706	0.3230	0.3075	1.3354
$\gamma = 0.7$	0.2157	0.1689	0.3557	0.9816	1.0804	4.2862
$\gamma = 0.8$	0.2937	1.8739	2.1154	13.041	11.279	13.946
$\gamma = 0.9$	2.6851	17.422	46.149	21.014	78.732	27.399
	$\sigma = 25$					
$\gamma = 0.5$	0.1081	0.0818	0.1364	0.1378	0.1975	0.8805
$\gamma = 0.6$	0.1312	0.1394	0.1902	1.0873	0.8545	11.930
$\gamma = 0.7$	0.2295	0.3719	0.5886	8.5978	7.3271	51.315
$\gamma = 0.8$	0.5810	16.429	20.889	142.35	117.37	98.386
$\gamma = 0.9$	18.903	184.29	391.14	157.09	540.10	179.46

We have several conclusions drawn from the results displayed in Table 1:
(a) When the coverage probability γ increases both the MSE's of the two interpercentile distances also increase. This shows that both techniques of interpercentile distances are less efficient when the coverage probability is large.
(b) When the coverage probability γ is a smaller value (< 0.7) the median type interpercentile distance seems to be more efficient in estimation. However, when it is relatively larger (> 0.7) the mode type interpercentile distance seems to be mostly more efficient than the other one.
(c) In practical applications of the use of interpercentile distance it is quite often that large coverage probability is adopted (for interpercentile distance $\gamma = 0.75$ is adopted). Then the mode type interpercentile distance is a good choice in defining this type scale.

In the next we conduct a simulation with a design of sample sizes $n = 50$ and 100 to compare the MSE's of the two interpercentile distances where the underlying distribution is the exponential one with pdf

$$f(x) = e^{-x}, x > 0.$$

This is a study for a situation that there is no contaminated outliers. The results of MSE's are displayed in Table 2.

Table 2. MSE's of the median and mode type interpercentile distances under the exponential distribution

γ	MSE_{mod}	MSE_{med}	MSE_{mod}	MSE_{med}
	$n = 50$		$n = 100$	
0.5	0.0208	0.0548	0.0096	0.0272
0.6	0.0321	0.0839	0.0149	0.0373
0.7	0.0469	0.1031	0.0234	0.0508
0.8	0.1010	0.1579	0.0433	0.0838
0.9	0.1503	0.2522	0.0819	0.1547
0.95	0.3507	0.5979	0.1503	0.3085

It is nice that the mode type interpercentile distance has MSE's all are less than the corresponding value of the median type interpercentile distance.

Let's conduct several more study by computing the efficiency defining by

$$Eff_{med} = \frac{\text{Min}\{MSE_{med}, MSE_{mod}\}}{MSE_{med}} \text{ and } Eff_{mod} = \frac{\text{Min}\{MSE_{med}, MSE_{mod}\}}{MSE_{mod}}.$$

The first is conducted for the Chi-square distribution $\chi^2(k)$ where the sample size is $n = 50$ and replication is $m = 10000$. The simulation results are displayed in Table 3.

Table 3. Efficiencies of two interpercentile distance for chi-square distribution

γ	Eff_{mod}	Eff_{med}	Eff_{mod}	Eff_{med}
	$k = 3$		$k = 10$	
0.6	1	0.2702	1	0.4445
0.7	1	0.3387	1	0.5881
0.8	1	0.3821	1	0.6318
0.9	1	0.4910	1	0.7181
0.95	1	0.5449	1	0.6661

The above table provides the clear results of the efficiencies of these two interpercentile distances where the median type one could have efficiency as small as 0.27 and are all less than 0.72. This support to use the mode type interpercentile distance when the distribution follows the Chi-square one.

In the next we consider a contaminated exponential distribution as follows

$$X = (1 - \delta)Exp(1) + \delta Uni(-20, 20).$$

where $Exp(1)$ is the exponential distribution we defined before and $Uni(-20, 20)$ represents the variable with probability 0.5 for either value 20 or -20. The sample size and replication number are again $n = 50$ and $m = 10000$. The simulation results are displayed in the following table.

Table 4. Efficiencies of two interpercentile distances for contaminated exponential distribution

γ	Eff_{mod}	Eff_{med}	Eff_{mod}	Eff_{med}	Eff_{mod}	Eff_{med}
	$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.3$	
0.5	1.0000	0.0192	1.0000	0.0003	1.0000	0.0001
0.6	1.0000	0.0043	1.0000	0.0056	1.0000	0.1280
0.7	1.0000	0.0013	1.0000	0.0655	1.0000	0.5413
0.8	1.0000	0.0886	1.0000	0.7248	0.9696	1.0000
0.9	1.0000	0.6623	0.9537	1.0000	0.9412	1.0000
0.95	0.9300	1.0000	0.9298	1.0000	0.9298	1.0000

The efficiencies of the mode type interpercentile range for this distribution are almost with few exceptions better than the median type interpercentile range.

5. Breakdown Point Analysis for Interpercentile Distance

The classical statistical techniques are designed to be the best possible when stringent assumptions apply. However, experience and further research have forced us to recognize that classical techniques can behave badly when the practical situation departs from the ideal described by such assumptions. The more recently developed robust and exploratory methods are broadening the effectiveness of statistical analysis. One aspect to evaluate the effectiveness is to compare the breakdown point for the estimators.

The breakdown point for an estimator, loosely speaking, is the largest proportion of gross errors that never can carry the estimator over all bounds. The sample size is 1000 with replication $m = 100$ and this sample are drawn from the following distribution model,

$$X_i = \begin{cases} Z_i & \text{if outlier does not occurs} \\ Z_i + v_i & \text{if outlier does occurs} \end{cases}$$

where z_i are iid drawn from an ideal distribution and $v_i = 1000 + 10 * i$. If $X_i = Z_i + v_i$ then this x represents an extreme point.

For replication number j, we generate a sample $z_1, ..., z_n$. Then we define the breakdown number bd^j_{med} and bd^j_{mod} as

$$bd^j_{med}(z_1, ..., z_n) = \frac{1}{n}\max\{k : \max_{i_1,...,i_k} |\hat{\tau}_{med}(x_1, ..., x_n) - \tau_{med}| \geq a\}$$
$$bd^j_{mod}(z_1, ..., z_n) = \frac{1}{n}\max\{k : \max_{i_1,...,i_k} |\hat{\tau}_{mod}(x_1, ..., x_n) - \tau_{mod}| \geq a\}$$

where the sample $x_1, ..., x_n$ is obtained by replacing the k data points $z_{i_1}, ..., z_{i_k}$ by the contaminated values $z_{i_1} + v_1, ..., z_{i_k} + v_k$. The average breakdown points are then defined as

$$BD_{med} = \frac{1}{m}\sum_{j=1}^{m} bd^j_{med}(z_1, ..., z_n) \text{ and } BD_{mod} = \frac{1}{m}\sum_{j=1}^{m} bd^j_{mod}(z_1, ..., z_n).$$

In the following table, we present the average breakdown points of mode type and median type interpercentile distances under the case that the ideal distribution is standard normal distribution.

Table 5. Breakdown points under normal distribution

γ	BD_{mod}	BD_{med}	γ	BD_{mod}	BD_{med}
0.95	0.051	0.027	0.5	0.5	0.251
0.9	0.101	0.052	0.4	0.6	0.301
0.8	0.2	0.101	0.3	0.7	0.351
0.7	0.301	0.152	0.2	0.8	0.401
0.6	0.4	0.201	0.1	0.9	0.451

In the next simulation, we consider the Gamma distribution with $\alpha = 2.5$ and $\beta = 2$ as the ideal distribution.

Table 6. Breakdown points under Gamma distribution ($Gamma(2.5, 2)$)

γ	BD_{mod}	BD_{med}	γ	BD_{mod}	BD_{med}
0.95	0.0502	0.0267	0.5	0.4968	0.2484
0.9	0.0992	0.0514	0.4	0.597	0.2984
0.8	0.1957	0.0996	0.3	0.6973	0.3485
0.7	0.2938	0.1502	0.2	0.7976	0.3987
0.6	0.3891	0.1986	0.1	0.8986	0.4497

From Tables 5 and 6, we have several conclusions:

(a) The breakdown points for the mode type interpercentile distance are about twice the values of the median type interpercentile distance. It for mode type interpercentile distance is about $1 - \gamma$ and it for median type interpercentile distance is about a half of $1 - \gamma$.

(b) The breakdown point for each type interpercentile distance is increasing when γ decreases.

(c) Hampel et al. (1986) claimed that the breakdown point for estimator may not be larger than 0.5. However, the breakdown point of the mode type interpercentile distance is available not only more than 0.5 but also close to 1. This interesting result has not been observed in the literature.

Now, we consider the exponential distribution as the ideal distribution and we display the simulation results of breakdown points of the two interpercentile distances in the following table.

Table 7. Breakdown points under Exponential distributions ($Exp(5)$)

γ	BD_{mod}	BD_{med}	γ	BD_{mod}	BD_{med}
0.95	0.0446	0.0236	0.5	0.4627	0.2267
0.9	0.0881	0.0451	0.4	0.5648	0.2762
0.8	0.1779	0.0891	0.3	0.6682	0.3271
0.7	0.2710	0.1344	0.2	0.7765	0.3816
0.6	0.3653	0.1802	0.1	0.8864	0.4385

The results displayed in Table 7 for this exponential distribution are similar to the results in Tables 5 and 6. However, they are less than the results in Tables 5 and 6.

The above results of breakdown points are all performed with $a = 10$. We may want to see if the breakdown points can be improved if increase the value a. In the following Table, we display a result for that a is set to be 25.

Table 8. Breakdown points under Exponential distributions ($Exp(5)$)

γ	BD_{mod}	BD_{med}	γ	BD_{mod}	BD_{med}
0.95	0.0507	0.0268	0.5	0.4983	0.2500
0.9	0.1003	0.0517	0.4	0.5984	0.2999
0.8	0.1987	0.1003	0.3	0.6986	0.3500
0.7	0.2996	0.1512	0.2	0.7989	0.4001
0.6	0.3984	0.1999	0.1	0.8994	0.4505

In this situation, the breakdown points for the two types of interpercentile distance are improved and their results are very close to the results in Tables 5 and 6.

6. Process Capability Index

In understanding what the process is actually doing and seeing if the process meets the quality requirements or the consumer's expectations, the manufacturer often needs to provides an index for process improvement and the certificate for customers. In a perfect world, all process data would be normally distributed. If only that were true! The usual process capability analysis has provided some very powerful tools to describe the capability of processes. However, the indices for the usual process capability analysis are designed to be used with normally distributed data.

Process data do not always follow a normal distribution. A one-sided specification limit is an immediate clue that the data might be non-normal. For example, a chemical may have an upper specification limit (USL) for a contaminant. The impurity concentration cannot be less than zero, and the normal distribution is unlikely to be a good model. In most of the earlier work for dealing with non-normal data, the authors

have tried to fit an appropriate probability distribution of the process from available data and then define indices based on the estimated distribution. Such an approach would require large amounts of data to have a clear understanding of the shape of the distribution and the analysis can also be very sensitive to departure from that distribution.

The most commonly used techniques to handle non-normal data are transformation and quantile estimation. Many practitioners are not comfortable with transformed data and may have difficulty in translating the results back to the original scale. Many a time, it will be also difficult to identify the correct transformation. Clements (1989) has proposed a pioneering approach to the modification of process capability indices for non-normality considering estimates of median type interpercentile distance $\tau_{med}(0.9973)$. Gilchrist (1993), Chang and Lu (1994) and Sundaraiyer (1996) have extended Clements method to incorporate various related situations.

The capability index generally defined as

$$\frac{\text{Specification limit}}{\text{Process spreading limit}}.$$

is the most popular one for the purposes. Numerous process capability indices, including the one in (1.1), dealing with normal and asymmetric distributions have been provided. From their formulation, we see that these indices are all constructed by median range $\tau_{med}(0.9973)$ no matter what the distribution is dealing for (see Clements (1989), Kotz and Lovelace (1998), Pearn and Chen (1997) and Yeh and Bhattacharya (1998)). We here will introduce an analogue of the simplest one in (1.1) through the mode range as an alternative process capability index where other indices involving median range may also be analogously developed.

Definition 6.1. The mode process capability index is defined as

$$C_p^{mod} = \frac{USL - LSL}{\tau(\gamma)}.$$

We also call it the mode C_p.

To investigate the efficiencies of the mode process capability, we will proceed a simulation to compare the mean square errors (MSE) of this mode process capability and the median type process capability. To do this, we randomly drawn a sample of size

$n = 100$ from a distribution and compute their corresponding sample process capabilities. With replication 1000 and we compute the average process capabilities. Without loss of generality, we set the true value of $USL - LSL = 1$. Furthermore we consider process capabilities various in coverage probability γ with $\gamma = 0.7, 0.8, 0.9, 0.95$.

Exponential Distribution

Let $X_1, ..., X_n$ be a random sample drawn from the exponential distribution $Exp(\lambda)$ with pdf

$$f(x, \lambda) = \frac{1}{\lambda} e^{-x/\lambda}, x > 0.$$

In this simulation, we let sample size $n = 1000$ and replications $m = 100$. The observations are randomly drawn from the exponential distribution $Exp(\lambda)$ with $\lambda = 5$. We consider process capability indices of $\gamma = 0.7, 0.8, 0.9, 0.95$ The extreme point setting is as what we have done and our purpose is to compute the average of breakdown points of these two process capability indices. Moreover, we define an estimator to be broken if the following occurs

$$C_p < a$$

where the value of a are 0.02 and 0.01.

Table 9. Breakdown points under Exponential distributions ($Exp(5)$)

γ	BD_{mod}	BD_{med}	BD_{mod}	BD_{med}
	$a = 0.02$		$a = 0.01$	
0.95	0.051	0.027	0.101	0.111
0.9	0.101	0.052	0.117	0.100
0.8	0.200	0.101	0.220	0.111
0.7	0.301	0.152	0.301	0.152

The mode type C_p has breakdown points better than those of the median type C_p. In case that $a = 0.02$, the breakdown point of the mode type C_p is about $1 - \gamma$ and it of the median type C_p is about a half of $1 - \gamma$. However, case for $a = 0.01$ is not so desirable.

The next we consider the normal distribution $N(0, 1)$ and conduct the same simulation. We display the simulation results of breakdown points in the following table.

Table 10. Breakdown points under Normal distribution

γ	BD_{mod}	BD_{med}	BD_{mod}	BD_{mod}
	$a = 0.02$		$a = 0.01$	
0.95	0.051	0.027	0.051	0.027
0.9	0.101	0.052	0.101	0.052
0.8	0.200	0.101	0.200	0.101
0.7	0.301	0.152	0.301	0.152

For this case of normal distribution, the breakdown points for these two interpercentile distances are with results desirable. The above simulation results support us to use the mode type interpercentile distance.

The efficiency of the process capability index is

$$Eff = \frac{\min\{\text{MSE}_{mod}, \text{MSE}_{med}\}}{\text{MSE}}$$

where $MSE = MSE_{mod}$ or MSE_{med}.

Table 11. Efficiencies of process capability indices

γ	Eff_{mod}	Eff_{med}	Eff_{mod}	Eff_{med}
	$\lambda = 5$		$\lambda = 10$	
0.7	1	0.5005	1	0.5194
0.8	1	0.4179	1	0.4232
0.9	1	0.5396	1	0.5407
0.95	1	0.6269	1	0.6272

In the next, we presents the simulation results for chi-square distribution $\chi^2(k)$.

Table 12. Efficiencies of process capability indices

γ	Eff_{mod}	Eff_{med}	Eff_{mod}	Eff_{med}
	$k = 3$		$k = 10$	
0.7	1	0.5445	1	0.8796
0.8	1	0.5993	1	0.8842
0.9	1	0.6660	1	0.8961
0.95	1	0.7048	1	0.8978

From the results displayed in Tables 11 and 12, it is interesting that the mode type interpercentile distance has simultaneously smaller MSE's than the median type interpercentile distance so that the mode type one has efficiencies all with values 1s.

7. Conclusions

In this paper, we proposed the mode type interpercentile distance. The most interesting result showing by this distance is that its breakdown point may be greater

than 0.5 which was claimed by Hampel et al. (1986) that the breakdown point for any estimator may not be larger than 0.5. We also made a simulation showing that this new interpercentile distance may be more efficient than the traditional interpercentile distance. We also introduce the mode type process capability index where simulation results of breakdown point and MSE for comparing it and the traditional one are also displayed.

References

Chang, P. and Lu, K. (1994). PCI calculations with any shape distribution with percentile. *QWTS*, 110-114.

Clements, J. A. (1989). Process capability calculations for non-normal distributions. *Quality Progress*, September, 95-100.

Gilchrist, W. G. (1993). Capability of the customer-supplier chain. *First Newcastle Conference on Quality and its Applications.* Penshaw Press, Newcastle-upon-Tyne, United Kingdom, 587-591.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based On Influence Function.* Wiley, New York.

Hodges, J. L., Jr. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics.* 34, 598-611.

Kotz, S. and Lovelace, C. R. (1998). *Process Capability Indices in and Practice.* Arnold: London.

Pearn, W. L. and Chen, K. S. (1997). Capability indices for non-normal distributions with an application in electrolytic capacitor manufacturing. *Microelectronic Reliability*, 1-6.

Rousseeuw, P. J. (1983). Multivariate estimation with high breakdown point, paper presented at fourth Pannonian symposium on mathematical statistics and probability. Bad Tatzmannsdorf, Australia.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association.* 79, 871-880.

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

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

Sundaraiyer, V. H. (1996). Estimation of a process capability index for inverse Gaussian distributions. *Communications in Statistics: Theory and Methods* 26, 2381-2396.

Vannman, K. (1995). A unified approach to capability indices. *Statistica Sinica*, 5, 805-820.

Yeh, A. B. and Bhattacharya, S. (1998). A robust process capability index. *Communications in Statistics - Simulation and Computation*, 26, 565-589.