On the efficiency of the Trimean and Q123

John R. Doyle *
Cardiff Business School
Cardiff University
Aberconway Building, Colum Drive
Cardiff CF10 3EU
United Kingdom

Catherine Huirong Chen
Middlesex University Business School
Middlesex University
Hendon Campus, The Burroughs
London NW4 4BT
United Kingdom

Abstract

A Monte Carlo study showed that, for all but small samples with high levels of contamination, the robust efficiency of Tukey's trimean may be improved by single-weighting the median.

Keywords and phrases : *Quartiles, contaminated Gaussian distribution, biweight, robust, median, Monte Carlo.*

1. Introduction

The purpose of this paper is to demonstrate that a simple alteration to the trimean improves its efficiency as a measure of central tendency across a variety of distributions. The trimean was originally defined by Tukey [4] to be TRI $= (H_1 + 2M + H_2)/4$, where M is the median and H_1 and H_2 are the lower and upper hinges, respectively. A search of the internet suggests that the trimean (and, incidentally, the boxplot) is now

E-mail: doylejr@cardiff.ac.uk

Journal of Statistics & Management Systems
Vol. ( ), No. , pp. 1–5

more widely defined by using Q_1 and Q_3, the lower and upper quartiles, instead of H_1 and H_2, respectively. We will follow this trend.

TRI is part of the Exploratory Data Analysis toolkit; it is intuitively appealing, simple to understand and compute, and it maintains its relative efficiency well in the presence of contamination. Our suggestion is to define a new measure, namely Q123 $= (Q_1 + M + Q_3)/3$. The quartiles are computed according to Definition 8 in Hyndman and Fan [3], which is the version the authors recommend from their review of competing definitions of quantiles. Our new statistic Q123 clearly does not sacrifice any of the simplicity of TRI, but we show that it is more efficient than TRI, both when the distribution is normal or symmetrically contaminated normal.

2. Analysis of data

To this end, we conducted a Monte Carlo study in which samples of 10, 15, 25 or 250 observations were drawn from a normal distribution $N(0,1)$ with either no contamination, 5% contamination, or 10% contamination. We varied the severity of the contaminating distribution, by drawing from $N(0,4)$, $N(0,9)$, $N(0,25)$, or $N(0,100)$. In each condition 100,000 samples were drawn, and for each sample the mean, median, TRI, and Q123 were calculated. Although our focus is on simple statistics, as a benchmark we also calculated the biweight with tuning constant $C = 6$, as suggested in Mosteller and Tukey [3], because it is known to be quite robustly efficient [2]. The relative efficiencies, being just the reciprocal of the variances of the sampling distributions with the mean for the Gaussian indexed to 100, are recorded in Table 1 for each of these statistics and for each of the conditions. Does TRI or Q123 more frequently find estimates that are closer to their common Expected Value of 0? The second last column is a count of the number of samples in which abs$(\text{TRI}) >$ abs(Q123). If this count exceeds 50000, then Q123 is more efficient in this sense. The last column shows for each condition the empirically derived weighting, w of the median in the statistic $(Q1 + wM + Q3)/(2 + w)$ that would minimize its sampling variance, based on 1 million samples per condition.

By the variance definition, Q123 was more efficient than TRI in 31 of the 36 conditions, and in 34 conditions by the count measure. Focusing on the conditions in which TRI performed better, which have been emboldened in Table 1, we see that they occur for small samples with large contaminations. Now, Q123 and TRI have the same breakdown point,

Table 1

The efficiency of estimators of central tendency for different degrees of Gaussian contamination. The second last column is the count of the number of times from 100,000 that $|\mathbf{TRI}| > |\mathbf{Q123}|$. The last column is the optimal Median Weight for that condition

n	% contam	s.d. of contam.	Mean	Median	Biweight	TRI	Q123	$\lvert TRI\rvert > \lvert Q123\rvert$	Optimal MedWt
10	0	n.a.	100.0	72.6	87.2	88.2	90.8	52357	0.92
10	5	2	87.2	68.6	81.3	82.4	84.6	52013	1.04
10	10	2	77.7	64.8	76.1	77.1	78.9	51467	1.16
10	5	3	72.2	67.4	79.9	80.6	82.0	51488	1.18
10	10	3	55.9	62.3	71.8	72.3	72.5	50447	1.56
10	5	5	45.5	65.5	78.5	**77.9**	**77.3**	51074	1.45
10	10	5	29.5	59.5	68.6	**67.0**	**63.2**	49276	2.85
10	5	10	16.7	64.5	80.0	**76.3**	**70.9**	50543	2.40
10	10	10	9.3	57.2	67.2	58.8	45.9	47350	12.51
15	0	n.a.	100.0	65.6	87.4	86.2	89.1	54336	0.74
15	5	2	87.5	62.3	81.9	81.1	83.4	54058	0.81
15	10	2	76.8	58.4	75.9	75.2	77.1	53807	0.89
15	5	3	71.1	60.9	80.1	78.3	80.2	53616	0.88
15	10	3	56.2	56.6	73.3	71.8	73.0	53350	1.07
15	5	5	45.8	59.8	79.9	76.4	78.0	53538	0.97
15	10	5	29.4	54.4	71.6	66.5	66.6	52904	1.41
15	5	10	16.9	59.2	82.1	74.9	76.0	53337	1.08
15	10	10	9.1	52.9	74.4	**60.0**	**57.7**	52424	2.53
25	0	n.a.	100.0	64.6	88.2	85.2	88.1	53819	0.72
25	5	2	86.5	61.0	82.3	79.8	82.3	53563	0.78
25	10	2	76.1	57.9	76.9	75.0	76.9	52931	0.84
25	5	3	70.9	59.6	80.5	77.4	79.6	53186	0.83
25	10	3	55.6	55.5	74.0	71.1	72.6	52665	0.97
25	5	5	44.9	58.6	80.2	75.2	76.9	53124	0.89
25	10	5	29.2	53.9	73.7	67.3	68.0	51641	1.14
25	5	10	16.7	58.6	82.7	74.7	76.2	52681	0.94
25	10	10	9.1	52.5	76.7	64.4	64.5	51054	1.37
250	0	n.a.	100.0	63.9	91.2	84.0	86.4	53155	0.86
250	5	2	865	60.6	85.1	78.7	80.5	52616	0.92
250	10	2	77.2	57.6	80.1	74.6	76.3	52171	0.98
250	5	3	71.6	59.7	84.6	77.5	79.3	52562	0.96
250	10	3	55.1	55.3	77.1	70.5	71.7	51846	1.09
250	5	5	45.2	58.7	84.1	75.3	76.6	52184	1.00
250	10	5	29.2	54.2	77.9	68.1	68.7	51023	1.19
250	5	10	16.9	58.1	86.0	74.5	75.8	52320	1.04
250	10	10	9.2	53.0	81.2	66.1	66.5	50718	1.29

which is at best .25 (when $n = 10$, for instance, because the quartiles are calculated by interpolation, we only need change 2 points to make either quartile arbitrarily large, implying a breakdown point of .2). The greater the percentage contamination, and/or the smaller the sample, the more likely it is that the contaminating distribution will breach this breakdown point, resulting in a loss of efficiency. With sufficient contamination, double-weighting the median, as in TRI, will tend to slightly dilute the effects of a contaminated quartile in TRI relative to Q123. However, as contamination farther increases it soon becomes an even better idea to place all weight on the median, whose breakdown point is .5. This trend is evident in the final column of Table 1, where the optimal weight to place on M is 12.51 when there is 10% contamination from $N(0, 100)$ and $n = 10$. One further simulation showed that the median is more efficient than TRI for 15% contamination from $N(0, 100)$ for sample size $n = 10$. Therefore, TRI has a limited domain in which it dominates the other simple estimates.

One final point is that if simulations are run with contaminating distribution that have variances < 1, thus moving probability mass towards the central region, rather than towards the tails, then double-weighting the median starts to become a good idea again. In these rather specialised circumstances, with approximately 10% contamination from $N(0, .25)$, TRI performs better than Q123. But go too far in this direction and, same old story, the median outperforms other measures anyway.

3. Summary

In summary, each of the simple estimates has its own preferred habitat. The mean enjoys the advantage in near-Gaussian distributions, the median for highly contaminated ones, Q123 for moderate to high contamination. TRI's preferred habitat is squeezed between these last two. Slightly surprising was that Q123 outperformed the biweight for small samples with moderate or no contamination. Given this, and that across all 36 conditions the median of the optimal weights on M was 1.02, and because Q123's habitats are more frequently encountered than TRI's, we recommend that it be adopted in preference to TRI.

References

[1] C. Goodall (1983), M-estimators of location: an outline of the theory, in *Understanding Robust and Exploratory Data Analysis*, F.D.C. Hoaglin (editor), ???.

[2] R. J. Hyndman and Y. Fan (1996), Sample quantiles in statistical packages, *The American Statistician*, Vol. 50 (4), pp. 361–365.

[3] F. Moseller and J. W. Tukey (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

[4] J. W. Tukey (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Received May, 2008