On Preliminary Test Estimator for Median

Okazaki, Takeo
Department of Information Systems, Interdisciplinary Graduate School of Engineering Sciences, Kyushu University

https://doi.org/10.15017/17207

出版情報：九州大学大学院総合理工学報告. 12 (3), pp.347-354, 1990-12-01. Interdisciplinary Graduate School of Engineering Sciences, Kyushu University
バージョン：
権利関係：

九州大学総合理工学研究科報告
第12巻 第3号 347-354頁 平成2年12月

Engineering Sciences Reports, Kyushu University
(KYUSHU DAIGAKU SOGORIKOGAKU KENKYUKA HOKOKU)
Vol. 12, No. 3, pp. 347-354, December 1990

On Preliminary Test Estimator for Median

Takeo OKAZAKI*

(Received August 31, 1990)

The purpose of the present paper is to discuss about estimation of median with a preliminary test. Two procedures are presented, one uses Median test and the other uses Wilcoxon two-sample test for the preliminary test. Sections 3 and 4 give mathematical formulations of such properties, including mean square errors with one specified case. Section 5 discusses their optimal significance levels of the preliminary test and proposes their numerical values by Monte Carlo method. In addition to mean square errors, mean absolute errors are used for the regret criterion.

1. Introduction

Estimation with a preliminary test that aims at pooling of data has been investigated under the parametric model mainly. Especially normal assumption was used frequently because of the easiness to formalize and to evaluate its performances. However it is rare that such assumption is satisfied in real case, as these procedures are intended for small sample cases.

Tamura (1965, 1967) discussed about some properties of preliminary test estimator for the population median and shift parameter in the nonparametric circumstances, and derived expected values and mean square errors and relative efficiency with respect to the never pooling estimator numerically, but didn't refer to the selection of significance level for the preliminary test, though it is quite necessary in practical situations.

This paper presents two procedures for estimating population median, one uses Median test and the other uses Wilcoxon two-sample test for the preliminary test. To our regret it is almost impossible to derive the general forms available for arbitrary sample size and significance level of the preliminary test, then one specified case is dealt with. Minimax regret principle is applied to determine the optimal significance level for the preliminary test. For the regret criterion, mean square errors and mean absolute errors are used. Monte Carlo method is adopted for numerical evaluation.

2. Procedures

Let $X=(X_1,\cdots, X_m)$ and $Y=(Y_1,\cdots, Y_n)$ be respectively the random samples of size m, n from the continuous distributions $F(x)$ and $F(x-\delta)$, where $F(\xi)=0.5$ and $\delta \geq 0$. Our main object is to estimate the value of the population median ξ and for this purpose, we first test the hypothesis $\delta=0$. The preliminary test is done by Median test or Wilcoxon two-sample test and the sample median is used as the estimator of ξ.

Let now the critical regions of size α for Median test and Wilcoxon two-sample test be

*Department of Information Systems

$R_M(\alpha)$, $R_W(\alpha)$ respectively. Then the estimator with Median test is defined as follows,

(a) $\tilde{\xi} = median(X)$, if $(X, Y) \in R_M(\alpha)$,

(b) $\tilde{\xi} = median(X, Y)$, if $(X, Y) \in R_M^c(\alpha)$.

And the estimator with Wilcoxon two-sample test is defined as follows,

(a) $\hat{\xi} = median(X)$, if $(X, Y) \in R_W(\alpha)$,

(b) $\hat{\xi} = median(X, Y)$, if $(X, Y) \in R_W^c(\alpha)$.

Here, *median* (X), *median* (X, Y) imply the sample median of X and the sample median of pooled data (X, Y) respectively.

3. Distribution

Since we cannot derive the general distribution of $\tilde{\xi}$ and $\hat{\xi}$ in the forms available for arbitrary values of m, n and α, we shall deal with the case $m=3$, $n=4$, and set the significance level of the preliminary test to 0.1143, because it is the exact level that exists in both tests.

The critical regions $R_M(0.1143)$, $R_W(0.1143)$ are determined by the following orderings,

$R_M(0.1143) = \{(XXXYYYY), (XXYXYYY), (XYXXYYY), (YXXXYYY)\}$,

$R_W(0.1143) = \{(XXXYYYY), (XXYXYYY), (XYXXYYY), (XXYYXYY)\}$.

If (X, Y) is in the region of other orderings, the hypothesis $\delta = 0$ is accepted. The density of the estimator $\tilde{\xi}$ is given as follows,

$$
\begin{aligned}
g_M(z)\,dz = Pr\ & [z<X<z+dz,\ X_1<X<X_2<\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\
& + Pr\ [z<X<z+dz,\ X_1<X<Y_1<X_2<\begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\
& + Pr\ [z<X<z+dr,\ \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}<X<X_2<\begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\
& + Pr\ [z<X<z+dz,\ \begin{pmatrix} X_1 \\ Y_1 \\ Y_2 \end{pmatrix}<X<\begin{pmatrix} X_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\
& + Pr\ [z<X<z+dz,\ \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}<X<\begin{pmatrix} X_1 \\ X_2 \\ Y_4 \end{pmatrix}] \\
& + Pr\ [z<Y<z+dz,\ \begin{pmatrix} X_1 \\ Y_1 \\ Y_2 \end{pmatrix}<Y<\begin{pmatrix} X_2 \\ X_3 \\ Y_3 \end{pmatrix}] \\
& + Pr\ [z<Y<z+dz,\ \begin{pmatrix} X_1 \\ X_2 \\ Y_1 \end{pmatrix}<Y<\begin{pmatrix} X_3 \\ Y_2 \\ Y_3 \end{pmatrix}]
\end{aligned}
$$

$$+Pr\ [z<Y<z+dz, \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} < Y < \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}].$$

And the density function $g_M(z)$ may be expressed as follows,

$$\begin{aligned} g_M(z) =& 6F(z)f(z)\int_z^\infty f(x)\,(1-F(x-\delta))^4 dx \\ &+24F(z)f(z)\int_z^\infty (F(x-\delta)-F(z-\delta))f(x)\,(1-F(x-\delta))^3 dx \\ &+24F(z)F(z-\delta)f(z)\int_z^\infty f(x)\,(1-F(x-\delta))^3 dx \\ &+36F(z)F(z-\delta)^2 f(z)\,(1-F(z))\,(1-F(z-\delta))^2 \\ &+12F(z-\delta)^3 f(z)\,(1-F(z))^2(1-F(z-\delta)) \\ &+36F(z)F(z-\delta)^2 f(z-\delta)\,(1-F(z))^2(1-F(z-\delta)) \\ &+36F(z)^2 F(z-\delta)f(z-\delta)\,(1-F(z))\,(1-F(z-\delta))^2 \\ &+4F(z-\delta)^3 f(z-\delta)\,(1-F(z))^3. \end{aligned}$$

On the other hand, the density of the estimator $\hat{\xi}$ is given as follows,

$$\begin{aligned} g_W(z)\,dz =& Pr\ [z<X<z+dz,\ X_1<X<X_2< \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz,\ X_1<X<Y_1<X_2< \begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz,\ X_1<Y_1<X<X_2< \begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz,\ X_1<X< \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} <X_2< \begin{pmatrix} Y_3 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz, \begin{pmatrix} X_1 \\ Y_1 \\ Y_2 \end{pmatrix} <X< \begin{pmatrix} X_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz, \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} <X< \begin{pmatrix} X_1 \\ X_2 \\ Y_4 \end{pmatrix}] \\ &+Pr\ [z<X<z+dz,\ Y_1<X_1<X<X_2< \begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}] \end{aligned}$$

$$+Pr\ [z<Y<z+dz, \begin{pmatrix} X_1 \\ Y_1 \\ Y_2 \end{pmatrix}<Y<\begin{pmatrix} X_2 \\ X_3 \\ Y_3 \end{pmatrix}$$

$$+Pr\ [z<Y<z+dz, X_1<Y_1<X_2<Y<\begin{pmatrix} X_3 \\ Y_2 \\ Y_3 \end{pmatrix}]$$

$$+Pr\ [z<Y<z+dz,\ Y_1<\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}<Y<\begin{pmatrix} X_3 \\ Y_2 \\ Y_3 \end{pmatrix}]$$

$$+Pr\ [z<Y<z+dz, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}<Y_1<Y<Y_2<\begin{pmatrix} X_3 \\ Y_3 \end{pmatrix}]$$

$$+Pr\ [z<Y<z+dz, \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}<Y<\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}].$$

And the density function $g_W(z)$ may be experssed as follows,

$$g_W(z)=6F(z)f(z)\int_z^\infty f(x)\,(1-F(x-\delta))^4dx$$

$$+24F(z)f(z)\int_z^\infty (F(x-\delta)-F(z-\delta))f(x)\,(1-F(x-\delta))^3dx$$

$$+24f(z)\int_0^z f(x)\,(F(z-\delta)-F(x-\delta))\,dx\int_z^\infty f(x)\,(1-F(x-\delta))^3dx$$

$$+36F(z)f(z)\int_z^\infty (F(x-\delta)-F(z-\delta))^2f(x)\,(1-F(x-\delta))^2dx$$

$$+36F(z)F(z-\delta)^2f(z)\,(1-F(z))\,(1-F(z-\delta))^2$$

$$+12F(z-\delta)^3f(z)\,(1-F(z))^2(1-F(z-\delta))$$

$$+24f(z)\int_0^z f(x-\delta)\,(F(z)-F(x))\,dx\int_z^\infty f(x)\,(1-F(x-\delta))^3dx$$

$$+36F(z)F(z-\delta)^2f(z-\delta)\,(1-F(z))^2(1-F(z-\delta))$$

$$+72f(z-\delta)\int_0^z F(x)f(x-\delta)\,(F(z)-F(x))\,dx(1-F(z))\,(1-F(z-\delta))^2$$

$$+36f(z-\delta)\int_0^z f(x-\delta)\,(F(z)-F(x))^2dx(1-F(z))\,(1-F(z-\delta))^2$$

$$+72f(z-\delta)\int_0^z F(x)^2f(x-\delta)\,dx\int_z^\infty f(x-\delta)\,(1-F(x))\,(1-F(x-\delta))\,dx$$

$$+4F(z-\delta)^3f(z-\delta)\,(1-F(z))^3.$$

In the concrete, assume that $F(x)$ be exponential distribution with mean 1 and variance 1, then the density function $g_M(z)$ and $g_W(z)$ may be expressed as follows,

$$g_M(z)=\begin{cases}-\frac{12}{5}e^{-3z-\delta}(e^z-1)(4e^z-5e^{\delta}), & 0\leq z\leq\delta,\\ \frac{12}{5}e^{-7z+4\delta}(e^z-1)-2e^{-7z+\delta}(e^z-e^{\delta}) & \\ \quad\cdot(-59e^{2\delta}-8e^{2z}+52e^{z+\delta}+69e^{z+2\delta}-36e^{2z+\delta}-18e^{2z+2\delta})\,\varepsilon(z), & \delta\leq z,\end{cases}$$

$$g_W(z)=\begin{cases}-\frac{6}{5}e^{-3z-\delta}(e^z-1)(3e^z-5e^{\delta}), & 0\leq z\leq\delta,\\ \frac{6}{5}e^{-7z+2\delta}(e^z-1)(3e^{2\delta}+10e^{2z}-10e^{z+\delta})+6e^{-7z+2\delta}(-e^z+e^{\delta}) & \\ \quad\cdot(-17e^{\delta}-6e^{2\delta}+20e^z-14e^{2z}+15e^{z+\delta}+12e^{z+2\delta}-10e^{2z+\delta})\,\varepsilon(z), & \delta\leq z,\end{cases}$$

where $\varepsilon(z)=0$ for $z=\delta$ and $\varepsilon(z)=1$ for $z>\delta$.

4. Expected value and MSE

The expected values and mean square errors for the exponential distribution with mean 1, variance 1, $m=3$, $n=2$ and $\alpha=0.1143$ are given as follows,

$$E(\tilde{\xi})=\frac{5}{3}-\frac{81}{70}e^{-3\delta}+\frac{33}{5}e^{-2\delta}-\frac{639}{100}e^{-\delta}-\delta e^{-3\delta}+3\delta e^{-2\delta}\frac{9}{5}\delta e^{-\delta},$$

$$E(\hat{\xi})=\frac{5}{6}-\frac{4294}{3675}e^{-3\delta}+\frac{8637}{2450}e^{-2\delta}-\frac{223}{100}e^{\delta}-\frac{16}{7}\delta e^{-3\delta}+\frac{27}{7}\delta e^{-2\delta}+\frac{3}{5}\delta e^{-\delta},$$

$$\begin{aligned}MSE(\tilde{\xi})=&\frac{19}{9}-(\frac{17393}{11025}-\frac{81\log_e 2}{35}+(\log_e 2)^2)e^{-3\delta}+(\frac{1639}{100}-\frac{66\log_e 2}{5}+3(\log_e 2)^2)e^{-2\delta}\\&-(\frac{16251}{1000}-\frac{639\log_e 2}{50}+3(\log_e 2)^2)e^{-\delta}-\frac{81}{35}\delta e^{-3\delta}\\&+\frac{66}{5}\delta e^{-2\delta}+\frac{81}{50}\delta e^{-\delta}-\delta^2 e^{-3\delta}+3\delta^2 e^{-2\delta}+\frac{9}{5}\delta^2 e^{-\delta}+2\log_e 2\,\delta e^{-3\delta}\\&-6\log_e 2\,\delta e^{-2\delta}-\frac{18\log_e 2}{5}\delta e^{-\delta}-\frac{10\log_e 2}{3}+2(\log_e 2)^2,\end{aligned}$$

$$\begin{aligned}MSE(\hat{\xi})=&\frac{19}{18}-(\frac{392293}{385875}-\frac{8588\log_e 2}{3675}+\frac{16(\log_e 2)^2}{7})e^{-3\delta}+(\frac{575733}{85750}-\frac{8637\log_e 2}{1225}\\&+\frac{27(\log_e 2)^2}{7})e^{-2\delta}-(\frac{17611}{3000}-\frac{223\log_e 2}{50}+\frac{6(\log_e 2)^2}{5})e^{-\delta}\\&-\frac{8588}{3675}\delta e^{-3\delta}+\frac{8637}{1225}\delta e^{-2\delta}+\frac{47}{50}\delta e^{-\delta}-\frac{16}{7}\delta^2 e^{-3\delta}+\frac{27}{7}\delta^2 e^{-2\delta}+\frac{3}{5}\delta^2 e^{-\delta}\\&+\frac{32\log_e 2}{7}\delta e^{-3\delta}-\frac{54\log_e 2}{7}\delta e^{-2\delta}-\frac{6\log_e 2}{5}\delta e^{-\delta}-\frac{5\log_e 2}{3}+(\log_e 2)^2,\end{aligned}$$

Remark

The distributions of never pooling estimator $\hat{\xi}_1$, *i.e.* $\alpha=1$ and always pooling estimator $\hat{\xi}_0$, *i.e.* $\alpha=0$ are expressed as follows,

$$
\begin{aligned}
g_1(z) &= 6F(z)f(z)\,(1-F(z)) \\
g_0(z) &= 12F(z)^2F(z-\delta)f(z)\,(1-F(z-\delta))^3 \\
&\quad +36F(z)F(z-\delta)^2f(z)\,(1-F(z))\,(1-F(z-\delta.))^2 \\
&\quad +12F(z-\delta)^3f(z)\,(1-F(z))^2(1-F(z-\delta)) \\
&\quad +4F(z)^3f(z-\delta)\,(1-F(z-\delta))^3 \\
&\quad +36F(z)^2F(z-\delta)f(z-\delta)\,(1-F(z))\,(1-F(z-\delta))^2 \\
&\quad +36F(z)F(z-\delta)^2f(z-\delta)\,(1-F(z))^2(1-F(z-\delta)) \\
&\quad +4F(z-\delta)^3f(z-\delta)\,(1-F(z))^3
\end{aligned}
$$

Assume that $F(x)$ be exponential distribution with mean 1 and variance 1, then the density function $g_1(z)$ and $g_0(z)$ are given as follows,

$$
\begin{aligned}
g_1(z) &= 6(1-e^z)e^{-2z} \\
g_0(z) &= -140e^{-7z+4\delta}+240e^{-6z+4\delta}+180e^{-6z4\delta}-120e^{-5z+2\delta}-240e^{-5z+3\delta} \\
&\quad -60e^{-5z+4\delta}+16e^{-4z+\delta}+72e^{-4z+2\delta}+48e^{-4z+3\delta}+4e^{-4z+4\delta}
\end{aligned}
$$

Mean square errors are given as follows,

$$
MSE(\hat{\xi}_1)=\frac{19}{18}-\frac{5}{3}\log_e 2+(\log_e 2)^2
$$

$$
\begin{aligned}
MSE(\hat{\xi}_0) &= \frac{1}{8}-\left(\frac{311}{22050}+\frac{2\log_e 2}{105}\right)e^{-3\delta}+\left(\frac{23}{300}+\frac{\log_e 2}{5}\right)e^{-2\delta}+\left(\frac{27}{50}-\frac{6\log_e 2}{5}\right)e^{-\delta}+\frac{\delta}{2} \\
&\quad +\frac{2}{105}\delta e^{-3\delta}-\frac{1}{5}\delta e^{-2\delta}+\frac{6}{5}\delta e^{-\delta}+\delta^2-2(\log_e 2)\,\delta-\frac{\log_e 2}{2}+(\log_e 2)^2
\end{aligned}
$$

Figure 1 shows mean square errors of Median test estimator, Wilcoxon test estimator, never pooling estimator and always pooling estimator.

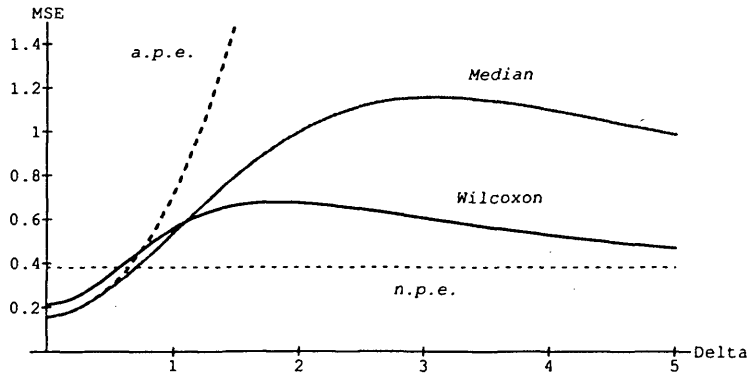

Figure 1 MSE of $\bar{\xi}$, $\hat{\xi}$, $\hat{\xi}_1$ and $\hat{\xi}_0$

5. Optimal Significance Level

Let a minimax regret be a criterion for optimal significance level. The regret R is defined as follows,

$$R=err(\hat{\xi}_p)-\min(err(\hat{\xi}_1), err(\hat{\xi}_0))$$

where *err* is for mean square errors or mean absolute errors, and $\hat{\xi}_p$ is for preliminary test estimatiors, and $\hat{\xi}_1$ is for never pooling estimator, and $\hat{\xi}_0$ is for always pooling estimator. Then the optimal significance level α_{opt} may be determined on the minimax regret.

Since it is too complicated to get exact regret values for each significance level, every numerical evaluation is done by Monte Carlo method. **Table 1** and **2** show the optimal significance levels and minimax regret values at the time.

Table 1. Optimal significance levels based on MSE

	Median test	Wilcoxon two-sample test
	α_{opt} (regret)	α_{opt} (regret)
$m=5, n=5$	0.5000 (0.0041)	0.4206 (0.0064)
$m=5, n=10$	0.5734 (0.0147)	0.2567 (0.0139)
$m=5, n=15$	0.5000 (0.0228)	0.1974 (0.0197)

Table 2. Optimal significance levels besed on MAE

	Median test	Wilcoxon two-sample test
	α_{opt} (regret)	α_{opt} (regret)
$m=5, n=5$	0.5000 (0.0183)	0.2783 (0.0179)
$m=5, n=10$	0.5734 (0.0319)	0.1855 (0.0289)
$m=5, n=15$	0.5000 (0.0451)	0.1528 (0.0360)

6. Conclusion

In this paper, we derived the optimal significance level of the preliminary test for estimating median. At each case, it will be clear that significance levels on MAE are relatively smaller than those on MSE, *i.e.* $\alpha_{opt}(MSE) \geq \alpha_{opt}(MAE)$. As these tests take discrete values for significance level, optimal significance levels of Median test are fairly large. And it is well known that minimax regret criterion detemination is conservative. So if we want to utilize prior information actively, we should use other criterion, for instance use of prior distribution.

References

1) Tamura, R. (1965) Nonparametric inference with a preliminary test, *Bull. Math. Statist.* vol. 11, 39-61.

2) Tamura, R. (1967) Small sample properties of a sometimes pool estimate in the nonparametric case, *Bull. Math.*

Statist. vol. 12, 75-83.

3) Okazaki, T. and Asano, Ch. (1989) On Sometimes Pooling of Data in View of L_1-Measure, *Quality for Progress and Development* 645-653.

4) Okazaki, T. (1989) On Estimation of Median with a Preliminary Test, *Proc. 3rd Japan-China Symposium Statist.* 351-354.