A Simple Confidence Interval for the Median

David J. Olive*

Southern Illinois University

June 3, 2005

Abstract

Large sample confidence intervals often have the form $D_n \pm z_{1-\alpha/2}\ SE(D_n)$ where D_n is an estimator of the parameter and $P(Z \leq z_\alpha) = \alpha$ when Z has a normal N(0,1) distribution. Replacing $z_{1-\alpha/2}$ by $t_{p,1-\alpha/2}$ can be viewed as multiplying $z_{1-\alpha/2}\ SE(D_n)$ by a finite sample correction factor $t_{p,1-\alpha/2}/z_{1-\alpha/2}$ in order to improve the performance of the interval for small sample sizes. This technique is used to modify a large sample confidence interval for the population median. This interval is compared to the intervals based on the sample mean and 25% trimmed mean.

KEY WORDS: Outliers; Trimmed Mean.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This work was supported by the National Science Foundation under grant DMS 0202922.

1 Introduction

The population median MED(Y) is a measure of location and is any value that satisfies

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \quad (1)$$

This population quantity can be estimated from the sample $Y_1, \ldots, Y_n$. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \quad (2)$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n) = \text{MED}(Y_i,\ i = 1, \ldots, n)$ will be useful.

Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$). Then the β *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \quad (3)$$

where $L_n = \lfloor n\beta \rfloor$ and $U_n = n - L_n$.

The trimmed mean is estimating a truncated mean μ_T. Assume that Y has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let y_β be the number satisfying $P(Y \leq y_\beta) = \beta$. Then

$$\mu_T = \frac{1}{1 - 2\beta} \int_{y_\beta}^{y_{1-\beta}} y f_Y(y) dy. \quad (4)$$

Notice that the 25% trimmed mean is estimating

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

Section 2 modifies the Bloch and Gastwirth (1968) confidence interval for the population median. The modified interval is compared to the intervals based on the sample mean and 25% trimmed mean.

2 A Simple Confidence Interval for MED(Y).

The large sample theory of the β trimmed mean T_n has been examined by Bickel (1965), Stigler (1973), Tukey and McLaughlin (1963), Yuen (1974), and Shorack and Wellner (1986, pp. 680-683). First, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{[(U_n - L_n)/n]^2}. \tag{5}$$

Let $p = U_n - L_n - 1$. Then the standard error (SE) of T_n is $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ and a large sample 100 $(1 - \alpha)$% confidence interval (CI) for μ_T is

$$T_n \pm t_{p,1-\alpha/2}\, SE(T_n) \tag{6}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom.

The 100 $(1 - \alpha)$% confidence interval for the population mean μ is

$$\overline{Y} \pm t_{p,1-\alpha/2}\, S/\sqrt{n}$$

where $p = n - 1$ and S is the sample standard deviation. Notice this interval can be found using (6) with $L_n = 0$ and $U_n = n$.

Several confidence intervals for the population median have been proposed. Price and Bonnett (2001), McKean and Schrader (1984) and Bloch and Gastwirth (1968) are useful references for estimating the SE of the sample median.

The following confidence interval provides considerable resistance to gross outliers while being very simple to compute. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$). Let $U_n = n - L_n$ where $L_n = [n/2] - \lceil \sqrt{n/4}\ \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$ ($\approx \lceil\ \sqrt{n}\ \ \rceil$). Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2}\ SE(\text{MED}(n)).$$

This SE is due to Bloch and Gastwirth (1968), but the degrees of freedom p is motivated by the confidence interval for the trimmed mean.

For large samples, the CIs based on the trimmed mean, mean and median could be written as $D_n \pm z_{1-\alpha/2}\ SE(D_n)$ where $P(Z \leq z_\alpha) = \alpha$ when Z has a normal N(0,1) distribution. Notice that

$$D_n \pm t_{p,1-\alpha/2}\ SE(D_n) = D_n \pm \frac{t_{p,1-\alpha/2}}{z_{1-\alpha/2}} z_{1-\alpha/2}\ SE(D_n),$$

and the term

$$a_n = \frac{t_{p,1-\alpha/2}}{z_{1-\alpha/2}} \to 1$$

as $n \to \infty$. We can regard a_n as a finite sample correction factor that makes the coverage of the CI more accurate for small samples.

Example 1. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height is $\overline{Y} = 1598.862$ and the classical 95% CI is (1514.206, 1683.518). MED(n) = 1693.0 and the resistant 95% CI based on the median is (1678.517, 1707.483). The 25% trimmed mean $T_n = 1689.689$ with 95% CI (1672.096, 1707.282).

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then $\overline{Y} = 1692.356$ and the classical 95% CI is (1678.595, 1706.118). Now MED(n) = 1694.0 and the 95% CI based on the median is (1678.403, 1709.597). The 25% trimmed mean $T_n = 1693.200$ with 95% CI (1676.259, 1710.141). Notice that when the outliers are corrected, the three intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm while the outliers greatly increased the length of the classical t–interval.

Table 1 presents the results from a small simulation study. In order for a location estimator to be used for inference, there must exist a useful SE and a useful cutoff value t_p where the degrees of freedom p is a function of n. Two criteria will be used to evaluate the CI's. First, the observed coverage is the proportion of the $K = 500$ runs for which the CI contained the parameter μ_D estimated by D_n. This proportion should be near the nominal coverage 0.95. Notice that if W is the proportion of runs where the CI contains the parameter, then KW is a binomial random variable. Hence the SE of W is $\sqrt{\hat{p}(1-\hat{p})/K} \approx 0.013$ for the observed proportion $\hat{p} \in [0.9, 0.95]$, and an observed coverage between 0.92 and 0.98 suggests that the observed coverage is close to

the nominal coverage of 0.95.

The second criterion is the scaled length of the CI $= \sqrt{n}$ CI length $=$

$$\sqrt{n}(2)(t_{p,0.975})(SE(D_n)) \approx 2(1.96)(\sigma_D)$$

where the approximation holds if $p > 30$, if $\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2)$, and if $SE(D_n)$ is a good estimator of $\sigma_D/\sqrt{n}$ for the given value of n.

Table 1 can be used to examine the three different interval estimators. A good estimator should have an observed coverage $\hat{p} \in [.92, .98]$, and a small scaled length. In Table 1, coverages were good for normal $N(0,1)$ data, except the median interval where $SE(\text{MED}(n))$ is slightly too small for $n \approx 100$. The coverages for the Cauchy C(0,1) and double exponential DE(0,1) data were all good even for $n = 10$. The exponential EXP(1) distribution is skewed, so the central limit theorem is not a good approximation for $n = 10$. For this skewed distribution, the estimators $\overline{Y}$, MED(n) and the 25% trimmed mean are estimating the population parameters 1, log(2) and 0.73838 respectively.

Examining Table 1 for N(0,1) data shows that the median interval and 25% trimmed mean interval are noticeably larger than the classical interval. Since the degrees of freedom $p \approx \sqrt{n}$ for the median interval, $t_{p,0.975}$ is considerably larger than $1.96 = z_{0.975}$ for $n \leq 100$. The rows labeled ∞ give the scaled length $2(1.96)(\sigma_D)$ expected if $\sqrt{n}SE$ is a good estimator of σ_D.

The intervals for the C(0,1) and DE(0,1) data behave about as expected. The classical interval is very long at C(0,1) data since the first moment of C(0,1) data does not exist. Notice that the two resistant intervals are shorter than the classical intervals for DE(0,1) data.

Table 1: Simulated 95% CI Coverages and lengths, 500 Runs

F	n	$\overline{Y}$	MED	25% TM	$\overline{Y}$len	MEDlen	25% TMlen
N(0,1)	10	0.960	0.948	0.938	4.467	7.803	5.156
N(0,1)	50	0.948	0.936	0.926	4.0135	5.891	4.419
N(0,1)	100	0.932	0.900	0.938	3.957	5.075	4.351
N(0,1)	1000	0.942	0.940	0.936	3.930	5.035	4.290
N(0,1)	∞	0.95	0.95	0.95	3.920	4.913	4.285
DE(0,1)	10	0.966	0.970	0.968	6.064	7.942	5.742
DE(0,1)	50	0.948	0.958	0.954	5.591	5.360	4.594
DE(0,1)	100	0.956	0.940	0.938	5.587	4.336	4.404
DE(0,1)	1000	0.948	0.936	0.944	5.536	4.109	4.348
DE(0,1)	∞	0.95	0.95	0.95	5.544	3.920	4.343
C(0,1)	10	0.974	0.980	0.962	54.590	12.682	9.858
C(0,1)	50	0.984	0.960	0.966	94.926	7.734	6.794
C(0,1)	100	0.970	0.940	0.968	243.4	6.542	6.486
C(0,1)	1000	0.978	0.952	0.950	515.9	6.243	6.276
C(0,1)	∞	0.95	0.95	0.95	∞	6.157	6.255
EXP(1)	10	0.892	0.948	0.916	4.084	6.012	3.949
EXP(1)	50	0.938	0.940	0.950	3.984	4.790	3.622
EXP(1)	100	0.938	0.930	0.954	3.924	4.168	3.571
EXP(1)	1000	0.952	0.926	0.936	3.914	3.989	3.517
EXP(1)	∞	0.95	0.95	0.95	3.92	3.92	3.51

For EXP(1) data, $2(1.96)(\sigma_D) = 3.9199$ for $\overline{Y}$ and MED(n) while $2(1.96)(\sigma_D) \approx 3.51$ for the 25% trimmed mean. The 25% trimmed mean may be the best of the three intervals for these four distributions since the scaled length was small with good coverage.

The median interval was chosen so that $L_n \approx n/2$ outliers are needed to drive $SE(\text{MED}(n))$ to ∞. (This resistant interval is not a high breakdown method since about $\sqrt{n}$ maliciously placed outliers can drive $SE(\text{MED}(n))$ to zero.) Since the two resistant intervals are easy to compute, they can be included with computer output along with the warning to examine the data for outliers if the classical and resistant intervals differ greatly. This application is useful since statistical consulting clients all too often obtain their software output without plotting the data.

The median interval is useful for symmetric distributions and for one parameter families such as the exponential, power, and truncated extreme value distributions. See Patel, Kapadia and Owen (1976). The median interval may not need to be adjusted if there are censored observations present. Suppose that $Y_{(R+1)}, ..., Y_{(n)}$ have been right censored (similar results hold for left censored data). Then create a pseudo sample $Z_{(i)} = Y_{(R)}$ for $i > R$ and $Z_{(i)} = Y_{(i)}$ for $i \leq R$. Then compute the median interval based on $Z_1, ..., Z_n$. This CI will be identical to the CI based on $Y_1, ..., Y_n$ (no censoring) if $R + 1 > U_n$.

References

Bickel, P.J. (1965). On some robust estimates of location. *The Annals of Mathematical Statistics,* 36, 847-858.

Bloch, D.A. and Gastwirth, J.L. (1968). On a simple estimate of the reciprocal of the density function. *The Annals of Mathematical Statistics,* 39, 1083-1085.

Buxton, L.H.D. (1920). The anthropology of Cyprus. *The Journal of the Royal*

Anthropological Institute of Great Britain and Ireland, 50, 183-235.

McKean, J.W. and Schrader, R.M. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics: Simulation and Computation,* 13, 751-773.

Patel, J.K., Kapadia C.H., and Owen, D.B. (1976), *Handbook of Statistical Distributions,* Marcel Dekker, NY.

Price, R.M. and Bonett, D.G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation,* 68, 295-305.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes With Applications to Statistics*. Wiley, New York.

Stigler, S.M. (1973). The asymptotic distribution of the trimmed mean. *The Annals of Mathematical Statistics,* 1, 472-477.

Tukey, J.W. and McLaughlin, D.H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization 1. *Sankhya A,* 25, 331-352.

Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika,* 61, 165-170.