Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 973~980

An Improved Method for Constructing Confidence Interval of Median : Small Sample Case

Sang-Gue Park[1] · Jiyun Choi[2]

Abstract

Phase I clinical trials are often pharmacologically oriented and usually attempt to find the best dose of drug to employ. However, other purposes like determination of sizes and types of side effects and toxicity and organ system involved are equally important. Estimation of treatment effects or side effects is usually ignored since it is usually based on too small sample, even though Phase II clinical trials would be designed based on the Phase I studies. Statistical methods for constructing the approximate confidence interval for population median in case of small sample are considered and an improved method is proposed. The proposed estimator is compared with current methods through simulation studies.

Keywords : Interpolated order statistics, Population median, Sign test, Wilcoxon signed rank test

1. Introduction

Phase I clinical trials are the first studies in which a new drug is administered to human subjects. The primary purpose of phase I studies of new drugs is to establish a safe dose and schedule of administration. Other purposes are to determine the types or sizes of side effects and toxicity and organ systems involved, to assess evidence for efficacy, and to investigate basic clinical pharmacology of the drug. Not all of these goals can be met completely in any phase I trial, in part because the number of patients treated is small. However, well-conducted phase I studies can achieve substantial progress toward each of

1) First Author : Professor, Department of Statistics, Chung-Ang University, Seoul, 156-756, Korea
E-mail : spark@cau.ac.kr
2) Ph.D Candidate, Department of Statistics, Chung-Ang University, Seoul, 156-756, Korea
E-mail : jiyun1023@hotmail.com

these goals.

Some randomized phase I studies often report the confidence intervals for the population median of treatment effect under the given administered doses. It is said that researchers do not count it on much because of the small sample size, but they seems to keep it up to phase III study. It does actually provide some intuition about the treatment effect and design for its phase II study and some statistics produced in its phase I study also used for validation for its phase II study.

This paper aims for developing an approximate confidence interval for the population median in case of small sample, which is less than or equal to 10. In section 2, we present a statistical method of constructing the confidence interval for median given by Hettmansperger and Sheather(1986), which seems to be useful in small sample case. In section 3, we propose the extension of Hettmansperger and Sheather(1986)'s interpolated method based on Hodges and Lehmann estimator. In section 4, we discuss the statistical method based on normal approximation given by Sheather(1986). We compare the empirical coverage probabilities of the discussed methods through simulation study.

2. Confidence intervals based on sign test

Suppose $X_1, \cdots, X_n$ is a random sample of size n from a distribution with absolute continuous distribution function $F(x-\theta)$ and density $f(x-\theta)$. Further, suppose $F(0) = 0.5$, uniquely, so that θ is the unique median; no shape assumption is imposed on F. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the order statistics. Then the interval $[X_{(d)}, X_{(n-d+1)}]$ is a simple distribution-free confidence interval for θ. The confidence coefficients $\gamma = 1-2P(S < d)$ where S has a binomial distribution with parameters n and 0.5; that is, $P(X_{(1)} \leq \theta \leq X_{(5)}) = 0.938$ from the binomial table with $n = 5$ and $p = 0.5$. If S denotes the sign test statistic for testing $H_0: \theta = 0$ versus $H_0: \theta \neq 0$, then the interval corresponds to inverting the acceptance region of a size $\alpha = 2P(S < d)$ test(See Hettmansperger, 1984).

This confidence interval is quite versatile since it makes no shape assumption on the underlying distribution, is easy to compute, and requires only a binomial table to establish the confidence coefficient.

Because of the discreteness of the binomial distribution, for small to moderate sample sizes the available set of possible confidence coefficients is rather sparse. Sometimes we need to construct a 95% confidence interval rather than 93.8% confidence interval because we want to compare it other 95% confidence intervals from different sources.

Hettmansperger and Sheather(1986) consider the problem of interpolating adjacent order statistics to form confidence intervals with intermediate values of the confidence coefficients. The interpolated intervals are no longer distribution-free in general; however, they showed that the confidence coefficient depends only slightly on the underlying F for a broad collection of distributions.

Suppose $[X_{(d)}, X_{(n-d+1)}]$ and $[X_{(d+1)}, X_{(n-d)}]$ are $\gamma_d = 1-\alpha_d$ and $\gamma_{d+1} = 1-\alpha_{d+1}$ confidence interval for θ, respectively. From the binomial distribution, we have,

$$\frac{\alpha_{d+1}}{2} = \frac{\alpha_d}{2} + \binom{n}{d}\left(\frac{1}{2}\right)^n$$

or

$$\gamma_{d+1} = \gamma_d - 2\binom{n}{d}\left(\frac{1}{2}\right)^n.$$

This links the successive intervals based on d and $d+1$.

Define, for $0 \le \lambda < 1$,

$$\hat{\theta}_{HS,L} = (1-\lambda)X_{(d)} + \lambda X_{(d+1)},$$
$$\hat{\theta}_{HS,U} = (1-\lambda)X_{(n-d+1)} - \lambda X_{(n-d)}$$

and let γ be the confidence coefficient for $[\hat{\theta}_{HS,L}, \hat{\theta}_{HS,U}]$. Then $\gamma_{d+1} \le \gamma < \gamma_d$ and

$$\begin{aligned}\gamma &= P(\hat{\theta}_{HS,L} \le \theta \le \hat{\theta}_{HS,U}) \\ &= 1 - P(\theta < \hat{\theta}_{HS,L}) - P(\theta > \hat{\theta}_{HS,U}) = 1-\alpha_L-\alpha_U\end{aligned}.$$

Hettmansperger and Sheather established the connection between λ and γ. Given γ, we present a simple interpolation formula for finding λ.

Theorem 2.1 (Hettmansperger and Sheather(1986)) Let $\beta = \lambda/(1-\lambda)$. Then

$$\alpha_L = \frac{\alpha_{d+1}}{2} - (n-d)\binom{n}{d}\int_0^\infty F(-\beta y)^d[1-F(y)]^{n-d-1}dF(y),$$
$$\alpha_U = \frac{\alpha_{d+1}}{2} - (n-d)\binom{n}{d}\int_{-\infty}^\theta F(1-\beta y)^d[F(y)]^{n-d-1}dF(y).$$

Theorem 2.2 (Hettmansperger and Sheather(1986)) Suppose f is symmetric about 0. Then

(i) $\alpha_L = \alpha_U$;
(ii) if $\lambda = 0.5$ it follows that

$$\alpha_L = \frac{\alpha_{d+1}}{2} - \frac{n-d}{n}\binom{n}{d}\left(\frac{1}{2}\right)^n = \frac{\alpha_d}{2} + \frac{d}{n}\binom{n}{d}\left(\frac{1}{2}\right)^n.$$

Theorem 2.3 (Hettmansperger and Sheather(1986)) Let $\beta = \lambda/(1-\lambda)$. Suppose f is symmetric about 0. Then

(i) $I(\lambda) = 1-(n-d)2^n\int_0^\infty [F(-\beta y)]^d[1-F(y)]^{n-d-1}dF(y)$, with $I(0) = 0$, $I(0.5) = d/n$ and $I(\lambda)\to 1$ as $\lambda \to 1$.

(ii) If F is sufficiently regular so that differentiation can be carried out under the integral and if $f(x) > 0$ for $x \le 0$, then $I(\lambda)$ is a continuous and strictly increasing convex function of λ.

Hettmansperger and Sheather suggested that linear interpolation is inappropriate from these theorems and recommended the following interpolation formula by assuming that f is symmetric about θ

$$\lambda = \frac{(n-d)I}{d+(n-2d)I}, \tag{1}$$

where the interpolation factor I is

$$I = \frac{\gamma_d - \gamma}{\gamma_d - \gamma_{d+1}}. \tag{2}$$

Hettmansperger and Sheather showed that the confidence interval with (1) and (2) has exact γ confidence coefficient when f is doubly exponential distributed.

3. Confidence intervals based on signed rank test

Hettmansperger and Sheather's idea is somewhat fascinating. However, it does still get affected and limited since the values of γ_d and γ_{d+1} are finite in small sample size. One might consider improving it by extending the finiteness. Hodges

and Lehmann(1963) proposed the median estimator by using Walsh averages when f is symmetric about θ. Given a random sample $X_1, \cdots, X_n$ the $n(n+1)/2$ Walsh averages are defined by $W_{ij} = \frac{X_i + X_j}{2}$, $i \leq j$. Then from the signed rank statistics W^+ and the probability table of signed rank test $w^+(n, \alpha/2)$ we obtain c such that

$$c = \frac{n(n+1)}{2} + 1 - w^+(n, \alpha/2),$$
$$P(c \leq W^+ \leq \frac{n(n+1)}{2} - c) = 1 - \alpha.$$

Then the confidence interval for θ with the confidence coefficient $1-\alpha$ is

$$[W_{(c)}, W_{(n_w - c+1)}],$$

where $n_w = \frac{n(n+1)}{2}$.

Suppose $[W_{(c)}, W_{(n-c+1)}]$ and $[W_{(c+1)}, W_{(n-c)}]$ are $\gamma_c = 1 - \alpha_c$ and $\gamma_{c+1} = 1 - \alpha_{c+1}$ confidence interval for θ, respectively. Then we could improve Hettmansperger and Sheather's interpolation method.

For $0 \leq \lambda < 1$,

$$\hat{\theta}_{HL,\,L} = (1-\lambda) W_{(c)} + \lambda W_{(c+1)},$$
$$\hat{\theta}_{HL,\,U} = (1-\lambda) W_{(n-c+1)} - \lambda W_{(n-c)}$$

and let γ be the confidence coefficient for $[\hat{\theta}_{HL,\,L}, \hat{\theta}_{HL,\,U}]$. Then $\gamma_{c+1} \leq \gamma < \gamma_c$ and

$$\gamma = P(\hat{\theta}_{HL,\,L} \leq \theta \leq \hat{\theta}_{HL,\,U}),$$

with $\lambda = \frac{(n_w - c)I}{c + (n_w - 2c)I}$ and the interpolation factor $I = \frac{\gamma_c - \gamma}{\gamma_c - \gamma_{c+1}}$.

4. Confidence intervals from Sheather

Sheather(1986) considered the problem of estimating the variance of sample median $\hat{\theta}$ in small sample. Maritz and Jarrett(1978) have shown that the asymptotic formula for the variance of $\hat{\theta}$ (that is, $1/[4nf^2(\theta)]$) provides a poor approximation to the variance of $\hat{\theta}$ when n is small. Also, Brown and Wolfe(1983) have shown that the asymptotic formula can have large coefficient of variation even when the form of the density is known and only θ has to be estimated. Sheather proposed a direct method for estimating the small sample variance of $\hat{\theta}$.

For the case of sample size $n = 2m+1$

$$\widehat{var}(\hat{\theta}) = \sum_{i=1}^{n} v_i X_{(i)}^2 - \left\{ \sum_{i=1}^{n} v_i X_{(i)} \right\}^2,$$

where $v_i = \dfrac{J_1\left\{\dfrac{i-1/2}{n}\right\}}{\sum_{j=1}^{n} J_1\left(\dfrac{j-1/2}{n}\right)}$ and $J_1(y) = \dfrac{n!}{m!}\, y^m (1-y)^m$.

For the case of sample size $n = 2m$

$$\widehat{var}(\hat{\theta}) = \frac{(D_2 + E_{12})}{2} - D_1^2,$$

where $D_1 = \sum_{i=1}^{n} u_i X_{(i)}$, $D_2 = \sum_{i=1}^{n} u_i X_{(i)}^2$, $u_i = \dfrac{J_2\left\{\dfrac{i-1/2}{n}\right\}}{\sum_{j=1}^{n} J_2\left(\dfrac{j-1/2}{n}\right)}$,

$J_2(y) = \dfrac{n!}{2m!(m-1)!}\, y^{m-1}(1-y)^{m-1}$, $E_{12} = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} X_{(i)} X_{(j)}$, $(i \le j)$,

$u_{ii} = u_i - \dfrac{n!}{2m!^2}\left\{ \left(\dfrac{i}{n}\right)^m \left(\dfrac{i-i}{n}\right)^m + \left(\dfrac{i-1}{n}\right)^m \left(\dfrac{n+1-i}{n}\right)^m - 2\left(\dfrac{i-1}{n}\right)\left(\dfrac{n-i}{n}\right)^m \right\}$,

$u_{ij} = u_i - \dfrac{n!}{m!^2}\left\{ \left(\dfrac{i}{n}\right)^m - \left(\dfrac{n-i}{n}\right)^m \right\}\left\{ \left(\dfrac{n+1-i}{n}\right)^m - \left(\dfrac{n-i}{n}\right)^m \right\}$ $(i < j)$.

Table 1 shows v_i and u_i values of various sample sizes for the application.

[Table 1] Weights values for various sample sizes

i \ n	3	4	5	6	7	8	9
1	0.26316	0.15909	0.04853	0.02916	0.00584	0.00352	0.00145
2	0.47368	0.34091	0.26423	0.17566	0.09554	0.06191	0.02892
3	0.26316	0.34091	0.37448	0.29518	0.24225	0.17363	0.11447
4		0.15909	0.26423	0.29518	0.31256	0.26095	0.22066
5			0.04853	0.17566	0.24225	0.26095	0.26899
6				0.02916	0.09554	0.17363	0.22066
7					0.00584	0.06191	0.11447
8						0.00352	0.02892
9							0.00145

The confidence interval for the population median with confidence coefficient γ is

$$\hat{\theta}_{SL} = \hat{\theta} - z_{\alpha/2}\sqrt{\widehat{var}(\hat{\theta})},$$
$$\hat{\theta}_{SU} = \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{var}(\hat{\theta})}.$$

5. Simulation and conclusion

We design the simulation to compare the performance of 3 estimators by calculating the empirical coverage probabilities. We consider 3 well-known probability distributions such as standard normal, standard double exponential and standard Cauchy distribution with sample sizes 5, 7 and 9 and calculate the 90% or 95% confidence intervals. For the case of sample size of 5, interpolation methods can not be applied to 95% confidence interval calculation so that 90% confidence intervals would be obtained instead. The results are based on 10,000 repetitions and run by SAS IML program.

[Table 2] Empirical coverage probabilities for 90% and 95% confidence intervals

Methods	n=5(90%)			n=7(95%)			n=9(95%)		
	Normal	D-Exp	Cauchy	Normal	D-Exp	Cauchy	Normal	D-Exp	Cauchy
1	.9293	.9183	.9227	.9631	.9431	.9602	.9523	.9427	9475
2	.9032	.8855	.8993	.9490	.9458	.9438	.9522	.9480	.9503
3	.8764	.9371	.9480	.9186	.8720	.9765	.9275	.8688	.9760

Method 1 stands for Hettmansperger and Sheather(1986)'s interpolation one,

method 2 for interpolation one based on Hodges and Lehmann estimator and method 3 for Sheather one. We can observe some interesting points. Firstly, the proposed method maintains the given nominal level the best regardless the sample sizes and the probability distributions. Secondly, Hettmansperger and Sheather method seems to maintain the nominal level reasonably well but it overestimates the level when the sample sizes are 5 and 7. Thirdly, Sheather(1986)'s estimator heavily depend on the population distributions and it does not maintain the nominal level good enough.

We did not examine the case of asymmetry probability distributions, because the response usually assumed symmetric probability distribution like normal in phase I study. However, the proposed method is expected to maintain the nominal level reasonably well in practical cases since Hettmansperger and Sheather(1986) reported that their method does not matter much even in the case of the mild asymmetry.

Phase I study usually does not require statistical analysis since the number of patients treated is small. However, some well-intended phase I studies need to produce some meaningful numbers or statistics for the phase II study. The current study could be useful in such cases and some refined statistical methodologies for the various phase I studies would be needed to research further, especially which could be used in small samples.

References

1. Brown, M.B. and Wolfe, R.A.(1983). Estimation of the variance of percentile estimates. *Computational Statistical Data Analysis* 1, 167-174.
2. Hettmansperger, T.P.(1984). *Statistical Inference Based on Ranks*. John Wiley, New York.
3. Hettmansperger, T.P. and Sheather, S.J.(1986). Confidence intervals based on interpolated order statistics. *Statistics and Probability Letters* 4, 75-79.
4. Hodges, J.L. and Lehmann, E.L.(1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34, 598-611.
5. Maritz, J.S. and Jarrett, R.G.(1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association* 73, 409-411.
6. Sheather, S.J.(1986). A finite estimate of the variance of the sample median. *Statistics and Probability Letters* 4, 337-342.

[received date : Sep. 2004, accepted date : Nov. 2004]