应用概率统计 第二十六卷
第三期 2010年6月

Chinese Journal of Applied Probability
and Statistics Vol.26 No.3 Jun. 2010

Optimal Ranked Set Sampling Design for the Sign Test

Zhang Liangyong Dong Xiaofang

(*College of Sciences, Qiqihar University, Qiqihar, 161006*)

Abstract

The sign test based on ranked set sampling is proposed for testing hypotheses concerning the quantiles of a population characteristic. Both balance and selective designs are considered and the relative performance of different designs is assessed in terms of Pitman's asymptotic relative efficiency. For each quantile, the sampling allocation that maximizes the efficacy of sign statistic is identified and shown to not depend on the population distribution.

Keywords: Nonparametric test, ranked set sampling, optimal design, efficacy.

AMS Subject Classification: 62G10, 62G30.

§1. Introduction

Ranked set sampling (RSS) is a sampling protocol to improve the cost efficiency of an experiment. It is appropriate for situations in which quantification of sampling units is costly or difficult but ranking of the units in a small set is easy and inexpensive.

For an integer $t \in \{1, \cdots, m\}$, let $D = \{d_1, \cdots, d_t\}$ be a set of integers that contains the ranks of the observation to be quantified in each cycle. For example, if $D = \{1, 3, 5\}$ and the set size is m, we would quantify two extreme and one middle observations in each cycle. Note that, for $t \in \{1, \cdots, m\}$, $D = \{d_1, \cdots, d_t\}$ is arbitrary to allow all possible designs. The set D will be called a selective design. The design is said to balance (or the standard ranked set sampling) design when $D = \{1, 2, \cdots, m\}$. To collect data with design $D = \{d_1, \cdots, d_t\}$, we draw mt units from an infinite population. These units are partitioned into t sets each having m units. Each set is judgment ranked without actually measuring the units. The observation with judgment rank d_1 is quantified from the first set, the observation with judgment rank d_2 is quantified from the second set and so forth until the observation with judgment rank d_t is quantified from the final set. The process is repeated n times (called cycles) to have nt quantified observations. Thus, our ranked set sample with design $D = \{d_1, \cdots, d_t\}$ is $X_{(d_i)j}$, $i = 1, \cdots, t$ and $j = 1, \cdots, n$.

Received April 7, 2006. Revised December 18, 2009.

Although RSS was first introduced in the context of estimating the population mean (see Takahasi and Wakimoto (1968)), it is also very natural for nonparametric methods where order statistics plays a fundamental role. The sign test for median, one of the fundamental methods in nonparametric, has been studied for analyzing RSS data by a few authors (See Bohn and Wolfe (1992), Hettmansperger (1995), Ozturk and Wolfe (2000), Wang and Zhu (2005)). This paper extends these results to the RSS sign test for population quantiles and identifies the optimal design which maximizes the efficacy of sign statistic, we find that optimal sampling depends on the quantile but not on the parent population. Quantile testing has frequent application in environmental assessment. For instance, testing may make safety decisions at a hazardous waste site if an upper quantile (say the 95th percentile) is less than safe level prescribed by regulation. In Section 2, we discuss the RSS sign test for quantiles under selective designs. In Section 3, we compare the performance of the sign test under selective design with that under balance design. Section 4 gives the optimal design for the sign test.

§2. Testing Quantiles under Selective Designs

Let ξ_p be the pth quantile of infinite population having cumulative distribution function (cdf) $G(x)$ and probability density function (pdf) $g(x)$. Thus, we have $G(\xi_p)=p$. The null hypothesis asserts that ξ_p is equal to some known constant, but, without loss of generality, we may take that constant to equal zero. Thus, we want to test $H_0:\xi_p=0$ against either a one-sided or a two-sided alternative. Let us write $G(x)=F(x-\xi_p)$ where $F(t)$ is a distribution function whose pth quantile is zero. For $p=0.5$, this scenario reduces to a test of the median.

Let $X_{(d_i)j}$, $i=1,\cdots,t$ and $j=1,\cdots,n$ be a ranked set sample with design $D=\{d_1,\cdots,d_t\}$. We assume perfect ranking, so that $X_{(d_i)j}$ is the jth observation on d_ith order statistic of G. The distribution of $X_{(d_i)j}$, which depends on the rank order d_i but not on j, has its pdf and cdf given by (see Mao et al (1998), Fang et al (2006))

$$g_{(d_i)}(x)=\frac{1}{B(1;d_i,m+1-d_i)}[G(x)]^{d_i-1}[1-G(x)]^{m-d_i}g(x), \tag{2.1}$$

$$G_{(d_i)}(x)=\int_{-\infty}^{x}g_{(d_i)}(t)\mathrm{d}t=\frac{B(G(x);d_i,m+1-d_i)}{B(1;d_i,m+1-d_i)}, \tag{2.2}$$

where $B(u;a,b)=\displaystyle\int_0^u y^{a-1}(1-y)^{b-1}\mathrm{d}y$.

Note that $X_{(d_i)j}$, for $j=1,2,\cdots,n$, are independent of each other because they are selected from independent ranked set. The sign test statistic based on data from design

D is

$$S_D^+ = \sum_{i=1}^{t}\sum_{j=1}^{n} I(X_{(d_i)j} > 0) = \sum_{i=1}^{t} \eta_i,$$

where $\eta_i = \sum_{j=1}^{n} I(X_{(d_i)j} > 0) \sim \text{binomial}(n, 1 - G_{(d_i)}(0))$.

The mean and variance of S_D^+ are

$$\mathsf{E}(S_D^+) = \sum_{i=1}^{t} n[1 - G_{(d_i)}(0)], \tag{2.3}$$

$$\mathsf{Var}\,(S_D^+) = \sum_{i=1}^{t} nG_{(d_i)}(0)[1 - G_{(d_i)}(0)] = \frac{n}{4}\delta_D^2, \tag{2.4}$$

where $\delta_D^2 = t - 4\sum_{i=1}^{t}[G_{(d_i)}(0) - 0.5]^2 = \sum_{i=1}^{t} D_{d_ig}$ with $D_{d_ig} = 4G_{(d_i)}(0)[1 - G_{(d_i)}(0)]$.

Under the null hypothesis, for simplicity we write D_{d_ig} as D_{d_i}, thus

$$D_{d_i} = 4F_{(d_i)}(0)[1 - F_{(d_i)}(0)]. \tag{2.5}$$

As a special case, when $D = \{1, 2, \cdots, m\}$ the test statistic under balance design will be denoted by

$$S_{\text{RSS}}^+ = \sum_{i=1}^{m}\sum_{j=1}^{n} I(X_{(i)j} > 0),$$

having the mean and variance as

$$\mathsf{E}(S_{\text{RSS}}^+) = nm(1 - G(0)), \qquad \mathsf{Var}\,(S_{\text{RSS}}^+) = \frac{n}{4}\delta_R^2,$$

where $\delta_R^2 = m - 4\sum_{i=1}^{m}[G_{(i)}(0) - 0.5]^2$.

Under the null hypothesis, $\xi_p = 0$ and $G(0) = F(0) = p$. Thus from (2.1) and (2.2) we have

$$g_{(i)}(0) = f_{(i)}(0) = \frac{1}{B(1; i, m+1-i)} p^{i-1}(1-p)^{m-i} f(0) = C_i f(0), \tag{2.6}$$

$$G_{(i)}(0) = F_{(i)}(0) = \frac{B(p; i, m+1-i)}{B(1; i, m+1-i)}, \tag{2.7}$$

where

$$C_i = \frac{p^{i-1}(1-p)^{m-i}}{B(1; i, m+1-i)}. \tag{2.8}$$

Since S_D^+ is a sum of binomial random variables with different parameters, its exact distribution does not have a simple closed from, except in very special case. However, the null distribution does not require the knowledge of the population distribution, implying that the test is distribution-free. The following theorem establishes the asymptotic normality of S_D^+.

Theorem 2.1 For fixed m and $n \to +\infty$, the test statistic has an asymptotic normal distribution

$$n^{-1/2}[S_D^+ - \mathsf{E}(S_D^+)] \to N(0, \delta_D^2/4).$$

Proof The result follows immediately by using the independence of η_i for $i = 1, 2, \cdots, t$ and the central limit theorem. □

The above theorem allows one to carry out an asymptotically size test. With a two-sided alternative, for example, one would reject $H_0 : \xi_p = 0$ in favor of $H_0 : \xi_p \neq 0$ when $|S_D^+ - \mathsf{E}_{H_0}(S_D^+)| > (1/2) \cdot \sqrt{n} U_{\alpha/2}$ where $U_{\alpha/2}$ is the $(1-\alpha/2)$th upper percentile of the standard normal distribution.

§3. Relative Efficient

In this section, we compare the performance of the test under selective design with that under balance design using the criterion of Pitman's asymptotic relative efficiency (ARE).

Pitman's ARE of S_D^+ versus $S_{\rm RSS}^+$ is defined as

$$\mathrm{ARE}(S_D^+, S_{\rm RSS}^+) = \frac{\mathrm{eff}\,(S_D^+)}{\mathrm{eff}\,(S_{\rm RSS}^+)}, \tag{3.1}$$

the efficacy $\mathrm{eff}\,(T)$ of a test statistic T is given by

$$\mathrm{eff}\,(T) = \lim_{n\to+\infty} \frac{[u'(T)]^2}{nt\mathsf{Var}\,(T)}\Big|_{H_0}, \tag{3.2}$$

where $u'(T) = \partial \mathsf{E}(T)/\partial \xi_p$.

By using (2.3), (2.4) and noting that $G(0) = F(-\xi_p)$ and $g(0) = f(-\xi_p)$, the efficacy based on data from selective and balance are obtained as

$$\mathrm{eff}\,(S_D^+) = \frac{4f^2(0)\Big(\sum\limits_{i=1}^{t} C_{d_i}\Big)^2}{t\delta_D^2\,|_{H_0}}, \tag{3.3}$$

$$\mathrm{eff}\,(S_{\rm RSS}^+) = \frac{4mf^2(0)}{\delta_R^2|_{H_0}}. \tag{3.4}$$

Similarly, the efficacy of the simple random sample (SRS) sign test is

$$\mathrm{eff}\,(S_{\rm SRS}^+) = \frac{f^2(0)}{p(1-p)}. \tag{3.5}$$

On substituting (3.3) and (3.4) in (3.1), further simplification yields

$$\text{ARE}(S_D^+, S_{\text{RSS}}^+) = \frac{\left(\sum\limits_{i=1}^{t} C_{d_i}\right)^2}{mt\sum\limits_{i=1}^{t} D_{d_i}}\delta_R^2\Big|_{H_0}.$$

Comparing balance design with simple random sampling, we see that

$$\text{ARE}(S_{\text{RSS}}^+, S_{\text{SRS}}^+) = \frac{\text{eff}\,(S_{\text{RSS}}^+)}{\text{eff}\,(S_{\text{SRS}}^+)} = m\delta_R^{-2}|_{H_0} \geq 1.$$

Table 1 The asymptotic relative efficiencies of S_{RSS}^+ versus S_{SRS}^+

	$p=0.05$	$p=0.10$	$p=0.20$	$p=0.50$	$p=0.80$	$p=0.90$	$p=0.95$
$m=2$	1.05	1.10	1.19	1.33	1.19	1.10	1.05
$m=3$	1.10	1.20	1.37	1.60	1.37	1.20	1.10
$m=4$	1.15	1.29	1.53	1.83	1.53	1.29	1.15
$m=5$	1.20	1.38	1.68	2.03	1.68	1.38	1.20
$m=8$	1.34	1.64	2.08	2.55	2.08	1.64	1.34
$m=10$	1.43	1.80	2.31	2.84	2.31	1.80	1.43
$m=20$	1.85	2.45	3.21	3.99	3.21	2.45	1.85

For a give quantile, Table 1 compares the asymptotic relative efficiencies of balance ranked set sample with simple random sample. It is observed that balance allocated ranked set sampling is more efficient than simple random sampling for all quantiles. The relative advantage of RSS is greatest at the median and tapers off in the tails. Increasing set-size m further enhances the performance of RSS. These findings suggest that a suitable selective allocation may further enhance the performance of RSS and make it more attractive for tests on the more extreme quantiles.

§4. Determination of the Optimal Design

We want to find the design D that maximizes ARE. Since δ_R^2/m is independent of D, the only factor of $\text{ARE}(S_D^+, S_{\text{RSS}}^+)$ that depends upon D is

$$h(d_1, \cdots, d_t) = \frac{\left(\sum\limits_{i=1}^{t} C_{d_i}\right)^2}{t\sum\limits_{i=1}^{t} D_{d_i}} = \frac{\left(\sum\limits_{i=1}^{t}\left(\frac{C_{d_i}}{t}\right)\right)^2}{\sum\limits_{i=1}^{t}\left(\frac{D_{d_i}}{t}\right)}.$$

To obtain the optimal design, we state the following lemma (See Kaur and Patil (2002)).

Lemma 4.1 Let j be fixed. If $C_j^2/D_j \geq C_i^2/D_i$ for all i, then $\left(\sum\limits_{i=1}^{m}\rho_i C_i\right)^2 \Big/ \sum\limits_{i=1}^{m}\rho_i D_i \geq C_i^2/D_i^2$ for all $\sum\limits_{i=1}^{m}\rho_i = 1$, $\rho_i \geq 0$, $i = 1, \cdots, m$.

The above lemma implies that the maximum of $h(d_1, \cdots, d_t)$ occurs at vertex of the simplex. Thus, the search of the optimum is narrowed down to examination of the m possible corner solutions. The question, then, is which rank order i maximizes C_i^2/D_i or, equivalently, minimizes D_i/C_I^2 for fixed $p \in (0,1)$. Using (2.5), (2.6), (2.7) and (2.8), we obtain

$$\frac{D_i}{C_i^2} = 4\int_0^p \left(\frac{u}{p}\right)^{i-1}\left(\frac{1-u}{1-p}\right)^{m-i}\mathrm{d}u \int_p^1 \left(\frac{u}{p}\right)^{i-1}\left(\frac{1-u}{1-p}\right)^{m-i}\mathrm{d}u = 4h(i,p),$$

let us define

$$h(x,p) = \int_0^p \left(\frac{u}{p}\right)^{x-1}\left(\frac{1-u}{1-p}\right)^{m-x}\mathrm{d}u \int_p^1 \left(\frac{u}{p}\right)^{x-1}\left(\frac{1-u}{1-p}\right)^{m-x}\mathrm{d}u, \qquad x \in [1,m].$$

Lemma 4.2 For fixed p, $h(x,p)$ is a convex function which has a unique minimum value.

In light of the above Lemma, one of the following must hold:

- $h(i,p)$ is increasing, in which case the smallest rank-order, $i = 1$, minimizes the function.
- $h(i,p)$ is decreasing, in which case the largest rank-order, $i = m$, minimizes the function.
- $h(i,p)$ is first decreasing and then increasing. Here, the minimum is attained at the first value of i for which $h(i,p) \leq h(i+1,p)$.

We have established the existence of an essentially unique rank-order that minimizes D_i/C_i^2. The only possible non-uniqueness occur when $h(i,p)$ is equal for two consecutive values of i. This optimal rank-order depends upon the value of p.

To actually locate the minimum, we go back to previous results and recall $D_i/C_i^2 = 4F_{(i)}(0)[1-F_{(i)}(0)]/C_i^2$. For simplification of notation, we write $F_i = F_{(i)}(0)$. Using the relationship between the incomplete beta function and cumulative binomial probabilities, we can write (2.7) as

$$F_i = b_0 + b_1 + \cdots + b_{i-1}, \qquad i = 1, 2, \cdots, m,$$

where $b_i = [m!/(i!(m-i)!)] \cdot p^i(1-p)^{m-i}$ are the binomial probabilities. The same approach can be used to get the optimal design in terms of p for other values of m. Mathematically, this entails finding the roots of polynomials. For ith rank-order to be optimal, we need $(i-1)$th rank-order to be non-optimal and

$$\frac{F_i(1-F_i)}{C_i^2} \le \frac{F_{i+1}(1-F_{i+1})}{C_{i+1}^2} \qquad \text{or} \qquad 0 \le C_i^2 F_{i+1}(1-F_{i+1}) - C_{i+1}^2 F_i(1-F_i).$$

Since $C_i = ib_i/p$ and $C_{i+1} = (m-1)b_i/(1-p)$, the above condition simplifies to

$$0 \le i^2(1-p)^2 F_{i+1}(1-F_{i+1}) - p^2(m-1)^2 F_i(1-F_i).$$

We have used to *Mathematica* determine the roots of the above series of polynomials. The results are summarized in Table 2. As a rough approximation, the ith rank-order is optimal for p in the range of $((i-1)/m, i/m)$. It appears that the optimal rank-order is a monotone increasing (step) function of the quantile(p) under test.

Table 2 Range of values of p over which the rank-order i^* is optimal for given m

	$i^*=1$	$i^*=2$	$i^*=3$	$i^*=4$	$i^*=5$
$m=2$	[0,0.50]	[0.50,1.0]			
$m=3$	[0,0.31]	[0.31,0.69]	[0.69,1.0]		
$m=4$	[0,0.22]	[0.22,0.50]	[0.50,0.78]	[0.78,1.0]	
$m=5$	[0,0.17]	[0.17,0.39]	[0.39,0.61]	[0.61,0.82]	[0.82,1.0]
$m=6$	[0,0.14]	[0.14,0.31]	[0.31,0.50]	[0.50,0.68]	[0.68,0.86]
$m=7$	[0,0.12]	[0.12,0.27]	[0.27,0.42]	[0.42,0.58]	[0.58,0.73]
$m=8$	[0,0.10]	[0.10,0.23]	[0.23,0.36]	[0.36,0.50]	[0.50,0.63]
$m=9$	[0,0.09]	[0.09,0.21]	[0.21,0.32]	[0.32,0.44]	[0.44,0.56]
$m=10$	[0,0.08]	[0.08,0.18]	[0.18,0.29]	[0.29,0.39]	[0.39,0.50]
	$i^*=6$	$i^*=7$	$i^*=8$	$i^*=9$	$i^*=10$
$m=6$	[0.86,1.0]				
$m=7$	[0.73,0.88]	[0.88,1.0]			
$m=8$	[0.63,0.77]	[0.77,0.90]	[0.90,10]		
$m=9$	[0.56,0.68]	[0.68,0.79]	[0.79,0.91]	[0.91,1.0]	
$m=10$	[0.50,0.61]	[0.61,0.71]	[0.71,0.81]	[0.81,0.92]	[0.92,1.0]

For given m and p, let i^* be the optimal rank-order on which all quantifications are to be made. The test statistic under optimal protocol for the p quantile simplifies to

$$S^+_{\text{opt}} = \sum_{j=1}^{n} I(X_{(i^*)j} > 0),$$

and its exact distribution is seen to be

$$S_{\text{opt}}^{+} \sim \text{binomial}(n, 1 - G_{(i^*)}(0)).$$

Further, $\mathsf{E}(S_{\text{opt}}^{+}) = n[1 - G_{(i^*)}(0)]$ and $\mathsf{Var}\,(S_{\text{opt}}^{+}) = nG_{(i^*)}(0)[1 - G_{(i^*)}(0)]$.

From the definition (3.2), the efficacy of S_{opt}^{+} is

$$\text{eff}\,(S_{\text{opt}}^{+}) = \frac{g_{(i^*)}^2(0)}{G_{(i^*)}(0)[1 - G_{(i^*)}(0)]}\Big|_{H_0}. \tag{4.1}$$

By using (2.6), (2.7), (3.5) and (4.1), we have

$$\begin{aligned} \text{ARE}(S_{\text{opt}}^{+}, S_{\text{SRS}}^{+}) &= \frac{\text{eff}\,(S_{\text{opt}}^{+})}{\text{eff}\,(S_{\text{SRS}}^{+})} \\ &= \frac{p^{2i^*-1}(1-p)^{2m-2i^*+1}}{B(p; i^*, m+1-i^*)[B(1; i^*, m+1-i^*) - B(p; i^*, m+1-i^*)]}. \end{aligned}$$

Table 3 The asymptotic relative efficiencies of S_{opt}^{+} versus S_{SRS}^{+}

	$p=0.05$	$p=0.10$	$p=0.20$	$p=0.50$	$p=0.80$	$p=0.90$	$p=0.95$
$m=2$	1.95	1.89	1.78	1.33	1.78	1.89	1.95
$m=3$	2.85	2.69	2.36	2.25	2.36	2.69	2.85
$m=4$	3.70	3.39	2.78	2.62	2.78	3.39	3.70
$m=5$	4.50	4.01	3.46	3.52	3.46	4.01	4.50
$m=8$	6.64	5.37	5.52	5.17	5.52	5.37	6.64
$m=10$	7.85	6.95	6.68	6.45	6.68	6.95	7.85
$m=20$	13.92	13.39	13.07	12.82	13.07	13.39	13.92

For a give quantile, Table 3 gives the asymptotic relative efficiencies of the sign test under optimal ranked set sample with that under the simple random sample. The performance of optimal design improves as p moves away from 0.5, and as the set-size, m, increases. As observed from Table 3, The relative advantage of optimal design is quite high for testing the more extreme quantiles.

References

[1] Takahasi, K. and Wakimoto, K., On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of Institute of Statistical Mathematics*, **20**(1968), 1–31.

[2] Bohn, L.L. and Wolfe, D.A., Nonparametric two-sample procedures for ranked-set samples data, *Journal of the American Statistical Association*, **87**(1992), 552–561.

[3] Hettmansperger, T.P., The ranked-set sample sign test, *Journal of the Nonparametric Statistics*, **4**(1995), 263–270.

[4] Ozturk, O. and Wolef, D.A., Alternative ranked-set sampling protocols for the sign test, *Statistics & Probability Letters*, **47**(2000), 15–23.

[5] Wang, Y.G. and Zhu, M., Optimal sign tests for data from ranked set samples, *Statistics & Probability Letters*, **72**(2005), 13–22.

[6] Mao, S.S., Wang, J.L. and Pu, X.L., *Advanced Mathematical Statistics*, China Higher Education Press and Springer Press, Beijing, 1998.

[7] Fang, Z.B., Hu, T.Z., Wu, Y.H. and Zhuang, W.W., Multivariate stochastic orderings of spacings of generalized order statistics, *Chinese Journal of Applied Probability and Statistics*, **3**(2006), 295–303.

[8] Kaur, A. and Patil, G.P., Ranked set sample sign test for quantiles, *Journal of Statistical Planning and Inference*, **100**(2002), 351–370.

符号检验的最优排序集抽样

张良勇　　董晓芳

(齐齐哈尔大学理学院, 齐齐哈尔, 161006)

提出检验总体分位数的基于排序集抽样的符号检验, 分析了不同挑选抽样相对于均衡抽样的Pitman渐近效率. 针对不同分位数, 具体给出使符号统计量的效率达到最大的抽样设计, 并且证明了最优抽样不依赖于总体分布.

关键词: 非参数检验, 排序集抽样, 最优设计, 功效.

学科分类号: O212.7.