Ann. Inst. Statist. Math.
Vol. 40, No. 1, 111-117 (1988)

CONVERGENCE RATES FOR TWO-STAGE CONFIDENCE INTERVALS BASED ON U-STATISTICS

N. MUKHOPADHYAY[1] AND G. VIK[2]

[1] *Department of Statistics, University of Connecticut, 196 Auditorium Road, Storrs, CT 06268, U.S.A.*
[2] *Norwegian Defense Research Establishment, 2007 Kjeller, Norway*

(Received March 9, 1987; revised September 29, 1987)

Abstract. We consider a modified two-stage procedure for constructing a fixed-width confidence interval for the mean of a U-statistic. First, we discuss a few asymptotic results with the associated rates of convergence. The main result gives the rate of convergence for the coverage probability of our proposed confidence interval which is seen to be slower than that for the purely sequential procedure.

Key words and phrases: Fixed-width confidence intervals, two-stage procedure, rate of convergence, mean of a U-statistic.

1. Introduction

Various problems related to rates of convergence for randomly stopped U-statistics have received much attention lately since the appearance of the works of Landers and Rogge (1976). One may note this trend from Ghosh (1980), Ghosh and DasGupta (1980), Callaert and Janssen (1981) and Aerts and Callaert (1982) among others. Csenki (1980) used the results of Landers and Rogge (1976) to derive the rate of convergence of the coverage probability for Chow and Robbins' (1965) fixed-width sequential confidence interval procedure. Mukhopadhyay (1981) generalized this particular result by applying the tools from Ghosh and DasGupta (1980) to derive the rate of convergence of the coverage probability for Sproule's (1969, 1974) fixed-width sequential confidence interval procedure. The recent paper of Mukhopadhyay and Vik (1985) considers a nonparametric approach to a parametric problem and develops the rates of convergences of various characteristics of the proposed sequential procedures in general. In this present note, we continue the type of program started in Mukhopadhyay and Vik (1985) for suitable modified two-stage procedures.

We first propose a general stopping time along the lines of the modified two-stage procedures of Mukhopadhyay (1980, 1981, 1982) based on U-statistics. We briefly present results on various aspects of rates of convergences for such stopping times, and these are applied to obtain the rate of convergence of the coverage probability of a fixed-width confidence interval for the mean of a U-statistic followed by a few examples. For most of the groundwork and other references, the reader is referred to Mukhopadhyay and Vik (1985).

2. A modified two-stage procedure

In this section we propose a general modified two-stage procedure based on U-statistics and study some of its asymptotic properties. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables having a distribution function $F(\cdot)$ where $F(\cdot)$ belongs to a family $\boldsymbol{F}$ of distribution functions. Let $\phi(X_1, \ldots, X_r)$ be a symmetric kernel of degree r. Now for $n \geq r$, Hoeffding (1948) defined U-statistics as follows:

$$U_n = \binom{n}{r}^{-1} \sum_{n,r} \phi(X_{\alpha_1}, \ldots, X_{\alpha_r}) ,$$

where $\sum_{n,r}$ stands for the summation over all combinations $\{\alpha_1 < \cdots < \alpha_r\}$ formed from the integers $\{1, 2, \ldots, n\}$. Furthermore, we define

$$\phi_c(x_1, \ldots, x_c) = E\{\phi(X_1, \ldots, X_r) | X_1 = x_1, \ldots, X_c = x_c\} ,$$

$$\xi_c = \mathrm{Var}\{\phi_c(X_1, \ldots, X_c)\}, \quad c = 1, \ldots, r .$$

At this stage we refer the reader to Subsection 1.2 in Mukhopadhyay and Vik (1985) for some examples and further comments.

We will very shortly connect these $\{U_n: n \geq r\}$ to an estimation problem for the mean of another sequence of U-statistics denoted by $\{V_m: m \geq s\}$. We assume that

$$V_m = \binom{m}{s}^{-1} \sum_{m,s} g(X_{\beta_1}, \ldots, X_{\beta_s}) ,$$

corresponding to some symmetric kernel $g(x_1, \ldots, x_s)$. We further assume that $E\{|g|^2\} < \infty$ and define $\eta_c = \mathrm{Var}\{g_c(X_1, \ldots, X_c)\}$ where $g_c(x_1, \ldots, x_c) = E\{g(X_1, \ldots, X_s) | X_1 = x_1, \ldots, X_c = x_c\}$ for $c = 1, \ldots, s$. Let us write $\mu = E\{g(X_1, \ldots, X_s)\}$.

Given d (>0) and $q \in (0, 1)$, we would like to construct a confidence interval I_m for μ such that the length of I_m is $2d$ and $P\{\mu \in I_m\} \approx 1 - q$. We propose to consider the natural confidence interval $I_m = [V_m \pm d]$ for the parameter μ. Now,

$$\begin{aligned} P\{\mu \in I_m\} &= P\{|V_m - \mu| \leq d\} \\ &= P\{|m^{1/2}(V_m - \mu)| \leq m^{1/2} d\} . \end{aligned}$$

From Hoeffding's (1948) results it follows that $m^{1/2}(V_m-\mu) \xrightarrow{L} N(0, s^2\eta_1)$ as $m\to\infty$ if $0<\eta_1<\infty$. Thus, $P\{\mu \in I_m\}\approx(1-q)$ for large m if $2\Phi((m^{1/2}d)(s\eta_1^{1/2})^{-1})-1\approx 1-q$. Let $2\Phi(a)-1=1-q$. Then, m needs to be the smallest integer $\geq a^2s^2d^{-2}\eta_1=m_d$, say.

Of course, m_d would be unknown in most applications. Motivated by the developments in Mukhopadhyay (1980, 1981, 1982) we now define a modified two-stage procedure giving rise to the following stopping variable. For arbitrary but fixed $0<\eta<2$, let the starting sample size m_0 be defined by

$$m_0 = \max\{[(a/d)^\eta]^* + 1, s\}\ ,$$

where $[x]^*$ stands for the largest integer smaller than x. We assume that the sequence $\{U_m\colon m\geq r\}$ is such that $\{E(U_m)\}^\alpha=k_1^2\eta_1$ for some known positive numbers α and k_1. Now, we define the stopping variable as

$$M_d = \max\left\{m_0,\ \left[\left(\frac{as}{dk_1}\right)^2 U_{m_0}^\alpha\right]^* + 1\right\}. \tag{2.1}$$

That is, having observed $X_1,\ldots, X_{m_0}$, we determine M_d and we sample the difference at the second stage, if necessary. Having recorded $X_1,\ldots, X_{M_d}$, we obtain V_{M_d} and we thus propose the confidence interval I_{M_d} for μ.

LEMMA 2.1. *Assume that $E(\phi^2)<\infty$ and $\eta_1>0$. Then, for the stopping variable M_d defined in* (2.1) *we have*

(i) $P(M_d < \infty) = 1$,

(ii) $M_d\to\infty$ *w.p.* 1 *as* $d\to 0$,

(iii) $E\{(M_d/m_d)^\delta\}\to 1$ *as* $d\to 0$ *if* $E\{|\phi|^{2\xi}\} < \infty$ *for* $\xi > \max(\delta/2, 1/2)$,

where $m_d=(as/d)^2\,\eta_1$ *and* δ *is positive.*

Parts (i) and (ii) follow directly from the definition of M_d. A proof of part (iii) can be constructed along the lines of Lemma 2.4 in Mukhopadhyay and Vik (1985). Further details are omitted.

We have somewhat stronger assertions in the following lemma for $\alpha=1$.

LEMMA 2.2. *Let $\alpha=1$ in the definition* (2.1) *of the stopping variable M_d. If $E(\phi^2)<\infty$, then we have as $d\to 0$*

(i) $E\{M_d d^2(a^2s^2\eta_1)^{-1}\} = 1 + O(d^2)$,

(ii) $\mathrm{Var}\{M_d\}[\mathrm{Var}\{U_{m_0}\}\{as(dk_1)^{-1}\}^4] = 1 + O(d^{2-\eta})$,

where k_1 is defined by $E(U_{m_0})=k_1^2\eta_1$ and $\eta \in (0, 2)$ appears in the definition of m_0.

Its proof is omitted for brevity. The reader is referred to Vik (1984).

Now we consider the coverage probability when I_{M_d} is proposed as the confidence interval for μ. From Sproule (1969, 1974) it follows that $\lim_{d\to 0} P\{\mu \in$

$I_{M_d}\} = 1 - q$; however, we wish to study the rate of this convergence along the lines of Csenki (1980), Mukhopadhyay (1981), and Mukhopadhyay and Vik (1985). In order to do that, we will need the following result for M_d.

THEOREM 2.1. *Consider M_d as defined by (2.1). Assume that $E\{|\phi|^{2\xi}\}<\infty$ for $\xi \geq (1-2\lambda)/(2\eta-4+8\lambda)$ where $\lambda \in (l, 1/2)$ for $l=(2-\eta)/4$. Then we have as $d \rightarrow 0$*

$$P\{|(M_d/m_d) - 1| > kd^{2(1/2-\lambda)}\} = O(d^{1/2-\lambda}) ,$$

where $m_d=(as/d)^2\eta_1$ and $k(>0)$ is arbitrary.

PROOF. Let k be a generic positive constant independent of d and $c=(as/k_1)^2$. Now, we have

$$\begin{aligned} P\{M_d > m_d + km_d^{1/2+\lambda}\} & \\ \leq P\{cd^{-2}U_{m_0}^{\alpha} + 1 > m_d + km_d^{1/2+\lambda}, cd^{-2}U_{m_0}^{\alpha} > m_0\} & \\ + P\{cd^{-2}U_{m_0}^{\alpha} \leq m_0\} , & \end{aligned} \tag{2.2}$$

and we also have

$$P\{cd^{-2}U_{m_0}^{\alpha} > m_d + km_d^{1/2+\lambda} - 1\} = O(d^{\xi\{4(\lambda-1/2)+\eta\}}) , \tag{2.3}$$

by the Corollary 1.2 in Mukhopadhyay and Vik (1985). For the term in (2.3) to have the specific order, we need $\xi \geq (1-2\lambda)(2\eta-4+8\lambda)^{-1}$ and $\lambda > (2-\eta)/4$. Next, we note that

$$P\{cd^{-2}U_{m_0}^{\alpha} \leq m_0\} \leq km_0^{-\xi} = O(d^{1/2-\lambda}) ,$$

if $\xi \geq (1/2-\lambda)/\eta$. This will always hold if the conditions for (2.3) to have the specific order hold. So far we have shown that under the conditions of Theorem 2.1,

$$P\{(M_d/m_d) - 1 > kd^{2(1/2-\lambda)}\} = P\{M_d > m_d + km_d^{1/2+\lambda}\} = O(d^{1/2-\lambda}) . \tag{2.4}$$

Again we have

$$P\{1 - (M_d/m_d) > kd^{2(1/2-\lambda)}\} \leq P\{\theta^{\alpha} - U_{m_0}^{\alpha} > km_d^{\lambda+1/2}\} . \tag{2.5}$$

By comparing (2.5) with (2.3), we see that the same kind of arguments that led to (2.4) will be also applicable to obtain the specific order in (2.5). This concludes the proof of Theorem 2.1.

The following theorem establishes our main result.

THEOREM 2.2. *Consider M_d as defined by* (2.1). *Assume that* $E\{|\phi|^{2\xi}\}<\infty$ *for* $\xi\geq(1-2\lambda)/(2\eta-4+8\lambda)$, $E(g^4)<\infty$ *and* $\lambda\in(l, 1/2)$ *for* $l=(2-\eta)/4$. *Then we have as* $d\to 0$

$$P\{\mu\in I_{M_d}\} = (1 - q) + O(d^{1/2-\lambda}) .$$

In view of our Theorem 2.1, we can construct a proof of Theorem 2.2 quite easily along the lines of Csenki (1980). The reader may also look at the proof of Theorem 3.1 in Mukhopadhyay and Vik (1985) for some clarification. We omit all further details.

Remark 1. If we compare this last result about the rate of convergence of the coverage probability with the corresponding result for the sequential procedure (see Theorem 3.1 in Mukhopadhyay and Vik (1985)), we readily see that for the modified two-stage procedure (2.1), λ is bounded below by a positive constant. This gives us a slower rate of convergence for the modified two-stage procedure (2.1) in comparison with that for the sequential one. In the terminology of Mukhopadhyay (1981), this two-stage procedure is only "first-order asymptotically consistent", while the corresponding sequential procedure is also "second-order asymptotically consistent". Note that the rate for the two-stage procedure gets better as $\eta\in(0, 2)$ gets larger in the definition of m_0 in (2.1).

Remark 2. Here, we will take the opportunity to correct the requirements for Theorem 5 in Mukhopadhyay (1981). In the context of that paper, Theorem 5 holds for $(1-\eta)/4<\gamma<1/2$. Thus, the sharper rate of convergence of the coverage probability is obtained in Mukhopadhyay's (1981) Theorem 5, if we choose larger η in $[0, 1)$. One may note that this observation is completely in agreement with our findings here in Theorem 2.2 and Remark 1.

3. Examples

As in Mukhopadhyay and Vik's (1985) Section 5, we can easily illustrate the result obtained in Theorem 2.2 by specializing F to Bernoulli (p), Poisson (δ), Gamma (δ, β), etc. To be candid, we refrain from doing that; instead, we give an example from the $N(\mu, \sigma^2)$ population.

Let $X_1, X_2, \ldots$ be i.i.d. $N(\mu, \sigma^2)$ random variables with $\mu\in(-\infty,\infty)$, $\sigma\in(0, \infty)$. We consider the problem of estimating σ^2 when μ is unknown. For $m\geq 2$, let us use $S_m^2=(m-1)^{-1}\sum_{i=1}^{m}(X_i-\bar{X}_m)^2$, which is a U-statistic as an estimator of σ^2. Now, from a result of Hoeffding (1948) we get $4\eta_1=\lim_{m\to\infty}\{m \operatorname{Var}(S_m^2)\}$, that is $\eta_1=\sigma^4/2$. We then suggest $I_M=[S_M^2\pm d]$ as the fixed-width confidence interval for σ^2 where $M=M_d$ is defined as in (2.1); that is, we let

$$m_0 = \max\{2, [(a/d)^{\eta}]^* + 1\}$$

and

$$M = \max\{[2a^2d^{-2}S_{m_0}^4]^* + 1, m_0\} \quad \text{for} \quad \eta \in (0, 2) .$$

Now, Lemma 2.1 gives

$$E\{Md^2/(2a^2\sigma^4)\} \to 1 ,$$

as $d \to 0$. Also, for $\lambda \in ((2-\eta)/4, 1/2)$, Theorem 2.2 gives

$$P\{\sigma^2 \in I_M\} = (1 - q) + O(d^{1/2-\lambda}) .$$

Let us now consider the two-stage procedure leading to the sample size M^* proposed by Graybill and Connell (1964). This M^* is defined by $M^* = m_1 + M_1$ where $m_1 (\geq 2)$ is the starting sample size and

$$M_1 = [2 + \pi\{q^{-2/(m_1-1)} - 1\}^2 (m_1 - 1)^2 S_{m_1}^4 (4d^2)^{-1}]^* + 1 .$$

Then, $P(\sigma^2 \in I_{M_1}) \geq 1-q$ and

$$E(M^*) \simeq m_1 + 2 + h(q, m_1)\, \sigma^4 d^{-2} ,$$

where $h(q, m_1) = (1/4)\pi\{q^{-2/(m_1-1)} - 1\}^2 (m_1^2 - 1)$. Let us now compare our M with Graybill and Connell's (1964) M^* by considering

$$e(q, m_1) = \lim_{d \to 0} \{E(M^*)/E(M)\} = h(q, m_1)/2a^2 .$$

As an illustration, let $q = .05$. Now, the following table gives the values of $e(.05, m_1)$ for some values of m_1. The quantity $e(.05, m_1)$ being larger than unity will signify the superiority of our two-stage procedure. In the context of this particular problem, Table 1 shows that the procedure through M^* will need

Table 1. Values of $e(.05, m_1)$.

m_1	$e(.05, m_1)$
10	9.05
50	4.32
100	3.98
500	3.73
1000	3.70
10000	3.67
100000	3.67

more than 3.5 times the sample size required by the procedure through our M, over the range of m_1 considered here. However, the larger average sample size required by the procedure through M^* is expected to provide us with higher coverage probability than our asymptotic target, namely, $(1-q)$.

REFERENCES

Aerts, M. and Callaert, H. (1982). The convergence rate of sequential fixed-width confidence intervals for regular functionals, Tech. Report, Limburgs Universitair Centrum.

Callaert, H. and Janssen, P. (1981). The convergence rate of fixed-width sequential confidence intervals for the mean, *Sankhyā Ser. A*, **43**, 211–219.

Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean, *Ann. Math. Statist.*, **36**, 457–462.

Csenki, A. (1980). On the convergence rate of fixed-width sequential confidence intervals, *Scand. Actuar. J.*, 107–111.

Ghosh, M. (1980). Rate of convergence to normality for random means: Applications to sequential estimation, *Sankhyā Ser. A*, **42**, 231–240.

Ghosh, M. and DasGupta, R. (1980). Berry-Esseen theorems for *U*-statistics in the non i.i.d. case, *Proc. Nonparam. Statist. Inference*, 293–313, Budapest, Hungary.

Graybill, F. A. and Connell, T. L. (1964). Sample size for estimating the variance within *d* units of the true value, *Ann. Math. Statist.*, **35**, 438–440.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.*, **19**, 293–325.

Landers, D. and Rogge, L. (1976). The exact approximation order in the central-limit-theorem for random summation, *Z. Wahrsch. Verw. Gebiete*, **36**, 269–283.

Mukhopadhyay, N. (1980). A consistent and asymptotically efficient two-stage procedure to construct fixed-width confidence intervals for the mean, *Metrika*, **27**, 281–284.

Mukhopadhyay, N. (1981). Convergence rates of sequential confidence intervals and tests for the mean of a *U*-statistic, *Comm. Statist. A—Theory Methods*, **10**, 2231–2244.

Mukhopadhyay, N. (1982). Stein's two-stage procedure and exact consistency, *Scand. Actuar. J.*, 110–122.

Mukhopadhyay, N. and Vik, G. (1985). Asymptotic results for stopping times based on *U*-statistics, *J. Sequential Analysis*, **4**, 83–109.

Sproule, R. N. (1969). A sequential fixed-width confidence interval for the mean of a *U*-statistic, Ph.D. Dissertation, University of North Carolina, Chapel Hill.

Sproule, R. N. (1974). Asymptotic properties of *U*-statistics, *Trans. Amer. Math. Soc.*, **199**, 55–64.

Vik, G. (1984). Asymptotic results for stopping times based on *U*-statistics, Ph.D. Dissertation, Oklahoma State University, Stillwater.