

## Asymptotic Optimality in Sequential Interval Estimation

MICHAEL WOODROOFE\*

*Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109*

DEDICATED TO HERBERT ROBBINS ON THE OCCASION  
OF HIS 70TH BIRTHDAY

The problem of setting a fixed width confidence interval for the mean of a normal distribution with unknown variance is considered. Several procedures are reviewed. An asymptotic lower bound for the expected sample size of any sequential sampling plan with the specified confidence coefficient is obtained as the width of the interval decreases to 0. © 1986 Academic Press, Inc.

### 1. INTRODUCTION

Beginning with Stein's [9] two stage procedure, there has been substantial interest in the use of sequential methods to set a fixed width confidence interval for the mean of a distribution with unknown variance. This interest was stimulated by the fully sequential procedures of Anscombe [1] and Chow and Robbins [2] for the normal and non-parametric cases, and continues today. The recent work of Hall [5] and Finster [4] is especially noteworthy and may be consulted for further references.

The emphasis of much of this work has been the construction of sequential sampling plans for which the probability that the sample and population means differ by less than a prescribed  $h > 0$  is large, at least of prescribed  $\gamma$ , and the comparison of the expected sample sizes of such plans with the hypothetical fixed sample size which would be required if the variance were known. The plans effectively estimate the hypothetical fixed sample size as data accumulate and stop when the sample size exceeds the

\*Research supported by the National Science Foundation under DMS-8413452.

estimate; and the comparisons take the form of limit theorems as  $h \rightarrow 0$  with fixed  $\gamma$ . Some of this work is reviewed in Section 2.

While there is considerable research which studies specific sampling plans and compares them to the hypothetical fixed sample size plan, there is little research which seeks to find optimal procedures. Stein and Wald [10] obtain a lower bound on the sample size needed for the case in which  $\sigma$  is known; and the asymptotic expression for the expected sample size of the fully sequential procedure compares favorably with it; but it is not known whether this bound is the best possible for the case in which  $\sigma$  is unknown.

The main result of this paper is an inequality which relates the expected sample size of a sequential procedure to its coverage probabilities for the case of unknown  $\sigma$ . From this inequality, the fully sequential sampling plan is shown to be the best possible up to terms which are small compared to the cost of a single observation.

The paper is organized as follows: A brief account of previous research on the problem is given in Section 2. The main result and its corollary are stated in Section 3, and proved in Section 4.

## 2. THE PROBLEM

In this section the problem is stated, and several solutions are described. Of these, only (5) and (6) are needed below. The reader who is uninterested in the historical development of the problem may simply note their statement and proceed to the next section. For simplicity, only the one sample problem is described, though many of the ideas extend to more complicated designs. Also, second order asymptotics are emphasized, since the effect of optional stopping only appears in these terms.

To be specific, let  $X_1, X_2, \dots$  denote independent normally distributed random variables with common unknown mean  $\mu$ ,  $-\infty < \mu < \infty$ , and common unknown variance  $\sigma^2$ , and let

$$\bar{X}_n = (X_1 + \dots + X_n)/n, \quad n \geq 1,$$

and

$$S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 / (n - 1), \quad n \geq 2,$$

denote the sample means and variances. Further, let  $h > 0$  denote the desired half width of the interval; let  $0 < \gamma < 1$  denote the desired confidence coefficient; and let  $c$  denote the  $(1 + \gamma)/2$  quantile of the standard

normal distribution,

$$2\Phi(c) - 1 = \gamma, \quad (1)$$

where  $\Phi$  denotes the standard normal distribution function.

If the variance  $\sigma^2$  was known, then an elementary calculation shows that the interval

$$I_n = [\bar{X}_n - h, \bar{X}_n + h]$$

covers  $\mu$  with probability at least  $\gamma$  if  $n$  is a (non-random) integer for which

$$n \geq N = N(h, \sigma) = c^2\sigma^2/h^2. \quad (2)$$

So,  $N$  is the hypothetical fixed sample size which would be required if  $\sigma$  were known. Observe that  $N \rightarrow \infty$  as  $h \rightarrow 0$  for each fixed  $\sigma > 0$ .

For the case in which  $\sigma$  is unknown, Stein's [9] two-stage procedure takes an initial sample of size  $m \geq 2$  and then a second sample, if necessary, to bring the total sample size to

$$s = s_h = \max\{m, [c_{m-1}^2 S_m^2/h^2] + 1\}, \quad (3)$$

where  $c_{m-1}$  denotes the  $(1 + \gamma)/2$  quantile of the  $t$  distribution on  $m - 1$  degrees of freedom and  $[x]$  denotes the greatest integer which does not exceed  $x$ . It is easily seen that the distribution of  $s$  and the coverage probabilities depend on  $\mu$ ,  $\sigma$  and  $h$  only through  $N$ . Stein showed that  $I_s$  covers  $\mu$  with probability of at least  $\gamma$  for all  $N > 0$  and that  $E_N(s) = Nc_{m-1}^2/c^2$  as  $h \rightarrow 0$ .

Observe that the two-stage procedure does not use the information from the second sample to estimate the variance and that the expected sample size is asymptotically bigger than  $N$  by a constant factor. The fully sequential procedures of Anscombe [1] and Chow and Robbins [2] avoid these problems. Again, let  $m$  denote an initial sample size and now let  $c_n$ ,  $n \geq m$ , denote any sequence for which

$$c_n = c[1 + b/n + o(1/n)], \quad \text{as } n \rightarrow \infty,$$

for some finite constant  $b$ . Define stopping times by

$$\tau = \tau_h = \inf\{n \geq m: n > c_n^2 S_n^2/h^2\} \quad (4)$$

for  $h > 0$ . Then the distribution of  $\tau$  and the related coverage probabilities depend only on  $N$ ; and if  $m \geq 4$ , then the following relations hold as  $N \rightarrow \infty$ :

$$E_N(\tau) = N + \nu + 2b - 2 + o(1) \quad (5)$$

and

$$P_N\{\mu \in I_s\} = \gamma + N^{-1}c\phi(c)\{E_N(\tau - N) - (1 + c^2)/2\} + o(N^{-1}),$$

where  $\nu \simeq .82$  denotes a constant (an expected excess over the boundary in a non-linear renewal problem). Thus, if

$$\nu + 2b - 2 = (1 + c^2)/2, \tag{6}$$

then

$$E_N(\tau) = N + (1 + c^2)/2 + o(1)$$

and

$$P_N\{\mu \in I_\tau\} = \gamma + o(1/N), \quad \text{as } N \rightarrow \infty.$$

These results appear in Anscombe [1], although without complete proofs. Such were provided by Woodroffe [11], under the additional condition  $m \geq 7$ . Starr [7] gives numerical calculations with which the asymptotics may be compared. Simons [6] and Starr and Woodroffe [8] give bounds which complement the asymptotic relations. The corollary in the next section shows that the expected sample size following (6) is best possible, subject to the conditions on the coverage probabilities.

With Hall [5], interest returned to multistage procedures, which have the practical advantage of not requiring the data to be monitored continuously. Hall's procedure takes an initial sample of size  $m$ , a second sample (if necessary) to bring the sample size to

$$s = \max\{m, [uc^2S_m^2/h^2] + 1\},$$

where  $0 < u < 1$ , and then a third sample (if necessary) to bring the total sample size to

$$t = \max\{s, [c^2S_s^2/h^2 + (1 + c^2)/2u - 1/2] + 1\}$$

As above, it is easily verified that the coverage probabilities and expected sample size depend only on  $\sigma$  and  $\mu$  only through  $N$ . Hall showed that if  $m = m(h) \rightarrow \infty$  at a suitable rate, then

$$E_N(t) = N + (1 + c^2)/2u + o(1)$$

and

$$P_N\{\mu \in I_t\} = \gamma + o(1/N),$$

as  $h \rightarrow 0$ ; and Hall's results remain valid if  $u = u(h) \rightarrow 1$  as  $h \rightarrow 0$ , sufficiently slowly.

## 3. THE THEOREM

In the sequel  $t$  denotes a stopping time with respect to the sequence of sample variances. Thus,  $t$  is a positive integer or  $\infty$  valued random variable for which  $t < \infty$  w.p.1 for all  $\sigma > 0$  and the event  $\{t = n\}$  depends only on  $S_k^2, k = 2, \dots, n$  for all  $n = 1, 2, \dots$ .

Let  $Y_1, Y_2, \dots$  denote the Helmert transformation

$$Y_k = \left[ \sum_{i=1}^k X_i - kX_{k+1} \right] / \sqrt{k(k+1)}$$

for  $k = 1, 2, \dots$ . Then  $Y_1, \dots, Y_{n-1}$  are independent of  $\bar{X}_n$  for each  $n = 2, 3, \dots$ , and

$$S_k^2 = \left( \sum_{i=1}^{k-1} Y_i^2 \right) / (k-1)$$

for all  $k = 2, 3, \dots$ . It follows that the event  $\{t = n\}$  is independent of  $\bar{X}_n$  for all  $n = 2, 3, \dots$ . Let

$$\gamma_h(t; \sigma) = P_\sigma\{|\bar{X}_t - \mu| \leq h\}.$$

Then

$$\gamma_h(t; \sigma) = E_\sigma\{2\Phi(h\sqrt{t}/\sigma) - 1\}$$

follows easily for all  $\sigma > 0$  and all  $h > 0$ , by conditioning on the value of  $t$ . See, for example, Woodroffe ([12], Section 10.2).

The key to relating the expected sample size and coverage probability of sequential sampling plans is to consider a decision problem in which the statistician selects a stopping time  $t$  (with respect to the sample variances) and loses the amount

$$L_h(t; \sigma) = NK(t/N), \quad (7)$$

where

$$K(x) = 2ac^{-2}[1 - \Phi(cx)] + x - [ac^{-2}(1 - \gamma) + 1],$$

$$a = c/\phi(c),$$

$\phi$  denotes the standard normal density, and  $2\Phi(c) - 1 = \gamma$ , and  $x > 0$ . Here the first term in  $L_h(t; \sigma)$  represents the probability that  $I_t$  does not cover  $\mu$ ; the second term represents a cost of sampling; and the third, which does not depend on  $t$ , is included for normalization.

The risk function and average risk functional of a stopping time  $t$  are defined by

$$r_h(t; \sigma) = E_\sigma\{L_h(t; \sigma)\} = NE_\sigma\{K(t/N)\} \tag{8}$$

and

$$R_h(t; \pi) = \int_0^\infty r_h(t; \sigma) d\pi(\sigma)$$

for  $\sigma > 0, h > 0$ , and prior distributions  $\pi$  on  $(0, \infty)$  for which  $\sigma^2$  has a finite expectation. The following asymptotic solution to the decision problem provides inequalities which relate the coverage probabilities and expected sample size.

**THEOREM.** *If  $\pi$  is any prior distribution on  $(0, \infty)$  having a twice continuously differentiable density with compact support in  $(0, \infty)$ , then*

$$\liminf_{h \rightarrow 0} \inf_t R_h(t; \pi) \geq (1 + c^2)/2, \tag{9}$$

where the infimum extends over all stopping times (with respect to the sample variance sequence); moreover, if  $\tau = \tau_h$  denotes the fully sequential procedure (4) with  $m \geq 4$  and  $b$  as in (6), then

$$\lim_{h \rightarrow 0} R_h(\tau; \pi) = (1 + c^2)/2. \tag{10}$$

The proof of the theorem is presented in the next section. A simple corollary restates the asymptotic optimality in classical terms.

**COROLLARY.** *Let  $t = t_h, h > 0$ , be any family of stopping times for which*

$$P_\sigma\{|\bar{X}_t - \mu| \leq h\} \geq \gamma + o(h^2), \tag{11}$$

uniformly in  $\sigma$  on compact subsets of  $(0, \infty)$  as  $h \rightarrow 0$ . Then, for any compact subinterval  $J \subset (0, \infty)$  with non-empty interior,

$$\liminf_{h \rightarrow 0} \sup_{\sigma \in J} E_\sigma[t - N] \geq (1 + c^2)/2. \tag{12}$$

Moreover, the limits exist and there is equality in (11) and (12) when  $t = \tau$  is the fully sequential procedure (4) with  $b$  as in (6) and  $m \geq 4$ .

*Proof.* Suppose that (11) holds, but (12) fails for some compact subinterval  $J$  with non-empty interior. Then there is a  $b < (1 + c^2)/2$  for which  $E_\sigma(t - N) \leq b$  for all  $\sigma \in J$ , for all small  $h > 0$ . Consequently, if  $\pi$  has

support  $J$ , then

$$R_h(t; \pi) \leq \int_0^\infty \{ a(1 - \gamma)\sigma^2/h^2 + (N + b) \\ - [a(1 - \gamma) + c^2]\sigma^2/h^2 \} d\pi + o(1) \leq b + o(1)$$

for all sufficiently small  $h$ , contradicting the theorem. This establishes the first assertion, and the second simply restates (5) and (6).

#### 4. THE PROOF

The second assertion of the theorem, relation (10), follows directly from (5) and (6). In the proof of the first,  $\pi$  denotes a prior distribution with compact support in  $(0, \infty)$ ;  $P$  denotes probability in the Bayesian model, where  $\sigma$  has prior distribution  $\pi$  and  $Y_k, k = 1, 2, \dots$ , are conditionally independent normally distributed random variables with mean 0 and variance  $\sigma^2$ , given  $\sigma$ ; and  $E$  denotes expectation with respect to  $P$ .

The first step in the proof is to observe that

$$R_h(t; \pi) = E[L_h(t; \sigma)]$$

for stopping times  $t$  and  $h > 0$ . Thus, the problem of minimizing  $R_h(t; \pi)$  with respect to  $\pi$  is an optimal stopping problem. Such problems are admirably described in the monograph by Chow *et al.* [3]; and it follows from their Theorem 4.5.1' that there is an optimal stopping time  $t = t(h, \pi)$ , one which minimizes  $R_h(t; \pi)$ , for fixed  $\pi$  and  $h$ .

The next step is to examine the loss function  $L_h(n, \sigma) = NK(n/N)$ , defined in (7). Simple properties of  $K$  may be found from the derivatives

$$K'(x) = -(a/c\sqrt{x})\phi(c\sqrt{x}) + 1$$

and

$$K''(x) = (ac/2\sqrt{x})(1 + 1/c^2x)\phi(c\sqrt{x})$$

for  $x > 0$ . Thus,  $K$  is a convex function, since  $K''(x) > 0$  for all  $x > 0$ . The minimum value of  $K$  may be found by solving the equation  $K'(x) = 0$ . The solution is  $x = 1$ , in view of the definition of  $a$ . Also,  $K(1) = 0$  and  $K''(1) = (1 + c^2)/2$ .

For fixed  $\pi$  and  $h > 0$ , let  $t = t(h, \pi)$  denote the optimal stopping time. Then

$$t/N \rightarrow 1 \text{ in probability, as } h \rightarrow 0. \quad (13)$$

To see this, let  $\tau$  denote the fully sequential procedure (4) with  $b$  as in (6)

and  $m \geq 4$ . Then, by (6) (see also (10))

$$R_h(t; \pi) \leq R_h(\tau, \pi) = 0(1)$$

as  $h \rightarrow 0$ . Let  $\sigma_0 > 0$  denote the lower endpoint of the support of  $\pi$ . If  $\delta > 0$ , then

$$h^2 R_h(t; \pi) \geq c^2 \sigma_0^2 \min\{K(1 - \delta), K(1 + \delta)\} P\{|t/N - 1| > \delta\}$$

for all  $h > 0$ , so,  $P\{|t/N - 1| > \delta\} \rightarrow 0$  as  $h \rightarrow 0$ . This establishes (13), since  $\delta > 0$  was arbitrary.

Let  $\sigma_0$  and  $\sigma_1$  denote the upper and lower endpoints of the support of  $\pi$ ; let  $B_h$  be the event,

$$B_h = \{\sigma_0 < h\sqrt{t}/c < \sigma_1\}, \quad h > 0;$$

and let  $2K_0 = 2K_0(\pi) = \min\{K''(x): \sigma_1^{-2}\sigma_0^2 \leq x \leq \sigma_0^{-2}\sigma_1^2\}$ . Then  $P(B_h) \rightarrow 1$  as  $h \rightarrow 0$  by (13); and  $B_h$  implies that  $K''(x) \geq 2K_0$  for all  $x$  with  $|x - 1| < |t/N - 1|$ , w.p.1 ( $P$ ). So,

$$\begin{aligned} R_h(t; \pi) &\geq K_0 \int_{B_h} N(t/N - 1)^2 dP \\ &= K_0 c^2 h^{-2} \int_{B_h} \sigma^2 (s\sigma^{-2} - 1)^2 dP, \end{aligned} \tag{14}$$

where

$$s = h^2 t / c^2.$$

Let  $\theta = \sigma^{-2}$ ; and let  $\mathcal{D}$  denote the sigma-algebra generated by  $Y_i, i = 1, \dots, t$ . Then the conditional expectation of the integrand in (14) is

$$E\left[\theta^{-1}(s\theta - 1)^2 | \mathcal{D}\right] \geq -\text{Cov}(\theta, 1/\theta | \mathcal{D}) / E(\theta | \mathcal{D}) \tag{15}$$

by a simple minimization (with respect to  $s$ ).

Now suppose that  $\pi$  has a twice continuously differentiable density. Let  $\rho$  denote the (prior) density of  $\theta$ ; and let  $\ell$  denote the likelihood function given  $t$  and  $S_t^2 = y$ , say

$$\ell(\theta) = \theta^{(t-1)/2} \exp[-(t-1)y\theta/2]$$

for  $\theta > 0$ . Then, letting  $C = C(t, y)$  denote a normalizing constant and letting

$$\theta_t = E(\theta | \mathcal{D}) = C \int_0^\infty \theta \ell(\theta) \rho(\theta) d\theta$$



denote the conditional expectation of  $\theta$  given  $\mathcal{D}$ , the conditional covariance may be written

$$\begin{aligned} -\text{Cov}(\theta, 1/\theta|\mathcal{D}) &= -C \int_0^\infty (\theta - \theta_t) \left( \frac{1}{\theta} - y \right) \ell(\theta) \rho(\theta) d\theta \\ &= - \left( \frac{2}{t-1} \right) C \int_0^\infty \ell'(\theta) (\theta - \theta_t) \rho(\theta) d\theta \\ &= \left( \frac{2}{t-1} \right) C \int_0^\infty \ell(\theta) [\rho(\theta) + (\theta - \theta_t) \rho'(\theta)] d\theta \\ &= \left( \frac{2}{t-1} \right) \{1 + E[(\theta - \theta_t) \rho_1(\theta)|\mathcal{D}]\}, \end{aligned}$$

where  $\rho_1 = \rho'/\rho$ , by a simple integration by parts. Now  $\theta_t \rightarrow \theta$  in prob. by the Martingale Convergence Theorem; and  $E(\rho_1|\mathcal{D}) \rightarrow \rho_1$  prob. for the same reason. So,

$$(-t/2) \text{Cov}(\theta, 1/\theta|\mathcal{D}) \rightarrow 1 \quad (16)$$

in prob. as  $h \rightarrow 0$ . Since this term is non-negative,  $t/N \rightarrow 1$  in prob., and  $P(B_h) \rightarrow 1$  as  $h \rightarrow 0$ , it follows from (14), (15), and (16), and Fatou's Lemma that

$$\liminf_{h \rightarrow 0} R_h(t; \pi) \geq 2K_0. \quad (17)$$

To complete the proof, let  $\varepsilon > 0$  be given. Then there is a  $\delta > 1$  for which  $K'(x) \geq (1 - \varepsilon)K''(1)$  whenever  $1/\delta^2 < x < \delta^2$ ; and there is a finite open cover of the support of  $\pi$  by intervals whose upper and lower endpoints differ by a factor of at most  $\delta$ . Let  $k$  denote the number of intervals in this cover; let  $\gamma_i$ ,  $i = 1, \dots, k$ , be a smooth partition of unity, subordinate to the cover; and let

$$d\pi_i \propto \gamma_i d\pi, \quad i = 1, \dots, k.$$

Then

$$\pi = \sum_{i=1}^k a_i \pi_i$$

for some non-negative coefficient  $a_1, \dots, a_k$  with unit sum. Now, by construction,

$$2K_0(\pi_i) \geq (1 - \varepsilon)K''(1), \quad \text{for } i = 1, \dots, k,$$

and (17) is applicable to each  $\pi_i$ ,  $i = 1, \dots, k$ , since each has a twice

continuously differentiable density. Now,

$$\inf_t R_h(t; \pi) \geq \sum_{i=1}^k a_i \inf_t R_h(t; \pi_i);$$

for all  $h > 0$ , since  $R(t, \pi)$  is linear in  $\pi$  for each fixed  $t$ ; and

$$\liminf_{h \rightarrow 0} \inf_t R_h(t; \pi) \geq \sum_{i=1}^k 2a_i K_0(\pi_i) \geq (1 - \varepsilon) K''(1)$$

by (17) applied to  $\pi_i, i = 1, \dots, k$ . The theorem follows since  $\varepsilon > 0$  was arbitrary and  $K''(1) = (1 + c^2)/2$ .

### REFERENCES

1. F. J. ANSCOMBE, Sequential estimation, *J. R. Statist. Soc., Ser. B*, **15** (1953) 1–21.
2. Y. S. CHOW AND H. ROBBINS, On the asymptotic theory of fixed width confidence intervals, *Ann. Math. Statist.* **36** (1965), 457–462.
3. Y. S. CHOW, H. ROBBINS, AND D. SIEGMUND, “Great Expectations,” Houghton-Mifflin, Boston, 1971.
4. M. FINSTER, Estimation in the general linear model when accuracy is specified before data collection, *Ann. Statist.*, **13** (1985), 663–675.
5. P. HALL, Asymptotic theory of triple sampling for sequential estimation of the mean, *Ann. Statist.* **9** (1981), 1229–1238.
6. G. SIMONS, On the cost of not knowing the variance when making a fixed width confidence interval for the mean, *Ann. Math. Statist.* **39** (1968), 1946–1952.
7. N. STARR, The performance of a sequential procedure for fixed width confidence interval estimation of a mean, *Ann. Math. Statist.* **37** (1966), 36–50.
8. N. STARR AND M. WOODROOFE, Remarks on a stopping time, *Proc. Nat. Acad. Sci.* **61** (1968), 1215–1218.
9. C. STEIN, A two-sample test for a linear hypothesis whose power is independent of the variance, *Ann. Math. Statist.* **16** (1945), 243–258.
10. C. STEIN AND A. WALD, Sequential confidence intervals for the mean of a normal distribution with known variance, *Ann. Math. Statist.* **18** (1947), 427–433.
11. M. WOODROOFE, Second order approximations for sequential point and interval estimation, *Ann. Statist.* **5** (1977), 984–995.
12. M. WOODROOFE, “Non-Linear Renewal Theory in Sequential Analysis,” SIAM, 1982.