COMPARISONS OF SOME TWO STAGE SAMPLING METHODS[1]

By Aaron S. Goldman[2] and R. K. Zeigler

Gonzaga University and University of California, Los Alamos Scientific Laboratory

1. Introduction. The use of multistage sampling procedures has been of great value in providing a solution to the problem of estimating a parameter with a prescribed precision. There are several two-stage methods available so that either (A) the estimate of a parameter has a specified variance, or (B) a $(1 - \alpha)$ confidence interval placed on a parameter has a specified width. Of the methods available that provide a solution to (A) or (B), the techniques of Birnbaum and Healy [2] (henceforth called BH), Stein [11], and Graybill [6] appear easiest to apply. The purpose of this paper is to present a general result that holds under certain conditions for obtaining the expected sample size in Graybill's method and to compare results where feasible with the techniques of Stein and BH. A review of Graybill's theorem is given. Brief explanations of the applications of the three methods are presented when estimating the mean or the variance from a normal population.

2. The expected sample size using Graybill's method. Suppose w is the width of a confidence interval on a parameter ξ with confidence coefficient $1 - \alpha$. Suppose further that it is desired that the probability that w be less than d lie between β^2 and $2\beta - \beta^2$. The problem is to determine k, the number of observations, on which to base w.

The Graybill [6] technique will be described for a two-stage procedure. The first stage yields a random variable z from which is determined a sample size k on which to base the confidence interval of random width w. Suppose that the distribution of w depends on k and an unknown parameter θ (θ may be the parameter ξ). Suppose also there exists a function g such that the distribution of $Y = g(w; \theta, k)$ depends only on k (and not on the unknown parameter) and g is monotonic increasing in w for every k and θ. Then a function $f(k)$ may be obtained so that $P[Y < f(k)] = \beta$; $0 < \beta < 1$. Let the solution for $g(w; \theta, k) = f(k)$ for w be $w = h(\theta, k)$ such that $h(\theta, k)$ is monotonic increasing for every k and monotonic decreasing in k for every θ.

Let n be defined as a random variable such that $h(t(z), n) = d$; consequently k is the smallest positive integer such that $k \geqq n$ and $h(t(z), k) \leqq d$. Then the following inequality is true:

$$\beta^2 \leqq P(w \leqq d) \leqq 2\beta - \beta^2.$$

At this point an expression for $E(k)$ shall be presented.

Received 22 March 1965; revised 15 February 1966.

[1] Work performed under the auspices of the U. S. Atomic Energy Commission.

[2] A part of this paper was written while A. Goldman was sponsored by a National Science Foundation fellowship at Oklahoma State University.

It is readily seen that:

$$(2.1)\quad E(k) = \sum_{i=1}^{\infty} iP\{k = i\} = \sum_{i=1}^{\infty} P\{k \geqq i\} = 1 + \sum_{i=1}^{\infty} P\{n > i\}.$$

Assume that $h[t(z), n] = d$ can be solved for z by $z = f_1(n)$ where $f_1(n)$ is monotonic increasing in n. Then $P\{n > i\} = P\{z > f_1(i)\}$ and if z has the probability density $g(z)$,

$$(2.2)\quad E(k) = 1 + \sum_{i=1}^{\infty} \int_{f_1(i)}^{\infty} g_1(z)\, dz.$$

This expectation could diverge; therefore the following sufficient conditions for convergence are also listed:

(2.3) (a) for some $s \geqq 2$, $g_1(z) = O(z^{-s})$ as $z \to \infty$.

(b) for some i_0, $f_1(i_0) > 0$ and $\sum_{i=i_0}^{\infty} [f_1(i)]^{-(s-1)}$ converges.

3. Procedures for estimating the mean. Stein's [11] classic technique is applicable to estimating the mean of a normal population so that a $(1 - \alpha)$ confidence interval has width less than or equal to "d" specified units. Briefly, the procedure is as follows:

Select a sample of size m and place a confidence interval on μ in the usual way when σ is unknown. If the interval width is less than d units, then a solution is obtained in just one step; however, if the width is greater than d, then a sample of size n is needed. The value of k is determined by the smallest integer value $k \geqq n$ such that

$$(3.1)\quad k \geqq 4s_1^2 t_{\alpha/2}^2(m)/d^2 - m;$$

where the upper γ points of Student's t distribution with $(v - 1)$ degrees of freedom are denoted as $t_\gamma(v)$. The confidence interval on μ is given by the quantity in the brackets:

$$P[\bar{x}_c - t_{\alpha/2}(m)s_1/(k + m)^{\frac{1}{2}} \leqq \mu \leqq \bar{x}_c + t_{\alpha/2}(m)s_1/(k + m)^{\frac{1}{2}}] \geqq 1 - \alpha;$$

where $\bar{x}_c$ is the overall sample mean, and s_1^2 is the variance of the first sample.

Graybill's method is used in determining a $(1 - \alpha)$ confidence interval on the mean such that $\beta^2 \leqq P(w \leqq d) \leqq 2\beta - \beta^2$ and

$$(3.2)\quad P[\bar{x}_2 - t_{\alpha/2}(n)s_2/n^{\frac{1}{2}} \leqq \mu \leqq \bar{x}_2 + t_{\alpha/2}(n)s_2/n^{\frac{1}{2}}] = 1 - \alpha,$$

where β, d, and α are specified, and $\bar{x}_2$ and s_2 are the mean and standard deviation of the second sample.

The method is as follows: Choose an initial sample of size m and compute s_1^2. Then k is the smallest integer value $k \geqq n$ such that

$$(3.3)\quad k(k - 1)/t_{\alpha/2}^2(k)\chi_{1-\beta}^2(k) \geqq 4s_1^2(m - 1)/\chi_\beta^2(m)\, d^2.$$

4. The expected sample size for the mean. Seelbinder [10] has provided tables for finding expected values of $m + n$ for various values of d/σ in Stein's method. A portion of these results may be found in Table I.

TABLE I

Expected total sample size under Graybill's and Stein's methods to obtain desired width 95% confidence intervals on the mean of a normal population

	d/σ									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$m = 61$										
S	400	100	62	61						
$G_{0.90}$	2122	597	309	207						
$G_{0.95}$	2297	648	334	223						
$G_{0.99}$	2677	756	388	256						
$m = 51$										
S	403	101	53	52						
$G_{0.90}$	2169	602	307	201						
$G_{0.95}$	2364	657	333	218						
$G_{0.99}$	2799	780	394	255						
$m = 41$										
S	408	102	48	42	41					
$G_{0.90}$	2240	613	306	197	145					
$G_{0.95}$	2470	677	337	215	159					
$G_{0.99}$	2984	820	407	258	187					
$m = 31$										
S	417	104	47	32	31					
$G_{0.90}$	2355	635	311	195	141					
$G_{0.95}$	2639	713	349	218	156					
$G_{0.99}$	3295	893	436	271	192					
$m = 21$										
S	435	109	49	29	22	21				
$G_{0.90}$	2583	686	328	201	141	108				
$G_{0.95}$	2965	789	378	230	161	122				
$G_{0.99}$	3941	1052	502	305	211	157				
$m = 11$										
S	496	124	56	32	21	16	13	11		
$G_{0.90}$	3269	853	398	236	160	118	93	76		
$G_{0.95}$	4055	1059	494	293	198	146	113	92		
$G_{0.99}$	6263	1637	762	450	302	220	169	134		
$m = 6$										
S	661	165	74	42	47	19	14	11	10	8
$G_{0.90}$	4900	1262	579	337	224	161	123	99	81	69
$G_{0.95}$	6869	1769	811	471	312	224	161	136	111	94
$G_{0.99}$	14341	3684	1683	973	640	457	346	262	221	184

The expected value of n using Graybill s technique may be approximated by using the results of Section 3.

From Equation (3.3)

$$z \leqq [k(k-1)/t^2_{\alpha/2}(k)\chi^2_{1-\beta}(k)]\,[\chi_\beta^2(m)\, d^2/4] = f_1(k); \tag{4.1}$$

where $z = s_1^2(m-1)$.

The function $g_1(z/\sigma^2)$ is the chi-square distribution with $(m - 1)$ degrees of freedom. Let z^* denote the mode. Then $g_1(z)$ is monotonically decreasing for $z > z^*$. For $s = 3$, $.8 < \beta < 1$, conditions in Equation (2.3) are satisfied and

$$(4.2) \qquad E(n) \approx E(k) = 1 + \sum_{i=1}^{\infty} \int_{f_1(i)/\sigma^2}^{\infty} [g_1(z/\sigma^2)\, d(z/\sigma^2)].$$

Computations of some values of $E(n)$ for choices of d/σ and m are given in Table I. Graybill's method is denoted by G_β where β is the width coefficient and Stein's method is denoted by S. Table I represents a $1 - \alpha = 0.95$ confidence interval. Each term in Equation (4.2) was summed for all values of the integral greater than 10^{-17}.

5. Procedures for estimating the variance. It has been demonstrated [7] that Graybill's method may be applied to obtaining a $(1 - \alpha)$ confidence interval for the variance of a normal distribution such that

$$(5.1) \qquad \beta^2 \leqq P(w \leqq d) \leqq 2\beta - \beta^2$$

and

$$(5.2) \qquad P(s_2^2/\chi^2_{\alpha/2}(n) \leqq \sigma^2 \leqq s_2^2/\chi^2_{1-\alpha/2}(n)) = 1 - \alpha.$$

The procedure in obtaining a value of n so Equation (5.1) holds is as follows: Choose a sample of size m and compute s_1^2. Then k is the smallest integral value of n so that

$$(5.3) \qquad \chi^2_{1-\beta}(k)\{[\chi^2_{1-\alpha/2}(k)]^{-1} - [\chi^2_{\alpha/2}(k)]^{-1}\} \geqq d\chi_\beta^2(m)/s_1^2(m - 1).$$

BH's method is directly applicable in finding $\hat{\sigma}^2$, an estimate of the variance of a normal distribution such that the estimate has a specified variance, $B_{\hat{\sigma}^2}$. A sample of size $m > 5$ is chosen and s_1^2 is computed. Then k is the smallest integer value of n so that

$$(5.4) \qquad k \geqq 2s_1^4(m - 1)^2/B_{\hat{\sigma}^2}(m - 3)(m - 5) + 1.$$

Because the two methods are used differently, a comparison is rather difficult. The comparison will be based upon utilizing a confidence interval approach in BH's method. A rather crude confidence interval may be developed using Tchebycheff's inequality.

$$(5.5) \qquad P(s_2^2 - d/2 \leqq \sigma^2 \leqq s_2^2 + d/2) \geqq 1 - 4B_{\hat{\sigma}^2}/d^2.$$

Let $\alpha = B_{\hat{\sigma}^2}/d^2$. Then Equation (5.4) may be written

$$(5.6) \qquad k \geqq 8s_1^4(m - 1)^2/\alpha d^2(m - 3)(m - 5).$$

6. The expected sample size for the variance. In BH's method the expected sample size has been found [2] for Equation (5.6) and is expressed as:

$$(6.1) \quad E(n) = 1 + [8(m + 1)(m - 1)/(m - 3)(m - 5)\alpha](\sigma^2/d)^2.$$

By using a technique that is analogous to that given in Section 4, the expected sample size for Graybill's procedure may be obtained.

From Example 1 in Reference [6]

$$\chi_{1-\beta}^2(k)G(k)z/\chi_\beta^2(m) \leqq d, \tag{6.2}$$

where $G(k) = [\chi_{1-\alpha/2}^2(k)]^{-1} - [\chi_{\alpha/2}^2(k)]^{-1}$. Solving for z,

$$z = d\chi_\beta^2(m)/\chi_{1-\beta}^2(k)G(k) = f_1(k). \tag{6.3}$$

As noted earlier z/σ^2 is distributed as the chi-square distribution with $(m - 1)$ degrees of freedom.

Let $s = 4$ in (2.3a). The convergence of

$$\sum_{i=i_0}^\infty [f_1(i)]^{-4} \tag{6.4}$$

can be demonstrated by substituting Fisher's approximation [8] for the chi-square deviates in $f_1(i)$ and examining the summation of terms involving k for convergence. Using Fisher's approximation and ignoring the constants d and $\chi_\beta^2(m)$, Equation (6.4) may be written

$$\sum_{i=i_0}^\infty \{[(2i-1)^{\frac{1}{2}} - v_{1-\beta}]^2[((2i-1)^{\frac{1}{2}}v_{1-\alpha/2})^{-2} - ((2i-1)^{\frac{1}{2}} + v_{1-\alpha/2})^{-2}]\}^4; \tag{6.5}$$

where v_γ is the upper γ point of the normal distribution. Gauss' test may be used to show that Equation (6.5) converges. The convergence of (6.4) may be demonstrated as follows. As $i \to \infty$, the numerator and denominator of Equation (6.5) are both of order $O(i^{-\frac{1}{2}})$. By the fact that the difference of two chi-square fractiles increase in proportion with $i^{\frac{1}{2}}$ (see page 295 of Reference [9]) and $\chi_{1-\beta}^2(i)$ is of order i then Equation (6.4) is of order $O(1)$.

Thus

$$E(n) \approx E(k) = 1 + \sum_{i=1}^\infty \int_{[f_1(i)]/\sigma^2}^\infty g_1(z/\sigma^2)\, d(z/\sigma^2) \tag{6.6}$$

where $f_1(i) = d\chi_\beta^2(m)/\chi_{1-\beta}^2(i)G(i)$.

Table II compares the expected sample size of the two described methods for estimating the variance with a desired width confidence interval. In the table G_β denotes Graybill's technique for $\beta = 0.90, 0.95, 0.99$ and B denotes Birnbaum and Healy's method. Values of $1 - \alpha$ are 0.90, 0.95, and 0.99. Computations of the sum in Equation (6.6) were carried out for all terms greater than 10^{-15}.

7. Summary. A perusal of Tables I and II would indicate that Stein's method is far superior to Graybill's for the mean whereas the latter technique is better than BH's for variance. One possible occasion when Graybill's method might be utilized occurs when there is a change of variance and a more precise confidence coefficient is desired. Stein's method relies solely on s^2 computed on the first step of the two stages. If a second sample is necessary and if σ^2 should change, then the value of α would be incorrect. On the other hand, Graybill's method relies on s^2 calculated on the second sample; therefore the confidence coefficient remains exact. The width coefficient, β, equals 1 in Stein's method but is unknown in Graybill's technique when the change of variance takes place. A time delay between samples could possibly result in a change of variance of the population.

TABLE II

Expected size of second sample under Graybill's and BH's methods for desired width confidence interval on the variance of a normal population

	d/σ^2								
	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	3.0
			$1-\alpha=0.99$						
$m=21$									
B	4890	3397	2496	1911	1510	1224	545	307	136
$G_{0.90}$	206	152	118	96	80	68	39	27	18
$G_{0.95}$	272	201	156	126	105	90	50	35	23
$G_{0.99}$	469	344	267	216	179	152	84	57	35
$m=31$									
B	4221	2932	2154	1650	1304	1056	470	265	119
$G_{0.90}$	170	126	99	80	67	58	34	24	16
$G_{0.95}$	215	160	125	102	85	73	42	30	20
$G_{0.99}$	334	248	194	158	132	113	64	45	29
$m=61$									
B	3667	2547	1871	1433	1133	918	409	231	103
$G_{0.90}$	136	102	80	65	55	48	29	21	15
$G_{0.95}$	163	122	96	79	67	58	35	25	17
$G_{0.99}$	224	168	132	108	91	79	47	33	22
			$1-\alpha=0.95$						
$m=21$									
B	979	681	500	383	303	246	110	63	29
$G_{0.90}$	126	93	73	60	50	43	25	18	12
$G_{0.95}$	168	125	98	80	67	57	33	24	15
$G_{0.99}$	291	216	169	137	115	98	56	39	24
$m=61$									
B	735	511	375	288	228	185	83	47	22
$G_{0.90}$	85	64	51	42	36	31	19	14	10
$G_{0.95}$	102	77	62	51	44	38	23	17	12
$G_{0.99}$	143	108	86	71	61	54	33	24	16
			$1-\alpha=0.90$						
$m=21$									
B	502	349	257	197	157	127	57	33	15
$G_{0.90}$	93	69	55	45	38	33	26	14	10
$G_{0.95}$	124	93	73	60	51	44	33	19	12
$G_{0.99}$	216	161	127	104	87	75	43	30	20
$m=61$									
B	114	79	59	45	36	30	14	9	5
$G_{0.90}$	63	48	38	32	27	24	15	11	8
$G_{0.95}$	77	59	47	39	34	29	18	14	10
$G_{0.99}$	108	83	66	55	47	42	26	19	13

In general though, Stein's method appears to be preferable. Table II describes BH's method as being inferior to Graybill's; however, BH utilizes a conservative confidence interval that lends itself to a larger sample size. The simplicity in applying BH's method should be considered an advantage.

Other methods that could be compared with these techniques include that of

Cox [4] and Anscombe [1]. A general article on existence theorems has been given by Blum and Rosenblatt [3].

The authors are indebted to F. A. Graybill, R. H. Moore, and the reviewers for their assistance in this work.

REFERENCES

[1] ANSCOMBE, F. J. (1953). Sequential estimation. *J. Roy. Statist. Soc. Ser. B* **15** 1–29.

[2] BIRNBAUM, A. and HEALY, W. C., JR. (1960). Estimates with prescribed variance based on two-stage sampling. *Ann. Math. Statist.* **31** 662–76.

[3] BLUM, J. R. and ROSENBLATT, J. (1963). On Multistage estimation. *Ann. Math. Statist.* **34** 1452–58.

[4] COX, D. R. (1952). Estimation by double sampling. *Biometrika* **39** 217–27.

[5] GOLDMAN, A. (1963). Sample size for a specified width confidence interval on the ratio of variances from two independent normal populations. *Biometrics* **19** 465–77.

[6] GRAYBILL, F. (1958). Determining sample size for a specified width confidence interval. *Ann. Math. Statist.* **29** 282–87.

[7] GRAYBILL, F. A. and MORRISON, R. (1960). Sample size for a specified width interval on the variance of a normal distribution. *Biometrics* **16** 636–41.

[8] HALD, ANDERS (1952). *Statistical Tables and Formulas.* Wiley, New York.

[9] KENDALL, MAURICE (1947). *The Advanced Theory of Statistics*, **1** (3rd edition). Griffin, London.

[10] SEELBINDER, B. M. (1953). On Stein's two-stage sampling scheme. *Ann. Math. Statist.* **24** 640–47.

[11] STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16** 243–58.