ON STEIN'S TWO-STAGE SAMPLING SCHEME

By B. M. SEELBINDER

University of North Carolina

Summary. This paper gives a method for determining the size of the first part of a two-stage sample for estimating the population mean with a given accuracy. The method is based on a scheme of Stein [4]. The tables necessary for application of this method have been given. A more detailed summary will be found at the end of the Introduction.

1. Introduction. When it is desired to investigate the characteristics of a specified population on the basis of a sample, the size of the sample depends on the accuracy which it is desired to attain. Thus, in order to estimate the mean or average of the population we may fix an allowable discrepancy d, and a risk or significance level α , such that the chance of the absolute difference between the true mean and the estimate exceeding the allowable discrepancy d is not greater than α . Thus

$$(1) P\{|T-m| \ge d\} \le \alpha,$$

where m is the true mean and T is its estimate.

One approach to the problem of sample size is to use a two-stage sampling plan, the size of the second part of the sample depending on the information supplied about the variance of the population by the first part of the sample. Stein has suggested such a two-stage plan, but in his work the size n_1 of the first part of the sample is left to the discretion of the experimenter. If n is the total size of the sample (including both parts), then the expected value of n depends on n_1 . It would thus be worthwhile to have some clues for the proper determination of n_1 . In this connection Cochran [1] states:

"The average value of n that is required in a given situation depends on the choice of n_1 . Exact information about the optimum value of n_1 is not yet available, the optimum being that value which leads to the *smallest* average n. It appears however that the optimum n_1 would be such that a second part will usually be necessary. In other words, if it is convenient to take the sample in two parts, n_1 should be chosen as somewhat less than the size that seems to be needed."

It is the object of this study to throw further light on the choice of the value of n_1 , the size of the first part of the sample. For this purpose we first compute the expected value of the total sample size for given n_1 , when α and $c = d/\sigma$ are given. The necessary formulae for computing tables of these expected values are derived.

www.jstor.org

Received 2/21/51, revised 4/6/53.

¹ Part of a Master's thesis presented to the Department of Mathematical Statistics, University of North Carolina.

It was shown by Stein that the computation can be made to depend on the knowledge of Pearson's Incomplete Gamma Function I(u, p). An approximation whereby the computation can be made to depend only on the knowledge of the normal distribution function $\Phi(L)$ is also developed. Numerical evidence for the adequacy of the approximation for moderately large values of $n_1(n_1 \ge 61)$ is adduced. Limiting values for the expected value of the total sample size E(n) are given for fixed n_1 and α with varying c.

The use of the tables in choosing n_1 is discussed. If we have an approximate knowledge of σ , enabling us to fix an interval in which σ might be assumed to lie, then the value of n_1 can be determined with the help of the tables by using the minimax principle (i.e., minimizing the maximum loss in observations due to ignorance of σ). Reasons are given which point to 250 as a practical upper limit for the size of the first part of the sample in a two-stage sampling scheme when we have no precise knowledge of σ . Tables for four different significance levels of α are included at the end of the paper.

2. Stein's method. Stein's two-stage plan for estimating the population mean may be stated as follows. Given d and α , we start with a random sample of size n_1 . An estimate of the population variance σ^2 is given by the sample variance

(2)
$$s_1^2 = \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 / n_0,$$

where $n_0 = n_1 - 1$ and \bar{x}_1 is the sample mean. A confidence interval for m may now be calculated. The half-width of this interval is given by

$$s_1 t / \sqrt{n_1}$$

where $t = t(\alpha, n_0)$ is the value of t corresponding to the given significance level α , for n_0 degrees of freedom. If

$$(4) s_1 t / \sqrt{n_1} \le d,$$

the sample is already sufficiently large and we stop here, saying that the estimate T of m is given by $T = \bar{x}_1$. If

$$(5) s_1 t / \sqrt{n_1} > d,$$

then additional observations are taken so that the total sample size is the smallest integer not less than n, where n is given by

(6)
$$n = s_1^2 t^2 / d^2.$$

The estimate of m is now given by $T = \bar{x}$, where \bar{x} is the mean of the total sample.

When this procedure is adopted (1) is satisfied. Thus given n_1 , α and d, if the event (4) occurs then the total sample size is n_1 . On the other hand, if (5) occurs, we shall make our calculations as if the total sample size were n given by (6), neglecting the small discrepancy introduced by the fractional nature of n.

3. Expected sample size E(n). Now

$$\chi^2 = n_0 s_1^2 / \sigma^2$$

obeys the χ^2 distribution with n_0 degrees of freedom. Setting $d^2/\sigma^2 = c^2$, we see that when (4) happens

$$(8) 0 < \chi^2 \le c^2 n_0 n_1 / t^2,$$

and $n = n_1$. When (5) happens

(9)
$$\chi^2 > c^2 n_0 n_1 / t^2,$$

and $n=t^2\chi^2/c^2n_0$. Recalling that the χ^2 distribution with n_0 degrees of freedom is

(10)
$$f(\chi^2, n_0) d\chi^2 = \frac{e^{-\chi^2/2} (\chi^2)^{n_0/2-1}}{2^{n_0/2} \Gamma(n_0/2)} d\chi^2,$$

we may write

(11)
$$E(n) = \int_0^\infty nf(\chi^2, n_0) \ d\chi^2 = \int_0^{\chi_0^2} n_1 f(\chi^2, n_0) \ d\chi^2 + \int_{\chi^2}^\infty \frac{t^2 \chi^2}{c^2 n_0} f(\chi^2, n_0) \ d\chi^2,$$

where $\chi_0^2 = c^2 n_0 n_1/t^2$. Evaluating the second integral above by parts we have

(12)
$$E(n) = n_1 \int_0^{x_0} f(\chi^2, n_0) d\chi^2 + \frac{t^2}{c^2} \int_{\chi_0^2}^{\infty} f(\chi^2, n_0) d\chi^2 + K,$$

where

(13)
$$K = \frac{(\chi_0^2/2)^{n_0/2}}{\frac{n_0}{2} \Gamma(n_0/2)e^{\chi_0^2/2}}.$$

Let

(14)
$$F(\chi_0^2) = \int_0^{x} f(\chi^2/n_0) d\chi^2.$$

Now Karl Pearson's Incomplete Gamma Function is defined as

(15)
$$I(u, p) = \int_0^{u\sqrt{p+1}} \frac{e^{-\nu_p p}}{\Gamma(p+1)} d\nu.$$

Putting $\nu = \chi^2/2$, $p = n_0/2 - 1$, $u = \chi_0^2/\sqrt{2n_0}$, we see that

(16)
$$F(\chi_0^2) = I(\chi_0^2/\sqrt{2n_0}, n_0/2 - 1).$$

Using (16) in (12) we have

(17)
$$E(n) = \left[(n_0 + 1) - \frac{t^2}{c^2} \right] F(\chi_0^2) + \frac{t^2}{c^2} [1 + K].$$

(This formula may be compared with formula (16), p. 247 of Stein's paper.)

4. Normal approximation for expected sample size. For large values of n_1 we may use the fact that

$$z = \sqrt{2\chi^2} - \sqrt{2n_1 - 3}$$

is asymptotically normally distributed with zero mean and unit variance. We shall discuss later, on the basis of the evidence provided in the numerical calculations, what constitutes a sufficiently large value of n_1 for the approximation derived below to hold.

When (4) happens, (8) may be written

$$(18) -\infty \leq -\sqrt{2n_1-3} \leq z \leq L,$$

where

(19)
$$L = \frac{c}{t} \sqrt{2n_0 n_1} - \sqrt{2n_1 - 3}.$$

The sample size is in this case $n = n_1$.

When (5) holds, (9) may be written

$$(20) z > L.$$

Thus the sample size in this case is

$$n = \frac{t^2}{2c^2n_0} \left(z + \sqrt{2n_1 - 3}\right)^2.$$

Hence the expected sample size can be written

(21)
$$E(n) = \frac{n_1}{\sqrt{2\pi}} \int_{-\infty}^{L} e^{-z^2/2} dz + \frac{1}{\sqrt{2\pi}} \int_{L}^{\infty} \frac{t^2}{2c^2 n_0} (z + \sqrt{2n_1 - 3})^2 e^{-z^2/2} dz,$$

where in the first integral in (21), the lower limit has been put as $-\infty$ instead of $-\sqrt{2n_1-3}$. For moderately large values of n_1 , say $n_1 > 10$, this will cause only a negligible difference. On evaluating the second integral in (21) we get

(22)
$$E(n) = \left\lceil (n_0 + 1) - \frac{t^2}{c^2} \right\rceil \Phi(L) + \frac{t^2}{c^2} \left\lceil 1 + \frac{L + 2\sqrt{2n_0 - 1}}{2n_0 e^{L^2/2} \sqrt{2\pi}} \right\rceil,$$

where as usual

$$\Phi(L) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{L} e^{-z^2/2} dz.$$

5. Make-up of tables and use of normal approximation. Formulae (17) and (22) were used to determine E(n) for assigned values of $n_0 = n_1 - 1$ with four different significance levels $\alpha = .01$, .02, .05 and .10 and values of $c = d/\sigma$ ranging from .01 to 1.0. For each combination of c and α , a value of E(n) is listed in the tables. The size of a sufficiently large value of n_0 for the normal approximation to be valid was determined on the basis of the numerical calcula-

tions. For $10 \le n_0 \le 60$, the values of E(n) were computed both by the χ^2 method and by using the normal approximation, which served primarily as a check in this interval. A comparison for the results for E(n) is given below for a portion of the data as computed for Table IIB.

	α/σ										
		.1		.2		.3		.4		.5	
		χ²	N	χ²	N	χ²	N	χ2	N	χ²	N
n_0	60 30 20	416.98	400.00 416.98 435.14	104.25	100.04 104.28 108.78	46.65	46.71	32.02		31.01	30.94

Expected sample size for $\alpha = .05$

A similar comparison table for $\alpha = .01$ was computed and the same degree of agreement was found to exist. In no instances did any pair of values of E(n) computed by (17) and (22) differ by more than one observation. In fact, for $n_0 = 60$, the two computed values of E(n) were so close as to warrant use of only the normal approximation when $n_0 > 60$ (for all four significance levels). Figures in Tables I to IV are correct to the penultimate digit. The last digit shown may be in error by unity (e.g., a value listed as 20.8 might actually be 20.7 or 20.9).

6. Limiting values of E(n)**.** Let us consider the value of E(n) given by (17). For fixed n_0 and α , t is fixed. As c^2 increases, both χ_0^2 and $F(\chi_0^2)$ increase. Differentiating E(n) with respect to χ_0^2 , we get

(23)
$$E'(n) = \frac{n_0(n_0+1)}{\gamma_0^4} [F(\chi_0^2) - (1+K)].$$

Now $F(\chi_0^2) \leq 1$ and $K \geq 0$. Hence E'(n) is negative, which shows that E(n) is a monotonically decreasing function of χ_0^2 or of c^2 for fixed n_0 and α . With t fixed, as c^2 increases $F(\chi_0^2) \to 1$ and $K \to 0$, causing E(n) to approach $n_0 + 1$. It should be noted in the tables that for any row the value of E(n) for increasing c has been given until it sensibly becomes equal to $n_0 + 1$. In the blank space left thereafter in the Tables IB, IIIB, IIIB and IVB values of E(n) will sensibly remain equal to $n_0 + 1$, since obviously,

$$(24) E(n) \ge n_0 + 1.$$

Let us set

(25)
$$E(n) - t^2/c^2 = \phi(\chi_0^2).$$

Integrating (14) by parts we find

(26)
$$F(\chi_0^2) = K + \frac{1}{n_0} \int_0^{\chi_0^2} \frac{e^{-\chi^2/2} (\chi^2)^{n_0/2}}{2^{n_0/2} \Gamma(n_0/2)} d\chi^2.$$

Hence

(27)
$$\phi(\chi_0^2) = (n_0 + 1) \int_0^{\chi_0^2} f(\chi^2, n_0) \left[1 - \frac{\chi^2}{\chi_0^2}\right] d\chi^2.$$

Also,

(28)
$$\phi'(\chi_0^2) = E'(n) + \frac{n_0(n_0+1)}{\chi_0^4} = \frac{(n_0+1)}{\chi_0^4} \int_0^{\chi_0^2} \frac{e^{-\chi^2/2}(\chi^2)^{n_0/2}}{2^{n_0/2}\Gamma(n_0/2)} d\chi^2.$$

Now $0 \le \chi^2 \le \chi_0^2$. Hence $0 \le (1 - \chi^2/\chi_0^2) \le 1$. It follows from (27) that

(29)
$$0 \leq \phi(\chi_0^2) \leq (n_0 + 1)F(\chi_0^2).$$

Hence $\phi(\chi_0^2)$ is positive and tends to zero as $\chi_0^2 \to 0$. Thus

$$(30) E(n) \ge t^2/c^2,$$

and

(31)
$$\lim_{n \to \infty} \left[E(n) - t^2/c^2 \right] = 0.$$

Also from (28), $\phi'(\chi_0^2)$ is positive, which shows that $E(n) - t^2/c^2$ monotonically decreases as c decreases. In calculation of the tables we have used the fact that when E(n) sensibly becomes equal to t^2/c^2 for any value of c, it will remain so for smaller values of c (n_0 and α remaining fixed).

7. Use of tables for a two-stage scheme. When an approximate estimate of σ is available, and the sample size is determined by one-stage sampling, the mean eventually determined may have less accuracy than what is desired, for it may turn out that the estimate of σ is in error. This situation is avoided by using Stein's two-stage plan. Our Tables I, II, III and IV give a good guidance for choosing a suitable size for the first part of the sample when used in conjunction with the minimax principle (i.e., minimizing the maximum loss in observations due to ignorance of σ). Of course, the tables allow for only four significance levels, namely $\alpha = ... 1, .05, .02$ and .01, but these are likely to suffice in practice.

Let $E(n \mid n_1)$ denote the value of E(n) corresponding to n_1 . For each value of c in the tables we have a minimum value for $E(n) = t_{\infty}^2/c^2 = E(n \mid t_{\infty}^2/c^2)$, which is the total sample size if σ were known. Let $D = E(n \mid n_1) - E(n \mid t_{\infty}^2/c^2)$. This difference D represents the loss in observations brought about by our ignorance of σ .

If it is believed or assumed that σ lies within a certain interval, say $\sigma_1 \geq \sigma \geq \sigma_2$, we can calculate, for fixed d, an interval for c, say $c_1 \leq c \leq c_2$. In choosing a starting sample size n_1 , we want to select that value of n_1 which gives us an optimum D, that is, that value of D which causes us to lose the least number of observations.

Let n_1^* be the value of n_1 corresponding to the optimum D. The following procedure may be employed to find an optimum D and n_1^* . For each value of c in our guessed interval there is a smallest $E(n) = E(n \mid n_1)$ and its corre-

sponding n_1 . We take these values of n_1 along with our values for c and tabulate values for d. From this tabulation we select the two n_1 values which would lead to the smallest value of d0 over the interval for d0. Using interpolation between these two values of d1 we arrive at an optimum d2 and d1.

Thus $n_1 = [n_1^*]$, where $[n_1^*]$ indicates the smallest integer $\ge n_1^*$, gives us the first part of our sample. After the estimated variance s_1^2 from this part of the sample has been calculated, we proceed to take the second part of our sample.

The above discussion may be illustrated by considering a specific example based on data by Cochran [1]. Suppose d=10, $\alpha=.05$ and σ is believed to lie in the interval $100 \ge \sigma \ge 25$. Thus $.1 \le c \le .4$. In Table II for c=.1, .2, .3 and .4 we find that the n_1 corresponding to the smallest $E(n \mid n_1)$ for these values of c are $n_1=241$, 61, 31 and 21. Using these values of n_1 with the values of c, we get the following tabulated values of D.

Starting sample		D									
$\mathrm{size}\ n_1$	c = .1	c = .2	c = .3	c = .4							
241	4	145	198.3	217							
61	16	4	18.3	37							
$n_1^* = 47.3$	23.7	5.8	11.7	23.7							
31	33	8	3.8	7.9							
21	51	13	5.3	4.1							

From the above figures we see that if we started with $n_1 = 61$ we might expect to lose no more than 37 observations, or if $n_1 = 31$ we might expect to lose no more than 33 observations. Interpolating between these two values of n_1 , we find that the optimum D is 23.7 and $n_1^* = 47.3$. Thus $n_1 = 48$ should be our starting sample size and should cause us to lose no more than 24 observations if σ lies between 100 and 25.

Thus when the sample is to be taken in two parts and we can assume an interval for σ , use of the minimax principle in conjunction with our tables is recommended as the procedure to be adopted in selecting a starting sample size.

8. An upper limit for n_1 . The minimax principle is very satisfactory if the true value of σ lies within our assumed interval. However, if we do not feel safe in assuming an interval for σ , we still want to limit our loss in observations if at all possible.

An examination of the tables reveals that if $c \leq .1$ or $\sigma \geq 10d$, then for all four significance levels the expected sample size for $n_1 = 241$ differs from the minimum E(n) given by t_{∞}^2/c^2 by comparatively few readings. For fixed α , n_1 and c, the value of E(n) could never be smaller than the value given by t_{∞}^2/c^2 . If we define percentage loss as $[E(n \mid n_1) - E(n \mid t_{\infty}^2/c^2)]/E(n \mid n_1)$, we see that this ratio is less than or equal to .02 for all four significance levels when

 $c \le .1$ and $n_1 = 241$. Since $E(n \mid n_1 = 241) > E(n \mid n_1 > 241)$ our percentage loss could never be greater than .02 for $n_1 > 241$ and $c \le .1$.

If we have no precise knowledge of σ we could go far wrong in choosing n_1 . For example, suppose $\alpha = .10$ and we choose $n_1 = 2400$. Using formula (22) we can compute $E(n \mid n_1 = 2400)$ for any given c. Let us consider the following tabulation.

Expected sample size for $\alpha = .10$

					<i>c</i>							
		.01	.02	.03	.04	.05	.06	.07	.08	.09	.1	.2
n_1	$t_{\infty}^{2}/c^{2} \ 2400 \ 240$	27060 27090 27290	6772	3010	2400	2400	2400	2400	2400		2400	67.6 2400 241

We see from the above figures that if the true σ leads to a value of $c \leq .03$, $n_1 = 2400$ would be a slightly better starting sample size than $n_1 = 241$. But, if the true σ leads to a value of c > .03, $n_1 = 241$ is far more efficient than $n_1 = 2400$ since our values of $E(n \mid n_1 = 241)$ are considerably smaller than the $E(n \mid n_1 = 2400)$.

There is of course no peculiar virtue in the precise number 241 and we may round off our figures to 250 stating the following rule. When using Stein's two stage sampling scheme and the value of σ is uncertain, but there is reason to believe that σ is not so small as to make $c=d/\sigma$ appreciably greater than .1, then the size of the first part of the sample should be taken to be 250 or thereabouts. Thus we may regard 250 as a sort of practical upper limit for the size of the first part of our sample.

I am indebted to Professor R. C. Bose under whose guidance this research was carried out.

REFERENCES

- [1] W. G. Cochran, "Sample Survey Techniques," Notes prepared by the Institute of Statistics at North Carolina State College, 1948.
- [2] R. A. FISHER AND F. YATES, Statistical Tables for Biological, Agricultural and Medical Research, 3rd ed., Hafner Publishing Co., 1948.
- [3] KARL PEARSON, Tables of the Incomplete Γ-Function, Cambridge University Press, 1946.
- [4] Charles Stein, "A two-sample test for a linear hypothesis whose power is independent of the variance," Ann. Math. Stat., Vol. 16 (1945), pp. 245-258.
- [5] J. V. USPENSKY, Introduction to Mathematical Probability, McGraw-Hill Book Co., 1937.

(Tables I to IV follow)

TABLE IA Expected sample size for $\alpha = .10$

 d/σ

	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
n_0 240	27,060 27,290 27,500	6,822	3,032	1,706	1,092	758	557		337	271 273 275

TABLE IB

 d/σ

	-	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
	240	273	241		***************************************						
	120	275	121								
	80	277	81.0								
	60	279	71.6	61.0							
n_0	50	281	70.6	51.0							
	40	284	71.0	41.4	41.0						
	30	288	72.0	34.8	31.0						
	20	298	74.4	33.5	22.4	21.0					
	10	328	82.1	36.5	20.8	14.6	12.0	11.2	11.0		
	5	406	102	45.1	25.4	16.2	11.4	8.8	7.3	6.6	6.2
	t_{∞}^2/c^2	271	67.6	30.1	16.9	10.8	7.5	5.5	4.2	3.3	2.7

TABLE IIA

Expected sample size for $\alpha = .05$

 d/σ

	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
n_0	38,410 38,810 39,200									

TABLE IIB

 d/σ

		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
	240	388	241								
	120	392	121			·					
	80	396	101	81.0							
	60	400	100	61.1	61.0						
ι_0	50	403	101	52.6	51.0						
	40	408	102	47.9	41.1	41.0					
	30	417	104	46.6	32.0	31.0					
	20	435	109	48.4	28.2	21.9	21.0				
	10	496	124	55.1	31.1	20.2	15.1	12.4	11.0		
	5	661	165	73.4	41.3	46.4	18.5	13.9	10.9	9.1	7.9
	t_{∞}^2/c^2	384	96	42.7	24.0	15.4	10.7	7.8	6.0	4.7	3.8

TABLE IIIA Expected sample size for $\alpha = .02$

			•		d/d	т					
		.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
	t_{∞}^2/c^2	54,100	13,520	6,011	3,381	2,164	1,503	1,104	845	668	541
n_0	240	54,850	13,712		3,428	2,194	1,524	1,119	857	677	548
	120	55,600		6,178	3,475	2,224	1,544	1,135	869	686	556
				7	F ABLE	IIIB					
					d/a	r					
		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
	240	548	241								
	120	556	139	121							
	80	564	141	81.1	81.0						
	60	571	143	67.3	61.1	61.0					
n_0	50	577	144	63.9	51.2	51.0					
	40	587	147	65.4	43.1	41.1	41.0				
	30	604	151	69.1	39.0	31.8	31.2	31.0			
	20	639	160	71.0	40.1	26.9	22.1	21.2	21.0		
	10	764	191	84.9	47.8	30.6	21.5	16.4	13.5	12.0	11.4
	t_{∞}^2/c^2	541	135	60.1	33.8	21.6	15.3	11.0	8.4	6.7	5.4
				,	TABLE	IVA					
			Ex_{i}		sample s		$\alpha = .01$	L			
					d/a	r					
		.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
	$t_{\infty}^{f 2}/c^2$	66,360		7,373	4,148	2,654	1,843	1,354	1,037	819	664
n_0	240	67,440	16,860	7,493	4,215	2,698	1,873	1,376	1,054	833	674
	120	68,500	17,125	7,611	4,281	2,740	1,905	1,398	1,070	846	685
				7	FABLE	IVB					
					d/σ						
		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
	240	674	241								
	120	685	171	121	,						
	80	696	174	84.5	81.0						
	60	708	177	79.1	61.1	61.0					
n_0	50	717	179	79.9	53.1	51.1	51.0				
	40	791	109	00.2	40 1	41 0	41 0				

48.1

47.6

50.6

62.8

41.5

41.2

33.7

32.8

40.2

26.5

41.0

31.1

24.7

27.9

18.4

31.0

21.7

20.8

13.5

21.2

16.5

10.4

21.0

13.9

8.2

12.4

6.6

40

30

20

10

 $t_{\,\infty}^{2}/c^{2}$

731

756

809

664

1,004

183

189

202

251

166

82.3

84.0

89.9

73.7

112