SAMPLE SIZE REQUIRED FOR ESTIMATING THE VARIANCE WITHIN d UNITS OF THE TRUE VALUE[1]

By Franklin A. Graybill and Terrence L. Connell

Colorado State University

1. Introduction. The problem of estimating the variance (σ^2) of a normal density arises in many experimental situations. J. A. Greenwood and M. M. Sandomire [3] have presented a means of obtaining the sample size required to estimate the variance of a normal density within a given per cent of its true value. An investigator may prefer instead to estimate the variance within a given number of units. This paper will provide a two step sampling procedure to solve that problem.

Assume a preliminary sample of size m; $z_1, z_2, \cdots, z_m$, is taken from a normal density with variance σ^2. The unbiased estimator of the variance s_m^2 is computed by the formula $s_m^2 = (m-1)^{-1}\sum (z_i - \bar{z})^2$, and d and $1-\alpha$ are specified in advance. It is desired to determine n, on the basis of the preliminary sample, such that

$$P[|s_{n+1}^2 - \sigma^2| < d] > 1 - \alpha \tag{1.1}$$

where s_{n+1}^2 is equal to $(1/n)\sum_{i=1}^{n+1}(y_i - \bar{y})^2$ and where $y_1, y_2, \cdots, y_{n+1}$ is a random sample of size $n+1$, from a normal density with variance σ^2.

Table I in Section 3 provides the sample size $n+1$, such that (1.1) is true, for $1-\alpha = .90, .95, .99$; $m = 5, 10, 15, 20, 50, 100, 200, 500, 1000$. The only other known method for solving this problem is given in [1], which requires the use of Tchebycheff's inequality. It can be shown that the method presented in this paper provides a significantly smaller second sample size than does [1]. For some comparisons with [1], see Table III.

2. Solution. Equation (1.1) may be written as

$$P[|s_{n+1}^2 - \sigma^2| < d] = E_n\{P[(1-a) < v < (1+a) \mid n]\}$$
$$= \int_1^\infty g(n) \int_{1-a}^{1+a} f_1(v \mid n)\, dv\, dn$$

where E_n is expectation with respect to n; $a = d/\sigma^2$; $v = s_{n+1}^2/\sigma^2$; $g(\cdot)$ is the density of n, and $f_1(\cdot \mid n)$ is the density of a chi-square variable divided by n, its degrees of freedom. We shall restrict n such that $n \geqq 1$. By definition

$$f_1(v \mid n) = [(n/2)^{(n/2)}/\Gamma(n/2)]v^{(n/2-1)}e^{-(n/2)v}, \qquad 0 < v < \infty$$
$$= 0 \qquad\qquad , \qquad -\infty < v \leqq 0.$$

Received 22 July 1963.

[1] This research was supported by the Air Force Office of Scientific Research.

Given that

$$f_2(v \mid n) = [(n-1)^{\frac{1}{2}}/2\pi^{\frac{1}{2}}] \exp[-(n-1)^{\frac{1}{2}}|v-1|/\pi^{\frac{1}{2}}], \quad -\infty < v < \infty$$

it has been shown by Connell and Graybill [2] that

$$\int_{1-a}^{1+a} f_1(v \mid n)\, dv > \int_{1-a}^{1+a} f_2(v \mid n)\, dv = 1 - \exp[-(n-1)^{\frac{1}{2}}a/\pi^{\frac{1}{2}}].$$

If a were known, we might set n equal to $1 + [\pi \log^2 \alpha]/a^2$, since in that case we would have

$$P[|s_{n+1}^2 - \sigma^2| < d] > E_n \int_{1-a}^{1+a} f_2(v \mid n)\, dv = E_n(1-\alpha) = 1 - \alpha.$$

Because a is assumed unknown let

$$n = 1 + [\pi \log^2 \alpha] k^2 s_m^4/d^2 \tag{2.1}$$

where k is some constant, independent of a, such that

$$E_n \int_{1-a}^{1+a} f_2(v \mid n)\, dv = 1 - \alpha. \tag{2.2}$$

The density of s_m^4, and consequently of n, is a known function of σ^2. Also, the expectation of $1 - \exp[-(n-1)^{\frac{1}{2}}a/\pi^{\frac{1}{2}}]$ for n given in (2.1), clearly does not involve σ^2.

The value of k in (2.1) such that (2.2) is true is

$$k = (m-1)[(1/\alpha)^{2/(m-1)} - 1]/2 \log(1/\alpha).$$

Thus, if the sample size

$$n + 1 = (\pi/4)[(1/\alpha)^{2/(m-1)} - 1]^2 (m-1)^2 s_m^4/d^2 + 2 \tag{2.3}$$

is used for the second step sample, the inequality in (1.1) is satisfied. The expected second sample size in (2.3) is

$$E_n(n+1) = (\pi/4)[(1/\alpha)^{2/(m-1)} - 1]^2 (m^2 - 1)\sigma^4/d^2 + 2.$$

3. Sample size tables. The second sample size $n + 1$ in (2.3) insures that (1.1) is true. To find $n + 1$, compute s_m^4/d^2, where s_m^2 is available from the preliminary sample of the procedure and d is the desired allowable deviation from the true variance, multiply by the entry in Table I which corresponds to the appropriate $1 - \alpha$ level and m (the size of the preliminary sample), and add 2.

Table II gives $n + 1$ for some particular values of s_m^4/d^2, $1 - \alpha$, and m.

Table III shows some comparisons between the sample size given in (2.3) and the sample size obtained in [1]. The quantities tabled are

$$\begin{aligned} h(m, \alpha) &= (n-1)/(n'-1) \\ &= (\pi/8)\alpha(m-3)(m-5)[(1/\alpha)^{2/(m-1)} - 1]^2; \quad m \geq 6 \end{aligned}$$

TABLE I

Entries are $(\pi/4)[(1/\alpha)^{2/(m-1)}-1]^2(m-1)^2$

$1-\alpha$	$m = 5$	10	15	20	50	100	200	500	1000
.90	58.75	28.40	23.35	21.33	18.31	17.45	17.05	16.81	16.73
.95	151.50	56.92	43.92	38.97	31.90	29.96	29.06	28.53	28.36
.99	1017.88	202.14	133.34	110.32	80.64	73.17	69.79	67.87	67.24

TABLE II

Sample size $n+1$ *such that* $P[|s_{n+1}^2 - \sigma^2| < d] > 1-\alpha$

s_m^4/d^2	$1-\alpha =$ .90 $m =$ 10	.90 100	.90 1000	.95 10	.95 100	.95 1000	.99 10	.99 100	.99 1000
.25	10	7	7	17	10	10	53	21	19
.5	17	11	11	31	17	17	104	39	36
1.0	31	20	19	59	32	31	205	76	70
2.0	59	37	36	116	62	59	407	149	137
5.0	144	90	86	287	152	144	1013	368	339
10.0	256	177	170	572	302	286	2024	734	675

TABLE III

Comparison of sample size: $n+1$ *given in* (2.3), n' *given in* [1]

$h(m, \alpha) = (n-1)/(n'-1) = E(n-1)/E(n'-1)$

m	$1-\alpha =$ .90	.95	.99
10	.613	.615	.437
100	.820	.704	.344
1000	.832	.705	.334

where $n+1$ is given in (2.3) and n' is the sample size given in [1]. It is noted that $h(m, \alpha) = E(n-1)/E(n'-1)$. It can be demonstrated that

$$h(m, \alpha) < h(m, \alpha_0) < \lim_{m\to\infty} h(m, \alpha_0) = 2\pi e^{-2} \cong .85$$

where $\alpha_0 = [(m-5)/(m-1)]^{(m-1)/2}$. With minor modifications, the results of this paper can be used to estimate the mean of the gamma distribution.

REFERENCES

[1] Birnbaum, A. and Healy, W. C., Jr. (1960). Estimates with prescribed variance based on two-stage sampling. *Ann. Math. Statist.* **31** 662–676.

[2] Connell, T. L. and Graybill, F. A. A Tchebycheff type inequality for chi-square. Submitted for publication.

[3] Greenwood, J. S. and Sandomire, M. M. (1950). Sample size required for estimating the standard deviation as a per cent of its true value. *J. Amer. Statist. Assoc.* **45** 257–260.