ESTIMATES WITH PRESCRIBED VARIANCE BASED ON TWO-STAGE SAMPLING

By Allan Birnbaum[1] and William C. Healy, Jr.[2]

New York University and Ethyl Corporation

1. Summary. A method is given which provides, under conditions satisfied by many common distributions, rules for sampling in two stages so as to obtain an unbiased estimator of a given parameter, having variance equal to, or not exceeding, a prescribed bound. The method is applied to estimation of the means of binomial, Poisson, and hypergeometric distributions; scale-parameters in general and of the Gamma distribution in particular; the variance of a normal distribution; and a component of variance. The use of such estimators to achieve homoscedasticity is discussed. Optimum sampling rules are discussed for some of these estimators, and some tables are given to facilitate their use. The efficiency of the method is shown to be high in many cases.

2. Introduction. In most problems of estimation, estimators based on samples of fixed sizes have precisions which depend on unknown parameters, and estimators with prescribed precision are not available without resort to sequential sampling in two or more stages, as in Stein's procedure [1] for estimation of the mean of a normal distribution with unknown variance. For problems other than those of the type treated by Stein the only available general methods which are both fairly practicable and efficient seem to be the double-sampling method of Cox [2], [3] and the sequential method of Anscombe [4]. The latter methods, however, are approximate, being based on asymptotic theory, and there seems to be no easily applicable method available for determining in a given case the closeness of the approximations involved. An approach employing a different concept of prescribed precision is described by Graybill [16].

The method to be described below (developed independently by the authors) is a simple one which provides, in a number of problems, procedures for two-stage sampling leading to estimators which are exactly unbiased; in certain problems these estimators have exactly a prescribed variance, while in other problems they have variances never exceeding but generally close to a prescribed bound. Under certain conditions, primarily that the precision prescribed is sufficiently high, these estimators are shown to have generally high efficiency.

3. General discussion of the method.

3.1 Statement of problems. Let $S = \{x\}$ be the sample space for a single random observation, X, on which a density or discrete elementary probability function,

Received February 12, 1959; revised October 16, 1959.

[1] Work supported by the Office of Naval Research.

[2] Work supported in part by the Office of Ordinance Research, U. S. Army, while this author was at the University of Illinois.

$f(x, \theta)$, is defined for each θ in a given parameter space Ω. Suppose that it is desired to estimate with prescribed precision a real-valued function $\rho = \rho(\theta)$. We adopt the following formalization of this requirement: it is required to find find an unbiased estimator of ρ having variance not exceeding a given positive function $B(\theta)$.

3.2 Assumptions.

I. Assume that, for each non-sequential sample size n not less than a known n_0, there exists an unbiased estimator $t = t(x_1, \cdots x_n)$ of ρ, i.e.,

$$E_\theta t(X_1, \cdots X_n) = \rho(\theta). \tag{1}$$

II. Let $\sigma^2(\theta, n)$ denote the variance of t for sample size n. Assume that, for each non-sequential sample size m not less than a known m_0, there exists a measurable "second sample size function" $n = n(x_1, \cdots x_m)$ taking integer values not less than n_0, and such that either

$$E_\theta \sigma^2(\theta, n(X_1, \cdots X_m)) = B(\theta) \tag{2a}$$

or

$$E_\theta \sigma^2(\theta, n(X_1, \cdots X_m)) \leqq B(\theta) \tag{2b}$$

holds for each θ.

3.3 Estimation Procedure. Under the above assumptions, a simple unbiased estimator of ρ, having variance not exceeding $B(\theta)$, is given by any procedure of the following form:

A) Take a sample of m observations ($m \geqq m_0$), $x_1, \cdots x_m$, and compute $n = n(x_1, \cdots x_m)$.

B) Take a second independent sample of $n = n(x_1, \cdots x_m) \geqq n_0$ additional observations, $x_{m+1}, \cdots x_{m+n}$.

C) Estimate ρ by $t = t(x_{m+1}, \cdots x_{m+n})$, ignoring at this stage the first sample observations $x_1, \cdots x_m$.

The fact that this procedure seems to involve gross waste of information in the first sample suggests at first sight that its efficiency must be low. It will be shown, however, that the efficiency of the method, with a suitable choice of sampling rule, is so high in a number of cases that the search for more efficient methods (generally not known at present) would seem to be of more theoretical than practical interest for those cases.

3.4 Properties of the Method. We first verify that, when functions t and n can be found satisfying conditions (1) and (2b) above, the method gives unbiased estimators with variances not exceeding the prescribed bound. Let

$$N = n(X_1, \cdots X_m).$$

Then the estimate is $T = t(X_{m+1}, \cdots X_{m+n})$, and

$$E_\theta(T) = E_N(E_T[T \mid N]) = E_\theta(\rho) = \rho, \tag{3}$$

since, for each fixed $n \geqq n_0$, we have $E_\theta t(X_{m+1}, \cdots X_{m+n}) = \rho$ by (1). Also

$$\text{(4)} \qquad \text{Var}_\theta(T) = E_N \text{Var}_T(T \mid N) = E_\theta \sigma^2(\theta, N) \leqq B(\theta)$$

by (2b); if (2a) holds, $\text{Var}_\theta(T) = B(\theta)$.

3.5 Efficiency Considerations. A measure of efficiency for any sequential estimator satisfying (3) and (4), and *not* restricted to the use of only two stages of sampling, may be devised as follows: It has been shown by Wolfowitz [5], under certain regularity conditions on $f(x, \theta)$ and $\rho(\theta)$ and certain broad conditions on the sequential sampling rule, that each unbiased sequential estimator t of ρ, together with its total random sequential sample size N', satisfies

$$\text{(5)} \qquad \text{Var}_\theta(T) \geqq E_\theta \left(\frac{\partial \log f(X, \theta)}{\partial \rho} \right)^2 \Big/ E_\theta(N').$$

From (4) and (5) we obtain, under the conditions mentioned, the following lower bound for the expected total sample size required by any sequential estimator meeting our conditions (3) and (4):

$$\text{(6)} \qquad E_\theta(N') \geqq E_\theta \left(\frac{\partial \log f(X, \theta)}{\partial \rho} \right)^2 \Big/ B(\theta)$$

As will be shown by specific examples below, there does not necessarily exist an estimator which attains this lower bound. Nevertheless it is useful to define as an index of efficiency the function

$$\text{(7)} \qquad R(\theta) = E_\theta \left(\frac{\partial \log f(X, \theta)}{\partial \rho} \right)^2 \Big/ [B(\theta) E_\theta(N')],$$

where $E_\theta(N')$ is computed for any given estimate satisfying (3) and (4). As an example of the interpretation of this index, suppose that for a given estimator we find that $R(\theta) \geqq 0.90$ for all θ; then we can assert that for every estimator meeting the conditions (3), (4), and the general conditions of [5], the required expected total sample size function $E_\theta(N^*)$ will satisfy $E_\theta(N^*) \geqq 0.90\, E_\theta(N')$ for all θ; hence average savings in sample sizes of at most 10% might be achieved. It is known that in general the savings actually possible are less than indicated by such bounds, e.g. less than the 10% indicated here.) Such efficiency bounds are given in the following sections for various specific problems.

The estimation methods of the present paper are roughly similar to the method of Stein [1] for estimation of a normal mean. For most purposes the prescribed-length confidence interval formulation adopted by Stein seems preferable to the prescribed-variance formulation adopted here; the present formulation is akin to a decision-theoretic one with mean-squared error loss function, but the restriction of unbiasedness which provides essential simplications of calculations also generally entails some inefficiency from this standpoint. While Stein was able to give exact confidence intervals by determination of the exact (Student's) distribution of the point estimator implicit in his method, the exact distributions of the estimators given here are not known. Consequently this paper makes

no contribution to the theory of exact interval estimation comparable with Stein's, apart from the following crude use of Tchebycheff's inequality: If $\hat{\theta}$ is an unbiased estimator of θ with variance not exceeding a constant B, then the interval estimator $\hat{\theta} \pm \epsilon$ covers θ with probability at least $1 - \alpha$ where $\alpha = B/\epsilon^2$. For many of the problems considered below, even such confidence intervals have not previously been available. (For a number of problems, a method of constructing confidence intervals of fixed length and confidence coefficient, but probably poor efficiency, was given in [6]).

In many cases, particularly those in which high precision is specified, the estimators given here have approximately normal distributions. This is illustrated in the Poisson case below. To the extent that this is true, all methods for confidence regions and significance tests based on assumptions of normality with known variance may be applied. Useful approximations to the distributions of some of the estimators can probably be based on Student's t distributions with the number of degrees of freedom determined by a fitting of fourth moments; further investigation of this possibility is required.

It should be noted that, with the methods of this paper, there will sometimes occur samples which on inspection strongly suggest that some modification of the estimators given here would be more appropriate and efficient. A similar comment applies, with somewhat less force, to Stein's procedure and some other sequential procedures. These features seem symptomatic of possible improvements in efficiency of these methods which have not yet been found. They seem also to point to more basic problems in the foundations of statistical inference which lie outside the scope of the present paper. The estimators given in this paper have variance and efficiency properties which are valid within the unconditional two-stage sampling probability framework; these properties are not considered here (except in some computational steps) conditionally on a given first or second sample size. The unbiasedness properties of these estimators generally hold both conditionally and unconditionally.

4. Estimation of a mean. Suppose that X is real-valued and that the mean

$$\theta = \rho(\theta) = E_\theta(X)$$

is the parameter to be estimated. Then Assumption I is obviously satisfied if we take

$$t = t(x_{m+1}, \cdots x_{m+n}) = \frac{1}{n}\sum_{i=1}^{n} x_{m+i} .$$

Letting $\sigma_\theta^2 = \mathrm{Var}_\theta(X)$, we have $\sigma^2(\theta, n) = \sigma_\theta^2/n$. Condition (4) becomes

$$E_\theta(1/N) \equiv E_\theta 1/n(X_1, \cdots X_m) \leqq B(\theta)/\sigma_\theta^2 . \tag{8}$$

Then any integer $m \geqq m_0$, and any function $n = n(x_1, \cdots x_m)$ satisfying (8), may be used to define an estimator, which will then automatically satisfy (3) and (4). Such an estimator has expected total sample size

$$E_\theta(N') = m + E_\theta[n(X_1, \cdots X_m)] = m + E_\theta(N), \tag{9}$$

and efficiency bound

$$R(\theta) = E_\theta\left(\frac{\partial \log f(X,\theta)}{\partial\theta}\right)^2 \Big/ B(\theta)\,[m + E_\theta(N)]. \tag{10}$$

In the special case of constant prescribed precision, $B(\theta) \equiv B$, (8) becomes $(1/B)E_\theta[1/n(X_1, \cdots X_m)] \leqq 1/\sigma_\theta^2$, and the problem of finding a suitable second-sample-size function $n(x_1, \cdots x_m)$ may be stated as the problem of finding an estimator $1/\hat{\sigma}^2$ of $1/\sigma_\theta^2$, based on m observations, which is unbiased (condition (2a)), or which has positive bias at no $\theta \,\varepsilon\, \Omega$ (condition (2b)). Then the sequential sampling rule may be stated as:

A′) Observe $x_1, \cdots x_m$, and compute $\hat{\sigma}^2 = \hat{\sigma}^1(x_1, \cdots x_m)$.

B′) Take a second sample of $n = \hat{\sigma}^2/B$ observations $x_{m+1}, \cdots x_{m+n}$.

C′) Estimate the mean $\theta = E_\theta(X)$ by the mean $t = 1/n\sum_{i=1}^{n} x_{m+i}$ of the second sample only.

It is sometimes convenient to define $\hat{\sigma}^2(x_1, \cdots, x_m)$ formally in such a way that $\hat{\sigma}^2/B$ is not always an integer. Then for most applications it will suffice to take n as the smallest integer not less than $\hat{\sigma}^2/B$. A calculation like that above shows that this gives again $\mathrm{Var}_\theta(T) \leqq B$. Alternatively, given $\hat{\sigma}^2/B$, we could use a random device to choose $n = [\hat{\sigma}^2/B]$ = the largest integer not exceeding $\hat{\sigma}^2/B$, with a probability γ, and $n = [\hat{\sigma}^2/B] + 1$ with probability $1 - \gamma$, where γ is determined by the equation

$$\gamma[\hat{\sigma}^2/B]^{-1} + (1 - \gamma)([\hat{\sigma}^2/B] + 1)^{-1} = B/\hat{\sigma}^2.$$

The latter procedure, which is perhaps of primarily theoretical interest, gives $\mathrm{Var}_\theta(T) = B$ exactly if $E_\theta[1/\hat{\sigma}^2(X_1, \cdots X_m)] = 1/\sigma^2$ exactly. Henceforth we write $n = \hat{\sigma}^2/B$ to indicate that one of these procedures is used in defining n. It follows that calculations based on the equation $n = \hat{\sigma}^2/B$, such as the equation

$$E_\theta n(X_1, \cdots X_m) = E_\theta\hat{\sigma}^2(X_1, \cdots X_m)/B$$

used below, may involve an error whose magnitude is in any case less than one. Similar remarks apply to cases of the method other than those of estimation of a mean.

For any such procedure we have expected total sample size

$$E_\theta(N') = m + E_\theta n(X_1, \cdots X_m) = m + (1/B)E_\theta\hat{\sigma}^2(X_1, \cdots X_m),$$

and efficiency bound given by

$$1/R(\theta) = (B/\sigma_\theta^2)E_\theta(N') = (Bm/\sigma_\theta^2) + (1/\sigma_\theta^2)E_\theta\hat{\sigma}^2.$$

If B is sufficiently small, and/or if θ is such that σ_θ^2 is sufficiently large, it is true in many cases (as illustrated below) that $(1/\sigma_\theta^2)E_\theta\hat{\sigma}^2 \doteq 1$ and that $Bm/\sigma_\theta^2 \doteq 0$, and hence that $R(\theta) \doteq 1$; in such cases, for the indicated range of θ, no appreciable improvements in efficiency are possible even by resort to fully sequential estimators.

Such estimators have been found and investigated quantitatively for a number of common problems. These results are summarized in the following paragraphs.

4.1 Poisson Mean. If X has the Poisson distribution $f(x, \theta) = e^{-\theta}\theta^x / x!$ for $x = 0, 1, \cdots$, we may take

$$\hat{\sigma}^2 = \hat{\sigma}^2(x_1, \cdots x_m) = \left(\sum_1^m x_i + 1\right) \Big/ m,$$

since $y = \sum_1^m x_i$ has the Poisson distribution $f(y, m\theta) = e^{-m\theta}(m\theta)^y/y!$, $y = 0, 1, 2, \cdots$, and

$$E_\theta(1/\hat{\sigma}^2) = m \sum_{y=0}^{\infty} f(y, m\theta)/(y + 1) = me^{-m\theta} \sum_{y=0}^{\infty} (m\theta)^y/(y + 1)!$$

$$= (e^{-m\theta}/\theta) \sum_{y=0}^{\infty} (m\theta)^{y+1}/(y + 1)! = (1 - e^{-m\theta})/\theta < 1/\theta = 1/\sigma_\theta^2.$$

When the second sample size is determined by $n = \hat{\sigma}^2/B = (y + 1)/mB$, the expected total sample size is

$$E_\theta(N') = m + E_\theta(n) = m + (1/mB)E_\theta(y + 1) = m + (m\theta + 1)/mB.$$

This is minimized by taking $m \doteq 1/B^{1/2}$, regardless of the value of θ. (In other examples, an optimal first sample size is not so simple to determine.) Then $E_\theta(N') = \theta/B + 2/B^{1/2}$.

This estimator has efficiency bound $R(\theta)$ given by $1/R(\theta) = 1 + 2B^{1/2}/\theta$. If, for example, $\theta = 8B^{1/2}$, then $R(\theta) = 0.8$, and a decrease of at most

$$100(1 - R(\theta))\,\% = 20\,\%$$

in $E_\theta(N')$ might be possible by resort to some (unknown) more refined sequential procedure; for $\theta \gg 8B^{1/2}$, the possible gains are negligible.

An alternative two-stage estimator (of the mean of a Poisson process) employing "inverse" sampling in the first stage, given in [10], has exactly variance B, but can be shown to be less efficient.

The following discussion illustrates that such estimators can have approximately normal distributions. For any fixed $\theta > 0$, $B > 0$, and k, we may write $\text{Prob}\{(T - \theta)/B^{1/2} < k\} = E_N U(N, \theta, k, B)$, where

$$U(N, \theta, k, B) = \text{Prob}\{(T - \theta)/B^{1/2} < k \mid N\}.$$

For sufficiently large fixed N,

$$U(N, \theta, k, B) \equiv \Pr\{(T - \theta)/(\theta/N)^{1/2} < kB^{1/2}/(\theta/N)^{1/2}\} \doteq \Phi(k(BN/\theta)^{1/2}),$$

where $\Phi(u)$ is the standard normal c.d.f. As $B \to 0$, the random variable

$$\Phi(k(BN/\theta)^{1/2})$$

converges in probability to the constant $\Phi(k)$, as does the random variable

$U(N, \theta, k, B)$ with N random. Since $0 \leqq U \leqq 1$, we have

$$E_N(U) = \text{Prob}\{(T - \theta)/B^{1/2} < k\} \to \Phi(k)$$

as B decreases, proving the asymptotic normality of T.

4.2 Binomial Mean. If X has the binomial distribution

$$f(x, \theta) = \theta^x(1 - \theta)^{1-x}, \qquad x = 0, 1$$

we may take

$$\hat{\sigma}^2 = (1 - 2^{-(m+1)})(\sum_1^m x_i + 1)(m + 1 - \sum_1^m x_i)/(m + 1)(m + 2),$$

since (by a calculation similar to that in the preceding section)

$$E_\theta(1/\hat{\sigma}^2) = (1 - \theta^{m+2} - (1 - \theta)^{m+2})/(1 - 2^{-(m+1)})\theta(1 - \theta)$$
$$\leqq 1/\theta(1 - \theta) = 1/\sigma_\theta^2.$$

The expected total sample size is

$$E_\theta(N') = m + (1/B)E_\theta\hat{\sigma}^2 = m + (1 - 2^{-(m+1)})\frac{m(m-1)\theta(1-\theta) + m + 1}{B(m+1)(m+2)}$$
$$= m + (1 - 2^{-(m+1)})\frac{1}{B(m+2)} + \theta(1-\theta)\left(1 - \frac{4m+2}{(m+1)(m+2)}\right).$$

The latter expression does not yield a minimizing value of m independent of the unknown θ, but for any chosen B and guessed value θ a minimizing value

$$m = m(\theta, B)$$

can be found by numerical solution of the equation $\frac{\partial}{\partial\theta} E_\theta(N') = 0$. Table 1 provides some such values.

TABLE 1

Best Binomial First Sample Sizes $m(\theta, B)$

θ	B			
	$(0.05)^2$	$(0.02)^2$	$(0.01)^2$	$(0.005)^2$
0.5	0	0	0	0
0.4 or 0.6	0	0	26	47
0.3 or 0.7	0	20	40	81
0.2 or 0.8	11	29	59	119
0.0 or 1.0	18	48	98	198

The value $m = 0$ indicates use of a single sample procedure with $n = 1/4B$ observations. However, calculations of $E_\theta(N')$ for various values of B and m, such as those given in Table 2, indicate that for $B \leqq (0.05)^2$ a choice of m such

as $m(0.2, B)$ provides appreciable savings as compared with $m = 0$ over a wide range of θ at the cost of a relatively small loss as compared with use of a best value $m(\theta', B)$ based on any guessed value θ' which happens to be correct.

TABLE 2

Values of $E_\theta(N')$ for Binomial Estimates

(a) $B = (0.05)^2$

θ	m		
	0	11	18
0.5	100	112.3	118.5
0.4 or 0.6	100	109.4	115.3
0.3 or 0.7	100	101.0	105.6
0.2 or 0.8	100	86.9	89.5
0.1 or 0.9	100	67.1	67.0
0.0 or 1.0	100	41.8	38.0

(b) $B = (0.005)^2$

θ	m				
	0	47	81	119	198
0.5	10,000	10,056	10,084	10,120	10,199
0.4 or 0.6	10,000	9,688	9,703	9,734	9,806
0.3 or 0.7	10,000	8,585	8,561	8,573	8,630
0.2 or 0.8	10,000	6,746	6,656	6,639	6,670
0.1 or 0.9	10,000	4,173	3,990	3,931	3,926
0.0 or 1.0	10,000	863	563	450	398

The efficiency bound $R(\theta)$ of such estimators is given by

$$1/R(\theta) = Bm/\theta(1-\theta) + (1 - 2^{-(m+1)}) \cdot\{1/[(m+2)\theta(1-\theta)] + 1 - (4m+2)/[(m+1)(m+2)]\};$$

For any given B, the values of m to be considered are $0 \leqq m \leqq m(0, B)$. If we take $m = m(0, B) = 1/B^{1/2} - 2 \doteq 1/B^{1/2}$ (for $B \leqq (0.05)^2$), we have

$$1/R(\theta) \doteq 1 + B^{1/2}(2/\theta(1-\theta) - 4).$$

For any B, as $\theta \to 0$ or 1, $R(\theta) \to 0$; but these are values of θ for which the lower bound σ_θ^2/B on $E_\theta(N')$ cannot be attained by any estimator with the desired properties. For any fixed θ, $0 < \theta < 1$, as $B \to 0$, $R(\theta) \to 1$; thus the efficiency of $\hat{\theta}$ cannot be much improved upon when high precision is required. Analogous statements hold if we take, for example, $m(0.2, B)$.

The formula for the second sample size is

$$n = c\left(\sum_1^m x_i + 1\right)\left(m + 1 - \sum_1^m x_i\right),$$

where $c = c(B, m) = (1 - 2^{-(m+1)})/B(m + 1)(m + 2)$. Table 3 provides some values of $c(B, m)$.

TABLE 3

Values of $c(B, m)$ for Binomial Second Sample Sizes

m	B			
	$(0.05)^2$	$(0.02)^2$	$(0.01)^2$	$(0.005)^2$
10	3.03	18.94	75.76	303.04
15	1.47	9.19	36.76	147.04
20	0.87	5.41	21.64	86.56
25	0.57	3.56	14.24	56.96
30	0.40	2.52	10.08	40.32
35	0.30	1.88	7.52	30.08

The variance of $\hat{\theta}$ is

$$\mathrm{Var}_\theta(\hat{\theta}) = B(1 - \theta^{m+2} - (1 - \theta)^{m+2})/(1 - 2^{-(m+1)}) \leqq B;$$

this is appreciably less than B only when θ is very near 0 or 1, provided m is not very small.

There exists, as in the Poisson case, a procedure employing "inverse" sampling to yield a binomial estimator having exactly constant variance as follows:

Let M be a fixed positive integer. Make successive independent Bernoulli trials (samples of size one) until min (total successes, total failures) $= M$. Let x be the number of trials up to and including the Mth success. Let y be the number of trials up to and including the Mth failure. Take an additional sample of size $n = M/B(x + y)$ and let z denote the number of additional successes observed. Then $\tilde{\theta} = Bz(x + y)/M$ is an unbiased estimator of θ having variance exactly equal to B. It seems clear that the expected sample size will be larger for this "inverse" sampling plan, and that as a practical matter exactly prescribed variance would seldom be worth the cost in additional observations.

4.3 Hypergeometric mean. In a finite population of known size M, let $\theta = D/M$ be the unknown proportion of items having a given trait, e.g. being defective. Let X denote the number of defectives in a first sample of size m, $n = n(x)$ the size of a second sample, and Y the number of defectives in the second sample. (All sampling is without replacement.) Then it is readily verified that an unbiased estimator of θ is

$$\hat{\theta} = [(Y(M - m)/n) + X]/M.$$

This estimator will have variance bounded by B if we take

$$n = n(x) = \frac{(M - m)^2(x + 1)(m - x + 1)}{BM^2(m + 1)(m + 2) + (x + 1)(m - x + 1)(M - m)}$$

since

$$\mathrm{var}(\hat{\theta} \mid x) = \frac{(D - x)(M - m - D + x)(M - m - n)}{(M - m - 1)M^2n}$$

and hence

$$\operatorname{var}\hat{\theta} = E\operatorname{var}(\hat{\theta} \mid x)$$

$$= \sum_{x=a}^{b} \frac{(D-x)(M-m-D+x)(M-m-n)}{(M-m-1)M^2n} \binom{D}{x}\binom{M-D}{m-x}\Big/\binom{M}{m}$$

$$= \sum_{x=a}^{b} \frac{(M-m-n)(x+1)(m-x+1)(M-m)}{M^2n(m+1)(m+2)} \cdot \binom{D}{x+1}\binom{M-D}{m-x+1}\Big/\binom{M}{m+2}$$

$$= \sum_{w=a+1}^{b+1} \frac{(M-m-n)w(m+2-w)(M-m)}{M^2n(m+1)(m+2)} \cdot \binom{D}{w}\binom{M-D}{m+2-w}\Big/\binom{M}{m+2}$$

$$= B\sum_{w=a+1}^{b+1} \binom{D}{w}\binom{M-D}{m+2-w}\Big/\binom{M}{m+2} \leqq B.$$

Exact results on expected sample sizes are not available, but an indication of the possible savings is given by regarding the results in the binomial case to be limits approached as M becomes infinite. Additional information is given by the range of $n(x)$: for m even,

$$m + \frac{(M-m)^2}{M^2B(m+2)+(M-m)} \leqq m + n(x) \equiv N' \leqq m + \frac{(M-m)^2(m+3)^2}{4BM^2(m+1)(m+2)+(M+3)^2(M-m)}.$$

With a single sample of size r, the best unbiased estimate has variance not exceeding B for all θ provided r is at least $M/(4B(M-1)+1)$. If for example $M = 1{,}000, B = .0001$, and $m = 80$, then $r = 714$ while $173 \leqq m + n(x) \leqq 728$; when $\theta = 0$ or 1, the two-stage estimate saves 541 observations (76%), while when $\theta = \frac{1}{2}$ its maximum sample size exceeds r by less than 14 observations (2%).

4.4 Mean of a Normal Distribution with Unknown Variance. If X has the normal density function $f(x, \theta, \sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(x-\theta)^2/2\sigma^2)$, with σ^2 unknown, we may apply the present method with an advantageous modification based on the independence of $\bar{x}_1 = (1/m)\sum_{i=1}^{m} x_i$ and

$$s^2 = (1/(m-1))\sum_{i=1}^{m}(x_i - \bar{x})^2.$$

Take any $m > 3$, let $\hat{\sigma}^2 = (m-1)s^2/(m-3)$, $n = \max[\theta, \hat{\sigma}^2/B - m]$, and $\hat{\theta} = (1/(m+n))\sum_{i=1}^{m+n} x_i$. It is easily verified that the latter estimator is unbiased and has variance not exceeding B. For most purposes Stein's procedure [1] which gives confidence intervals of prescribed length will probably be preferred; optimal choice of m for this procedure has been extensively investigated [7] and [17].

4.5 Estimation of a Scale Parameter. Let X have a density function

$$f(x, \theta) = (1/\theta)g(x/\theta), \qquad x \geqq 0,$$

and $f(x, \theta) = 0$ otherwise, where $g(u)$ is a known function with

$$c_1 = \int_0^\infty ug(u)\, du = E(X \mid \theta = 1),$$

$$c_2 = \int_0^\infty u^2 g(u)\, du = E(X^2 \mid \theta = 1) < \infty, \qquad \text{and}$$

$$c_3 = \int_0^\infty u^{-2} g(u)\, du = E\,(1/X^2 \mid \theta = 1) < \infty.$$

Then $E(X/c_1) = \theta$, Var $(X/c_1) = \theta^2(c_2/c_1^2 - 1) = \sigma^2$, say, and $E(1/X^2) = c_3/\theta^2$. Letting $\hat{\sigma}^2 = mc_3(c_2/c_1^2 - 1)[\sum_1^m 1/x_i^2]^{-1}$, we have

$$E(1/\hat{\sigma}^2) = 1/[c_3(c_2/c_1^2 - 1)]E(1/X^2) = 1/\sigma^2.$$

Thus an unbiased estimator of the scale-parameter θ, having variance B, is $\hat{\theta} = \sum_{i=m+1}^{m+n} x_i/c_1 n$. The choice of m may be made so as to minimize, at any guessed value of θ,

$$E_\theta(N') = m + E_\theta(n) = m + E_\theta(\hat{\sigma}^2)/B.$$

For any specific density function $f(x, \theta)$, it may be possible to find an estimate $\tilde{\sigma}^2$ preferable to the $\hat{\sigma}^2$ given above: $\tilde{\sigma}^2 = \tilde{\sigma}^2(x_1, \cdots x_m)$ is clearly preferable to $\hat{\sigma}^2$ if it (has the essential property $E(1/\tilde{\sigma}^2) \leqq 1/\sigma^2$ and also) makes $\hat{\theta}$ more efficient, that is, if $E(\tilde{\sigma}^2) < E(\hat{\sigma}^2)$. (This remark may be applied also to the estimators $\hat{\sigma}^2$ discussed in other sections of this paper.)

For example, if X has the gamma density

$$f(x, \theta) = (1/\theta\alpha!)(x/\theta)^\alpha e^{-x/\theta}, \qquad x \geqq 0,$$

where α is known, $\alpha > -1$, then

$$c_1 = \frac{1}{\alpha!}\int_0^\infty u^{\alpha+1} e^{-u}\, du = \alpha + 1,$$

$$c_2 = \frac{1}{\alpha!}\int_0^\infty u^{\alpha+2} e^{-u}\, du = (\alpha + 1)(\alpha + 2),$$

and, provided we require $\alpha > 1$,

$$c_3 = \frac{1}{\alpha!}\int_0^\infty u^{\alpha-2} e^{-u}\, du = 1/\alpha(\alpha - 1).$$

Thus $\hat{\sigma}^2 = m[(\alpha - 1)\alpha(\alpha + 1)\sum_{i=1}^m 1/x_i^2]^{-1}$. The evaluation of $E_\theta(\hat{\sigma}^2)$, required to compute $E_\theta(N')$, appears difficult for $m > 1$ and has not been carried out. For $m = 1$,

$$E_\theta(\hat{\sigma}^2) = E_\theta(X^2)/(\alpha - 1)\alpha(\alpha + 1) = \theta^2 c_2/(\alpha - 1)\alpha(\alpha + 1)$$
$$= \theta^2(\alpha + 2)/\alpha(\alpha - 1),$$

and $E_\theta(N') = 1 + \theta^2(\alpha + 2)/\alpha(\alpha - 1)B$. Presumably the estimator

$$\tilde{\sigma}^2 = K\left(\sum_1^m x_i\right)^2,$$

where K is such that $E_\theta(1/\tilde{\sigma}^2) = 1/\sigma^2$, is preferable to $\hat{\sigma}^2$ in this example, since $\tilde{\sigma}^2$ is a function of the sufficient statistic $\sum_1^m x_i$ (based on the first sample) and $\hat{\sigma}^2$ is not. Results of Ghurye [15] lend support to this conjecture. An estimator based on $\tilde{\sigma}^2$ is given in Section 5.3 below.

The estimation of any given power θ^p of a scale-parameter θ, $p = \pm 1, \pm 2, \cdots$, may be treated similarly.

5. Other estimation problems.

5.1 Variance of a normal distribution with unknown mean. Let X have the normal density function with unknown mean and variance as in Section 4.4, but let θ now denote the unknown *variance*. For any $m > 5$, let

$$n = 2s^2(m - 1)^2/B(m - 3)(m - 5) + 1,$$

where s^2 is the first sample variance defined as above. Then it is readily verified that an unbiased estimator of θ, with variance not exceeding B, is given by the second sample variance.

$$\hat{\theta} = \sum_{m+1}^{m+n}\left(x_i - \frac{1}{n}\sum_{m+1}^{m+n} x_j\right)^2 \Big/ (n - 1),$$

and that

$$E_\theta(N') = m + 1 + \frac{2(m + 1)(m - 1)\theta^2}{(m - 3)(m - 5)B}.$$

For given B and a guessed value of θ, m may be chosen so as to minimize $E_\theta(N')$.

5.2 Estimation of a "between classes" variance component. Consider the usual assumptions for a one-way analysis of variance, with n observations from each of k classes: $Y_{ij} = \mu + c_i + e_{ij}$, $i = 1, \cdots k, j = 1, \cdots n$, with μ an unknown constant, and the c_i's and e_{ij}'s all independently normally distributed with means zero and unknown variances

$$\text{var}\,(c_i) = \sigma_0^2, \qquad \text{var}\,(e_{ij}) = \sigma^2, \qquad i = 1, \cdots k, j = 1, \cdots n.$$

The usual between classes mean square s_0^2 has expected value $\sigma^2 + n\sigma_0^2$ and $k - 1$ degrees of freedom. The usual within classes mean square s^2 has expected value σ^2 and $k(n - 1)$ degrees of freedom. Then $(s_0^2 - s^2)/n$ is an unbiased estimator of σ_0^2, with variance

$$2[(\sigma^2 + n\sigma_0^2)^2/(k - 1) + \sigma^4/k(n - 1)]/n^2$$

when k and n are fixed.

Alternatively, suppose a first sample of r classes and n observations per class has been taken. Let T_0^2 and T^2 respectively denote the between and within classes mean squares, based respectively on $\nu_0 = r - 1 > 4$ and $\nu = r(n - 1) > 4$

degrees of freedom. Then it is easily verified that

$$E(\nu_0 - 2)(\nu_0 - 4)/\nu_0^2 T_0^4 = 1/(\sigma^2 + n\sigma_0^2)^2$$

and

$$E(\nu - 2)(\nu - 4)/\nu^2 T^4 = 1/\sigma^4.$$

This leads to the choice of k defined by $k = \max(2, k', k'')$ where

$$k' = 1 + 2T_0^4\nu_0^2/(\nu_0 - 2)(\nu_0 - 4)(B - b)$$

$$k'' = 2T^4\nu^2/(\nu - 2)(\nu - 4)bn^2(n - 1)$$

and b is any constant, $0 < b < B$. To see that with k so defined, the sampling variance of $\hat{\sigma}_0^2$ is less than B, observe that

$$\begin{aligned} B &= 2[(\sigma^2 + n\sigma_0^2)^2 E(1/(k' - 1)) + (\sigma^4/(n - 1))E(1/k'')]/n^2 \\ &\leqq 2[(\sigma^2 + n\sigma_0^2)^2 E(1/(k - 1)) + (\sigma^4/(n - 1))E(1/k)]/n^2 \\ &= \text{var}\,(\hat{\sigma}_0^2). \end{aligned}$$

The choice of n would ordinarily be influenced by practical limitations on the experiment, and the choice of both n and b could also be governed by an a priori estimate of σ^2/σ_0^2.

An alternative approach to the present problem is to apply twice the method of the preceding Section 5.1 as follows: Estimate $(\sigma_0^2 + \sigma^2)$ by a two-stage estimator s_1^2 having variance not exceeding $B_1 < B$, based on observations $Y_{11}, Y_{21}, \cdots Y_{m,1} Y_{(m+1),1}, \cdots Y_{(m+n),1}$, so that only one observation is taken from each class. Secondly, estimate σ^2 by a two-stage estimator s_2^2 having variance not exceeding B_2, where $B_2 = B - B_1$, based on additional observations within any one class (or on additional "within degrees of freedom" from several classes). Then $s^2 = s_1^2 - s_2^2$ is the required estimate, for $E(s^2) = \sigma^2$, and

$$\text{var}\,(s^2) \leqq B_1 + B_2 = B.$$

Rules for optimal choice of B_1, and comparisons with the preceding method, remain to be developed.

5.3 Scale parameter of a gamma distribution. If X has the Gamma density defined in Section 4.5 above,

$$\text{Var}\,(X/c_1) = \text{Var}\,(X/(\alpha + 1)) = \theta^2(c_2/c_1^2 - 1) = \theta^2/(\alpha + 1) = \sigma^2,$$

and we may take

$$\hat{\sigma}^2 = \left(\sum_1^m x_i\right)^2/(\alpha + 1)(m\alpha + m - 2)(m\alpha + m - 1),$$

for all α and m such that $(m\alpha + m - 2) > 0$. This gives $E(1/\hat{\sigma}^2) = 1/\sigma^2$, and $E(\hat{\sigma}^2) = \theta^2(m\alpha + m + 1)(m\alpha + m)/(\alpha + 1)(m\alpha + m - 1)(m\alpha + m - 2)$. For any guessed value of θ, m may be chosen, subject to $m > 2/(\alpha + 1)$, so as to minimize $E_\theta(N') = m + E_\theta(\hat{\sigma}^2)/B$.

A modification analogous to that in Section 4.5 above, replacing $\sum x_1$ by

$(\sum x_i)^p$ throughout, with a corresponding modification of constants, gives an estimator of θ^p with variance B.

6. Applications to achieve homoscedasticity. Many standard techniques for comparing means, related to the analysis of variance (Model I), are seriously dependent for validity on the assumption that observations have (approximately) equal variances, but much less seriously dependent on the usual assumption of normality (see, for example, [8] and references therein). It is frequently desired to apply such methods to means of observations having some of the distributions considered in Section 4 above; but in such cases the unknown variances are functions of the unknown means of observations, and hence the assumption of equal variances generally holds only when the unknown means happen to be equal.

The methods of the present paper provide a way of meeting this difficulty which may be considered in cases where it is feasible to use a two-stage sampling method providing (approximately) a common prescribed variance B for the observation in each cell of any Model I experimental design. Techniques related to analysis of variance will be used taking, formally, the case of an infinite number of degrees of freedom for the error mean square; the latter, of course, will not be calculated from data, but the known variance B will be used instead. The methods which are usually considered for meeting this difficulty are variance-stabilizing transformations of the observations (see, for example, [9]). Concerning the relative advantages and disadvantages of these approaches, it should be noted that the goal of (a) variance-stabilization for application of standard inference techniques is usually of interest simultaneously with certain goals of (b) precision of estimation (or power of tests), (c) efficient utilization of data obtained, and (d) simplicity of interpretation. Concerning (d), use of the methods of this paper offers some advantages over use of transformations since the former provides inferences directly about the means of interest with prescribed precision on their original scale, rather than inferences about functions of those means (e.g., $E(\sin^{-1}(x/n)^{\frac{1}{2}})$ in the binomial case) which are often harder to interpret and perhaps less meaningful. Furthermore, the latter estimators lose their constant-precision property when interpreted in the original units of the parameters.

In cases like the Poisson there is no single-sample procedure which provides even bounded, let alone prescribed, precision in the original scale. Hence if such prescribed precision is one goal of interest, sequential methods more or less like those of this paper are required, and the simultaneous achievement of simplicity, and of exactly or approximately known common variances of estimators of means, may be regarded as convenient desirable by-products of the method.

In cases like the binomial, the goals of bounded precision and homoscedasticity are attainable by use of transformed single-sample estimates. In the binomial case, we have seen above that when high constant precision is desired, the two-sample estimate is on the whole rather efficient, and in this case again affords the properties of homoscedasticity and simplicity. If only low precision is required, there is some conflict between the goals mentioned. For example, for

binomial estimation with $B = (.05)^2$ it was shown above that a first-sample size of $m \geqq 20$ gives an inefficient estimate, but $m \geqq 20$ is required for a good degree of homoscedasticity. In such cases efficiency considerations may be weighed against considerations of simplicity of application and interpretation. If it can be assumed that $.2 \leqq \theta \leqq .8$, then $m \geqq 10$ suffices to give a variation of at most 7% in variances of $\hat{\theta}$'s. If it can be assumed that $.1 \leqq \theta \leqq .9$, the variation is at most 10% if $m \geqq 20$.

7. Acknowledgements. The authors are grateful to Dr. Donald Guthrie and to the Applied Mathematics and Statistics Laboratory of Stanford University for permission to reprint some of the tables above from [11] and [12]. Mr. B. B. Bhattacharya [14] of the Indian Statistical Institute independently found some estimates of the type given in this paper.

REFERENCES

[1] STEIN, C. M., "A two-sample test for a linear hypothesis whose power is independent of the variance," *Ann. Math. Stat.*, Vol. 16 (1945), pp. 243–258.

[2] COX, D. R., "A note on the sequential estimation of means," *Proc. Camb. Phil. Soc.* Vol. 48 (1952), pp. 447–450.

[3] COX, D. R., "Estimation by double sampling," *Biometrika*, Vol. 39 (1952), pp. 217–227.

[4] ANSCOMBE, F. J., "Sequential estimation," *J. Roy. Stat. Soc., Ser. B*, Vol. 15 (1953), pp. 1–29.

[5] WOLFOWITZ, J., "The efficiency of sequential estimates and Wald's equation for sequential processes," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 215–230.

[6] BIRNBAUM, A., "Sequential probability ratio confidence sets" (Abstract), *Ann. Math. Stat.*, Vol. 24 (1953), p. 686.

[7] SEELBINDER, B. M., "On Stein's two stage sampling scheme," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 640–647.

[8] BOX, G. E. P., "Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 290–302; and "—II. Effects of inequality of variance and of correlation between errors in the two-way classification," *ibid.*, pp. 484–498.

[9] FREEMAN, M. F., AND TUKEY, J. W., "Transformations related to the angular and the square root," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 607–611.

[10] BIRNBAUM, A., "Statistical methods for Poisson processes and exponential populations," *J. Amer. Stat. Assn.*, Vol. 49 (1954) pp. 254–266.

[11] GUTHRIE, D., "Some two-sample procedures," unpublished M. A. Thesis, Columbia University, 1955.

[12] BIRNBAUM, A., AND GUTHRIE, D., "Tables for estimating a proportion or a Poisson mean with prescribed precision," *Technical Report No. 25*, Jan. 18, 1956, Applied Mathematics and Statistics Laboratory, Stanford University.

[13] HEALY, W. C., "Multiple sampling to estimate a proportion," *O.O.R. Technical Report*, Dec. 15, 1954, Statistical Research Laboratory, University of Illinois.

[14] BHATTACHARYA, B. B. Unpublished note on sequential estimation.

[15] GHURYE, S. G., "Note on sufficient statistics and two-stage procedures," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 155–166.

[16] GRAYBILL, FRANKLIN A., "Determining sample size for a specified width confidence interval," Ann. Math. Stat., Vol. 29 (1958), pp. 282–287.

[17] MOSHMAN, JACK, "A method for selecting the size of the initial sample in Stein's two sample procedure," *Ann. Math. Stat.*, Vol. 29 (1958) pp. 1271–1275.