Journal of the KSQC
Vol. 8, No. 2 (1980)

On the Estimation of Fraction Defectives **

Seong-in Kim*

Abstract

This paper is concerned with the design of an appropriate sampling plan or stopping rule and the construction of estimate for the estimation of process or lot fraction defective. Various sampling plans which are well known or have potential applications are unified into a generalized sampling plan. Under this sampling plan sufficient statistic, probability distribution, moment, and minimum variance unbiased estimate are obtained. Results for various sampling plans can be derived as special cases. Then, under given parameter values, the relative efficiencies of the various sampling plans are compared with respect to expected sample sizes and variances of estimates.

I. INTRODUCTION

In many application of quality control techniques, one is often concerned with the fraction defective of a lot or a process. However, it is usually unknown and has to be estimated from a sample. Since every item in a sample can be classified into two catetories of defective and non-defective, the estimation fraction defective can be viewed as a problem of parameter (success probability p) estimation of independent Bernoulli trials.

For the estimation of unknown parameters various estimates under various sampling plans or stopping rules may be useful or conceivable. Probability distribution and statistical inference under fixed sample size sampling plan (S_1) where preassigned number of observations are taken, have been studied extensively in many statistical literatures. The inverse binomial sampling plan (S_2) where observations are to be continued until a preassigned number of successes are obtained was first treated formally by Haldane [4]. The generalized inverse binomial sampling plan where observations are to be continued until at least m_1 successes and m_2 failures are obtained, where m_1 and m_2 are preassigned numbers, has been studied by Bai, et al. [1].

*Department of Industrial Engineering, Korea University

**Supported by the Asan Foundation(峨山社会福祉事業財団)

In addition to the above three sampling plans various sampling plans can be devised. It is therefore necessary to choose an approptiate sampling plan and estimate considering the cost of samples incurred with the stopping rule and the loss such as biases and mean squared errors incurred with the estimation of a parameter. This may be a serious problem in many applications especially in expensive sampling or destructive inspection.

In Section 2 a sampling plan is developed which will be called a generalized sampling plan (S_0) since it includes almost all of the conceivable sampling plans in practice. In Section 3 the probability distribution of a sufficient statistics for the parameter is studied under S_0 and a simple unbiased estimate based on this sufficient statistic is presented. In Section 4 the relative efficiencies of various sampling plans in unbiased estimation are discussed for various values of the parameter.

II. SAMPLING PLAN

We use the nomenclatures following Girshick et al. [3] and DeGroot [2]. Consider the sampling plan whose boundary points are determined by the stopping rule of addition of successive observations to n_1 observations until at least m_1 successes and m_2 failures are obtained or until the total number of observations equals n_2 where n_1, n_2, m_1, and m_2 are preassigned numbers. This sampling plan is denoted as S_0 $(\cdot; n_1, n_2, m_1, m_2)$ whose boundary points are depicted in Figure 1. We note that a point (x,y) in xy-plane with nonnegative integral coordinates represent an outcome of $x+y$ independent Bernoulli trials where x and y denote the number of successes and failures respectively.

If, in sampling plan S_0 $(\cdot; n_1, n_2, m_1, m_2)$, there is no restriction on the minimum sample size, n_1 will be replaced by 0. Likewise, n_2 will be replaced by ∞ if there is no restriction on the maximum sample size. The subscript and argument will be dropped whenever it is not needed or there is no possibility of confusion.

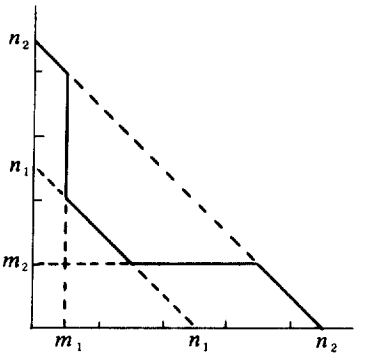

Figure 1
Boundary points under
$S_0(\cdot; n_1, n_2, m_1, m_2)$

Since, as seen below, sampling plan S_0 $(\cdot; n_1, n_2, m_1, m_2)$ includes almost all of the conceivable sampling plans in practice, it will be called a generalized sampling plan. For example, fixed sample size sampling plan is S_1 $(\cdot; n_1, n_1, 0, 0)$ whose boundary points are depicted in Figure 2, inverse binomial sampling plan S_2 $(\cdot; 0, \infty, m_1, 0)$ is depicted in Figure 3, and generalized inverse binomial sampling plan S_3 $(\cdot; 0, \infty, m_1, m_2)$ is depicted in Figure 4.

Three sampling plans can be devised as modifications of the inverse binomial sampling plan S_2 by assigning some positive values to n_1 and/or n_2. If, in S_2, the sampling terminates at an early stage but taking more samples cost nothing or slightly more, there is no reason why more samples should not be taken until sample size becomes n_1 to gain more low cost information. This sampling plan is S_4 $(\cdot; n_1, \infty, m_1, 0)$ whose boundary points are depicted in Figure 5. On the other hand, if the sampling does not terminate over considerable stages, additional samples are not justified costwise. This sampling plan is S_5 $(\cdot; 0, n_2, m_1, 0)$ and depicted in Figure 6. Combination of the above two sampling plans may also be useful. If the sampling terminates at

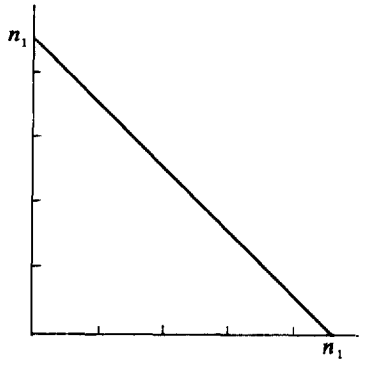

Figure 2
$S_1(\cdot; n_1, n_1, 0, 0)$

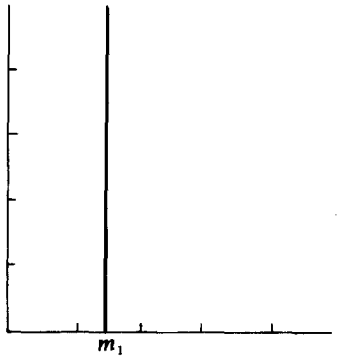

Figure 3
$S_2(\cdot; 0, \infty, m_1, 0)$

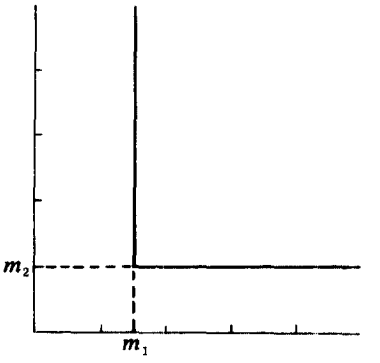

Figure 4
$S_3(\cdot; 0, \infty, m_1, m_2)$

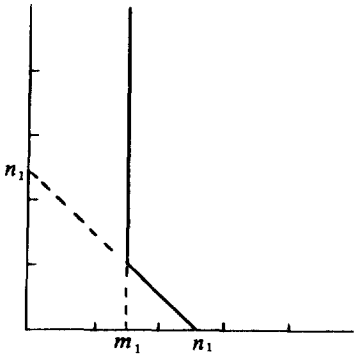

Figure 5
$S_4(\cdot; n_1, \infty, m_1, 0)$

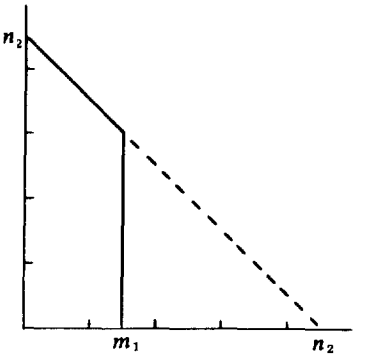

Figure 6
$S_5(\cdot; 0, n_2, m_1, 0)$

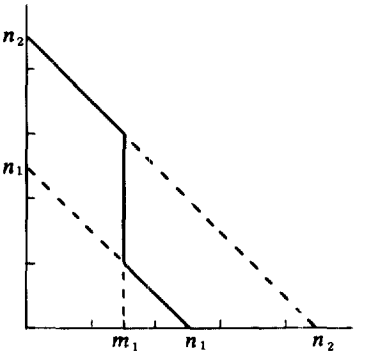

Figure 7
$S_6(\cdot; n_1, n_2, m_1, 0)$

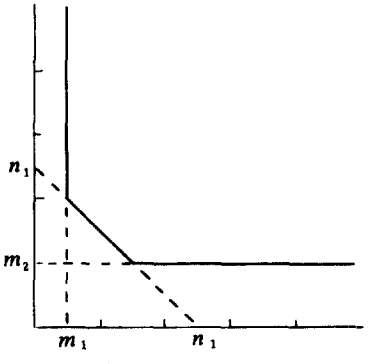

Figure 8
$S_7(\cdot; n_1, \infty, m_1, m_2)$

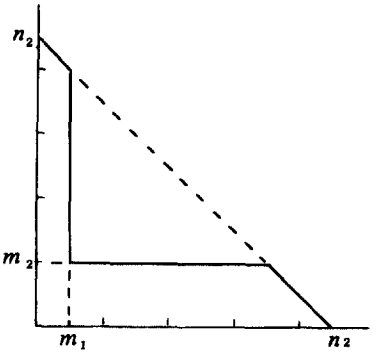

Figure 9
$S_8(\cdot; 0, n_2, m_1, m_2)$

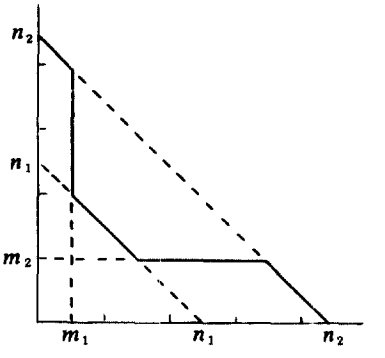

Figure 10
$S_9(\cdot; n_1, n_2, m_1, m_2)$

early stage, we may want to take some more samples and if the sampling does not terminate until a considerable later stages, we may want to curtail the sampling. Then the sampling plan is S_6 $(\cdot;\ n_1,\ n_2,\ m_1,\ 0)$ and depicted in Figure 7.

The above three modifications can also be performed for generalized inverse binomial sampling plan S_3. The corresponding sampling plans are S_7 $(\cdot;\ n_1,\ \infty,\ m_1,\ m_2)$, S_8 $(\cdot;\ 0,\ n_2,\ m_1,\ m_2)$, and S_9 $(\cdot;\ n_1,\ n_2,\ m_1,\ m_2)$ and depicted in Figure 8-10, respectively.

Now consider the mathematical expression of sampling plan S. Note that a sampling plan is a function S defined on points t in 2-dimension Euclidean space as

$$S(t) = \begin{cases} 1, & t \in B \\ 0, & \text{otherwise,} \end{cases}$$

where B is the set of boundary points. For every point t there corresponds sample size $N = N(t) = X(t) + Y(t)$ where $X = X(t)$ is the number of successes. Thus the sampling plan is completely determined by N and X.

Define function ϵ and δ by

$$\epsilon(y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\delta(y) = \begin{cases} 1, & \text{if } y \geqslant 0 \\ 0, & \text{otherwise.} \end{cases}$$

Then the generalized sampling plan S_0 $(\cdot;\ n_1,\ n_2,\ m_1,\ m_2)$ can be expressed as

$$\begin{aligned} & S_0(n, x;\ n_1,\ n_2\ m_1,\ m_2) \qquad (1) \\ = \ & \delta(n-n_2)\,[\epsilon(m_1-1-x) + (x-n_2+m_2-1)] \\ + \ & \epsilon(n_2-n)\,\epsilon(n-n_1-1)\,[\delta(x-m_1) + \delta(x-n+m_2)] \\ + \ & \delta(n-n_1)\,\epsilon(x-m_1)\,\epsilon(n_1-m_2-x). \end{aligned}$$

The mathematical forms of various sampling plans can be derived from (1) by giving corresponding values of n_1, n_2, m_1, and m_2.

III. PROBABILITY DISTRIBUTION AND PARAMETER ESTIMATION

Consider independent Bernoulli random variables, U_1, U_2, . . . such that

$$P[U_i=1] = p, \qquad 0 \leq p \leq 1$$
$$P[U_i=0] = q = 1 - P[U_i=1] = 1 - p, \quad i = 1,2,\ldots$$

Thus the parameter is the success probability p.

When the sampling terminates, we are usually interested in the sample size N and the number of successes X. In addition, suppose we observe another random variable U_N, i.e., observe how the boundary point is reached by its last trial. Though observing another random variable U_N is superfluous in some cases, it serves to simplify and generalize the probability distributions and estimates under various sampling plans as seen below.

The generalized sampling plan defined on the space of (n,x_n) in (1) can also be modified and defined on the space of $(n,x_n,\ u_n)$:

$$\begin{aligned} & S_0(n, x_n, u_n;\ n_1,\ n_2,\ m_1,\ m_2) \qquad (2) \\ = \ & \delta(n-n_2)\,[\epsilon(m_1-1-x_n) + \epsilon(x_n-n+m_2-1)] \\ + \ & \delta(n-n_1)\,\epsilon(x_n-m_1)\,\epsilon(n-m_2-x_n) \\ + \ & \epsilon(n-n_1-1)\,\epsilon(n_2-n) \\ & \cdot\{\delta[u_n(x_n-m_1)] + \delta[(1-u_n)(n-x_n+m_2)]\} \end{aligned}$$

It can easily be seen that the joint distribution of (N,X) under monotone sampling plans is

$$\begin{aligned} & p(n,x) \\ = \ & P[N=n,\ X=x] \\ = \ & \begin{cases} \binom{n-1}{x-1} p^x q^{n-x}, & u_n = 1 \\ \binom{n-1}{x} p^x q^{n-x}, & u_n = 0. \end{cases} \end{aligned}$$

Hence the joint distribution of (N,X,U_N) under generalized sampling plan S_0 is obtained as

$$\begin{aligned} & P_{S_0}(n, x, u\ ;n_1, n_2, m_1, m_2, p) \qquad (3) \\ = \ & P[N=n,\ X=x,\ U_N=u_n \,/\, S_0] \\ = \ & \binom{n-1}{x-u_n} p^x q^{n-x}, \qquad (n, x, u_n) \in B_0 \end{aligned}$$

where B_0 is the set of boundary points under S_0.

From the form of the joint distribution in (3), it can be seen that (N, X, U_N) forms a suffici-

ent statistic for parameter p. The marginal distribution $p_{S_0}(x; n_1, n_2, m_1, m_2, p)$ of X, the marginal distribution $p_{S_0}(n; n_1, n_2, m_1, m_2, p)$ of N and k-th ascending factorial moment $E_{S_0}(N^{[k]}; n_1, n_2, m_1, m_2, p)$ of N will be derived. These equations will be denoted by $p_{S_0}(x)$, $p_{S_0}(n)$, and $E_{S_0}(N^{[k]})$, respectively, when there is no possibility of confusion.

We have

$$p_{S_0}(x) = P[X=x \mid S_0] = \sum_{u_n=0}^{1} \sum_{n=0}^{\infty} S_0(n, x, u_n) \binom{n-1}{x-u_n} p^x q^{n-x}$$

and

$$p_{S_0}(n) = P[N=n \mid S_0] = \sum_{u_n=0}^{1} \sum_{n=0}^{n} S_0(n, x, u_n) \binom{n-1}{x-u_n} p^x q^{n-x}$$

Let $n^{[k]} = n(n+1)\ldots(n+k-1)$.

The k-th ascending factorial moment of N is given by

$$\begin{aligned} &E_{S_0}(N^{[k]}; n_1, n_2, m_1, m_2, p) \\ &= \sum_{n=0}^{\infty} n^{[k]} p_{S_0}(n) \\ &= E_{S_3}(N^{[k]}; 0, \infty, m_1, m_2, p) \\ &\quad - \Big[\sum_{n=m_1+m_2}^{n_1} + \sum_{n=n_2+1}^{\infty}\Big] n^{[k]} p_{S_3}(n; 0, \infty, m_1, m_2, p) \\ &\quad + \sum_{x=m_1}^{n_1-m_2} n_1^{[k]} p_{S_1}(x; n_1, n_1, 0, 0) \\ &\quad + \Big[\sum_{x=0}^{m_1-1} + \sum_{x=n_2-m_2-1}^{n_2}\Big] n_2^{[k]} p_{S_1}(x; n_2, n_2, 0, 0) \end{aligned}$$

where

$$\begin{aligned} &E_{S_3}(N^{[k]}; 0, \infty, m_1, m_2, p) \qquad (4) \\ &= E_{S_2}(N^{[k]}; 0, \infty, m_1, 0, p) + E_{S_2}(N^{[k]}; 0, \infty, \\ &\quad m_2, 0, q) - \sum_{n=0}^{n_1+m_2-1} n^{[k]} [p_{S_2}(n; 0, \infty, m_1, 0, p) \\ &\qquad + p_{S_2}(n; 0, \infty, m_2, 0, 1)] \end{aligned}$$

and

$$E_{S_2}(N^{[k]}; 0, \infty, m, 0, p) = m^{[k]}/p^k$$

which is the quantity given by Haldane [4]. In fact it can be seen that substuting k = 1 in (4) we obtain the expected sample size under S_3 as

$$\begin{aligned} &E_{S_3}(N; 0, \infty\ m_1, m_2, p) \\ &= \frac{m_1}{p}[1 - I_p(m_1, m_2)] + \frac{m_2}{p}[1 - I_q(m_1, m_2)], \end{aligned}$$

where $I_w(a,b)$ is the incomplete beta function as tabulated by Pearson [5].

Utilizing a method of Girshick, et al. [3], we obtain an unbiased estimate of p as

$$\hat{p}_{S_0}(n_1, n_2, m_1, m_2) = \frac{x-u_n}{n-1}. \quad (n,x,u_n) \in B_0. \quad (5)$$

In particular, from (2) and (5) we have, under S_3,

$$\hat{p}_{S_3}(0, \infty, m_1, m_2) = \frac{m-1}{n-1} u_n + \frac{n-m}{n-1}(1-u_n), \quad (n,x,u_n) \in B_3,$$

which is the same as the one proposed by Bai, et al. [1]. We also have, under S_2,

$$\hat{p}_{S_2}(0, \infty, m, 0) = \frac{m-1}{n-1} \quad (n,m,u_n) \in B_2,$$

which is the same as the one obtained by Haldane [4].

The variance of $\hat{p}_{S_0}$ is obtained as follows.

$$\begin{aligned} &Var_{S_0}(\hat{p}; n_1, n_2, m_1, m_2, p) \\ &= Var_{S_3}(\hat{p}) - \Big(\sum_{n=m_1+m_2}^{n_1} \sum_{u_n=0}^{1} + \sum_{n=m_2+1}^{\infty} \sum_{u_n=0}^{1}\Big) \hat{p}_{S_3}^2 p_{S_3}(n,m,u_n) \\ &\quad + \sum_{x=m}^{n_1-m_1} \hat{p}_{S_1}^2(n_1, n_1, 0, 0) p_{S_1}(x; n_1, n_1, 0, 0, p) \\ &\quad + \Big(\sum_{x=0}^{m_1-1} + \sum_{x=n_1-m_2+1}^{\infty}\Big) \hat{p}_{S_1}^2(n_2, n_2, 0, 0) p_{S_1}(x; n_2, n_2, \\ &\quad 0, 0) \end{aligned}$$

where

$$\begin{aligned} &Var_{S_3}(\hat{p}; n_1, n_2, m_1, m_2, p) \\ &= Var_{S_2}(\hat{p}; 0, \infty, m_1, 0, p) + Var_{S_2}(\hat{p}; 0, \infty, m_2, \\ &0, q) - \sum_{n=0}^{m_1+m_2-1} [\hat{p}_{S_2}(o, \infty, m_1, 0) p_{S_2}(n; 0, \infty, m_1, \\ &0, p) + \hat{p}_{S_2}(0, \infty, m_2, 0) p_{S_2}(n; 0, \infty, m_2, 0, q)] \\ &+ p^2 \end{aligned}$$

and

$$\begin{aligned} &Var_{S_2}(\hat{p}; 0, \infty, m, 0, p) \\ &= p \sum_{j=0}^{\infty} \binom{m-1+j-1}{j} q^j \\ &= \frac{p^2 q}{m}\left[1 + \frac{2}{m+1} q + \frac{3}{(m+1)(m+2)} q^2 + \ldots\right] \end{aligned}$$

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

which is given by Haldane [4], whereas DeGroot [2] expresses it in the form of

$$\frac{(m-1)p^m}{q^{m-1}}\left[(-1)^{m-1}\log p + \sum_{i=1}^{m-2}\frac{(-1)^{m-1}}{i}\left(\frac{q}{p}\right)^i\right] - p^2.$$

IV. RELATIVE EFFICIENCIES

Restriction of S_2 on minimum number of failures m_2 gives rise to the sampling plan S_3. In the same way restrictions of S_4, S_5, and S_6 on minimum number of failures give rise to sampling plans, S_7, S_8, and S_9, respectively. For the relative efficiency of S_3 to S_2, see Bai, et al. [1]. Similar results are expected to hold in comparing S_4 with S_7, S_5 with S_8, and S_6 with S_9.

Restriction of S_2 on minimum sample size n_1 gives rise to the sampling plan S_4. In the same way restrictions of S_3, S_5 and S_8 on minimum sample size give rise to the sampling plans, S_4, S_6 and S_9, respectively. Only sampling plans S_2 and S_4 will be compared. Similar results are expected to hold in comparing S_3 with S_7, S_5 with S_6, and S_8 with S_9.

In Figure 11 and Figure 12, $E(N)$ and $Var(\hat{p})$ are plotted as functions of p and selected values of n_1 and for given values m. Figure 13-17 show the values of $E(N)$ and $Var(\hat{p})$ as n_1 varies for selected values of m and p.

From Figure 11 to Figure 17, it can be seen that adding the requirement of minimum sample size n_1, particularly small values of n_1, tends to make sampling plan S_4 more efficient for all values of p, especially for intermediate values of p. The figures also show that for small values of p,

$$E_{S_2}(N; 0, \infty, m^*, 0) \approx E_{S_4}(N; n_1, \infty, m^*, 0)$$

and

$$Var_{S_2}(\hat{p}; 0, \infty, m^*, 0) \approx Var_{S_4}(\hat{p}; n_1, \infty, m^*, 0)$$

and for large values of p,

$$E_{S_2}(N; n_1^*, \infty, m, 0) \approx E_{S_4}(N; n_1^*, \infty, m^*, 0)$$

and

$$Var_{S_2}(\hat{p}; n_1^*, \infty, m, 0) \approx Var_{S_4}(\hat{p}; n_1^*, \infty, m^*, 0)$$

where stars indicate the parameters fixed. For intermediate values of p, $Var_{S_4}(\hat{p})$ decreases sharply as n_1 increases while $E_{S_4}(N)$ increases moderately.

Therefore, $S_4(\cdot; n_1, \infty, m, 0)$ can be advantageously used, especially when n_1 is small and p is moderate, in place of $S_2(\cdot; 0, \infty, m, 0)$ if i) high precision is required in estimating p, ii) sampling is not expensive, and iii) it is known a priori that p is not small.

Now consider the effects of maximum sample size n_2. Restriction of S_2 on maximum sample size gives rise to the sampling plan S_5. In the same way restrictions of S_3, S_4 and S_7 on maximum sample size give rise to the sampling plans, S_5, S_6 and S_9, respectively. Only sampling plans S_2 and S_5 will be compared. Similar results are expected to hold in comparing S_3 with S_8, S_4 with S_6, and S_7 with S_9.

In Figure 18 and Figure 19, $E(N)$ and $Var(\hat{p})$ are plotted as functions of p and selected values of n_2 and for given values m. Figure 19-24 show the values of $E(N)$ and $Var(\hat{p})$ as n_2 varies for selected values of m and p.

From Figure 18 to Figure 24, it can be seen that adding the requirement of maximum sample size n_2, particularly small values of n_2, tends to make sampling plan S_2 less efficient for all values of p, especially for intermediate values of p. The figures also show that for large values of p,

$$E_{S_2}(N; 0, \infty, m^*, 0) \approx E_{S_5}(N; 0, n_2, m^*, 0)$$

and

$$Var_{S_2}(\hat{p}; 0, \infty, m^*, 0) \approx Var_{S_5}(\hat{p}; 0, n_2, m^*, 0)$$

and for small values of p

$$E_{S_2}(N; 0, n_2^*, m, 0) \approx E_{S_5}(n; 0, n_2^*, m^*, 0)$$

and

$$Var_{S_2}(\hat{p}; 0, n_2^*, m, 0) \approx Var_{S_5}(\hat{p}; 0, n_2^*, m^*, 0$$

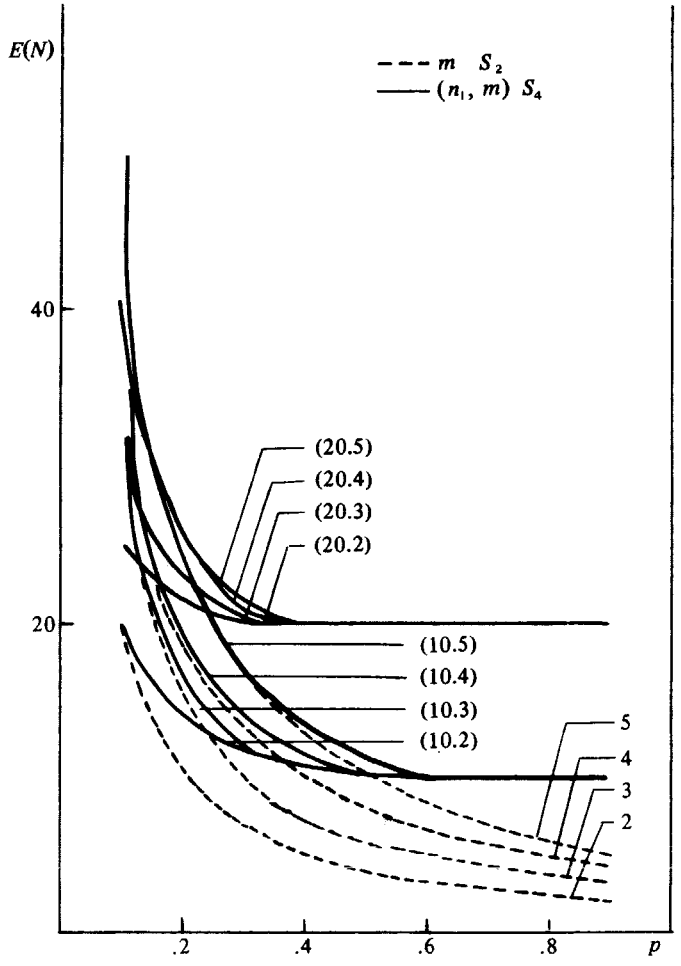

Figure 11
E(N) under S_2 and S_4

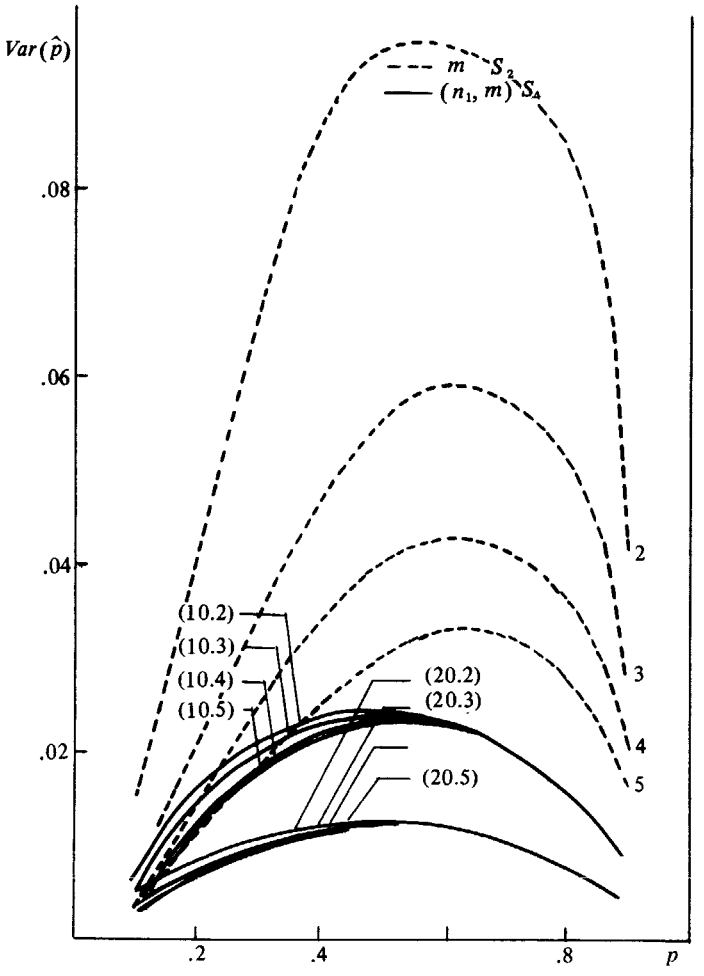

Figure 12
Var $(\hat{p})$ under S_2 and S_4

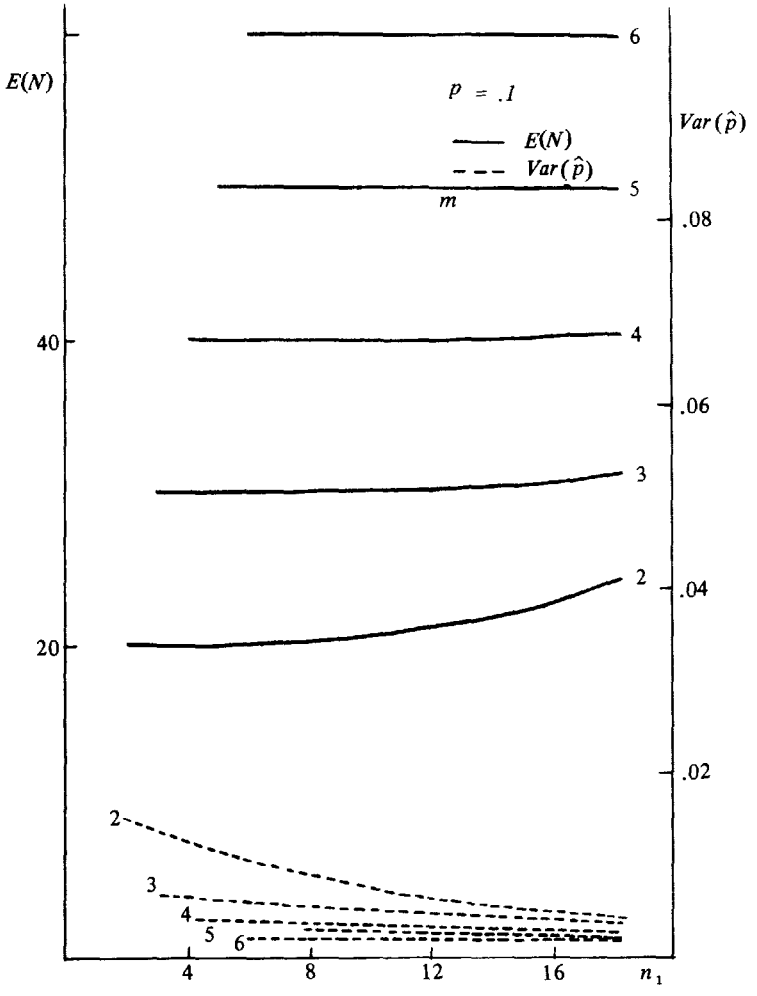

Figure 13
$E(N)$ and $Var(\hat{p})$ under S_2 *and* S_4
for $p = .1$

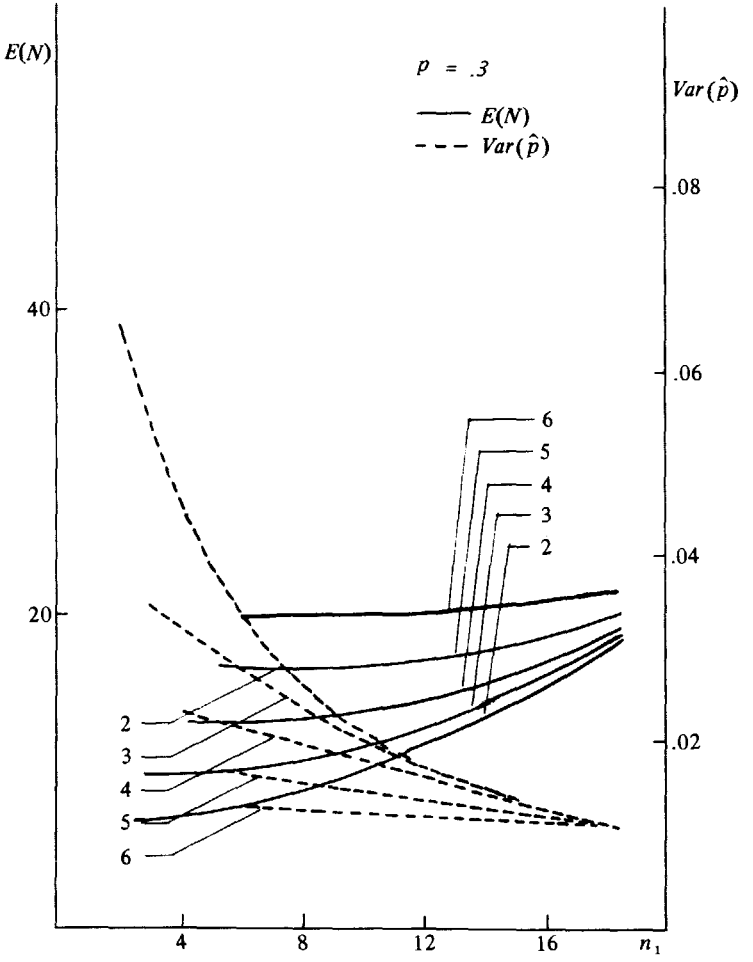

Figure 14
$E(N)$ and $Var(\hat{p})$ under S_2 *and* S_4
for $p = .3$

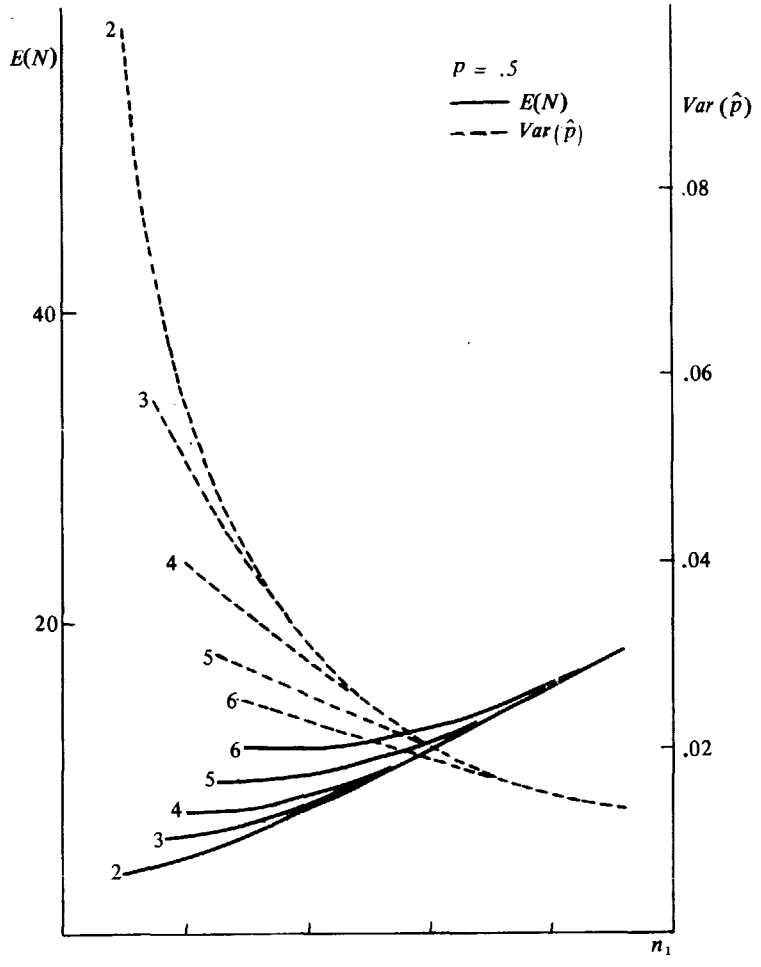

Figure 15
$E(N)$ and $Var(\hat{p})$ under S_2 and S_4
for $p = .5$

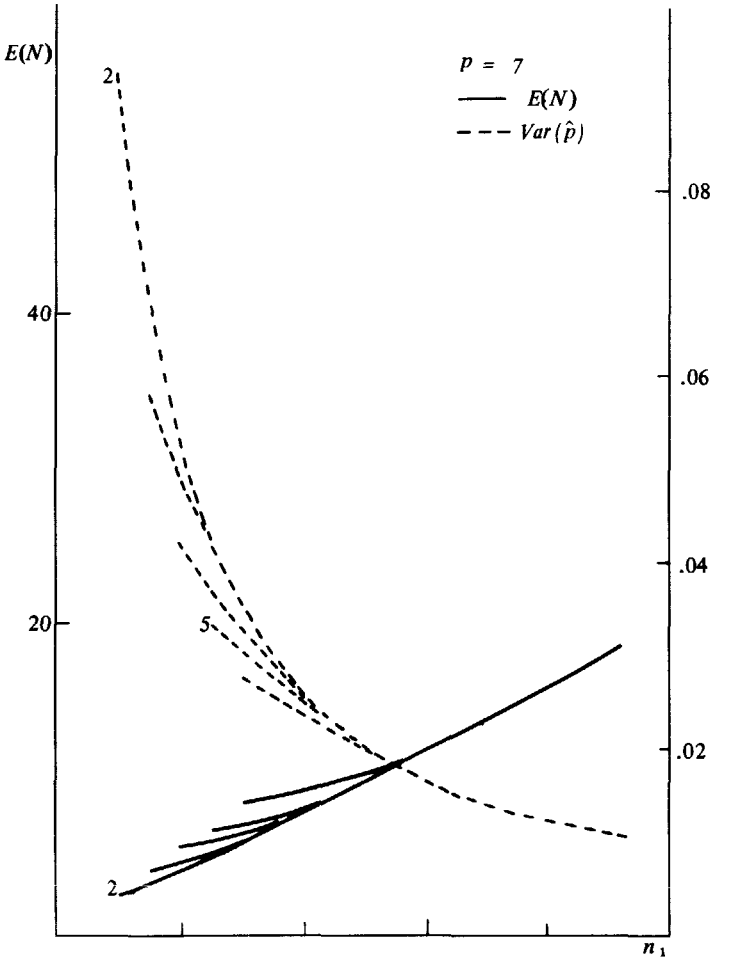

Figure 16
$E(N)$ and $Var(\hat{p})$ under S_2 *and* S_4
for $p = .7$

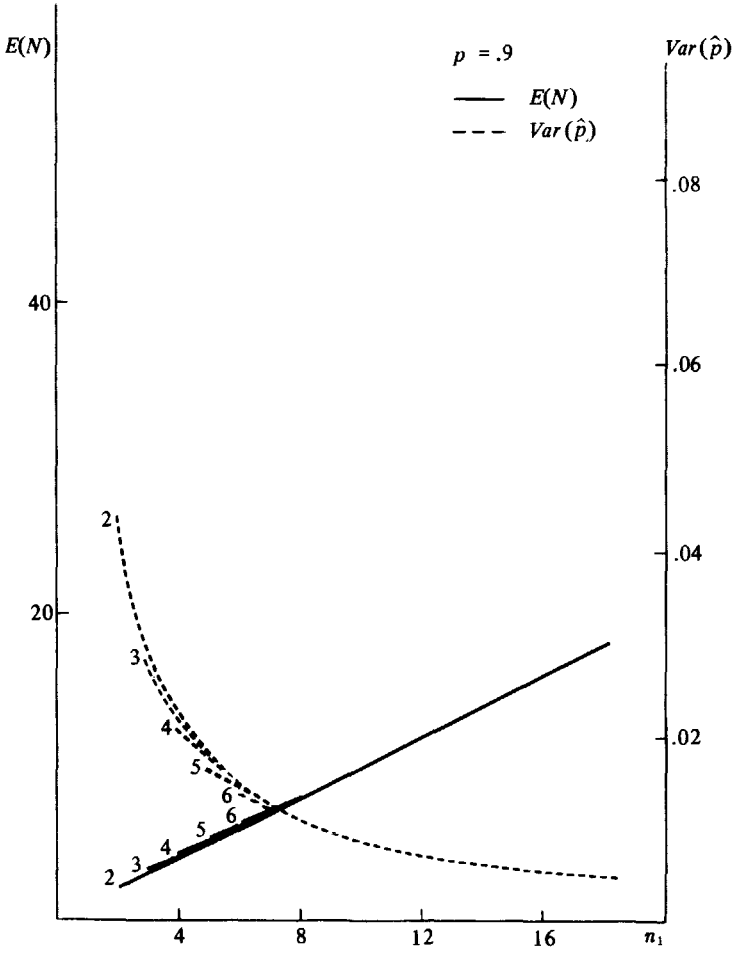

Figure 17
$E(N)$ and $Var(\hat{p})$ under S_2 and S_4 for $p = .9$

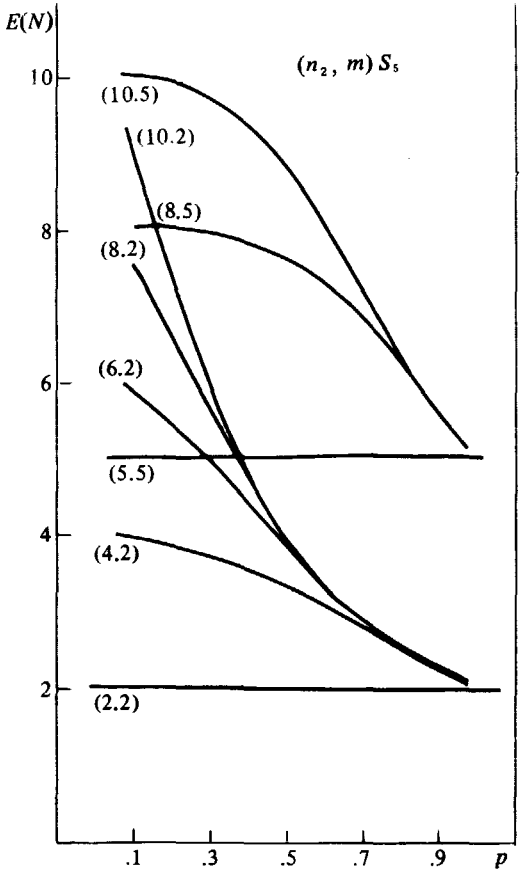

Figure 18
$E(N)$ under S_2 and S_5

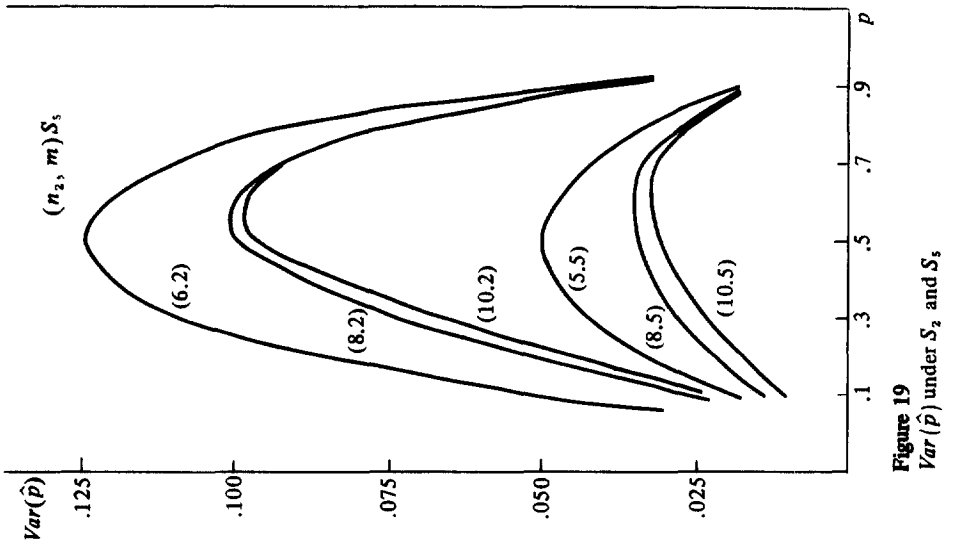

Figure 19
$Var(\hat{p})$ under S_2 and S_5

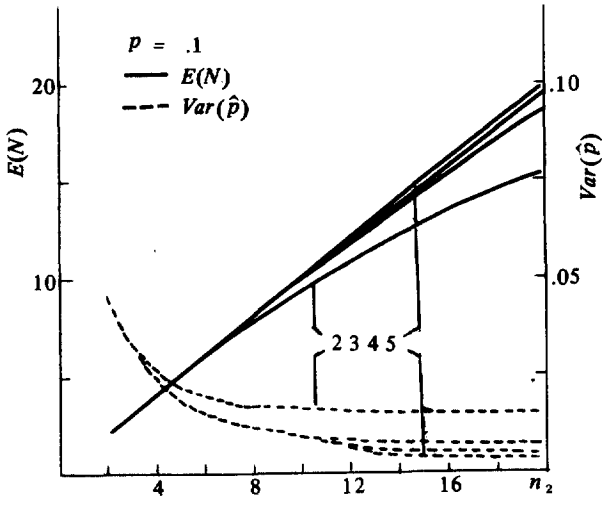

Figure 20
$E(N)$ and $Var(\hat{p})$ under S_2 and S_5 for p = .1

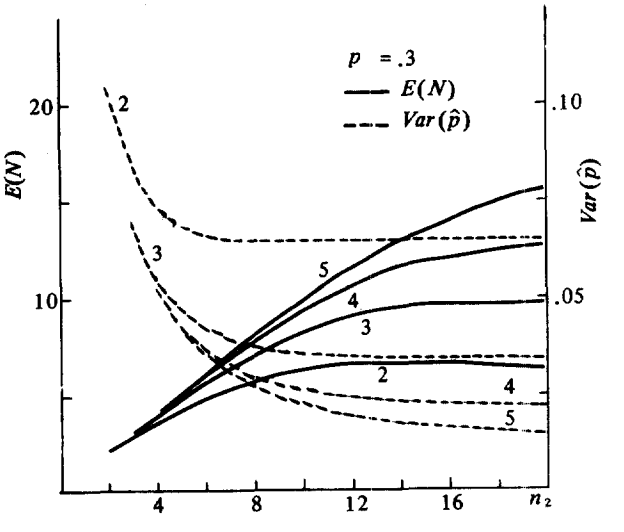

Figure 21
$E(N)$ and $Var(\hat{p})$ under S_2 and S_5 for p = .3

and $E_{S_5}(N)$ decreases sharply as n decreases while $Var_{S_5}(\hat{p})$ is relatively constant. Therefore, S_5 (·; 0, n_2, m, 0) can be advantageously used, especially n is moderate and p is small, in place of S_7 (·; 0, ∞, m, 0) if i) high precision is not required in estimating p, ii) sampling is expensive, and iii) it is known a priori that p is not large.

V. CONCULSIONS

The generalized sampling plan is characterized by four conditions, n_1 minimum sample size, n_2 maximum sample size, m_1 minimum number of successes and m_2 minimum number of failures. Combinations of the values of n_1, n_2, m_1 and m_2

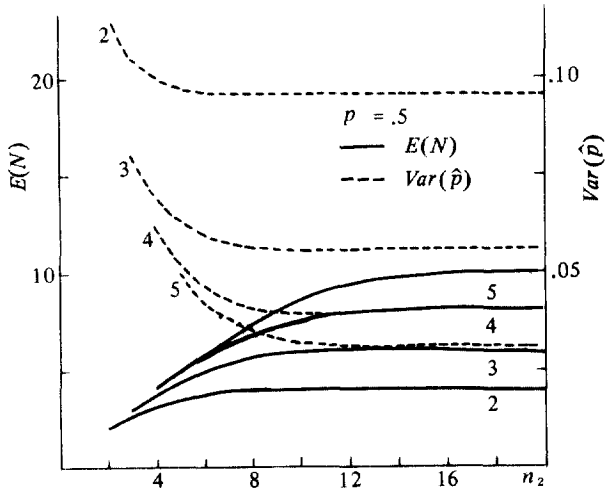

Figure 22
$E(N)$ and $Var(\hat{p})$ under S_2 and S_5 for $p = .5$

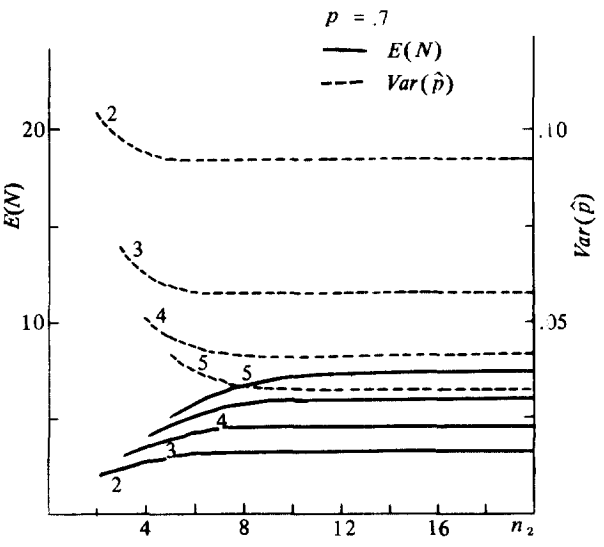

Figure 23
$E(N)$ and $Var(\hat{p})$ under S_2 and S_5 for $p = .7$

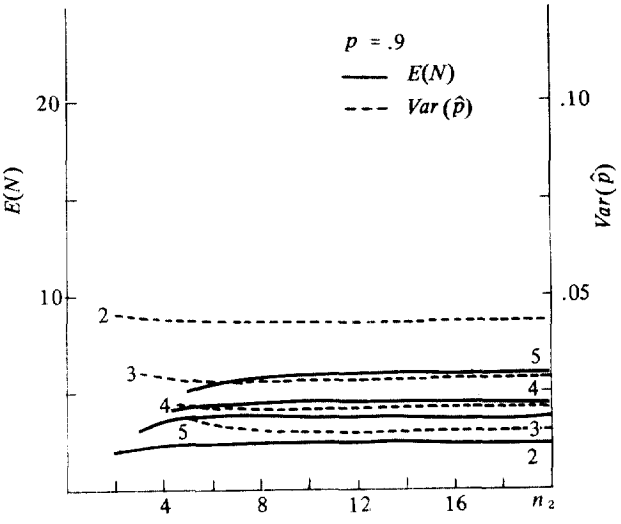

Figure 24
$E(N)$ and $Var(\hat{p})$ under S_2 and S_5 for $p = .9$

yeild various sampling plans.

Sufficient statistics, its probability distributions, moments, unbiased estimates and variances under various sampling plans can be obtained as special cases of the generalized sampling plan.

All of the restrictions, n_1, n_2, m_1 and m_2 serve to make a sampling plan more efficient. These values can be determined by finding the break-even point between factors such as the accuracy and precision of estimates, sample costs, efforts, time, applicability, etc.

REFERENCES

1. Bai, D.S., Kim, S.I. and Lee, J.K.; "On a Generalized Inverse Binomial Sampling Plan"; *The Journal of the Korean Statistical Society*, Vol. 1, 1977, pp. 1-20.
2. DeGroot, M.H.; "Unbiased Sequential Estimation for Binomial Population"; *Annals of Mathematical Statistics*, Vol. 30, 1959, pp. 80-101.
3. Girshick, M.A., Mosteller, F. and Savage, L. T.; "Unbiased Estimation for Certain Binomial Sampling Problems with Applications"; *Annals of Mathematical Statistics*, Vol. 13, 1946, pp. 13-23.
4. Haldane, J.B.S.; "On a method of Estimating Frequencies"; *Biometrika*, Vol. 33, 1945, pp. 222-225.
5. Pearson, K.; *Tables of Incomplete Beta Function;* Cambridge University Press; London: 1934.