POLYCHOTOMY SAMPLING[1]

By Sakti P. Ghosh

Thomas J. Watson Research Center, New York

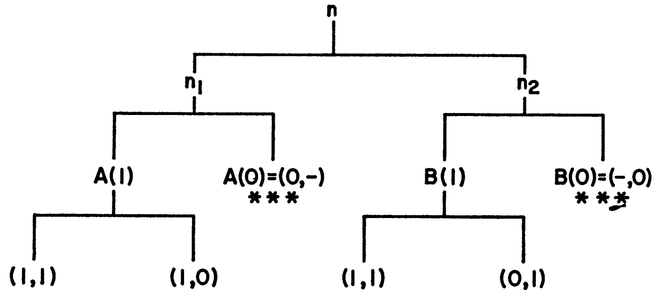
1. Introduction. In sampling, one sometimes has to deal with a particular type of binary character known as semi-observation-destructive (SOD) i.e., the action of observing the particular character in the unit may result in destroying the unit or leaving the unit unchanged. Such situations are encountered very often in quality control or biological experiments, etc. So long as one is interested in only one character of the unit, the situation does not present much of a problem; but when one is interested in more than one SOD character, the situation becomes more complex because no further observations can be made on units which are destroyed while a character is observed. Thus some combination of characters cannot be observed and yet an estimate of their proportion may be of interest. The aim of this paper is to illustrate how by splitting a single sample into a number of subsamples and then making different types of observations on different subsamples and by using the properties of Boolean algebra, estimates of all possible combinations of characters can be built up. This problem with two characters was first treated by Dalenius (1959). In that paper, the bivariate case was discussed in detail, and it was also pointed out that the design could be generalized to the case with three or more variates. The aim of the present paper is to derive the variances of the different estimates possible. In order to make the paper self-contained, an account is given of the content of Dalenius (1959), which is written in Swedish and thus not generally available. Here the bivariate situation will be treated first and the variances of the estimates given by Dalenius will be presented. Some other new estimates with their variances will also be presented. The sampling scheme will then be generalized to trivariate situations. The multivariate generalization will not be treated because it will be a simple extension of trivariate.

2. Bivariate sampling scheme and the estimates. Suppose A and B are two binary characters of a bivariate population (A, B), i.e., there are only four different types of units viz (1, 1), (1, 0), (0, 1) and (0, 0). When one character is being observed, the outcome will be denoted by $A(0)$ or $A(1)$ or $B(0)$ or $B(1)$ as the case may be. If $A(0)$ is observed, then the unit is destroyed and no further observation can be made on B. Similarly if $B(0)$ is observed then the unit is destroyed and no further observation can be made on A. The problem is to estimate the proportion of (0, 0) in the population.

Suppose the total sample size available is n. n is divided into two parts n_1 and n_2, i.e., $n_1 + n_2 = n$. Then two samples of sizes n_1 and n_2 are drawn from the population without replacement. In the sample n_1, the character A is first ob-

Received 12 May 1964; revised 19 August 1965.

[1] This work was supported in part by Air Force Research Contract AF 19(626)-10.



(*** INDICATES TERMINATION OF OBSERVATION)

FIG. 1

served, i.e., $A(1)$ and $A(0)$. All the units which had $A(0)$ are destroyed and hence no further observations can be made on them. Hence, among the units which had $A(1)$, B is observed and thus the number of units which have (1, 0) and (1, 1) are obtained. Similarly, in the sample n_2, B is first observed, i.e., $B(0)$ and $B(1)$ and then among the $B(1)$'s (1, 1) and (0, 1) are obtained. The sampling scheme is given in Figure 1. Suppose

(1)
$n(1, 1)$ is the frequency of (1, 1) in the combined sample n.
$n_1(1, 0)$ is the frequency of (1, 0) in n_1.
$n_1(0, -)$ is the frequency of (0, −) in n_1.
$n_2(0, 1)$ is the frequency of (0, 1) in n_2.
$n_2(-, 0)$ is the frequency of (−, 0) in n_2.

$$p(1, 1) = n(1, 1)/n, \qquad p_1(1, 0) = n_1(1, 0)/n_1,$$
$$p_2(0, 1) = n_2(0, 1)/n_2, \qquad p_1(0, -) = n_1(0, -)/n_1, \tag{2}$$
$$p_2(-, 0) = n_2(-, 0)/n_2.$$

Similarly $P(i, j)$ shall denote the proportion of the character (i, j) in the population where the symbols i, j can represent 1, 0 or −. In the next few lines a little of set theory will be used and there (i, j) will mean the set of points which have the character (i, j).

From the self-evident pictorial diagram of sets in Figure 2, the estimates will be derived.

The following set unions

$$(0, -) = (0, 1) \cup (0, 0),$$
$$(-, 0) = (1, 0) \cup (0, 0),$$
$$\Omega = (1, 1) \cup (0, -) \cup (-, 0),$$
$$\Omega = (1, 1) \cup (1, 0) \cup (0, 1) \cup (0, 0),$$

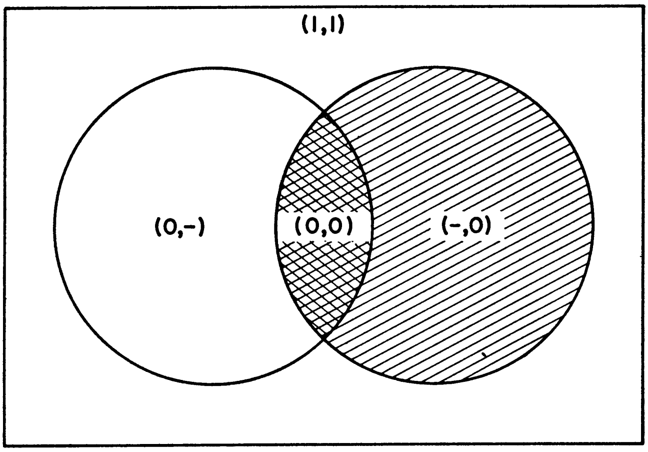


FIG. 2

suggest four estimates which were given by Dalenius (1959):

$$\hat{p}_{D1} = p_1(0, -) - p_2(0, 1), \tag{3}$$

$$\hat{p}_{D2} = p_2(-, 0) - p_1(1, 0), \tag{4}$$

$$\hat{p}_{G1} = p(1, 1) + p_1(0, -) + p_2(-, 0) - 1, \tag{5}$$

$$\hat{p}_{G2} = 1 - p(1, 1) - p_2(0, 1) - p_1(1, 0). \tag{6}$$

3. Variances of the estimates of $P(0, 0)$. The Equations (3), (4), (5) and (6) give four different estimates of $P(0, 0)$ and the form of their variances will be discussed in this section. $n_1(1, 1)$, $n_1(1, 0)$ and $n_1(0, -)$ will follow a multinomial distribution in the sample of size n_1 and similarly $n_2(1, 1)$, $n_2(0, 1)$ and $n_2(-, 0)$ will follow a multinomial distribution in the sample of size n_2. When the population is assumed to be infinite then

$$\begin{aligned} V(p(1, 1)) &= P(1, 1)Q(1, 1)/n, \\ V(p_1(1, 0)) &= P(1, 0)Q(1, 0)/n_1, \\ V(p_2(0, 1)) &= P(0, 1)Q(0, 1)/n_2, \\ \operatorname{Cov}(p(1, 1), p_1(1, 0)) &= -P(1, 1)P(1, 0)/n, \\ \operatorname{Cov}(p(1, 1), p_2(0, 1)) &= -P(1, 1)P(0, 1)/n, \end{aligned} \tag{7}$$

where $Q = 1 - P$.

If the population is finite of size N then the above variances have to be multiplied by proper finite population corrections (fpc). The fpc for $V(p(1, 1))$ will be $(N - n)/(N - 1)$. The fpc for the others will depend on whether n_1 was drawn first or n_2 was drawn first. If n_1 was drawn first then

$$\begin{aligned} V(p_1(1, 0)) &= [P(1, 0)Q(1, 0)/n_1]\cdot(N - n_1)/(N - 1), \\ V(p_2(0, 1)) &= [P(0, 1)Q(0, 1)/n_2] \\ &\quad\cdot(N - n_1 - n_2)/(N - n_1 - 1), \\ \text{Cov}\,(p(1, 1), p_1(1, 0)) &= [-P(1, 1)P(1, 0)/n]\cdot(N - n_1)/(N - 1), \\ \text{Cov}\,(p(1, 1), p_2(0, 1)) &= [-P(1, 1)P(0, 1)/n] \\ &\quad\cdot(N - n_1 - n_2)/(N - n_1 - 1). \end{aligned} \tag{8}$$

If n_2 was drawn first then

$$\begin{aligned} V(p_1(1, 0)) &= [P(1, 0)Q(1, 0)/n_1] \\ &\quad\cdot(N - n_1 - n_2)/(N - n_2 - 1), \\ V(p_2(0, 1)) &= [P(0, 1)Q(0, 1)/n_2]\cdot(N - n_2)/(N - 1), \\ \text{Cov}\,(p(1, 1), p(1, 0)) &= [-P(1, 1)P(1, 0)/n] \\ &\quad\cdot(N - n_1 - n_2)/(N - n_2 - 1), \\ \text{Cov}\,(p(1, 1), p_2(0, 1)) &= [-P(1, 1)P(0, 1)/n]\cdot(N - n_2)/(N - 1). \end{aligned} \tag{9}$$

For algebraic simplicity the effect of fpc will not be discussed in the subsequent results. In deriving the variances of the estimates a few more symbols will be needed. Though $(0, 0)$ could not be observed in the sample they were present all the time. Suppose

$n_1(0, 0)$ is the frequency of $(0, 0)$ in n_1,

$n_2(0, 0)$ is the frequency of $(0, 0)$ in n_2,

$n(0, 0)$ is the frequency of $(0, 0)$ in n,

$p_1(0, 0) = n_1(0, 0)/n_1, \qquad p_2(0, 0) = n_2(0, 0)/n_2.$

The following lemma will also be of much use in deriving the variances:

LEMMA 1. *Partition of a binomial variable leads to two uncorrelated variables.*

PROOF. Suppose X is $b(n, \pi)$ and n is partitioned into two parts such that $n_1 + n_2 = n$. Suppose X_1 and X_2 are the random variables denoting the binomial character in n_1 and n_2, respectively. Thus $X = X_1 + X_2$. Hence

$$V(X) = V(X_1) + V(X_2) + 2\,\text{Cov}\,(X_1, X_2)$$

or

$$n\pi(1 - \pi) = n_1\pi(1 - \pi) + n_2\pi(1 - \pi) + 2\,\text{Cov}\,(X_1, X_2) \Leftrightarrow \text{Cov}\,(X_1, X_2) = 0.$$

The lemma is proved.

It is easy to see that (3) can be written as

$$\hat{p}_{D1}(0, 0) = p_1(0, 0) + p_1(0, 1) - p_2(0, 1). \tag{10}$$

By application of Lemma 1 it can be shown that $p_1(0, 1)$ and $p_2(0, 1)$ are uncorrelated.

From (10) it follows that

$$\begin{aligned} V(\hat{p}_{D1}(0, 0)) &= P(0, 0)Q(0, 0)/n_1 + P(0, 1)Q(0, 1)/n_1 \\ &\quad + P(0, 1)Q(0, 1)/n_2 - 2P(0, 0)P(0, 1)/n_1 \\ &= P(0, 0)Q(0, 0)/n_1 + (n/n_1n_2)P(0, 1)Q(0, 1) \\ &\quad - 2P(0, 0)P(0, 1)/n_1 . \end{aligned} \tag{11}$$

Proceeding exactly similarly the variance of $\hat{p}_{D2}(0, 0)$ can be obtained as follows:

$$\begin{aligned} V(\hat{p}_{D2}(0, 0)) &= P(0, 0)Q(0, 0)/n_2 \\ &\quad + (n/n_1n_2)P(1, 0)Q(1, 0) - 2P(0, 0)P(1, 0)/n_2 . \end{aligned} \tag{12}$$

It is easy to show that

$$\begin{aligned} p(0, -) + p(-, 0) &= n_1(0, 0)/n_1 + n_1(0, 1)/n_1 + n_2(0, 1)/n_2 + n_2(0, 0)/n_2 \\ &= p_1(0, 0) + p_2(0, 0) + p_1(0, 1) + p_2(1, 0). \end{aligned}$$

Hence, from (5) it follows,

$$\hat{p}_{G1}(0, 0) = p(1, 1) + p_1(0, 0) + p_2(0, 0) + p_1(0, 1) + p_2(1, 0) - 1.$$

Taking the variance and simplifying one gets

$$\begin{aligned} V(\hat{p}_{G1}(0,0)) &= (n/n_1n_2)P(0,0)Q(0,0) + P(1,0)[Q(1,0) - 2P(0,0)]/n_2 \\ &\quad + P(0, 1)[Q(0, 1) - 2P(0, 0)]/n_1 \\ &\quad - P(1, 1)Q(1, 1)/n - 2P(1, 1)P(0, 0)/n. \end{aligned} \tag{13}$$

The variance of $\hat{p}_{G2}(0, 0)$ follows directly from (6) by taking variance of both sides and is given by

$$\begin{aligned} V(\hat{p}_{G2}(0, 0)) &= P(1, 1)Q(1, 1)/n + P(0, 1)Q(0, 1)/n_2 \\ &\quad + P(1, 0)Q(1, 0)/n_1 - 2P(1, 1)[P(0, 1) + P(1, 0)]/n. \end{aligned} \tag{14}$$

Another form for $V(\hat{p}_{G2}(0, 0))$: It is obvious that $p(1, 1) = 1 - p(0, 1) - p(1, 0) - p(0, 0)$. Substituting in (6) one gets $\hat{p}_{G2}(0, 0) = p(0, 1) + p(1, 0) + p(0, 0) + p_2(0, 1) - p(1, 0)$. Now

$$p(0, 1) - p_2(0, 1) = (n_1/n)p_1(0, 1) - (n_1/n)p_2(0, 1),$$

$$p(1, 0) - p_1(0, 1) = (n_2/n)p_2(1, 0) - (n_2/n)p_1(1, 0).$$

Hence

$$\hat{p}_{G2}(0, 0) = (n_1/n)[p_1(0, 1) - p_2(0, 1)] + (n_2/n)[p_2(1, 0) - p_1(1, 0)].$$

Taking variance of both sides, making use of Lemma 1 and after simplification it follows,

$$(14') \quad V(\hat{p}_{G2}(0, 0)) = P(0, 0)Q(0, 0)/n + (n_1/nn_2)P(0, 1)Q(0, 1) + (n_2/nn_1)P(1, 0)\, Q(1, 0) + 2P(0, 1)P(1, 0)/n.$$

On comparing (11), (12), (13) and (14) it is difficult to say in general terms which of the four estimates is superior but it is obvious that the variances of (3), (4) and (5) contains $P(0, 0)$, whose magnitude is unknown and even cannot be estimated from conventional sampling procedure. In this respect $\hat{p}_{G2}$ is superior to the other three. In the trivial situation when $n_1 \approx n_2$ and $P(0, 1) \approx P(1, 0) \approx P(1, 1) \approx P(0, 0)$ it is expected

$$(15) \quad V(\hat{p}_{G2}(0, 0)) \approx V(\hat{p}_{G1}(0, 0)) \leqq V(\hat{p}_{D1}(0, 0)) \approx V(\hat{p}_{D2}(0, 0)).$$

Various remarks can be made in special cases about the relative superiority of these four estimates, but from all practical purposes it appears that $\hat{p}_{G2}(0, 0)$ is the best.

Maximum likelihood estimate (mle) *for bivariate sampling scheme.* The mle of $P(0, 0)$ will be discussed here. The likelihood function is given by

$$(16) \quad \begin{aligned} L = {} & [n_1 !/n_1(1, 1)!\, n_1(1, 0)!\, n_1(0, -)!][P(1, 1)]^{n_1(1,1)}[P(1, 0)]^{n_1(1,0)} \\ & \cdot[P(0, 0) + P(0, 1)]^{n_1(0,-)}\cdot[n_2 !/n_2(1, 1)!\, n_2(0, 1)!\, n_2(-, 0)!] \\ & \cdot[P(1, 1)]^{n_2(1,1)}[P(0, 1)]^{n_2(0,1)}[P(0, 0) + P(1, 0)]^{n_2(-,0)}. \end{aligned}$$

The likelihood equations are obtained by differentiating $\ln L - \lambda[(P(0, 0) + P(0, 1) + P(1, 0) + P(1, 1) - 1]$ with respect to the parameters and equating them to zero. The likelihood equations are given by

$$(17) \quad \begin{aligned} n_1(0, -)/[\hat{P}(0, 0) + \hat{P}(0, 1)] + n_2(-, 0)/[\hat{P}(0, 0) + \hat{P}(1, 0)] - \lambda &= 0, \\ n_1(0, -)/[\hat{P}(0, 0) + \hat{P}(0, 1)] + n_2(0, 1)/\hat{P}(0, 1) - \lambda &= 0, \\ n_2(-, 0)/[\hat{P}(0, 0) + \hat{P}(1, 0)] + n_1(1, 0)/\hat{P}(1, 0) - \lambda &= 0, \\ n(1, 1)/\hat{P}(1, 1) - \lambda &= 0, \\ \hat{P}(0, 0) + \hat{P}(0, 1) + \hat{P}(1, 0) + \hat{P}(1, 1) &= 1. \end{aligned}$$

For simplicity the algebraic details of solving (17) will not be given here. The mle are as follows:

$$(18) \quad \begin{aligned} \hat{P}(1, 1) &= n(1, 1)/n, \\ \hat{P}(1, 0) &= n_1(1, 0)[n - n(1, 1)]/n[n_1 - n_1(1, 1)], \\ \hat{P}(0, 1) &= n_2(0, 1)[n - n(1, 1)]/n[n_2 - n_2(1, 1)], \\ \hat{P}(0, 0) &= \{[n - n(1, 1)]/n\}\{n_2(-, 0)/[n_2 - n_2(1, 1)] \\ & \qquad - n_1(1, 0)/[n_1 - n_1(1, 1)]\}. \end{aligned}$$

The variance and covariances can be calculated by using standard techniques. The method used here was to find the matrix of the expectations of the second derivatives of the logarithm of the likelihood function with respect to the parameters and then finding the inverse of the matrix with changed signs. (For notational simplicity, P_{ij} shall be used for $P(i, j)$.) They are as follows:

$$\begin{aligned} V(\hat{P}(0, 1)) &= P_{01}(1 - P_{11} - P_{01})/n_2(1 - P_{11}) \\ &\quad + P_{01}^2 P_{11}/n(1 - P_{11}), \\ V(\hat{P}(0, 0)) &= [nn_1 P_{01}(1 - P_{11} - P_{01}) + n_1 n_2 P_{01}^2 P_{11} \\ &\quad + 2n_1 n_2 P_{10} P_{01} P_{11} - 2n_1 n_2 (1 - P_{11}) P_{01} P_{11} \\ &\quad + nn_2 P_{10}(1 - P_{11} - P_{10}) + n_1 n_2 P_{10}^2 P_{11} \\ &\quad + n_1 n_2 P_{11}(1 - P_{11})^2 \\ &\quad - 2n_1 n_2 P_{10} P_{11}(1 - P_{11})]/nn_1 n_2(1 - P_{11}), \\ \text{Cov}\,(\hat{P}(0, 1), \hat{P}(1, 0)) &= P_{10} P_{01} P_{11}/n(1 - P_{11}). \end{aligned} \tag{19}$$

The other second order moments are standard or can be obtained by interchange of subscripts.

REMARK 1. Due to sampling fluctuation sometimes the estimates of $P(0, 0)$ by any of the methods discussed, may come out to be negative. In such cases, zero may be taken as the estimate.

REMARK 2. In the special case when all the $P(i, j)$'s are equal to $\frac{1}{4}$ and $n_1 = n_2$, then $V(\hat{p}_{G1}(0, 0)) = V(\hat{p}_{G2}(0, 0)) = V(\hat{P}(0, 0)) = .34375/n_1$.

4. Trivariate sampling scheme. In this section the situation where each unit has three (A, B, C) SOD binary characters will be discussed. The problem is to estimate the proportion of (0, 0, 0) in the population on the basis of a sample of size n. The total sample size is divided into three parts n_1, n_2, n_3 such that $n = n_1 + n_2 + n_3$. Then three samples of sizes n_1, n_2, n_3 are drawn without replacement from the population. The three characters, A, B, C, are then observed in a cyclic pattern as indicated in Figure 3. No attempt will be made to explain the figure in words because it is self-evident from the explanations given for Figure 1.

Let $n(1, 1, 1)$ be the frequency of (1, 1, 1) in the combined sample n. $n_l(i, j, k)$ be the frequency of (i, j, k) in the sample n_l where i, j, k represents symbols 0, 1 or $-$ and $l = 1, 2, 3$.

Any triplet (i, j, k) is observable provided it does not have two or more 0's. The observed $n_l(i, j, k)$'s correspond to the triplets shown in Figure 3.

$$\begin{aligned} p(1, 1, 1) &= n(1, 1, 1)/n, \\ p(i, j, k) &= n_l(i, j, k)/n_l \text{ where } n_l = n_1 \text{ or } n_2 \text{ or } n_3 \text{ depending on the} \\ &\qquad \text{sample in which } (i, j, k) \text{ is observed,} \\ P(i, j, k) &= \text{proportion of } (i, j, k) \text{ in the population.} \end{aligned} \tag{20}$$

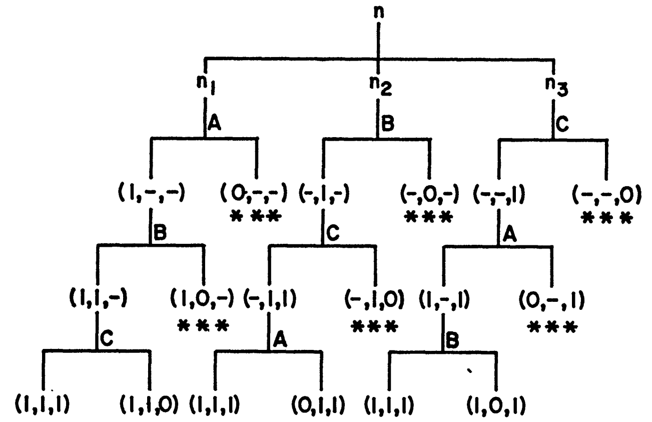


FIG. 3

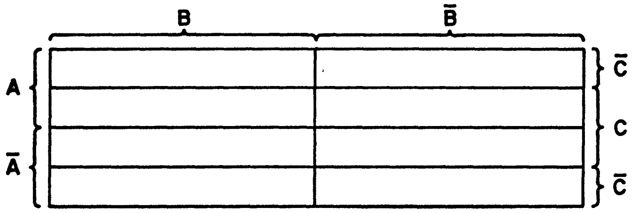


FIG. 4

The pictorial representation of Boolean functions of three or more variables can be given neatly by Veitch diagrams. For three variables it is as in Figure 4.

Using standard notations of Boolean algebra it follows from the above Vitch diagram, $A\bar{B} + B\bar{C} + \bar{A}C + ABC + \bar{A}\bar{B}\bar{C} = \Omega$. The left hand side represents a simple Boolean function, hence

$$P(1, 0, -) + P(-, 1, 0) + P(0, -, 1) + P(1, 1, 1) + P(0, 0, 0) = 1.$$

Hence

$$\hat{p}(0, 0, 0) = 1 - p(1, 0, -) - p(-, 1, 0) - p(0, -, 1) - p(1, 1, 1). \tag{21}$$

Usually, it would be possible to build more estimates of $P(0, 0, 0)$ under this sampling scheme but because of the SOD nature of the characters it can be shown by a simple enumeration method that $p(0, 0, 0)$ is the only estimate of $P(0, 0, 0)$ possible by using Boolean algebra.

The variance of $\hat{p}(0, 0, 0)$ is given by

$$\begin{aligned} V(\hat{p}(0, 0, 0)) = {} & P(1, 1, 1)Q(1, 1, 1)/n + P(1, 0, -)Q(1, 0, -)/n_1 \\ & + P(-, 1, 0)Q(-, 1, 0)/n_2 + P(0, -, 1)Q(0, -, 1)/n_3 \\ & - 2P(1, 1, 1)[P(1, 0, -) + P(-, 1, 0) + P(0, -, 1)/n]. \end{aligned} \tag{22}$$

In case of finite population, a fpc term will have to be attached to each term on the right hand side, but that situation will not be discussed here. The proportion of various other non-observable triplets can be estimated from this scheme.

5. Optimum design for trivariate scheme. The free parameters in the sampling design are n_1, n_2, and n_3 satisfying the restriction $n_1 + n_2 + n_3 = n$. The optimum choice of the triplet (n_1, n_2, n_3) can be defined as the one which minimizes the variance of the estimate subject to the restriction about the sum. Neyman's well-known technique can be applied to minimize (22) subject to $n_1 + n_2 + n_3 = n$. The solutions are given by

$$\begin{aligned} n_1 &= [\{P(1, 0, -)Q(1, 0, -)\}^{\frac{1}{2}}/\Sigma]n, \\ n_2 &= [\{P(-, 1, 0)Q(-, 1, 0)\}^{\frac{1}{2}}/\Sigma]n, \\ n_3 &= [\{P(0, -, 1)Q(0, -, 1)\}^{\frac{1}{2}}/\Sigma]n, \end{aligned}$$

where

$$\Sigma = [P(1, 0, -)Q(1, 0, -)]^{\frac{1}{2}} + [P(-, 1, 0)Q(-, 1, 0)]^{\frac{1}{2}} + [P(0, -, 1)Q(0, -, 1)]^{\frac{1}{2}}.$$

A model sampling for bivariate was simulated on the IBM 7094 computer and the results are given below:

The variance of each of the estimates had been calculated for 38 combinations of $P(0,0)$, $P(1,0)$, $P(0,1)$, $P(1,1)$ with $n_1 = 70$, $n_2 = 30$. The combinations vary from (.10, .10, .20, .60) to (.70, .10, .10, .10). The variances vary from .0047 (Estimate $G1$ with .10, .10, .20, .60 probabilities and estimate ML with .30, .10, .10, .50 probabilities) to .0183 (estimate ML with .20, .10, .60, .10 probabilities).

6. Acknowledgment. The author wishes to thank Professor Tore Dalenius of the University of Stockholm for suggesting the problem. The author is grateful to Dr. Rolf Bargmann for some valuable comments and suggestions.

REFERENCES

[1] Dalenius, T. (1959). Ett mätfelsproblem. *Nordisk Tidskrift Industriel Statistik* **4** 65–74.
[2] Phister, M., Jr. (1961). *Logical Design of Digital Computers.* Wiley, New York.