Commun. Fac. Sci. Univ. Ank. Ser. A_1
V. 36, Number 2, pp. 69–81 (1987)

THREE - PHASE SAMPLING FOR NUMERICAL AND FOR BINARY DATA

İSMAİL ERDEM

Department of Statistics, METU, Ankara - Turkey.

ABSTRACT

A sample drawn in several phases refers to selection of a large-sized preliminary sample in which some quick and inexpensive method of measurement is applied. A subsample is drawn and a somewhat more elaborate method of measurement is used and then successively smaller subsamples are drawn and successively more expensive, elaborate and accurate methods of measurement are applied. This paper furnishes estimator and variance expressions for the three-phase design. The examples, arising from remote sensing applications, illustrate estimation of the population mean of a numerical characteristic and the population of a binary variable of interest. Formulas for optimum sampling rates are shown to be of the same form for the the two types of variables.

Key Words: Measurement intensity, misclassification, optimal sampling rates

INTRODUCTION

The multi–phase feature of sample design refers to drawing a large preliminary sample, with a subsample and possibly with sub–subsamples. Data are collected on the variable of interest only from the smallest sub–subsample. The more inclusive samples have data on surrogate or auxiliary variables that are closely related to the variable of interest but are less expensive to collect. In order for three–phase sampling to be worthwhile there must be two surrogate or auxiliary variables to the variable of interest; while two–phase sampling requires only one auxliary variable. The three variables need to be well ordered by cost of measurement and correspondingly by their reliabilities in measuring the vaiable of interest.

Because sample design terminology is not always uniform, let's review some designs that are conceptually akin to the multi–phase

ISSN 0251–0871, A. Üniversitesi Basımevi.

one and perhaps make clearer what the multi–phase feature offers. Multi–phase designs can easily be distinguished from multi–stage designs. With multi–stage sampling there are a number of large first stage units and each is broken into several second stage units, and so forth. Multi–stage sampling involves selecting a few first–stage units and in these a few second stage units are selected, and so forth. Two–phase sampling designs should also be distinguished from two–step designs. Here the first step is a small preliminary sample and the size of the second step depends on the variability found in the first–step as well as on some pre–set amount of precision required. Both multi–stage and two–step sampling theory may involve only the one variable of interest while the multi–phase feature always has the other auxiliary variables.

There are two variants of multi–phase sampling that are helpful to keep in view even though our case differs from them. The optimum preliminary sample size may be calculated to be larger than N, the population size. Using auxiliary information available on all members of the population, such as with a regression estimator, is thus seen to be a variant of multi–phase sampling. Another variant is a design whereby the preliminary sample is disjoint from the smaller sample and there are thus no paired observations. There are cases when the calibration factor is well known and there is thus no great help in this regard from having paired data. It may also be physically impossible to have paired data or it may be very costly. Having spent some time on what we won't be discussing we'd better get on with our topic.

The archtypical case of multi–phase sampling arises when measurement of some variable occurs naturally with differing levels of intensity. It may be possible to obtain vague impressions by visual scanning or by long–distance reconnaissance; more precise information can be obtained by use of more elaborate equipment and closer approach. Need for multi–phase sampling arises in various fields. In agriculture an estimate of land under irrigation in California involved satellite photo data, then aerial photos and finally ground determination (Colwell 1977). In atmospheric pollution studies of CO levels, the investigators used space shuttle radiometry, aircraft radiometry and, finally, aircraft-collected air samples (Reichle et al. 1982). A study of diseases of tobacco included eye–scan in the field, close–up visual inspection of plants, and laboratory study of leaf samples (Main and Proctor 1980).

The corbon monoxide variable was essentially numerical (parts per billion by volume) while the irrigation example was also numerical in giving percentage of land irrigated on land areas. The tobacco diseases data involved a dichotomy when the individual plant was used as the unit of analysis. That is, the plant is disseased or not. Our derivation of variance formulas, estimation formulas and optimization formulas has been done separately for these two cases the one where one might suppose some similarity in distribution to the multivariate normal and the other where the variable is a dichotomy. We will first review the results that appeared some years ago for the multivariate normal case as background to some new results on the dichotomy case.

THREE–PHASE SAMPLING TO ESTIMATE $\bar{Y}$

As a fist approach we will suppose simple random sampling is done at all phases. Also we suppose the population is so large as to be effectively infinite. Denote the first phase observations as w_1, w_2, ..., w_{n_1}, the second phase ones as x_1, x_2, ..., x_{n_2} and the third phase ones as y_1, y_2, ..., y_{n_3}. The overlapping observations are taken to be the first ones in each set. Since the first observations in each set are from the same unit we can describe the multivariate normal distri-distributions for w_1, x_1 and y_1 as:

$$w_1 \sim N(\mu_w, \sigma^2_w),\ x_1 \sim N(\nu + \beta_{xw}\ w_1,\ \sigma^2_{x\cdot w})$$

and

$$y_1 \sim N(\lambda + \beta_{yx\cdot w}\cdot x_1 + \beta_{yw\cdot x} w_1,\ \sigma^2_{y\cdot wx}).$$

The population mean, defined as $\bar{Y} = \Sigma Y_i / N$, is, in accord with our supposition of large N, very nearly equal to $E(y) = \mu_y$, and so we can take μ_y to be the parameter of interest.

Nine parameters have been introduced. They are: μ_w, σ^2_w, ν, β_{xw}, $\sigma^2_{x\cdot w}$, λ, $\beta_{yx\cdot w}$ $\beta_{yw\cdot x}$, and $\sigma^2_{x\cdot xw}$. They can be estimated as follows: (1) w_1, w_2, ..., w_{n_1} are used to estimate μ_w and σ^2_w; (2) one then takes the first n_2 w's to be fixed and, along with x_1, x_2, ..., x_{n_2} estimates ν, β_{xw} and $\sigma^2_{x\cdot w}$; and (3) the first n_1 w's and x's are treated as fixed while the y's are used to estimata λ, $\beta_{yx.w}$, $\beta_{yw.x}$ and $\sigma^2_{y\cdot wx}$.

The estimate of μ_w is denoted $\bar{w}''$ and is based on all n_1 observations on the first phase. In order to estimate $E(x) = \mu_x$, the population mean of the x's, one notes that

$$E(x_i) = E(\nu + \beta_{x_w} w_i) \tag{2.1a}$$

$$= \nu + \beta_{x_w}\mu_w \tag{2.1b}$$

$$\triangleq \bar{x}' - b_{x_w}\bar{w}' + b_{x_w}\bar{w}'' \tag{2.1a}$$

where the single prime denotes use of the n_2 second phase observations. The last expression (2.1a) can be written

$$\bar{x}_{lr,2} = \bar{x}' + b_{x_w}(\bar{w}'' - \bar{w}') \tag{2.2}$$

which is the two–phase regression estimator (Cochran 1977, p. 339) of μ_x. The final step is to express μ_y as

$$E(y_i) = E(\lambda + \beta_{yx\cdot w} x_i + \beta_{yw\cdot x} w_i) \tag{2.3a}$$

$$= \lambda + \beta_{yx\cdot w}\mu_x + \beta_{yw\cdot x}\mu_w \tag{2.3b}$$

$$\triangleq \bar{y} - b_{yx\cdot w}\bar{x} - b_{yw\cdot x}\bar{w} + b_{yx\cdot w}\bar{x}_{lr,2} + b_{yw\cdot x}\bar{w}'' \tag{2.3c}$$

$$= \bar{y} + b_{yx\cdot w}(\bar{x}' - \bar{x}) + b_{yx\cdot w}b_{x_w}(\bar{w}'' - \bar{w}') + b_{yw\cdot x}(\bar{w}'' - \bar{w}) \tag{2.3d}$$

$$= \bar{y}_{lr,3} \text{ say,} \tag{2.3e}$$

the three–phase estimator of $\hat{\mu}_y$. It is oftentimes reasonable in applications to use the ordinary least squares (OLS) estimators of the b's. Whenever the sample sizes are small or when other information is clearly superior, there may be quantities better than these OLS estimators to insert for the b's. Under repeated sampling all the differences of means in $y_{lr,3}$ (see (2.3d)) average to zero and the estimator $y_{lr,3}$ is seen to be unbiased. It's variance can be easily obtained from the three–by–three covariance matrix of the variates y, x, and w.

The entries of this matrix are σ^2_y, σ^2_x and σ^2_w on the diagonal with σ_{yx}, σ_{xw} and σ_{xw} in the appropriate off–diagonal positions. The total variance of y is σ^2_y. Using knowledge of w to predict y leaves a residual variance of

$$\sigma^2_{y\cdot w} = \sigma^2_y - \sigma^2_{yw}/\sigma^2_w \triangleq s^2_y = (1 - r^2_{yw}) \tag{2.4}$$

The reduction in variance here will be denoted $V_1 = \sigma^2_{yw}/\sigma^2_w$. The additional reduction in the residual variance due to adding knowledge of x is found to be:

$$V_2 = (\sigma_{xy} - \sigma_{wy}\sigma_{xw}/\sigma^2_w)^2 / (\sigma_x^2 - \sigma^2_{xw}/\sigma^2_w).$$

$$\triangleq s^2_y (1 - r_{xy} - r_{yw} - r_{xw} + 2r_{xy}r_{yw}r_{xw}) / (1 - r^2_{xw}) \tag{2.5}$$

To make the notation most convenient let $V_3 = \sigma^2_y - V_2 - V_1$. The variance of $y_{lr,3}$ then becomes:

$$V(\bar{y}_{lr,3}) = V_1/n_1 + V_2/n_2 + V_3/n_3 . \quad (2.6)$$

An interesting example of three-phase sampling was furnished by Colwell (1977). The objective was to estimate the percentage of irrigated land among farm land in a portion of California. The data are shown in the Table. Although there weren $n_1 = 1292$ pixels in the first phase sample we only have individual data for the $n_2 = 88$ cases where data are available from both satellite and aerial photos. The table also shows the $n_2 = 16$ ground-based observations. The sample was stratified by county and thus we should subtract county means from all observations. However, some counties have only one ground-based observation. Rather than loose this third phase information we defined two combined couties (conunties 2,5 and 6 versus the others) and subtracted these means before computing coveriances.

In computing the means $\bar{y}$, $\bar{x}$, $\bar{x}'$, $\bar{w}$ and $\bar{w}'$ we used stratum weights of .3125, .0660, .1351, .0369, .0229, .0919, .1607, .0942 and .0798 for the nine counties. The value for $\bar{w}''$ was found as $\bar{w}'' = 80.84$ (from Colwell 1977, p. 26). One may verify from the data in the table that $\bar{y} = 75.09$, $\bar{x} = 74.34$, $\bar{x}' = 71.70$, $\bar{w} = 71.94$ and $\bar{w}' = 71.19$. The estimator $\bar{y}_{lr\cdot 3}$ awas thus computed as:

$$\begin{aligned}\bar{y}_{lr,3} &= 75.09 + .4332\ (71.70 - 74.34) \\ &\quad + .4332 \times .8586\ (80.84 - 71.19) + .5992\ (80.84 - 71.94) \\ &= 82.87 .\end{aligned}$$

The values for $b_{yx\cdot w} = .4332$ and $b_{yw\cdot x} = .5992$ were found by ordinary least squares regression calculations on residuals after subtracting combined county means while $b_{xw} = .8586$ was based on residuals from separete country means. The correlations among residuals from combined county means were:

$r_{xw} = .910$, $r_{yw} = .815$ and $r_{yx} = .823$, with standard deviations $s_w = 21.67$, $s_x = 21.75$ and $s_y = 16.34$. Other estimates of these quantities are available but will be in the same neighborhood. This shows that $V_1 \cong 177.34$, $V_2 \simeq 25.50$ and $V_3 \cong 64.15$, so that an estimate of $V(\bar{y}_{lr,3})$ is

$$\begin{aligned}V(\bar{y}_{lr,3}) &= 177.34/1292 + 25.50/88 + 64.15/16 \\ &= 4.44 . \quad (2.7)\end{aligned}$$

The estimate and its standard error, both in percent, are thus 82.9 $\pm$ 2.1. The sampling coefficient of variation is about 3 % which is fairly good.

EXTENDED CONSIDERATION WHEN USING $\bar{y}_{lr,3}$

The cost such a survey can oftentimes be reasonably represented by a linear cost function namely

$$C_T = C_1 n_1 + C_2 n_2 + C_3 n_3 . \tag{3.1}$$

The per unit costs increase as we go from satellite photos to ground-based observations. If both the variance factors V_1, V_2 and V_3 and the cost coefficients C_1, C_2 and C_3 are known, then optimum sampling rates can be calculated as:

$$\text{opt } (n_1 / n_2) = \sqrt{V_1 C_2 / V_2 C_1}, \text{ and}$$

$$\text{opt } (n_2 / n_3) = \sqrt{V_2 C_3 / V_3 C_2}. \tag{3.2}$$

If both C_3 / C_2 and C_2 / C_1 are judged to be about 10, then the optimum sampling rates are found as 13 and 1. The actual rates were 14 and 5. If this judgement is correct then the aircraft was somewhat overused and could perhaps be dispensed with. This might become more apparent if the various fixed costs of the three types of information had been included in our cost function.

The underlying distributional assumptions have included independence as well as normality and in the present application some attention needs to be paid to the possible violation of these assumptions. There would likely be some negative skewness in the distribution of $\bar{y}_{lr,3}$ because the values of $\bar{y}$ as seen in the Table are percentages near 80 The skewness coefficient for a binomial binary variate is $G_1 = (Q-P) / \sqrt{PQ}$ where P is the proportion of ones an $Q = 1-P$. For $P = .8$, one finds $G_1 = -1.5$ which is fairly skewed. A sample size of 30 should yield a sample mean whose confidence intervals at a nominal 95 % coefficient have coverage probabilities between 94 % and 96 % (see Barrett and Goldsmith 1976). The sample size in this example is $n_3 = 16$ and thus some caution is needed. If the 82.87 estimate is low it may be fairly far below but if it is high it is not very far above. We're not just sure how knowledge of skewness should affect the use of an estimate but it should probably be announced.

The other aspect is idependence of the observations from one point to another. The issue here involves adjacency correlations on the map and sample selection methods. Even if there is no adjacency correlation the finiteness of the population, along with sampling without repla-

cement, may induce negative correlations among the observations. Sampling in the present case was simple random and thus some negative correlation will appear. This, however, will only affect the ground–based observations which are supposed to be measured without error. The sampling fraction at this last phase is about 16–from–1292 and is thus too small to affect the sampling variance.

The aircraft and satellite quantities are produced by a stochastic process (a, so called, interpretation operation) which refers to an effectively infinite population. For these processes the possibility of adjacency correlation is quite a likely one. Given that one pattern of sample selections is more clustered than another, one could expect larger variance in one case than the other. However, under simple random sampling these effects would average to zero and thus we ignore them.

THREE–PHASE SAMPLING FOR BINARY DATA

We turn now to the case of a binary variable of interest. This may represent simply a change from area sampling to point sampling. That is, we could estimate the percentage of irrigated land from a sample of points as well as from the pixel–areas that were used above. Measurement is, of course, a somewhat different operation when dealing with points rather than areas. Working with clusters of adjacent points would be more akin to the use of areas. We will not at present, attempt any comparison of points to areas as sampling units. We mention the two approaches to the same question by way of transition. It is our purpose now to furnish variance expressions and optimum allocation formulas for three–phase sampling with a binary variable of interest.

The binary two–phase case has been worked out by Aaron Tenesnbein (1970). The three–phase case that we will be treating here present no great novelties but one can see how complicated it would be to formulate a genaral treatment for te multi–phase case. It seems unlikely that a sample with more than three phases would be a practical survey design, although successive occasions data may lead one to such a possibility.

For theree phases of binary data the n_3 triply–measured observations have the true measurements on the variable of interest plus a less fallible classifier (LFC) and a more fallible classifier (MFC). The eight possible frequencies here will be denoted f_{iik} where $i = 0$ or 1

represents true measurements, $j = 0$ or 1 is the less fallible classifier and $k = 0$ or 1 tells about the more fallible classifier. There are $n_2 - n_3$ observations for wnich we have both LFC and MFC and these four frequencies will be denoted f_{jk}. Finally, the $n_1 - n_2$ observations with only SFC give frequencies f_k, where $f_0 + f_1 = n_1 - n_2$. Most of the following results appear in Erdem (1984).

If one views the simple random sampling method as being applied to a finite population of N points of which PN are found to be ones when measured by the true measurement method, then the frequencies $f_{0..} = \sum_{j=0}^{1} \sum_{k=0}^{1} f_{0jk}$ and $f_{1..} = n_3 - f_{0..}$ are seen to be distributed in accord with the hypergeometric distribution. In most of the applications this hypergeometric distribution will be indistinguishable from the binomial distribution with sample sizen n_3 and parameter P. We will thus suppose that all of the frequencies f_{ijk} are multinominally distributed. This also requires one to verify that the methods of making the fallible classifications do not induce too many dependencies in measurement errors that could upset these assumptions.

The underlying probabilities can be denoted as P_{ijk} for the f_{ijk}, as P_{jk} for the f_{jk} and as P_1 and $P_0 = 1-P_1$ for the f_k. Whith apologies for adding to our notation, a more convenient parameterization is in terms of the following Θ's plus P_1:

$$\Theta_1 = P_{000}/P_{00}\ \Theta_2, = P_{010}/P_{10}$$
$$\Theta_3 = P_{001}/P_{01},\ \Theta_4 = P_{011}/P_{11}$$
$$\Theta_5 = P_{00}/P_0, \quad \Theta_6 = P_{01}/P_1. \tag{4.1}$$

In this way Θ_1, Θ_2, Θ_3 and Θ_4 can be estimated directly from the f_{ijk}, the next two, Θ_5 and Θ_6, from the f_{jk}, while P_1 can be estimated from f_1. In fact the likelihood function can be found to factor and this makes these estimates maximum likelihood (ML) estimators. There is a similarity in approach here to the above multivariate normal case.

The parameter of interest, P, can now be expressed in terms of the Θ's and P_1 as

$$P = (1-P_1)\ \Theta_5\ (1-\Theta_1) + (1-P_1)\ (1-\Theta_2)\ (1-\Theta_5) + P_1\ (1-\Theta_3)\ \Theta_6 + P_1\ (1-\Theta_4)\ (1-\theta_6). \tag{4.2}$$

By substituting the ML estimates of the Θ's and of P_1 into this expression one obtains the ML estimate of P, under the proviso of $f_{\cdot jk} > 0$ for all j and k, as:

$$\hat{P}_{3p} = \sum_{j=0}^{1} \sum_{k=0}^{1} \frac{f_{\cdot\cdot k} + f_{\cdot k} + f_{k}}{n_1} \cdot \frac{f_{\cdot jk} + f_{jk}}{f_{\cdot\cdot k} + f_{\cdot k}} \cdot \frac{f_{ijk}}{f_{\cdot jk}} \qquad (4.3)$$

If any of the $f_{\cdot jk}$ should be zero then we suggest that the contribution from that combination of j and k to the sum be set to zero.

When one visulizes the data being produced by the initial random selection and subsequent independent stochastic classifications, a further reparametirization seems logical. The new parameters along with P are α_0, α_1, β_0 and β_1 where:

$$\begin{aligned} \alpha_0 &= \Pr(\text{LFC Misclassifies} / i = 0) \\ \alpha_1 &= \Pr(\text{LFC Misclassifies} / i = 1) \\ \beta_0 &= \Pr(\text{MFC Misclassifies} / i = 0) \\ \beta_1 &= \Pr(\text{MFC Misclassifies} / i = 1). \end{aligned} \qquad (4.4)$$

The links back to the underlying probabilities are:

$$\begin{aligned} P_{000} &= Q(1-\alpha_1)(1-\beta_1) & P_{100} &= P\,\alpha_0\beta_0 \\ P_{010} &= Q\,\alpha_1(1-\beta_1) & P_{110} &= P(1-\alpha_0)\beta_0 \\ P_{001} &= Q(1-\alpha_1)\beta_1 & P_{101} &= P\,\alpha_0(1-\beta_0) \\ P_{011} &= Q\,\alpha_1\beta_1 & P_{111} &= P(1-\alpha_0)(1-\beta_0) \end{aligned} \qquad (4.5)$$

There are only five parameters in this version as compared to the seven we used earlier and one should probably examine the data in deciding on the appropriateness of this formulation.

In actual applications one would study the measurement operation under better controlled conditions than during the survey itself. Thus we may well suppose this separate investigation has furnished estimates of α_0, α_1, β_0 and β_1. Our interest is in calculating the variance of $\hat{P}_{3p}$ as a function of these, presumed known, misclassification parameters.

By returning to the joint multinomial distribution of the frequencies, finding their covariances and then applying a Taylor series expansion to the expression for $\hat{P}_{3p}$ one sees that:

$$\text{Var}(\hat{P}_{3p}) = A_1/n_1 + A_2/n_2 + A_3/n_3 \qquad (4.6a)$$

$$= PQ\ (K_1/n_1 + K_2/n_2 + K_3/n_3) \qquad (4.6b)$$

where

$$A_1 = P_1(1-P_1)\ [\Theta_6(1-\Theta_3) + (1-\Theta_6)(1-\Theta_4) - \Theta_5(1-\Theta_1) - (1-\Theta_5)(1-\Theta_1)]^2,$$

$$A_2 = \Theta_6(1-\Theta_6)\,P_1(\Theta_4-\Theta_3)^2 + \Theta_5(1-\Theta_5)(1-P_1)(\Theta_2-\Theta_1)^2, \text{ and}$$

$$A_3 = \Theta_1(1-\Theta_1)\,\Theta_5(1-P_1) + \Theta_2(1-\Theta_2)(1-\Theta_5)(1-P_1) + \Theta_3(1-\Theta_3)\,\Theta_6P_1 + \Theta_4(1-\Theta_4)(1-\Theta_6)\,P_1,$$

while

$$K_1 = PQ\ (1-\beta_0-\beta_1)^2 / P_1(1-P_1),$$

$$K_2 = PQ\,(1-\alpha_0-\alpha_1)^2\ \left[\frac{\beta^2_0(1-\beta_1)^2}{(1-P_1)\,P_{10}\,P_{00}} + \frac{\beta^2_1(1-\beta_0)^2}{P_1P_{01}\,P_{11}}\right], \text{ and}$$

$$K_3 = \beta_0(1-\beta_1)\ \left[\frac{\alpha_0(1-\alpha_1)}{P_{00}} + \frac{\alpha_1(1-\alpha_0)}{P_{10}}\right] + \beta_1(1-\beta_0)\ x$$

$$\left[\frac{\alpha_0(1-\alpha_1)}{P_{10}} + \frac{\alpha_1(1-\alpha_0)}{P_{11}}\right] .$$

The variance expression for binary data thus breaks conveniently into the same three parts as the earlier one (2.4) for numerical data did.

This offers the same possibilities for gauging the relative effort one should put on each of the phases whenever one has some notion of the relative cost of the three classifiers. In sor far as the costs follow the same simple function as (3.1) the optimum sample size formulas of (3.2) will hold. Notice that usefulness of a measurement method is effectively summarizaed by the ratio K_i/C_i, where C_i is the cost, in time or other resources, per measurement for the more fallible classifier when $i = 1$, less fallible classifier when $i = 2$ and for the third phase or true classifier when $i = 3$.

NUMERICAL ILLUSTRATION

Lacking more realistic data, let's use some generated in the course of a Monte Carlo simulation to investigate the properties of $\hat{P}_{3p}$ (4.3) and of an estimator of its variance based on (4.6a). The preliminary

sample was of size $n_1 = 500$, with $n_2 = 50$ and $n_3 = 25$. The data were found to be:

$$f_{111} = 7,\ f_{110} = 2,\ f_{101} = 0,\ f_{100} = 0,$$
$$f_{011} = 0,\ f_{010} = 1,\ f_{001} = 3,\ f_{000} = 12,$$
$$f_{11} = 3,\ f_{10} = 2,\ f_{01} = 4,\ f_{00} = 16,$$
$$f_{0} = 313, \text{ and } f_{1} = 137.$$

The first step is to estimate Θ_1 through Θ_7 as:

$$\hat{\theta}_1 = f_{000}/f_{\cdot 00},\ \hat{\theta}_2 = f_{010}/f_{\cdot 10}$$
$$\hat{\theta}_3 = f_{001}/f_{\cdot 01},\ \hat{\theta}_4 = f_{011}/f_{\cdot 11}$$
$$\hat{\theta}_5 = (f_{\cdot 00} + f_{00})/(f_{\cdot\cdot 0} + f_{\cdot 0}), \text{ and}$$
$$\hat{\theta}_6 = (f_{\cdot 01} + f_{01})/(f_{\cdot\cdot 1} + f_{\cdot 1}), \text{ with}$$
$$\hat{\theta}_7 = \hat{P}_1 = (f_{\cdot\cdot 1} + f_{\cdot 1} + f_{1})/(f... + f.. + f.).$$

These estimates, given here to just two decimals, are: 1.00, .33, 1.00, .00, .85, and .41. Entering them into expression (4.2) produces $\hat{P}_{3p} = .251075$. Next by substituting them into (4.6a) we find $V(\hat{P}_{3p}) = .003316$. The conclusion is that there is a population proportion of $.25 \pm .06$.

If there had been separate knowledge of α_1, α_0, β_1 and β_0, this could be incorporated into $V(\hat{P}_{3p})$. Such information could stabilize the variance estimate. However, we believe it could be a little risky using it in estimating p because it would create bias if it was mistaken knowledge. When $\alpha_1 = \alpha_0 = .1$ and $\beta_2 = \beta_0 = .2$, as somewhat reasonable values, are used in expression (4. 6b) then $V(\hat{P}_{3p}) = .003610$.

SUMMARY

The basic expressions for any sampling method are the estimator, its bias and variance, a variance estimator and optimum sampling rates. For three–phase sampling and for numerical data these expressions are (2.3d), (2.6), (2.7) and (3.2) respectively. Actually formula (2.5), the estimated variance, is only a numerical example but the text tells how to use sample variances and correlations to obtain this variance estimate. For binary data the basic formulas are (4.3 and (4.6a). Either formula (4.6a) or (4.6b) can be used for variance estimation as illustrated in the example just above. The optimization formula is again (3.2) for binary data with the K's in place of V's.

Table. Percent of Land Irrigated for Sampled Areas as Measures from Landsat (Phase I), from Aircraft (Phase II) and from Ground (Phase III) in Nine Counties of California

County 1	I, w_i	II, x_i	III, y_i
	91	91	94
	84	83	83
	90	92	93
	97	98	
	100	100	
	88	100	
	72	71	
	68	70	
	100	100	
	65	70	
	90	100	
	81	79	
	81	96	
	95	94	
	94	100	
	100	99	
	85	88	
	92	88	
	66	99	
	73	78	
	67	54	
	85	66	
	93	76	
	96	95	
	95	94	
County 2			
	48	69	56
	90	92	
	51	51	
	79	67	
	78	90	
	35	43	
	76	76	
County 3			
	91	87	88
	68	89	86
	77	75	
	82	73	
	44	41	
	86	79	
	87	86	
	68	79	
	63	76	
County 4			
	81	88	86
	82	77	82
	81	79	
	67	73	
	62	64	
	10	8	
	82	83	
County 5			
	27	30	27
	6	5	5
County 6			
	51	51	54
	34	50	
	2	6	
	45	42	
	75	77	
	37	36	
	58	50	

ACKNOWLEDGEMENT

A partion of this work is based on the author's Ph.D. thesis submitted to North Carolina State Universty, Raleigh, N, C., U.S.A.

REFERENCES

ANDERSON, T.W. 1957. "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing", *Journal of the American Statistical Association*, 52, 200–203.

BARRETT, J.P. and GOLDSMITH, L. 1976. "When is n Sufficiently Large?", *The American Statistician*, 30, 67–70.

COLWELL, R.N. 1977. "An Inventory of Irrigated Lands for Selected Counties Within the State of California Based on Landsat and Supporting Aircraft Data," Space Sciences Laboratory Series 18, Issue 50, Berkeley, CA, University of California.

ERDEM, İSMAİL 1984. "Three-Phase Sampling for Misclassified Binary Data", Unpublished Ph. D. Dissertation, Raleigh, NC, North Carolina State University.

MAIN, C. E., and PROCTOR, C.H. 1980. "Developing Optimal Strategies for Disease-Loss Sample Survey", in "Crop Loss Assessment", Misc. Publication 7, St. Paul, MN, Agricultural Experiment Station, University of Minnesota, pp. 118–123.

REICHLE, H.G., BECK, S.M., HAYNES, R.E., HESKETH, W.D., HYPES, W.D., ORR, H. D. III, SHERRİLL, R.T., WALLIS, H.A., CASAS, J.C., SAYLOR, M.S. and GORMSEN, B.B. (1982. "Carbon Monoxide Measurements in the Troposphere", *Science*, 218, 1024–1026.

TENENBEIN, A. 1970. "A Double Sampling Scheme for Estimating from Misclassified Binomial Data", *Journal of the American Statistical Association*, 65, 1350–1361.