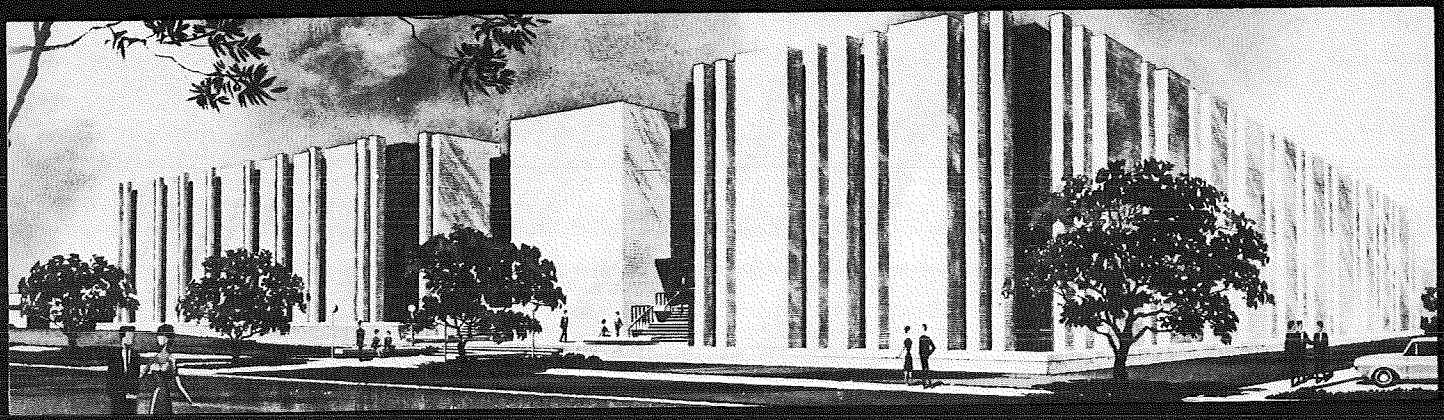
TRAINING & RESEARCH

TRAINING & RESEARCH

TRAINING & RESEARCH

TRAINING & RESEARCH

GRADUATE INSTITUTE OF STATISTICS



TEXAS A&M UNIVERSITY · COLLEGE STATION

OPTIMUM INCOMPLETE MULTINORMAL SAMPLES

by

R. R. Hocking and W. B. Smith

Institute of Statistics
Texas A&M University

Technical Report #1
National Aeronautics and Space Administration
Research Grant NGR 44-001-095

March, 1970

OPTIMUM INCOMPLETE MULTINORMAL SAMPLES

by

R. R. Hocking and W. B. Smith

Texas A&M University

1. Introduction

When sampling from a multivariate normal population, the researcher is frequently forced to cope with incomplete data. The data may be incomplete because not all elements of the multivariate observation vectors are recorded on every occasion (the 'missing data' problem) or because, in some cases, certain linear combinations of the elements of the observation vectors are recorded rather than the individual elements. Such data may arise by chance due to the peculiaritie of the data collecting procedure or there may be economical or physical reasons for collecting the data in incomplete form. For example, it may be costly to make certain measurements on the experimental unit so these measurements are not taken every time whereas the less costly measurements are made.

In a recent series of papers, [1], [2], [3] a maximum likelihood solution was described for the problem of estimating the parameters in the multivariate normal distribution with incomplete data. In addition to developing the estimates, expressions were developed for the asymptotic variance of these estimates. These expressions make it possible to assess the effect of the incomplete data on the precision of the estimates.

In this paper we consider the problem of designing the data collection procedure to intentionally yield incomplete data but at the same time give desired precision while satisfying certain other requirements. Specifically, we shall consider the problem of minimizing the cost of gathering the data subject to the

requirement that the parameters or functions of the parameters are estimated with desired precision. We shall, for simplicity, restrict our discussion to a special, but useful, class of incomplete data problems but this will serve to illustrate how one might procede in the more general situations described in the papers mentioned above.

In Section 2 we give a brief summary of the notation used in [1] and then develop specific large sample variances formulas. The minimum cost sample allocation problem and its solution are described in Section 3. Some extensions are suggested in Section 4.

2. Asymptotic Variance Expressions.

The special class of incomplete data problems to be considered in this development is that in which n_1 observations are taken from the p-variate normal distribution, $N(\mu_1, \Sigma_1)$, and n_2 observations are taken on the q-variate normal distribution, $N(\mu_2, \Sigma_2)$. It is assumed that $q < p$ and further, that Σ_2 is a principal minor of Σ_1 and that the vector μ_2 consists of the corresponding q-components of μ_1. This is just the 'missing data' situation in which all p measurements are made on n_1 occasions but only q of the measurements are made on n_2 occasions.

The elements of the matrices Σ_1 and Σ_2 may also be written as vectors σ_1 and σ_2 with dimension $p(p+1)/2$ and $q(q+1)/2$, respectively. The relations between μ_1, σ_1, Σ_1 and μ_2, σ_2, Σ_2 may be described by introducing matrices C and D such that

$$\mu_2 = D\,\mu_1$$

$$\Sigma_2 = D\,\Sigma_1\,D'$$

$$\sigma_2 = C\,\sigma_1 \ .$$

The form of D and C should be clear from the above. The large sample covariance matrices for maximum likelihood estimates of μ_1 and Σ_1 using only the n_1 incomplete observations are denoted by $V_{\mu 1}$ and $V_{\sigma 1}$ and given by

$$V_{\mu 1} = \frac{1}{n_1} \Sigma_1$$

$$V_{\sigma 1} = \frac{1}{n_1} U_1$$

Here the matrix U_1 is of dimension $p(p+1)/2$ and is a function of the elements of Σ_1. The rows and columns of U_1 are indicated by double subscripts, say (i, j) for $1 \leq i \leq j \leq p$. Specifically, the element in row (u, v), column (i, j) for $1 \leq u \leq v \leq p$, $1 \leq i \leq j \leq p$ is given by

$$\sigma_{iu}\sigma_{jv} + \sigma_{iv}\sigma_{ju} .$$

The analogous quantities for estimating μ_2 and Σ_2 using only the n_2 partial observations are

$$V_{\mu 2} = \frac{1}{n_2} \Sigma_2$$

$$V_{\sigma 2} = \frac{1}{n_2} U_2 .$$

The matrix U_2 is defined as is U_1 in terms of the elements of Σ_2 and may be obtained directly from U_1 by the relation

$$U_2 = C\, U_1\, C' .$$

Denoting by V_μ and V_σ the large sample covariance matrices for estimating μ_1 and Σ_1 from the combined sample, it is shown in [1] that

$$V_\mu = (I + BD)V_{\mu 1}$$

$$V_\sigma = (I + AC)V_{\sigma 1} .$$

Here we have introduced the matrices A and B defined as

$$B = -\frac{n_2}{N}\Sigma_1 D' \Sigma_2^{-1}$$

$$A = -\frac{n_2}{N}U_1 C' U_2^{-1}$$

with $N = n_1 + n_2$.

From these expressions we see that the gain in precision attainable, asymptotically, by using the n_2 partial observations is given by $BDV_{\mu 1}$ and $ACV_{\sigma 1}$ for μ_1 and Σ_1, respectively. We note that the gain in precision depends on the values of the unknown parameter Σ and is proportional to the fraction n_2/N. The simple way in which this partial data fraction, n_2/N, enters into these expressions suggests the possibility of intentionally collecting incomplete data in such a way as to obtain prescribed large sample variances on the estimates. Since the gain in precision offered by the incomplete data depends on the unknown parameters, it is clear that we can not hope to design an optimal data collecting procedure without some prior knowledge of these parameters. This is usually the case in sample allocation problems and appears in this case to be asking for prior knowledge of the elements of Σ. We shall now show precisely what prior knowledge is required and will note that it is a function of the elements of Σ whose magnitude may be known, at least approximately.

Let the elements of μ_1 be denoted by m_i, $i = 1, \ldots, p$ and the elements of μ_2 by m_i, $i \in I$ where I is the appropriate subset of the integers $1, \ldots, p$. Denote the elements of Σ_1 by σ_{ij}, $i, j = 1, \ldots, p$ and of Σ_2 by σ_{ij}, $i, j \in I$. The elements of Σ_1^{-1} are denoted by σ^{ij} and those of Σ_2^{-1} by ω^{ij}. Denote by δ_i the vector of regression coefficients if the i^{th} variable for $i \notin I$ is regressed on the variables indexed by I. Thus the u^{th} component of δ_i is given by

$$\delta_{iu} = \sum_{r \in I} \sigma_{ri}\omega^{ur} .$$

The corresponding multiple correlation coefficients are denoted by R_i^2 and given by

$$R_i^2 = \delta_i' \Sigma_2 \delta_i / \sigma_{ii} = \sum_{u \in I} \delta_{iu} \sigma_{iu} / \sigma_{ii} .$$

We shall now develop specific expressions for the variances of estimates of certain parameters and functions of parameters as obtained from the results described above. In particular, we want to look at the asymptotic variances for estimates of m_i and σ_{ii}, $i = 1, \ldots, p$ and β_i, $i = 2, \ldots, p$ where the β_i are regression coefficients obtained by regressing the first variable on variables 2 through p.

2.1. Variance of Estimates of m_i

If we denote by $V(m_i)$ the large sample variance of the maximum likelihood estimate of the parameter m_i, then from above we see that the gain in precision is given by

$$BDV_{\mu 1} = - \frac{n_2}{n_1 N} \Sigma_1 D' \Sigma_2^{-1} D \Sigma_1$$

and we obtain

$$V(m_i) = \begin{cases} \sigma_{ii}/N & i \in I \\ (1 - \frac{n_2}{N} R_i^2)\sigma_{ii}/n_1 & i \notin I . \end{cases}$$

2.2. Variance of Estimates of σ_{ij}

The gain in precision is determined by

$$ACV_{\sigma 1} = - \frac{n_2}{n_1 N} U_1 C' U_2^{-1} C U_1 .$$

In [1] it is shown that the element in row (i, j) column (u, v) of the matrix $U_1 C' U_2^{-1}$ is

$$\frac{1}{k_{uv}}\left(\sum_{r\epsilon I}\sigma_{ri}\omega^{ur}\sum_{t\epsilon I}\sigma_{tj}\omega^{vt}+\sum_{r\epsilon I}\sigma_{rj}\omega^{ur}\sum_{t\epsilon I}\sigma_{ti}\omega^{vt}\right)$$

for $1 \le i \le j \le p$ and $u \le v \epsilon I$ where $k_{uv} = 2$ for $u = v$ and $k_{uv} = 1$ for $u \neq v$.

Note that, for example

$$\sum_{r\epsilon I}\sigma_{ri}\omega^{ur} = \begin{cases} 0 & i \neq u,\ i \epsilon I \\ 1 & i = u \\ \delta_{iu} & i \notin I. \end{cases}$$

Multiplying on the right by $C\,U_1$ and simplifying shows that the matrix $U_1\,C'\,U_2^{-1}\,CU_1$ has in row (ij), column (ij) the quantity

$$\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2 \quad \text{for } ij \epsilon I$$

and

$$\sum_{u\epsilon I}\delta_{iu}\sigma_{iu}\sum_{v\epsilon I}\delta_{jv}\sigma_{jv} + \sum_{u\epsilon I}\delta_{iu}\sigma_{ju}\sum_{v\epsilon I}\delta_{jv}\sigma_{iv} \quad \text{for } ij \notin I.$$

In particular, for $i = j$ we have

$$2\,\sigma_{ii}^2 \qquad \text{for } i \epsilon I$$

$$2(R_i^2\,\sigma_{ii})^2 \qquad \text{for } i \notin I\ .$$

If we let $V(\sigma_{ii})$ denote the large sample variance of the maximum likelihood estimate of σ_{ii} we have

$$V(\sigma_{ii}) = \begin{cases} 2\,\sigma_{ii}^2/N & i \epsilon I \\ \left(1 - \frac{n_2}{N}R_i^4\right) 2\,\sigma_{ii}^2/n_1 & i \notin I\ . \end{cases}$$

2.3. Variance of Estimates of the Regression Coefficients, β_i.

In many applications, we are not interested in the σ_{ij} as such but, rather, certain functions of them, the most common being regression coefficients. If we write the matrix Σ and its inverse in partioned form as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad \Sigma^{-1} = \begin{bmatrix} \sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} ,$$

then the vector of regression coefficients for regressing the first variable on the remaining p-1 variables is given by

$$\beta = \Sigma_{22}^{-1} \Sigma_{21} = - \frac{1}{\sigma^{11}} \Sigma^{21} .$$

The elements of this vector will be denoted by β_i, $i = 2, \ldots, p$. The large sample covariance matrix for the maximum likelihood estimates of β from the incomplete data is given by

$$\tau'(I + AC)V_{\sigma 1}\tau .$$

Here τ is the matrix of derivatives of β with respect to the σ_{ij}. In particular, the element in row (u, v) column k for $1 \leq u \leq v \leq p$, $k = 2, \ldots, p$ is $\partial\beta_k/\partial\sigma_{uv}$. Alternately, we may write

$$\tau = \Delta^{-1}\Gamma$$

where Δ^{-1} is the matrix of partial derivatives of the elements of Σ^{-1} with respect to the elements of Σ and Γ is the matrix of derivatives of β with respect to the elements of Σ^{-1}. In particular, from [1] we have in row (i, j) column (s, t) of Δ^{-1} the quantity

$$- \frac{1}{k_{ij}} \left(\sigma^{is}\sigma^{jt} + \sigma^{it}\sigma^{js} \right)$$

for $1 \leq i \leq j \leq p$ and $1 \leq s \leq t \leq p$. The matrix Γ is given by

$$\Gamma = \begin{cases} \sigma^{1k}/(\sigma^{11})^2 & \text{in row 11, column k, } k = 2, \ldots, p \\ -\dfrac{1}{\sigma^{11}} & \text{in row 1k, column k} \\ 0 & \text{elsewhere.} \end{cases}$$

Combining we see that row (i, j), column k of τ is given by

$$- \frac{1}{k_{ij}\sigma^{11}} \left(2\sigma^{1i}\sigma^{1j}\sigma^{1k}/\sigma^{11} - \sigma^{1i}\sigma^{kj} - \sigma^{1j}\sigma^{ki} \right) .$$

As shown in [1],

$$V_{\sigma 1} = - \frac{1}{n_1} K\Delta$$

where K is the diagonal matrix with k_{uv} in the $(u, v)^{th}$ diagonal position. Thus if the incomplete data is ignored, the large sample variance is given by

$$\tau' V_{\sigma 1} \tau = - \frac{1}{n_1} \tau' K\Gamma .$$

In view of the above development, we see that row **k**, column h of this matrix is given by

$$\frac{1}{n_1\sigma^{11}} \left(\sigma^{kh} - \sigma^{1k}\sigma^{1h}/\sigma^{11} \right)$$

or in matrix form we have the familiar result

$$\tau' V_{\sigma 1} \tau = \frac{1}{n_1\sigma^{11}} \Sigma_{22}^{-1} .$$

To investigate the gain in precision afforded by the incomplete data we consider the matrix $\tau' AC V_{\sigma 1} \tau$. Looking first at the matrix $\tau' AC$ we see from the expressions for τ, A and $V_{\sigma 1}$ that we may write

$$\tau' AC = - \frac{n_2}{N} \Gamma' \, KC' \, U_2^{-1} \, .$$

Using the expression for U_2^{-1} from [1] we see that in row (u, v), column (i, j) of the matrix $KC' \, U_2^{-1}$ we have

$$- \frac{1}{k_{ij}} \left(\omega^{iu} \omega^{jv} + \omega^{iv} \omega^{ju} \right) \qquad \text{for } (u, v) \in I$$

$$0 \qquad \text{for } (u, v) \notin I \, .$$

Recalling the form of Γ developed earlier we see that if the index set I does not contain the integer 1, then $\tau' AC = - \frac{n_2}{N} \Gamma' \, KC' \, U_2^{-1}$ is the zero matrix. That is, if the partial data does not include the dependent variable in the regression problem then no gain in precision is achieved for estimating the regression coefficients. An investigation of the estimation procedure described in [1] shows that the estimates of the regression coefficients from the combined data are, in fact, identical to those obtained by ignoring the partial data if the partial data does not include the dependent variable. We remark in passing that the restriction does not affect the gain in precision described earlier for estimating m_i and σ_{ij}.

Since the only interesting situation is that in which the integer one is in I we hereafter assume, without loss of generality, that $I = \{1, 2, \ldots, q\}$. That is, the additional observations are made on the first q of the p variables. To compute the reduction in variance due to the incomplete data we have

$$\tau' A C V_{\sigma 1} \tau = - \frac{n_2}{n_1 N} \Gamma' \, KC' \, U_2^{-1} \, K \Gamma$$

which is readily computed from the preceding results. After some simplification we obtain in row h, column k,

$$- \frac{2n_2}{n_1 N} \left(\omega^{11}/\sigma^{11}\right)^2 \left\{(\beta_h - \alpha_h)(\beta_k - \alpha_k) + \left(\omega^{hk} - \omega^{1h}\omega^{1k}/\omega^{11}\right)/2\omega^{11}\right\}$$

for $h, k = 2, \ldots, q$

$$- \frac{2n_2}{n_1 N} \left(\omega^{11}/\sigma^{11}\right)^2 \left\{\beta_k(\beta_h - \alpha_h)\right\}$$

for $h = 2, \ldots, q$

$k = q+1, \ldots, p$

$$- \frac{2n_2}{n_1 N} \left(\omega^{11}/\sigma^{11}\right)^2 \left\{\beta_h\beta_k\right\}$$

for $h, k = q+1, \ldots, p$.

Here we have introduced the vector α with components α_j, $j = 2, \ldots, p$, which are the regression coefficients for regressing the first variable on variables 2 through q.

Denoting by $V(\beta_i)$ the large sample variance of the maximum likelihood estimate of β based on the N observations we have

$$V(\beta_i) = \frac{1}{n_1}\left\{\left(\sigma^{11}\sigma^{ii} - (\sigma^{1i})^2\right)/(\sigma^{11})^2\right\} \cdot \left\{1 - \frac{2n_2}{N}(\omega^{11})^2\left[(\beta_i - \alpha_i)^2 + \left(\omega^{11}\omega^{ii} - (\omega^{1i})^2\right)/2(\omega^{11})^2\right] \Big/ \left(\sigma^{11}\sigma^{ii} - (\sigma^{1i})^2\right)\right\}$$

for $i = 2, \ldots, q$

and

$$V(\beta_i) = \frac{1}{n_1}\left\{\left(\sigma^{11}\sigma^{ii} - (\sigma^{1i})^2\right)/(\sigma^{11})^2\right\}\left\{1 - \frac{2n_2}{N}(\omega^{11})^2\,\beta_i^2/\left(\sigma^{11}\sigma^{ii} - (\sigma^{1i})^2\right)\right\}$$

for $i = q+1, \ldots, p$.

In summary, we see that the large sample variance expressions for estimating m_i, σ_{ii} or β_i are all of the form

$$\frac{1}{n_1} V_1(\sigma)\left\{1 - \frac{n_2}{N} f(\sigma)\right\}$$

where $\frac{1}{n_1} V_1(\sigma)$ represents the variance if only the n_1 complete vectors are used and $f(\sigma)$ is a function of the elements of Σ with the property that $0 \leq f(\sigma) \leq 1$. In the case of m_i and σ_{ii} we have seen that for $i = 1, \ldots, q$, or more generally, $i \in I$ that $f(\sigma) = 1$ and for $i = q+1, \ldots, p$, or $i \notin I$, that $f(\sigma)$ depends only on R_i^2, the multiple correlation coefficients obtained when regressing variable $i \notin I$ on the variables in I. We have just noted that when considering β_i, the expressions for $f(\sigma)$ are more complex, requiring additional knowledge about the parameters, σ_{ij}. There are numerous alternate expressions for these $f(\sigma)$ which may be obtained. The choice of these depends on the form of the prior information available. For example, expressions involving only multiple R^2 for various regressions appear to be most palatable. To illustrate, let R_{1p}^2, R_{1q}^2, and $R_{1p.i}^2$ denote the multiple correlation coefficients for regressing variable one on variables 2 through p, 2 through q, and 2 through p excluding variable i, respectively. Let $\rho_{1i.p}^2$ denote the partial correlation between variables 1 and i given the remaining p-2 variables and let R_{ip}^2 and R_{iq}^2 denote multiple correlation coefficients for regressing variable i on variables 1 through p and 1 through q, respectively, excluding variable i. The following relations then enable us to obtain alternate expressions for $V(\beta_i)$.

$$\beta_i = -\sigma^{1i}/\sigma^{11} \qquad \alpha_i = -\omega^{1i}/\omega^{11}$$

$$(\sigma^{11}\sigma_{11})^{-1} = 1 - R_{1p}^2 \qquad (\omega^{11}\sigma_{11})^{-1} = 1 - R_{1q}^2$$

$$(\sigma^{ii}\sigma_{ii})^{-1} = 1 - R_{ip}^2 \qquad (\omega^{ii}\sigma_{ii})^{-1} = 1 - R_{iq}^2$$

$$\rho_{1i.p}^2 = (\sigma^{1i})^2/\sigma^{11}\sigma^{ii} \qquad \rho_{1i.p}^2 = \frac{R_{1p}^2 - R_{1p.i}^2}{1 - R_{1p.i}^2} .$$

To illustrate, the expression for $f(\sigma)$ in the expression for $V(\beta_i)$ for $i = q+1, \ldots, p$ is given by

$$f(\sigma) = 2(1 - R^2_{1p})\,(1 - R^2_{1q})\,(R^2_{1p} - R^2_{1p.i})\ .$$

For $i = 2, \ldots, q$, the expression is more complex, involving also the quantities R^2_{ip} and R^2_{iq}.

In the next section we discuss the problem of determining optimal sample allocations based on prior knowledge of the $f(\sigma)$. Since these quantities arise from asymptotic variance expressions, it is natural to ask how appropriate these expressions are for small samples. This is, in general, quite difficult but Monte Carlo studies for $p = 3$ indicate that the asymptotic variances are very close to the small sample variances for n_1 and n_2 of the order of 15. This limited evidence is offered only to suggest that the sample allocations to be developed may be reasonably good especially in view of the, less than exact, prior information.

3. The Optimum Sample Allocation Problem.

In this section we consider the problem of minimizing the cost of collecting the data subject to meeting requirements on the variances of the parameter estimates. It is assumed that the variance requirements are specified in terms of what would be obtained if all complete vectors are observed. Thus, in general the restrictions are of the form

$$\frac{1}{n_1} V_1(\theta)\left(1 - \frac{n_2}{N} f(\sigma)\right) \leq \frac{1}{m} V_1(\theta)$$

where $\frac{1}{m} V_1(\theta)$ is the variance obtained if m complete vectors are observed.

To be more specific, consider the problem of estimating only the m_i and the σ_{ii}. It is natural to ask for higher precision for $i \in I$, say $i = 1, \ldots, q$, that for $i \notin I$. Thus we shall require $V_1(\theta)/M$ for $i = 1, \ldots, q$ and $V_1(\theta)/m$

for $i = q+1, \ldots, p$, where $M > m$. Since $f(\sigma) = 1$ for $i = 1, \ldots, q$ the constraints for this problem are

$$N \geq M$$

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} R_i^2\right) \leq \frac{1}{m}$$

$$\frac{1}{n}\left(1 - \frac{n_2}{N} R_i^4\right) \leq \frac{1}{m} . \qquad i = q+1, \ldots, p$$

Assume further that the cost of gathering the data is C_1 for observing $X_1, \ldots, X_q$ and C_2 for observing $X_{q+1}, \ldots, X_p$, where C_2 is considerably greater than C_1. This is just the situation in which one would be tempted to gather some incomplete data. The cost of gathering the data is thus given by

$$C = n_1(C_1 + C_2) + n_2 C_1 = NC_1 + n_1 c_2 .$$

To determine the optimal values of n_1 and n_2 we must solve the constrained optimization problem of minimizing C subject to restrictions of the type illustrated above and the obvious requirements that n_1 and n_2 be non-negative. Mathematical programming algorithms are available for solving such problems but we shall now see that the problem can be easily solved analytically. To develop the solution we first discuss a special case in sub-section 3.1 then develop the general solution in sub-section 3.2.

3.1. A Single Complete Sample Size Specification.

Consider the problem of minimizing C subject to the constraints

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} f_j(\sigma)\right) \leq \frac{1}{m} \qquad j = 1, \ldots, J$$

$$N \geq M$$

$$n_1, n_2 \geq 0$$

where the $f_j(\sigma)$ are obtained from the expressions for $V(m_i)$, $V(\sigma_{ii})$ and $V(\beta_i)$. The essential point to notice is that, as in our previous illustration, the right-hand side of the inequalities, $j = 1, \ldots, J$, is always $1/m$. This is just the situation in which the variance requirements on all parameters for which $f(\sigma) < 1$ are based on the same complete sample size, namely m.

It can be shown in this case that only one of the inequalities, $j = 1, \ldots, J$, is ever active, namely that constraint for which $f_j(\sigma)$ is maximum. Letting

$$F = \max_j f_j(\sigma)$$

the problem described above is thus equivalent to that of minimizing C subject to the constraints

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} F\right) \leq \frac{1}{m}$$

$$N \geq M$$

$$n_1, n_2 \geq 0 .$$

The solution of the problem is quite simple and is shown in Table I for various situations involving the relative magnitudes of C_2 and C_1 as well as M and m. To simplify Table I we have introduced the symbols

$$n_M = m\,M(1 - F)/(M - m\,F)$$

$$n_o = m\left((1 - F) + (F(1 - F)C_2/C_1)^{\frac{1}{2}}\right)$$

and

$$S_F = m^2 F(1 - F)/\left(n_m - m(1 - F)\right)^2 .$$

One advantage of being able to solve the problem analytically as opposed to a numerical solution is that the user can easily see how sensitive his solution is to his prior knowledge of F.

Magnitudes of C_1, C_2, M, m	Solution
1. $m \leq M$ $\quad C_2/C_1 \leq S_F$	$n_1 = n_M \quad n_2 = M - n_M$
2. $m \leq M$ $\quad C_2/C_1 \geq S_F$ or $m > M$ $\quad C_2/C_1 > (1 - F)/F$	$n_1 = n_o$ $n_2 = n_o(m - n_1)/(n_o - m(1 - F))$
3. $m > M$ $\quad C_2/C_1 \leq (1 - F)/F$	$n_1 = m \quad n_2 = 0$

TABLE 1. Solutions for Single, Complete Sample Size Specification

3.2. Different Complete Sample Size Specifications.

If the variance requirements are based on different, complete sample sizes, then no single constraint dominates the others. In this case we wish to minimize C subject to the constraints

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} f_j(\sigma)\right) \leq 1/m_j \qquad j = 1, \ldots, J$$

$$N \geq M$$

$$n_1, n_2 \geq 0$$

where the m_j are generally different.

An approximate solution to this problem can be obtained by computing the solution described in Table I replacing F by $f_j(\sigma)$ and m by m_j for $j = 1, \ldots, J$ and selecting from these solutions the one which gives minimum cost.

This approximate solution may well be optimal. The optimality is easily verified by checking to see if the selected values of n_1 and n_2 satisfy all of the

constraints. If so, the solution is optimal. If not, the solution is at the intersection of two constraints, say,

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} f_1(\sigma)\right) \leq \frac{1}{m_1}$$

$$\frac{1}{n_1}\left(1 - \frac{n_2}{N} f_2(\sigma)\right) \leq \frac{1}{m_2} \quad .$$

The point of intersection is given by,

$$n_1 = m_1 m_2 (f_1 - f_2)/(m_1 f_1 - m_2 f_2)$$

$$n_2 = n_1 (m_1 - n_1)/(n_1 - m_1(1 - f_1)) \quad .$$

Evaluating C for each possible intersection for which n_1 and $n_2 \geq 0$ and $n_1 + n_2 \geq M$ and selecting that solution which minimizes C will then yield the optimum solution.

The general solution described in this section is conceptually quite simple but its practicality depends on the number of constraints, J, and the degree of belief in the prior estimates of the $f_j(\sigma)$. The approximate solution indicated above may well be adequate in most cases.

4. Extensions.

We have developed the large sample variance formulas and discussed the optimal sample allocation problem for a particular case of the incomplete data problem. That is, the situation in which we make n_1 observations on variables 1 through p and n_2 observations on variables 1 through q for $q < p$. In [1] and [2] much more general incomplete data problems are considered but in general the role of the sample size is not so simple. One class of problems, called 'nested',

does suggest further consideration. To illustrate, suppose that in addition to the above we make n_3 observations on variables 1 through r for $r < q$. This is a 'nested' situation. In this case, the variance expressions are of the form

$$\left(1 - \frac{n_2}{N_2} f_1(\sigma) - \frac{n_3}{N_3}\left(f_2(\sigma) - \frac{n_2}{N_2} f_3(\sigma)\right)\right) V(\theta)/n_1$$

where $N_2 = n_1 + n_2$, $N_3 = N_2 + n_3$, and the $f_i(\sigma)$ are functions depending on the σ_{ij}. The development of the variance formulas proceeds as in Section 2 and the sample allocation problem is the natural extension of that described in Section 3. The solution is generally more complicated requiring, usually, a numerical solution at least for the case of different, complete sample size specifications.

5. Acknowledgments.

The authors wish to acknowledge partial support during the course of this research by the National Aeronautics and Space Administration, Manned Spacecraft Center (Grant No. NGR 44-001-095).

References

[1] Hocking, R. R., Oxspring, H. H. and Waldron, B. R., "Maximum Likelihood Estimation with Incomplete Normal Data: Part I," submitted to J. American Statistical Association.

[2] Hocking, R. R., Oxspring, H. H. and Waldron, B. R., "Maximum Likelihood Estimation with Incomplete Normal Data: Part II," submitted to J. American Statistical Association.

[3] Hocking, R. R., and Smith, W. B., "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations," J. American Statistical Association, Vol. 63, 1968, 159-173.