Published in final edited form as:

Commun Stat Theory Methods. 2019 ; 48(5): 1149–1165. doi:10.1080/03610926.2018.1425447.

On the Distribution of Summary Statistics for Missing Data

B.M. Ringham*,

Department of Biostatistics and Informatics, University of Colorado Denver

S.M. Kreidler,

Neptune and Company

K.E. Muller,

Department of Health Outcomes and Policy, University of Florida, Gainesville

D.H. Glueck

Department of Biostatistics and Informatics, University of Colorado Denver

Abstract

Under an assumption that missing values occur randomly in a matrix, formulae are developed for the expected value and variance of six statistics that summarize the number and location of the missing values. For a seventh statistic, a regression model based on simulated data yields an estimate of the expected value. The results can be used in the development of methods to control the Type I error and approximate power and sample size for multilevel and longitudinal studies with missing data.

Keywords

missing data; power; multilevel; longitudinal; multinomial

1. Introduction

Missing data are a common challenge in studies with repeated measures of outcomes. Here, we use the term ‘repeated measures’ to indicate any combination of longitudinal, spatial, multilevel or multivariate designs (Ringham *et al.*, 2016). For repeated measures designs, missing data can lead to low rates of mixed model convergence (Catellier and Muller, 2000; Fouladi and Shieh, 2004; Serrano, 2008; Gurka, Edwards and Muller, 2011), inflated Type I error rate (Gurka, Edwards and Muller, 2011) and low statistical power (Ringham *et al.*, 2016).

For a concrete example of missing data in a repeated measures study, suppose a study investigator plans to measure the levels of three biomarkers on each of N_i participants. The participants are the independent sampling units (Muller and Stewart, 2006) and the biomarker levels are three repeated measurements taken on each independent sampling unit. One possible realization of the experiment involves complete data. Complete data occurs if

*Corresponding author, 13001 E 17th Pl, Mail Stop B119, Aurora, CO 80045, U.S.A., brandy.ringham@ucdenver.edu (B.M.Ringham).

every planned measurement is made on every planned independent sampling unit. Yet, in other realizations of the experiment, one or more assays might fail, causing some participants to be missing one, two or all three biomarker measures.

In the present work, we assume that the location and extent of the missing data is determined by a stochastic process that is independent of the observed data. The stochastic process assumption aligns with the missing completely at random process and a general missing data pattern, as defined by Little and Rubin (pgs. 4–19 2002). Continuing the example from above, biomarker levels may be missing completely at random if the assay used to assess the biomarker levels fails due to random chance. Assay failures are a common missing data problem in biomarker studies. See, for instance, the gene expression and DNA microarray studies by DeRisi, Iyer and Brown (1997), Spellman *et al.* (1998) and Alizadeh *et al.* (2000) as referenced by Aittokallio *et al.* (2010).

Each observed set of data is a particular realization generated by the missing data process. A data analyst is charged with drawing inference from an already observed single set of data, with a particular pattern of missing data. For data analysts, Barton and Cramer (1989) and Catellier and Muller (2000) introduced missing data summary statistics to quantify an observed pattern of missing data from a single realization of an experiment.

In contrast to a data analyst, a study planner must imagine all possible missing data patterns, and account for their likelihood. For study planners, we derive the expected value and variance of several missing data summary statistics. The expected value describes the amount and location of the missing data on average. The variance describes the spread of the missing data summary statistic over all possible realizations of the experiment. We have not found a derivation of the moments in any published work.

The distributional results for the missing data summary statistics can be used in the following ways. In the design stages of a study, researchers can use the expected value of the amount of missing data to make appropriate adjustments to the study design. Study investigators facing large amounts of missing data may choose to target phases of data collection for additional investments in quality control, use analytic approaches to control Type I error or select robust analysis strategies to avoid convergence difficulties. In addition, investigators who understand the potential for missing data may choose to increase sample size, while considering the ethical balance between the desire to prove a scientific result and the possibility of harm for study participants.

Missing data summary statistics can be used as an analytic foundation for power approximations for a variety of study designs, models and hypotheses. Ringham *et al.* (2016) utilized three missing data summary statistics to develop a power approximation for general linear multivariate or balanced linear mixed models with Gaussian outcomes. Tu *et al.* (2007) used a missing data summary statistic in their power approximations for the general linear mixed model and generalized estimating equation (GEE) models under compound symmetry. The Tu work applies to binary, categorical or ordinal outcomes as well.

The seven missing data summary statistics considered in the present work appear in Table 1, with the notation mirroring that of Table 1 in Catellier and Muller (2000). Here, we provide

an intuitive definition of each missing data summary statistic. A strict, mathematical definition follows in Section 2. The first missing data summary statistic, N_{m1} , is the number of independent sampling units for which all measurements are observed, *i.e.*, the number of complete cases. The N_{m3} , N_{m9} and N_{m10} statistics are functions of the number of observed measurements for each independent sampling unit: N_{m3} is the minimum, N_{m10} is the maximum and N_{m9} is the mean number of observed measurements across all independent sampling units. Statistics N_{m2} and N_{m6} are functions of the number of non-missing pairs of measurements observed on each sampling unit. The N_{m2} statistic is the minimum number of non-missing pairs across all independent sampling units while N_{m6} is the mean. The last missing data summary statistic, N_{m11} , is technically not a missing data summary statistic at all. Rather, it is the planned number of independent sampling units, or the number of independent sampling units if there were no missing data.

The manuscript is divided into six sections. Section 1 gives the context of the derivations. Section 2 lists definitions and demonstrates the notation using an example. Section 3 contains the results of the derivations and Section 4 discusses extensions of the results to future work.

2. Definitions and Example

2.1. Definitions

We begin with a brief discussion of general notation. For p and r integers and $p > r$, we use ${}_pC_r$ to indicate $p!/[r!(p-r)!]$. We denote a matrix as $A = \{a_{ij}\}$ with dimensions $(r \times c)$ and its transpose as $A' = \{a_{ji}\}$. The i th row of A is indicated as A_i and the j th column of A as a_j (Muller and Stewart, 2006). We write $\mathbf{1}_{(r,c)}$ to denote $(r \times c)$ matrices with all elements equal to 1. We denote the expected value of the random variable X as $E(X)$, the variance as $V(X)$ and the covariance of random variables X and Y as $\text{Cov}(X, Y)$. Finally, $F_X(x)$ is the cumulative density function of the random variable X evaluated at x .

For $i \in \{1, 2, \dots, N_t\}$, $j \in \{1, 2, \dots, p\}$, consider an $(N_t \times p)$ matrix $\mathbf{Y} = \{y_{ij}\}$ containing the outcomes of a repeated measures experiment. Each row of $\mathbf{Y}$ corresponds to p planned repeated measurements on independent sampling unit i . Characterize the location of the missing data in $\mathbf{Y}$ with the $(N_t \times p)$ matrix $\mathbf{D} = \{d_{ij}\}$ so that

$$d_{ij} = \begin{cases} 1 & y_{ij} \text{ is non-missing} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For $i \in \{1, 2, \dots, N_t\}$, $j \in \{1, 2, \dots, p\}$, assume that d_{ij} are independently and identically distributed Bernoulli($1-\pi$) and do not depend on y_{ij} . The assumption corresponds to data that are missing completely at random (Little and Rubin, 2002). Equiprobable missing data might occur, for example, when a biochemical assay fails at random. Suppose that d_j is column j of $\mathbf{D}$. The number of missing values for independent sampling unit i is

$$r_i = p - \sum_{j=1}^p d_{ij}. \text{ The number of non-missing values in column } j \text{ of } \mathbf{D}, d_j, \text{ is } N_{jj} \equiv \mathbf{1}'_{N_t} d_j.$$

In a repeated measures study with j planned measurements, N_{jj} is the number of times that the measurement j is non-missing. Note that $N_{jj} \in \{0, 1, \dots, N_t\}$. See Figure 1 for a hypothetical depiction of $\mathbf{Y}$, $\mathbf{D}$, r_i and N_{jj} .

In order to derive the covariance of the missing data summary statistics, we must look at pairs of missing measurements for the same independent sampling unit i . For $j \neq j'$, $(y_{ij}, y_{ij'})$ is a non-missing pair if neither y_{ij} nor $y_{ij'}$ are missing (Catellier and Muller, 2000). For $q = p(p-1)/2$, define the $(N_t \times q)$ matrix of non-missing pair indicators $\mathbf{P} = \{p_{ik(j, j')}\} = \{d_{ij}d_{ij'}\}$. Here, the index k is defined in quasi-lexicographic order so that $k(1,3) = 2, \dots$, and $k(p-1, p) = q$. One can calculate the number of non-missing pairs in columns j and j' of $\mathbf{D}$ by $N_{jj'} = \mathbf{1}'_{N_t} \mathbf{P}_{k(j, j')}$. Figure 2 shows the transformation from the hypothetical matrix $\mathbf{D}$ to the corresponding matrix $\mathbf{P}$.

Using the notation above, we now define the missing data summary statistics examined in the present work (Table 1). The missing data summary statistics are listed in increasing order, with the exception of N_{m3} . The value of N_{m3} can sometimes be greater than that of N_{m6} (Catellier and Muller, 2000).

In order to derive the expectation and variance of the missing data summary statistics, we must describe the number, location and correlation between the missing values. To simplify the derivation, we will consider sub-parts of the data that contain the same number of missing values per independent sampling unit. We sort the rows of $\mathbf{D}$ into submatrices set up so that every row in a submatrix has the same number of missing values, say r . Explicitly, for $r \in \{0, 1, \dots, p\}$, $N_t(r) \in \{0, 1, 2, \dots, N_t\}$ and $\sum_{r=0}^p N_t(r) = N_t$, let $N_t(r)$ be the number of rows in $\mathbf{D}$ such that $r_i = r$. Then the $[N_t(r) \times p]$ matrix $\mathbf{D}(r)$ is the submatrix of $\mathbf{D}$ such that $r_i = r$ for each row in $\mathbf{D}(r)$. Let the submatrix $\mathbf{P}(r) = \{p_{ik(j, j')}\} = \{d_{ij}d_{ij'}\}$, where the d_{ij} are elements of $\mathbf{D}(r)$ and $k(j, j')$ is defined as above. We can then calculate the number of non-missing pairs in columns j and j' of $\mathbf{D}(r)$ by $N_{jj'}(r) \equiv \mathbf{1}'_{N_t(r)} \mathbf{P}_{k(j, j')}(r)$. Figure 3 gives an example of a $\mathbf{P}(r)$ matrix and the corresponding calculation of $N_{jj'}$.

In order to calculate the expected value and variance of the missing data summary statistics, we need to define notation for all possible realizations of the pattern of missingness. Each of the $N_t(r)$ rows of $\mathbf{D}(r)$ will take on one of $\theta(r)$ possible patterns of 0's and 1's, with

$$\theta(r) = \begin{cases} pC_r & p-r > 1 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Consider one independent sampling unit with r values missing out of p repeated measures. Define $\mathbf{D}^*(r)$ to be the $[\theta(r) \times p]$ matrix containing one row for each possible pattern of missingness that may occur. Further, define $\mathbf{P}^*(r)$ to be the corresponding $[\theta(r) \times q]$ matrix of non-missing pairs containing each possible pattern of non-missing pairs that may occur.

The columns of $\mathbf{P}^*(r)$ are related in one of two ways. The two columns of $\mathbf{P}^*(r)$, $p_{k(j,j')}(r)$ and $p_{k(j,j'')}(r)$ share a common subscript, j , and are referred to as *siblings*. Conversely, $p_{k(j,j')}(r)$ and $p_{k(j'',j''')}(r)$ share no common subscript, and are referred to as *strangers* (Figure 4).

Just as we did for $\mathbf{P}(r)$, we can calculate the number of non-missing pairs for each column of $\mathbf{P}^*(r)$. For the derivation of the covariance of the missing data summary statistics, we also require the number of non-missing pairs calculated across rows and across the specific groupings of the columns of $\mathbf{P}^*(r)$ that we have defined as *siblings* and *strangers*. Let $\psi(r) \equiv \mathbf{P}^*(r) \mathbf{1}_{\theta(r)}$ be the number of non-missing pairs in a row of $\mathbf{P}^*(r)$. Let $k(r) \equiv \mathbf{1}'_{\theta(r)} \mathbf{P}^*(r)$ be the number of non-missing pairs in a column of $\mathbf{P}^*(r)$. For $j \neq j' \neq j'' \neq j'''$, define the following. Let $\gamma(r) \equiv p_{k(j,j')}^* p_{k(j,j'')}^*$ be the number of rows of $\mathbf{P}^*(r)$ where the *siblings* $p_{k(j,j')}^*$ and $p_{k(j,j'')}^*$ both contain a non-missing pair. Finally, let $\omega(r) \equiv p_{k(j,j')}^* p_{k(j'',j''')}^*$ be the number of rows of $\mathbf{P}^*(r)$ where the *strangers* $p_{k(j,j')}^*$ and $p_{k(j'',j''')}^*$ both contain a non-missing pair. Figure 4 shows an example $\mathbf{P}^*(r)$ matrix and demonstrates the concept of *siblings* and *strangers*.

2.2. Example

We give an example computation of the terms defined in Section 2.1. As shown in Figure 1, the matrix $\mathbf{D}$ is a transformation of the matrix $\mathbf{Y}$. For every observed data point in $\mathbf{Y}$, we record a value of 1 at the parallel locations in $\mathbf{D}$. For every missing data point in $\mathbf{Y}$, we record a value of 0 at the parallel locations in $\mathbf{D}$. An example transformation from the data indicator matrix, $\mathbf{D}$, to the matrix of non-missing pairs, $\mathbf{P}$, is shown in Figure 2. The number of non-missing values for each column of $\mathbf{D}$ appear as marginal totals for the matrix $\mathbf{D}$. The number of non-missing pairs for each pair of columns in $\mathbf{D}$ are the marginal totals for the example matrix $\mathbf{P}$.

In Figure 3, we show the submatrices $\mathbf{P}(r)$ for the example matrices $\mathbf{D}$ and $\mathbf{P}$ given in Figure 2. We can calculate $N_{jj'}$ for each pair of $j \neq j' \in \{1, 2, 3, 4\}$. Consider, specifically, the example submatrices

$$\mathbf{D}(1) = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (3)$$

and the corresponding

$$\mathbf{P}(1) = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \quad (4)$$

The solutions for $N_{jj'}$ are $N_{12}(1)=1$, $N_{13}(1)=0$, $N_{14}(1)=1$, $N_{23}(1)=1$, $N_{24}(1)=2$, and $N_{34}(1)=1$.

The matrix $\mathbf{D}(1)$ in Equation 3 and the corresponding matrix $\mathbf{P}(1)$ in Equation 4 contain one possible realization of the pattern of missing data for each of the $N_t = 2$ independent sampling units. However, in order to calculate the expectation and variance of the missing data summary statistics, we need to consider all possible realizations of the pattern of missing data. For $p = 4$ and $r = 1$, the total number of possible realizations of the missing data pattern is $\theta(r) = 4$. We can write the matrix

$$\mathbf{D}^*(1) = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (5)$$

and the corresponding matrix

$$\mathbf{P}^*(1) = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad (6)$$

to describe the four possible realizations of the missing data pattern.

The columns of $\mathbf{P}^*(1)$ are related in two different ways: as *siblings* or as *strangers*. In Figure 4, columns $\mathbf{p}^*_1(1)$, $\mathbf{p}^*_2(1)$, $\mathbf{p}^*_3(1)$, $\mathbf{p}^*_4(1)$ and $\mathbf{p}^*_5(1)$ are considered *siblings* because they are derived from a shared column of $\mathbf{D}^*(1)$, column $\mathbf{d}^*_1(1)$. The columns $\mathbf{p}^*_1(1)$ and $\mathbf{p}^*_6(1)$ are considered *strangers* because they are derived from non-overlapping columns of $\mathbf{D}^*(1)$: columns $\mathbf{d}^*_1(1)$ and $\mathbf{d}^*_2(1)$ for column $\mathbf{p}^*_1(1)$ and columns $\mathbf{d}^*_3(1)$ and $\mathbf{d}^*_4(1)$ for column $\mathbf{p}^*_6(1)$.

We continue the example by computing $\psi(r)$, $\kappa(r)$, $\gamma(r)$ and $\omega(r)$, functions of the number of non-missing pairs in the columns of *siblings* and *strangers* (Figure 4). The solutions are as follows: $\psi(1) = 3$, $\kappa(1) = 2$, $\gamma(1) = 1$ and $\omega(1) = 0$. Note that $\gamma(r)$ and $\omega(r)$ are the same regardless of which set of *siblings* or *strangers* are chosen for the calculation.

3. Results

We calculated the expectation and variance of the missing data summary statistics using two approaches. An analytic approach was used for N_{m1} , N_{m3} , N_{m6} , N_{m9} and N_{m10} . The exact derivation of $E(N_{m2})$ proved intractable. Instead we used a Monte Carlo simulation and a general linear univariate model to estimate $E(N_{m2})$. The statistic N_{m11} , the planned number of independent sampling units, is a constant and thus required no derivations.

The analytic results are presented in Tables 2 and 3, with proofs appearing in the Appendix. All analytic results were computed using a series of conditioning arguments (Ross, 1984,

pgs. 273, 279). The statistics N_{m3} and N_{m10} are the minimum and maximum of N_{jj} . We show that N_{jj} are independently and identically distributed binomial random variables with parameters N_t and $1 - \pi$. The cumulative distribution function is given by

$$F_{N_{jj}}(N_{jj}) = \begin{cases} 1 & N_{jj} > N_t \\ \sum_{i=0}^{N_{jj}} \binom{N_t}{i} (1-\pi)^i \pi^{N_t-i} & 0 \leq N_{jj} \leq N_t \\ 0 & N_{jj} < 0 \end{cases} \quad (7)$$

We use previously derived formulae to describe the distributions of the minimum and maximum (David, 1981, Equations 2.1.1, 2.1.2, 2.1.4). The expectation and variance of N_{m6} are defined in terms of the constant

$$h = \sum_{i=3}^{p=1} i [(i-1)C_2] \quad (8)$$

and the variance and covariance of $N_{jj'}$. Recall that

$$\theta(r) = \begin{cases} p^{C_r} & p-r > 1 \\ 1 & \text{otherwise} \end{cases}$$

With $\lambda(r) = 1/\theta(r)$ and $\alpha(r) = p^{C_r} \pi^r (1-\pi)^{p-r}$, it can be shown that the variance of $N_{jj'}$ is

$$\begin{aligned} V(N_{jj'}) &= N_t \sum_{r=0}^p \alpha(r) \kappa(r) \lambda(r) [1 - \kappa(r) \lambda(r)] \quad (9) \\ &+ N_t \sum_{r=0}^p \alpha(r) [1 - \alpha(r)] \kappa(r)^2 \lambda(r)^2 \\ &- 2N_t \sum_{r=1}^p \sum_{r'=1}^p \alpha(r) \alpha(r') \kappa(r) \kappa(r') \lambda(r) \lambda(r'). \end{aligned}$$

There are two different types of covariances for $N_{jj'}$: the covariance between $N_{jj'}$ and $N_{jj''}$ and the covariance between $N_{jj'}$ and $N_{j''}N_{jj''}$. The covariances can be written as

$$\text{Cov}(N_{jj'}, N_{jj''}) = N_t \sum_{r=0}^p \{ \alpha(r) \lambda(r) \gamma(r) [1 - \lambda(r)] \quad (10)$$

$$- \alpha(r) \lambda(r)^2 [\kappa(r) - \gamma(r)]$$

$$+ \alpha(r) \kappa(r)^2 \lambda(r)^2 [1 - \alpha(r)] \}$$

$$- 2N_t \sum_{r=1}^p \sum_{r'=1}^p \alpha(r) \alpha(r') \kappa(r) \kappa(r') \lambda(r) \lambda(r')$$

and

$$\text{Cov}(N_{jj'}, N_{j''j'''}) = N_t \sum_{r=0}^p \{ \alpha(r) \lambda(r) \omega(r) [1 - \lambda(r)] \quad 11$$

$$- \alpha(r) \lambda(r)^2 [\kappa(r) - \omega(r)]$$

$$+ \alpha(r) \kappa(r)^2 \lambda(r)^2 [1 - \alpha(r)] \}$$

$$- 2N_t \sum_{r=1}^p \sum_{r'=1}^p \alpha(r) \alpha(r') \kappa(r) \kappa(r') \lambda(r) \lambda(r').$$

Since the exact analytic derivation of $E(N_{m2})$ proved intractable, we used a Monte Carlo simulation and a general linear univariate modeling approach to provide approximate estimates. Recall $N_{m2} = \min(\{N_{jj'}\})$. For the Monte Carlo simulation, we specified a range of values for N_t , p and π . Specifically, N_t was an even integer between 8 and 50, p was an integer between 2 and 12 and $\pi \in \{0, 0.02, \dots, 0.3\}$. All possible combinations of the parameter values yields 2, 112 unique experimental conditions. For each experimental condition, we generated ten thousand realizations of the $(N_t \times p)$ matrix $\mathbf{D}$. The elements of $\mathbf{D}$ were sampled from a Bernoulli($1 - \pi$) distribution. The 10, 000 realizations were indexed by $g \in \{1, 2, \dots, 10,000\}$. We supposed that j and j' were elements of $S = \{1, 2, \dots, p\}$, and that $j \neq j'$. The realization-specific number of non-missing pairs was defined as

$N_{jj',g} = \mathbf{1}'_{N_t} p_{k(j,j'),g}$, $g \in \{1, 2, \dots, 10,000\}$. We calculated $N_{m2,g} = \min_S(\{N_{jj',g}\})$. The expected value of N_{m2} , here denoted as $E[N_{m2}(N_t, p, \pi)]$, was estimated as the sample mean over the 10,000 realizations, so that $\bar{N}_{m2}(N_t, p, \pi) = (1/10,000) \sum_{i=1}^{10,000} N_{m2,g}$.

We fit a general linear univariate model with the sample mean $\bar{N}_{m2}(N_t, p, \pi)$, as the outcome and N_t , p and π , second degree polynomials in N_t , p and π and all two- and three-way interaction terms as the predictors. Model fit was assessed using established principles of assumption diagnostics (Muller and Fetterman, 2000, Ch. 7). The final model had $R^2 = 0.9982$. All terms were significant.

With regression coefficients given in Table 3, the following formula provides an estimate of $E[N_{m2}(N_t, p, \pi)]$:

$$E[N_{m2}(N_t, p, \pi)] \approx \begin{cases} N_t & \pi = 0 \\ \sum_{i=1}^{15} \hat{\beta}_i X_i & 0 < \pi < 1 \\ 0 & \text{otherwise} \end{cases} \quad 12$$

In order to assess the accuracy of the estimation, we calculated the absolute percent deviation, defined as the absolute value of

$$100 \times \{E[N_{m2}(N_t, p, \pi)] - \bar{N}_{m2}(N_t, p, \pi)\} / \bar{N}_{m2}(N_t, p, \pi).$$

The maximum absolute percent deviation across all experimental conditions was 2.4.

For example, $E[N_{m2}(N_t = 12, p = 4, \pi = 0.05)] \approx 10.24$. Note that, as expected, $E(N_{m2}) \approx 10.24$ falls between $E(N_{m1}) = 9.77$ and $E(N_{m3}) = 10.60$. Both $E(N_{m1})$ and $E(N_{m3})$ were calculated using the analytic formulas in Table 2.

4. Discussion

The assumptions about missing data in the present work arise from a different point of view than the classic Little and Rubin (2002) approach. For data analysis, as considered by Little and Rubin (2002), the pattern of missing data is quite important, as is the assumption that the chance that the data are missing is unrelated to the values of the data. For power analysis, the application of interest here, one thinks instead of multiple possible realizations of the data, each generated by a probability process. The specific probability process assumed for the present work induces data that are missing completely at random in the Little and Rubin sense. As a result of the assumption about the probability process, the observed data will follow a “general” pattern of missingness (Little and Rubin, 2002, Figure 1.1).

We assume that every data point for every independent sampling unit is missing with equal probability. More complex missing data processes, however, are common in longitudinal

epidemiological studies and in randomized controlled clinical trials. The probability that data are missing may vary across repeated measures and independent sampling units. The approach of using conditioning arguments provides a general strategy for deriving the moments of missing data summary statistics. In future work, the general strategy can be applied to more complex missing data processes. One could introduce more stringent assumptions about the missing data process. The probability of missing data could increase over time, depend on observed values or be associated with observed or unobserved covariates.

The results of the present work will allow researchers to predict potential missing data pitfalls in the design stages of their study. Researchers aware of possible problems can make appropriate adjustments before the study begins. For example, researchers could increase the sample size, alter sampling strategies or plan statistical analyses to control the Type I error rate. In addition, the results directly inform the development of power approximations for the general linear multivariate model and general linear mixed models when data are missing (Ringham *et al.*, 2016).

Acknowledgments

The research presented in this paper was supported by three grants: NCI 1R03CA136048-01A1, NIGMS 9R01GM121081-05 and NIDCR 3R01DE020832-01A1S1. The NCI and NIGMS grants were awarded to the Colorado School of Public Health, D. H. Glueck, Principal Investigator. The NIDCR grant was a minority supplement awarded to the University of Florida, K. E. Muller, Principal Investigator, with a subaward to the Colorado School of Public Health, D. H. Glueck, Site Principal Investigator. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institute of General Medical Science, the National Institute of Dental and Craniofacial Research nor the National Institutes of Health. This manuscript was submitted to the Department of Biostatistics and Informatics at the Colorado School of Public Health, University of Colorado Denver, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics for B. M. Ringham.

Appendix: Proofs

In the following section, we derive the results presented in Tables 2 and 3. The proofs are indexed to match the results as they appear in Tables 2 and 3. For instance, Proof 1.1 provides the derivation of Result 1.1.

Recall that $\alpha(r) = {}_p C_r \pi^r (1 - \pi)^{p-r}$ and $\lambda(r) = 1/\theta(r)$. The constants $\kappa(r)$, $\psi(r)$, $\gamma(r)$ and $\omega(r)$ were defined in Section 2.1. We now provide formulae to evaluate the functions. Let

$$\psi(r) = {}_{(p-r)} C_2, \quad \kappa(r) = {}_{(p-2)} C_{(p-r-2)}, \quad \gamma(r) = {}_{(p-3)} C_{(p-r-3)} \text{ and } \omega(r) = {}_{(p-4)} C_{(p-r-4)}.$$

In addition let $h = \sum_{i=3}^{p-1} i \left[{}_{(i-1)} C_2 \right]$, the number of *strangers* relative to any column of $P^*(r)$.

Proof 1.1 and 1.2 (N_{m1}).

Given the definitions and assumptions in Section 2, it follows that R_i are independently and identically distributed as binomial random variables with parameters N_t and $1 - \pi$. Then,

$$E(N_{m1}) = E \left[\sum_{i=1}^{n_t} I_{(R_i=0)} \right] \quad (13)$$

$$= N_t E[I_{(R=0)}]$$

$$= N_t (1 - \pi)^P$$

and

$$V(N_{m1}) = V \left[\sum_{i=1}^{n_t} I_{(R_i=0)} \right] \quad (14)$$

$$= N_t^2 V[I_{(R=0)}]$$

$$= N_t (1 - \pi)^P [1 - (1 - \pi)^P].$$

Proof 2.1, 2.2, 5.1 and 5.2 (N_{m3} and N_{m10}).

Given the definitions and assumptions in Section 2, it follows that N_{jj} , $j \in \{1, 2, \dots, p\}$, j, j' , are independently and identically distributed binomial random variables with parameters N_t and $1 - \pi$. Then,

$$E(N_{jj}) = N_t (1 - \pi), \quad (15)$$

$$V(N_{jj}) = N_t \pi (1 - \pi) \quad (16)$$

and

$$\text{Cov}(N_{jj}, N_{j'j'}) = 0. \quad (17)$$

Using distribution theory for the order statistics of discrete random variables (David, 1981, Equations 2.1.1, 2.1.2, 2.1.4), the cumulative distribution function of N_{m3} , the minimum N_{jj} is

$$F_{N_{m3}}(n_{jj}) = 1 - \left[1 - F_{N_{jj}}(n_{jj}) \right]^P \quad (18)$$

$$= 1 - \left[1 - \sum_{i=0}^{n_{jj}} \binom{N_t}{i} (1-\pi)^i \pi^{(N_t-i)} \right]^P$$

and the probability density function is

$$f_{N_{m3}}(n_{jj}) = F_{N_{m3}}(n_{jj}) - F_{N_{m3}}(n_{jj} - 1). \quad (19)$$

The cumulative distribution function of N_{m10} , the maximum N_{jj} is

$$F_{N_{m10}}(n_{jj}) = F_{N_{jj}}(n_{jj})^P \quad (20)$$

$$= \left[\sum_{i=0}^{n_{jj}} \binom{N_t}{i} (1-\pi)^i \pi^{(N_t-i)} \right]^P$$

and the probability density function is

$$f_{N_{m10}}(n_{jj}) = F_{N_{m10}}(n_{jj}) - F_{N_{m10}}(n_{jj} - 1). \quad (21)$$

Using the definition of the expected value and variance of a random variable, the results follow.

Proof 4.1 and 4.2 (N_{m9}).

Recall that N_{jj} , $j \in \{1, 2, \dots, p\}$, $j \neq j'$, are independently and identically distributed binomial random variables with parameters N_t and $1 - \pi$. Then,

$$E(N_{m9}) = E(\bar{N}_{jj}) \quad (22)$$

$$= E\left(\frac{1}{q} \sum_{j=1}^p N_{jj}\right)$$

$$= E(N_{jj})$$

and

$$V(N_{m9}) = V(\bar{N}_{jj}) \quad (23)$$

$$= V\left(\frac{1}{q} \sum_{j=1}^p N_{jj}\right)$$

$$= p^{-2} \sum_{j=1}^p [V(N_{jj})].$$

After substitution of $E(N_{jj})$ and $V(N_{jj})$, the results follow.

Proof 3.1 and 3.2 (N_{m3}).

We derive the expected value and variance of N_{m6} , the minimum number of non-missing pairs in three steps.

Step 1. Given the definitions and assumptions in Section 2,

$$E(N_{m6}) = E(\bar{N}_{jj'}) \quad (24)$$

$$= E\left(\frac{1}{q} \sum_{jj'=1}^{(p-1)p} N_{jj'}\right)$$

$$= E(N_{jj'})$$

and

$$\begin{aligned}
 V(N_{m6}) &= V(\bar{N}_{jj'}) \\
 &= V\left(\frac{1}{q} \sum_{jj'=12}^{(p-1)p} N_{jj'}\right) \\
 &= q^{-2} \sum_{\{j,j':j \neq j'\}} \sum V(N_{jj'}) \\
 &\quad + 2q^{-2} \left[\sum_{\{j,j',j'':j \neq j' \neq j''\}} \sum \text{Cov}(N_{jj'}, N_{jj''}) \right. \\
 &\quad \left. + \sum_{\{j,j',j'',j''':j \neq j' \neq j'' \neq j''' \}} \sum \text{Cov}(N_{jj'}, N_{j''j'''}) \right] \\
 &= q^{-2} [qV(N_{jj'}) + 2[C_2 - h] \text{Cov}(N_{jj'}, N_{jj''}) \\
 &\quad + 2h \text{Cov}(N_{jj'}, N_{j''j'''})].
 \end{aligned} \tag{25}$$

Step 2. The above formulae require substitution of the expected value, variance and covariance of the number of non-missing pairs, $N_{jj'}$. Using the theorems of conditional expectation, variance and covariance,

$$E(N_{jj'}) = E_{N_t(r)}\{E[N_{jj'}|N_t(r)]\}, \tag{26}$$

$$V(N_{jj'}) = E_{N_t(r)}\{V[N_{jj'}|N_t(r)]\} + V_{N_t(r)}\{E[N_{jj'}|N_t(r)]\} \tag{27}$$

and

$$\text{Cov}(N_{jj'}, N_{jj''}) = E_{N_t(r)}\{\text{Cov}[N_{jj'}|N_t(r), N_{jj''}|N_t(r)]\} \tag{28}$$

$$+ \text{Cov}_{N_t(r)}\{E[N_{jj'}|N_t(r)], E[N_{jj''}|N_t(r)]\}.$$

By the definition of the number of non-missing pairs, $N_{jj'}$,

$$E(N_{jj'}) = E_{N_t(r)}\left\{E\left[\sum_{r=0}^p N_{jj'}(r)|N_t(r)\right]\right\}, \tag{29}$$

$$V(N_{jj'}) = E_{N_t(r)} \left\{ V \left[\sum_{r=0}^p N_{jj'}(r) | N_t(r) \right] \right\} \quad (30)$$

$$+ V_{N_t(r)} \left\{ E \left[\sum_{r=0}^p N_{jj'}(r) | N_t(r) \right] \right\} \quad (31)$$

and

$$\text{Cov}(N_{jj'}, N_{jj''}) = E_{N_t(r)} \left\{ \text{Cov} \left[\sum_{r=0}^p N_{jj'}(r) | N_t(r), \sum_{r=0}^p N_{jj''}(r) | N_t(r) \right] \right\} \quad (32)$$

$$+ \text{Cov}_{N_t(r)} \left\{ E \left[\sum_{r=0}^p N_{jj'}(r) | N_t(r) \right], E \left[\sum_{r=0}^p N_{jj''}(r) | N_t(r) \right] \right\}.$$

Using the formulas for the expected value, variance and covariance of a sum of random variables,

$$E(N_{jj'}) = E_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'}(r) | N_t(r)] \right\}, \quad (33)$$

$$V(N_{jj'}) = E_{N_t(r)} \left\{ \sum_{r=0}^p V[N_{jj'}(r) | N_t(r)] \right\} \quad (34)$$

$$+ 2 \sum_{\{r, r': r \neq r'\}} \text{Cov}[N_{jj'}(r) | N_t(r), N_{jj'}(r') | N_t(r')]$$

$$+ V_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'}(r) | N_t(r)] \right\}$$

and

$$C(N_{jj'}, N_{jj''}) = E_{N_t(r)} \left\{ \sum_{r=0}^p \sum_{r'=0}^p C[N_{jj'}(r)|N_t(r), N_{jj''}(r)|N_t(r)] \right\} \quad 35$$

$$+ Cov_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'}(r)|N_t(r)], \sum_{r'=0}^p E[N_{jj''}(r)|N_t(r)] \right\}$$

$$= E_{N_t(r)} \left\{ \sum_{r=0}^p Cov[N_{jj'}(r)|N_t(r), N_{jj''}(r)|N_t(r)] \right.$$

$$\left. + 2 \sum_{\{r, r': r \neq r'\}} \sum_{\Sigma\Sigma} Cov[N_{jj'}(r)|N_t(r), N_{jj''}(r')|N_t(r')] \right\}$$

$$+ Cov_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'}(r)|N_t(r)], \sum_{r'=0}^p E[N_{jj''}(r)|N_t(r)] \right\}.$$

Since $D_i \perp D_{i'}$, and, thus, $P(r) \perp P(r')$,

$$\sum_{\{r, r': r \neq r'\}} \sum_{\Sigma\Sigma} Cov[N_{jj'}(r)|N_t(r), N_{jj''}(r')|N_t(r')] = 0, \quad 36$$

which gives,

$$V(N_{jj'}) = E_{N_t(r)} \left\{ \sum_{r=0}^p V[N_{jj'}(r)|N_t(r)] \right\} \quad (37)$$

$$+ V_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'}(r)|N_t(r)] \right\}$$

and

$$Cov(N_{jj'}, N_{jj''}) = E_{N_t(r)} \left\{ \sum_{r=0}^p Cov[N_{jj'}(r)|N_t(r), N_{jj''}(r)|N_t(r)] \right\} \quad (38)$$

$$+Cov_{N_t(r)} \left\{ \sum_{r=0}^p E[N_{jj'(r)}|N_t(r)], \sum_{r=0}^p E[N_{jj''(r)}|N_t(r)] \right\}.$$

Step 3. To evaluate the expected value, variance and covariance of $N_{jj'}$, we derive the

expected value, variance and covariance of $N_{jj'(r)}$ conditional on $N_t(r)$. Let

$b_u(r) \in \{0, 1, \dots, N_t(r)\}$ be the number of rows of $\mathbf{P}^*(r)$ that contain pattern u , where

$u \in \{1, 2, \dots, \theta(r)\}$ and $\sum_{u=1}^{\theta(r)} b_u(r) = N_t(r)$. Define a $[\theta(r) \times \theta(r)]$ matrix

$B_d(r) = \text{diag}[b_1(r), b_2(r), \dots, b_{\theta(r)}(r)]$. One can write the set of numbers $N_{jj'(r)}$ as elements in the $(1 \times q)$ matrix $N_{jj'(r)} \equiv \mathbf{1}'_{\theta(r)} B_d(r) \mathbf{P}^*(r)$.

Using the definitions in Section 2 it follows that $b_1(r), b_2(r), \dots, b_{\theta(r)}(r)$ are distributed as multinomial with parameters $N_t(r)$ and $\lambda(r) = [\lambda(r), \lambda(r), \dots, \lambda(r)]$, with $\lambda(r)$ a $[\theta(r) \times 1]$ matrix. Then,

$$E[b_u(r)|N_t(r)] = N_t(r)\lambda(r), \quad (39)$$

$$V[b_u(r)|N_t(r)] = N_t(r)\lambda_u(r) [1 - \lambda(r)] \quad (40)$$

and

$$C[b_u(r), b_{u'}(r)|N_t(r)] = -N_t(r)\lambda(r)^2. \quad (41)$$

Define the set $U_{jj'(r)} = \{u : \sum_{u \in U_{jj'(r)}} b_u(r) = N_{jj'(r)}\}$ containing $\kappa(r)$ elements. Recall that $N_{jj'(r)} \equiv \mathbf{1}'_{\theta(r)} B_d(r) \mathbf{P}^*(r)$. Completing the matrix multiplication for one element of the set $U_{jj'(r)}$, we have $N_{jj'(r)} = \sum_{u \in U_{jj'(r)}} b_u(r)$. Then,

$$E[N_{jj'(r)}|N_t(r)] = E \left[\sum_{u \in U_{jj'(r)}} b_u(r) | N_t(r) \right] \quad (42)$$

$$= \kappa(r) E[b_u(r)|N_t(r)]$$

$$= N_t(r) \kappa(r) \lambda(r),$$

$$V[N_{jj'}(r)|N_t(r)] = V\left[\sum_{u_{jj'}(r)} b_u(r)|N_t(r)\right] \quad (43)$$

$$= \kappa(r) V[b_u(r)|N_t(r)]$$

$$+ 2\binom{\kappa(r)}{2} \text{Cov}[b_u(r)|N_t(r)]$$

$$= \kappa(r) N_t(r) \lambda(r) [1 - \lambda(r)]$$

$$- \kappa(r) [\kappa(r) - 1] N_t(r) \lambda(r)^2$$

$$= N_t(r) \kappa(r) \lambda(r) [1 - \kappa(r) \lambda(r)]$$

and

$$\text{Cov}[N_{jj'}(r)|N_t(r), N_{jj''}(r)|N_t(r)] = \text{Cov}\left[\sum_{u_{jj'}(r)} b_u(r)|N_t(r), \sum_{u_{jj''}(r)} b_{u'}(r)|N_t(r)\right] \quad (44)$$

$$= \sum_{u_{jj'}(r)} \sum_{u_{jj''}(r)} \text{Cov}[b_u(r)|N_t(r), b_{u'}(r)|N_t(r)]$$

$$= \gamma(r) V[b_u(r)|N_t(r)] + [\kappa(r) - \gamma(r)]$$

$$\times \text{Cov}[b_u(r)|N_t(r), b_{u'}(r)|N_t(r)]$$

$$= \gamma(r) N_t(r) \lambda(r) [1 - \lambda(r)]$$

$$-N_t(r) \lambda(r)^2 [\kappa(r) - \lambda(r)].$$

The conditional covariance between the *siblings*, and $N_{jj''}$, is different than the conditional covariance between the *strangers*, $N_{jj'}$ and $N_{j'j''}$. We derived the conditional covariance for the *siblings* above. The conditional covariance for the *strangers* is a parallel proof.

After substitution,

$$E(N_{jj'}) = E \left[\sum_{r=0}^p N_t(r) \kappa(r) \lambda(r) \right] \quad (45)$$

$$= \sum_{r=0}^p \kappa(r) \lambda(r) E[N_t(r)],$$

$$V(N_{jj'}) = E \left\{ \sum_{r=0}^p N_t(r) \kappa(r) \lambda(r) [1 - \kappa(r) \lambda(r)] \right\} \quad (46)$$

$$+ V \left[\sum_{r=0}^p N_t(r) \kappa(r) \lambda(r) \right]$$

$$= \sum_{r=0}^p \kappa(r) \lambda(r) [1 - \kappa(r) \lambda(r)] E[N_t(r)]$$

$$+ \sum_{r=0}^p \kappa(r)^2 \lambda(r)^2 V[N_t(r)]$$

$$+ \sum_{\{r, r': r \neq r'\}} 2 \Sigma \kappa(r) \kappa(r') \lambda(r) \lambda(r') \text{Cov}[N_t(r), N_t(r')]$$

and

$$\text{Cov}(N_{jj'}, N_{jj''}) = E \left(\sum_{r=0}^p \left\{ \gamma(r) \lambda(r) N_t(r) [1 - \lambda(r)] \right. \right. \quad 47$$

$$\left. \left. - \lambda(r)^2 N_t(r) [\kappa(r) - \gamma(r)] \right\} \right)$$

$$+ \text{Cov} \left\{ \sum_{r=0}^p N_t(r) \kappa(r) \lambda(r), \sum_{r=0}^p N_t(r) \kappa(r) \lambda(r) \right\}$$

$$= \sum_{r=0}^p \left\{ \gamma(r) \lambda(r) [1 - \lambda(r)] E[N_t(r)] \right.$$

$$\left. - [\kappa(r) - \gamma(r)] \lambda(r)^2 E[N_t(r)] \right\}$$

$$+ \sum_{r=0}^p \kappa(r)^2 \lambda(r)^2 V[N_t(r)]$$

$$+ \sum_{\{r, r': r \neq r'\}} 2 \Sigma \Sigma \kappa(r) \kappa(r') \lambda(r) \lambda(r') \text{Cov}[N_t(r), N_t(r')].$$

Given the definitions and assumptions in Section 2, it follows that

$N_t(r=0), N_t(r=1), \dots, N_t(r=p)$ are distributed as multinomial with parameters N_t and

$\alpha(r) = [\alpha(r), \alpha(r), \dots, \alpha(r)]$ with $\alpha(r)$ a $[(p+1) \times 1]$ matrix. Then,

$$E[N_t(r)] = N_t \alpha(r), \quad (48)$$

$$V[N_t(r)] = N_t \alpha(r) [1 - \alpha(r)] \quad (49)$$

and

$$\text{Cov}[N_t(r), N_t(r')] = -N_t \alpha(r) \alpha(r'). \quad (50)$$

After substituting in the distributional results for $N_i(r)$, the result follows.

References

- Aittokallio T (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics* 11(2):253–264. doi:10.1093/bib/bbp059. [PubMed: 19965979]
- Alizadeh AA, Eisen MB, Davis ER, Ma C, Lossos IS, Rosenwald A, Boldrick JC, ..., Staudt LM (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769):503–511. doi:10.1038/35000501. [PubMed: 10676951]
- Barton CN, Cramer EC (1989). Hypothesis testing in multivariate linear models with randomly missing data. *Communications in Statistics - Simulation and Computation* 18(3):875–895.
- Casella G, Berger RL (2002). *Statistical Inference, Second Edition* Pacific Grove: Duxbury Press/Thomson Learning Inc.
- Catellier DJ, Muller KE (2000). Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine* 19(8):1101–1114. [PubMed: 10790683]
- David HA (1981). *Order Statistics, Second Edition*. New York: Wiley-Interscience.
- DeRisi JL, Iyer VR, Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(5338):680–686. [PubMed: 9381177]
- Fouladi RT, Shieh Y-Y (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics - Simulation and Computation* 33(3):807–824.
- Gurka MJ, Edwards LJ, Muller KE (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine* 30(22):2696–2707. [PubMed: 21751227]
- Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data, Second Edition*. New York: Wiley-Interscience.
- Muller KE, Fetterman BA (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software* Cary: SAS Institute.
- Muller KE, Stewart PW (2006). *Linear Model Theory: Univariate, Multivariate, and Mixed Models, First Edition*. New York: Wiley-Interscience.
- Ringham BM, Kreidler SM, Muller KE, Glueck DH (2016). Multivariate test power approximations for balanced linear mixed models in studies with missing data. *Statistics in Medicine* 35(17):2921–2937. [PubMed: 26603500]
- Ross SM (1984). *A First Course in Probability* New York: Macmillan.
- Serrano D (2008). *Error of Estimation and Sample Size in the Linear Mixed Model* Ann Arbor, Michigan: ProQuest, UMI Dissertation Publishing.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, ..., Futcher B (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9(12):3273–3297. [PubMed: 9843569]
- Tu XM, Zhang J, Kowalski J, Shults J, Feng C, Sun W, Tang W (2007). Power analyses for longitudinal study designs with missing data. *Statistics in Medicine* 26(15):2958–2981. [PubMed: 17154250]

	y_1	y_2	y_3	y_4		d_1	d_2	d_3	d_4	r_i
Y_1	.	.	y_{13}	y_{14}	D_1	0	0	1	1	2
Y_2	.	.	.	.	D_2	0	0	0	0	4
Y_3	y_{31}	y_{32}	y_{33}	y_{34}	D_3	1	1	1	1	0
Y_4	y_{41}	y_{42}	.	y_{44}	D_4	1	1	0	1	1
Y_5	.	y_{52}	y_{53}	y_{54}	D_5	0	1	1	1	1
Y_6	y_{61}	y_{62}	y_{63}	y_{64}	D_6	1	1	1	1	0
Y_7	y_{71}	y_{72}	y_{73}	y_{74}	D_7	1	1	1	1	0
Y_8	y_{81}	.	y_{83}	.	D_8	1	0	1	0	1
					N_{jj}	5	5	6	6	

Figure 1:

Example data matrix and data indicator matrix. In the data matrix (left), missing values are represented by ., while non-missing values are labeled as y_{ij} .

	d_1	d_2	d_3	d_4		p_1	p_2	p_3	p_4	p_5	p_6
D_1	0	0	1	1	P_1	0	0	0	0	0	1
D_2	0	0	0	0	P_2	0	0	0	0	0	0
D_3	1	1	1	1	P_3	1	1	1	1	1	1
D_4	1	1	0	1	P_4	1	0	1	0	1	0
D_5	0	1	1	1	P_5	0	0	0	1	1	1
D_6	1	1	1	1	P_6	1	1	1	1	1	1
D_7	1	1	1	1	P_7	1	1	1	1	1	1
D_8	1	0	1	0	P_8	0	1	0	0	0	0
N_{jj}	5	5	6	6	$N_{jj'}$	4	4	4	4	5	5

Figure 2:
 Example transformation from the data indicator matrix to the matrix of non-missing pairs.
 Here, $p_{13} = 0$, since $d_{11} d_{13} = 0.1 = 0$. All other entries are calculated in

Figure 3:
Rows of P with the same color were derived from rows of D with the same number of missing values, r .

	$\boldsymbol{p}^*_1(1)$	$\boldsymbol{p}^*_2(1)$	$\boldsymbol{p}^*_3(1)$	$\boldsymbol{p}^*_4(1)$	$\boldsymbol{p}^*_5(1)$	$\boldsymbol{p}^*_6(1)$
$\boldsymbol{P}^*_1(1)$	1	0	1	0	1	0
$\boldsymbol{P}^*_2(1)$	0	0	0	1	1	1
$\boldsymbol{P}^*_3(1)$	0	1	1	0	0	1
$\boldsymbol{P}^*_4(1)$	1	1	0	1	0	0

Figure 4:
Siblings and strangers. $\boldsymbol{P}^*(1)$ indicates all patterns of non-missing pairs that can occur when a row of $\boldsymbol{Y}$ has one missing value. Column $\boldsymbol{p}^*_1(1)$ has four *siblings* (grey) and one *stranger* (blue).

Table 1:

Missing data summary statistics.

Statistic	Function	Description
N_{m1}	$\sum_{i=1}^{N_t} I(r_i = 0)$	number of complete cases
N_{m2}	$\min(\{N_{jj'}\})$	minimum number of non-missing pairs
N_{m3}	$\min(\{N_{jj}\})$	minimum number of non-missing measures
N_{m6}	$\bar{N}_{jj'}$	mean number of non-missing pairs
N_{m9}	$\bar{N}_{jj}$	mean number of non-missing measures
N_{m10}	$\max(\{N_{jj}\})$	maximum number of non-missing measures
N_{m11}	N_t	planned number of independent sampling units

Table 2:

Exact expected values of the missing data summary statistics.

Result ID	Statistic ID	$E(N_m)$
Result 1.1	N_{m1}	$N_t(1 - \pi)^p$
Result 2.1	N_{m3}	$\sum_{N_{jj}=0}^{N_t} N_{jj} \left\{ \left[1 - F_{N_{jj}}(N_{jj} - 1) \right]^p - \left[1 - F_{N_{jj}}(N_{jj} - 1) \right]^p \right\}$
Result 3.1	N_{m6}	$N_t \sum_{r=0}^p \alpha(r) \kappa(r) \lambda(r)$
Result 4.1	N_{m9}	$N_t(1 - \pi)$
Result 5.1	N_{m10}	$\sum_{N_{jj}=0}^{N_t} N_{jj} \left[F_{N_{jj}}(N_{jj})^p - F_{N_{jj}}(N_{jj} - 1)^p \right]$
Result 6.1	N_{m11}	N_t

Table 3:

Exact variances of the missing data summary statistics.

Result ID	Statistic ID	$V(N_m)$
Result 1.2	N_{m1}	$N_t(1-\pi)^p[1-(1-\pi)^p]$
Result 2.2	N_{m3}	$\sum_{N_{jj}=0}^{N_t} N_{jj}^2 \left\{ \left[1 - F_{N_{jj}}(N_{jj}-1) \right]^p - \left[1 - F_{N_{jj}}(N_{jj}) \right]^p \right\}$ $- \left(\sum_{N_{jj}=0}^{N_t} N_{jj} \left\{ \left[1 - F_{N_{jj}}(N_{jj}-1) \right]^p - \left[1 - F_{N_{jj}}(N_{jj}) \right]^p \right\} \right)^2$
Result 3.2	N_{m6}	$q^{-2} \times \left\{ qV(N_{jj'}) + 2 \left[hCov(N_{jj'}, N_{jj''}) \right] \right.$ $\left. + (qC_2 - h)Cov(N_{jj'}, N_{jj''}) \right\}$
Result 4.2	N_{m9}	$N_t \pi(1-\pi)/p$
Result 5.2	N_{m10}	$\sum_{N_{jj}=0}^{N_t} N_{jj}^2 \left[F_{N_{jj}}(N_{jj})^p - F_{N_{jj}}(N_{jj}-1)^p \right]$ $- \left\{ \sum_{N_{jj}=0}^{N_t} N_{jj} \left[F_{N_{jj}}(N_{jj})^p - F_{N_{jj}}(N_{jj}-1)^p \right] \right\}^2$
Result 6.2	N_{m11}	0

Table 4:

Regression coefficients for the regression of $E(N_{m2})$ on $N_t/10$, p and π .

i	X_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	Intercept	62.7318676	1.39997758
2	$N_t/10$	-5.5156768	0.48554341
3	p	-1.6042196	0.19829228
4	$1 - \pi$	-147.7255861	2.42449211
5	$N_t^2/100$	-0.1324363	0.03086649
6	p^2	0.0640387	0.00512660
7	$(1 - \pi)^2$	87.3472243	1.05263141
8	$pN_t/10$	-0.4981166	0.05868018
9	$(1 - \pi)N_t/10$	14.8545994	0.56777048
10	$p(1 - \pi)$	1.2421550	0.23187504
11	$p(1 - \pi)N_t/10$	0.4137540	0.06862130
12	$p^2N_t^2/100$	0.0019218	0.00035756
13	$(1 - \pi)^2N_t^2/100$	0.1812043	0.04124270
14	$p^2(1 - \pi)^2$	-0.0423721	0.00685029
15	$p^2(1 - \pi)^2N_t^2/100$	-0.0013272	0.00047793