Randomization Tests with Missing Data

Rafe Donahue, PhD[*] Robert Small, PhD[†]

Abstract
We discuss randomization tests for use with missing data and show that failing to account for missing data produces underestimates of the variance of the test statistic.

Key Words: Randomization tests, missing data

1. Introduction

A researcher is planning to conduct a trial on ten experimental units, five in group A and five in group B. During collection of the data, two experiment units in each group are lost, resulting in six observed data and four unobserved (missing) data. The complete outcome data are presented below, where the M_i represent the planned-for, but not-obtained, data in the experiment.

Group A	Group B
1	4
2	5
3	6
M_1	M_2
M_3	M_4

The researcher decides to carry out a randomization test on the non-missing observations, examining all the $\binom{6}{3} = 20$ combinations of the ways the observed data could be assigned to the two groups.

Using the difference of the means (Group A minus Group B) as the summary statistic under consideration, the researcher notes that the observed mean difference of $2 - 5 = -3$ for the observed data is more extreme than any of the other 19 combinations of data available under the randomization test and computes the p-value as $1/20 = 0.05$.

A more careful examination of the problem, however, recognizes that the missing data are data nonetheless and should be considered in the collection of possible combinations. Under the null hypothesis of no difference between the treatments, *all* the data are just as likely to fall into either of the treatment groups and should thus be accounted as such.

In reality, there are $\binom{10}{5} = 252$ combinations that must be examined. These combinations include several more combinations that will yield a difference of the means of the (non-missing) data between the two groups of -3.

Computing the differences in the means of the non-missing data in each group under the 252 possible combinations shows that there are now 16 different ways to achieve the minimum value. For example, one such way is to place the values $1, 2, 3, 4, 5$ in Group A and the values $6, M_1, M_2, M_3, M_4$ in Group B. Thus, using *all* the data — both non-missing and missing — the randomization p-value becomes $16/252 = 0.0635$.

What is apparent then is that when one accounts for missing data, one's conclusion becomes less certain. This will be shown to be a consequence of an increase in the variance of the sampling distribution of the test statistic being used in the statistical test when one allows for "missing" to be an outcome.

[*] Biomimetic Therapeutics and Vanderbilt University School of Medicine.

[†] Sanofi-Pasteur.

2. Computing variance inflation to account for missing data

A general form for adjustment of the variance of the difference in the treatment group means can be attained as follows.

First, assume an experiment with $2n$ experimental units will be run, with n in group A and n in group B and further suppose that each experimental unit will generate a missing datum with probability p and an observed datum with probability $1-p$.

Let J_A and J_B be the number of missing data in groups A and B, respectively.

Let $X_{i,A}, i = 1, \ldots, n - j_A$ be the i.i.d. observed data from group A and $X_{i,B}, i = 1, \ldots, n - j_B$ be those from group B.

The distributions of both J_A and J_B, the numbers of missing data in the two groups, are binomial with parameters n and p.

Suppose futher that the observed data are independent and identically distributed random variables such that the expectation $E(X) := E(X_{i,T})$ and variance $V(X) := V(X_{i,T})$ both exist and are the same for both $T \in \{A, B\}$.

We seek to compute $E(D)$ and $V(D)$ where $D = \bar{X}_A - \bar{X}_B$ and the means $\bar{X}_A$ and $\bar{X}_B$ are computed using only the non-missing data in each group.

The expected value of the D can be shown to be zero, as each of two groups have the same expectation for the means of the nonmissing data. Computation of the variance is more tricky.

We start with the computation of the variance of D conditional on the numbers of missing data in the two groups J_A and J_B.

$$\begin{aligned} V(D|j_A, j_B) &= \frac{1}{n-j_A} V(X) + \frac{1}{n-j_B} V(X) \\ &= V(X) \left[\frac{1}{n-j_A} + \frac{1}{n-j_B} \right] \end{aligned}$$

Note that this is simply the variance of D ignoring the missing data. Also note the obvious issue if either j_A or j_B is equal to n, the number of experimental units in each group. Indeed, the form of the variance above implies that the variance is not defined if either $j_A = n$ or $j_B = n$, as (at least) one of the means cannot be computed. As such, the investigation cannot procede if one (or both) of the groups produces only missing data.

Computing the unconditional variance for D involves integrating (summing) out the j_A and j_B terms to yield

$$\begin{aligned} V(D) &= E\left(V(D \mid j_A, j_B)\right) + V\left(E(D \mid j_A, j_B)\right) \\ &= \sum_{j_A=0}^{n} \sum_{j_B=0}^{n} V(D \mid j_A, j_B) p(j_A, j_B) + 0 \\ &= \sum_{j_A=0}^{n} \sum_{j_B=0}^{n} V(X) \left[\frac{1}{n-j_A} + \frac{1}{n-j_B} \right] p(j_A, j_B) \\ &= V(X) \left[\sum_{j_A=0}^{n} \frac{1}{n-j_A} \binom{n}{j_A} p^{j_A} (1-p)^{n-j_A} + \sum_{j_B=0}^{n} \frac{1}{n-j_B} \binom{n}{j_B} p^{j_B} (1-p)^{n-j_B} \right] \\ &= V(X) \left[E\left(\frac{1}{n-J_A} \right) + E\left(\frac{1}{n-J_B} \right) \right]. \end{aligned}$$

The unconditional variance for D reveals an anomaly for which we must account. Recall that we supposed the number of missing data in a particular group were distributed as a binomial random variable with parameters n and p. This implies that there is a non-zero probability that the number of missing data in a group could be equal to n; that is, there is a possibility that all the data in one group might be missing. The consequence of such an event would be the inability of being able to compute the difference in the means of the observed data rendering the value of D missing and the value of the variance undefined. As such, we must make the further assumptions that there exists at least one observed datum in each group and that the total number of missing data does not exceed n, lest the problem degenerates into anarchy.

Conditioning instead on the total number of missing data fixed at $m := j_A + j_B < n$, we can compute the conditional variance as

$$\begin{aligned} V(D \mid m) &= E\left(V(D \mid m, j_A)\right) + V\left(E(D \mid m, j_A)\right) \\ &= \sum_{j_A=0}^{m} V(D \mid m, J_A = j_A) P(J_A = j_A) + 0 \\ &= \sum_{j_A=0}^{m} V(X) \left[\frac{1}{n-j_A} + \frac{1}{n-(m-j_A)}\right] P(J_A = j_A) \\ &= \sum_{j_A=0}^{m} V(X) \left[\frac{1}{n-j_A} + \frac{1}{n-(m-j_A)}\right] \frac{\binom{m}{j_A}\binom{2n-m}{n-j_A}}{\binom{2n}{n}} \\ &= \frac{V(X)}{\binom{2n}{n}} \sum_{j_A=0}^{m} \left[\frac{1}{n-j_A} + \frac{1}{n-(m-j_A)}\right] \binom{m}{j_A}\binom{2n-m}{n-j_A}. \end{aligned}$$

Again, notes from this specification of the variance of the sampling distribution of the difference in the means of the observed data are in order.

First, if $m = 0$, then we get the simple variance of the difference in the means, $V(D \mid m = 0) = \frac{2}{n} V(X)$, as would be anticipated if there are no missing data.

Second, the hypergeometric probabilities partition the $\binom{2n}{n}$ elements of the randomization distribution based on the number of missing data in (without loss of generality) the first group. Carrying out a randomization test without treating the missing data as valid data results in examining only a portion of the complete randomization distribution, that portion that results from fixing j_A at the value observed in the trial.

3. Returning to the example

We can compute the proper sampling distribution in the case of $n = 5$ and $m = 4$, the situation observed in our introductory example, and compare it to the situation that would come about if $n = 3$ and $m = 0$, a situation that would arise if one used only the non-missing data in the introductory example.

Using our formula for the variance of the difference of the observed mean in the two situations, one can show that

$$\begin{aligned} V(D \mid n = 5, m = 4) &= \frac{V(X)}{\binom{10}{5}} \sum_{j_A=0}^{4} \left[\frac{1}{5-j_A} + \frac{1}{1+j_A}\right] \binom{4}{j_A}\binom{6}{5-j_A} \\ &= \frac{461}{630} V(X), \end{aligned}$$

while

$$\begin{aligned} V(D \mid n=3, m=0) &= \frac{V(X)}{\binom{6}{3}} \sum_{j_A=0}^{0} \left[\frac{1}{3-j_A} + \frac{1}{3+j_A} \right] \binom{0}{j_A} \binom{6}{3-j_A} \\ &= \frac{2}{3} V(X) = \frac{420}{630} V(X). \end{aligned}$$

Thus, failing to account for the fact that the missing data are actually data underestimates the variance of the difference. To get the proper variance, one must scale the "naïve" varince by a factor of $\frac{461}{420} = 1.098$.

The following table presents the scale factors that show the appropriate scaling of the naïve variance to yield the proper missing-data-adjusted variance based on various planned sample sizes in each of the two groups (n) and the total number of missing data (m), assuming equal numbers of observed data in each group.

Number planned for each group (n)	Total number missing (m)	Variance inflation factor
3	2	1.13
4	2	1.05
5	2	1.03
5	4	1.10
6	2	1.02
6	4	1.05
7	2	1.01
7	4	1.03
7	6	1.07
8	2	1.01
8	4	1.02
8	6	1.05
12	2	1.00
12	4	1.01
12	6	1.02
12	8	1.02
12	10	1.03
20	2	1.00
20	4	1.00
20	8	1.01
20	10	1.01
20	12	1.01
20	16	1.02
20	18	1.02

4. Conclusions

Missing data in a planned, randomized experiment are still data. Failure to account for the missing data when performing a randomization test underestimates the variance of the test statistic and thus leads to overestimating the precision with which one makes inference based on the data.

For large studies, this underestimate of the variance is small but for small studies the underestimate can be substantial.