Missing Data in Experiments: Challenges and Solutions

Robin Gomila and Chelsey S. Clark
Princeton University

Abstract

Missing data is a common feature of experimental datasets. Standard methods used by psychology researchers to handle missingness rely on unrealistic assumptions, invalidate random assignment procedures, and bias estimates of effect sizes. We describe different classes of missing data typically encountered in experimental datasets, and we discuss how each of them impacts researchers' causal inferences. In this tutorial, we provide concrete guidelines for handling each class of missingness, focusing on 2 methods that make realistic assumptions: (a) inverse probability weighting (IPW) for mild instances of missingness, and (b) double sampling and bounds for severe instances of missingness. After reviewing the reasons why these methods increase the accuracy of researchers' estimates of effect sizes, we provide lines of R code that researchers may use in their own analyses.

Translational Abstract

Researchers rarely manage to collect every piece of information about each participant in their study. For instance, participants sometimes refuse to answer questions that they consider sensitive (e.g., income, political orientation, sexual practices) or quit the study before completing it. If ignored or handled inappropriately, this phenomenon referred to as "missingness" generally compromises researchers' ability to make causal inferences based on their experiments. Specifically, missingness biases researchers' estimates of the effect size of the treatment. In this tutorial, we review the different ways in which missingness impacts the results of experimental studies and provide researchers with concrete steps for addressing each type of missingness they may encounter. For mild cases of missingness, we recommend using a method called inverse probability weighting (IPW). For severe instances of missingness, we recommend that researchers recontact a sample of participants with missing values to fill the gaps. This method, which involves recollecting data, is called double sampling and bounds. For both methods, we provide lines of R code that researchers may use in their own analyses.

Keywords: missing data, attrition, experiment, inverse probability weighting, double sampling and bounds

Experimental datasets often involve some degree of data missingness. This issue arises when one or more variables from a dataset include missing values for some participants but not for others. The primary focus of this tutorial is to introduce psychology researchers to the issue of and corrective methods for *attrition*, defined as missingness in the dependent variable. Attrition is the most pervasive and critical type of missingness in psychology

Robin Gomila and Chelsey S. Clark, Department of Psychology, Princeton University.

Correspondence concerning this article should be addressed to Robin Gomila, Department of Psychology, Princeton University, Peretsman-Scully Hall, Princeton, NJ 08544. E-mail: rgomila@princeton.edu

studies. We also review missingness in pretreatment covariates (e.g., gender, race, income), which researchers use in their regression analyses to increase statistical power and the precision of their estimate of experimental treatment effects (Gerber & Green, 2012; Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017). The presence of missingness in pretreatment covariates is problematic but easily solved with simple imputation methods. In this tutorial, we describe one of these corrective methods. A final possible type of missingness is missingness in the treatment assignment variable. However, in the case of experimental studies, this possibility is ruled out by design because researchers can always know (at least in principle), who was randomly assigned to the treatment versus control conditions.

Missingness is pervasive in psychology studies because researchers rarely manage to gather all of the information that they need from everyone in their sample. First, participants may be unwilling to provide certain responses, which is often the case when questions are perceived to be sensitive. For instance, questions about participants' mental health, employment status, attitudes toward controversial topics, or sexual practices may induce anxiety and cause attrition. Second, missing data can result from participants dropping out of the study. Motivations to dropout may be boredom, having other priorities, no longer needing the pay-

ment, or simply deciding to use free time in a different way. Third, participants may not be able to complete the study. This may happen, for example, when participants move out of the area in which the study is being conducted. Finally, data missingness may be due to administrative errors, such as accidental deletion of some values in a dataset.

Missingness is a serious issue that psychologists need to address in their data analyses. Specifically, missing data compromise causal inferences by invalidating random assignment procedures (Zhou & Fishbach, 2016) and introducing bias into researchers' estimates of effect sizes (Gerber & Green, 2012). If not handled appropriately, missingness will turn a well-designed experiment into a correlational study.

Although powerful methods have been developed to account for missing data in experimental studies, psychologists rarely use them. Instead, psychologists typically ignore the presence of missing values entirely and simply conduct their analyses on the data that are not missing. In some cases, this practice is accompanied by comparisons of rates of missingness in different experimental conditions or demographic groups using statistical tests (e.g., t tests). Unfortunately, these widespread strategies are inadequate and rely on unrealistic assumptions. In this tutorial, we recommend that researchers use different methods to handle attrition in their data. Specifically, we review methods that make weaker, more realistic assumptions: inverse probability weighting (IPW; Gerber & Green, 2012) and double sampling and bounds (DSB; Coppock, Gerber, Green, & Kern, 2017). We do not discuss statistical methods that make stronger model assumptions such as multiple imputation (MI) or multiple overimputation (MO). Researchers interested in learning about these methods may consult a large literature on the topic (e.g., Blackwell, Honaker, & King, 2017; Enders, 2010; Graham, 2009, 2012).

Our objective is to provide researchers who encounter missing data in their experimental studies with concrete guidelines. In the presence of missingness, researchers need to carefully think about the possible reasons why data are missing for some participants but not for others. This will lead researchers to make a key assumption about the class of missingness in their data, and this assumption will determine the appropriate statistical or design-based procedure to handle attrition. To be clear, any assumptions made during this process are based on human judgment, and researchers should be ready to justify their decisions in their articles.

The remaining parts of this tutorial aim to help researchers go through these different steps. First, we review a framework to understand how missing data affect the results of experimental studies. Specifically, we introduce the *potential outcomes* framework (Holland, 1986; Neyman, 1923, 1938; Rubin, 1974, 1977) and define missingness as a potential outcome (Gerber & Green, 2012). This framework allows us to distinguish between different classes of missing data. The first class of missingness, called *missingness completely at random* (MCAR), is extremely unlikely in psychology studies. Yet, common practices inappropriately make the assumption that missing values are MCAR. The second and third classes of missingness, called *missingness completely at random conditional on observed variables* (MCAR|X) and *missingness not at random* (MNAR), are much more plausible. This tutorial focuses on methods targeting these more realistic cases. Specifically, we explain the theoretical underpinnings of IPW and

DSB, and we provide lines of R code that researchers may use as templates for their own analyses.

Potential Outcomes: A Framework to Understand Missingness in Experimental Datasets

Analytic Strategies for Experimental Designs

For concreteness, imagine that we conducted an experiment testing the causal effect of a treatment on a dependent variable (DV) at *Fictional School*, a middle school for sixth and seventh graders. For this hypothetical experiment, we recruited a sample of students, randomly assigned each student to a treatment or control condition, and collected the data. Our final dataset includes the DV and the three pretreatment covariates: race, gender, and grade. Table 1 displays a hypothetical dataset ($N = 8$) for this study.

We now consider two common analytic strategies to test the effect of the treatment on the DV in experimental studies: (a) simple regression analysis (equivalent to ANOVA); and (b) multiple regression analysis, in which we control for pretreatment covariates.

The simple regression model can be formally written:

$$Y_i = \beta_0 + \tau Z_i + \epsilon_i \tag{1}$$

in which i indexes participants in the sample, Y is the dependent variable, τ is the treatment effect, Z_i is a binary treatment assignment indicator returning 1 if a participant was assigned to the treatment condition and 0 if a participants was assigned to the control condition, and ϵ is an error term.

To run this analysis in R, we could write the following line of code:

```
lm(DV ~ treatment,
  data = data)
```

The multiple regression model can be formally written:

$$Y_i = \beta_0 + \tau Z_i + X_i\beta + \epsilon_i \tag{2}$$

in which i indexes participants in the sample, Y is the dependent variable, τ is the treatment effect, Z_i is a binary treatment assignment indicator returning 1 if a participant was assigned to the treatment condition and 0 if a participants was assigned to the control condition, X is a matrix of pretreatment covariates, β is a vector of covariate effects, and ϵ is an error term.

To run this analysis in R, we could write the following line of code:

Table 1
Illustration of the Fictional School Student Dataset Without Missingness

ID	Treatment	DV	Race	Gender	Grade
1	1	76	White	Male	Sixth grade
2	0	70	Black	Female	Seventh grade
3	0	68	White	Female	Sixth grade
4	0	59	White	Female	Sixth grade
5	0	76	Black	Male	Seventh grade
6	1	86	White	Female	Sixth grade
7	1	90	Black	Female	Sixth grade
8	1	84	White	Male	Seventh grade

Note. DV = dependent variable.

```
lm(DV ~ treatment + race + gender + grade,
  data = data)
```

If no data points are missing from the dataset, such as in the data displayed in Table 1, both analytic strategies yield unbiased estimates of the average treatment effect (ATE) at Fictional School. Note that the multiple regression analysis generally performs better because including pretreatment covariates such as gender, race, and grade into the regression model improves the precision of $\widehat{ATE}$, the estimated average treatment effect (Gerber & Green, 2012). This is true even when researchers do not specify the "right" underlying model linking the covariates to the dependent variable, or when covariates are measured with error (e.g., via imputation).

Potential Outcomes and Average Treatment Effects

Researchers conduct experiments to estimate the average causal effect of a treatment of interest (e.g., an intervention, a training) on a dependent variable (e.g., belonging, IQ), in a population (e.g., the students of Fictional School). To do so, individuals are randomly sampled from the population of interest and randomly assigned to one of two experimental conditions: treatment versus control. In essence, this procedure aims to answer a question that is difficult to directly test. Suppose that we could observe all individuals from the population of interest simultaneously in two parallel worlds that differ in one dimension: the presence versus absence of a treatment. What would be the average difference in the DV between these two worlds?

This question posits that each individual i has two *potential outcomes* for the dependent variable: an outcome $Y_i(0)$ in a world *without* the treatment and an outcome $Y_i(1)$ in an otherwise identical world *with* the treatment. Under this framework, the treatment has a causal effect τ_i for each individual i, which can be written:

$$\tau_i = Y_i(1) - Y_i(0) \tag{3}$$

The ATE across all individuals from a population of size N is equal to the average value of τ_i, which can be expressed:

$$ATE = \frac{1}{N}\sum_{i=1}^{n} \tau_i \tag{4}$$

We illustrate the concept of *potential outcomes* in Table 2, which displays hypothetical potential outcomes of Fictional School students. We observe, for instance, that the treatment has a causal effect of Size 4 on Student 1 whereas it has a causal effect

Table 2
Illustration of Potential Outcomes for Students From the Fictional School

ID	$Y_i(0)$	$Y_i(1)$	τ_i
1	72	76	4
2	70	76	6
3	68	68	0
4	59	61	3
5	76	77	1
6	90	86	−4
7	90	90	0
8	74	84	10

Note. DV = dependent variable.

of Size 0 on Student 7. The average treatment effect for the 8 students from Table 2 is equal to 2.5, that is, the sum of τ_i divided by eight (number of students).

In reality, the causal effect of a treatment τ_i for an individual i is impossible to measure because can never observe both potential outcomes $Y_i(0)$ and $Y_i(1)$ for the same individual. Instead, in the absence of missingness, we observe either $Y_i(0)$ or $Y_i(1)$ depending on which experimental condition z_i (treatment or control) individual i was assigned to. The *observed* potential outcome for each individual i can be written:

$$Y_i = Y_i(1)z_i + Y_i(0)(1 - z_i) \tag{5}$$

in which z_i takes the value 1 when individual i was assigned to the treatment condition and 0 when individual i was assigned to the control condition. Because $z_i \in (0, 1)$, Equation 5 implies that we observe $Y_i(1)$ for participants assigned to the treatment condition, and $Y_i(0)$ for participants assigned to the control condition.

Estimating the ATE With Experimental Designs

If experiments do not allow us to derive τ_i, how do experimental designs allow us to practically estimate the ATE?

To understand how, let's use Equation 4 to derive the ATE in terms of potential outcomes:

$$\begin{aligned} ATE &= \frac{1}{N}\sum_{i=1}^{n} \tau_i \\ &= \frac{1}{N}\sum_{i=1}^{n} (Y_i(1) - Y_i(0)) \\ &= \frac{1}{N}\sum_{i=1}^{n} Y_i(1) - \frac{1}{N}\sum_{i=1}^{n} Y_i(0) \\ &= \mu_{Y_i(1)} - \mu_{Y_i(0)} \end{aligned} \tag{6}$$

in which $\mu_{Y_i(1)}$ is the average of $Y_i(1)$ and $\mu_{Y_i(0)}$ is the average of $Y_i(0)$.

By randomly assigning individuals to experimental condition, we eliminate—in expectation—the presence of any systematic difference between those in the treatment group and those in the control group. This implies that under the conditions of no data missingness, experimental designs allow us to derive unbiased estimates of $\mu_{Y_i(1)}$ using the outcomes of those assignment to treatment and of $\mu_{Y_i(0)}$ using the outcomes of those assigned to control. Specifically, $\mu_{Y_i(1)}$ is estimated by averaging the observed values of $Y_i(1)$ and $\mu_{Y_i(0)}$ is estimated by averaging the observed values of $Y_i(0)$. In this sense, the estimated average treatment effect $\widehat{ATE}$ is best expressed using conditional expectations:

$$\begin{aligned} \widehat{ATE} &= E[\hat{Y}_i(1)\,|\,Z_i = 1] - E[\hat{Y}_i(0)\,|\,Z_i = 0] \\ &= \underbrace{E[\hat{Y}_i(1)]}_{\mu_{\widetilde{Y}_i(1)}} - \underbrace{E[\hat{Y}_i(0)]}_{\mu_{\widetilde{Y}_i(0)}} \end{aligned} \tag{7}$$

A Note on Bias

Importantly, *bias* does not refer to the difference between a single estimate of a quantity of interest (e.g., ATE, $\mu_{Y_i(0)}$, $\mu_{Y_i(0)}$) and the true value of this quantity of interest in the population. Instead, bias is a product of the process used to generate an estimate. Therefore, bias in $\widehat{ATE}$ can be thought of as the difference between estimates of the ATE on average across all possible

random assignments and the true population ATE. This implies that an estimate is unbiased if it is correct in *expectation*, that is, on average over all assignments. In the context of Equation 7, if the random assignment procedure was not corrupt, estimates of $\mu_{Y_i(0)}$ and $\mu_{Y_i(0)}$ are correct in expectation, and therefore, $\widehat{ATE}$ is unbiased.

In practice, the unbiased estimate of the ATE along with its standard error and corresponding p value are derived with one line of code using simple or multiple linear regression, as described in the code following Equations 1 and 2. Let's now explore how these analyses lead to bias in the presence of missing data.

Missingness Leads to Bias

Statistical Software Discard Individuals With Missing Values From the Analyses

When values are missing from experimental datasets, statistical software (e.g., R, Stata, SPSS) disregard entirely all participants displaying missing values in at least one of the variables included in the analysis, without warning. Table 3 displays a new version of the simulated Fictional School dataset, which includes missing values in the DV and pretreatment covariates, indicated with "NA." We will use Table 3 to illustrate how missingness impacts the results of experimental data analysis.

Missingness in pretreatment covariates. Missing values in pretreatment covariates affect multiple linear regression but not simple linear regression. Specifically, statistical software removes all the participants who have at least one missing value in one of the pretreatment covariates included in the model. For instance, in the Fictional School dataset, none of the students displayed in Table 3 would be taken into account in a multiple regression analysis that includes the pretreatment covariates race and gender in the model. Some students, such as Students 1 and 7, would be removed by the software because their gender is missing. Other students would be removed because their race (e.g., Students 3 and 6) or DV is missing (e.g., Student 2).

This implies that when researchers ignore the presence of missing values in covariates and use multiple linear regression, they introduce missingness in the dependent variable for their analysis. It follows that missingness in covariates, if not corrected for, generates attrition. The main distinction between these two forms of missing data is that missingness in covariates, contrary to attrition, is easy to correct. Researchers can (and should always!)

Table 3
Illustration of Missingness in the Fictional School Dataset

ID	Treatment	DV	Race	Gender	Grade
1	1	76	White	NA	Sixth grade
2	0	NA	Black	Female	Seventh grade
3	0	68	White	NA	Sixth grade
4	0	NA	NA	NA	Sixth grade
5	0	76	Black	NA	Seventh grade
6	1	86	NA	Female	Sixth grade
7	1	90	Black	NA	Sixth grade
8	1	NA	White	NA	Seventh grade

Note. DV = dependent variable.

use a simple strategy such as mean substitution to prevent the statistical software from excluding observations because of covariate missingness. This method consists of replacing the missing value of each covariate by the mean of that covariate. In this tutorial, we provide a simple code that does just that (see Scenario 2). Importantly, this substitution method (or comparable ones) does not introduce bias in $\widehat{ATE}$ and can be used to correct for any type of missingness in covariates. However, this method is never appropriate to correct for attrition.

Attrition. Missingness in the dependent variable Y_i impacts simple and multiple linear regression analyses in the exact same way. Statistical software discard any individuals whose dependent variable is missing, without warning, and conduct the analysis on only the remaining participants.

Correcting for attrition is critical but not as straightforward. To understand how attrition affects experimental results, we go back to the potential outcomes framework and define attrition as a potential outcome (Gerber & Green, 2012, p. 215).

Attrition as a Potential Outcome

In an experiment, individuals assigned to an experimental condition z_i have two potential outcomes for attrition: their dependent variable is either reported or missing. Let $r_i(z)$ indicate the potential outcomes of individual i assigned to an experimental condition z, such that $z_i = 1$ if Individual i was assigned to the treatment condition and $z_i = 0$ if Individual i was assigned to the control condition. Let $r_i = 1$ when the dependent variable is reported and $r_i = 0$ when the dependent variable is missing. As a result, $r_i(0)$ indicates whether the dependent variable is reported for Individual i when Individual i was assigned to the control condition. Conversely, $r_i(1)$ indicates whether the dependent variable is reported for individual i when individual i was assigned to the treatment condition (see Table 4). The observed potential outcome r_i can be written:

$$r_i = r_i(0)(1 - z_i) + r_i(1)z_i \qquad (8)$$

As illustrated in Table 4, Equation 8 implies four possible types of participants with regard to attrition in an experiment (Gerber & Green, 2012, p. 228). Participants can be *always responders*, in which case we observe their DV independent of treatment assignment. Participants can be *never responders*, in which case they are missing independent of treatment assignment. Finally, some participants' potential outcomes for attrition may depend on treatment assignment: If being assigned to the treatment condition (but not control) causes them to be missing, we call them *if-untreated responders*, and if being assigned to the control condition (but not treatment) causes them to be missing, we call them *if-treated responders*. At this point, it is important to keep in mind that attrition can cause bias in the estimate of the population ATE even in the absence of if-treated responders and if-untreated responders. That is, attrition can lead to bias even when treatment assignment does not cause attrition. As explained in more details subsequently, this happens when always responders and never responders are impacted by the treatment differently. However, when treatment assignment does not cause attrition, we can get an unbiased estimate of the ATE for a specific—and often relevant subset of the population: *always responders* (Gerber & Green, 2012, pp. 224–

Table 4

Illustration of Potential Outcomes for Students From the Fictional School

$z = 0$	$z = 1$	Type of participant
$r_i(0) = 1$	$r_i(1) = 1$	always responder
$r_i(0) = 0$	$r_i(1) = 1$	if-treated responder
$r_i(0) = 1$	$r_i(1) = 0$	if-untreated responder
$r_i(0) = 0$	$r_i(1) = 0$	never responder

226). Note that the ATE for always responders may or may not be different from the population ATE.

When Is Attrition Innocuous?

The short answer: Attrition is *rarely* innocuous. It is difficult to think of any psychology studies in which researchers may safely assume that missingness was produced in a way that does not bias estimates of the population ATE. The presence of missing values does not bias researchers' causal inferences only if these values are MCAR.

MCAR is the strongest possible assumption that researchers can make about attrition in their dataset. This type of missingness is extremely unlikely and difficult to prove because it implies that attrition is unrelated to any variables that one can imagine, including variables that were not collected as part of the study (also known as *unobservables*). For instance, MCAR implies that missingness is independent of treatment assignment, as well as participants' mood, values, income, gender, race, political orientation, religious beliefs, sleeping patterns, hair color, visual acuity, and neighborhood. Put differently, if missing data are MCAR, each participant in the study has the exact same probability of missingness. This could be the case if some values were accidentally deleted by a computer program in a perfectly random way.

Formally, when missingness is MCAR, R_i is independent of both the treatment assignment Z_i and the potential outcomes for the dependent variable Y_i.[1] Looking back at Equation 7, this implies that in expectation, neither $\mu_{Y_i(0)}$ nor $\mu_{Y_i(1)}$ nor the difference between $\mu_{Y_i(0)}$ and $\mu_{Y_i(1)}$ is affected by missingness. A related implication of Equation 7 is that attrition caused by experimental conditions is innocuous as long as missingness is *completely random* within each condition. Specifically, if values are deleted completely randomly from each experimental condition separately—even at different rates–$\mu_{Y_i(0)}$ and $\mu_{Y_i(1)}$ will be correct in expectation, therefore the difference between $\mu_{Y_i(0)}$ and $\mu_{Y_i(1)}$ will also be correct in expectation.

To summarize, missingness does not lead to bias if:

$$Y_i(z) \perp\!\!\!\perp R_i(z) \tag{9}$$

When to Assume That Data Are Missing Completely at Random (MCAR)

We urge psychologists to not assume that values are MCAR. When participants decide to drop out of the study or to not answer some questions, which is the most common cause of missingness in psychology studies, it is impossible to demonstrate that missing values are MCAR because it would require proving the null for an infinite number of unknown unobservables. In the rare occasions

in which researchers have good reasons to believe that missingness was generated in a completely random way (e.g., by a computer program that has no knowledge of the participants characteristics or responses) and can prove it, missingness is considered *ignorable* (Little & Rubin, 1987) and limiting the analysis to complete cases does not lead to bias.

Statistical Analyses Cannot Justify the Relaxing of Assumptions About Attrition

Contrary to common misconceptions, statistical analyses comparing rates of missing values between different subgroups, such as men and women or treatment and control condition participants, are often uninformative because they cannot prove that missingness is MCAR. There is only one circumstance under which these analyses can be informative: if they reveal differential rates of attrition between subgroups. In this case, they confirm that missingness is *not* MCAR and researchers can use them to speculate about the possible reasons why values are missing in their data. However, these analyses are not informative in the opposite scenario: when researchers find no significant difference in rates of missingness between different subgroups or experimental conditions. In this case, researchers should not conclude that missingness is MCAR. First, researchers generally lack suitable measures to predict missingness in their sample. That is, they may have not measured enough variables to predict missingness. For example, researchers rarely have access to all the variables that could explain missingness, such as mood, caffeine intake, or number of hours of sleep in the last few nights. Second, researchers may lack statistical power to detect asymmetrical missingness. For instance, researchers may compare rates of missingness between males and females, Black participants and White participants, or liberals and conservatives and find nonsignificant differences. However, these nonsignificant results do not prove that missingness is symmetrical or that missingness is innocuous. In fact, these nonsignificant results may be due to low statistical power, especially in small samples.

How Does Nonrandom Attrition Lead to Bias?

We described that data are rarely missing completely at random (MCAR) because it is unlikely that attrition truly occurs completely randomly. Most of the time, some participants have a higher probability of being missing than other participants. Consider a hypothetical study on the impact of a new policy on workers' perception of gender discrimination in the workplace. In this setting, gender minorities may perceive some questions as more sensitive than members of the majority gender identity group, and as a result, be less likely to answer them. Or, think about a longitudinal study on the impact of a therapy on patients' well-being, patients who are doing particularly poorly may drop out of the study at higher rates than others.

Nonrandom attrition introduces bias when it is not independent of the potential outcomes for Y_i. This can be the case when attrition is caused by treatment assignment z_i, but there are other cases as

---

[1] As a convention, we use capital letters to describe random variables and lower-case letters to describe specific values that random variables may take.

well. Let's illustrate how nonrandom attrition biases the results of experimental studies using the Fictional School study.

Bias from nonrandom attrition caused by treatment assignment. Suppose that in this study, being assigned to treatment caused attrition in seventh graders and that seventh graders have a tendency for lower values of the dependent variable Y_i. This implies that the estimate of $\mu_{Y_i(1)}$ (Equation 7) will be biased upward, which leads to bias in $\widehat{ATE}$ (i.e., the difference: $\mu_{Y_i(1)} - \mu_{Y_i(0)}$; Equation 7). In sum, nonrandom attrition that is caused by treatment assignment produces bias, even when there is no heterogeneity in the treatment effect, that is, even when the treatment has the same effect size τ_{attrit} on the individuals with missing data.

Bias from nonrandom attrition independent of treatment assignment. Now suppose that in the Fictional School study, seventh graders still have a tendency for lower values of the dependent variable Y_i and are more likely to be missing. However, in this scenario, nonrandom attrition is not a function of treatment assignment, such that the population does not include any if-treated responders or if-untreated responders. This could happen if, for instance, a random group of seventh graders went to volunteer at the soup kitchen on data-collection day, without knowing their treatment assignment. If the treatment effect is the same for seventh graders as on average in the population (Fictional School in this case), $\mu_{Y_i(1)}$ and $\mu_{Y_i(0)}$ will both be biased upward, but the difference $\mu_{Y_i(1)} - \mu_{Y_i(0)}$ will remain unbiased for the ATE. On the contrary, anytime the treatment effect τ_{attrit} is different for the subgroup of participants who are missing, the estimate of the ATE will be biased. To understand why, imagine what would happen if the treatment effect on the participants who τ_{attrit} is 10 but the treatment effect on the rest of the population is one? In this case, missingness would create a downward bias in the estimated ATE for the population. However, because missingness is not caused by treatment assignment, the estimated ATE is unbiased for always responders (see Table 4). Researchers who can demonstrate that missingness is not caused by treatment assignment may decide to redefine their quantity of interest as the ATE on always responders (Gerber & Green, 2012, pp. 224–226).

Handling Different Forms of Attrition in Experimental Studies

In the presence of nonrandom missingness, researchers need to make an assumption about the type of missingness, which will guide the method that they will use to address it. Researchers may make one of two realistic assumptions. These assumptions depend on researchers' beliefs about the causes of attrition on the one hand, and the presence of key pretreatment covariates in the dataset on the other hand. Researchers may decide to assume that they have collected the variables that explain missingness in their sample (e.g., gender, ethnicity, mood), in which case missingness is assumed to be completely random *conditional on observed variables* (MCAR|X). On the contrary, researchers may believe that they do not have access to the covariates that explain missingness in their dataset, in which case they assume that missingness is *conditional on unobserved variables*, and missing values are said to be MNAR.

In the remaining parts of this tutorial, we review the details of these two assumptions, and provide concrete strategies to correct for missingness based on each of these assumptions.

Outcome Data Missing Completely at Random Conditional on Observed Covariates (MCAR|X)

If attrition is not completely random, it may be a function of *observed variables*, that is, pretreatment covariates that were collected by the researcher. Thus, we refer to missingness conditional on observed variables as MCAR|X.[2] If data in the Fictional School study are missing because a random sample of the students from a specific grade were sent to volunteer at the soup kitchen on the day of data collection, missing data are MCAR|grade. In this specific example, the remaining sample only includes always responders and never responders. A related but different scenario could be that assignment to the treatment condition makes seventh graders more likely to volunteer at the soup kitchen on data collection day. In this case, missing data are MCAR|grade as long as the seventh graders who volunteer are similar to other seventh graders on every other dimension. That is, there is no reason to believe that any observed or unobserved variables other than grade explain missingness. This implies that, in expectation, the students who are missing are the same as the students in the same grade who are not missing.

When treatment assignment causes missingness, it is more likely for missingness to be conditional on a set of covariates. For instance, being assigned to the treatment condition may cause Black male seventh graders to volunteer at the soup kitchen. Perhaps something about the treatment is particularly off-putting to this demographic, causing an otherwise random group of these students to seek other ways to spend their time. In this case, missingness is conditional on grade, race, and gender.

In this tutorial, we illustrate further MCAR|X with simulated data and we review IPW, a method that leverages nonmissing data from individuals that are similar to those who are missing to recover the average treatment effect.

Outcome Data Missing Not at Random (MNAR)

When attrition is dependent on *unobserved variables*, that is, variables that researchers do not have access to, data are considered MNAR. Imagine that instead of being randomly assigned to volunteer at the soup kitchen, we learn that the students who are missing *decided* to stay at home on data collection day. It may be the case that the students who are most insecure about tests or live further away from school did not show up. In this situation, we do not know what triggered missingness or we have not collected the covariates that explain missingness. As a result, we cannot leverage existing data to correct for attrition and the missing data mechanism is said to be *nonignorable* (Little & Rubin, 1987). When missing data are MNAR, correcting for bias due to attrition requires more effort. In this tutorial, we review a method suggested by Coppock, Gerber, Green, and Kern (2017), which combines double sampling and bounding procedures to account for attrition dependent on unknown variables.

---

[2] Traditionally, this type of missing data is said to be *missing at random* (MAR), a term that can be misleading because MAR missingness is, in fact, only conditionally random (Graham, 2009).

Solutions to Missingness in Experimental Studies

We are going to walk you through a series of imaginary scenarios to demonstrate how to handle different types of missingness in your dataset. Specifically, we will focus on the following issues: (a) fixing missing values in covariates, (b) implementing IPW; and (c) using the DSB method. To do so, we create a population dataset from scratch, draw a sample from this dataset, introduce different types and amounts of missing values in the sample data, and provide code for the relevant method.

We are going to work with the following imaginary setting: Consider a company interested in introducing a mandatory diversity training program for all of their 20,000 employees. Before introducing the program, they would like to test it on a random sample of 2,000 employees for effectiveness. Prior to the treatment, the company administers a survey to all employees, requesting their race, gender, and education level. For simplicity, imagine that all employees identify as either Black or White, female or male, and hold a highest degree either from college or graduate school. As part of this survey, the company also measured each employee's baseline views toward diversity. This variable, labeled `pretest`, is coded "high" for employees more likely to perceive diversity in a complex way and "low" for those more likely to perceive diversity in a simple way. Half of the employees in the sample are randomly assigned to sit through the full-day diversity training (treatment condition) and the other half are randomly assigned to sit through a full-day training on sustainability (control condition). Apart from content, the structure of these trainings is identical. At the end of the day, all employees in the sample are asked to complete a series of tasks measuring the effectiveness of the training. This constitutes the outcome measure of interest, that is, the DV.

Description of the Population Data

Using simulation,[3] we generated a complete dataset (i.e., without missing values) that includes demographic and pretest variables for all 20,000 employees of this hypothetical company. We then generated the DV by assigning potential outcome values to employees under the treatment and control conditions. The true average treatment effect of the program in the simulated population is:

$$ATE = 1.00$$

The objective of experimental studies is to collect data from a random sample of the population of interest and estimate the true ATE as precisely as possible. This is what we are going to do in the next parts of this tutorial. As displayed in Table 5, this population average treatment effect (ATE = 1.00) is not constant across subgroups of employees. In other words, the effect size of the `treatment` on the `DV` varies depending on employees' pretest scores, as well as their demographics. For instance, the program has, on average, an effect size of 0.64 on black employees, and 1.09 on white employees (see Table 5). As previously described, when causal effects are heterogeneous and groups attrit at different rates, we can obtain biased estimates of the ATE.

From now on, our objective is twofold. First, we aim to illustrate how different classes of missing data are generated, which will clarify how to make the correct assumption about missingness in

Table 5

Description of the Simulated Population Data

Demographic	N	Proportion	Treatment effect
Female	5,795	29.00%	0.92
Male	14,205	71.03%	1.04
Black	4,283	21.42%	0.65
White	15,717	78.59%	1.10
College	7,792	39.76%	1.00
Graduate school	12,048	60.24%	1.00
Pretest low	6,412	32.06%	0.53
Pretest high	13,588	67.94%	1.23

your data. Second, we provide concrete steps and lines of R code that will allow you to compute unbiased estimates of the population average treatment effect for each type of missingness.

To do so, we draw a random sample of 2,000 employees from the population data (N = 20,000), which we use in four different scenarios. The variable names and first few rows of this sample dataset, which we call `dat`, are displayed in Figure 1. In each scenario, we use this exact same sample of 2,000 employees, but we introduce different amounts and types of missing data. We then use the appropriate method to correct for missingness in R and estimate the ATE ($\widehat{ATE}$).

Scenario 1: No Missing Data

In Scenario 1, we work on a sample that has no missing data at all. This allows us to introduce our analysis strategy and retrieve an estimate of the ATE from an imaginary study that was fortunate enough to not involve any missingness.

In this hypothetical scenario, all of the employees from our random sample of 2,000 employees provided their demographic information and completed their posttraining survey. To estimate the ATE in the population of 20,000 employees based on this sample of 2,000 employees, we use one of the following two linear regression models:

```
Model 1:
lm(DV ~ treatment, data = dat)

Model 2:
lm(DV ~ treatment + race + gender +
education + pretest,
data = dat)
```

In the absence of missingness, both the simple linear regression estimator (Model 1) and the multiple linear regression estimator (Model 2) are unbiased for the average treatment effect. In the following scenarios, we will use Model 2 exclusively, which is often preferred in the experimental framework since controlling for covariates typically increases precision (Gerber & Green, 2012). In the files that we share on the Open Science Framework, we provide R codes for both models for all missingness scenarios.

---

[3] All of our R code files and simulated data can be found on the Open Science Framework: https://osf.io/9sva5/.

Figure 1. Dataset of the simulated sample. DV = dependent variable. See the online article for the color version of this figure.

In the context of Scenario 1, Model 2 provides, on average,[4] the following unbiased estimate of the average treatment effect:

$$\widehat{ATE} = 1.00$$

$$\widehat{SE} = 0.11$$

Scenario 2: Missing Covariates

The procedure for correcting for missing covariate data is always the same. It is simple, highly effective, and does not depend on the class of missingness involved. In the experimental framework, including covariates in the analysis serves one purpose: increasing the precision of your estimate of the effect of the treatment on the DV (the ATE). The objective of this procedure is to ensure that the statistical software not drop any subjects based on covariate missingness. To do so, we simply substitute the missing values of each covariate by the mean value of that covariate. This way, the dependent variable remains unchanged and we do not introduce bias in the estimate of the effect of the `treatment` on the `DV`.

Imagine that no values are missing in the `DV`, but that at least one value from `race`, `gender`, `education`, or `pretest` is missing for a total of 500 out of 2,000 employees in the sample. Because no covariates are included in the simple linear regression analysis (Model 1), this hypothetical case of missing covariates would only affect the multiple linear regression analysis (Model 2). Specifically, Model 2 would estimate the average treatment

effect based on only 1,500 observations (instead of 2,000 observations), which reduces the statistical power of the analysis. Furthermore, the subsample of complete observations—in this case, 1,500 observations—is usually nonrepresentative of the population. This is the case because some subgroups are often more likely than others to have missing values, both in the DV and covariates, and this nonrepresentative feature of the subsample usually biases the estimate of the ATE.

As an illustration, we used simulation to create 1,000 random samples of 2,000 employees, 500 of which have missing `race`, `gender`, or `education` data if they scored "low" (but not "high") at pretest. These simulations revealed that, on average, Model 2 yields the following biased estimate of the average treatment effect:

$$\widehat{ATE} = 1.16$$

$$\widehat{SE} = 0.14$$

---

[4] We used simulation to generate 1,000 random samples of 2,000 individuals from the entire population. For each of these 1,000 samples, we randomly assigned each individual to the treatment or control condition and estimate the ATE using Model 2. This procedure allowed us to derive the sampling distribution of the average treatment effect. In the body of this tutorial, we provide: (a) the average value of the sampling distribution of the average treatment effect, denoted by $\widehat{ATE}$, and (b) the empirical standard error, denoted by $\widehat{SE}$, which is the standard deviation of the sampling distribution of the average treatment effect.

Solution: Mean Substitution in Covariates

Correcting for bias in the ATE when covariate values are missing is straightforward and does not depend on the type of missingness at play. When values are missing from a covariate, simply replace all the missing values by the mean of the available values of that covariate. This method works under the assumption that covariate missingness is not a function of treatment assignment.

For instance, to replace missing values from the `race` variable by the mean of the `race` variable in R, we use the following code:

```
dat$race[is.na(dat$race)] <- mean(dat$race,
na .rm = T)
```

After using this procedure to replace missing values from each of the covariates that we use in Model 2, regressing the DV on the treatment variable, covariates, and pretest (Model 2) yields, on average,[5] the following unbiased estimate of the ATE:

$$\widehat{ATE} = 1.00$$

$$\widehat{SE} = 0.11$$

Scenario 3: Outcome Data Missing Completely at Random Conditional on Observed Covariates (MCAR|X)

Now let's imagine that 750 of the 2,0000 employees from the sample decided to drop out of the study. After taking a closer look at these employees, you realize that for all of them, without exception, the variable `pretest` returns high. If you are willing to assume that these participants who dropped out of the study constitute a *completely* random sample of the participants who scored high at `pretest`, you are ready to make the assumption that missingness in the DV is MCAR|X. To be precise, in this case, your missing outcome is MCAR|pretest. In other words, data are not missing completely at random in the whole dataset, but they are missing completely at random among the subset of participants who scored high on `pretest`.

Given that in the population, the size of the treatment effect is different for employees who scored high and low at `pretest` (see Table 5), underrepresentation of one of these two subsets of the population in the sample biases your inferences. Specifically, because the treatment effect is larger, on average, for employees who scored high at pretest (see Table 5), the estimated ATE from Model 2 should be biased downward. This is illustrated by the mean of the sampling distribution of the ATE, calculated using multiple linear regression for 1,000 simulated samples of 2,000 employees, prior to correcting for missingness. This procedure returns, on average, the following biased estimates, illustrated in Figure 2):

$$\widehat{ATE} = 0.86$$

$$\widehat{SE} = 0.12$$

Solution: Inverse Probability Weighting (IPW)

If outcome data is missing for some participants, and if we are willing to assume that one or more *observed* variables fully explain patterns of outcome missingness, we can use IPW to correct for bias (Gerber & Green, 2012; Seaman & White, 2013).
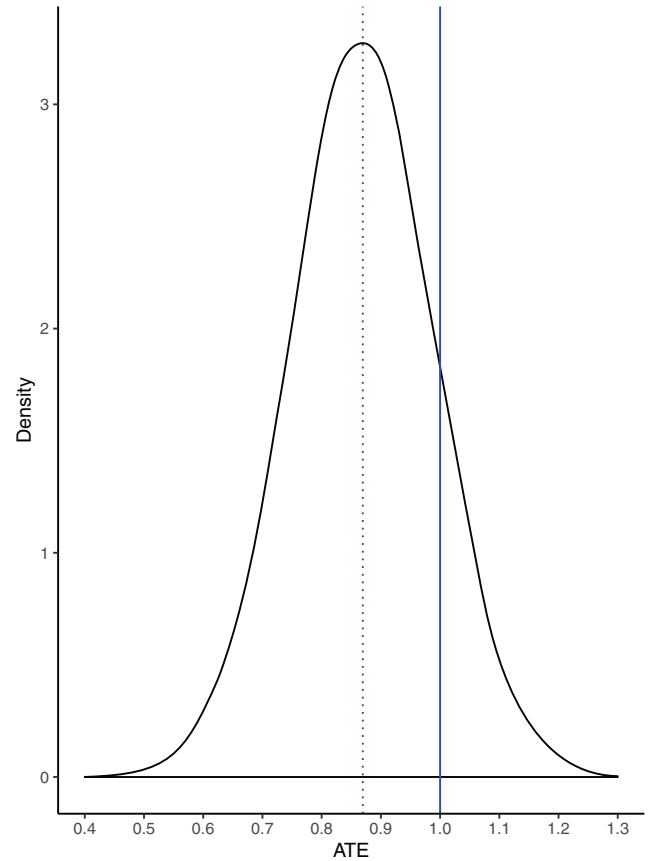

Figure 2. Sampling distribution of the average treatment effect (ATE) estimator. We used 10,000 randomizations to estimate the density of the sampling distribution. The solid line indicates the ATE in the population of interest. The dotted line represents the expected value of the ATE estimator, computed from the empirical distribution. The distance between solid and dotted lines represents bias. See the online article for the color version of this figure.

This approach, often used in medical research (Austin & Stuart, 2015; Hofler, Pfister, Lieb, & Wittchen, 2005) or longitudinal studies in the social sciences to account for drop-outs (e.g., Tankard & Paluck, 2017), assigns a weight to each value of the DV that is available in the data (i.e., nonmissing outcomes). Larger weights are assigned to observations that have larger probabilities of being missing, and smaller weights are assigned to observations that have lower chances of being missing.

For instance, in Scenario 3 described above, all 750 values that are missing from the DV come from an assumed random sample of employees with high `pretest` scores. Using IPW, we assign a

---

[5] We used simulation to create 1,000 random samples of 2,000 employees in which we generated missingness in `race`, `gender`, or `education` for 500 employees who scored "low" (but not "high") at pretest. For each of these samples, we used the mean substitution procedure to correct for missingness in covariates and estimated the average treatment effect using Model 2. This procedure allowed us to derive the sampling distribution of the average treatment effect and its standard error. The values displayed in the body of this tutorial, $\widehat{ATE}$ and $\widehat{SE}$, constitute the average value of the sampling distribution of the average treatment effect and its standard error.

greater weight to available observations from employees with high `pretest` scores. In this case, missingness can be fully explained by one observed variable (e.g., `pretest`), but in other cases, it might be explained by multiple observed variables (e.g., `pretest`, `gender`, and `education`). In the latter case, the weights will be calculated using all the variables that are required to fully explain missingness in the outcome variable.

Once the weights are calculated for each available observation, we run weighted multiple linear regression instead of regular multiple linear regression. Here are the details of the procedure to follow in `R` for missingness conditional on: (a) one variable and (b) multiple variables.

Step 1: Create a Response Dummy Variable

This variable returns 1 for available observations and 0 for missing outcomes. In `R`, use the following code as a template:

```
dat$response <- as.numeric(!is.na(dat$DV))
```

Step 2: Predict the Probability of Response of Each Employee in the Sample

Now that we created the `response` dummy variable, we can use the relevant covariates to predict the probability that each observation is nonmissing in the sample. Put differently, we can predict the probability that the `response` dummy variable returns 1.

We approach this prediction problem by using logistic regression. Specifically, we regress the `response` dummy variable on the relevant covariates interacted with the treatment indicator:

(a) For missingness conditional on one variable, such as `pretest`:

```
fit_p_resp <- glm(response ~ pretest*treatment,
        family = binomial(link = "logit"),
        data = dat)
```

(b) For missingness conditional on multiple variables, such as `pretest`, `race`, `gender`, and `education`:

```
fit_p_resp <- glm(response ~ pretest*treatment
  + race*
    treatment + gender*treatment +
      education*treatment,
        family = binomial(link = "logit"),
        data = dat)
```

Step 3: Probabilities of Response

Next, we use the logistic regression output to create a new variable `p_resp`, which returns each employee's probability of having a nonmissing outcome value:

```
p_resp <- fit_p_resp$fitted
```

Step 4: Generate the Weights

Now that we have created the `p_resp` variable, we can calculate the weights. To do so, we simply generate a variable `gen_weights`, which returns the inverse of the probability of response of each employee:

```
gen_weights <- 1/p_resp
```

Step 5: Weighted Linear Regression

We are now ready to use Model 2 in a weighted multiple linear regression:

```
fit_ipw <- lm(DV ~ treatment + race +
gender + education + pretest,
        weights = gen_weights,
        data = dat)
```

This weighted multiple linear regression allows us to recover the true ATE of the sample. That is, on average,[6] this procedure yields the following unbiased estimate of the population average treatment effect in both (a) and (b) cases (see Figure 3):

$$\widehat{ATE} = 1.00$$

$$\widehat{SE} = 0.14$$

Scenario 4: Outcome Data Missing Not at Random (MNAR)

Imagine that 300 participants drop out of the study posttreatment such that those with high `pretest` are more likely to drop out in treatment than control condition. At this point, because we have access to the `pretest` variable from the dataset, this situation seems to be MCAR|X. It is! Now, imagine that the company never introduced the `pretest` question in the survey. In that case, `pretest` is unobserved, which makes patterns of missingness MNAR.

We simulated such patterns of missingness in our sample of 2,000 employees, and removed the `pretest` variable from the dataset. In this new version of the dataset, that is, without information about the missingness generating process, we have no way to recover the source of missingness in the sample. The typical practice of comparing rates of missingness across subgroups is insufficient for diagnosing MNAR. Simply checking for significant differences in missingness among observed variables cannot rule out the possibility that missingness is being determined by a relevant unmeasured variable. Thus, confirming that your data missingness is not MCAR|X is not the end of your missing data problem-solving, but rather the beginning.

Because we lack information on the cause of missingness, we cannot use IPW to correct for bias. If we were to analyze this data as is, using Model 2, we would obtain, on average across a large number of samples, the following biased estimates (see Figure 4):

$$\widehat{ATE} = 0.56$$

$$\widehat{SE} = 0.12$$

---

[6] We used simulation to create 1,000 random samples of 2,000 employees in which we generated missingness in the DV for 750 employees who scored "high" (but not "low") at pretest. For each of these samples, we used IPW to correct for missingness and estimated the average treatment effect using Models 1 and 2. This procedure allowed us to derive the sampling distribution of the average treatment effect and its standard error. The values displayed in the body of this tutorial, $\widehat{ATE}$ and $\widehat{SE}$, constitute the average value of the sampling distribution of the average treatment effect and its standard error.
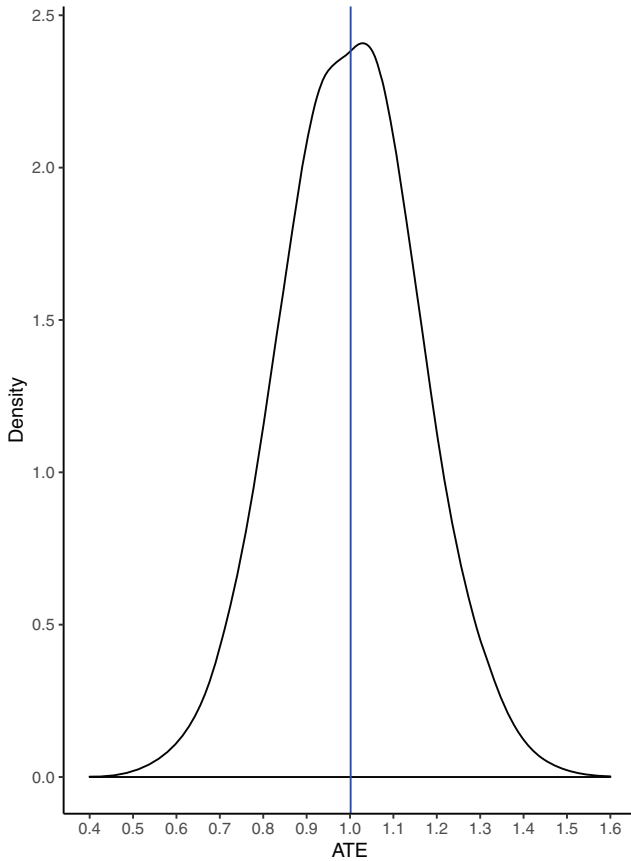
Figure 3. Sampling distribution of the average treatment effect (ATE) estimator. We used 10,000 randomizations to estimate the density of the sampling distribution. The solid line indicates the ATE in the population of interest, which is equal to the expected value of the ATE estimator computed from the empirical distribution. This figure illustrates that the corrected ATE estimator is unbiased. See the online article for the color version of this figure.

Solution: Combine Worst-Case Bounds and Double Sampling Methods

This method was developed by Coppock et al. (2017), who combined two approaches to estimate the ATE under MNAR with minimal assumptions.: worst-case bounds (Horowitz & Manski, 1998, 2000; Manski, 1990, 2009) and double sampling (Gerber & Green, 2012; Hansen & Hurwitz, 1946; Neyman, 1923).

Broadly speaking, worst-case bounds is a method that yields upper and lower bounds of the estimate of the ATE by replacing the missing values in the DV by the most extreme values that the DV could possibly take. This method usually produces bounds that are too wide to be informative, making it so punishing that researchers are reluctant to use it. In response, Coppock et al. (2017) proposed a strategy to make the procedure more realistic and informative by combining worst-case bounds with another method that involves additional data collection: double sampling.

The double sampling method requires researchers to obtain more data from the participants whose outcome values are missing. In the case of fewer missing values or more resources, the researcher may attempt to collect data from all of the participants

with missing data. Alternatively, in the case of large amounts of missing values or fewer resources, the researcher may attempt to collect data from a random sample of the participants with missing data. In either case, the goal of double sampling is to obtain data for the greatest proportion of participants with missing data.

This recruitment process can be achieved by simply offering these participants to take the study again, or by increasing incentives to participate. All in all, the objective is to gather additional knowledge about the outcome values of the participants who dropped out of the study. Importantly, the success of this procedure lies in the researchers' ability to achieve high rates of responses from the follow-up sample. Put differently, researchers need to limit missing data from the follow-up sample at all costs. In an ideal world, the follow-up sample would include responses from all or most of the participants with missing data. However, when resources are limited, researchers need to carefully think about the size of the follow-up sample and the incentive that they can offer to the participants to ensure that the largest possible proportion of response.

Keep in mind that by offering different incentives to these subjects than to those who initially reported their outcomes, you
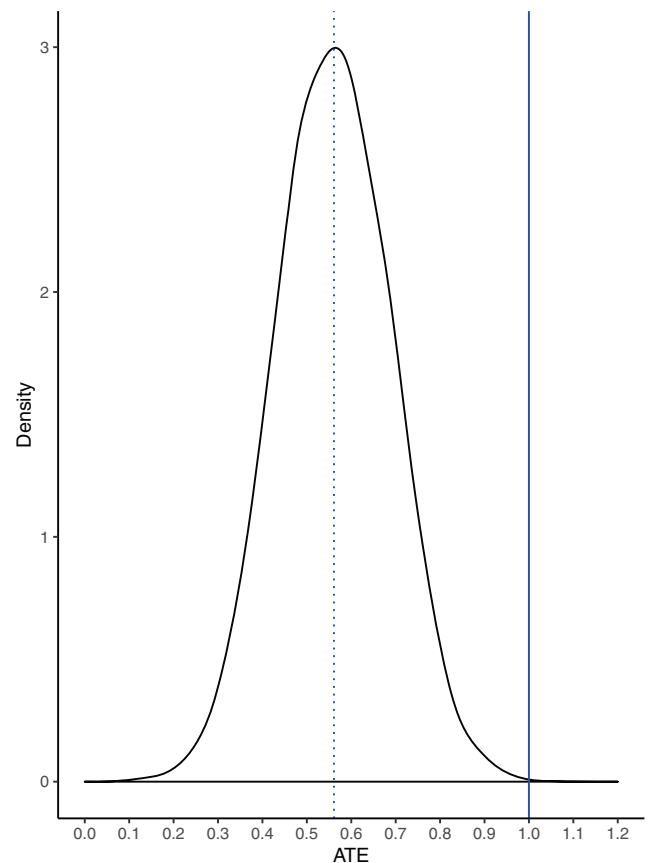

Figure 4. Sampling distribution of the average treatment effect (ATE) estimator. We used 10,000 randomizations to estimate the density of the sampling distribution. The solid line indicates the ATE in the population of interest. The dotted line represents the expected value of the ATE estimator, computed from the empirical distribution. The distance between solid and dotted lines represents bias. See the online article for the color version of this figure.

run the risk of making these samples incomparable. For example, the same person may respond differently to a question or situation when offered \$12 versus \$80 to participate, especially if the study measures outcomes such as generosity or cooperation. When using a double sampling procedure, make sure to consider the extent to which different incentives may impact the participants' responses. However, if researchers manage to collect responses from all of the participants with missing data with incentives that do not make Round 1 and Round 2 samples incomparable, the ATE for the recontacted sample would be point identified.

Assuming that through double random sampling you are able obtain outcome data for at least some participants whose outcomes were initially missing, you may proceed with the following code from the R package "attrition," developed by Coppock et al. (2017).

Step 1: Load the Attrition Package

If you have never installed the `attrition` package, run the two lines of code below to install the package "attrition," one line at a time.

```
Install.packages("devtools")
devtools::install_github("acoppock/attrition")
```

Now that the package is installed, you do not need to repeat the above installation. Load the package:

```
library(attrition)
```

Step 2: Create the `r1` Dummy Variable

Create a dummy variable `r1`, which returns 0 for participants whose data are missing, 1 otherwise. We use the following code:

```
dat$r1 <- as.numeric(!is.na(dat$DV))
```

Step 3: Recontact Participants for Whom `r1 = 0` and Create the Attempt Variable

Decide the proportion *P*(*attempt*) of participants to recontact from your `r1 = 0` sample. Based on this decision, randomly select the participants to recontact and offer them another opportunity to complete the study. Make sure to keep track of the participants who were recontacted by creating a dummy variable called `attempt`, which returns 1 if a participants was recontacted, 0 otherwise. In this hypothetical case, a total of 250 participants were recontacted (out of 300 initially missing). As mentioned above, the success of this procedure depends on your ability to convince a large proportion of the participants that you recontact to take the study. This implies that, when resources are limited, recontacting fewer participants would be most effective (e.g., if it allows you to use more resources to convince each of them to take the study).

Step 4: Create the `r2` Dummy Variable

Create a dummy variable `r2`, which returns 0 for participants whose data are still missing after you recontacted some (or all) participants who had missing outcome, 1 otherwise. In this hypothetical case, 10 out of the 250 participants that were recontacted are still missing at Round 2.

Step 5: Calculate the Lower and Upper Bounds of the Estimate of the Average Treatment Effect

Use the `estimator_ds` function from the `attrition` package to estimate the upper and lower bounds of the estimate of the ATE, as well as its confidence intervals:

```
est <- estimator_ds(Y = DV,
            Z = treatment,
            R1 = r1,
            Attempt = attempt,
            R2 = r2,
            minY = min(dat$DV, na.rm = T),
            maxY = max(dat$DV, na.rm = T),
            alpha = .05,
            data = dat)
```

After going through these different steps, Model 2 produces, on average,[7] the following lower and upper estimates of the true average treatment effect:

$$\widehat{ATE}_{lower} = 0.80$$

$$\widehat{ATE}_{upper} = 1.16$$

Although recontacting participants does represent a cost, this approach considerably reduces uncertainty around the average treatment effect compared to a traditional worst-case bounds analysis (Manski, 1990). For comparison purposes, using worst-case bounds on the Round 1 dataset (i.e., without recontacting missing participants) yields, on average, a lower bound of -3.77 and an upper bound of 4.88.

Conclusion

The presence of missing data in experimental studies has consequential implications for causal inference. Attrition invalidates random assignment of participants to the treatment versus control conditions and introduces bias in the estimate of the average treatment effect.

In this tutorial, we urge researchers to make realistic assumptions about missingness in their data, and we provide concrete guidelines for two methods that make more realistic and weaker assumptions to handle missingness in experimental datasets. After addressing all missingness in covariates included in their analysis using mean substitution, researchers may correct for missingness in the dependent variable with inverse probability weighting or double sampling and bounds. Inverse probability weighting is purely statistical, which makes it less costly: it can be implemented immediately by researchers after they finished collecting their data. Double sampling and bounds relies on a weaker assumptions, but requires that researchers collect addi-

---

[7] We used simulation to create 1,000 random samples of 2,000 employees in which we generated MNAR missingness for 300 employees in the DV. For each of these samples, we used the double sampling and bounds method to correct for missingness and estimated the lower and upper bounds average treatment effect. This procedure allowed us to derive the sampling distribution of the lower and upper bounds of average treatment effect. The values displayed in the body of this tutorial constitute the average value of the sampling distribution of the lower and upper bounds of average treatment effect.

tional data. Assumptions and decisions about which method to use is based on human judgment, and researchers should justify their choices in their articles. Finally, we strongly recommend that researchers refrain from assuming that missingness was generated completely randomly. This implies that researchers should not limit their analysis to the available data without correcting for missingness.

References

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34,* 3661–3679. http://dx.doi.org/10.1002/sim.6607

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research, 46,* 303–341. http://dx.doi.org/10.1177/0049124115585360

Coppock, A., Gerber, A. S., Green, D. P., & Kern, H. L. (2017). Combining double sampling and bounds to address nonignorable missing outcomes in randomized experiments. *Political Analysis, 25,* 188–206. http://dx.doi.org/10.1017/pan.2016.6

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation.* New York, NY: Norton.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). Simulations with missing data. *Missing data* (pp. 229–251). New York, NY: Springer.

Hansen, M. H., & Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association, 41,* 517–529. http://dx.doi.org/10.1080/01621459.1946.10501894

Hofler, M., Pfister, H., Lieb, R., & Wittchen, H.-U. (2005). The use of weights to account for non-response and drop-out. *Social Psychiatry and PsychiatricEpidemiology, 40,* 291–299. http://dx.doi.org/10.1007/s00127-005-0882-5

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960. http://dx.doi.org/10.1080/01621459.1986.10478354

Horowitz, J. L., & Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics, 84,* 37–58. http://dx.doi.org/10.1016/S0304-4076(97)00077-8

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95,* 77–84. http://dx.doi.org/10.2307/2669526

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York, NY: Wiley.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review, 80,* 319–323. Retrieved from https://www.jstor.org/stable/2006592

Manski, C. F. (2009). *Identification for prediction and decision.* Cambridge, MA: Harvard University Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles, Section 9. *Statistical Science, 5,* 465–472. http://dx.doi.org/10.1214/ss/1177012031

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association, 33,* 101–116. http://dx.doi.org/10.1080/01621459.1938.10503378

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701. http://dx.doi.org/10.1037/h0037350

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26. http://dx.doi.org/10.2307/1164933

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22,* 278–295. http://dx.doi.org/10.1177/0962280210395740

Tankard, M. E., & Paluck, E. L. (2017). The effect of a supreme court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science, 28,* 1334–1344. http://dx.doi.org/10.1177/0956797617709594

Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. *Journal of Experimental Social Psychology, 72,* 118–124. http://dx.doi.org/10.1016/j.jesp.2017.04.011

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111,* 493–504. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2016-29224-001&site=ehost-live