# Supplementary File 2: Mathematical proofs

First, we provide proof that the weights obtained with a traditional non-response analysis and Alt. 1 are equivalent in expectation if the invited individuals constitute a random sample of the target population. To re-weight the participant data so that it matches the distribution of *Z* in the population, we exploit the following expressions, which follow from Bayes’ theorem. Under the traditional non-response scenario, we use

|  |  |  |
| --- | --- | --- |
|  | , | (S.1) |

where *I* denotes invitation to participate. Random invitation implies that *Z* is independent of *I*,and therefore. As *S* implies *I*, we also have that, and . Thus, the expression in Eq. (S.1) is equivalent to

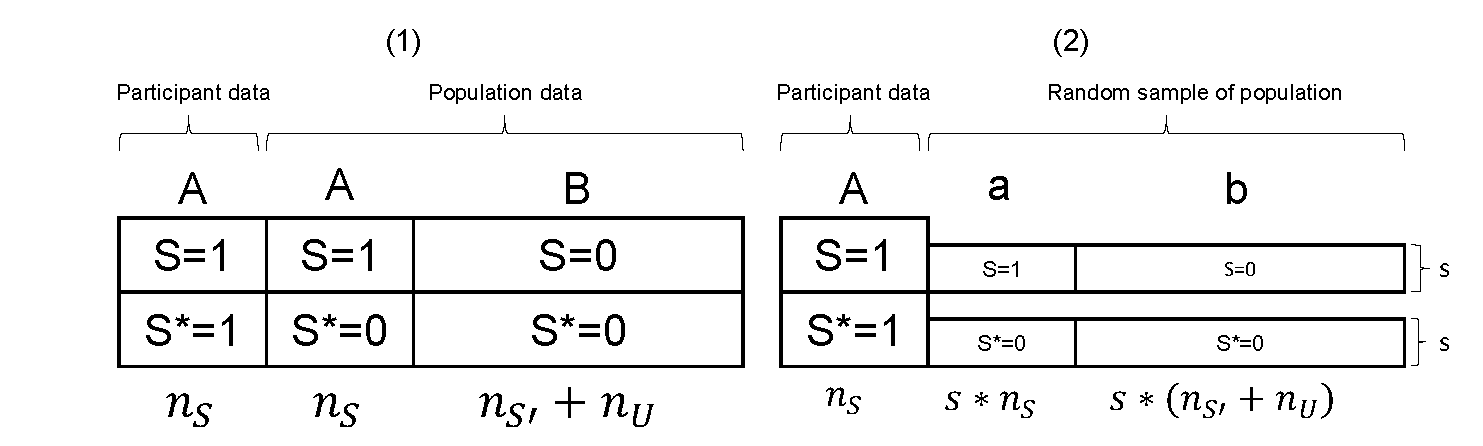
|  |  |  |
| --- | --- | --- |
|  |  | (S.2) |

where the last expression is what we obtain with Alt. 1.

Next, we provide proof that the odds-form weights suggested for Alt. 2A are equal to the probability-form weights for the other alternatives. First, note that , where is the number of cohort participants and *N* is the population size. Let *S\*=1* denote the known cohort members and *S\*=0* the population *including* duplicates of the cohort members under the data scenarios described in Alt. 2 (see main text). For the moment, let denote probabilities from an event space that includes duplicates of the participants. The probability we estimate is then instead , which may be approximately equal to the desired if , but otherwise not. However, it is not necessary to rely on this assumption, because transforming to gives the desired probability. We illustrate this point in the left plot of Figure S1, in which the top row shows the true participation status for the subpopulations A and B, where A represents the participants and B represents the rest of the population. We observe A twice, and falsely code the duplicates as non-participants. Together, these subpopulations sum to one in our data, i.e., . The bottom row then shows how these subpopulations are observed (and coded) in the data. Using this notation, we can express as , which is our target. However, we instead observe . To recover in this case, we re-write the expression as . Thus, we get

|  |  |  |
| --- | --- | --- |
|  |  | (S.3) |

where constitutes the generalization weight under Alt. 2A.



**Figure S1.** The event space and subpopulations when the analysis contains duplicates of the participants (S=1) under (1) and a version based on a random sample with probability s under (2). The top row shows the true underlying data, and the bottom row shows the data as observed by the investigator. The sample sizes (n) for each subpopulation are shown at the margins (S=Participants, S’=Non-participants, U=Uninvited).

These results extend to the use of random samples of the population under Alt. 2B. The right plot in Figure S1 shows a random sample of the population instead of the entire population in the boxes *a* and *c*. The sampling probability (for the random sample of the population) is . In this case, we have that A + a + b = 1, where a = s\*A and b = s\*B. We use the subscript *R* denote the event space under the random sampling version of Alt. 2. In this case, , i.e., the expression from before is multiplied by a scaling factor equal to *1/s*. However, this does not matter for the construction of the generalization weights, since the scaling factor occurs in both the numerator and denominator of the weight (and therefore cancels out). To see this, let Az/(Az+Bz) denote the propensity score for participation given covariate value(s) *Z=z*, where, e.g., Az denotes the share of individuals in the population data with a particular covariate value combination. We then get

|  |  |  |
| --- | --- | --- |
|  |  | (S.4) |

Hence, using random samples of the population as the comparison group under Alt. 2B, together with odds-form weights, will lead to the same result as using the entire population (in expectation; realized samples will suffer from sampling error that decreases with larger sample sizes).