Attempting to Prove that the Minimum Ratio of the Final Weight to the Selection Selection Weight is the Survey Response Rate

Jonathan Mellon

13/12/2021

In my [previous post](https://www.filedrawer.blog/post/minimum-post-strat-weight-is-response-rate/) I showed that the minimum inverse-probability weight in a simple random sample should be equal to the response rate of the survey. However, simple random samples are basically non-existent in the real world, so I need to work out the equivalent constraint for complex survey designs.

This post summarizes my progress so far towards doing that.

In the previous post I speculated that the constraint for ’s weight should be:

tl;dr that constraint is close but not exactly correct. I think I have now identified the correct constraint, but I haven’t yet proved it.

# Does it work in practice?

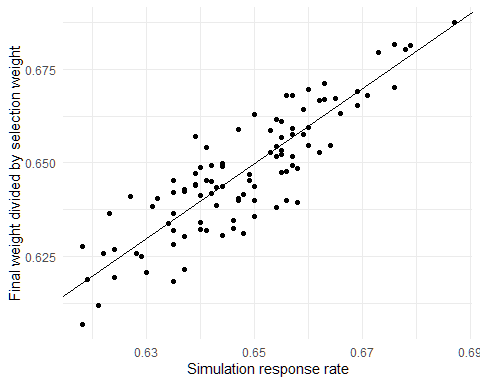
After an hour of failing to work out an analytic proof, I thought it would be worth doing a simulation to see whether the claim is at least empirically true.

So let’s do a simple two-stage sample where households are randomly selected, then a respondent is randomly selected from that household. This means that people in larger households have a lower probability of selection. If you live alone there’s a 100% chance of selection conditional on your household being selected, whereas if you live in a two-person household your selection probability is only 50%.

Conditional on being selected for the survey, rich people respond at 100% and poor people respond at 30%.

So after simulating that, I look at the response rate for the simulation (0.641) and the minimum value of the ratios of the final true weight to the selection weight (0.659). Not far off, but also not identical to each other. So, it looks like my initial guess about the constraint might be on the right track but not exactly right.

Let’s try this a few more times and see whether it’s robust.



After 100 simulations, it’s clear there’s a very strong relationship, but this isn’t an exact match.

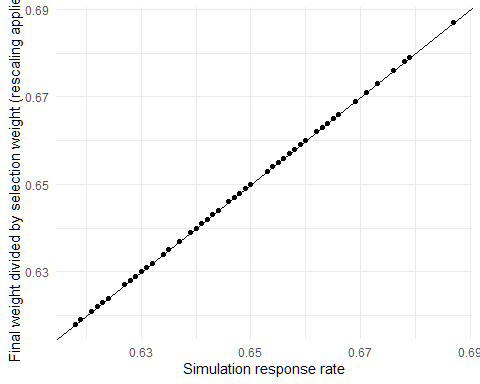
OK, so what’s going on with those ratios? The mean of the scaled inverse-probability weights is always exactly 1 in the simulation and the mean of the scaled selection weights is always 1 to 10 decimal places (probably a floating point problem).

However, the ratios themselves are not always this close to one. In fact, on average, they are 0.0103 away from 1. But should we actually expect these ratios to be near one on average? If I simulate two random vectors and rescale them to both have a mean of 1, the average ratio of them isn’t usually anywhere near 1.

set.seed(1298)  
A <- abs(rnorm(1000))  
B <- abs(rnorm(1000))  
A <- A / mean(A)  
B <- B / mean(B)  
mean(A/B)

## [1] 4.292947

OK, well what if I rescale the ratios so that they *actually* have a mean of 1 rather than just being close to that?

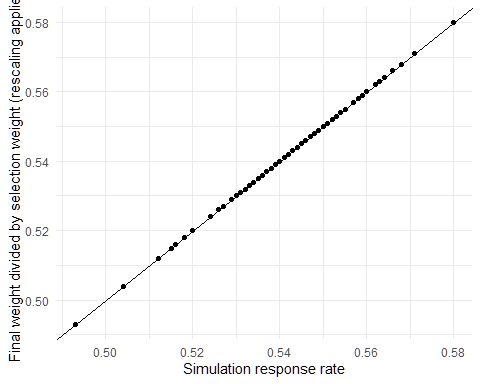


Apparently, that nails the minimum weight exactly. The mean absolute difference between this adjusted ratio and the response rate for the survey is just 0.0000000000000000643929354282591. At that point I’m willing to believe any remaining differences are just about floating point precision.

# Does this work with different simulation parameters?

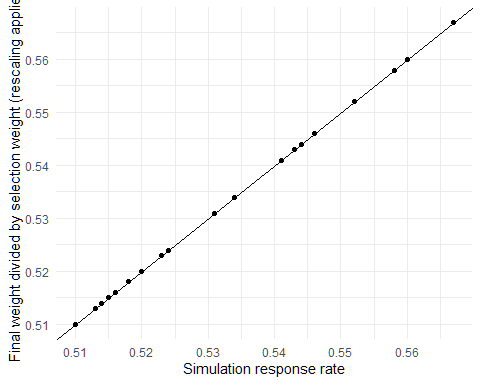
First of all, let’s make sure this actually holds up in a more complex simulation. I’ll set the poor response rate to 40% and say that people in bigger households are more likely to be rich (so that we have a correlation between selection probability and group response rates).

Apparently everything still works and the adjusted ratios exactly match the response rate.



What about if I make the population 10 times larger?

Apparently this still works too:



# Summarizing the claim

So to formalize, the constraint is as follows. For each respondent, , define:

The adjusted ratio is defined as follows, where is the mean ratio across all respondents to the survey:

Based on my simulations, it appears that:

# But why does this work?

🤷