Survey Weighting

Zach Markovich

March 18, 2022

Why Weight?

- Suppose there's a population: $Y_1 \ldots Y_N$

Why Weight?

- Suppose there's a population: $Y_1 \dots Y_N$
- So you sample a subset of them, uniformly at random, and ask their opinion

Why Weight?

- Suppose there's a population: $Y_1 \ldots Y_N$
- So you sample a subset of them, uniformly at random, and ask their opinion
- If D_i is a dummy variable indicating whether unit i was sampled, then the simple mean of your sample is an unbiased estimator for the true mean:

Why Weight?

- Suppose there's a population: $Y_1 \dots Y_N$
- So you sample a subset of them, uniformly at random, and ask their opinion
- If D_i is a dummy variable indicating whether unit i was sampled, then the simple mean of your sample is an unbiased estimator for the true mean:

$$\mathbb{E}\left(\sum_{i=1}^{N} \frac{D_i Y_i}{\sum_{i=1}^{N} D_i}\right) = \mathbb{E}\left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)$$

Why Weight?

- Suppose there's a population: $Y_1 \dots Y_N$
- So you sample a subset of them, uniformly at random, and ask their opinion
- If D_i is a dummy variable indicating whether unit i was sampled, then the simple mean of your sample is an unbiased estimator for the true mean:

$$\mathbb{E}\left(\sum_{i=1}^{N} \frac{D_i Y_i}{\sum_{i=1}^{N} D_i}\right) = \mathbb{E}\left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)$$

- Unfortunately, in the real world we don't get a true random sample

Why Weight?

- Suppose there's a population: $Y_1 \dots Y_N$
- So you sample a subset of them, uniformly at random, and ask their opinion
- If D_i is a dummy variable indicating whether unit i was sampled, then the simple mean of your sample is an unbiased estimator for the true mean:

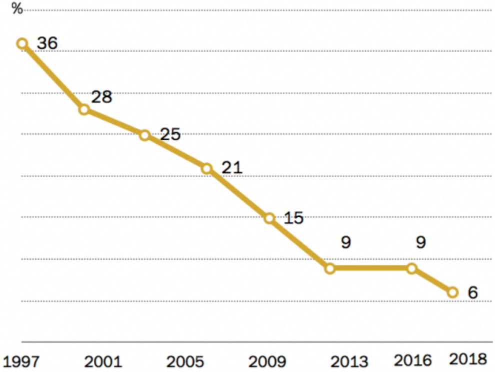
$$\mathbb{E}\left(\sum_{i=1}^{N} \frac{D_i Y_i}{\sum_{i=1}^{N} D_i}\right) = \mathbb{E}\left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)$$

- Unfortunately, in the real world we don't get a true random sample
- Specifically, if some variable influences both $P(D_i = 1)$ and Y_i, the simple mean will be a biased estimator

Declining response rates have intensified this problem

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

Quota Sampling

- Declining response rates have led many survey researchers to abandon random (or probability) sampling entirely

Quota Sampling

- Declining response rates have led many survey researchers to abandon random (or probability) sampling entirely
- Instead, they've turned to online panels – sometimes they will require the sample conform to some demographic quotas (quota sampling) – this can reduce the need for weights

Quota Sampling

- Declining response rates have led many survey researchers to abandon random (or probability) sampling entirely
- Instead, they've turned to online panels – sometimes they will require the sample conform to some demographic quotas (quota sampling) – this can reduce the need for weights
- Other times, you're just stuck with what there is (i.e. MTurk)

Quota Sampling

- Declining response rates have led many survey researchers to abandon random (or probability) sampling entirely
- Instead, they've turned to online panels – sometimes they will require the sample conform to some demographic quotas (quota sampling) – this can reduce the need for weights
- Other times, you're just stuck with what there is (i.e. MTurk)
- The same general idea applies – want to draw conclusions about a population that's different from our sample

Weighting to the rescue

In an ideal world, we'd know the true probability that a unit was sampled from the population: $P_i = P(D_i = 1)$. Then:

$$\mathbb{E}\left(\sum_{i=1}^{N} \frac{P_i^{-1} Y_i}{\sum_{i=1}^{N} P_i^{-1}}\right) = \mathbb{E}\left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)$$

Of course, in the real world, we don't actually know the weights, so instead we guess

Designing Survey Weights

- Observe a random sample drawn from the *sample frame* with known probabilities

Designing Survey Weights

- Observe a random sample drawn from the *sample frame* with known probabilities
- Wish to use the sample to draw inferences about a different population

Designing Survey Weights

- Observe a random sample drawn from the *sample frame* with known probabilities
- Wish to use the sample to draw inferences about a different population
- Specifically, will use covariates with a known frequency in the target population to produce estimates that will be representative of the target population

Post-Stratification

- Most common and straightforward approach to weighting

Post-Stratification

- Most common and straightforward approach to weighting
- Assumes that for some set of demographic covariates, $X_1 \ldots X_N$

$$w_i = \frac{P(X_i \text{ in Population})}{P(X_i \text{ in Sample Frame})} P(\text{i drawn from sample frame})$$

Note, often use value of $P(X_i \text{ in Sample Frame})$ estimated from the sample.

Post-Stratification

- Most common and straightforward approach to weighting
- Assumes that for some set of demographic covariates, $X_1 \ldots X_N$

$$w_i = \frac{P(X_i \text{ in Population})}{P(X_i \text{ in Sample Frame})} P(\text{i drawn from sample frame})$$

Note, often use value of $P(X_i \text{ in Sample Frame})$ estimated from the sample. Then estimate mean as,

Post-Stratification

- Most common and straightforward approach to weighting
- Assumes that for some set of demographic covariates, $X_1 \ldots X_N$

$$w_i = \frac{P(X_i \text{ in Population})}{P(X_i \text{ in Sample Frame})} P(\text{i drawn from sample frame})$$

Note, often use value of $P(X_i$ in Sample Frame) estimated from the sample. Then estimate mean as,

$$\frac{\sum_{i=1}^{N} w_i Y_i}{\sum_{i=1}^{N} w_i}$$

Multi-level Regression and Post-stratificiation (MrP)

- It is common to replace Y_i in the post-stratification estimator with the predictions from a hierarchical model

Multi-level Regression and Post-stratificiation (MrP)

- It is common to replace Y_i in the post-stratification estimator with the predictions from a hierarchical model
- The estimator then becomes:

$$\frac{\sum_{i=1}^{N} w_i \widehat{\mathbb{E}(Y_i|X_i)}}{\sum_{i=1}^{N} w_i}$$

Which simplifies to:

$$\sum_{x \in \mathcal{X}} P(X_i = x \text{ in population}) \widehat{\mathbb{E}(Y_i|X_i = x)}$$

Multi-level Regression and Post-stratificiation (MrP)

- It is common to replace Y_i in the post-stratification estimator with the predictions from a hierarchical model
- The estimator then becomes:

$$\frac{\sum_{i=1}^{N} w_i \widehat{\mathbb{E}\left(Y_i|X_i\right)}}{\sum_{i=1}^{N} w_i}$$

Which simplifies to:

$$\sum_{x \in \mathcal{X}} P(X_i = x \text{ in population}) \widehat{\mathbb{E}\left(Y_i|X_i = x\right)}$$

- There are two advantages of this:
  - Allows us to make predictions about $\widehat{\mathbb{E}\left(Y_i|X_i = x\right)}$ for empty cells
  - Bayesian methods regularize estimates for $\widehat{\mathbb{E}\left(Y_i|X_i = x\right)}$ for small cells

Raking

- Very often, we don't have knowledge of the joint frequency of all demographic traits

Raking

- Very often, we don't have knowledge of the joint frequency of all demographic traits
- Instead we might just have the marginal means

Raking

- Very often, we don't have knowledge of the joint frequency of all demographic traits
- Instead we might just have the marginal means
- *Raking* is a procedure to produce weights will match the population marginal means, while maintaining the sample joint distribution

Raking

- Very often, we don't have knowledge of the joint frequency of all demographic traits
- Instead we might just have the marginal means
- *Raking* is a procedure to produce weights will match the population marginal means, while maintaining the sample joint distribution
- There are many variations of the exact raking procedure, but they generally involve iteratively balancing the the marginal distribution of each variable

Weight Trimming

- Sometimes we'll end up with very extreme weights for rare combinations of covariates

Weight Trimming

- Sometimes we'll end up with very extreme weights for rare combinations of covariates
- In this case, such extreme weights may lead to an increase in the variance of their estimates that is out of step with the reduction in bias they facilitate

Weight Trimming

- Sometimes we'll end up with very extreme weights for rare combinations of covariates
- In this case, such extreme weights may lead to an increase in the variance of their estimates that is out of step with the reduction in bias they facilitate
- The only common solution here is to "trim" the weights by setting all weights greater than some value to a fixed level

Weight Trimming

- Sometimes we'll end up with very extreme weights for rare combinations of covariates
- In this case, such extreme weights may lead to an increase in the variance of their estimates that is out of step with the reduction in bias they facilitate
- The only common solution here is to "trim" the weights by setting all weights greater than some value to a fixed level
- This level is usually chosen in an ad hoc way (e.g. mak esure no weights are greater than a 10 or a 100)

Weight Trimming

- Sometimes we'll end up with very extreme weights for rare combinations of covariates
- In this case, such extreme weights may lead to an increase in the variance of their estimates that is out of step with the reduction in bias they facilitate
- The only common solution here is to "trim" the weights by setting all weights greater than some value to a fixed level
- This level is usually chosen in an ad hoc way (e.g. mak esure no weights are greater than a 10 or a 100)
- More principled techniques exist, but they're not frequently used

Go Over Code

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data
- Often, the population target is an estimated quantity as well:

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data
- Often, the population target is an estimated quantity as well:
  - Might have to interpolate quantities that are only observed at infrequent time intervals (e.g. deciennial census)

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data
- Often, the population target is an estimated quantity as well:
  - Might have to interpolate quantities that are only observed at infrequent time intervals (e.g. deciennial census)
  - Might impute joint distribution for some variable – e.g. ecological inference

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data
- Often, the population target is an estimated quantity as well:
  - Might have to interpolate quantities that are only observed at infrequent time intervals (e.g. deciennial census)
  - Might impute joint distribution for some variable – e.g. ecological inference
- Generally, we treat weights as known constants, but this isn't right if the weights are modeled. Bayesian or bootstrap procedures will be the best way to incorporate this into variance estimates, although this isn't often done in practice

Choosing the Target Population

- Sometimes we know the full target population – say we're working with census data
- Often, the population target is an estimated quantity as well:
  - Might have to interpolate quantities that are only observed at infrequent time intervals (e.g. deciennial census)
  - Might impute joint distribution for some variable – e.g. ecological inference
- Generally, we treat weights as known constants, but this isn't right if the weights are modeled. Bayesian or bootstrap procedures will be the best way to incorporate this into variance estimates, although this isn't often done in practice

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use
- This represents a bias-variance tradeoff – basing weights on more variables reduces bias, but leads to more extreme cells which inflate the variance of the estimate

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use
- This represents a bias-variance tradeoff – basing weights on more variables reduces bias, but leads to more extreme cells which inflate the variance of the estimate
- A variable that is important to weight with will be a strong predictor of both the *response* and the *sampling probability*

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use
- This represents a bias-variance tradeoff – basing weights on more variables reduces bias, but leads to more extreme cells which inflate the variance of the estimate
- A variable that is important to weight with will be a strong predictor of both the *response* and the *sampling probability*
- Weight with a variable that is only related to the sampling probability, but not the outcome will inflate the variance, but not lead to bias

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use
- This represents a bias-variance tradeoff – basing weights on more variables reduces bias, but leads to more extreme cells which inflate the variance of the estimate
- A variable that is important to weight with will be a strong predictor of both the *response* and the *sampling probability*
- Weight with a variable that is only related to the sampling probability, but not the outcome will inflate the variance, but not lead to bias
- There are formal methods for optimizing this choice (e.g Caughey and Hartman), but we won't get into them now

Choosing Variables to Weight With

- Often there are more available covariates than we can practically use
- This represents a bias-variance tradeoff – basing weights on more variables reduces bias, but leads to more extreme cells which inflate the variance of the estimate
- A variable that is important to weight with will be a strong predictor of both the *response* and the *sampling probability*
- Weight with a variable that is only related to the sampling probability, but not the outcome will inflate the variance, but not lead to bias
- There are formal methods for optimizing this choice (e.g Caughey and Hartman), but we won't get into them now

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population
- The two most common approaches to generating these weights are post-stratification and raking

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population
- The two most common approaches to generating these weights are post-stratification and raking
- Both techniques assume that the researcher knows some feature of the target population – this is not true in practice

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population
- The two most common approaches to generating these weights are post-stratification and raking
- Both techniques assume that the researcher knows some feature of the target population – this is not true in practice
- Researchers should focus weighting on variables that are likely to influence both the outcome and the sampling probability

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population
- The two most common approaches to generating these weights are post-stratification and raking
- Both techniques assume that the researcher knows some feature of the target population – this is not true in practice
- Researchers should focus weighting on variables that are likely to influence both the outcome and the sampling probability
- This is a very active area of methodological research – the number covariates available for weighting has increased while samples have become increasingly less representative

Conclusion

- Survey weighting is about designing weights so that estimates generated from the sample will conform to those generated for some target population
- The two most common approaches to generating these weights are post-stratification and raking
- Both techniques assume that the researcher knows some feature of the target population – this is not true in practice
- Researchers should focus weighting on variables that are likely to influence both the outcome and the sampling probability
- This is a very active area of methodological research – the number covariates available for weighting has increased while samples have become increasingly less representative
- Nonetheless, these basic principles will be sufficient for designing weights for most applied projects.