Sampling-based vs. Design-based Uncertainty in Regression Analysis

by Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge
June (2019)

James J. Heckman

THE UNIVERSITY OF CHICAGO

Econ 312, Spring 2021

1. Introduction

- Differences between sampling-based inference and design-based inference.
- Consider two simple examples.
- Table 1: finite population consisting of n units with each unit characterized by a pair of variables, Y_i and Z_i.

Table 1: Sampling-based Uncertainty (√ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	...
1	✓	✓	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	✓	✓	1	✓	✓	1	...
4	?	?	0	✓	✓	1	?	?	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	✓	1	?	?	0	?	?	0	...

- Consider an estimand that is function of the full set of pairs $\{(Y_i, Z_i)\}_{i=1}^{ni}$.

- Uncertainty about such an estimand arises when we observe the values (Y_i, Z_i) only for a sample, that is, for a subset of the population.

- In Table 1, inclusion of unit i in a sample is coded by the binary variable $R_i \in \{0, 1\}$.

- Sampling-based inference uses information about the process that determines the sampling indicators $R_1, \ldots, R_n$ to assess the variability of estimators across different samples.

- The second and third sets of columns in Table 1 depict such alternative samples.

- Table 2 depicts a different scenario in which we observe for each unit in the population the value of one of two potential outcome variables, either $Y_i^*(1)$ or $Y_i^*(0)$, but not both.

Table 2: Sampling-based Uncertainty (√ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	...
1	✓	?	1	✓	?	1	?	✓	0	...
2	?	✓	0	?	✓	0	?	✓	0	...
3	?	✓	0	✓	?	1	✓	?	1	...
4	?	✓	0	?	✓	0	✓	?	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	?	1	?	✓	0	?	✓	0	...

- The binary variable $X_i \in \{0, 1\}$ indicates which potential outcome we observe.
- Consider an estimand that is a function of the full set of triples $\{(Y_i^*(1), Y_i^*(0), X_i)\}_{i=1}^n$.
- As before, an estimator is a function of the observed data, the pairs (X_i, Y_i), for $i = 1, \ldots, n$, where $Y_i = Y_i^*(X_i)$ is the realized value.
- Design-based inference uses information about the process that determines the assignment $X_1, \ldots, X_n$ to assess the variability of estimators across different samples.
- The second and third sets of columns in Table 2 depicts such alternative samples.

- More generally, can have missing data processes that combine features of these two examples, with some units not included in the sample at all, and with some of the variables not observed for the sampled units.

- Articulating both the exact nature of the estimand of interest and the source of uncertainty that makes an estimator stochastic is a crucial first step to valid inference.

- For this purpose, it will be useful to distinguish.

- Descriptive estimands, where uncertainty stems solely from not observing all units in the population of interest.

- Causal estimands, where the uncertainty stems, at least partially, from unobservability of some of the potential outcomes.

2. A Simple Example

- A finite population of size n.
- We sample N units from this population, with $\{R_i \in \{0,1\}$ indicating whether a unit was sampled $(R_i = 1)$ or not $(R_i = 0)$ so that $N = \sum_{i=1}^{n} R_i$.
- There is a single binary regressor, $X_i \in \{0,1\}$, and n_x $(resp. N_x)$ is the number of units in the population (resp. the sample) with $X_i = x$.

- We view the regressor X_i not as a fixed attribute or characteristic of each unit, but as a cause or policy variable whose value could have been different from the observed value.

- This generates missing data of the type shown in Table 2, where only some of the states of the world are observed, implying that there is design-based uncertainty.

- For the RTC example, $Y_i^*(1)$ and $Y_i^*(0)$ could be state-level crime rates with and without RTC.

- Realized outcome:

$$Y_i = Y_i^*(X_i) = \begin{cases} Y_i^*(1) & \text{if } X_i = 1, \\ Y_i^*(0) & \text{if } X_i = 0, \end{cases}$$

which is the observed state-level crime rate in the RTC example.

- Potential outcomes are viewed as non-stochastic attributes for unit i, irrespective of the realized value of X_i.
- They, as well as the additional observed attributes, remain fixed in repeated sampling thought experiments, whereas R_i and X_i are stochastic.
- As a result, so are the realized outcomes in the sample, Y_i.
- Let $\boldsymbol{Y}, \boldsymbol{Y}^*(1), \boldsymbol{Y}^*(0), \boldsymbol{R}$, and $\boldsymbol{X}$ be the population *n*-vectors with *i*-th element equal to $Y_i, Y_i^*(1), Y_i^*(0), R_i$, and X_i respectively.
- For sampled units (units with $R_i = 1$) we observe X_i and Y_i. For all units we observe R_i.

- In general, estimands are functions of the full set of values $(\boldsymbol{Y}^*(1), \boldsymbol{Y}^*(0), \boldsymbol{X}, \boldsymbol{R})$ for all units in the population, both those in the sample and those not in the sample.

- We consider two types of estimands, **descriptive** and **causal**.

- If an estimand can be written as a function of $(\boldsymbol{Y}, \boldsymbol{X})$, free of dependence on $\boldsymbol{R}$ and on the potential outcomes beyond the realized outcome, we label it a descriptive estimand.

- Intuitively a descriptive estimand is an estimand whose value would be known with certainty if we observe the realized values of all variables for all units in the population.

- If an estimand cannot be written as a function of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{R})$ because it depends on the potential outcomes $\boldsymbol{Y}^*(1)$ and $\boldsymbol{Y}^*(0)$, then we label it a causal estimand.

- Consider three closely related estimands, one descriptive and two causal:

$$\theta^{\text{descr}} = \theta^{\text{descr}}(\boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{n_1}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - X_i)Y_i,$$

$$\theta^{\text{causal,sample}} = \theta^{\text{causal,sample}}(\boldsymbol{Y}^*(1), \boldsymbol{Y}^*(0), \boldsymbol{R}) = \frac{1}{N}\sum_{i=1}^{n} R_i \left(Y_i^*(1) - Y_i^*(0)\right),$$

and

$$\theta^{\text{causal}} = \theta^{\text{causal}}(\boldsymbol{Y}^*(1), \boldsymbol{Y}^*(0)) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^*(1) - Y_i^*(0)\right).$$

- Focus on the properties of a particular estimator:

$$\widehat{\theta} = \frac{1}{N_1}\sum_{i=1}^{n} R_i X_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n} R_i(1 - X_i)Y_i.$$

- This is the least squares estimator of the coefficient on X_i for the regression in the sample of Y_i on X_i and a constant.
- There are two sources of randomness in this estimator:
- A sampling component arising from the randomness of $\boldsymbol{R}$.
- A design component arising from the randomness of $\boldsymbol{X}$.
- Uncertainty generated by the randomness in the sampling component as sampling-based uncertainty.
- Uncertainty generated by the design component as design-based uncertainty.

Assumption 1. (RANDOM SAMPLING)

$$\Pr(\boldsymbol{R} = \boldsymbol{r}) = 1 \Big/ \binom{n}{N},$$

for all n*-vectors* $\boldsymbol{r}$ *with* $\sum_{i=1}^{n} r_i = N$.

Assumption 2. (RANDOM ASSIGNMENT)

$$\Pr(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{R}) = 1 \Big/ \binom{n}{n_1},$$

for all n*-vectors* $\boldsymbol{x}$ *with* $\sum_{i=1}^{n} X_i = n_1$.

- $N_1 \geq 1$ and $N_0 \geq 1$.

- Taking the expectation only over the random sampling, or taking the expectation only over the random assignment, or over both, we find:

$$E\left[\widehat{\theta} \mid \boldsymbol{X}, N_1, N_0\right] = \theta^{\text{descr}}, \tag{2.1}$$

$$E\left[\widehat{\theta} \mid \boldsymbol{R}, N_1, N_0\right] = \theta^{\text{causal,sample}}, \tag{2.2}$$

$$E\left[\widehat{\theta} \mid N_1, N_0\right] = E\left[\theta^{\text{descr}} \mid N_1, N_0\right] = E\left[\theta^{\text{causal,sample}} \mid N_1, N_0\right] = \theta^{\text{causal}}.$$

- Next, we look at the variance of the estimator, maintaining both the random assignment and random sampling assumption.

- Define the population variances

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^*(x) - \frac{1}{n}\sum_{j=1}^{n} Y_j^*(x)\right)^2, \qquad \text{for } x = 0, 1,$$

and

$$S_\theta^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^*(1) - Y_i^*(0) - \frac{1}{n}\sum_{j=1}^{n}(Y_j^*(1) - Y_j^*(0))\right)^2.$$

3. Design

- We define the "design variance" conditional on ***R***, so that only the design uncertainty is taken into account.

- To make the different variances interpretable, look at the expected value of the variances, taking the expectation both over the assignment and the sampling.

$$V^{\text{total}}(N_1, N_0, n_1, n_0) = \text{var}\big(\widehat{\theta} \,|\, N_1, N_0\big) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}, \tag{2.3}$$

$$V^{\text{sampling}}(N_1, N_0, n_1, n_0) = E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{X}, N_1, N_0\big) \,\middle|\, N_1, N_0\right] = \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right),$$

$$V^{\text{design}}(N_1, N_0, n_1, n_0) = E\left[\text{var}\big(\widehat{\theta} \,|\, \boldsymbol{R}, N_1, N_0\big) \,|\, N_1, N_0\right] = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}.$$

Comment 1. NEYMAN VARIANCE

The variance $V^{\text{total}}(N_1, N_0, n_1, n_0)$ is the one derived by Neyman (1990) for randomized experiments. □

Comment 2. CAUSAL VERSUS DESCRIPTIVE ESTIMANDS

In general the variances $V^{\text{sampling}}(N_1, N_0, n_1, n_0)$ and $V^{\text{design}}(N_1, N_0, n_1, n_0)$ cannot be ranked: the sampling variance can be very close to zero if the sampling rate $(N_0 + N_1)/(n_0 + n_1)$ is close to one, but it can also be larger than the design variance if the sampling rate is small and the variance of the treatment effect is substantial. □

Comment 3. Infinite Population Case

If $n_0, n_1 \to \infty$, the total variance and the sampling variance are equal:

$$\lim_{n_0,n_1,\to\infty} V^{\text{total}}(N_1, N_0, n_1, n_0) = \lim_{n_0,n_1,\to\infty} V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}.$$

In this case taking the design-based uncertainty into account does not matter. This result will be seen to carry over to more general cases in Section 3. □

Comment 4. Finite Population Correction

Whether the estimand is θ^{causal} or θ^{descr}, ignoring the fact that the population is finite generally leads to an overstatement of the variance on average because it ignores the fact that we observe a non-negligible share of the population:

$$V^{\text{total}}(N_1, N_0, \infty, \infty) - V^{\text{total}}(N_1, N_0, n_1, n_0) = \frac{S_\theta^2}{n_0 + n_1} \geq 0,$$

$$V^{\text{sampling}}(N_1, N_0, \infty, \infty) - V^{\text{sampling}}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \geq 0.$$

If the estimand is $\theta^{\text{causal,sample}}$, however, the population size is irrelevant because units in the population but not in the sample do not matter for the estimand:

$$V^{\text{design}}(N_1, N_0, \infty, \infty) = V^{\text{design}}(N_1, N_0, n_1, n_0).$$ □

- In this single binary regressor example the EHW variance estimator can be written as

$$\widehat{V}^{\text{ehw}} = \frac{N_1 - 1}{N_1^2}\widehat{S}_1^2 + \frac{N_0 - 1}{N_0^2}\widehat{S}_0^2, \quad \text{where} \quad \widehat{S}_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{n} R_i X_i \left(Y_i - \frac{1}{N_1}\sum_{i=1}^{n} R_i X_i Y_i\right)^2,$$

and $\hat{S}_0^2$ is defined analogously.

- Adjusting the degrees of freedom, using the modification proposed in MacKinnon and White (1985) specialized to this binary regressor example, we obtain $\tilde{V}^{ehw} = \frac{\hat{S}_1^2}{N_1} + \frac{\hat{S}_0^2}{N_0}$, which is identical to the variance estimator proposed by Neyman (1990), with the expectation of this modified EHW variance estimator $\tilde{V}^{ehw}$ (conditional on $N_0$and N_1) equal to the sampling variance in the infinite population case, $V^{sampling}(N_1, N_0, \infty, \infty)$.

- In the infinite population case the design-based uncertainty does not matter, so the EHW variance can be interpreted as implicitly taking into account design-uncertainty by focusing on the infinite population case.

- We could also estimate the variance using resampling methods, which would give us variance estimates close to $\hat{V}^{ehw}$.

- To be precise, suppose we use the bootstrap where we draw N_1 bootstrap observations from the N_1 treated units and N_0 bootstrap units from the N_0 control units.

- In that case the bootstrap variance would in expectation (over the bootstrap replications) be equal to $\hat{V}^{ehw}$.

Comment 6. CAN WE IMPROVE ON THE EHW VARIANCE ESTIMATOR?

- The difference between $E[\tilde{V}^{ehw}|N_1, N_0]$ (or the Neyman variance) and the total variance is equal to S_θ^2/n.

- The term S^2 is difficult to estimate because it depends on the unobserved differences $Y_i^*(1) - Y_i^*(0)$.

- As a result, S_θ^2/n is typically ignored in analyses of randomized experiments
- (see Imbens and Rubin, 2015).

- In particular, the EHW variance estimator implicitly sets the estimator of S_θ^2 to be equal to zero, resulting in conservative inference.

- For the case of a randomized experiment with a binary treatment Aronow et al. (2014) provide a lower bound for S_θ^2 based on the Frechet-Hoefffding inequality.

- In Section 3, we propose an improved variance estimator that exploits the presence of fixed attributes.