CSH Forum on p-Values

Sources of Uncertainty in Estimation Results

Prof. Dr. Aderonke Osikominu

Faculty of Business, Economics, and Social Sciences
University of Hohenheim

July 6, 2023

Motivation I

Why do we do inference (i.e., compute standard errors, test statistics, p-values, ...) in the first place?

- Incomplete information $\rightarrow$ uncertainty

Where does the uncertainty in estimation results come from?

- Observe only a (small) subset of the population $\rightarrow$ sampling-based uncertainty
- Could implement an intervention in many different ways $\rightarrow$ design-based uncertainty
- Do not know the true model for the outcome variable as a function of the explanatory variable(s), i.e., the conditional expectation function $\rightarrow$ model uncertainty

Sampling-Based Uncertainty I

The goal is to obtain information about the relationship between a set of variables, (Y, X, Z) say

Full information on the joint distribution $F(Y, X, Z)$ would require knowledge of the values of (Y, X, Z) of all members of the population

The population is large, potentially infinite, so that one can never know the values of (Y, X, Z) of all members of the population

The solution is to sample a subset of the population in a probabilistic way

The assumption of random sampling provides the mathematical foundation for applying the tools of statistics that allow us to infer features of the population distribution, $F(Y, X, Z)$, from a sample (generalizing inference)

Sampling-Based Uncertainty II

	Actual Sample			Alternative Sample I			Alternative Sample II			...
Unit	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	...

Notes: Sampling from a population with n units, '✓' denotes observed, '?' denotes not observed. Source: Abadie et al. (2020), Table 1.

- How does an estimator behave across samples, if one drew a large number of independent random samples?
- How does an estimator behave, if the sample size is increased so that eventually the full population is observed?

Design-Based Uncertainty I

Consider a data set with information on traffic accidents for all 400 districts in Germany $\rightarrow$ sample $=$ population, no natural superpopulation

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.

The binary variable $X_i \in \{0, 1\}$ indicates which potential outcome we observe.

Example

- $Y_i^*(1)$ the number of traffic accidents with a universal speed limit of 30km/h in place and $Y_i^*(0)$ the number of traffic accidents without
- $X_i = 1$ if a universal speed limit of 30km/h is in place and zero else
- Research question: What is the effect of a universal speed limit of 30km/h on traffic accidents?

Design-Based Uncertainty II

	Actual Sample			Alternative Sample I			Alternative Sample II			...
Unit	$Y_i^*(1)$	$Y_i^*(0)$	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	...

Notes: Sampling from a population with n units, '✓' denotes observed, '?' denotes not observed. Source: Abadie et al. (2020), Table 2.

Design-based inference uses information about the process that determines the assignments $X_1, \ldots, X_n$ to assess the variability of an estimator across different samples.

Design-based uncertainty can also arise in convenience samples

Concept generalizes to the multiple regression framework (Abadie, Athey, Imbens, and Wooldridge, 2020, *Econometrica*).

Sampling-based standard errors are in general too conservative

Model Uncertainty I

The question of how to select an appropriate model is often not treated formally in econometrics and empirical economics

Traditional view: model specification firmly guided by economic theory

Moreover, the consequences of data-driven model selection on the sampling properties of post-selection estimators are complicated

Model Uncertainty II

Leeb and Pötscher (2005, *Econometric Theory*) review important implications of some commonly used model selection procedures for inference

- Standard consistent model selection procedures, i.e., procedures that asymptotically select the correct model with probability approaching one, do not affect the asymptotic distribution of the post-model selection estimator
- However, in finite samples of any given size, the sampling properties of the post-model selection estimator are very different from the standard finite-sample or asymptotic distributions arising under the assumption of a fixed model
- The finite-sample distribution of a post-model selection estimator is not uniformly close to its asymptotic distribution because the probability of a model selection mistake depends on the unknown values of the population parameters

Model Uncertainty III

Literature on causal machine learning proposes a principled approach to deal with model uncertainty

Mitigating model selection mistakes is an important concern, but also regularization and overfitting bias

Belloni, Chernozhukov, and Hansen (2014, *Journal of Economic Perspectives*) and Chernozhukov et al. (2018, *Econometrics Journal*) propose a generic approach for causal inference on low-dimensional target parameters in the presence of high-dimensional nuisance parameters

The nuisance parameters can be estimated using a broad array of machine learning methods including regularized regression, boosted trees and random forests, neural nets, as well as ensembles of these methods.

Post-model selection estimators of target parameters have standard asymptotic distribution as if their was no prior model selection and estimation of nuisance parameters

Evidence on finite-sample performance limited

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis, *Econometrica*, 88, 265-296. https://doi.org/10.3982/ECTA12675.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects, *Journal of Economic Perspectives*, 28(2), 29-50.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters, *Econometrics Journal*, 21(1), C1-C68.

Leeb Hannes and Benedikt M. Pötscher (2005). Model Selection and Inference: Facts and Fiction, *Econometric Theory*, 21(1), 21-59.