Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration

by Sonya K. Sterba
Multivariate behavioral research 44.6 (2009): 711-740.

James J. Heckman

Econ 312, Spring 2021

1. Design-Based Inferential Framework

- Neyman and Pearson (1933) deemed the construction of hypothetical infinite populations, and construction of models, to be fallible and subjective.

- “A model is a set of invented assumptions regarding invented entities such that if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypothesis constituting the model are expected to agree with observations” – Neyman

- Did not want models to have a mediating role in the validity of inference

- Neyman developed alternative *design-based* inferential framework

- Target parameters of design-based framework are *finite* population parameters
- Step 1: Specify a random sampling frame, design, and scheme that generates y in the finite population
  - Sampling frame: List of primary sampling units in the finite population
  - Sampling design: Assigns nonzero probabilities of selection to each sample that could be drawn from the frame
  - Sampling scheme: draw-by-draw mechanism for implementing the sampling design

- Example: Estimating total numbers of drinking and driving episodes by high school students in a particular region
  - Stratify region on a geographical variable correlated with outcome (rural vs. urban) will create H=2 strata
  - Select $5=n_h$ clusters with unequal probabilities and with replacement separately in each strata
  - Want unequal probabilities (π_{ht} with h corresponding to stratum and I to cluster) to be proportional to a cluster-level covariate correlated with outcome (e.g. percentage of students qualifying for free lunch)

- Sampling frame would be a list of primary sampling units (schools) in the region along with each school's urban/rural location and percentage free lunch qualifiers

- At second state of selection, want to sample $m_{ht} = 20$ students (secondary sampling unit) from M_{ht} students in cluster i with equal probabilities

- This stratified clustered sampling design would assign selection probabilities $\pi_{ht} \times \frac{m_{ht}}{M_{ht}}$ to students in cluster i of stratum h.

- Various sampling schemes exist for implementing this design (Lohr, 1999, ch 6) which have been automated (SAS Proc Surveyselect, 2008)

- Step 2: Using only the known, nonzero probabilities of selection, cluster indicators, strata indicators, and observed *y*-values for sampled units – not a statistical model – a finite population parameter and its variance can be estimated.

- Example:
  - Calculate a sampling weight as the inverse of the first stage selection probability times the second state selection probability $\omega_{ht} = \frac{1}{\pi_{ht}} \times \frac{m_{ht}}{M_{ht}}$
  - The weight for a selected student indicates the number of students in the finite population that he or she represents
  - This weight contains all information needed to construct a point estimate for our finite population parameter:

$$\widehat{t} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hi} y_{hij}. \qquad (8)$$

- Disadvantage: by not specifying a model, *y*-values of sampled units in the finite population and *y*-values of unsampled units in the unsampled units in the finite population are not meaningfully related
    - None of these *y*-values in the finite population are meaningfully related to *y*-values outside the finite population
    - Only *descriptive inference* is possible with respect to the finite population parameters in the design-based framework

- “Do inferences made from nonrandom samples differ from those possible under random sampling?”
    - Different kinds of inference (descriptive rather than analytic) to different kinds of populations (finite rather than infinite) are possible exclusively under random sampling, and explicit models are not required to achieve these inferences

1.1 Implementation of the Design-Based Framework in Psychology

- Neyman's framework was utilized by observational survey researchers in epidemiology, sociology, health sciences, and government census and polling agencies
    - Target parameters often descriptive quantities
    - Often needed to produce thousands of estimates with little population knowledge
    - Thus hypothetical/infinite population models would be of questionable validity
- Was not adopted in observational psychology research
    - Less interest in enumerating particular finite populations
    - More interested in theory-driven models to explain causal mechanisms and predict future behavior
    - Thus preferred model-based framework over design-based

2. Limitations of the Pure Model- and Design-Based Frameworks

- Model-Based Framework Limitations:
  - Conditioning on all stratification and selection variables complicate model specification
  - Also complicates interpretation of model parameters and swallows needed degrees of freedom
  - Also error prone, especially if little was known about sample selection mechanism

- Design-Based Framework Limitations:
  - Limited by restrictions on types of parameters to be estimated and type of inference that can be obtained
  - Greatest advantage of inference free of all modeling assumptions is not necessarily true
  - No explicit attempts to write out a model for an infinite population, but sampling weight itself entails an implicit model relating probabilities of selection and outcome
  - Types of non-sampling errors requiring explicit models cannot be accommodated by design-based framework whatsoever

2.1 Integration of the Model-Based and Design-Based Frameworks

- Sampling statisticians viewed pure model-based framework as susceptible to bias from incomplete conditioning on sampling design
- Viewed design-based framework as incongruent with analytic statistics, causal inferences and certain non-sampling errors
- Limitations can be overcome by a hybrid, integrated framework

- Hybrid Framework Features:

  A. Can produce analytic statistics from complex random samples, adjusting for disproportionate selection, stratification, and clustering, without needing to condition on all of these complex sampling features during model specification

  B. It permits causal or descriptive inference about these analytic stasitics to infinite or finite population

  C. It is flexible enough to take into account measurement error

  D. Can accommodate situations in which researchers desire to condition on some complex sampling features but not others

- Key Developments:

    1. Account for the sampling design during model estimation not in model specification

    2. Make infinite and/or finite population inference

    3. Account for measurement error

    4. Account for the sampling design partially in model estimation, partially in model specification

2.2 Implementation of the Hybrid to Psychology

- Hybrid allows both kinds of inference (finite and infinite)
  - However, is applicable to *random* samples only
  - Given nonrandom sample, only choice is still pure model-based framework
  - Psychologists are analyzing complex random samples through electronically available public-use data sets that can use the hybrid framework
  - Software programs can also fit models under the hybrid framework

2.3 Illustrative Analysis with the Hybrid Model

- Example: Theoretical model from Raudenbush and Bryk (2002, chap. 4) and Singer (1998).

- This model stipulates that math achievement (MATHACH) varies across schools according to school average socioeconomic status (MEANSES), controlling for school SECTOR type (Catholic or public).

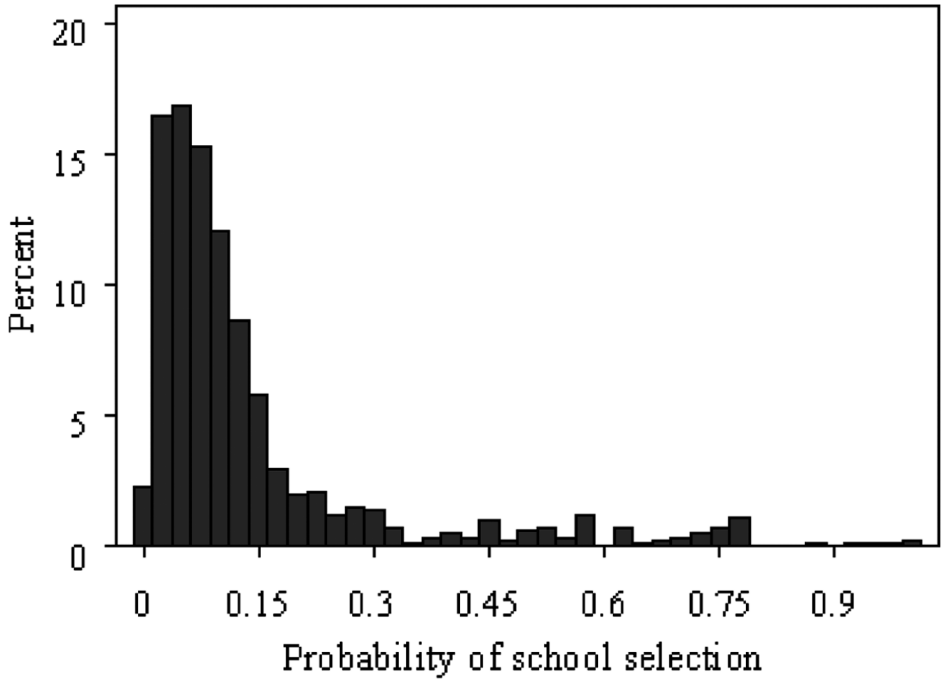
- This model also stipulates that the effect of school mean centered child socioeconomic status (CSES) on MATHACH varies across schools, but the strength of this relationship differs according to MEANSES

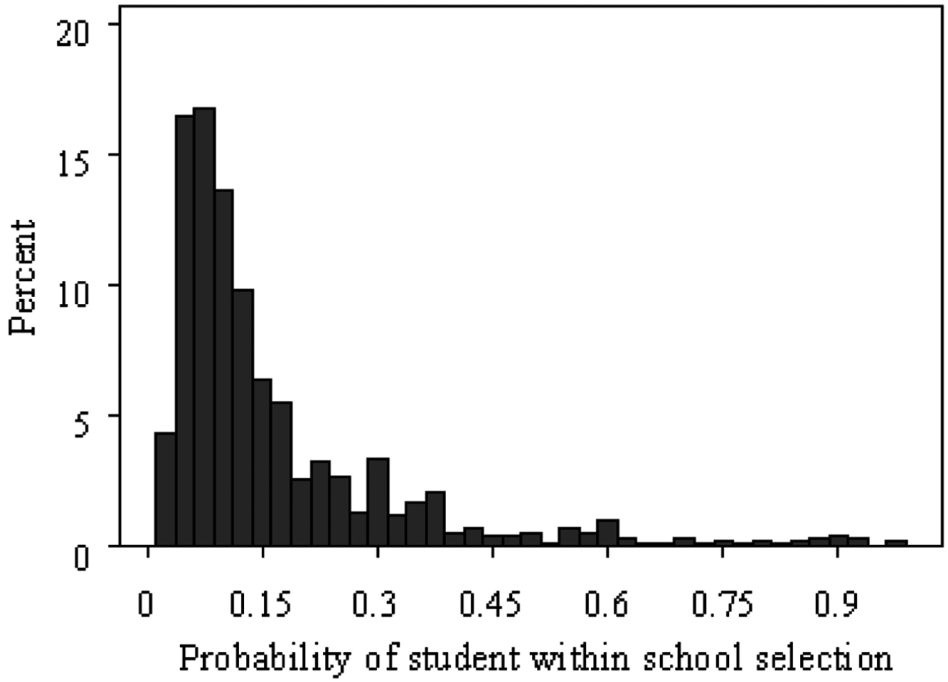
$$
\begin{aligned}
MATHACH_{ij} &= \beta_{0j} + \beta_{1j} CSES_{ij} + \varepsilon_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} MEANSES_j + \gamma_{02} SECTOR_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} MEANSES_j + \gamma_{12} SECTOR_j + u_{1j} \qquad (9) \\
&\text{where } \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)
\end{aligned}
$$

- This example uses the HSB (High School and Beyond) data set, whose sampling design includes clustering, stratification, and disproportionate selection
    - 26096 clusters (High schools)
    - 9 strata (based on school type: public, Catholic, private, and race composition)
    - Within some strata, schools were selected with probabilities proportional to estimated enrollment, but within other strata, schools were oversampled
    - 1,122 schools selected at primary stage

- At secondary stage, 36 seniors and 36 sophomores selected with equal probability at each selected school
- See Figure 3 for resultant variation in probabilities of selecting clusters and probabilities of selecting individuals within cluster
- Diagnostics showed that both sets of probabilities were significantly related to our outcome MATHACH after controlling for independent variables

Figure 1: Distributions of students' and schools' probabilities of selection in the High School and Beyond data set





- Model of Equation (9) had previously been fit to HSB data exclusively with model-based framework (Raudenbush & Bryk, 2002; Singer, 1998)
- Model specification does not account for HSB's disproportionate selection
- Partially accounts for HSB's stratification
- Fully accounts only for HSB's clustering
- Original, model-based analysis likely incurred bias due to incompletely conditioning on the sampling design.
- Hybrid analysis allows us to fully, and more flexibly, account for sampling design to avoid this problem

- Hybrid analysis allows for choice to account for each of HSB's complex sampling features the design-based or model-based way

- In this analysis, chose to adjust for disproportionate selection in design-based model (including sampling weights at both levels during estimation) rather than model-based way (including selection variables as model covariates)

- Accounted for stratification in model-based way (including strata variables as model covariates) rather than design-based way (standard error adjustments using the HSB-provided strata indicator SCHSAMP)

- Included fixed effects for SECTOR, high percentage Black enrollment (BLACK), high percentage Hispanic enrollment (HISPANIC) and their product terms (SECTOR x BLACK and SECTOR x HISPANIC)

- Made this choice as SECTOR was of substantive interest in the original model and was thought to interact with independent variables

- Accounted for clustering the model-based way (inclusion of random effects for cluster) rather than the design-based way (standard error adjustments using the HSB-provided cluster indicator SCHLID)

- Including sampling weights, we get the following table:

TABLE 2
Illustrative Hybrid Design/Model-Based Analysis Using the High School and Beyond (HSB) Data Set

Fixed Effects	*1. Original, Model-Based Analysis*[a] *Accounts for Clustering; Partially Accounts for Stratification*	*2. Hybrid Analysis, Intermediate Step* *Model 1 Plus Weights to Account for Disproportionate Selection*	*3. Hybrid Analysis, Final Step* *Model 2 Plus Covariates to Fully Account for Stratification*
INTERCEPT	7.27 (.06)**	7.59 (.11)**	7.76 (.12)**
CSES	2.09 (.07)**	2.16 (.11)**	2.16 (.11)**
MEANSES	4.43 (.14)**	4.37 (.25)**	3.60 (.28)**
SECTOR	−0.06 (.24)	−0.01 (.36)	0.15 (.36)
CSES × MEANSES	0.62 (.17)**	0.39 (.28)	0.39 (.28)
CSES × SECTOR	−1.50 (.19)**	−1.63 (.27)**	−1.63 (.28)**
HISPANIC			−0.96 (.32)**
BLACK			−1.83 (.29)**
HISPANIC × SECTOR			0.25 (.59)
BLACK × SECTOR			−1.42 (.74)
Variance Components			
τ_{00}	1.86 (.15)**	2.05 (.27)**	1.73 (.24)**
τ_{01}	0.31 (.13)*	0.27 (.16)	0.18 (.16)
τ_{11}	0.29 (.10)**	0.53 (.26)*	0.55 (.26)*
σ^2	21.89 (.25)**	21.09 (.24)**	21.08 (.34)**

[a]The model-based analysis results in Column 1 differ somewhat from those of Raudenbush and Bryk (2002, chap. 4) and Singer (1998) for two reasons. First, our variables were taken directly from HSB's 1982 public-use datafile for the sophomore cohort (see online Appendix and www.icpsr.umich.edu). Raudenbush and Bryk (2002) constructed and used factor score composites of 1980 and 1982 datafile variables for sophomore and senior cohorts (Lee & Bryk, 1989). Second, we used all public and Catholic schools and they used a random subset. MATHACH = math achievement; MEANSES = school average socioeconomic status; SECTOR = Catholic or public school; CSES = school mean centered child socioeconomic status; BLACK = high % Black enrollment; HISPANIC = high % Hispanic enrollment.

*$p < .05$. **$p < .01$.

- In columns 2 and 1 comparison indicates some bias was likely incurred in prior (model-based, unweighted) analyses due to ignoring disproportionate selection

- Conditional slope of CSES on MATHACT is still significant in column 2 and still varies across schools, but slopes for CSES no longer significantly differ according to school MEANSES

- Cross-level interaction of CSES by MEANSES is now nonsignificant

- There is now nonsignificant covariation between intercepts and slopes in Column 2, meaning that the effects of CSES on MATHACH no longer covary with the average MATHACH of the school

- Comparing columns 3 and 2 indicates more fully accounting for stratification does not markedly change conclusions in this case
- However, not only do standard errors change from column 2 to 3 but in this case several parameter estimates do as well
- Stratification variables should affect only standard errors, not parameter estimates when stratification variables neither interact with nor correlate with independent variables
- Not the case here, we do not explore here whether school racial composition interacts with student or school socioeconomic status

3. Conclusion

- In reviewing the design-based inferential framework, we showed that the different kinds of statistical inferences (descriptive rather than analytic) to different populations (finite rather than infinite) were possible exclusively under random sampling – and their accuracy was not dependent on the proper specification of a hypothetical model

- We provided reasons for the design-based framework's lack of implementation in psychology

- Showed the hybrid model's ability to overcome the limitations of the model- and design-based framework that can be used to analyze large, complex random samples from public-use data sets

    - This practice is becoming more common in psychology