Evolution of Models in Survey Sampling
by
Rick Valliant

Discussion: Did the use of models in survey sampling devolved to evolve?

Trivellore Raghunathan (Raghu)

University of Michigan

Fascinating History

- My travel down the memory lanes
  - Heavy influence of Godambe (founder of the department, along with Shrikhande), Sukathme and Mahalanobis
  - Same instructor for design of sample surveys and design of experiments
    - Design
      - The same role of Randomization
      - Blocking versus stratification
      - Replications versus allocations
    - Analysis
      - Study the variables and their correlates
    - Evaluate
      - Holistic
        - Sampling and non-sampling errors
  - What computing power?

(except for Facit mechanical calculator)

Model for Sampling Indicators, the I's

Models for the Y's

Text books:
Cochran, Sukhatme, Deming

Role of Models in Designs

$$Y_i : Outcome$$

$$I_i : Sampling\ \ Indicator$$

- The Pis don’t come from the skies

$$Pr\left(I_i = 1 \right) = \pi_i$$

  - Implicit/ explicit stratification –model relationship between Y_i and π_i
- Deming, Cochran, Horvitz, Thompson etc : For efficient estimates, inclusion probabilities should be nearly proportional to the outcome variable (for example, previous census or other information)
- Closely related to ratio estimates- feed back mechanism on the choice of pi’s
- Dalenius (1950) and Dalenius and Gurney (1951) use model for Y to create stratification

Cluster sampling

- Correlation of the I’s or Y’s?
  - Design makes the I’s correlated
  - Analysis treats the Y’s as correlated
- Models for within and between cluster variances
  - Smith and Fairfield (1938)

$$\log S_b^2 \approx \log S^2 - g \log M$$

  - Mahalanobis (1940), Jessen (1942)

$$\log S_w^2 \approx a + b \log M$$

Misconceptions

- Myth: Sample Design is irrelevant for modelers
- Fact: Need to be ignorable and model for Y should condition on design variables $\Pr(Y, I \mid Z) = \Pr(Y \mid Z)\Pr(I \mid Y, Z) \equiv \Pr(Y \mid Z)\Pr(I \mid Z)$

$$\Pr(Y, I \mid Z) = \Pr(Y \mid Z, I)\Pr(I \mid Z) \equiv \Pr(Y \mid Z)\Pr(I \mid Z)$$

$$\Pr(Y \mid Z, I = 0) = \Pr(Y \mid Z, I = 1)$$

- Rubin (1976, Biometrika), Rubin (1987, Chapter 2), Little (1982, JASA)
- (HMT) Perils of using

$$\Pr(Y) \; instead \;\; of \;\; \Pr(Y \mid Z) \; even \;\; if \;\; \| \Pr(Y) - \Pr(Y \mid Z) \| \;\; is \;\; small$$

- Reiter, Raghunathan and Kinney (2006)
- Synthetic populations to account for complex sample design and then model the synthetic populations (Dong, Elliott, Raghunathan (2014), Zhou, Elliott and Raghunathan (2016)

Survey Inference as a prediction problem, Missing data problem

- Ericson (1969), Smith (1974), Geisser (1993 Book), Rubin (1987), Little and Rubin (2022)
- Frequency or repeated sampling calculations are justifiable
  - Model Inference with Frequency calibration
  - Calibrated Bayes (Box, Rubin, Little)
- Computational infrastructure: Handle survey inference with full complexity using properly tuned and calibrated models
- Leverage auxiliary data and non-probability sources to make sample design more efficient
  - Models to inform designs
  - Designs to inform models
- Agree that the future is model based prediction based on smallish well designed probability sample surveys and leveraging organic data
- Not give up on probability sample designs in this era of DEI