Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 3: Modes of survey inference

University of Michigan

Approaches to Survey Inference

- Design-based (Randomization) inference
- Superpopulation Modeling
  - Specifies model conditional on fixed parameters
  - Frequentist inference based on repeated samples from superpopulation and finite population (hybrid approach)
- Bayesian modeling
  - Specifies full probability model (prior distributions on fixed parameters)
  - Bayesian inference based on posterior distribution of finite population quantities
  - argue that this is most satisfying approach

Design-Based Survey Inference

$Z = (Z_1, ..., Z_N) =$ design variables, known for population

$I = (I_1, \ldots, I_N)$ = Sample Inclusion Indicators

$$I_i = \begin{cases} 1, & \text{unit included in sample} \\ 0, & \text{otherwise} \end{cases}$$

$Y = (Y_1, ..., Y_N)$ = population values,

recorded only for sample

$Y_{\text{inc}} = Y_{\text{inc}}(I) =$ part of Y included in the survey

Note: here I is random variable, (Y, Z) are fixed

$Q = Q(Y, Z)$ = target finite population quantity

$\hat{q} = \hat{q}(I, Y_{\text{inc}}, Z)$ = sample estimate of Q

$\hat{V}(I, Y_{\text{inc}}, Z)$ = sample estimate of V

$\left(\hat{q} - 1.96\sqrt{\hat{V}}, \hat{q} + 1.96\sqrt{\hat{V}}\right) =$ 95% confidence interval for Q

I	Z	Y
1		Y_{inc}
1		
1		
0		$[Y_{\text{exc}}]$
0		
0		
0		
0		

Random Sampling

- Random (probability) sampling characterized by:
  - Every possible sample has known chance of being selected
  - Every unit in the sample has a non-zero chance of being selected
  - In particular, for simple random sampling with replacement: "All possible samples of size *n* have <u>same</u> chance of being selected"

$Z = \{1, \ldots, N\}$ = set of units in the sample frame

$$\Pr(I \mid Z) = \begin{cases} 1/\left(C_n^N\right), \sum_{i=1}^{N} I_i = n, \\ 0, \text{ otherwise} \end{cases}; \quad C_n^N = \frac{N!}{n!(N-n)!}, n! = 1 \times 2 \ldots \times n$$

$$E(I_i \mid Z) = \Pr(I_i = 1 \mid Z) = n / N$$

Example 1: Mean for Simple Random Sample

$$Q = \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \text{ population mean}$$

Random variable

$$\hat{q}(I) = \overline{y} = \sum_{i=1}^{N} I_i y_i / n, \text{ the sample mean}$$

Fixed quantity, not modeled

$$\text{Unbiased for } \overline{Y} : E_I\left(\sum_{i=1}^{N} I_i y_i / n\right) = \sum_{i=1}^{N} E_I(I_i) y_i / n = \sum_{i=1}^{N} (n/N) y_i / n = \overline{Y}$$

$$\text{Var}_I(\overline{y}) = V = {\color{red}(1-n/N)}\, S^2/n, \;\; S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$$

$${\color{red}(1-n/N)} = \text{ finite population correction}$$

$$\hat{V} = (1-n/N)s^2/n, \; s^2 = \text{sample variance} = \frac{1}{n-1}\sum_{i=1}^{N} I_i (y_i - \overline{y})^2$$

$$\text{95\% confidence interval for } \overline{Y} = \left(\overline{y} - 1.96\sqrt{\hat{V}}, \overline{y} + 1.96\sqrt{\hat{V}}\right)$$

Example 2: Horvitz-Thompson estimator

$$Q(Y) = T \equiv Y_1 + \ldots + Y_N$$

$$\pi_i = E(I_i \mid Y) = \text{inclusion probability } > 0$$

$$\hat{t}_{\text{HT}} = \sum_{i=1}^{N} I_i Y_i / \pi_i,\ \mathrm{E}_I(\hat{t}_{\text{HT}}) = \sum_{i=1}^{N} E(I_i) Y_i / \pi_i = \sum_{i=1}^{N} \pi_i Y_i / \pi_i = T$$

$$\hat{v}_{\text{HT}} = \text{Variance estimate, depends on sample design}$$

$$\left(\hat{t}_{\text{HT}} - 1.96\sqrt{\hat{v}_{\text{HT}}},\ \hat{t}_{\text{HT}} + 1.96\sqrt{\hat{v}_{\text{HT}}}\right) = 95\% \text{ CI for } T$$

- Pro: unbiased under minimal assumptions
- Cons:
  - variance estimator problematic for some designs (e.g. systematic sampling)
  - can have poor confidence coverage and inefficiency -- Basu “weighs in” with the following amusing example

Ex 2. Basu's inefficient elephants

$$\left(y_1,\ldots,y_{50}\right) = \text{ weights of } N = 50 \text{ elephants}$$

Objective: $T = y_1 + y_2 + \ldots + y_{50}$. Only one elephant can be weighed!

- Circus trainer wants to choose "average" elephant (Sambo)
- Circus statistician requires "scientific" prob. sampling:

  Select Sambo with probability 99/100

  One of other elephants with probability 1/4900

  Sambo gets selected! Trainer: $\hat{t} = y_{(\text{Sambo})} \times 50$

  Statistician requires unbiased Horvitz-Thompson (1952)

estimator: $$\hat{T}_{HT} = \begin{cases} y_{(\text{Sambo})} / 0.99 \text{ (!!);} \\ 4900 y_{(i)}, \text{if Sambo not chosen (!!!)} \end{cases}$$

HT estimator is unbiased on average but always crazy!

Circus statistician loses job and becomes an academic

Role of Models in Classical Approach

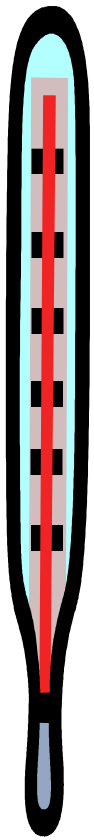
- Models are often used to motivate the choice of estimator. For example:
  - Regression model $\longrightarrow$ regression estimator
  - Ratio model $\longrightarrow$ ratio estimator
  - Generalized Regression estimation: model estimates adjusted to protect against misspecification, e.g. HT estimation applied to residuals from the regression estimator (Cassel, Sarndal and Wretman book).
- Estimates of standard error are then based on the randomization distribution
- This approach is design-based, model-assisted

Summary of design-based approach

- Avoids need for models for survey outcomes
- Robust approach for large probability samples
- Models needed for nonresponse, response errors, small areas
- Not well suited for small samples – inference basically assumes large samples, and models are needed for better precision in small samples
  - leading to “inferential schizophrenia”…

Inferential Schizophrenia

n-ometer



Design-based inference

Model-based inference

n_0 = "Point of inferential schizophrenia"

How do I choose n_0?
If $n_0 = 35$, should my entire statistical philosophy be different when n=34 and n=36?

Limitations of design-based approach

- Some raise theoretical objections to repeated-sampling inferences in general
  - Violates the likelihood principle (Birnbaum 1968)
  - Ambiguity about conditioning on ancillary statistics
- Inference based on probability sampling, but true probability samples are harder and harder to come by:
  - Noncontact, nonresponse is increasing
  - Face-to-face interviews increasingly expensive
- Can’t do “big data” (e.g. internet, administrative data) from the design-based perspective

Model-Based Approaches

- In the Bayesian approach models are used as the basis for the entire inference: estimator, standard error, interval estimation
- This approach is more unified, but models need to be carefully tailored to features of the sample design such as stratification, clustering.
- One might call this model-based, design-assisted
- Two variants:
  - Superpopulation Modeling
  - Bayesian (full probability) modeling
- Common theme is “Infer” or “predict” about non-sampled portion of the population conditional on the sample and model

Superpopulation Modeling

- Model distribution *M:*

  $Y \sim f(Y \mid Z, \theta), Z$ = design variables, $\theta =$ fixed parameters

- Predict non-sampled values $\hat{Y}_{\text{exc}}$:

$\hat{y}_i = E(y_i \mid z_i, \theta = \hat{\theta}), \hat{\theta} =$ model estimate of θ

$$\hat{q} = Q(\tilde{Y}), \tilde{y}_i = \begin{cases} y_i, & \text{if unit sampled;} \\ \hat{y}_i, & \text{if unit not sampled} \end{cases}$$

$\hat{v} = m\hat{s}e(\hat{q})$, over distribution of I and M

$(\hat{q} - 1.96\sqrt{\hat{v}}, \hat{q} + 1.96\sqrt{\hat{v}})$ = 95% CI for Q

I	Z	Y
1		Y_{inc}
1		
1		
0		$\hat{Y}_{\text{exc}}$
0		
0		
0		
0		

In the modeling approach, prediction of nonsampled values is central

In the design-based approach, weighting is central: “sample represents … units in the population”

Bayesian Modeling

Bayesian model adds a prior distribution for the parameters:

$$(Y,\theta) \sim \pi(\theta \mid Z) f(Y \mid Z,\theta), \;\; \pi(\theta \mid Z) = \text{ prior distribution}$$

Inference about θ is based on posterior distribution from Bayes Theorem:

$$p(\theta \mid Z, Y_{\text{inc}}) \propto \pi(\theta \mid Z) L(\theta \mid Z, Y_{\text{inc}}), L = \text{likelihood}$$

Inference about finite population quantitity $Q(Y)$ based on

$$p(Q(Y) \mid Y_{\text{inc}}) = \text{ posterior predictive distribution}$$

of Q given sample values Y_{inc}

$$p(Q(Y) \mid Z, Y_{\text{inc}}) = \int p(Q(Y) \mid Z, Y_{\text{inc}}, \theta) p(\theta \mid Z, Y_{\text{inc}}) d\theta$$

(Integrates out nuisance parameters θ)

I	Z	Y
1		
1		Y_{inc}
1		
0		
0		
0		$\hat{Y}_{\text{exc}}$
0		
0		

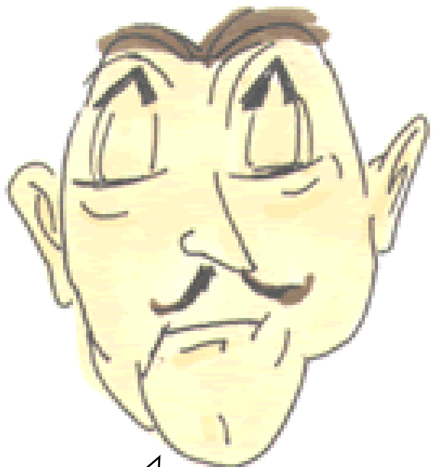
In the super-population modeling approach, parameters are considered fixed and estimated

In the Bayesian approach, parameters are random and integrated out of posterior distribution – leads to better small-sample inference

Advantages of Bayesian approach

- Unified approach for large and small samples, nonresponse and response errors, data fusion, “big data”.
- Frequentist superpopulation modeling has the limitation that uncertainty in predicting parameters is not reflected in prediction inferences
- Bayes propagates uncertainty about parameters, yielding better frequentist properties in small samples
- Statistical modeling is the standard approach to statistics in substantive disciplines – having a design-based paradigm for surveys is divisive and confusing to modelers

Models bring survey inference closer to the statistical mainstream



Follow my design-based statistical standards

Why? I am an economist, I build models!

Challenges of the model-based perspective

- Explicit dependence on the choice of model, which has subjective elements (but assumptions are explicit)
- Bad models provide bad answers – justifiable concerns about the effect of model misspecification
  - In particular, models need to reflect features of the survey design, like clustering, stratification and weighting
- Models are needed for all survey variables – need to understand the data
- Potential for more complex computations. Simulation techniques greatly facilitate implementation

Overarching philosophy: calibrated Bayes

- Survey inference is not fundamentally different from other problems of statistical inference
  - But it has particular features that need attention
- Statistics is basically prediction: in survey setting, predicting survey variables for non-sampled units
- Inference should be model-based, Bayesian
- Seek models that are “frequency calibrated” (Box 1980, Rubin 1984, Little 2006):
  - Incorporate survey design features
  - Properties like design consistency are useful
  - “objective” priors generally appropriate
    - Little (2004, 2006, 2012); Little & Zhang (2007)

Calibrated Bayes

"The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie."

"… frequency calculations are useful for making Bayesian statements scientific, … in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

Rubin (1984)