The Evolution of the Use of Models in Survey Sampling
Hansen Lecture 2022

Richard Valliant

University of Michigan & University of Maryland

2022

A few of Morris' contributions

- Led move to make probability sampling the standard for finite population estimation
- Improved statistical practice throughout US and foreign governments
- Trained many statisticians
- *Sample Survey Methods and Theory I & II* (1953)
- Innovations in specific surveys
  - First sample survey estimates of employment and unemployment in 1930s (which became the CPS)
  - Sample design of Consumer Price Index and related BLS surveys
  - Sample design of National Assessment of Education Progress (NAEP)
- Olkin interview (*Stat. Sci.* 1987). Waksberg, and Goldfield remembrance (NAS 1996).

A little background

- Worked at Westat 1975-1980
- For Morris, Joe Waksberg, Ben Tepping

Computing power

- Played key role in bringing UNIVAC to Census Bureau
- Huge increases since Morris was working; he died in 1990 (32 years ago)
- He was fascinated by gizmos
- When HP calculators came out, we all bought one

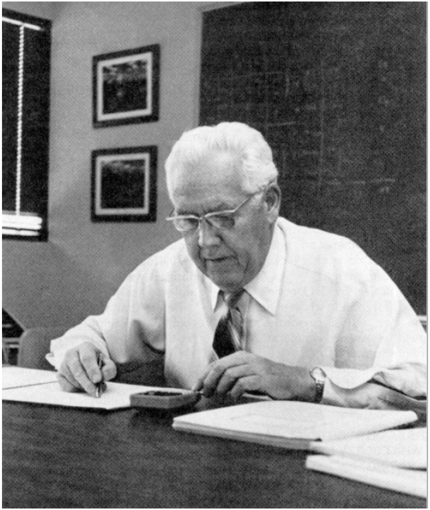
  HP25-C, programmable (49 program steps)
  Reverse Polish Notation
  Continuous memory

  $200 at Chafitz Calculators.

HP-25C

An improved version of the HP-25 featuring continuous memory. Programs and data were retained when the calculator was turned off, because the new low power CMOS memory was powered even in the "off" state. It was even possible to change the battery pack and retain the memory because of a capacitor inside. Users expect continuous memory today but it was an impressive upgrade at the time. This was the first example of HP upgrading rather than replacing a calculator.

Morris & an early laptop



Computing developments since 1970s

- Current computer options allow far more complex things to be done–commercial software, PCs, R
- How would Morris react to these things? I think he would dive in and learn all he could. He wasn't averse to taking advantage of new developments.
- He was a big believer in replication variance estimation which allowed adjustments like nonresponse to be reflected.

  Advanced theory for rep vars was developed in the mid-1980s by Rao, Shao, Wu and others, but MHH was using replication well before then.
- Would he treat developments in modeling like an innovation that he needed to learn?

  Probably, at least for model assisted estimation.

Computing developments since 1970s

- Current computer options allow far more complex things to be done–commercial software, PCs, R
- How would Morris react to these things? I think he would dive in and learn all he could. He wasn't averse to taking advantage of new developments.
- He was a big believer in replication variance estimation which allowed adjustments like nonresponse to be reflected.

  Advanced theory for rep vars was developed in the mid-1980s by Rao, Shao, Wu and others, but MHH was using replication well before then.
- Would he treat developments in modeling like an innovation that he needed to learn?

  Probably, at least for model assisted estimation.

His biggest fear about models ...

People would quit using probability sampling

His biggest fear about models ...

People would quit using probability sampling

Design-based inference

- All calculations of expectations and variances are made with respect to random sampling design used in selecting the sample
- Many departures from "by the book" procedures
- Systematic sampling from a list sorted by some auxiliary variable(s)
  - When list is sorted in a particular way, joint selection probs for some pairs of units are 0 $\Rightarrow$ unbiased variance estimation not possible
- Randomization analyses using the PISE method *("pretend it's something else")*
  - Systematic sampling from a sorted list treated as if the order of the list was randomized or that the sort provides implicit stratification

Design-based inference

- All calculations of expectations and variances are made with respect to random sampling design used in selecting the sample
- Many departures from "by the book" procedures
- Systematic sampling from a list sorted by some auxiliary variable(s)
  - When list is sorted in a particular way, joint selection probs for some pairs of units are 0 $\Rightarrow$ unbiased variance estimation not possible
- Randomization analyses using the PISE method ("*pretend it's something else*")
  - Systematic sampling from a sorted list treated as if the order of the list was randomized or that the sort provides implicit stratification

Design-based inference

- All calculations of expectations and variances are made with respect to random sampling design used in selecting the sample
- Many departures from "by the book" procedures
- Systematic sampling from a list sorted by some auxiliary variable(s)
  - When list is sorted in a particular way, joint selection probs for some pairs of units are 0 $\Rightarrow$ unbiased variance estimation not possible
- Randomization analyses using the PISE method ("*pretend it's something else*")
  - Systematic sampling from a sorted list treated as if the order of the list was randomized or that the sort provides implicit stratification

Model-based (superpopulation) estimation

- All calculations of expectations and variances are made wrt a model–not the randomization used in the sampling design.
- Introduced in Brewer (*AJS*, 1963) for ratio estimation
- But an earlier mention of the ratio model is in Cochran's *Sampling Techniques* (1st ed., 1953) and linear regression models for finite pops are in Cochran (*JASA* 1942); also Jessen (*Iowa Ag Exp Stat Rsch Bull*, 1942).

Model-based estimation

- Approach formulated in detail by Royall (*BMKA* 1970) and many later papers with co-authors (Eberhard, Herson, Cumberland)

- Formulation of estimating totals as prediction problem was a breakthrough in thinking that clarified the way calculations should be made
    - Compute bias as $E_M(\hat{t} - t_U)$ since pop total is a random variable
    - Coherent distinction between model-based and design-based approaches
    - Model-based calculations treat sample as fixed (not random); statistical distribution provided by model

Model-based estimation

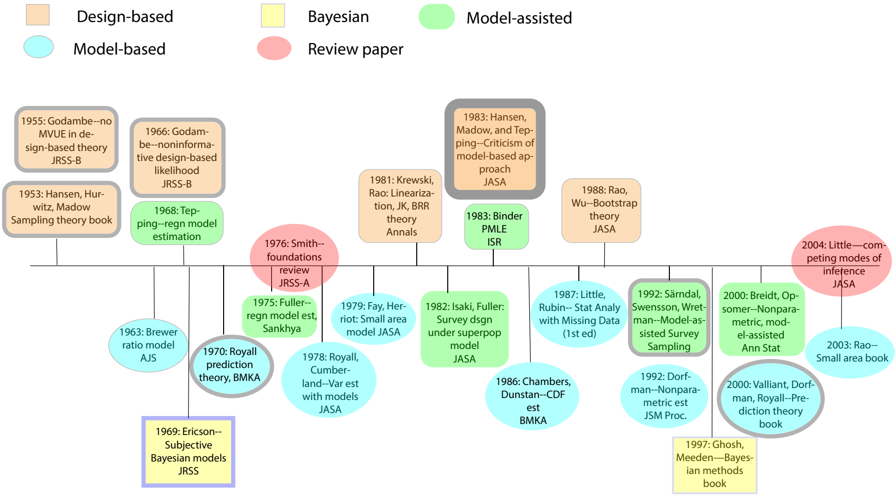
- Valliant, Dorfman, and Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach* collected RMRs work plus added new material on generalized inverses, nonparametric estimators, CDF estimators, and nonlinear models.

- Fundamental idea is that calculations of expectations and variances should be made wrt a superpopulation model

1983 Hansen, Madow, & Tepping paper

- "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys", *JASA*, 1983
- Showed by simulation that a small model misspecification leads to an important bias in a model-based estimator
- Ignorable sample design with full response
- Critiqued by discussants to 1983 paper and in Valliant et al (2000), Section 3.7

Timeline 1953-2004

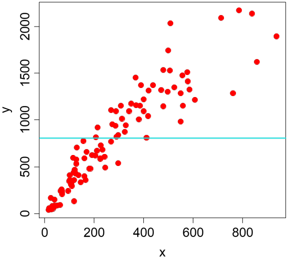


Why reject design-based inference and use model-based instead?

- **Ancillary Statistic.** A statistic whose probability distribution is completely known and does not depend on any unknown parameters.
- **Conditionality Principle** (Cox and Hinckley 1974). Inference should be made conditional on the value of any ancillary statistics.
  This principle says we should condition on the value of observed random variables whose distribution we know and does not depend on any parameters we want to make an inference about.
- In a pure probability design what do we know completely? The distribution of the indicators, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$, for whether units are in a sample or not $\Rightarrow \boldsymbol{\delta}$ is ancillary.
- Other arguments: uninformative likelihood, factorization theorem for sufficient statistics

Easy example of conditional bias

- Select simple random sample
- Estimate population average by sample mean
- Design bias of sample mean $\bar{Y}_s$ is 0
- Model-bias (if straight-line thru origin) is $E_M\left(\bar{Y}_s - \bar{Y}_U\right) \propto \left(\bar{x}_s - \bar{x}_U\right)$
- Model-bias has order $1/\sqrt{n}$ and so does $SE\left(\bar{Y}_s\right)$



- Confidence intervals will not have correct coverage in off-balance SRS's
- Conditional bias problem carries over to more complicated problems. *Every sample does not look like the "average" sample among all possible samples*

Model-assisted estimation

- General idea is to use a model to formulate an estimator but modify it so that the result is design consistent
- Särndal, Swensson, Wretman (1992), *Model Assisted Survey Sampling* came out after MHH's death but the idea of combining design-based randomization and models was in the literature prior to 1992.
    - PMLE: Binder (*ISR* 1983), contemporaneous with Hansen, Madow, and Tepping (*JASA* 1983) criticism of model-based estimation
    - Precursors: Tepping (*Proc. ASA* 1968), Fuller (*Sankhya* 1975, *SurvMeth* 2002)

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model   design-based
- Response model   quasi-randomization model or y model
- Coverage model   quasi-randomization model
- Imputation model   randomization model or y model
- Prior   model for parameters
- Hyper-prior   model for parameters
- Posterior   model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model   design-based
- Response model   quasi-randomization model or y model
- Coverage model   quasi-randomization model
- Imputation model   randomization model or y model
- Prior   model for parameters
- Hyper-prior   model for parameters
- Posterior   model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model   design-based
- Response model   quasi-randomization model or y model
- Coverage model   quasi-randomization model
- Imputation model   randomization model or y model
- Prior   model for parameters
- Hyper-prior   model for parameters
- Posterior   model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model design-based
- Response model quasi-randomization model or y model
- Coverage model quasi-randomization model
- Imputation model randomization model or y model
- Prior model for parameters
- Hyper-prior model for parameters
- Posterior model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model   design-based
- Response model   quasi-randomization model or y model
- Coverage model   quasi-randomization model
- Imputation model   randomization model or y model
- Prior   model for parameters
- Hyper-prior   model for parameters
- Posterior   model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Distributions used in sampling

In practice things are a lot more complicated than just design-based vs. model-based ...

- Superpopulation model for Y's y model
- Random selection model   design-based
- Response model   quasi-randomization model or y model
- Coverage model   quasi-randomization model
- Imputation model   randomization model or y model
- Prior   model for parameters
- Hyper-prior   model for parameters
- Posterior   model for parameters

Using models is unavoidable in finite population sampling because of all the things that are out of our control: nonresponse, non-coverage, missing item data

Standard form of an estimated total

- Standard practice in surveys is to compute one set of weights, then use them to estimate everything–means, totals, regression parameters, etc.
- Estimated total: $\hat{t} = \sum_{i \in s} w_i y_i$
- The weights are meant to produce design-unbiased, or at least, consistent estimators
- Same weights are used for quantitative or qualitative y's
- "Implied" model is one under which $\hat{t}$ model-unbiased or consistent.
  Typically, the implied model is linear (in simplest cases).

Model-based vs. model-assisted

Suppose underlying model is $y_i = \mu(\mathbf{x}_i) + \varepsilon_i$

Model can be linear or nonlinear in x's

- *Model-based*

  $\hat{t}_{MB} = \sum_{i \in U} \tilde{\mu}(\mathbf{x}_i) + \sum_{i \in s} \tilde{e}_{Mi}, \quad \tilde{e}_{Mi} = y_i - \tilde{\mu}(\mathbf{x}_i)$

- *Model-assisted* (Breidt & Opsomer, *Handbook of Stat 2009*)

  $\hat{t}_{MA} = \sum_{i \in U} \hat{\mu}(\mathbf{x}_i) + \sum_{i \in s} \frac{e_{MAi}}{\pi_i}, \quad e_{MAi} = y_i - \hat{\mu}(\mathbf{x}_i)$

- *Model calibrated* (Wu and Sitter *JASA* 2001)

  $\hat{t}_{MC} = \sum_{i \in s} \frac{\hat{\mu}(\mathbf{x}_i)}{\pi_i} + \sum_{i \in s} \frac{e_{MAi}}{\pi_i}$

Particular cases based on how $\mu(\mathbf{x}_i)$ is estimated

- $\hat{t}_{MB}$ is BLUP when $\mu(\mathbf{x}_i) = \mathbf{x}_i^T \beta$ (Royall & Cumberland *JASA* 1978)
- GREG is special case of $\hat{t}_{MA}$ with a linear model
- $\hat{t}_{MB}$ and $\hat{t}_{MA}$ are nonparametric if μ estimated by local polynomial regression (Dorfman *JSM Proc.* 1992; Chambers, et al. *JASA* 1993), neural networks (Montanari & Ranalli *JASA* 2005), GAM (Opsomer, et al. *JRSS-B* 2008)
- Regression trees are another option
  - Bayesian version in Wang, Rothschild, Goel, and Gelman (*Int.J.Forecasting*, 2015), multilevel regression and poststratification

Categorical y's

- Nonlinear models
- Example models are logistic for binary y and multinomial logistic for multi-category y's
- Logistic model: model est in Valliant (*JASA* 1985)
  MA estimator in Lehtonen and Veijanen (*SurvMeth* 1998)
- Multinomial MA in Kennel & Valliant (*JSSAM* 2021)

Empirical likelihood

- Pop composed of discrete set of values, $\{y_i\}_{i=1}^N$, some of which can be the same (first proposed by Hartley & Rao, *BMKA* 1968)
- $p_i = Pr\left(y = y_i\right)$ is mass assigned to y_i
- If y_i's are *iid*, the census likelihood is $L_N\left(\boldsymbol{p}\right) = \prod_{i=1}^N p_i$
- Pseudo-empirical log-likelihood (PELL, Chen & Sitter, *Stat Sinica* 1999; Wu & Rao, *CJS* 2006) is

  $l_n\left(\boldsymbol{p}\right) = n \sum_{i \in s} \widetilde{d}_i\left(s\right) \log\left(p_i\right)$

  where $\widetilde{d}_i\left(s\right) = \frac{d_i}{\sum_{i \in s} d_i}; \quad d_i = \pi_i^{-1}$
- Find $\{\hat{p}_i\}_{i \in s}$ to maximize the PELL

Empirical likelihood (continued)

- Calibration achieved by maximizing $l_n(\boldsymbol{p})$ subject to $p_i > 0$, $\sum_{i\in s} p_i = 1$, and $\sum_{i\in s} p_i x_i = \bar{x}_U$
- Estimator of pop mean is $\bar{y}_{PELL} = \sum_s \hat{p}_i y_i$
- $\hat{p}_i$ are normalized weights
- Extended by Wu & Sitter (*JASA* 2001) to case where underlying model is linear or nonlinear:

  $$E_M(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})\,;\, V_M(y_i \mid \boldsymbol{x}_i) = v(\boldsymbol{x}_i)\,\sigma^2$$

Advantages of empirical likelihood

- The $\{\hat{p}_i\}_{i\in s}$ are normalized weights that are always in (0,1)
- $\hat{F}(t) = \sum_{i\in s} \hat{p}_i I(y_i \leq t)$ is a CDF; quantiles estimated by inversion
- Works well in pops with many 0's, e.g., audit applications where most accounts have no errors but some have non-zero dollar-value errors
- CI's perform better than normal approximation intervals when estimating prevalence of rare characteristics
- But, likelihood being maximized depends on y

Some pros and cons for practice

- Pro: Some estimators lead to element-level weights (BLUP, GREG, PELL)
- Con: Element-level weights can be different for different y's (BLUP, PELL)
- Con: Some estimators do not yield element-level weights (trees, nonparametric, semiparametric, Bayesian)
- Con: Heavy computational burdens for some estimators that must be repeated for every y — Bayesian, some nonparametric & semiparametric, PELL

Designing samples using models

- Balanced sample: match sample moments to population moments for x's
- Cutoff samples: Single quantitative estimate with y variable closely related to an auxiliary on the frame; leads to cutoff sample being optimal
  Yorgason et al. (2011). *Cutoff Sampling in Federal Surveys*
- EIA Monthly Natural Gas Report is a cutoff sample of about 220 companies producing large volumes of natural gas. Companies in the sample account for 85% to 90% of all gas produced in lower 48 states.
- Anticipated variances
  - Godambe & Joshi (*AMS* 1965) optimal MOS in straight-line through origin model, $\sqrt{v}$
  - Extended by Isaki & Fuller (*JASA* 1982) to linear regression model: $E_M(y_i) = \boldsymbol{x}_i^T \beta$ and $V_M(y_i) = v_i$
  - Anticipated variances for variance components in multistage surveys (Valliant, Dever, Kreuter, *JOS* 2015)

Nonprobability sampling

- Other fields have used nonprobability samples for years
- Clinical trials in medical research are rarely (maybe never) based on probability samples from a well-defined finite population
  Lack of representation of some demographic groups (e.g., Blacks and women) is a recognized problem, but findings can still be useful.
- If we restrict ourselves only to cases where probability samples can be selected, we eliminate using some of the newer, readily available sources of data.

Nonprobability sampling

- Inferences are entirely model-based
- Problems
  - Selection bias (coverage error): characteristics of sample different from nonsample
  - Nonresponse (in panels)
  - Attrition (in panels)
  - Measurement error (Kennedy 2021 Hansen lecture)
- Even best probability surveys have coverage problems, e.g., Blacks have 75-80% coverage in the CPS. Coverage rates are worse for some subgroups like young Hispanic males, elderly Black men and women
- Coverage in nonprobability data sets is largely uncontrolled

Not all types of NP samples are equally good

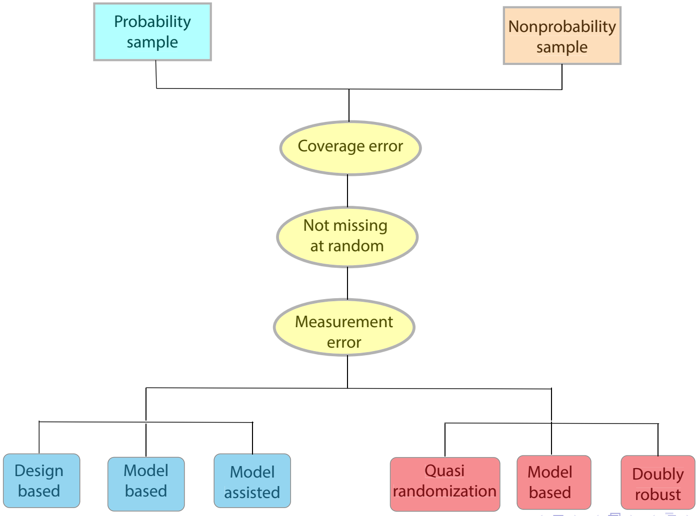
A few types of nonprobability samples

- Mall intercepts
- Volunteer panels of persons
- Panels recruited via addressed-based sampling (ABS)
- Incomplete administrative data because of, e.g., late or incomplete reporting (police crime reports, late tax return filers), lack of permission to link admin data to samples
- Data scraped from web
  - Airline prices used by BLS in CPI
  - MIT billion prices project 2008-2016
  - Twitter

Estimation from nonprobability samples

- Elliott & Valliant (*Stat Sci* 2017)
- Options
  - Quasi-randomization (QR)
    Estimate pseudo-inclusion probs using a reference prob sample
  - Superpopulation prediction (SP)
    Estimation based on model for y's
  - Doubly robust (DR)
    Combine QR and SP
- Theory: Likelihood formulation for estimating pseudo-inclusion probs + superpop model (Chen, Li, Wu; *JASA* 2019)
- Many other articles available

Parallels between nonprobability and probability samples



Integrating probability and nonprobability samples, s_p and s_{np}

- Worries in combining different data sources
  - Different modes of data collection
  - Different types of response errors
  - Different wordings, question contexts
- Lohr & Raghunathan review paper (*Stat Sci* 2017) and references
  - Concatenate data sets and impute missing values; could be applied if y's collected in both prob and nonprob samples; weights developed separately for s_p and s_{np}, composite estimation used (Kim & Rao *BMKA* 2012; Gelman, King, Liu *JASA* 1998)
- Mass imputation: y's collected only in nonprob sample. Impute y's to units in prob sample (Feder & Pfeffermann, 2015; Marella & Pfeffermann *ISR* 2022); can be used when nonresponse is non-ignorable

Summary

- Virtually all estimators used in finite population estimation depend on models (explicit or implicit)
  - Models for y's
  - Models for coverage
  - Models for response
  - Models used to create imputations
  - Models for parameters (Bayesian)
  - Models for small area estimation

Making clear what models underly statistical procedures is good practice

Future directions & issues

- Chasm between methods commonly used in practice and methods in literature
- Best procedures for estimation and imputation are y-specific
  - Standard of single-weight analysis prevents "best" being used
  - Limitations on time, effort, and cost that can be expended on any given survey
- Computing power becomes greater each year (we've been saying this for decades). This allows y-specific procedures to be more feasible.

  But, specialized software is required

Single purpose surveys

- Single purpose surveys can use most sophisticated and specialized estimators available
- Surveys done to support litigation
    - Identifying defective components in manufacturing
    - Locating victims of predatory lending practices
- Some election polls
- Audit samples to estimate $ amounts of depreciable items or items in error

Options for multipurpose surveys

- If implied model for an estimator is incorrect, model bias-squared and variance are same order of magnitude
  $\Rightarrow$ Important to get model as close to correct as possible
- Practical implications in multipurpose surveys
  - Select form of estimator that works reasonably well for many y's
  - Identify x's that are predictive of coverage rates, inclusion probabilities, and as many y's as feasible
  - Incorporate those x's in the estimator
  - An estimator like GREG, raking, or deep poststratification is still probably easiest to implement and yields element-level weights
- Result is "model assisted" in the sense of including estimates of coverage/inclusion probabilities and model for y
- Many refinements available to simultaneously account for inclusion rates, y model structure, and control extreme weights, e.g. raking with weight bounds; calibration with non-ignorable nonresponse (Kott & Chang, *JASA* 2010)