

Some History of Using Models in Survey Sampling

IASS seminar

Richard Valliant

University of Michigan & University of Maryland

2025

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

- 1 Outline
- 2 Background**
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

Early contributors to sampling theory & practice

- Review papers by T.M.F. Smith: (Smith, 1976; Smith, 1984; Smith, 1994)
- A. N. Kiaer (1903: ISI recommended *stsr*s with proportional allocation)
- A. L. Bowley (1906: assessed accuracy of estimates from large random samples drawn from larger finite populations, *JRSS*)
How to sample groups like census districts (clusters) was a worry. Balanced sampling proposed.
- **Jerzy Neyman** (Neyman, 1934) : randomization principle, confidence intervals, allocation to strata, cluster sampling; showed non-random samples from Italian census produced poor estimates for many variables
 - In discussion, R.A. Fisher argued that randomization was good for design but not analysis of results.

Other contributors

- P. C. Mahalanobis
- Frank Yates
- W. G. Cochran
- Leslie Kish
- Morris Hansen

A few of Hansen's contributions

- Led move to make probability sampling the standard for finite population estimation in late 1930s and 1940s (along with W.N. Hurwitz and others at US Census Bureau)
- Improved statistical practice throughout US and foreign governments
- *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 Current Pop Survey)
 - Sample design of Consumer Price Index and related economic surveys
 - Sample design of National Assessment of Education Progress (NAEP)
- Olkin interview (Olkin, 1987). Natl Acad of Sci. (Waksberg and Goldfield, 1996); Hansen Lecture 2022 (Valliant, 2024a)

Computing power

- Played key role in bringing UNIVAC to Census Bureau in 1951
- Huge increases since Hansen was working; he died in 1990 (35 years ago)
- Best personal computers available in 1970s were pocket calculators

HP25-C, programmable (49 program steps)

Reverse Polish Notation

Continuous memory

\$200 at Chafitz Calculators

HP-25C



MoRPG

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.



Computing developments since 1970s

- Current computer options allow far more complex things to be done—commercial software, PCs, R
- The early samplers were not averse to complex calculations, e.g., use of replication variances
- But current computing power allows much more complexity, especially in use of models
- Would the other early samplers treat developments in modeling as innovations that they 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
- The early samplers were not averse to complex calculations, e.g., use of replication variances
- But current computing power allows much more complexity, especially in use of models
- Would the other early samplers treat developments in modeling as innovations that they needed to learn?

Probably, at least for model assisted estimation.

Hansen's (and other design-based practitioners)
biggest fear about models ...

People would quit using
probability sampling

Hansen's (and other design-based practitioners)
biggest fear about models ...

People would quit using
probability sampling

- 1 Outline
- 2 Background
- 3 Approaches to inference**
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

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, 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, 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

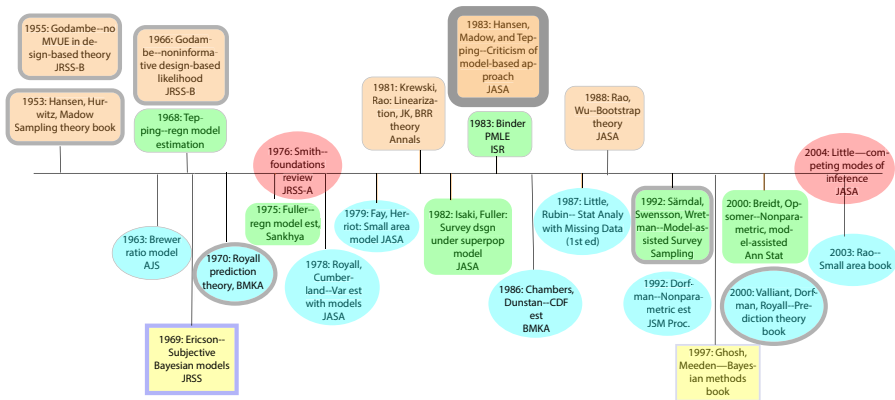
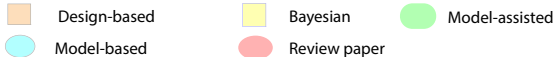
- Valliant, Dorfman, and Royall, 2000. *Finite Population Sampling and Inference: A Prediction Approach* collected Royall's work plus added new material on generalized inverses, nonparametric estimators, CDF estimators, and nonlinear models.
- Fundamental idea is that calculations of expectations and variances should be made wrt a superpopulation model
- Bayesian approach puts priors on model parameters (Ghosh and Meeden, 1997; Ghosh, 2009)

1983 Hansen, Madow, & Tepping paper

- "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys", *JASA*, (Hansen, Madow, and Tepping, 1983)
- Showed by simulation that a small model misspecification leads to an important bias in a model-based estimator
- HMT used ignorable sample design with full response
- Critiqued by discussants to 1983 paper and in (**sec. 3.7**; Valliant, Dorfman, and Royall, 2000)

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline**
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

Timeline 1953-2004



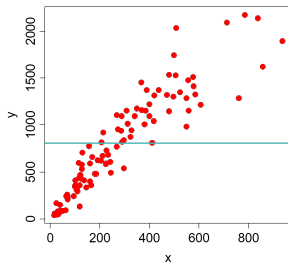
- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based**
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

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, $\delta = (\delta_1, \dots, \delta_N)$, for whether units are in a sample or not $\Rightarrow \delta$ is ancillary.
- Other arguments: uninformative likelihood, factorization theorem for sufficient statistics
- *Seminal Ideas and Controversies in Statistics* (Little, 2025)

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 $E_M(\bar{Y}_s - \bar{Y}_U) \propto (\bar{x}_s - \bar{x}_U)$
- Model-bias has order $1/\sqrt{n}$ and so does $SE(\bar{Y}_s)$
- 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, and 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, 1983 (*ISR*), contemporaneous with Hansen, Madow, and Tepping (*JASA* 1983) criticism of model-based estimation
 - Precursors: Tepping (*Proc. ASA* 1968), Fuller (*Sankhya* 1975, *SurvMeth* 2002)

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions**
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

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

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means**
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

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 and Opsomer, 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, 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 $\mu(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ (Royall and Cumberland, 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, 1992; Chambers, Dorfman, and Wehrly, 1993; Montanari and Ranalli, 2005; Opsomer et al., 2008)
- Regression trees are another option
 - Bayesian version in Wang et al., 2015, multilevel regression and poststratification (MRP)

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, 1985
MA estimator in Lehtonen and Veijanen, 1998
- Multinomial MA in Kennel and Valliant, 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 and Rao, 1968)
- $p_i = Pr(y = y_i)$ is mass assigned to y_i
- If y_i 's are *iid*, the census likelihood is $L_N(\mathbf{p}) = \prod_{i=1}^N p_i$
- Pseudo-empirical log-likelihood (PELL), (Chen and Sitter, 1999; Wu and Rao, 2006) is

$$l_n(\mathbf{p}) = n \sum_{i \in s} \tilde{d}_i(s) \log(p_i)$$

$$\text{where } \tilde{d}_i(s) = \frac{d_i}{\sum_{i \in s} d_i}; \quad d_i = \pi_i^{-1}$$

- Find $\{\hat{p}_i\}_{i \in s}$ to maximize the PELL

Empirical likelihood (continued)

- Calibration achieved by maximizing $l_n(\mathbf{p})$ 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 and Sitter, 2001 to case where underlying model is linear or nonlinear:

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

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

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design**
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models**
- 10 Nonprobability samples
- 11 Conclusion
- 12 References

Regression models

Valliant, 2024b review paper

- Balanced sample: match sample moments to population moments for x 's
 - Univariate results in (Royall and Herson, 1973a; Royall and Herson, 1973b)
 - Extended by Royall, 1992 to linear regression models
- Model $E_M(Y) = X\beta$; $V_M(Y) = V\sigma^2$; $V = \text{diag}(v_i)$
- When both v_i and $v_i^{1/2}$ are linear combinations of x_i , the sample that is optimal, i.e., has the **minimum error variance**, for the BLU predictor is a weighted-balanced sample:

$$\frac{N}{n} \bar{v}_U^{(1/2)} \sum_s \frac{x_i}{v_i^{1/2}} = N \bar{x}_U$$

- Left side is π -estimator if probability sample is selected with $\pi_i = n v_i^{1/2} / (N \bar{v}_U^{(1/2)})$
- Parallel to earlier, model-assisted results by (Godambe and Joshi, 1965) and (Isaki and Fuller, 1982)
- R package `sampling` will select weighted or unweighted balanced samples (Tillé and Matei, 2025)

Work on geographic balancing

- Grafström, Lundström, and Schelin, 2012 and Grafström, 2012: methods to control the geographic spread of a sample over a pop via distance between units to create small joint inclusion probabilities for nearby units, forcing samples to be well spread.
- Grafström and Tillé, 2013: method that is doubly balanced on auxiliary variables and topographical coordinates.
- All of the balancing methods are available in the R package `BalancedSampling`, (Grafström, Lisic, and Prentius, 2024).

Cutoff samples

- Single quantitative estimate with y variable closely related to an auxiliary on the frame; can lead to cutoff sample being optimal
 - $E_M(y) = \beta x$; $V_M(y) = \sigma^2 x$
 - Yorgason et al., 2011
- US EIA Monthly Natural Gas Report is a cutoff sample of about 220 large, gas producing companies. Sample companies account for 85% to 90% of all gas produced in lower 48 states. Small companies imputed.
 - In 2008 large (in-sample) natural gas well operators increased production rates faster than smaller (non-sample) operators, mostly due to shale gas extraction.
- Production for smaller producers was over-imputed \Rightarrow natural gas production overestimated for some states. Industry analysts claimed that overstated production estimates had artificially deflated market prices for natural gas.

Cutoff samples

- Problems with cutoff samples:
 - Miss economic turning points
 - Nonresponse
 - Model fit from past data may be wrong for current data
- Benedetti, Bee, and Espa, 2010 extended the cutoff idea by stratifying the population into three strata (take-all, take-some, take-none); developed algorithm for dividing pop into the strata and allocating sample to them.

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples**
- 11 Conclusion
- 12 References

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, Mercer, and Lau, 2024)
- Even best probability surveys have coverage problems, e.g., Blacks have 75-80% coverage in the US 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

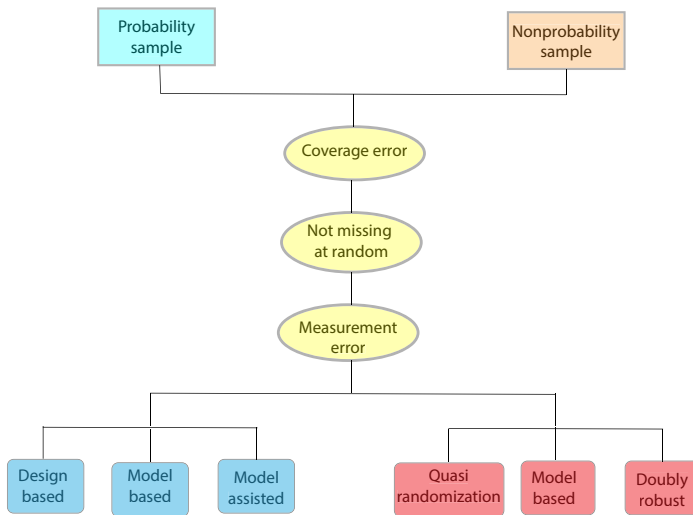
A few types of nonprobability samples

- 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 (X)

Estimation from nonprobability samples

- Elliott and Valliant, 2017 (*Stat Sci*)
- 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, and Wu, 2020)
- 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 and Raghunathan, 2017 *Stat Sci* review paper 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 and Rao, 2012; Gelman, King, and Liu, 1998)
- Mass imputation: y 's collected only in nonprob sample. Impute y 's to units in prob sample (Feder and Pfeffermann, 2015); (Marella and Pfeffermann, 2022); can be used when nonresponse is non-ignorable

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion**
- 12 References

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
⇒ 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 and Chang, 2010)

- 1 Outline
- 2 Background
- 3 Approaches to inference
- 4 Timeline
- 5 Design-based vs. Model-based
- 6 Statistical distributions
- 7 Alternatives for estimating totals and means
- 8 Models in sample design
- 9 Designing samples using models
- 10 Nonprobability samples
- 11 Conclusion
- 12 References**

-  Benedetti, R., M. Bee, and G. Espa (2010). “A framework for cut-off sampling in business survey design”. In: *Journal of Official Statistics* 26.4, pp. 651–671.
-  Binder, D. (1983). “On the Variances of Asymptotically Normal Estimators from Complex Surveys”. In: *International Statistical Review* 51, pp. 279–292.
-  Breidt, F. J. and J. D. Opsomer (2009). “Nonparametric and semiparametric estimation in complex surveys”. In: *Handbook of Statistics, Volume 29B Sample Surveys: Inference and Analysis*. Ed. by D. Pfeiffermann and C. R. Rao. Amsterdam: Elsevier. Chap. 27, pp. 103–120.
-  Brewer, K. R. W. (1963). “Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process”. In: *Australian Journal of Statistics* 5, pp. 93–105.



Chambers, R. L., A. H. Dorfman, and T. E. Wehrly (1993). “Bias Robust Estimation in Finite Populations Using Nonparametric Calibration”. In: *Journal of the American Statistical Association* 88.421. <https://doi.org/10.2307/2290722>, pp. 268–277.



Chen, J. and R. R. Sitter (1999). “A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys”. In: *Statistica Sinica* 9.2. <https://www.jstor.org/stable/24306590>, pp. 385–406.



Chen, Y., P. Li, and C. Wu (2020). “Doubly robust inference with non-probability survey samples”. In: *Journal of the American Statistical Association* 115.532. <https://doi.org/10.1080/01621459.2019.1677241>, pp. 2011–2021.



Dorfman, A. H. (1992). “Nonparametric regression for estimating totals in finite populations”. In: *Proceedings of the Survey Research Methods Section*. American Statistical Association, pp. 622–625.



Elliott, M. R. and R. Valliant (2017). “Inference for Nonprobability Samples”. In: *Statistical Science* 32, pp. 249–264.



Feder, M. and D. Pfeffermann (2015). *Statistical inference under non-ignorable sampling and non-response: An empirical likelihood approach*. Tech. rep.

<https://eprints.soton.ac.uk/378245/>. Southampton UK: University of Southampton Institutional Repository.





Gelman, A., G. King, and C. Liu (1998). “Not asked and not answered: Multiple imputation for multiple surveys”. In: *Journal of the American Statistical Association* 93.443.

<https://doi.org/10.2307/2669819>, pp. 846–857.



Ghosh, M. (2009). “Bayesian Developments in Survey Sampling”. In: *Handbook of Statistics, Volume 29B Sample Surveys: Inference and Analysis*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier. Chap. 29, pp. 153–188.

-  Ghosh, M. and G. Meeden (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall. ISBN: 0-412-98771-6.
-  Godambe, V. P. and V. M. Joshi (1965). “Admissibility and Bayes Estimation in Sampling Finite Populations - I”. In: *Annals of Mathematical Statistics* 36, pp. 1707–1723.
-  Grafström, A. (2012). “Spatially correlated Poisson sampling”. In: *Journal of Statistical Planning and Inference* 142, pp. 139–147.
-  Grafström, A., J. Lisic, and W. Prentius (2024). *BalancedSampling: Balanced and Spatially Balanced Sampling, R package version 2.1.1*. <https://CRAN.R-project.org/package=BalancedSampling>.
-  Grafström, A., N. L. Lundström, and L. Schelin (2012). “Spatially balanced sampling through the pivotal method”. In: *Biometrics* 68.2. <https://doi.org/10.1111/j.1541-0420.2011.01699.x>, pp. 514–520.



Grafström, A. and Y. Tillé (2013). “Doubly balanced spatial sampling with spreading and restitution of auxiliary totals”. In: *Environmetrics* 24, pp. 120–131.



Hansen, M. H., W. G. Madow, and B. J. Tepping (1983). “An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys”. In: *Journal of the American Statistical Association* 78.

<https://www.jstor.org/stable/2288182>, pp. 776–793.



Hartley, H. O. and J. N. K. Rao (1968). “A new estimation theory for sample surveys”. In: *Biometrika* 55, pp. 547–557.



Isaki, C. T. and W. A. Fuller (1982). “Survey Design under the Regression Superpopulation Model”. In: *Journal of the American Statistical Association* 77.377.

<https://doi.org/10.2307/2287773>, pp. 89–96.



Kennedy, C., A. Mercer, and A. Lau (2024). “Exploring the Assumption That Commercial Online Nonprobability Survey Respondents Are Answering in Good Faith”. In: *Survey Methodology* 50.

<https://www150.statcan.gc.ca/n1/pub/12-001-x/2024001/article/00013-eng.pdf>, pp. 3–21.



Kennel, T.L. and R. Valliant (2021). “Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples”. In: *Journal of Survey Statistics and Methodology* 9.4.

<https://doi.org/10.1093/jssam/smaa017>, pp. 856–890.



Kim, J. K. and J. N. K. Rao (2012). “Combining data from two independent surveys: A model-assisted approach”. In: *Biometrika* 99, pp. 85–100.



Kott, P. S. and T. Chang (2010). “Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse”. In: *Journal of the American Statistical Association* 105.491.

<https://www.jstor.org/stable/27920149>, pp. 1265–1275.



Lehtonen, R. and A. Veijanen (1998). “Logistic Generalized Regression Estimators”. In: *Survey Methodology* 24, pp. 51–55.



Little, R.J.A. (2025). *Seminal Ideas and Controversies in Statistics*. Boca Raton FL USA: Chapman & Hall.














Lohr, S.L. and T.E. Raghunathan (2017). “Combining Survey Data with Other Data Sources”. In: *Statistical Science* 32.2, pp. 293–312.









Marella, D. and D. Pfeffermann (2022). “Accounting for Non-ignorable Sampling and Non-response in Statistical Matching”. In: *International Statistical Review*.

<https://doi.org/10.1111/insr.12524>.

-  Montanari, G.E. and M.G. Ranalli (2005). “Combining Survey Data with Other Data Sources”. In: *Journal of the American Statistical Association* 100.472, pp. 1429–1442.
-  Neyman, J. (1934). “On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection”. In: *Journal of the Royal Statistical Society* 97.Part 4, pp. 558–625.
-  Olkin, I. (1987). “A Conversation with Morris Hansen”. In: *Statistical Science* 2.2. <https://www.jstor.org/stable/2245666>, pp. 162–179.
-  Opsomer, J.D. et al. (2008). “Nonparametric small area estimation using penalized spline regression”. In: *Journal of the Royal Statistical Society B* 70, pp. 265–286.
-  Royall, R. M. (1970). “On finite population sampling theory under certain linear regression models”. In: *Biometrika* 57.2, pp. 377–387.

-  Royall, R. M. (1992). "Robustness and Optimal Design Under Prediction Models for Finite Populations". In: *Survey Methodology* 18, pp. 179–185.
-  Royall, R. M. and W. G. Cumberland (1978). "Variance Estimation in Finite Population Sampling". In: *Journal of the American Statistical Association* 73.362, pp. 351–358.
-  Royall, R. M. and J. Herson (1973a). "Robust Estimation in Finite Populations I". In: *Journal of the American Statistical Association* 68.344, pp. 880–889.
-  — (1973b). "Robust Estimation in Finite Populations II". In: *Journal of the American Statistical Association* 68.344, pp. 890–893.
-  Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.
-  Smith, T. M. F. (1976). "The Foundations of Survey Sampling: A Review". In: *Journal of the Royal Statistical Society A* 139, pp. 183–204.

-  Smith, T. M. F. (1984). “Present position and Potential Developments: Some Personal Views, Sample Surveys”. In: *Journal of the Royal Statistical Society A* 147, pp. 208–221.
-  — (1994). “Sample Surveys 1975-1990: An Age of Reconciliation?” In: *International Statistical Review* 62, pp. 5–34.
-  Tillé, Y. and A. Matei (2025). *sampling: Survey Sampling*. R package version 2.11.
-  Valliant, R. (1985). “Nonlinear Prediction Theory and the Estimation of Proportions in a Finite Population”. In: *Journal of the American Statistical Association* 80, pp. 631–641.
-  — (2024a). “Hansen Lecture 2022: The Evolution of the Use of Models in Survey Sampling”. In: *Journal of Survey Statistics and Methodology* 12.2, pp. 275–304.
-  — (2024b). “Sample design using models”. In: *Survey Methodology* 50.2, pp. 149–183.



Valliant, R., A. H. Dorfman, and R. M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.



Waksberg, J. and E. Goldfield (1996). *Morris Howard Hansen 1910-1990: A Biographical Memoir*. Tech. rep. <https://nap.nationalacademies.org/read/5406/chapter/7#117>. Washington DC: National Academy of Sciences.



Wang, W. et al. (2015). “Forecasting Elections with Non-representative Polls”. In: *International Journal of Forecasting* 31, pp. 980–991.



Wu, C. and J. N. K. Rao (2006). “Pseudo empirical likelihood ratio confidence intervals for complex surveys”. In: *Canadian Journal of Statistics* 34, pp. 359–375.



Wu, C. and R. R. Sitter (2001). "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data". In: *Journal of the American Statistical Association* 96.453.

<http://www.jstor.org/stable/2670358>, pp. 185–193.
ISSN: 01621459.



Yorgason, D. et al. (2011). *Cutoff Sampling in Federal Surveys: An Inter-Agency Review*. Tech. rep.

<https://www.bls.gov/osmr/research-papers/2011/pdf/st110050.pdf>. Washington DC: Bureau of Labor Statistics.