The Analysis of Census and Survey Data: History, Current Approaches, and Research Topics

Roderick J. Little

UNIVERSITY OF MICHIGAN

Overview

- What distinguishes "survey sampling" from other branches of statistics?
- Design-based versus model-based inference
- Current orthodoxy: design-model compromise
  - Strengths and drawbacks
- A possible future paradigm
  - Calibrated Bayes
- A few topics of current research interest
- Decennial Census, in two slides

Distinguishing Statistical Features of Survey Sampling

- A simple and brilliant idea: simple random sampling
- Study of *complex sample designs*: designs that go beyond simple random sampling, including features like stratification, weighting and clustering
  - Simple random sampling, though simple, is not optimal or even practical in many settings
- Many practical real-world sampling issues: sampling frames, making use of administrative information, alternative modes of survey administration
- Side bar: current statistics training focuses on analysis methods, neglects important design issues like this

Distinguishing Statistical Features of Survey Sampling

- Major interest in *descriptive* inference about *finite population quantities*, as opposed to parameters of models (though *analytical inference* for parameters can also be of interest).
- Prevailing orthodoxy is *design-based* (randomization) inference: survey outcomes are treated as fixed quantities, and statistical uncertainty derives from the probability distribution that determines sample selection
  - Deep-seated distrust of models
- As a modeler who sees complex designs in other fields (e.g. epidemiology), I personally view survey sampling as not that different from other fields of statistics

Probability Sampling

- Definition of probability sampling:
  - every sample has a known probability of being selected
  - every individual in the population has a positive probability of being selected
- Initially, probability sampling was equated with its basic form, simple random sampling (SRS)
  - Every sample of size *n* has *equal* chance of being selected, hence an equal probability of selection method (*epsem*)
  - Samples of size other than *n* have no chance of being selected
  - With and without replacement

Complex Designs

- Neyman's (1934) paper considered stratified sampling, with selection probabilities varying across strata in an optimal way ("Neyman allocation")
- Helped to set the stage for extensions to cluster sampling, multistage sampling, greatly extending the practical feasibility and utility of probability sampling in practice
- E.g. simple random sampling of people in the US is not feasible – we do not have a complete list of everyone in the population from which to sample
- Work of Mahalanobis, Hansen, Cochran, Kish, ….

Is Probability Sampling Optimal?

- Simple random sampling (or equal probability sampling in general) is an all-purpose strategy for selecting units to achieve representativeness "on average"
  - compare with randomized treatment allocation in clinical trials
- However, statisticians like optimal properties, and SRS is very suboptimal for some specific purposes…
- E.g. if distribution of *X* is known in population, and objective is slope of linear regression of *Y* on *X*, it's obviously much more efficient to sample equally at the two extreme values of *X* – this minimizes the variance of the LS slope (Royall 1970)
- But this is not a probability sample– intermediate values of *X* have zero chance of selection!
- For linear regression through origin, optimal design is cut-off sampling, which is still applied in some business surveys

More on Neyman (1934)

- Neyman (1934) is even more celebrated for introducing confidence intervals as an approach to inference from a probability sample
- Foreshadows the "design-based" approach to survey inference, where population values are fixed and inferences are based on the randomization distribution in the selection of units …
- … Although Neyman never clearly states that he regards population values as fixed, and his references to Student's t distribution suggest that he had a distribution in mind
- This foreshadows the other topic of controversy, concerning design-based vs model-based inference

Link with Estimation

- *Design-based* inference: population values are fixed, inference is based on probability distribution of sample selection. Obviously this assumes that we have a probability sample (or "quasi-randomization", where we pretend that we have one)
- *Model-based* inference: survey variables are assumed to come from a statistical model: probability sampling is not the basis for inference, but useful for making the sample selection *ignorable*. (see e.g. Gelman et al., 2003)

Design vs Model-based Survey Inference

- Two main variants of model-based inference:
  - *Superpopulation models*: Frequentist inference based on repeated samples from a "superpopulation" model
  - *Bayes*: add prior distribution for parameters; inference about finite population quantities or parameters based on posterior distribution
- A fascinating part of the more general debate about frequentist versus Bayesian inference in statistics at large:
  - Design-based inference is inherently frequentist
  - Purest form of model-based inference is Bayes

Design-Based Survey Inference

$Y = (Y_1, ..., Y_N)$ = population values (fixed); Z = design variables

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

$I = (I_1, ..., I_N)$ = Sample Inclusion Indicators (random)

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

Y_{inc} = part of Y included in the survey

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

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

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

Choice of $\hat{q}$

Seek good design-based properties:

design unbiasedness: $E(\hat{q} \mid Y) = Q$ (too strong)

Or weaker: *design consistency*: $\hat{q} \to Q$ as sample size gets large

It is natural to seek an estimate that is *design - efficient*

However, this kind of optimality is not possible without a model (Horvitz and Thompson 1952, Godambe 1955)

There are many choices of design-consistent estimates ...

Many survey estimates are motivated by *implicit* models:

Regression model $\rightarrow$ regression estimator

Ratio model $\rightarrow$ ratio estimator, etc.

Example 1: Mean from Simple Random Sample

$Q = \bar{Y}$, population mean

$\hat{q} = \bar{y}$, sample mean, unbiased for $\bar{Y}$

$\hat{V} = (1 - n/N)s^2/n$, s^2 = sample variance

$\left(\bar{y} - 1.96\sqrt{\hat{V}}, \bar{y} + 1.96\sqrt{\hat{V}}\right)$ = 95% CI for $\bar{Y}$

- Note: lacks t correction for small samples (t based on a normal model for the outcomes)

Example 2: Design Weighting

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

$\pi_i = E(I_i \mid Y)$ = inclusion probability, by assumption > 0

$\hat{t}_{\mathrm{HT}} = \sum_{i=1}^{N} I_i Y_i / \pi_i$, unbiased for T, HT =Horvitz-Thompson

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

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

Weighting for a Stratified Sample

Weighting sampled cases in stratum j by $w_j = N_j / n_j$ leads to the stratified mean:

$$\bar{y}_{\mathrm{HT}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_h} w_j y_{ij} = \bar{y}_{\mathrm{st}} = \sum_{j=1}^{J} P_j \bar{y}_j,$$

$P_j = N_j / N, \bar{y}_j =$ sample mean in stratum j

Estimate of variance: $\hat{\sigma}^2_{\mathrm{st}} = \sum_{j=1}^{J} P_j^2 (1 - f_j) s_j^2 / n_j,$

$f_j = n_j / N_j, s_j^2 =$ sample variance in stratum j

95% Confidence interval for mean: $\bar{y}_{\mathrm{st}} \pm 1.96\hat{\sigma}_{\mathrm{st}}$

Weighting for PPS Samples

- A variable X measuring size of units is available for all units in the population
- Probability Proportional to Size (PPS) sampling: sample unit i with $X = x_i$ with probability

$$\pi_i = cx_i, c \text{ chosen to yield desired sample size}$$

- Clever and simple ways of implementation from lists of population units, with cumulated ranges of size
- Many applications, e.g. auditing
- HT estimation is standard:

$$\hat{T} = \sum_{i=1}^{N} I_i y_i / \pi_i = c \sum_{i=1}^{N} I_i y_i / x_i$$

A Tongue-in-Cheek Example of Weighting Gone Wrong: Basu's Elephants (Basu 1971)

$(y_1,...,y_{50}) =$ weights of $N = 50$ elephants

Objective: $T = y_1 + y_2 + ... + y_{50}$. 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:

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

HT estimator is unbiased on average but always crazy!

Circus statistician loses job and becomes an academic

Why does HT estimate fail?

HT estimator obtained from predictions for following "HT model":

$y_i / \pi_i \sim \text{Nor}(\mu, \sigma^2)$

If this model is not plausible (as in Basu's example), resulting estimator is highly inefficient

Moral: examine estimators from prediction perspective

Generalized Regression Estimation

- Create predictions of non-sampled units from a model, and "calibrate them" using weighted residuals from the model

$$\hat{T}_{\text{GREG}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} I_i (y_i - \hat{y}_i)/\pi_i, \hat{y}_i = \text{ model prediction}$$

- Uses the model to increase efficiency, but is design consistent even if model is misspecified
- Design-based, model-assisted
- Foreshadows "double robustness" in mainline statistics literature – recent interest in alternative weights for residuals (Cao, Tsiatis & Davidian 2009 Biometrika)

Model-Based Approaches

- In *model-based*, or *model-dependent*, approaches, models are the basis for the entire inference: estimator, standard error, interval estimation
- Two variants:
  - Superpopulation Modeling
  - Bayesian (full probability) modeling
- Common theme is to "infer" or "predict" about non-sampled portion of the population, conditional on the sample and model
- Focus here on Bayesian variant

Bayes Inference for Surveys

Model: $p(Y \mid Z)$ = prior distribution for Y

Data: Y_{inc} = sampled values of Y; Z = design variables

Inference about $Q = Q(Y, Z)$ are based on

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

In particular:

Estimate is posterior mean: $\hat{q} = E(Q \mid Y_{\text{inc}}, Z)$

Standard error is posterior sd: $\sqrt{Var(Q \mid Y_{\text{inc}}, Z)}$

95% posterior probability interval plays role

of confidence interval (with a simpler interpretation)

Parametric Models

Usually the prior distribution is specified via *parametric* models:

$$p(Y \mid Z) = \int p(Y \mid Z, \theta)\, p(\theta \mid Z) d\theta$$

$p(Y \mid Z, \theta)$ = parametric model, as in superpopulation approach

$p(\theta \mid Z)$ = prior distribution for θ

Inference about θ is then obtained from its posterior distribution, computed via Bayes' Theorem:

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

$$L(\theta \mid Y_{\text{inc}}, Z) = \text{Likelihood function}$$

That is: Posterior = Prior x Likelihood

Ex 1: Normal Model Bayesian analysis for Simple Random Sample

$$Y_i \sim_{\text{iid}} \text{Nor}(\mu, \sigma^2); i = 1, 2, ..., N$$

Flat "Jeffreys" prior: $\pi(\mu, \log \sigma^2) \propto \text{const.}$

simple random sample results in $Y_{\text{inc}} = (y_1, ..., y_n)$

A 95% Highest Posterior Density interval for $\bar{Y}$ is

$$\bar{y} \pm t_{n-1,1-\alpha/2} \sqrt{\frac{(1-f)\hat{\sigma}^2}{n}}$$

Equivalent to design-based interval in large samples

Recovers t correction for estimating variance

(Design-based programs often incorporate a t correction but justification is ad-hoc: see Valliant and Rust, 2010)

Bayes Inference for a Mean from a Stratified Sample

- Consider a model that includes stratum effects:

$$[y_i \mid z_i = j] \sim_{\text{ind}} \text{Nor}(\theta_j, \sigma_j^2)$$

- For simplicity assume the flat Jeffreys' prior:

$$p(\theta_j, \log \sigma_j^2 \mid Z) \propto \text{const.}$$

Posterior mean is stratified mean $\bar{y}_{\text{st}}$

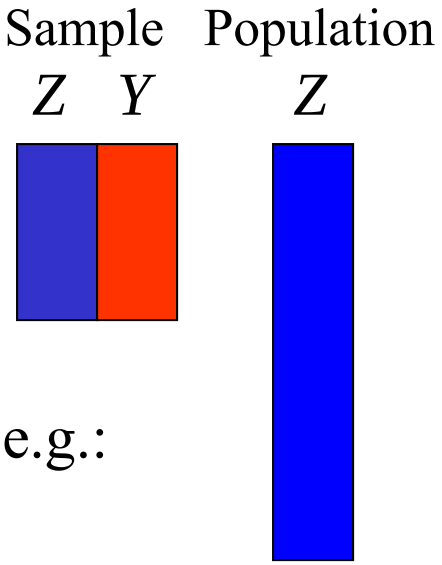
Posterior variance is stratified variance $\hat{\sigma}^2_{\text{st}}$

Posterior distribution is mixture of t distributions

HPD interval same as design-based confidence interval in large samples, but provides a t correction for estimating variances

Ex 3. One Continuous (Post)stratifier Z

$$\bar{y}_{\text{HT}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i / \pi_i\right); \pi_i = \text{ selection prob}$$



A modeling alternative to the HT estimator is create predictions from a more robust model relating Y to Z:

$$\bar{y}_{\text{mod}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{y}_i\right), \ \hat{y}_i \text{ predictions from model, e.g.:}$$

$y_i \sim \text{Nor}(\beta\pi_i, \sigma^2\pi_i^2)$; leads to $\bar{y}_{\text{HT}}$

$y_i \sim \text{Nor}(S(\pi_i), \sigma^2\pi_i^k)$; $S(\pi_i)$ = penalized spline of Y on Z

Simulations in Zheng and Little (2005) suggest better RMSE, confidence coverage for spline model compared with design-based approaches

The Design-Based Perspective- Pros

- Avoids dependence on a model for the population values
  - Models can help the choice of estimator, but the inference remains design-based, hence somewhat nonparametric
  - Design-based properties like design consistency confer robustness, since they apply regardless of the validity of a model
  - Weighting-based approaches can be applied uniformly to a set of outcomes, simplifying the computing

The Design-Based Perspective- Cons 1

- Inference is 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
  - High proportion of available information is now not based on probability samples (e.g. internet)
- Theory is basically asymptotic
  - Limited tools for small samples, e.g. small area estimation

Cons 2: Ancillarity Angst

- Thorny issues concerning what to condition on in "reference set" for repeated sampling

  E.g. consider regression estimate of mean $\bar{y}_{\text{reg}} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$,

  $\bar{y}$, $\bar{x}$ = sample means, $\bar{X}$ = population mean, $\hat{\beta}$ = slope

  Design-based SE averages over sampling distribution of $\bar{x}$,

  mixes in (irrelevant?) samples with other values of $\bar{x}$

  Model-based SE conditions on X, is higher when $|\bar{X} - \bar{x}|$ large

  (Cumberland and Royall 1988)

- Basic issue is whether to condition on ancillary statistics – if taken seriously, leads to likelihood principle, which design-based inference violates – without a model for predicting non-sampled cases, the likelihood is basically useless.

Design-Based Approach Needs Models

- Although not explicitly model-based, models are needed to motivate the choice of estimator
  - E.g. the HT estimator assumes an implicit HT model that y_i / π_i are "exchangeable" (iid conditional on parameters)
  - If implicit models are unreasonable, then the resulting inferences can be very poor in moderate samples (Basu's elephant being an extreme case)
- So models are needed in design-based approach, as in the "model-assisted" paradigm discussed above

The Model-Based Perspective- Pros

- Flexible, unified approach for all survey problems
  - Models for nonresponse and response errors, small area models, combining data sources
- Moves survey sample inference closer to mainstream statistics
  - Other disciplines like econometrics, demography, public health, rely on statistical modeling
- Models are now available that incorporate sample design features
- Bayesian approach is not asymptotic, provides better small-sample inferences
- Probability sampling justified as making sampling mechanism ignorable, improving robustness

The Model-Based Perspective- Cons

- More explicit dependence on the choice of model, which has subjective elements
- Concerns of lack of robustness to model misspecification
- Models needed for all outcomes – need to understand the data, and potential for more complex computations

The Design-Model Compromise

- Design-based for large samples, descriptive statistics
  - But may be *model assisted*, e.g. regression calibration: model estimates adjusted to protect against misspecification, (e.g. Särndal, Swensson and Wretman 1992).
- Model-based for small area estimation, nonresponse, time series,…
- Attempts to capitalize on best features of both paradigms, but at the expense of "inferential schizophrenia" (Little 2012)?
- Two examples of negative consequences follow …

1. Statistical Standards and the Bayes/Frequentist Gorilla

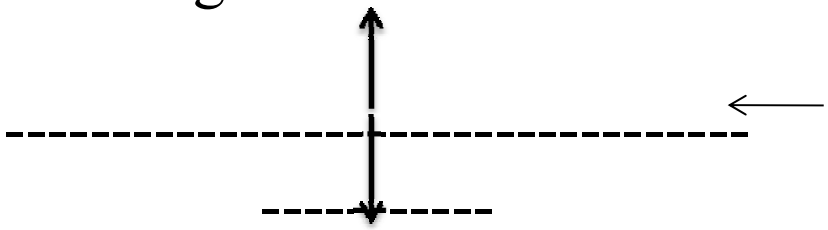
Follow my (frequentist) statistical standards

Why? I am an economist, I build models!

2. When Is An Area "Small"?

n-ometer

Design-based inference



n_0 = "Point of inferential schizophrenia"

Model-based inference

How do I choose n_0?

If $n_0 = 35$, should my entire statistical philosophy and inference be different when n=34 and n=36?

n=36, CI: [ ] (wider since based on direct estimate)

n=34, CI: [ ] (narrower since based on model)

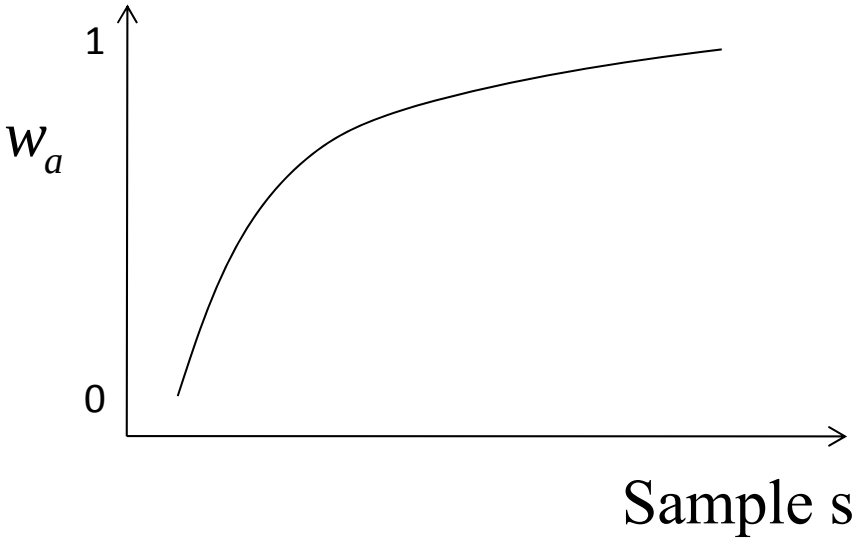
Multilevel (Hierarchical Bayes) Models

$$\tilde{\mu}_a = w_a \bar{y}_{\pi a} + (1 - w_a)\hat{\mu}_a$$

Model estimate

Direct estimate

n - o m e t e r



Bayesian multilevel model estimates borrow strength increasingly from model as *n* decreases

Example: SAIPE Project

- Objective: Provide estimates of poverty for various age groups and median household income for all *states*, *counties*, and *school districts* in the U.S.
- Problem: Direct survey estimates (from CPS or, later, ACS) too unreliable for many areas
  - CPS sample small for most states; no sample in ≈ 2/3 counties
  - ACS (single year) sample small for many counties and most school districts.
- Solution: U.S. Census Bureau uses small area model to integrate survey data with data from admin records (IRS, SNAP program) and previous census long form.

Posterior Variances from State Model for 2004 CPS 5-17 Poverty Rates

Results for four states

State	n_i	v_i	Var(Y_i|data)	approx. wt. on y_i in E(Y_i|data)
CA	5,834	1.1	0.8	.61
NC	1,274	4.6	2.0	.28
IN	904	8.1	2.0	.18
MS	755	12.0	3.9	.13

Calibrated Bayes Models for Surveys Should Incorporate Sample Design Features

- Seek robust models with good repeated sampling properties
  - Models that ignore features like survey weights are vulnerable to misspecification
- But Bayesian models that incorporate design features can yield inferences with good design-based properties:
  - Capture design weights as covariates in the prediction model (e.g. Gelman 2007)
  - Clustering via hierarchical random effects models
  - Example in Appendix slides

Some current survey research topics

- Generalized Regression vs Robust Modeling
- Combining information from diverse data sources, including administrative records
- Small area estimation: Bayes/EB, inferences with close to nominal coverage, reconciling small area estimates with direct estimates at higher levels of aggregation
- Methods for confidentiality protection in public use files: including fully synthetic data
- Responsive Design: tailoring alternative modes of data collection to optimize response, limit multiple call-backs of nonrespondents
- Use of "meta-data" to improve estimates

Decennial Census Research

- Several countries have moved or are moving towards an administrative census, using administrative records rather than traditional mail-out, mail-back plus interviews with households.
- This is much cheaper, but works better for "de jure" rather than "de facto" population counts
- U.S. does a "de facto" census – counts everyone who is there on Census day, regardless of legal status.
- Extremely complex operation: not just how many people, but where they are located, is important.
- Suffers from no explicit recognition of "error-cost" tradeoff
- Sampling for non-response follow-up is off the table
  - But coverage measurement survey is used to measure quality

Decennial Census Research

- Vast topic: would take another lecture (or a course) to cover this topic
- Some major areas of BoC research interest for 2020 are
- use of administrative records to replace call-backs for non-response follow-up (which costs $$$)
  - Confidentiality "big brother" versus "wasting people's time"
  - Combining information from multiple administrative data sources, with differential gaps and potentially conflicting information: alternatives to straight substitution, simple enough for the massive Census environment
- use of internet, handheld devices to transmit and record information – dealing with multiple platforms, threats to confidentiality
- continuous, targeted updating of Master Address File

References 1.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, eds. V.P. Godambe & D.A. Sprott), Toronto: Holt, Rinehart & Winston, 203-242.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *JASA*, 57, 269-326.

Box, G.E.P. (1980), Sampling and Bayes inference in scientific modeling and robustness (with discussion), *JRSSA*, 143, 383-430.

Cao, W., Tsiatis, A. and Davidian, M. (2009), Improved efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 96, 723-734.

Cumberland, W.G. & Royall, R.M. (1988). Does simple random sampling provide adequate balance? *JRSSB*, 50(1), 118-124.

Fay, R. & Herriot, R. (1979). Estimates of income for small places: an application of James-Stein procedures to Census data. *JASA*, 74, 366, 269-277.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.*, 22, 2, 153-164 (with discussion and rejoinder).

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003), *Bayesian Data Analysis*, 2nd. edition. New York: CRC Press.

References 2.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *JRSSB*, 17, 269-278.

Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *JASA*, 78, 776-793.

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *JASA*, 47, 663-685.

Kish, L. & Frankel, M. (1974). Inference from complex samples. *JRSSB*, 36, 1-37.

Little, R.J.A. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Am. Statist.*, 60, 3, 213-223

_____ (2012). Calibrated Bayes: an alternative inferential paradigm for official statistics (with discussion and rejoinder). *JOS*, 28, 3, 309-372.

_____ (2013). Survey Sampling: Past Controversies, Current Orthodoxies, and Future Paradigms. To appear in *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, X. Lin, D. L. Banks, C. Genest, G. Molenberghs, D.W. Scott, and J.-L. Wang, eds. CRC Press.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *JRSS*, 97, 558-606.

References 3.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M. & Herson, J.H. (1973). Robust estimation in finite populations, I and II. *JASA*, 68, 880-893.

Rubin, DB (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Annals of Statistics* 12, 1151-1172.

Särndal, C.-E., Swensson, B. & Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag: New York.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *JOS*, 24**,** 495–506.

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

Valliant, R. & Rust, K. F. (2010). Degrees of Freedom Approximations and Rules-of-Thumb. *JOS*, 26, No. 4, 585–602.

Zheng, H. & Little, R.J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *JOS*, 21, 1-20.

Appendix Example: Ratio Model Revisited

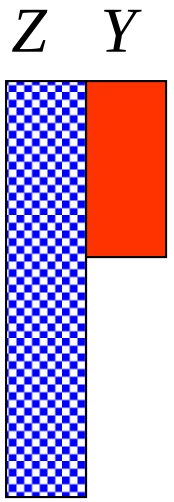
Ratio model for simple random sampling:

$$(Y_i / z_i \mid \beta, \sigma^2) \sim \text{ind } N(\beta_0, \sigma^2 / z_i),\ i = 1, 2, ..., N$$

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

$$\rightarrow \hat{\bar{Y}}_{\text{rat}} = (\bar{y} / \bar{z})\bar{Z},\ \text{where}$$

$\bar{y}$ and $\bar{z}$ sample means, $\bar{Z}$ population mean



Example: Ratio Model Revisited

Hansen, Madow and Tepping (HMT, 1983) compare the ratio model estimator $\bar{Y}_{\text{rat}}$ with the combined ratio estimator (CRE) $\bar{Y}_{\text{CRE}} = (\bar{y}_w / \bar{z}_w)\bar{Z}$,
where $\bar{y}_w, \bar{z}_w$ are weighted by the sampling weights.

HMT show that the ratio estimate can have poor RMSE and, with a model estimate of variance, poor confidence coverage, even with small deviations from the model that are not detectable by model checks.

Calibrated Bayes: need to condition on strata that define the weights, as in separate ratio estimator

Stratified Ratio Model

Extend ratio model by adding fully-observed stratum indicators x_i:

$$(Y_i / z_i \mid x_i = h, \beta, \sigma^2) \sim_{\text{ind}} N(\beta_h, \sigma_h^2 / z_i),\ i = 1, 2, ..., N$$

Posterior mean under flat prior is the separate ratio estimator:

$$\rightarrow \hat{\bar{Y}}_{\text{sep}} = \sum_{h=1}^{H} P_h (\bar{y}_h / \bar{z}_h) \bar{Z}_h$$

$\bar{y}_h$ and $\bar{z}_h$ sample means in stratum h

$P_h, \bar{Z}_h$ population proportion and mean in stratum h

Bayes Alternatives to Combined Ratio Estimate

$$(Y_i / z_i \mid x_i = h, \beta_h, \sigma^2) \sim_{\text{ind}} N(\beta_h, \sigma_h^2 / z_i),\ i = 1, 2, ..., N$$

$$\beta_h \sim_{\text{ind}} N(\beta, \tau^2)$$

Shrinks $\hat{\bar{Y}}_{\text{sep}}$ towards simple ratio estimator

Retains design consistency of $\hat{\bar{Y}}_{\text{sep}}$ with better small-sample properties

Or put more structure in prior for β_h (see Sedransk 2008)

Also could put more structured prior on $\{\sigma_h^2\}$