Answers to criticisms of using design-based methods for inference

David Binder
Statistics Canada (retired)
November 7, 2008

Quebec Inter-University Centre for Social Statistics

OUTLINE

1. What do we mean by design-based or model-based inference?
2. Example (informative sampling)
3. The rationale for design-based inference under the model-design-based randomization
4. Some problematic cases
5. An artificial example
6. Modeler's criticisms
7. Closing remarks

1. What do we Mean by Design-based or Model-based Inference?

We assume we have observed a sample obtained from a complex survey design, where the units sampled are taken from a finite population.

Most important are:

1. The target population(s) vs. the survey population
2. The randomization distribution
   a) Model-based framework
   b) Design-based framework
   c) Model-design-based framework

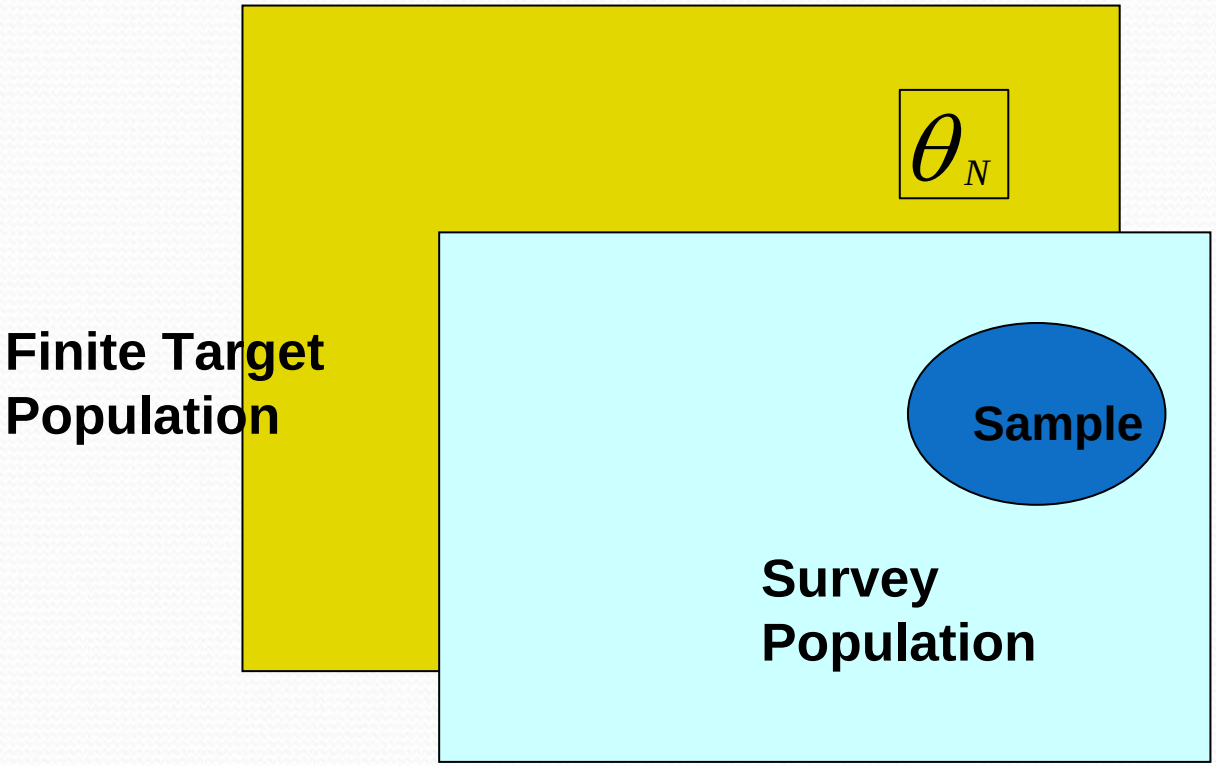
1. What do we mean by design-based or model-based inference?

The target population is the population about which the researcher wishes to make conclusions.

The survey population consists of all the units that are eligible for selection through the frame and survey design being used.

Finite Target Population vs. Survey Population

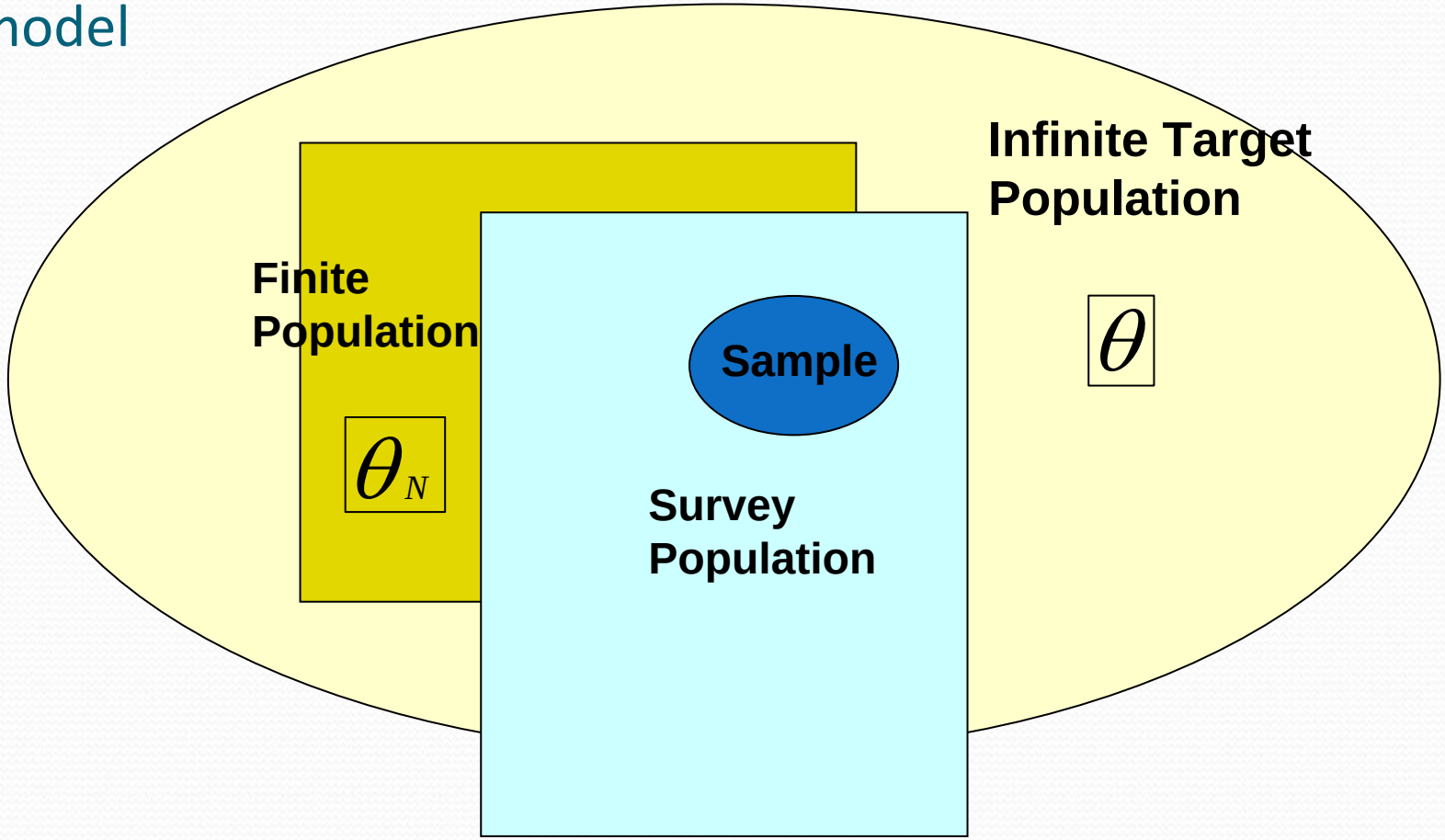
Quantities of interest are finite population quantities



The differences between the survey population and the target population has led to some criticisms of design-based methods; for example, for cut-off sampling.

Infinite Target Population

Quantities of interest are related to parameters of a statistical model

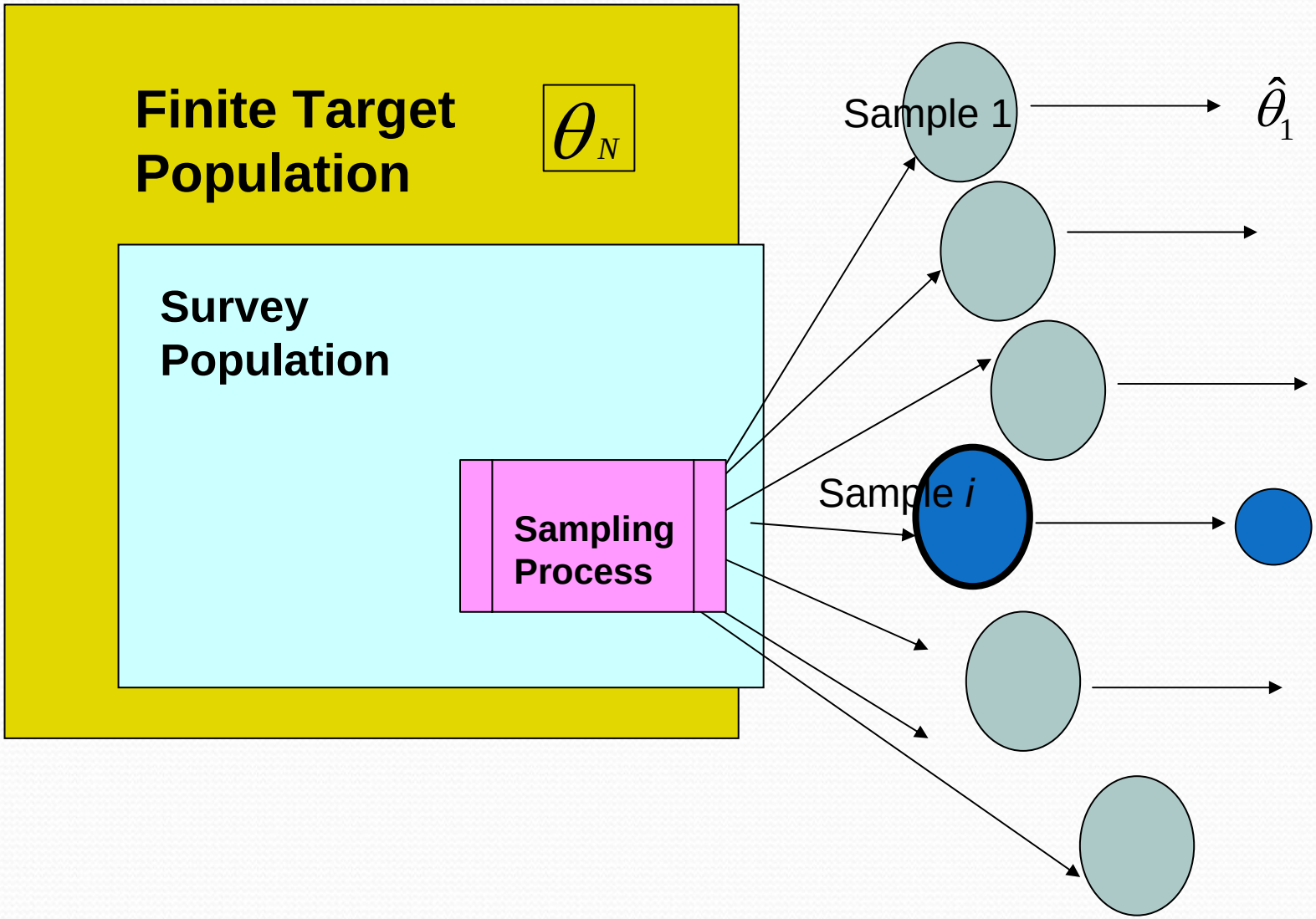


The Randomization Distribution

- For statistical inference you are interested not only what is actually observed in the selected sample, but also what could have been observed had other samples been selected under the particular randomization framework
- Of primary interest is the distribution of the estimates under hypothetical random repetitions
  - The distribution of the estimates depends on whether or not a statistical model is presumed to have generated the values of a finite population (***ξ*-randomization**)
  - The distribution of the estimates may or may not be affected by the sample design (***p*-randomization**)

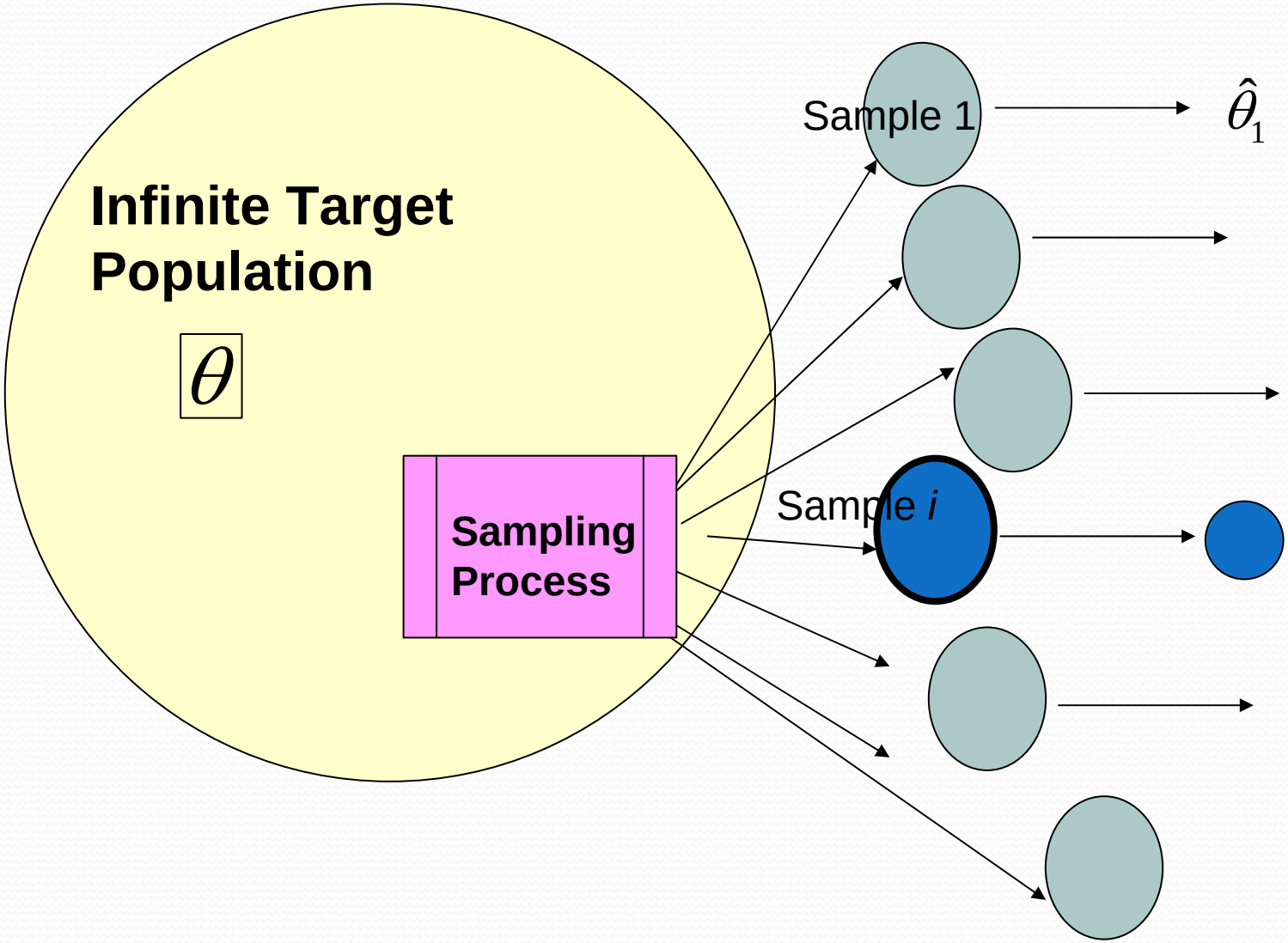
The Randomization Distribution

Finite target population – **design-based** (*p*-randomization)



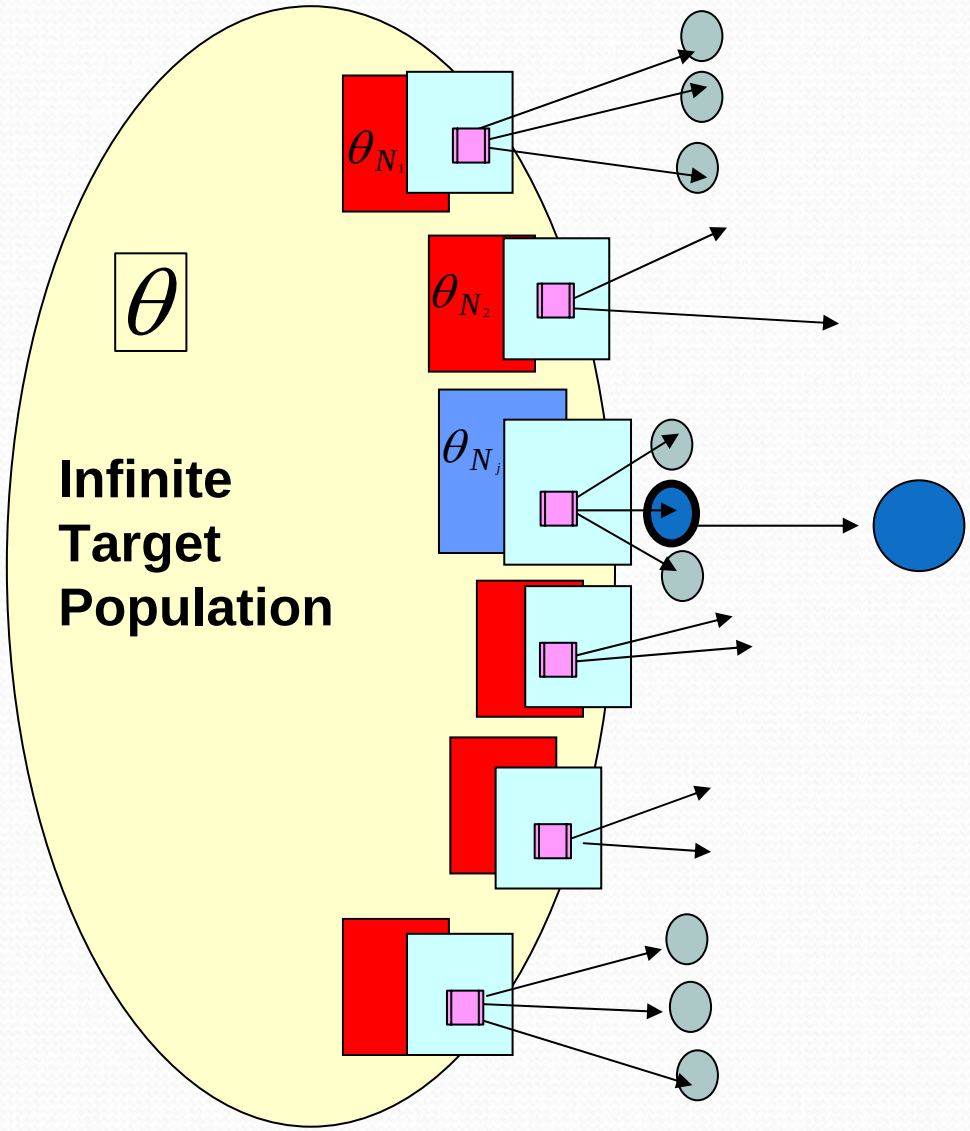
The Randomization Distribution

Infinite target population – **model-based** (ξ-randomization)



The Randomization Distribution

Infinite target population – **model-design-based** (ξp-randomization)



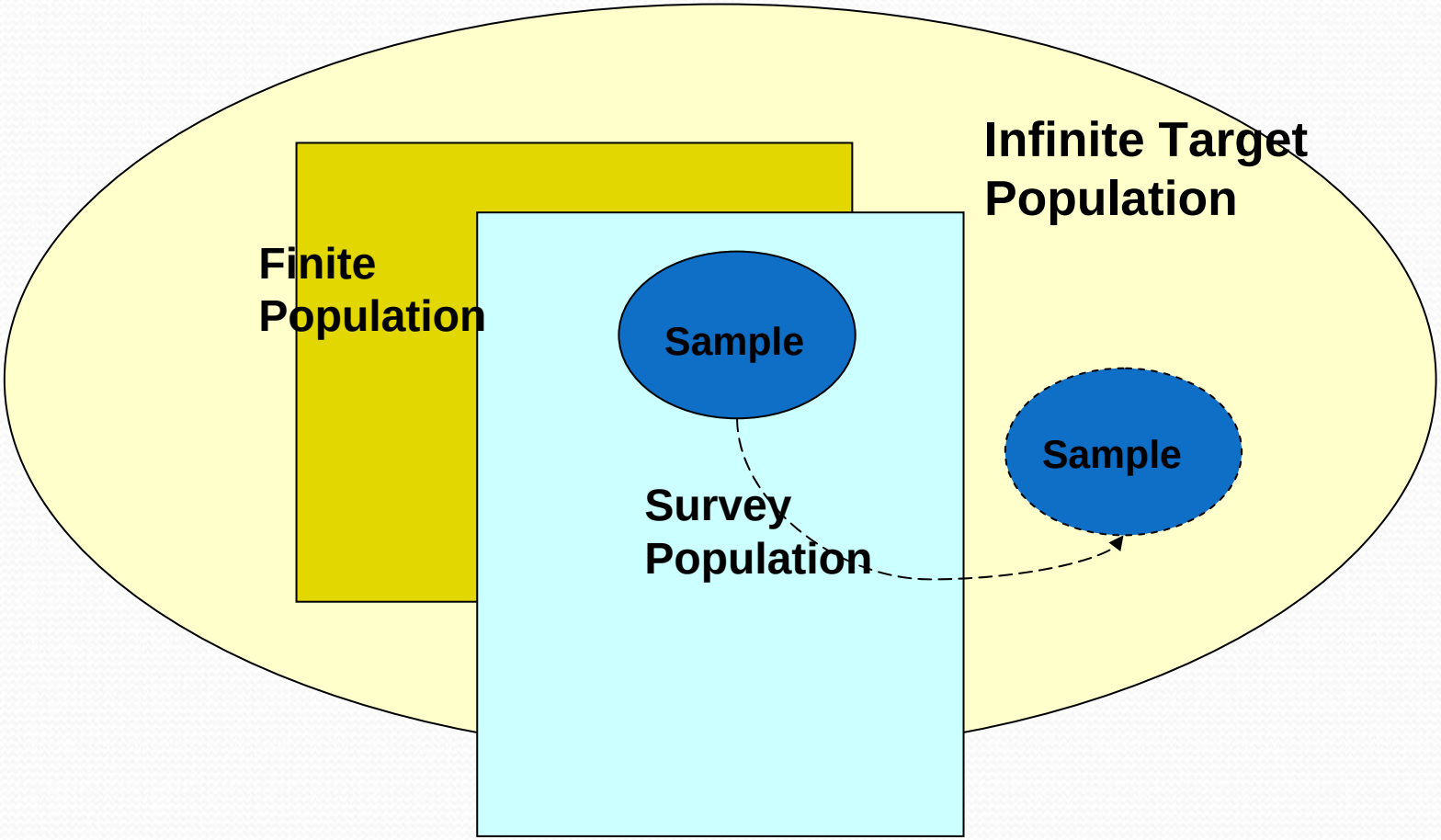
- **Finite populations are generated from the infinite population.**
- **Randomization for estimator is based on both the model and the design.**

Informative Sampling Chambers (2003, 2004)

- Broadly speaking – sampling is informative if distributions of population and observed sample values are different –More precisely, sampling is informative (non-informative) given some information if the two corresponding probabilities are unequal (equal).

- Informative/non-informative status depends on what is being conditioned on (e.g. design variables, sampling process outcome, response/nonresponse status).

Infinite target population – informative sampling



When the sample can be assumed to have been generated from the model, the sampling is not informative. Otherwise, it is informative.

2. Example (informative sampling)

Population values of *Y* are *iid* ~ one parameter exponential

$$f(y;\theta) = \theta \exp(-\theta y)$$

Target parameter of interest: $\mu = E[y] = \theta^{-1}$

Sample selected with known inclusion probabilities: $\pi_i = \frac{nY_i}{N\bar{Y}}$

Complete response, no other auxiliary information

2. Example (informative sampling)

Value $\bar{Y}$ is deducible from sample values of Y and their inclusion probabilities

$$\hat{\mu}_{ML} = \bar{Y}; \quad Var_{\xi p}[\hat{\mu}_{ML}] = \frac{\mu^2}{N}$$

Comment:

Full Information MLE becomes difficult in secondary analysis under informative sampling because much of the information desired for this modelling exercise is not available.

2. Example (informative sampling)

Pseudo-Likelihood Approach

Design-Based Methodology

$$\hat{\mu}_{PL} = \frac{\sum I_i \pi_i^{-1} Y_i}{\sum I_i \pi_i^{-1}}; \quad Var_{\xi p}[\hat{\mu}_{PL}] = o(n^{-1})$$

Clearly $\hat{\mu}_{PL}$ is suboptimal by a long shot!. We know the inclusion probabilities π_i, so we know $\overline{Y}$, which is the ML estimator of μ based on the entire population!

The PL estimator is approximately unbiased in large populations

2. Example (informative sampling)

Sample Likelihood Approach

Approximates distribution of sample values $\{Y_i;\ I_i = 1\}$ as a function of the population distribution and the sampling weights.

Chambers showed that SL estimator is approximately unbiased under size-biased sampling for the one-parameter exponential model in large populations, and that the variance reduction over the PL estimator could be very substantial!

2. Example (informative sampling)

In general, in terms of efficiency, ML dominates SL, which in turn dominates PL.

Both ML and SL require that the informative sampling mechanism be modelled.

Both SL and PL require values of the actual sample inclusion probabilities.

3. The Rationale for Design-based Inference under the Model-design-based Randomization

We use as a motivating example the standard linear regression model.

$$Y_i = \mathbf{x}_i'\boldsymbol{\theta} + \varepsilon_i \quad \text{, where } \varepsilon_i \sim IN(0, \sigma^2).$$

If the complete finite population could be observed, the usual estimating equations for estimating $\boldsymbol{\theta}$ are given by

$$\mathbf{U}_N(\boldsymbol{\theta}_N) = \sum_{i=1}^{N} \mathbf{x}_i (Y_i - \mathbf{x}_i'\boldsymbol{\theta}_N) = \sum_{i=1}^{N} \mathbf{u}_i(\boldsymbol{\theta}_N) = \mathbf{0},$$

where $\boldsymbol{\theta}_N$ is the finite population parameter associated with model parameter $\boldsymbol{\theta}$.

3. Design-based Rationale

It is important to note that

$$\mathbf{E}_{\xi}\mathbf{u}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{\xi}[\mathbf{x}_i(Y_i - \mathbf{x}_i'\tilde{\boldsymbol{\theta}})] = \mathbf{x}_i\mathbf{x}_i'(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

so that when $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, we have $\boxed{\mathbf{E}_{\xi}\mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0}.}$

However, if the true model is

$$Y_i = \mathbf{x}'_i\,\boldsymbol{\theta} + \mathbf{z}'_i\,\boldsymbol{\gamma} + \varepsilon_i,$$

then $\mathbf{E}_{\xi}\mathbf{u}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{x}_i\mathbf{x}_i'(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \mathbf{x}_i\mathbf{z}_i'\boldsymbol{\gamma},$

so that when $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, we have $\boxed{\mathbf{E}_{\xi}\mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{x}_i\mathbf{z}_i'\boldsymbol{\gamma}.}$

3. Design-based Rationale

We now define the model-based and the design-based estimators for $\boldsymbol{\theta}$. The model-based *maximum likelihood estimator* $\hat{\boldsymbol{\theta}}$ is the solution to

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} I_i \mathbf{x}_i (Y_i - \mathbf{x}_i' \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} I_i \mathbf{u}_i (\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

On the other hand, the design-based the *pseudo-maximum likelihood estimator* $\hat{\boldsymbol{\theta}}_p$ is the solution to

$$\hat{\mathbf{U}}_p(\hat{\boldsymbol{\theta}}_p) = \sum_{i=1}^{N} \pi_i^{-1} I_i \mathbf{x}_i (Y_i - \mathbf{x}_i' \hat{\boldsymbol{\theta}}_p) = \sum_{i=1}^{N} \pi_i^{-1} I_i \mathbf{u}_i (\hat{\boldsymbol{\theta}}_p) = \mathbf{0}.$$

(For simplicity here, we consider only a single-stage sample design with no weight adjustments.)

3. Design-based Rationale

This simple regression example can be easily extended to more general estimating equations, covering a wide range of estimators; for example,

- Generalized Linear Models
- Quasi-likelihood
- Generalized Estimating Equations for longitudinal data
- M-estimators

As well, Cox PH models can be adapted to this approach.

3. Design-based Rationale

To derive variances for large sample situations, we write the “kernel” of the estimating equations as

$$\mathbf{u}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{u}_i(\boldsymbol{\theta}) + \left[\left. \frac{\partial \mathbf{u}_i(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right|_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} \right] (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathbf{r}_i(\tilde{\boldsymbol{\theta}})$$

For the linear regression case, this is simply

$$\mathbf{u}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{x}_i(Y_i - \mathbf{x}_i'\tilde{\boldsymbol{\theta}}) = \mathbf{x}_i(Y_i - \mathbf{x}_i'\boldsymbol{\theta}) - \mathbf{x}_i\mathbf{x}_i'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

3. Design-based Rationale

We now consider the properties of $\boldsymbol{\theta}_N$ the finite population parameter associated with model parameter $\boldsymbol{\theta}$. Since

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{u}_i(\boldsymbol{\theta}_N) = \sum_{i=1}^{N} \left[\mathbf{x}_i(Y_i - \mathbf{x}_i'\boldsymbol{\theta}) - \mathbf{x}_i\mathbf{x}_i'(\boldsymbol{\theta}_N - \boldsymbol{\theta})\right]$$

we have

$$\boldsymbol{\theta}_N - \boldsymbol{\theta} = \left[\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \sum_{i=1}^{N} \left[\mathbf{x}_i(Y_i - \mathbf{x}_i'\boldsymbol{\theta})\right] = \mathbf{S}_{xx}^{-1} \sum_{i=1}^{N} \mathbf{u}_i(\boldsymbol{\theta}),$$

so that under the model,

$$\mathbf{E}_\xi[\boldsymbol{\theta}_N - \boldsymbol{\theta}] = \mathbf{0},$$

and

$$\mathbf{V}_\xi[\boldsymbol{\theta}_N - \boldsymbol{\theta}] = \mathbf{S}_{xx}^{-1}\left[\sum_{i,j=1}^{N} \sigma_{ij}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\mathbf{S}_{xx}^{-1} \stackrel{?}{=} \sigma^2\mathbf{S}_{xx}^{-1}.$$

3. Design-based Rationale

These standard results are very well known. We now apply a similar technique to the design-based and model-based estimators. For $\hat{\boldsymbol{\theta}}_p$, we have

$$\mathbf{0} = \sum_{i=1}^{N} \pi_i^{-1} I_i \left[\mathbf{x}_i (Y_i - \mathbf{x}_i' \boldsymbol{\theta}_N) - \mathbf{x}_i \mathbf{x}_i' (\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N) \right]$$

so that

$$\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N = \hat{\mathbf{S}}_{xx}^{-1} \sum_{i=1}^{N} \pi_i^{-1} I_i \left[\mathbf{x}_i (Y_i - \mathbf{x}_i' \boldsymbol{\theta}_N) \right].$$

Therefore, $\mathbf{E}_p[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N] \to \mathbf{0}$ and

$$\mathbf{V}_p[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N] \to \mathbf{S}_{xx}^{-1} \left\{ \sum_{i,j=1}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \mathbf{u}_i \mathbf{u}_j' \right\} \mathbf{S}_{xx}^{-1}.$$

3. Design-based Rationale

Under the model-design-based randomization framework, we have

$$\mathbf{E}_{\xi p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}] = \mathbf{E}_{\xi p}[(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N) + (\boldsymbol{\theta}_N - \boldsymbol{\theta})] \to \mathbf{0}$$

and

$$\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}] = \mathbf{V}_{\xi p}[(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N) + (\boldsymbol{\theta}_N - \boldsymbol{\theta})]$$

$$= \mathbf{E}_{\xi}\mathbf{V}_{p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N] + \mathbf{V}_{\xi}[\boldsymbol{\theta}_N - \boldsymbol{\theta}]$$

$$\to \mathbf{S}_{xx}^{-1}\mathbf{E}_{\xi}\left\{\sum_{i,j=1}^{N}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\mathbf{u}_i\mathbf{u}_j'\right\}\mathbf{S}_{xx}^{-1} + \mathbf{S}_{xx}^{-1}\sum_{i,j=1}^{N}\left[\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'\right]\mathbf{S}_{xx}^{-1}.$$

3. Design-based Rationale

Main result #1:

If $\mathbf{S}_{xx}^{-1}\sum_{i,j=1}^{N}\left[\sigma_{ij}\mathbf{x}_i\mathbf{x}'_j\right]\mathbf{S}_{xx}^{-1}=O(N^{-1})$

and $\mathbf{S}_{xx}^{-1}\mathbf{E}_{\xi}\left\{\sum_{i,j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\mathbf{u}_i\mathbf{u}'_j\right\}\mathbf{S}_{xx}^{-1}=O(n^{-1})$

then $\mathbf{V}_p[\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}_N]$ is model-design-unbiased for $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}]$

when n/N is negligible.

3. Design-based Rationale

Main result: #2

Since, if the sampling is non-informative,

$$\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}] = \mathbf{V}_{\xi}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}]$$

Then, under similar conditions, for large sample sizes, we have

$$\mathbf{V}_{\xi}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}] = \mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}] \approx \mathbf{V}_{p}[\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N].$$

3. Design-based Rationale

For the standard regression case, under non-informative sampling,

$$\mathbf{V}_{\xi}[\hat{\boldsymbol{\theta}}_{p} - \boldsymbol{\theta}] \rightarrow \sigma^{2}\mathbf{S}_{xx}^{-1}\left(\mathbf{E}_{\xi}\sum_{i=1}^{N}\pi_{i}^{-1}\mathbf{x}_{i}\mathbf{x}_{i}'\right)\mathbf{S}_{xx}^{-1}.$$

This is ***not*** what standard software estimates under WLS regression, even if using normalized weights.

(Note the distinction among probability weights, analytic weights and frequency weights.)

3. Design-based Rationale

Applying the same techniques to the model-based estimator $\hat{\boldsymbol{\theta}}$ does not give the same conclusions unless the sampling is non-informative

Main result #3

When the sampling is informative, the estimator may not be model-design unbiased, and the model-based variance can also be a biased estimate of $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]$.

3. Design-based Rationale

	Assumed first-phase model is valid and sampling is ignorable	Assumed first-phase model is misspecified or the sampling is nonignorable
Model-based estimator	1.Asymptotically unbiased 2.Efficient 3.Valid variance estimates 4.Valid inferences 5.May be best	1.May be inconsistent 2.Variance estimates may be invalid 3.Inferences may be invalid
Design-based estimator	1.Asymptotically unbiased 2.May be inefficient 3.Valid variance estimates 4.Valid inferences	If the mean of the estimating equation is zero under the model: 1. Asymptotically unbiased 2. Valid variance estimates 3. Valid inferences

4. Some Problematic Cases

- Non-negligible sampling fractions
- Rare characteristics
- Small sample sizes
- Models that include random effects
- Population-based case control studies
- Integrating data from more than one survey
- Event history analysis

5. An Artificial Example

The following data have been generated randomly using 100 independent and identically distributed trials (Bernoulli trials) taking the value (M or F) The "unknown" parameter is $\theta = \Pr(F) = .51.$

M M M M F F M M M F F F F F M F M F M M
F M F M F M F F M F F F F F F M F M M F
F M M M M M F M M M M M M F F M M F F M
F M F M M F M F F M F M M M M M M F F M
M M M F F F M F M M F F F M F F F M M M

We have 46F's and 54M's, so $\theta_N = .46.$

5. An Artificial Example

We now consider a finite consisting of these 100 observations as a starting point. The observations have been arranged into households using a non-random process. The results are as follows:

```
(M M M M F),(F M),(M M F),(F F F F M),
(F M),(F M),(M F),(M F),(M F),(M F),
(F M),(F F F F F F M),(F M),(M F),(F M),
(M M M M F),(M M M M M M F),(F M),(M F),
(F M),(F M),(F M),(M F),(M F),(F M),(F M),
(M M M M M F),(F M),(M M M F),(F F M),
(F M),(M F),(F F M),(F F F M),(M M)
```

5. An Artificial Example

We now take a cluster sample by selecting *m* households from the 35 households at random, with replacement, using selection probabilities proportional to household size, and then enumerating all persons in the selected households.

It turns out that $E_p[\hat{\theta}] = .445$, so that the design bias is only -.015.

5. An Artificial Example

However, for large samples, we have

$$E_p \hat{V}_\xi[\hat{\theta}_p] = .087/m$$

$$E_p \hat{V}_p[\hat{\theta}_p] = .049/m$$

$$E_p \hat{V}_\xi[\hat{\theta}] = .068/m$$

$$E_p \hat{V}_p[\hat{\theta}] = .114/m$$

Obviously, the sampling plan is informative.

We see from this simple example that the impact of informative sampling on model-based methods can be substantial.

6. Modeler's Criticisms

Hoem, Jan M. (1989)

"… sample-based analyses of individual longitudinal behavior can normally do well without sampling weights. Instead of worrying about such weights, it pays to concentrate on the modelling of behavior and on drawing inference about features of the model. One should not feel confined to finite population totals and means, finite population regression coefficients, and other finite population statistics. Also, some of the claims about the good properties of conventional weighting seem exaggerated."

6. Modeler's Criticisms

Hoem, Jan M. (1989)

- Use of weights is not appropriate when modeling probabilistic models of human behaviour because we are not estimating a finite population quantity.

- Regard sampling mechanism as part of the total model of the “random experiment” that produces the survey data, with normal consequences for the statistical analysis.

6. Modeler's Criticisms

Hoem, Jan M. (1989) – cont'd

- Informativeness is not relevant for the issue of robustness against behavioural model misspecification

- The finite population quantity associated with the model parameter may not always be a useful descriptive statistic.

6. Modeler's Criticisms

Fienberg (1989)

• Sampling weights, as they are usually constructed, are at best irrelevant to a likelihood-based approach to statistical inference.
• However, weights may be appropriate for outcome-based sampling.

Breckling et al. (1994) and others have mentioned the problem of using design-based methods for estimating finite population quantities with cut-off sampling.

- this is a clear-cut example of where the survey population and the target populations differ.

6. Modeler's Criticisms

Sugden (1998)

There is a danger of using purely asymptotic arguments to justify robustness. The asymptotic scheme requires an infinite sequence of samples and populations, usually replicating the given population and sampling scheme somewhat arbitrarily and incapable of verification. Instead of limit theorems, I would like to see approximations to the bias for use with moderate sample sizes.

6. Modeler's Criticisms

Little (2004)

The design-based approach to survey inference has a number of strengths that make it popular with practitioners. It automatically takes into account features of the survey design and provides reliable inferences in large samples, without the need for strong modeling assumptions.

6. Modeler's Criticisms

Little (2004) – cont'd

On the other hand, it is essentially asymptotic, and hence yields limited guidance for small-sample adjustments. Unlike models, which lead to efficient inferences based on likelihood or Bayesian principles, the design-based approach is not prescriptive for the choice of estimator. It lacks a theory for optimal estimation, and the estimates that it yields are potentially inefficient.

6. Modeler's Criticisms

Little (2004) – cont'd

- Modelling provides a unified approach to survey inference, aligned with mainline statistics approaches in other application areas, such as econometrics.
- In large samples and with uninformative prior distributions, results can parallel those from design-based inference.
- The Bayesian approach is well equipped to handle complex design features.
- The Bayesian approach may yield better inferences for small-sample problems where exact frequentist solutions are not available, by propagating error in estimating parameters.

6. Modeler's Criticisms

Little (2004) – cont'd

- The Bayesian approach allows prior information to be incorporated, when appropriate.
- The Bayesian approach avoids the ambiguities in the choice of reference distribution and has useful features of coherency not shared by frequentist approaches, such as satisfying the likelihood principle.
- Likelihood-based approaches like Bayes or maximum likelihood have the property of large-sample efficiency, and hence match or outperform design-based inferences if the model is correctly specified.

6. Modeler's Criticisms

The Question of Nonresponse Modelling

To this point, we have not considered the complexities which arise when making adjustments for non-response. Many of the criticisms of using design-based weights in making survey inferences are related to modeling for non-response.

As Kim and Kim (2007) note that the issue of whether or not to use the weights when modeling non-response propensities is not clear-cut, and that when one considers the estimates with the estimated non-response adjustment incorporated neither using the weights nor ignoring may be optimal. This is clearly an area of future research.

7. Closing Remarks

As we mentioned earlier, when successful at using models to incorporated the reasons for informative sampling, Maximum Likelihood or Sample Likelihood methods will usually outperform the Pseudo-likelihood approach using design-based methods.

When sample sizes (or number of psu's) are small, a model-based or even a Bayesian approach can be preferred to a design-based approach even if the model is not quite correct.

Models that include random effects are definitely problematic, even for large total sample sizes.

7. Closing Remarks

The use of the original weights when the weights are very variable may not be best, even when interested in finite population quantities. Some weight modifications may be suitable; for example in population-based case control studies, or in integrating data from more than one survey. Users may also find that a few observations are too influential due to the large weights, and it would make sense to make adjustments.

7. Closing Remarks

Weighted and unweighted point estimates may or may not be similar. If they are not, think about modifying the model to incorporate the fact that the sampling is informative, so that the model better explains the sampling distribution of the data.

Even if the point estimates are similar, consider modifying the model to account for reasons why the sampling could be informative.

7. Closing Remarks

Care must be exercised when modifying a model to include design variables. One needs to ensure that the model is being interpreted as was originally intended.

Standard errors based on a design-based approach may tend to be more robust in cases where the informativeness of the sample has not been fully incorporated into the model.

7. Closing Remarks

The design-based approach will still give correct inferences for the parameters of the model used to generate the finite population, when the assumed model for the finite population is true, even if the sampling is informative.

What do we mean by robustness? The design-based approach guards against misspecification of the model error structure. It is still important to get the best functional form for the model expectation - such as linearity and no missing variables.

FINAL WORDS

‘all models are wrong, some are useful’
- George Box

DISCUSSIONS WILL CONTINUE FOR A LONG TIME!

Contact information:

David Binder

dbinder49@hotmail.com