STATISTICAL INFERENCE FOR SURVEY DATA ANALYSIS

David A. Binder and Georgia R. Roberts
Methodology Branch, Statistics Canada, Ottawa, ON, Canada K1A 0T6

KEY WORDS: Design-based properties, Informative sampling, Phases of randomization

1. QUESTIONS TO BE CONSIDERED IN THIS PAPER

When a model generates a finite population, what would lead to a violation of that model in the sample (i.e. lead to informative sampling)?

When analysing survey data, what are the assumed "repetitions" when taking a frequentist approach to inference?

Why would approaches that have been developed for inferences about fixed finite populations be appropriate for inferences about model parameters?

2. AN ARTIFICIAL EXAMPLE

Model and data generation

We begin with a simple modelling problem. The following data have been generated randomly using **100 independent and identically distributed Bernoulli trials (M or F),** where the unknown Bernoulli 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
```

The data generated consist of 46 F's and 54 M's.

Model-based method for estimating θ

The standard method for estimating θ, the parameter of interest, from data generated in this manner is the following:

We let $y_i = \begin{cases} 1 \text{ if } F, \\ 0 \text{ if } M. \end{cases}$

Under the independent, identically distributed Bernoulli model, the loglikelihood function is:

$$\log f = \Sigma [y_i \log\theta + (1 - y_i)\log(1 - \theta)].$$

Maximization of this likelihood then leads to the estimator $\hat{\theta}_\xi = \Sigma y_i / n$, where n is the number of Bernoulli trials. Under this assumed model, it is well known that the expectation of $\hat{\theta}_\xi$ is $E_\xi[\hat{\theta}_\xi] = \theta = .51$, and the variance of $\hat{\theta}_\xi$ is

$$V_\xi(\hat{\theta}_\xi) = \frac{\theta(1 - \theta)}{n}.$$

As well, the usual estimated variance of $\hat{\theta}_\xi$ is given by:

$$\hat{V}_\xi(\hat{\theta}_\xi) = \frac{\hat{\theta}_\xi(1 - \hat{\theta}_\xi)}{n - 1}.$$

What is random here?

We are assuming that the randomness is due to **repetition** of "experiments", each consisting of **100 independent and identically distributed Bernoulli trials**, where $\theta = .51$ is the target parameter. *We denote this randomization as ξ-randomization*.

A simple example of sampling theory

The finite population

We now consider a finite population 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)
```

We consider persons in these 35 households as the finite population of interest. This population contains 46 F's and 54 M's clustered into households that are distributed as follows:

```
(1M,1F)-23; (2M,0F)-1; (2M,1F)- 1;
(3M,1F)-1; (4M,1F)-2; (5M,1F)-1; (6M,1F)-1;
(1M,2F)-2; (1M,3F)-1; (1M,4F)-1; (1M,6F)-1.
```

Note that even though the M's and F's were originally generated from an independent and identically distributed Bernoulli model, the larger households have more M's or F's than would be expected from a purely random assignment of units to households of given sizes.

Sampling from the finite population:

We cannot observe the complete finite population, but we are interested in estimating θ_p, the proportion of F's in this population. 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.

As an example, letting $m=10$, we select a sample, and observe: **(1M,1F)-6; (4M,1F)-1; (1M,6F)-1; (1M,2F)-1; (4M,1F)-1.**

We use this particular sample below for illustration.

What is random here?

Here it is assumed that the randomness is due to the **repetition** of "experiments", each consisting of **selecting the m households from the 35 households**. The number of M's and F's in each of the 35 households is **fixed**. The target parameter is the finite population quantity, $\theta_p = .46$. *We denote this randomization as p-randomization.*

It should be noted that there is a conceptual difference between the finite population parameter, $\theta_p = .46$, and the model parameter, $\theta = .51$. **But** under the Bernoulli model, $E_\xi[\theta_p] = \theta$. Therefore, if θ_p could be observed, it would provide a model-unbiased estimate for θ since the Bernoulli model was used to generate the finite population consisting of 100 units. In fact, for large finite populations, we have $\theta_p \approx \theta$. It then follows that a "good" estimate of θ_p would be expected to be a "good" estimate of θ.

Design-based method for estimating θ_p

Suppose that in the i-th selected household we observe h_i household members and x_i females. The standard probability-weighted estimate for the proportion of females in the finite population is

$$\hat{\theta}_p = \sum_s p_i / m ,$$

where $p_i = x_i / h_i$.

For our particular sample of 10 households, we have $\hat{\theta}_p = .492$.

Expectations and Variances

Expectations of design-based estimate, $\hat{\theta}_p$:

The estimator, $\hat{\theta}_p$, provides a design-unbiased estimate for the finite population proportion, under the random sampling plan used to select households and individuals; that is,

$$E_p[\hat{\theta}_p] = \theta_p = .46.$$

Under the assumption that the Bernoulli model is valid for the sample, the model-based expectation of $\hat{\theta}_p$ is $E_\xi[\hat{\theta}_p] = \theta = .51$, so that $\hat{\theta}_p$ would be model-unbiased for θ.

Expectations of model-based estimate, $\hat{\theta}_\xi$:

If we assume that the sample is generated from independent, identically distributed Bernoulli trials, then the standard estimator of θ based on this model is the unweighted estimator given by

$$\hat{\theta}_\xi = \frac{\sum_s x_i}{\sum_s h_i},$$

since $\sum_s x_i$ is the number of females in all the sampled households and $\sum_s h_i$ is the number of persons in all the sampled households.

For our particular sample of 10 households, we have $\hat{\theta}_\xi = .50$.

This estimator would be model-unbiased; that is, $E_\xi[\hat{\theta}_\xi] = \theta$, **if the assumed model were correct for the sample.**

However, it can be shown that for large m, we have $E_p[\hat{\theta}_\xi] = .445$, so that $\hat{\theta}_\xi$ is asymptotically design-biased as an estimate of θ_p. The asymptotic **design** bias of $\hat{\theta}_\xi$ as an estimate of θ_p is $E_p[\hat{\theta}_\xi] - \theta_p = .445 - .46 = -.015$, which is relatively small.

Since the selection probabilities are directly related to the household size, $\hat{\theta}_\xi$ would be design-unbiased for θ_p if the household size were uncorrelated with the number of females in the household. If the correlation is small - as is the case in this example - the design bias would be small.

The following table summarizes many of the statements made above.

Estimates of Expectation: Observed values from our sample of 10 households, and model-based and design-based expectations

Estimate	Sample of 10 households	E_p	$E_{\xi(false*)}$
$\hat{\theta}_\xi$	.500	.445	.51
$\hat{\theta}_p$	.492	.46	.51

*Note that because the number of females within each household is not binomially distributed (with θ constant), the ξ-model is false for the sample. (It is, however, valid for the finite population of 100 units, before the households were constructed.)

Variance Estimation

Under the **assumed Bernoulli model**, the usual estimate of the variance of $\hat{\theta}_\xi$ is

$$\hat{V}_\xi(\hat{\theta}_\xi) = \frac{\hat{\theta}_\xi(1-\hat{\theta}_\xi)}{n_s - 1} = \frac{\hat{\theta}_\xi(1-\hat{\theta}_\xi)}{\sum h_i - 1}.$$

In the case of $\hat{\theta}_p$, we consider both model-based and design-based variance estimates:

Model-based: $\hat{V}_\xi(\hat{\theta}_p) = \dfrac{\hat{\theta}_p(1-\hat{\theta}_p)}{m^2\left(\Sigma_s \dfrac{1}{h_i}\right)^{-1} - 1}$

Design-based: $\hat{V}_p(\hat{\theta}_p) = s^2/m$,

where $s^2 = \Sigma_s(p_i - \hat{\theta}_p)^2/(m-1)$.

The following table provides these estimates for our particular sample, and the large-sample expectations.

Estimates of Variance: Observed values from our sample of 10 households, and model-based and design-based expectations

Estimate	Sample of m=10 households	E_p (For large m)	$E_{\xi(false)}$ (For large m)
$\hat{V}_\xi(\hat{\theta}_\xi)$	.0081	.068/m	.069/m
$\hat{V}_p(\hat{\theta}_\xi)$*	.0119	.114/m	.069/m
$\hat{V}_\xi(\hat{\theta}_p)$	.0010	.087/m	.088/m
$\hat{V}_p(\hat{\theta}_p)$	.0056	.049/m	.087/m

*We include $\hat{V}_p(\hat{\theta}_\xi)$ for the sake of completeness only, even though it is seldom calculated, as $\hat{\theta}_\xi$ is generally design-biased.

Phases of randomization

So far we have discussed the notions of model-based randomness and design-based randomness separately. These two notions can be combined into one framework, comprising both the model and the design randomizations.

Randomization over both phases

We denote by ξp-randomization the situation where the randomness is due to **repetition** of "experiments" as follows:

First, we have **100 Bernoulli trials** where $\theta = .51$. Next we arrange these outcomes into households. Finally, we select m **households at random with replacement with probabilities proportional to household size.**

When the target parameters are defined by a model that generated the finite population, but these parameters are estimated from data from a survey, it would seem reasonable and desirable for the estimates to have good properties under this ξp-randomization. This framework has been examined by Hartley and Sielken (1975), Molina, Smith and Sugden (2001), and others.

Under this randomization, we have that $E_{\xi p}[\hat{\theta}_p] = \theta = .51$, so that $\hat{\theta}_p$ is ξp-unbiased for θ.

We define the *total variance* to be the variance over the two phases of sampling. Confidence intervals and tests of hypotheses using this total variance are then valid with respect to the double sampling described above; that is, on repetitions of both the generation of the finite population using the Bernoulli trials, and the sampling of households by selecting the households with probabilities proportional to household size.

We now examine how the properties of the variances of estimates under this ξp-randomization can be used to assess whether the model holds for the observed units.

Non-robustness of the model-based approach

- Taking into account the randomization due to both the model and the design, if the assumed model is true for the observed units, then it can be shown that the **design-based variance of $\hat{\theta}_p(\hat{\theta}_\xi)$** should be close to the **model-based variance of $\hat{\theta}_p(\hat{\theta}_\xi)$** for large samples and large populations.
- If the design-based and the model-based variances are not close, then the variances implied by the assumed model are wrong! Using an estimate of an incorrect model-based variance in analyses can result in poor coverage properties for confidence intervals and misleading tests of hypotheses.

In our example, we have that, for large values of m,

$$V_p(\hat{\theta}_p) = .049/m \text{ and } V_{\xi(false)}(\hat{\theta}_p) = .088/m.$$

$$V_p(\hat{\theta}_\xi) = .114/m \text{ and } V_{\xi(false)}(\hat{\theta}_\xi) = .069/m,$$

Therefore, we can surmise that the assumed the model is not true for the observed units!

In practice, it is not possible to compute the theoretical values of the variances given above, since the processes leading to the ultimate sample are not known. However, the estimates of the variances can be computed from the sample data. For large samples, the estimates will be approximately equal to their design expectations. Therefore, we consider the design expectations of the estimates of the variances of $\hat{\theta}_p(\hat{\theta}_\xi)$. The results are similar to those above, for our example. We have that, for large values of m,

$$E_p[\hat{V}_p(\hat{\theta}_p)] = V_p(\hat{\theta}_p) = .049/m \text{ and } E_p[\hat{V}_\xi(\hat{\theta}_p)] = .087/m.$$

If the assumed model is true for the observed units, then it can be shown that these design-based expectations should be close. This does not seem to be the case here, so again we conclude that assumed model is not true for the observed units. We see that,

for a particular sample, comparing $\hat{V}_p(\hat{\theta}_p)$ and $\hat{V}_\xi(\hat{\theta}_p)$ can be a useful diagnostic for determining if the sampling is possibly informative.

We also have a similar result for $\hat{\theta}_\xi$:

$$E_p[\hat{V}_p(\hat{\theta}_\xi)] = V_p(\hat{\theta}_\xi) = .114/m \text{ and } E_p[\hat{V}_\xi(\hat{\theta}_\xi)] = .068/m.$$

Our conclusions about the informativeness of the sample design would, therefore, be similar if we were to compare $\hat{V}_p(\hat{\theta}_\xi)$ and $\hat{V}_\xi(\hat{\theta}_\xi)$. However, this comparison is not as useful as a diagnostic tool since $\hat{V}_p(\hat{\theta}_\xi)$ is not commonly computed.

What does all this mean for the practitioner?

Comparing the design-based and model-based point estimates can be useful. If they are quite different, this is an indication that the assumption that the model for generating the finite population holds for the sample is incorrect.

Even if the point estimates are similar, if for large samples, $\hat{V}_p(\hat{\theta}_p)$ is not close to $\hat{V}_\xi(\hat{\theta}_p)$ the sampling may be "informative".

Pure modelers may prefer to account for the sample design in this example by modelling the process for creating the household structure, given the finite population. However, when the full model is hard to specify, the probability sampling mechanism must play an important role to avoid biases (Little and Rubin, 1987, p. 246).

What do we mean by "informative" sampling?

Informativeness of a sample is a model concept.

If, on repetitions of the ξp-randomization "experiments" described earlier, the distribution of the sampled units is different from the distribution that would be obtained by sampling directly from the model, then the sampling is said to be **informative**; see Binder and Roberts (2001).

3. PARAMETRIC MODEL ANALYSES FROM SURVEYS

In the example above, we generated our sample through a two-phase process. This artificial case can be generalized to most situations where an analyst is fitting models to survey data as follows:

Phases to Generate the Sample of Available Data

- First, the complete finite population is generated from a model.
- Then, conditional on these outcomes, design variables (e.g. household composition, geography, etc.) are added to the units of the population according to some **possibly unspecified** process.
- A sample of units is selected under a **known** probability design.

Other randomizations to allow for non-response, measurement error, etc. could also be added to this framework.

The case where the **assumed** model for generating the finite population is incorrect can also be incorporated into this framework, with the model generating the finite population in the first phase being specified as not being the same as the **assumed** model.

Design-based Approach to Inference for Target (Model) Parameters

A design-based approach is most useful when the sample sizes are large and the sampling fractions small. In this case, the design-based properties of the estimates will be close to those obtained under the ξp-randomization. Also, the design-based variances may be close to the true model-based variances even when the assumed model is incorrectly specified; see Binder and Roberts (2003) for details.

The procedure for implementing this approach is as follows:

- Define estimating equations (EE's) using a model-based procedure such as maximum likelihood estimation based on the complete finite population. Solutions to these EE's become the finite population "descriptive" quantities of interest.
- Determine finite population estimates for these quantities, using appropriate design-based methods.
- Use design-based randomization to provide measures of accuracy.

Warning: For mixed effects models such as hierarchical linear models and multilevel models, the usual asymptotic theory does not extend to some parameters. It is generally valid for the fixed effects, but not for the random effects; see, for example, Pfeffermann et al. (1998).

4. CONCLUSIONS

For the analysis of survey data where a model has generated the finite population and where the target parameters are based on this model:

- Typical complex survey designs often lead to informative samples.
- Weighted and unweighted point estimates may or may not be similar. If they are not, think about augmenting the model to incorporate the fact that the sampling is informative, so that the model better explains the sampling distribution.
- Even if the point estimates are similar, consider an augmented model that accounts for reasons why the sampling could be informative.
- Standard errors based on a design-based approach will

tend to be more robust because they account for the informativeness of the sampling design. For large samples and small sampling fractions, the design-based approach will still give correct inferences for the target parameters of the model, even when the sampling is informative.

- Care must be exercised when augmenting the model with design variables to ensure that the parameters being estimated are those of analytic interest.
- The model used to generate the finite population may be incorrectly specified and it is important to perform diagnostics on both the model means and the model variances and covariances.
- **If a pure model-based approach is used to analyse complex survey data, caveats about the fact that the design information has been ignored should be included in the analysis report.**

ACKNOWLEDGEMENTS

We are most grateful to John Eltinge and David Judkins for their useful comments and suggestions on an earlier draft of this paper.

REFERENCES

Binder, David A. and Georgia R. Roberts (2001). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section,* Issue 12, American Statistical Association.

Binder, David A. and Georgia R. Roberts (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data*, (eds. R.L. Chambers and Chris Skinner) Wiley, Chichester.

Hartley H. O. and Sielken, R. L., Jr. (1975) A "Super-population viewpoint" for finite population sampling. *Biometrics,* 31, 411-422.

Little, Roderick J. A and Donald B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Molina, E. A., T.M.F. Smith, and R.A. Sugden (2001). Modelling overdispersion for complex survey data. *International Statistical Review,* 69, 373-384.

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society, Series B,* 60, 23-56.