Session 2.4  How to restrict to the necessary?

Denise Silva
Challenge

How to get the right balance between:

The required information to lead the reader through your ideas and work

vs.

The ability to summarise technical details and results

Every journal addresses a specific audience: an author must write the paper with that audience in mind
The Audience

- **Journal of Official Statistics:** “The intended readers are researchers and practitioners in academia, government, business or research organisations with an interest in survey methodology and production of official statistics”
- **Survey Methodology:** “A key source of information for survey statisticians and methodologists”
- **Statistical Journal of the IAOS:** “The journal should publish papers of wide interest to both users and producers of official statistics”
- **International Statistical Review:** “Papers must be of interest to a sufficiently broad spectrum of the members of ISI and its family of Associations, who include researchers and practitioners in academia, government, business, and industry”
- **The Survey Statistician:** As the newsletter of the IASS, it includes general information about the activities of the Association (meetings, seminars, etc.) and welcome manuscripts “that are likely to be of interest to its members”.

Workshop Writing manuscripts for Official Statistics journals: Guidelines for practitioners and researchers: 8, 10 and 15 February 2022
The Hourglass

Consider the image of an hourglass to represent the organization (and balance) of your manuscript

Wide at top: The introduction presents the relevance of the study, then it narrows down the scope into a specific problem/question

Narrow towards the middle: The methods section should be restricted to a manageable focus (the neck of the hourglass)

Widening out again: Results – give more room to this section

Wide at the bottom: a final thin layer broadening the focus to relate your work with what is on and proposing future developments

https://www.enago.com/academy/academic-writing-in-science-overview/
Methods

The work may be completely new or may demonstrate an improvement of an existing method

Official statistics journals assume that the writer will demonstrate technical expertise

- If you have developed an entirely new method: write it out in detail
- If you are improving previous work:
  - There is no need to repeat all the details (use references to inform about the previous work)
  - Be clear about which methods should be compared with your proposal
Methods

• Clarity is the aim

• If proofs are included in the manuscript: they should be complete and presented in the easiest possible way

• You can omit details according to expected target audience knowledge

• Do not over-explain common scientific/technical procedures

• Provide some discussion/motivation to assist the reader decoding the formulas

• “An author must provide enough detail for a reader to be able to reconstruct his/her study, but not so much that the relevant points get buried”
  (Journal of Young Investigators, 2005)
Empirical Work/Application

Focus is on the data/analysis used and the specific results

It is important that the readers can follow the analysis and understand the debate

• The novelty is the conceptual or analytical approach or the new evidence to inform the debate

• You may need to present statistical theory to support your analysis

• The empirical work can be related to testing different methods or evaluating methodological procedures in a new domain or application

• Deciding how much detail is about judgment
Results

This section should get the lion’s share of your manuscript

• To show how your work fits in the context of existing literature and
• To explain “how your study adds to the body of knowledge”

• Inform the main findings
• Resist the temptation to include every result you have obtained
• Check the number of tables and figures that are allowed according to chosen journal
• Do not repeat in words everything that your tables and graphs convey (avoid redundancy)

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

Kees van Berkel, Suzanne van der Doef, and Barry Schouten

Adaptive survey design has attracted great interest in recent years, but the number of case studies describing actual implementation is still thin. Reasons for this may be the gap between survey methodology and data collection, practical complications in differentiating effort across sample units and lack of flexibility of survey case management systems. Currently, adaptive survey design is a standard option in redesigns of person and household surveys at Statistics Netherlands and it has been implemented for the Dutch Health survey in 2018. In this article, the implementation of static adaptive survey designs is described and motivated with a focus on practical feasibility.

2. Methodology

In this section, the four main elements of adaptive survey design, quality and cost criteria, design features, stratification and optimisation, are discussed. This is done from an operational perspective.
Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

Kees van Berkel¹, Suzanne van der Doef¹, and Barry Schouten²

2.4. Optimisation

Two approaches to optimisation are explored: case prioritisation and mathematical optimisation, including expected yield of the face-to-face follow-up.

https://doi.org/10.2478/jos-2020-0031
Methods – Example 1

2.4. Optimisation

Two approaches to optimisation are explored: case prioritisation and mathematical optimisation, including expected yield of the face-to-face follow-up.

https://doi.org/10.2478/jos-2020-0031
3. Application of Adaptive Survey Design to the Dutch Health Survey

3.1. The Dutch Health Survey

The aim of the Dutch Health Survey is to provide as complete an overview as possible of developments in health, medical contacts, lifestyle and preventive behaviour of the population in the Netherlands. The target population consists of all people living in the Netherlands who do not belong to the institutional population. The sample is a stratified two stage sample in which people with equal probabilities are selected. This sampling design is approximately the same as the simple random sampling design. The observation starts with CAWI and the re-approach mode is CAPL. As a response increasing measure, iPads are raffled among the sampled people.

https://doi.org/10.2478/jos-2020-0031
Empirical Work / Results – Example 1

3. Application of Adaptive Survey Design to the Dutch Health Survey

3.3. The Dutch Health Survey Optimisation Problem

The set $G$ of groups used to determine the target groups consists of 32 groups: ethnicity(2) × age(4) × income(2) × urbanity(2). The coefficient of variation of response probabilities $CV(p) = \frac{S_p}{\hat{p}}$ is estimated as described in Subsection 2.4.1. Minimising $CV(p)$ is carried out under the constraints:

- $n \leq n_{\text{max}}$ (CAWI sample size does not exceed $n_{\text{max}}$),
- $n \cdot \hat{p} \geq R$ [expected response size is at least $R$],
- $\sum_{d \in G} f_p(d) n_d \leq C$ [total CAPI sample size is at most $C$],
- For each target group $d$ and all groups $g_1, g_2 \subset d : f_p(g_1) = f_p(g_2)$ applies [one CAPI sampling fraction per target group].

Here $n_{\text{max}}$, $R$ and $C$ are constants to be filled in. The parameters with which the minimum can be found are the CAWI sample size $n$ and the CAPI sampling fractions for face-to-face observation $f_p(d)$ in the target groups $d$. Note that it follows from the first two constraints that $\hat{p} \geq R/n_{\text{max}}$.

In the case of the Health Survey 2018, the target groups with corresponding response rates per mode and probabilities of re-approachable CAWI nonresponse have been determined with data from the results of the Health Survey in January-June of 2017. The maximum CAWI sample size $n_{\text{CAWI}}$ has been set to 18,000 people. To be quite sure that 9,500 responses are achieved, the expected response size $R$ has been set to 9,631 people. The maximum CAPI sample size $C$ is 8,039 addresses, based on the available CAPI budget. The mean response rate must therefore be at least $631/18,000 = 53.5\%$.

3.4. Mathematical Optimisation

First, the mathematical optimisation approach is explored, as this approach may be used as a benchmark to the case prioritisation approach.

3.5. Case Prioritisation

Case prioritisation employs the same nine strata and sorts the strata after the CAWI phase by estimated response propensities. One practical complication is added. The Netherlands is divided into ten interviewer regions, each of which contains about one-tenth of the population. Each interviewer region employs 10 to 15 interviewers. Since 2016, CAPI

<table>
<thead>
<tr>
<th>Stratum</th>
<th>n CAWI</th>
<th>r CAWI</th>
<th>p CAWI</th>
<th>n elig</th>
<th>n CAWI</th>
<th>r CAWI</th>
<th>p CAWI</th>
<th>n elig</th>
<th>r CAWI</th>
<th>p CAWI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,542</td>
<td>1,936</td>
<td>42.6</td>
<td>2,470</td>
<td>1,472</td>
<td>59.6</td>
<td>714</td>
<td>48.5</td>
<td>2,650</td>
<td>58.3</td>
</tr>
<tr>
<td>2</td>
<td>785</td>
<td>194</td>
<td>24.8</td>
<td>567</td>
<td>508</td>
<td>89.5</td>
<td>844</td>
<td>48.0</td>
<td>438</td>
<td>55.6</td>
</tr>
<tr>
<td>3</td>
<td>7,361</td>
<td>2,765</td>
<td>37.6</td>
<td>4,375</td>
<td>3,480</td>
<td>79.5</td>
<td>1,473</td>
<td>42.3</td>
<td>4,239</td>
<td>57.6</td>
</tr>
<tr>
<td>4</td>
<td>727</td>
<td>168</td>
<td>23.1</td>
<td>538</td>
<td>538</td>
<td>100.0</td>
<td>137</td>
<td>25.5</td>
<td>305</td>
<td>41.9</td>
</tr>
<tr>
<td>5</td>
<td>1,472</td>
<td>580</td>
<td>39.4</td>
<td>848</td>
<td>410</td>
<td>48.4</td>
<td>625</td>
<td>64.6</td>
<td>845</td>
<td>57.4</td>
</tr>
<tr>
<td>6</td>
<td>332</td>
<td>121</td>
<td>36.5</td>
<td>201</td>
<td>149</td>
<td>74.3</td>
<td>72</td>
<td>48.0</td>
<td>193</td>
<td>58.1</td>
</tr>
<tr>
<td>7</td>
<td>1,274</td>
<td>245</td>
<td>19.3</td>
<td>991</td>
<td>991</td>
<td>100.0</td>
<td>304</td>
<td>30.7</td>
<td>550</td>
<td>43.1</td>
</tr>
<tr>
<td>8</td>
<td>411</td>
<td>83</td>
<td>20.3</td>
<td>315</td>
<td>315</td>
<td>100.0</td>
<td>122</td>
<td>38.6</td>
<td>205</td>
<td>49.9</td>
</tr>
<tr>
<td>9</td>
<td>963</td>
<td>105</td>
<td>28.9</td>
<td>247</td>
<td>176</td>
<td>71.4</td>
<td>103</td>
<td>58.3</td>
<td>208</td>
<td>57.3</td>
</tr>
<tr>
<td>Total</td>
<td>17,268</td>
<td>6,198</td>
<td>35.9</td>
<td>10,551</td>
<td>8,039</td>
<td>76.2</td>
<td>3,433</td>
<td>42.7</td>
<td>9,631</td>
<td>55.8</td>
</tr>
</tbody>
</table>

https://doi.org/10.2478/jos-2020-0031
Section 3.6 presents results of simulations and compare them with regular estimates.

This article describes and motivates choices that are made in the implementation of adaptive survey design at Statistics Netherlands. The focus is on sequential mixed-mode designs and the allocation of follow-up interviewer modes to nonrespondents of self-administered modes. The coefficient of variation of response propensities was adopted as the objective in optimisation of the designs. However, a range of logistical and cost constraints have been imposed and lead to a multifaceted optimisation problem. To facilitate easy management of the data collection, a case prioritisation approach was preferred over a mathematical optimisation. A case prioritisation approach is relatively easy to conduct and also is relatively robust to time change in survey design parameters such as costs and response propensities. However, improvement of the balance, that is a smaller coefficient of variation, is not guaranteed. A mathematical optimisation employing expected yield in follow-up interviewer modes does lead to improved balance, but is more sensitive to time change.

For the Health Survey case study, the implemented case prioritisation approach was compared to the mathematical optimisation approach. Results show that, as expected, on average the yield is smaller. However, balance is improved and the population strata that are allocated to face-to-face follow-up closely resemble each other. These results are promising.

There are a few limitations in this study: First, for ease of demonstration, the sampling design was restricted to simple random samples. Second, the role of mode-specific measurement bias was completely ignored. Third, the allocation of interviewer modes is posed as a simple yes-no decision, while it is clearly beneficial to also vary the amount of interviewer effort, for instance the number of contact attempts by the interviewers. Fourth,
Methods – Example 2

Abstract

In this paper, we consider the Fay-Herriot model for small area estimation. In particular, we are interested in the impact of sampling variance smoothing and modeling on the model-based estimates. We present methods of smoothing and modeling for the sampling variances and apply the proposed models to a real data analysis. Our results indicate that sampling variance smoothing can improve the efficiency and accuracy of the model-based estimator. For sampling variance modeling, the HB models of You (2016) and Sugasawa, Tamae and Kubokawa (2017) perform equally well to improve the direct survey estimates.

2. Fay-Herriot model using EBLUP approach

Under the Fay-Herriot model (1.3), assuming $\sigma^2_i$ and $\sigma^2$ known in the model, we obtain the best linear unbiased prediction (BLUP) estimator of $\theta_i$ as $
\hat{\theta}_i = y_i \gamma_i + (1-\gamma_i) x_i \beta, \n$ where $\gamma_i = \sigma^2_i / (\sigma^2_i + \sigma^2)$. And $\hat{\beta} = \left( \sum_i (\sigma_i^2 + \sigma^2)^{-1} x_i x_i^T \right)^{-1} \left( \sum_i (\sigma_i^2 + \sigma^2)^{-1} y_i x_i \right)$. To estimate the variance component $\sigma^2_i$, we have to first assume $\sigma^2_i$ known. There are several methods available to estimate $\sigma^2_i$, and we use REML method to estimate $\sigma^2_i$. Then the EBLUP of the small area parameter $\theta_i$ is obtained as

https://doi.org/10.2478/jos-2020-0031
Methods - Example 2

3. Fay-Herriot model using HB approach with sampling variance modeling

In this section we first present the Fay-Herriot model in a HB framework. Then we consider three models for the sampling variance modeling. The first model is the one considered in You and Chapman (2006) in which an inverse gamma model is used for the sampling variance $\sigma_y^2$ with known vague parameter values. The second model is introduced in You (2016) whereby a log-linear model with random error is used for $\sigma_y^2$. The third model is one proposed by Sugasawa et al. (2017) where an inverse gamma model is used for $\sigma_y^2$ but with different parameter settings.

**HB Model 1:** Fay-Herriot model in HB, denoted as FH-HB:

- $y_i | \theta, \sigma_y^2 \sim \text{iid } N(\theta_i, \sigma_y^2), \quad i = 1, \ldots, m;$
- $\theta_i | \mu, \sigma_\theta^2 \sim \text{iid } N(\mu, \sigma_\theta^2), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\theta) \propto 1, \quad \pi(\sigma_\theta^2) \propto 1.$

Note that in the FH-HB model, the sampling variance $\sigma_y^2$ is assumed to be known. Either a smoothed sampling variance $\hat{\sigma}_y^2$ or a direct sampling variance estimate $\hat{\sigma}_y^2$ will be used in place of $\sigma_y^2$.

**HB Model 2:** You-Chapman Model (You and Chapman, 2006), denoted as YCM:

- $y_i | \theta, \sigma_y^2 \sim \text{iid } N(\theta_i, \sigma_y^2), \quad i = 1, \ldots, m;$
- $d_\alpha_i | \sigma_y^2 \sim \text{iid } \text{IG}(a, b), \quad i = 1, \ldots, m;$
- $\theta_i | \beta, \sigma_\theta^2 \sim \text{iid } N(\beta, \sigma_\theta^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_y^2) \propto 1, \quad \pi(\sigma_\theta^2) \propto 1.$

The full conditional distributions for the Gibbs sampling procedure under both FH-HB and YCM can be found in You and Chapman (2006).

**HB Model 3:** You (2016) Log-linear model on sampling variances, denoted as YLLM:

- $y_i | \theta, \sigma_y^2 \sim \text{iid } N(\theta_i, \sigma_y^2), \quad i = 1, \ldots, m;$
- $d_\alpha_i | \sigma_y^2 \sim \text{iid } \text{IG}(a, b), \quad i = 1, \ldots, m;$
- $\theta_i | \beta, \sigma_\theta^2 \sim \text{iid } N(\beta, \sigma_\theta^2), \quad i = 1, \ldots, m;$
- $\log(\sigma_y^2) \sim N(c + d \log(n), c^2), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1, \quad \pi(\delta_1, \delta_2) \propto 1, \quad \pi(\sigma_\theta^2) \propto 1, \quad \pi(c) \propto 1.$

Note that model YLLM uses a log-linear model for the sampling variance $\sigma_y^2$, and extends the model proposed by Souza, Moura and Migon (2009) for sampling variances by using $\log(n)$ and adding a random effect to the regression part in the model. The full conditional distributions for the Gibbs sampling procedure are given in the Appendix.

https://doi.org/10.2478/jos-2020-0031
Methods - Example 2

3. Fay-Herriot model using HB approach with sampling variance modeling

HB Model 4: Sugasawa, Tamae and Kubokawa (2017) model shrinking both means and variances, denoted as STKM:

- $y_i | \theta, \sigma^2 \sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, m$
- $d_i | \sigma^2 \sim \sigma^2 \chi^2_i, \quad d_i = n_i - 1, \quad i = 1, \ldots, m$
- $\theta | \beta, \sigma^2 \sim N(x_i \beta, \sigma^2), \quad i = 1, \ldots, m$
- $\pi(\sigma^2) \sim \text{IG}(a_i, b_i)\gamma_i$, where $a_i$ and $b_i$ are known constants, $a_i = O(1)$, $b_i = O(n_i^{-1})$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1$, $\pi(\sigma^2) \propto 1$, $\pi(\gamma) \propto 1$

Note that in STKM, for the inverse gamma model of $\sigma^2_i$, we choose $a_i = 2$ and $b_i = n_i^{-1}$ as suggested by Sugasawa et al. (2017). Ghosh et al. (2018) also used the same setting in their study of comparing HB estimators. The full conditional distributions for STKM can be found in Sugasawa et al. (2017).

Note that the Chi-squared sampling variance modeling $d_1 \sigma^2 \sim \sigma^2 \chi^2_i$ in the above HB Models 2-4 is based on normality and simple random sampling (Rivest and Vandal, 2002). For complex survey designs, the degrees of freedom $d_i$ may need to be determined more carefully. There is no sound theoretical result for determining the degrees of freedom (Dass et al., 2012). The approximation formula based on non-normal unit level errors provided by Wang and Fuller (2003) and the simulation based guidelines of Maples, Bell and Huang (2009) could be useful but require unit level data and an extensive simulation study. A careful determination of the degrees of freedom may provide a reasonably useful approximation. Moreover, Bayesian model fit analysis can also be helpful for model determination.

https://doi.org/10.2478/jos-2020-0031
Appendix - Example 2

Full conditional distributions and sampling procedure for YLLM

\[ [\theta, \beta, \gamma, \delta, \tau] \sim N(\gamma, \beta, \gamma, \delta, \tau) \]
\[ [\gamma, \beta, \gamma, \delta, \tau] \sim N_p \left( \sum_{i=1}^{n} x_i \gamma \right) \left( \sum_{i=1}^{n} x_i \beta \right), \gamma, \delta, \tau \left( \sum_{i=1}^{n} x_i x_i \right) \]
\[ [\gamma, \beta, \gamma, \delta, \tau] \sim IG \left( \frac{\pi}{2}, \frac{1}{2}, \sum_{i=1}^{n} \left( \gamma - x_i \beta \right)^2 \right) \]
\[ [\gamma, \beta, \gamma, \delta, \tau] \sim IG \left( \frac{\pi}{2}, \frac{1}{2}, \sum_{i=1}^{n} \left( \gamma - x_i \beta \right)^2 \right) \]
\[ [\gamma, \beta, \gamma, \delta, \tau] \sim IG \left( \frac{\pi}{2}, \frac{1}{2}, \sum_{i=1}^{n} \left( \gamma - x_i \beta \right)^2 \right) \]

We use Metropolis-Hastings rejection step to update \( \sigma_i^2 \):

1. Draw \( \sigma_i^{(t+1)} \) from \( IG \left( \frac{\pi}{2}, \frac{1}{2}, \sum_{i=1}^{n} \left( \gamma - x_i \beta \right)^2 \right) \).
2. Compute the acceptance probability \( \alpha (\sigma_i^{(t)}, \sigma_i^{(t+1)}) = \min \left\{ h(\sigma_i^{(t)}) / h(\sigma_i^{(t+1)}), 1 \right\} \)
3. Generate \( u \) from Uniform(0, 1), if \( u < \alpha (\sigma_i^{(t)}, \sigma_i^{(t+1)}) \), the candidate \( \sigma_i^{(t+1)} \) is accepted, \( \sigma_i^{(t+1)} = \sigma_i^{(t)} \); otherwise \( \sigma_i^{(t+1)} \) is rejected, and set \( \sigma_i^{(t+1)} = \sigma_i^{(t+1)} \).

https://doi.org/10.2478/jos-2020-0031
Results - Example 2

Survey Methodology
Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

https://doi.org/10.2478/jos-2020-0031
5. Conclusion

In this paper, we compare the model-based estimates under the Fay-Herriot model when sampling variances are smoothed and modeled. As in Hidiroglou et al. (2019), our results indicate that the Fay-Herriot model can provide great improvement for the direct survey estimates for LFS rate estimation, even though more complex models such as unmatched models or time series models could be used (e.g., You, 2008). Among all the estimators, FH-EBLUP and FH-HB using smoothed sampling variances perform the best in terms of ARE and CV reduction. Both FH-EBLUP and FH-HB using direct sampling variance estimates perform the worst. For HB modeling approach, both YLLM and STKM perform very well and are better than YCM, and YLLM is slightly better than STKM in our study. Thus if direct sampling variance estimates are used, YLLM or STKM model is suggested. Alternatively, smoothed sampling variances should be used in the Fay-Herriot model to overcome the sampling variance modeling difficulty as discussed in Section 3. The smoothed sampling variances based on the GVF model given by (2.2) in Section 2 can perform very well as shown in our study.

https://doi.org/10.2478/jos-2020-0031
General Comments

• State the conclusion concisely and avoid overstatements

• Concise and informative **headings** (and subheadings) helps organizing the manuscript

• Be careful with notation and definition of variables

• Check the journal’s word limit for a manuscript
References


https://ugr.ue.ucsc.edu/sites/default/files/jyi_guide_to_scientific_writing.pdf