Optimal Sample Size Allocation to Mixed Modes: A Case Study Using the Residential Energy Consumption Survey

Peter Frechtel
Phillip S. Kott
Outline

- The general problem
- The problem as faced during the 2015 Residential Energy Consumption Survey
- Description of mixed mode allocation method
- Simulation study
- Discussion/Next Steps
The General Problem

- In many large-scale surveys, some variables are more expensive to collect than others
  - National Health and Nutrition Examination Survey (NHANES): physical examinations (expensive) and interview questions (less expensive)
  - National Survey on Drug Use and Health (NSDUH): clinical interviews (expensive) and non-clinical interviews (less expensive)
  - Residential Energy Consumption Survey (RECS): measured square footage (expensive) and interview questions (less expensive)

- If a mixed mode approach is adopted, how much sample should be allocated to each mode?
What happened during the 2015 RECS

- Original plan: face-to-face interviews only
- ~2,400 face-to-face interviews completed before this mode was abandoned
- ~3,000 web/mail interviews completed after face-to-face mode was abandoned
- Measured square footage was imputed for all w/m cases
- The imputation model was pretty good: there were plenty of good covariates with high item response rates. $R^2 = 0.73$.
- If we had planned the mixed mode approach all along, what would we have done?
Assume fixed data collection budget of $200K
Assume costs per completed case: $25 for web/mail and $500 for face-to-face
Minimize the variance of the estimate of mean measured square footage
- Not of primary interest to the Energy Information Administration (EIA), but it's a place to start.
(For RECS insiders only) Restrict analysis to single-family detached homes
Use double sampling and regression estimation (Legg & Fuller, 2009)¹ and Phil’s instructions:

Mixed Mode Allocation Method (2)

1) Regress measured square footage on a subset of the inexpensive variables, using 2005 RECS public-use data (reuse imputation model!)

2) Save predicted values and residuals
   1) Variance of predicted values = $\sigma_{\hat{y}}^2$
   2) Variance of residuals = $\sigma_{\varepsilon}^2$
Mixed Mode Allocation Method (3)

3) List possible values of sample size $n$ and subsample size $m$

1) Let $m$ vary from 100 to 400
2) Then $n = 8,000 - 19m$ (from cost equation)

4) Plot $\frac{\sigma^2_{\hat{m}}}{n} + \frac{\sigma^2_{\epsilon}}{m}$ as a function of $m$
Mixed Mode Allocation Method (4)

- Variance is minimized where \( \frac{m}{n} = \frac{\sigma^2}{\sqrt{500}} = \frac{\sigma^2_{\hat{\varepsilon}}}{\sigma^2_{\text{hat}}/\sqrt{25}} = 0.16 \)

Variance of Mean Measured Square Footage as a Function of Subsample Size
Simulation Study (1)

- Done to confirm that the results match the theory: the results depend on whether the model fits the population data

- For each value of $m$ and $n$ near the theoretical minimum:
  - Draw 10,000 samples and subsamples from the cases on the public-use dataset
  - Fit the regression model using the data in the subsample
Simulation Study (2)

- Estimate mean measured square footage as \( \bar{y} = \frac{1}{n} \left( \sum_{i \in M} y_i + \sum_{j \in (N - M)} \hat{y}_j \right) \), where
  - \( M \) is the set of observations in the subsample
  - \( y_i \) is the measured square footage of subsample unit \( i \)
  - \( N \) is the set of observations in the full sample
  - \( \hat{y}_j \) is the model-predicted measured total square footage for sample unit \( j \)

- Calculate the mean squared error and bias of the 10,000 point estimates (the bias turned out to be trivial)

- Plot the simulation-based variances on the same axes as the theoretical variances
Theoretical Variance vs. Empirical Variance

Theoretical Variance vs. Empirical Variance

Theoretical Variance
Empirical Variance

Subsample Size \( (m) \)
Conclusions

- In future RECS, should less data be collected face-to-face?
  - assumes that estimating the mean measured square footage is an important goal of the survey, and that the cost estimates were approximately correct
- Everything depends on the relative cost estimates and the quality of the model
- Playing with the model mattered!
- If we accounted for clustering, perhaps even a smaller proportion of the cases should have been face-to-face.
- If I can do this, you can do this. (It might not work as well for your survey, though.)
Next Steps

- How sensitive is this to the relative cost estimates?
- Minimize cost given a fixed variance
- Let both cost and variance vary and let the client decide where we belong on the curve
- Incorporating weights and sample design
Contact Information

Peter Frechtel: frechtel@rti.org
Phil Kott: pkott@rti.org