Estimation of Uncertainty: Confidence Intervals and Standard Errors

by the

Hard-to-Reach Population Methods Research Group*

Information available at
http://www.hpmrg.org/
http://www.hpmrg.org/workshop

*The project has been supported by the Presidents Emergency Plan for AIDS Relief (PEFPAR) through the US Centers for Disease Control and Prevention (CDC) under the terms of Cooperative Agreement U2GPS001468-5.
Hard-to-Reach Population Methods Research Group

- Ian E. Fellows, Fellows Statistics
- Lisa G. Johnston, Tulane University, UCSF
- Krista J. Gile, University of Massachusetts - Amherst
- Corinne M. Mar, University of Washington
- Mark S. Handcock, UCLA
Outline of Presentation

1. Link-Tracing Hard-to-Reach Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Discussion
Standard Survey Sampling

Stylized description

- Choose a population of interest and a population characteristic of interest $\mu$
- Determine the sampling frame: $i = 1, \ldots, N$ sample units.
- Choose variables to measure on them:
  \begin{itemize}
  \item outcome $z_i$, $i = 1, \ldots, N$
  \item control variables $x_i$, $i = 1, \ldots, N$
  \end{itemize}
- Choose a sampling design:
  e.g., simple random sampling, stratified sampling on $x$, stratified sampling on $z$
- Choose a sample of units $i = 1, \ldots, n$ and collect data on the sampled units
- Estimate the population characteristics of interest based on the sample
The level of certainty of the estimate for $\mu$ is determined by

- the true population from which the sample is drawn
- the chosen sampling design (e.g., sample size, seeds)
- **Sampling Variability**: the random or chance choice of sampled units
- **Representation** of the population by the sample:
  - the relationship between the defacto sampling frame and the population
  - the mechanism of non-observation
  - randomness in each sample
- **Measurement** of the variables of interest:
  - within the population
  - within the sample
Total Survey Error

Figure 3. Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (Groves et al. 2004).

The book offers examples of situations in Industry and Occupation coding where continuous quality improvement efforts decrease the total survey error. The import of this discussion in the book is the raising, for the first time within a survey error format, the notion that 'fitness for use' may be the most inclusive definition of quality. The authors offer no set of measurement approaches for 'fitness for use' and follow with a more traditional treatment of survey errors.

The reason that fitness for use is difficult to measure is that the notion encompasses not only the total survey error but also the qualitative nonstatistical dimensions mentioned earlier. Also, 'fitness for use' can vary within surveys since most of the surveys are multipurpose and multiestimate. 'Fitness for use' is a notion invented by Juran (1951) for quality in industry, while Deming used both 'usefulness' and 'fitness for purpose' for survey work. Also, Mahalanobis (1956) stated that 'statistics must have purpose.' This slight difference perhaps reflects the role of the customer in these two camps during the 1940s and 1950s.

The 2004 text Survey Methodology is organized around a total survey error framework, with an attempt to link the steps of survey design, collection, and estimation into the error sources. Figure 3 is a slight adaptation of the figure used in that text, which contains a term for the gap between the concept and the measure, labeled 'validity,' borrowing from psychometric notions of true score theory. In the original text, the figure is called 'Survey lifecycle from a quality perspective.' Chapters of the book link theory illuminating the causes of various error sources to design options to reduce the errors and practical tools that survey researchers use to implement the design options. The book attempts to note that two separate inferential steps are required in surveys — the first...
Estimation

• Goal: Estimate the population mean of $z$:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} z_i$$

where

$$z_i = \begin{cases} 1 & i \text{ has the characteristic} \\ 0 & i \text{ does not have the characteristic.} \end{cases}$$

• Sample indicators

$$S_i = \begin{cases} 1 & i \text{ sampled} \\ 0 & i \text{ not sampled} \end{cases}$$

• Inclusion probabilities

$$\pi_i = P(S_i = 1) \quad i = 1, \ldots, N$$

e.g. simple random sampling

$$\pi_i = n/N \quad i = 1, \ldots, N$$
Point Estimates from Design-Based Inference:

- Goal: Estimate proportion “infected”:
  \[
  \mu = \frac{1}{N} \sum_{i=1}^{N} z_i
  \]
  where
  \[
  z_i = \begin{cases} 
  1 & i \text{ infected} \\
  0 & i \text{ uninfected}.
\end{cases}
  \]

- Horvitz-Thompson Estimator:
  \[
  \hat{\mu} = \frac{1}{N} \sum_i S_i \pi_i z_i
  \]
  where
  \[
  S_i = \begin{cases} 
  1 & i \text{ sampled} \\
  0 & i \text{ not sampled}
\end{cases}
  \]
  \[
  \pi_i = P(S_i = 1).
  \]
Point Estimates from Design-Based Inference

• Goal: Estimate proportion “infected”:

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} z_i \]

where

\[ z_i = \begin{cases} 
1 & i \text{ infected} \\
0 & i \text{ uninfected}. 
\end{cases} \]

• Hajek Estimator:

\[ \hat{\mu} = \frac{\sum_{i} S_i z_i}{\sum_{i} S_i \pi_i} \]

where

\[ S_i = \begin{cases} 
1 & i \text{ sampled} \\
0 & i \text{ not sampled} 
\end{cases} \]

\[ \pi_i = P(S_i = 1). \]
Hajek Estimator

- The Hajek is useful when the population size $N$ is not known.
- The Hajek is better when $z$ is weakly or negatively correlated with $\pi_i$.
- **The key point:** Each estimator requires $\pi_i = P(S_i = 1) \quad \forall i : S_i = 1$
- We often need to model the sampling process to estimate these inclusion probabilities.
Volz-Heckathorn Estimator

- Approximate $\pi_i$ by $d_i$ based on a repeated-sampling model for RDS
- Assume $\pi$ is proportional to degree, $d_i$
- Volz-Heckathorn (RDS-II) Estimator:

$$\hat{\mu}_{VH} = \frac{\sum_i S_i z_i}{\sum_i S_i d_i}$$
Gile’s Sequential Sampling Estimator

- Approximate $\pi_i$ by $\hat{\pi}_i$ based on a successive-sampling model for RDS
- Gile’s Sequential Sampling (SS) Estimator:

$$\hat{\mu}_{SS} = \frac{\sum_i S_i z_i}{\sum_i S_i \pi_i}$$
Standard Error Estimation in Standard Surveys

Hajek Estimator:

\[ \hat{\mu} = \frac{\sum i \frac{S_i}{\pi_i} z_i}{\sum i \frac{S_i}{\pi_i}} \]

where

\[ S_i = \begin{cases} 1 & \text{i sampled} \\ 0 & \text{i not sampled} \end{cases} \quad \pi_i = P(S_i = 1). \]

- The only random thing is the \( S_i \).
- If we knew the (joint) distribution of \( S_1, S_2, \ldots, S_n \) we could compute the distribution of \( \hat{\mu} \).
- We can compute the standard error as the standard deviation of this distribution.
- For many standard survey designs, the \( S_i \) are independent, so \( \pi_i = P(S_i = 1) \) is enough.
- For standard errors have been worked out for many standard designs
More realistic designs

We need to know the (joint) distribution of $S_1, S_2, \ldots, S_n$:

- Clustered or multi-stage sampling designs
  - For these the clustering means the $S_i$ are dependent
  - In practice, software uses a simple first-stage-only approximation
  - Most large surveys do not release enough information on the design to improve on this
- Usually the formula is complicated to compute
  - They contain constants they themselves need to be estimated
  - Most software uses Taylor series expansion formulas to approximate the standard errors
An alternative: The Bootstrap

Idea: If we can simulate from the sampling process we can approximate the standard error from the simulations

Algorithm:

- Simulate $M = 10000$ sample (from the same process that generated the one we have)
- For each sample $m = 1, \ldots, M$, compute the estimate $\hat{\mu}_m$ (e.g., VH)
- Use the empirical standard deviation of $\{\hat{\mu}_m\}_{m=1}^M$ as an estimate of the standard error

$$s.e.(\hat{\mu}) = \sqrt{\frac{1}{M} \cdot \sum_{m=1}^{M} (\hat{\mu}_m - \bar{\hat{\mu}}_m)^2}$$
Bootstrap: Real world

Problem: We don’t know the true population and the actual sampling process, and so approximate them from the sample

Real Algorithm:

- Approximate the population its variables (e.g., $z_i, d_i$) from the sample
- Approximate the sampling process as best we can from what we know
- Simulate $M = 10000$ samples from approximate population using the approximate process
- For each sample $m = 1, \ldots, M$, compute the estimate $\hat{\mu}_m$ (e.g., VH)
- Use the empirical standard deviation of $\{\hat{\mu}_m\}_{m=1}^M$ as an estimate of the standard error
Application to RDS Error Estimation:

- Arithmetic mean
  - We can use the standard formula (assumes SRS)
  - More realistic to use the Gile bootstrap
- Salganik-Heckathorn (RDS-I)
  - Use a bootstrap where you divide the sample into recruiter-recruitee dyads:
    * Randomly select seeds (i.e., wave 0)
    * Randomly select a dyad where the recruiter has the same value of $z_i$ as the current wave. The next wave has the same value of $z_i$ as the recruitee in the dyad.
    * Repeat until the sample size is achieved
  - This is the bootstrapped sample. Repeat $M$ times.
- Volz-Heckathorn (RDS-II)
  - Approximate the RDS by with-replacement sampling
  - Use the Taylor Series expansion for that
Application to RDS Error Estimation:

- Gile’s Sequential Sampling (SS)
- Use a much more realistic bootstrap
  - Simulate Population
    * Estimate $z$ by $d$ distribution
    * Estimate infection mixing matrix by $z$
  - Simulate sequential without-replacement sampling
    * Choose recruit $z$ according to mixing matrix
    * Choose recruit $d$ by successive sampling
    * Update available population and mixing matrix
  - Compute SS Estimates
Performance of Gile’s Bootstrap

Table 1: Observed (simulation) standard errors of estimates, and average bootstrap standard error estimates, along with coverage rates of nominal 95% and 90% confidence intervals for procedure given in Section 1 for varying sample proportion and activity ratio $w$, and for initial sample selected either independent of infection (“No” bias) or all from within the infected subgroup (“Yes” bias). Observed standard errors are based on 1000 samples. Bootstrap standard errors are the average bootstrap standard error estimates over the same 1000 samples. Nominal confidence intervals are based on quantiles of the Gaussian distribution.

<table>
<thead>
<tr>
<th>%</th>
<th>homoph.</th>
<th>initial sample</th>
<th>SE observed</th>
<th>SE bootstrap</th>
<th>coverage 95%</th>
<th>coverage 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>R</td>
<td>$w$</td>
<td>bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>1</td>
<td>No</td>
<td>0.0212</td>
<td>0.0218</td>
<td>94.3%</td>
</tr>
<tr>
<td>70%</td>
<td>5</td>
<td>1.8</td>
<td>No</td>
<td>0.0087</td>
<td>0.0090</td>
<td>95.9%</td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>1</td>
<td>Yes</td>
<td>0.0211</td>
<td>0.0224</td>
<td>75.9%</td>
</tr>
</tbody>
</table>
Performance of Gile’s Bootstrap

- Performs well across differential activity ($w$) and sample fraction
- Performs well with homophily
- Unreliable when seeds biased.
Comparison of Variance Estimators:

Rules-of-thumb: How well do the estimators measure the actual sampling uncertainty?

- The analytic formulas tend to underestimate
- The Salganik bootstrap tends to underestimate if the sampling has not reached equilibrium
- The Gile bootstrap tends to underestimate if the homophily is large.
- In general the Gile bootstrap is the most credible and is preferred.
Other sources of uncertainty:

In measuring the total survey error we can discuss many possibilities:

- Sampling Variability: covered above
- Representation of the population by the sample
- Measurement of the variables of interest: