Discussion of

How Errors Cumulate: Two Examples

Jill A. Dever, PhD
jdever@rti.org

2018 Morris Hansen Award Lecture
Washington, DC

Key Theme from Presentation

“Weighting can improve things, but

representative data are better”

(Tourangeau, a few minutes ago)

The Dreaded "Bias" Word

Coverage Bias for Internet Surveys (Pew Research Center 2018a-c)

- ~2 of every 3 adults have broadband service at home
- Smartphone coverage ~77%
- Methods to access internet ⬆; ~11% with no use

Selection and Nonresponse Bias

- Confounded with nonprobability survey data
- Post-sampling nonresponse bias available for web panels

Commentary on “The Gold Standard”

Choice gold standard can be difficult:

- NHIS, BRFSS estimates for flu vaccination (Dever et al. forthcoming)
- CPS differences by month (Nadimpalli et al. 2004)
- Combine to strengthen (Schenker & Raghunathan 2007)

Differs by type of estimators:

- Estimated totals vs. ratio estimates (Dever & Valliant 2016)
- Univariate vs. multivariate statistics (Amaya & Presser 2016)

Estimation with Survey Data

General purpose weights:

- Known to produce efficient estimates for some <u>but not all</u> estimates for data from probability-based surveys
  - *AAPOR Task Force on An Evaluation of 2016 Election Polls in the U.S.* (Kennedy et al. 2017)

Variance estimation with nonprobability surveys:

- Speculation that replicate estimates are “the way to go”

Multiple Sources for Web Surveys

- **Opt-in web panels**
- **Pop-up Surveys**
- Twitter
- Facebook
- Snapchat
- Mechanical Turk
- SurveyMonkey
- Web-scraping
- Data warehouses

➡

Convenience, Matched, or Network (Baker et al. 2013)

Different TSE properties,

e.g., different coverage

Is One Source Adequate for Population Inference?

"Poor population coverage is difficult to overcome" (Valliant 2018)

Dual-frame estimation (e.g., Lohr & Raghunathan 2017)

- Landline random-digit-dial surveys no longer exist
- Targeted frames for specialized populations, e.g., surname lists

$$\hat{t}_y = \sum_{s_{A \cap B}} \lambda_k \hat{y}_{Ak} + \sum_{s_{B \cap A}} (1 - \lambda_k) \hat{y}_{Bk}$$

= *"A" weighted estimate* + *"B" weighted estimate*

where $\boldsymbol{\lambda_k}$ (≤1) is the composite factor

Hybrid Estimation for Population Inference

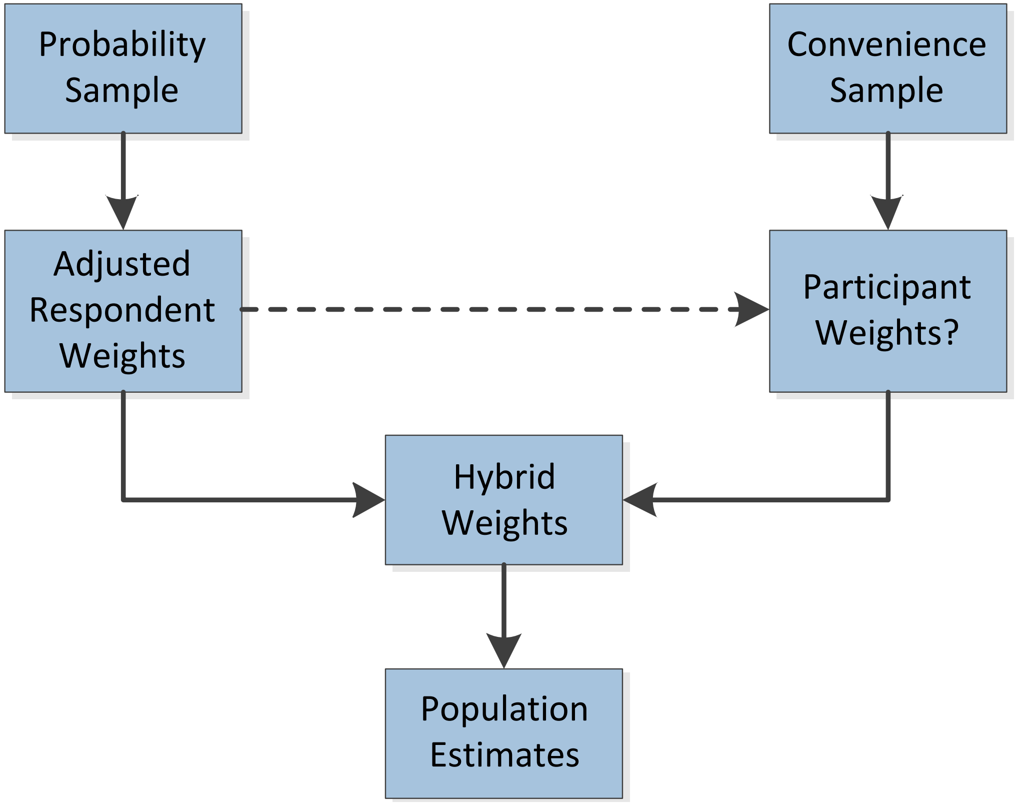
Combine Probability and Nonprobability Data

- Opinions of marijuana usage among adults (Allen et al. 2018)
- Methods for creating hybrid estimates using:
  - data on injury outcomes from vehicle crashes (Elliott 2009)
  - surveys of military caregivers (Robbins et al. 2017)
  - generic surveys characteristics (Elliott & Haviland 2007)

Combine Multiple Nonprobability Data Sources

- Social media to access LGBTQ youths (Berzofsky et al. in press)
- Social media to access marijuana users (Kott 2018)

Hybrid Estimation for Population Inference



Hybrid Estimation – Nonprobability Weights

- Quasi-randomization “pseudo” weights
  - Propensity scores (Valliant & Dever 2018, 2011)
  - Statistical matching (Ho et al. 2007, 2011; Dever 2018)
  - Bayes method (Robbins et al. 2017; Elliott 2009)
- Weight calibration
- Superpopulation “prediction” approach (Valliant et al. 2000)
- Multilevel regression & poststratification (Wang et al. 2015)

Informative covariates are critical (Mercer et al. 2018; Valliant 2018)

Hybrid Estimation – Additional Adjustments

- **Estimated-control calibration** (Dever 2010, 2018; Dever & Valliant 2016)
- **Adjustments for bias** (Brick et al. 2011)
- **Common support** (Dever 2018)

$$\hat{t}_y = \sum_{s_{A \cap B}} \lambda_k \hat{y}_{Ak} + \sum_{s_{B \cap A}} (1 - \lambda_k) \hat{y}_{Bk}$$

common support

$$+ \sum_{s_A} \hat{y}_{Ak}$$

survey-specific components

$$+ \sum_{s_B} \hat{y}_{Bk}$$

More Research is Needed into Hybrid Estimation

- Interplay between errors for each data source and among the data sources is critical
  - TSE for hybrid estimation
- Methods to maximize information from multiple sources
- Evaluate in the context of estimators

Congratulations Tex!