Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses

Stephen Fienberg
Department of Statistics, Machine Learning Department, and Cylab
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Outline

- What are weights for and where do they come from?
- To weight or not to weight: That is the question?
  - Issue for analysis of released microdata.
- What hazards do weights pose for confidentiality protection?
  - How can problems be averted or mitigated?
- Prescriptions for statisticians?

Survey Analysis

- **The core of official statistics activities in many countries.**
- **Frequentist motivation and analyses.**
  - **"Weighted" analyses are central to methodology for population estimates.**
  - **Debate over design- vs. model-based estimation.**
- **Statistical models are now commonplace in statistical agencies today.**

Sampling Weights

- **Weights arise in sampling settings when units are selected with unequal probabilities from a finite population:** $w_i = \frac{1}{p_i}$.
- **But in typical official statistics setting weight are typically product of 3 components:**
- $$w_i = \frac{1}{p_i} \times (\text{nonresponse}) \times (\text{poststratification})$$

Model based: MAR **Population controls**

To Weight or Not To Weight?

- **No debate about weight, at least for the probability of selection, about estimating population aggregates.**
- **Issue comes when we introduce statistical models:**
  - **e.g., weighted vs. unweighted regression models.**
  - **How do we combine models and sampling theory?**
    - **Naïve weighting**
    - **Estimating equations**

Arguments for Weighting

- **Almost always justified only from frequentist, design-based perspective:**
  - **Correct frequentist properties under finite population setup for aggregates.**
  - **Pays attention to sample space in evaluating contributions to estimation of model parameters.**
  - **Robustness.**

Hansen, et al (1983); Kalton; Pfeffermann (1988, 1996), Pfefferman et al. (1998).

Arguments Against Use of Weights

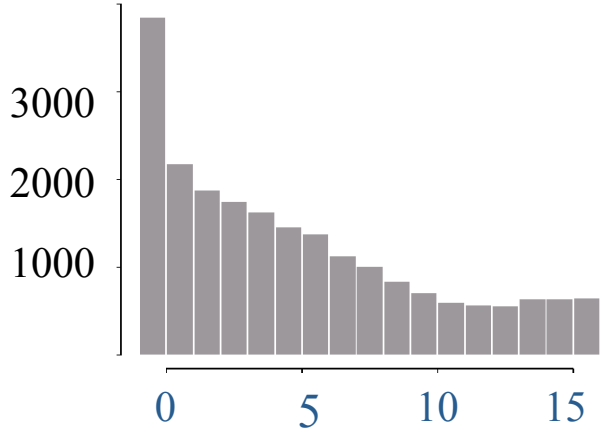
- **Stratification and clustering belong in model.**
  - **Weight components are insufficient for this purpose.**
- **Weights have non-sampling components, due to adjustment for *nonresponse*, and *post-stratification*.**
  - **Why are we incorporating these components into this way in likelihood?**
- **Estimating equation ideas collapse for complex hierarchically-structured models with latent variables.**

Disability Among Elderly

- **National Long Term Care Survey of US Medicare eligible population (aged 65+).**
  - **6 waves: 1982, 1984, 1989, 1994, 2004.**
- **Models for disability based on 16 binary measures called activities of daily living (ADL) and instrumental activities of daily living (IADL):**
  - **eating, getting in/out of bed, getting around inside, dressing, bathing, using a toilet, doing heavy house work, doing light house work, doing laundry, cooking, grocery shopping, getting about outside, traveling, managing money, taking medicine, telephoning.**

Disability Among Elderly—II

- **Data on *N*=21,574 individuals:**
  - **65,536 cells (3,152 non-zero):**
    - **82% of cell counts < 5.**
    - **4% of cell counts > 20.**
    - **18% with no disabilities.**
    - **3% with all 16.**



- **Bayesian hierarchical mixed membership models with informative latent structure:**
  - **Smoothing is crucial feature. Levels of hierarchy matter, as do choices of Dirichlet prior parameters.**
    - **Erosheva, Fienberg, Joutard (2007, *AOAS*)**

Grade of Membership Model

- ***N*** **individuals;** ***K*** **extreme profiles;** ***J*****=16 items.**
- **Membership scores** $\lambda = (\lambda_1, ..., \lambda_K)$ **define how close individual is to each extreme profile.**
- **Probability distribution of response *j*, given full membership in extreme profile *k*, is**

$$f(\mathbf{x}_{ij} \mid \lambda_{ik} = 1;\ \theta_{kj}) = Binomial(\theta_{kj}),$$

$$Pr(\mathbf{x}_j \mid \lambda) = \sum_k \lambda_k \cdot f(\mathbf{x}_j \mid \theta_k).$$

  - **Membership scores are random, i.e.,** $\lambda \sim D_\alpha$.
- **Weighted analyses fail here!**

Usual Privacy Questions

- **How do we protect the privacy of the responses represented by 2^{16} contingency table?**
- **What about the sampling variability in the sample design?**
- **What about the model variability? And the variability associated with model selection?**
  - **Cynthia and Adam's issue but not addressed here.**

Policy Issue and Full Question

- **Has disability been declining over time?**
  - **Implications for Social Security and Medicare planning.**
  - **How should we examine this question given the longitudinal nature of the data?**
  - **Role of sampling weights?**
  - **Privacy questions now relate to**
    - **Protecting privacy of $(2^{16})^6$ cell contingency table**
    - **Protecting privacy of weights**

Confidentiality and Weights

- **At last!**
- **Question 1: What information do weights provide to intruder intent on identifying individuals in the sample?**
- **Question 2: Do they increase the probability of disclosures? How?**

Ton de Waal and Leon Willenborg (1997) *J. Off. Statistics*, 13.

de Waal and Willenborg

- **Sampling weights can provide indirect identifying information regarding membership in substrata defined by sets of post-stratification variables.**
- **Idea is that intruder will have accurate information on post-stratification population counts, and can use numbers of sample people with given weights to *match* individuals with post-strata.**
- **How big is the problem? Is it different in Netherlands compared with the U.S.?**

De Wall/Willenborg Prescription

- **Subsampling**
  - **To reduce probability of correct matches**
- **Noise addition to weights**
  - **Ditto**
  - **Also messes up original notation of population controls!**

Rubin on Multiple Imputation

- **Compute relevant posterior distribution and generate multiple samples from it.**
- **These are synthetic samples and thus (he argues) they automatically solve the confidentiality problem.**
- **You might use weights to construct the posterior (although many of us wouldn't) but there is no requirement that the multiply imputed data be weighted!**

My Prescription

- **Get rid of population controls and thus the biggest part of the confidentiality concerns from sampling weights.**
- **Stop insisting that model-based analyses incorporate weights.**
- **Think about new approaches to survey design that deal *de novo* with confidentiality concerns as well as analytical goals; not just traditional sample efficiency goals.**
  - **Share real design information.**
  - **Address disclosure problems with cluster sampling at design stage.**

Summary

- What are weights for and where do they come from?
- To weight or not to weight: That is the question?
- What hazards do weights pose for confidentiality protection?
  - How can problems be averted or mitigated?
- Prescriptions for statisticians
  - De Waal and Willenborg
  - Rubin
  - Fienberg