Revista Colombiana de Estadística
Edición para autor(es)

Statistical Graphics for Survey Weights

Susanna Makela[1,a], Yajuan Si[1,b], Andrew Gelman[1,c]

[1]Department of Statistics, Columbia University, New York, USA

Resumen

Ponderación de la muestra se utilizan para corregir las diferencias conocidas entre la muestra y la población debido al diseño de la muestra, la falta de respuesta, la falta de cobertura, y otros factores. Sin embargo, consideraciones prácticas a menudo resultan en pesos que no se han construido de una manera sistemática. Los métodos gráficos pueden ser útiles en la comprensión de ponderaciones de la encuesta complejos y sus relaciones con otras variables del conjunto de datos, sobre todo cuando poca información sobre la construcción de los pesos está disponible. Herramientas gráficas también pueden ayudar en el diagnóstico, incluyendo la detección de valores atípicos y pesos extremos. Aplicamos nuestros métodos en el estudio de Familias Frágiles y Bienestar Infantil, un estudio longitudinal en curso.

***Palabras clave*:** Encuesta por muestreo, esquema de muestreo, gráficas, diagnósticos.

Abstract

Survey weights are used for correcting for known differences between the sample and population due to sampling design, nonresponse, undercoverage, and other factors. However, practical considerations often result in weights that are not constructed in a systematic fashion. Graphical methods can be useful in understanding complex survey weights and their relations with other variables in the dataset, particularly when little to no information on the construction of the weights is available. Graphical tools can also assist in diagnostics, including detection of outliers and extreme weights. We apply our methods to the Fragile Families and Child Wellbeing Study, an ongoing longitudinal survey.

***Key words*:** Sample survey, sampling scheme, graphics, diagnostics.

[a]. E-mail: susanna@stat.columbia.edu
[b]. E-mail: ysi@stat.columbia.edu
[c]. E-mail: gelman@stat.columbia.edu

Introduction

Sample survey data typically differ systematically from their target populations. The differences arise from sampling design, unbalanced coverage of certain subpopulations, and nonresponse. The standard approach to correct for the sample non-representativeness is weighting. Weights are typically constructed based on some combination of adjustment factors based on inverse selection probability (Horvitz & Thompson 1952), inverse response propensity (David, Little, Samuhel & Triest 1983), and poststratification ratios (Holt & Smith 1979) of the census distribution to the sample distribution. Hence an item's weight depends on the variables that affect its probability of inclusion in the survey. The design variables include, for example, number of telephones in the household (in a phone survey) and cluster size (in a probability-proportional-to-size design). The variables that affect the response propensity may include the individual's demographic and auxiliary information on previous waves, such as attrition rates, in a panel or longitudinal study. When it comes to poststratification, the key variables on demographics and design, such as income in a poverty measure survey that oversamples in poor neighborhoods, need to be accounted for the discrepancy with the target population.

In classical sampling theory, weights are used to generate design-consistent estimators. Assume w_i is the weight assigned to each unit i and y_i is one survey response variable of interest. The classical weighted ratio estimator (Horvitz & Thompson 1952, Hájek 1971) for the population mean is

$$\hat{\theta}^{\,\text{design}} = \sum_i w_i y_i \Big/ \sum_i w_i. \tag{1}$$

Adjustment for inverse response propensity is often done using the response propensity method, which applies the theory of propensity scores developed by Rosenbaum & Rubin (1983) for causal inference to survey nonresponse (David et al. 1983). Raking (Deming & Stephan 1940) is often used to match weighted sampling distribution in a survey to external population distribution, particularly when the joint population distribution across poststratification cells is unknown and only marginal distributions are available. Many recent developments in weighting methods and applications are discussed in Levy & Lemeshow (2013) and Lehtonen & Pahkinen (2004). Gelman (2007) and the associated discussions point out the problems of weights in practice.

In model-based approaches for sampling inference, the proper model specification of survey outcome should be conditional on the inclusion probabilities (Dunson 1983). It has been shown in various settings that taking account of the sample design in estimation will yield more efficient robust inferences (Zheng & Little 2003, Si, Pillai & Gelman 2014). If the weights are informative for the survey outcome, then ignoring the design information may result in biased conclusions. Moreover, model fitting and evaluation play important roles since survey practitioners are sensitive to model misspecification.

Therefore, to understand what weighting is doing in any particular example it can be useful to investigate the dependence between weights and the variables affecting inclusion into the sample, as well as the joint distribution of the weights and the survey responses of interest. In this article, we consider some graphical explorations of these relationships. When a few weights are extremely large (corresponding to a few units in the sample that represent a large share of the population), the weighted estimator will be highly noisy. Graphics can help reveal extreme cases and suggest potential transformations to improve precision. We are also interested in studying the weights' distributions for different waves in a longitudinal study, where the baseline weight serves as the anchor for all the follow-up weights. We demonstrate this visualization for a survey on which we are currently working.

Fragile Families and Child Wellbeing Study

The Fragile Families and Child Wellbeing Study is a longitudinal survey of children born in large U.S. cities between 1998 and 2000, many of whom were born to unmarried parents (Reichman, Teitler, Garfinkel & McLanahan 2001). The term "fragile families" refers to new unwed parents with their children and "the vulnerability of relationships within these families" (Reichman et al. 2001). The study provides data to help researchers understand the relationships in fragile families, particularly the role of the biological father, as well as the impacts on these relationships of public policies, especially those relating to welfare, paternity establishment, and child support. We briefly summarize the sampling scheme and construction of weights for the Fragile Families study; detailed descriptions can be found in Reichman et al. (2001) and Carlson (2008), respectively.

The sampling units for the Fragile Families study (Reichman et al. 2001) are live births between 1998 and 2000 in large U.S. cities. The sampling frame is multistage, sampling first cities, then hospitals, and finally births. The national sample consists of 16 cities selected from nine strata designed to generate a sample of cities varying widely in their public policies, such as welfare generosity and child support enforcement, that directly affect many fragile families. Hospitals within selected cities were sampled so as to ensure adequate coverage of non-marital births in the city, and within selected hospitals, all non-marital and marital births were selected until their respective quotas were achieved.

The weighting construction for the Fragile Families study (Carlson 2008) adjusts for three things: unequal probability of selection across the multistage sampling frame, nonresponse, and poststratification. Although the sampling unit is a birth, the weights are constructed at the mother level. One set of weights is generated for analysis at the city level and another for analysis at the national level. For the purposes of this paper, we use the national weights for the 16 sampled cities that allow analyses of births in the sample to be generalized to births in large U.S. cities. The weights are raked so that the weighted counts of births by mother's age, education, race/ethnicity, and marital status match known external counts: city counts for the city weights and national count for the national weights.

This process can approximate the adjustment when the joint cross-tabulation of these variables is available. The weights are also trimmed at an upper bound set as the mean weight plus four standard deviations to avoid extremely large values. To date, the study has collected five waves of data, and the weights include the baseline weights and four sets of follow-up weights.

Graphical Illustration

In this section, we investigate the contributions and patterns of survey weights via statistical graphics. We consider several variables that are used constructing the weights, several outcome variables of substantive interest, and the five sets of mother weights at the national level across the five waves in the Fragile Families study.

Plotting Weights vs. Raking Variables

The baseline national weights for mothers of the children in the sample are a product of inverse inclusion probabilities accounting for the sampling design, and nonresponse/poststratification factors accounting for unbalanced coverage. The poststratification variables include mother's age, race/ethnicity, education, and marital status. Only marginal distributions of these variables were available at the national level, and the adjustment was done by raking. We label these four variables as raking variables, and plot them vs. the survey weights.

Because the raking variables are categorical, simply plotting them against the weights would not yield easily interpretable or informative plots. Therefore, we divide the weights (on the logarithmic scale) into evenly-spaced bins. Within each bin, we calculate the proportion of respondents at each level of the given variable. We also include lowess fits to show overall trends and increase the readability of the plots, particularly for the categorical variables with more than two levels.

Intuitively, the weight values represent how many units each individual will represent in the population. The weights inside each bin have close values, but the correspondingly allocated sample sizes vary widely. The larger the weights, the more likely that the individuals with some characteristic are undercovered in the sample.

Figure 1 shows plots of the four raking variables and the log weights. In Figure 1a, we plot the binned weights against the proportion of mothers within each bin who are married to the father of their baby. Marital status is strongly related to the survey weights, with higher weights corresponding to higher proportions of mothers who are married. For mothers whose weights fall into the smallest bins, the proportion who are married is low; in fact, none of the mothers whose weights are in the two smallest bins are married. On the other hand, nearly all of the mothers whose weights are in the largest bins are married. This result agrees with our expectation since the Fragile Families study oversampled non-marital births.

In Figure 1c, we plot the binned weights against the proportion of mothers within each bin who are Hispanic, black Non-Hispanic, white non-Hispanic, and other. The racial/ethnic composition of mothers varies strongly across weight bins, with higher proportions of black non-Hispanic mothers in smaller weight bins and larger proportions of white non-Hispanic mothers in larger weight bins. Larger weight bins also contain a larger proportion of respondents aged 20-24 years (Figure 1d) than do smaller weight bins. These figures show that the Fragile Families study oversampled black non-Hispanic and young (under 18) mothers.

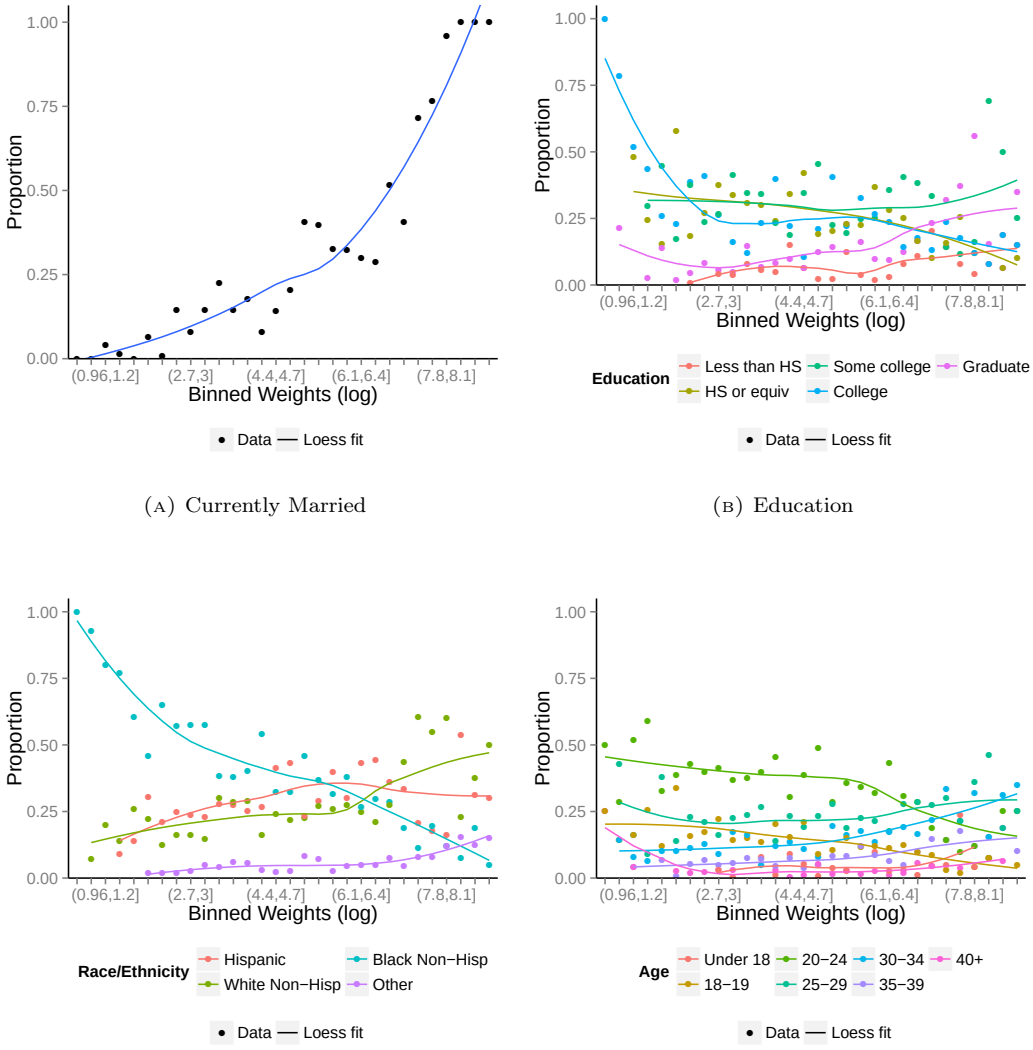
Figure 2 shows the square root of the sample size within each weight bin for the baseline survey when the weights are binned on the original scale. The sample sizes are roughly proportional to the inverse of the weights. There exist some extremely large weight values, and the corresponding sample sizes are extremely small. This warns us that the weighted estimator may be highly variable. The largest weights have already been trimmed: as explained in Carlson (2008), any weights exceeding the cutoff of four standard deviations away from the mean weight were set to this cutoff. Thus the largest weights observed in Figure 2 are smaller than the largest weights generated by the weight construction process.

Plotting Weights vs. Response Variables

To understand the relationship between the survey weights and the collected response or outcome variables of substantive interest, we again generate binned plots similar to those in Figure 1. In Figure 3, we bin the log weights and within each bin, calculate the proportion of respondents at each level of mother-reported health status for herself and the focal child. Similarly, in Figure 4, we bin the log weights and within each bin, calculate the proportion of respondents whose focal child is overweight or has asthma, the proportion of respondents who have received welfare benefits, and respondents' average annual household income. Whether the child is overweight is asked only in the nine-year follow-up wave, while asthma and health status are asked only in the one-year follow-up wave. Data on household income and whether the mother has received welfare benefits are available for every wave, but here we show them only for the one-year wave.

There is little statistical dependence between the weights and health-related variables, as can be seen in Figures 3, 4a, and 4b. This indicates that weighting should do little to change the distribution of health status within weight bins. Indeed, when we calculate the overall proportion of overweight children in the Fragile Families study, the unweighted proportion is 16.6 %, while the weighted proportion is 16.4 %. Similarly, the unweighted and weighted proportions of mothers reporting that their child's health is excellent are 65.5 % and 65.3 %, respectively.

However, the proportion of families who received welfare has a stronger relationship with the weights, with those at the extremes of the weight range being the least likely to receive welfare. Similarly, except for the high incomes in the smallest weight bins, larger weights are associated with higher incomes. In general, if the weights are informative for the survey outcome variables, they should be included in the analysis.



(A) Currently Married

(B) Education

(C) Race/Ethnicity

(D) Age

FIGURA 1: For four discrete raking variables in the Fragile Families study, we plot the proportion of respondents at each level of the given variable vs. binned baseline survey weights (log scale). The binned averages are smoothed by lowess curves. Sample size is high so we use a large number of bins (as indicated by the tick marks on the x-axes). A few of the tick marks are labeled to indicate the log weights in some of the bins; the total range of the weights is large, varying by a factor of approximately exp(8.5) or 5000.
HS = high school.

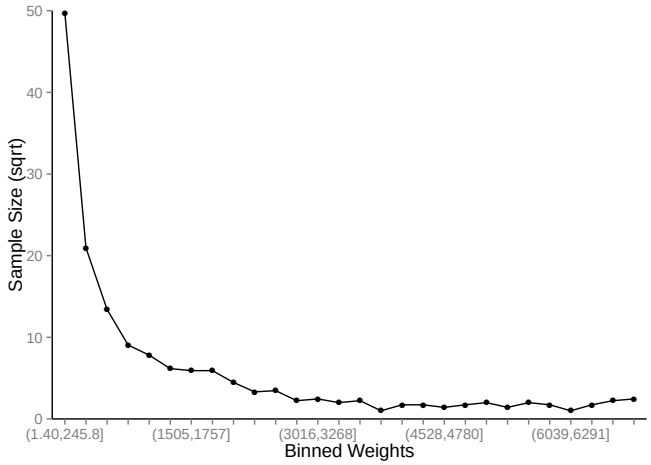


FIGURA 2: Square root of sample sizes by weight bin for baseline weights in the Fragile Families study.

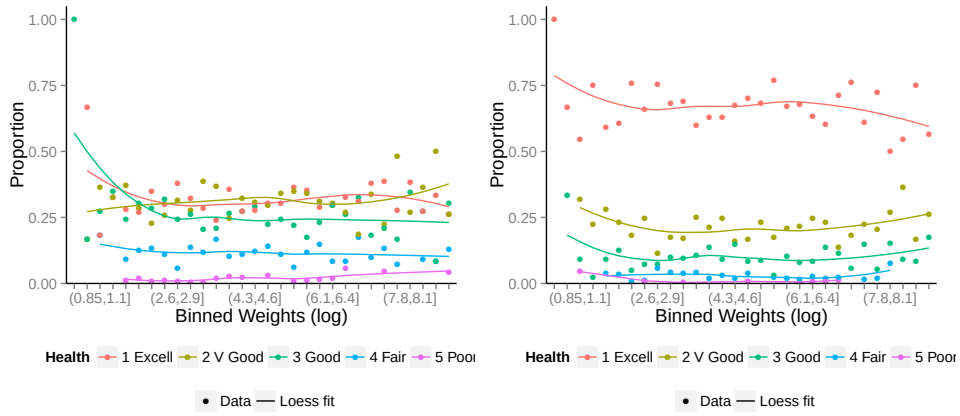
Weights Across Waves

In the Fragile Families study, because there is no subsampling at the various follow-up waves, constructing the follow-up weights consists mostly of nonresponse adjustments and re-raking the weights so that the weighted totals of births in subsequent waves add up to the baseline totals (Carlson 2008). Comparing summary statistics between the follow-up weights and the baseline weights can help detect extreme values or changes in the weights over time. Table 1 displays summary statistics for the weights at baseline and subsequent waves. The weights are subject to high variability. The sample size decreases, and the weights become larger in magnitude over time. This shift is reflected in the systematic increases in the mean, median, and 25 % and 75 % percent quantiles in Table 1. Figure 5 displays the density of the log weights, a plot that does not seem particularly revealing here but could be valuable in other settings. The one thing we do see here, in both Table 1 and Figure 5, is that the weights from Year 9 show relatively large shifts compared to those in the previous waves.

It might also be useful to track the changes in individual weights via scatterplots or a parellel coordinate plot. Further information could be obtained by constructing separate scatterplots or parallel coordinate plots for subsets of the population.

Discussion

This work is illustrative and preliminary rather than conclusive, but we feel it is valuable in illustrating the possibility of graphical methods for studying survey weights and their relationship with other survey variables. This provides an intuitive exploration on how the externally-supplied weights are constructed, the effects

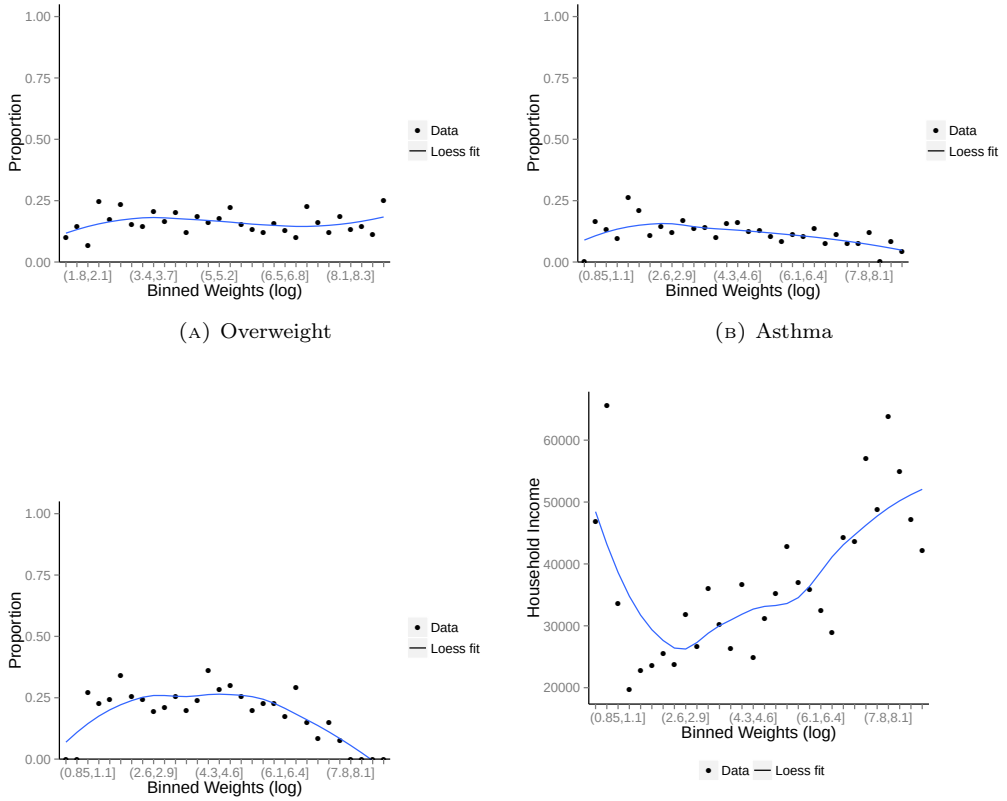


(A) Mother (B) Child

FIGURA 3: Mean response for each category of mother's report of her health status and the focal child's health status, plotted in bins of the log baseline weights, with lowess curves superimposed. See caption of Figure 1 for explanation of the x-axes.

Wave	N	min	25 %	50 %	mean	75 %	max	sd
Baseline	3442	1.5	22	82	330	290	7500	770
One-Year	3120	1.3	23	91	360	330	8100	850
Three-Year	3032	1.4	24	94	370	330	8400	860
Five-Year	3006	1.7	24	97	380	330	8000	880
Nine-Year	2655	3.8	31	111	430	370	9100	990

TABLA 1: Summary statistics of Fragile Family weights by survey wave. The average weight gradually increases to account for the decreasing number of people the sample. N = sample size.



(A) Overweight (B) Asthma

(C) Welfare (D) Income

FIGURA 4: Sample proportions of (A) children who are overweight, (B) children with asthma, and (C) families receiving welfare benefits, and (D) annual household income, all plotted vs. binned survey weights. Data for (A) are from the nine-year wave, and data for (B)–(C) are from the one-year wave.

of incorporating weights into analyses, and how to select an appropriate model. Weight trimming and smoothing are often applied arbitrarily. The relationship between weights and the design or raking variable will help find which variable significantly increases the variability and needs collapsing or pooling in its own distribution. When weights are informative for the response variables of interest, we should take them into account. Weighting adjustment in longitudinal surveys suffers from many practical issues, for example, how to build a robust model for response propensity score prediction. We recommend the use of data visualization in survey inference in concert with more complex analytical and computational approaches.

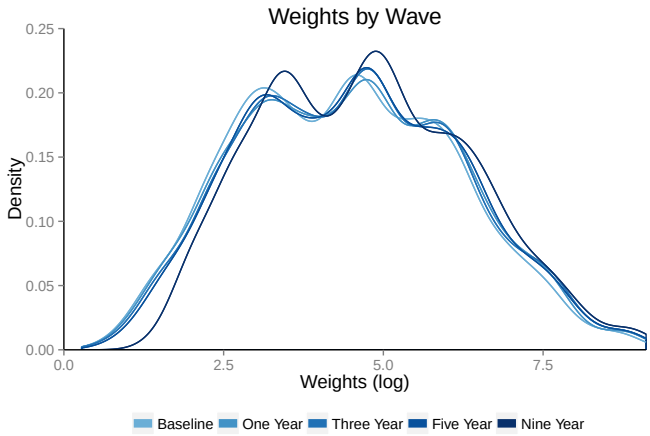


FIGURA 5: Probability density estimates of the log weights at each wave for the Fragile Families study. In this specific case we do not find this graph particularly revealing but we imagine such a plot could be useful in other contexts where a user is comparing the distribution of weights from two or more different surveys.

Referencias

Carlson, B. L. (2008), Fragile families & child wellbeing study: Methodology for constructing mother, father, and couple weights for core telephone surveys waves 1-4, Technical report, Mathematica Policy Research.

David, M., Little, R., Samuhel, M. E. & Triest, R. K. (1983), Nonrandom nonresponse models based on the propensity to respond, *in* 'Proceedings of the Business and Economic Statistics Section, American Statistical Association', pp. 168–173.

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics* **11**(4), 427–444.

Dunson, D. B. (1983), 'Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping', *Journal of the American Statistical Association* **78**, 803–805.

Gelman, A. (2007), 'Struggles with survey weighting and regression modeling', *Statistical Science* **22**(2), 153–164.

Hájek, J. (1971), Comment on "An essay on the logical foundations of survey sampling by D. Basu"., *in* V. P. Godambe & D. A. Sprott, eds, 'The Foundations of Survey Sampling', Holt, Rinehart and Winston, p. 236.

Holt, D. & Smith, T. M. F. (1979), 'Post stratification', *Journal of the Royal Statistical Society Series A* **142**(1), 33–46.

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite university', *Journal of the American Statistical Association* **47**(260), 663–685.

Lehtonen, R. & Pahkinen, E. (2004), *Practical methods for design and analysis of complex surveys*, 2nd edn, Wiley, West Sussex.

Levy, P. S. & Lemeshow, S. (2013), *Sampling of populations: methods and applications*, 4th edn, Wiley, New York.

Reichman, N. E., Teitler, J. O., Garfinkel, I. & McLanahan, S. S. (2001), 'Fragile Families: Sample and design', *Children and Youth Services Review* **23**(4/5), 303–326.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Si, Y., Pillai, N. & Gelman, A. (2014), 'Bayesian nonparametric weighted sampling inference', *Submitted to Bayesian Analysis* .

Zheng, H. & Little, R. J. (2003), 'Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples', *Journal of Official Statistics* **19**(2), 99–107.