Submitted to the Annals of Applied Statistics

CONDITIONAL RAKING

By Emily M. Kotnik

Loyola University Chicago

and

By Gregory J. Matthews

Loyola University Chicago

Some abstract

1. Introduction. Accurately analyzing data from samples with an unequal probability of being sampled provides a unique challenge that has been a topic of concern for statisticians for decades when stratified sampling is not an option. Stratified sampling has been one of the primary methods used in order to produce a more precise representation of the population (Lohr, 2010, p. 26) since it was first brought to the forefront of survey methodology in 1934 by Neyman. Neyman proposed stratified sampling and purposive selection as a method to ensure a sample that was representative of the population (Neyman, 1934). Stratified sampling is used to ensure the proportions of the sample are equal to those of the population to obtain very precise data on the population of interest, and as method in increasing efficiency of administering a survey (Lohr, 2010, p.74). Though stratified random sampling is a satisfactory method to increase precision, marginal population totals are necessary and stratification must be performed prior to sampling. In the event that stratified sampling is not possible or the data was collected at a previous date, several methods are used to correct samples of subjects with unequal probabilities of selection. The two primary methods of increasing precision are post stratification and raking. Both methods require additional data from a secondary data source. Post stratification partitions the sample based on demographic variables of interest and weights each partition by the known proportions from the population (Gelman, Little, 1997). Raking, first proposed in 1940 by Deming and Stephan under the name ”least squares adjustment” (Deming, 1940), was first used as a method to bring the marginal distributions from sample data to those of the known population distributions from the 1940 US Census (Brick, Montaquila, Roth, 2003) through a series of adjustments based on the known marginal totals until the sample frequencies converge with the population frequencies (Brick, Montaquila, Roth, 2003). The resulting frequencies from both post stratification and raking offered more precise joint and marginal totals while minimizing sample bias. Both post stratification and raking have limitations that have been acknowledged though few solutions have been proposed. Regarding post stratification, Holt and Smith argue that a post stratified estimate is not always more accurate than a simple random

Keywords and phrases: post-stratification, raking, sampling

sample. In the event that the sample is stratified into many categories, each with many levels, it may be not be possible to adjust each cell due to the large number of calculations required or empty cells in the frequency table (Holt, Smith 1979). Additionally, the population distribution of each variable in question must be known for accurate post stratification. If this is not the case, or if the population distribution is inaccurate, additional calculations must be performed to effectively model the variable in question (Little and Gelman, 1997) Brick, Montaquila, and Roth identified several problems with raking. The first arises from inconsistent population marginal frequencies. If multiple external data sources are used to provide population data in the event the necessary information is not available from one source, inconsistent counts may be found and result in inaccurate adjustments. The second problem identified with raking results from a large number of variables. Raking across tables with a large number of cells may result slow convergence or a failure to converge (2003). Additionally, raking may inaccurately adjust cells with few or no observations (Brick, Montaquila, Roth, 2003). Though post stratification and raking have their uses in correcting data collected from survey subjects with unequal probabilities of selection, there is an intermediate case that is not accounted for in either of these methods. In the event that some, but not all, joint population distributions are known from additional data sources, there is no readily available method to rake or post stratify the data. This paper proposes a method to accurately adjust the data to be reflective of the population.

2. Methods.

Consider a data set with n observations and p variables that is a sample from a population P. There are $k \leq p$ stratifying variables where there is at least some information about the joint population distribution of these k variables. That is we know based on auxilliary information the population counts of specified combinations of these k variables.

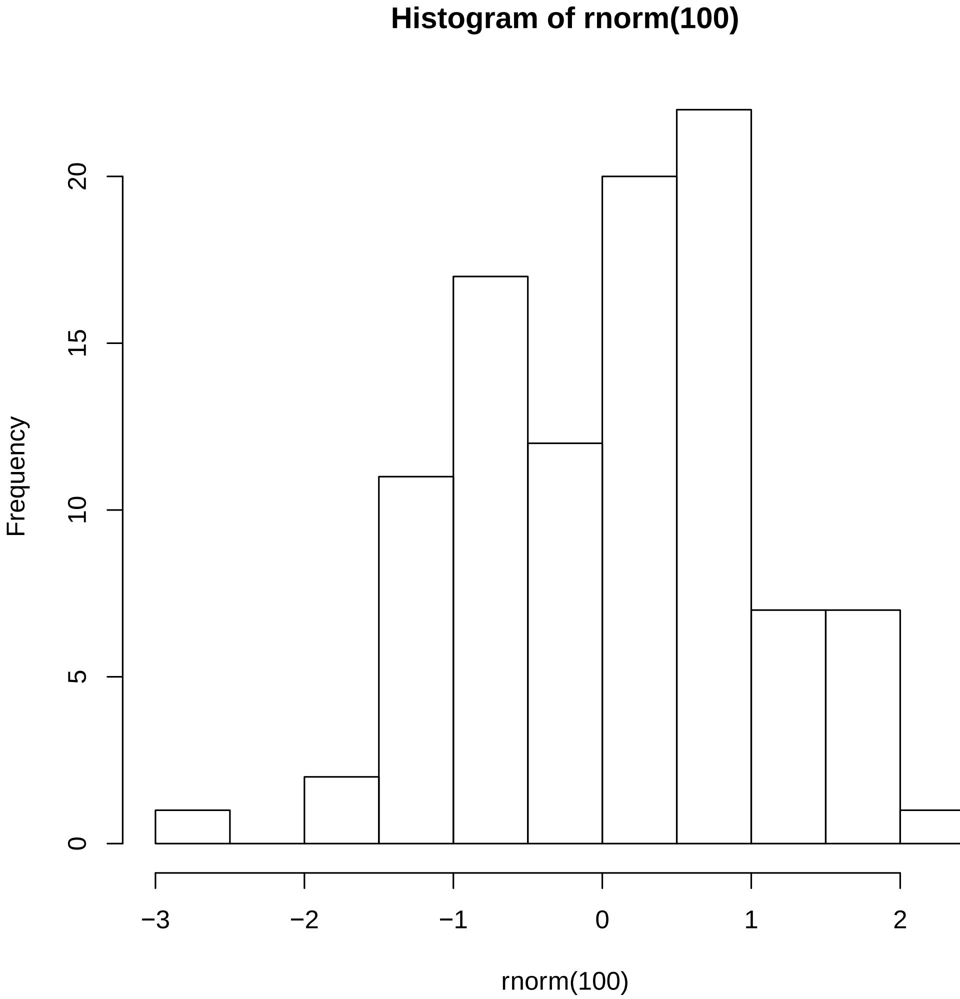
Partition the p variables into two sets: stratifiers and non-stratifers. Denote the non-stratifiers as $Y_1, \cdots, Y_{p-k}$ and the stratifiers as $Z_1, \cdots, Z_k$. Let A_j consist of the set of variables in the j-th known joint population distribution which will be used to post-stratifywhere $j = 1, \cdots, \ell$.

2.1. *Theoretcal Results.*

Expected value of post -stratified mean? How much bias is there? Variance estimate?

3. Simulation Study.

```
hist(rnorm(100))
```



3.1. *Simple k=3, l=2 case.*

3.2. *More complicated case.*

4. Real Data Example.

Hospital Datta

5. Conclusion and Future Work.

Cummings, Eftekhary and House (2003)

References.

Cummings, A. B., Eftekhary, D. and House, F. G. (2003). The accurate determination of college students coefficients of friction. *Journal of Sketchy Physics* **13** 46–129.

E-mail:

E-mail: gmatthews1@luc.edu