The Journal of Open Source Software

# Combining a Probability and a Non-Probability Sample in a Capture-Recapture Setting

**Benjamin Williams**[1]

**1** Department of Statistical Science, Southern Methodist University

**DOI:** 10.21105/joss.00886

**Software**

- Review
- Repository
- Archive

**Submitted:** 06 August 2018
**Published:** 13 August 2018



Statistical sampling plays a vital role in understanding and making inferences with respect to all types of populations and is especially salient in a world where populations are large and big data is the new standard. In an ideal situation, samplers have access to a list of all the units in a population, called a sampling frame, from which to draw a sample. This allows them to select individual population units with known probabilities, producing a probability sample (Lohr 2010). A sample for which probabilities of selection are not known is called a non-probability sample. Probability samples are preferred to non-probability samples because the sampling variance of the estimators calculated from probability samples can be determined using standard sampling theory. The primary disadvantage of non-probability samples is the potential for biased estimation stemming from undercoverage and a lack of representativeness in the samples. Without external sources of information, these sample deficiencies cannot be detected.

It is usually easier to obtain a non-probability sample than a probability sample. For example, a frame may not be available, which complicates the selection of a probability sample. A non-probability sample, such as one using data from volunteers, does not require an expensive nonresponse follow-up. The larger the sample, the more information it may contain. This is an attractive option for analysts working with limited resources while studying elusive populations. As a result, statisticians have begun to investigate methods for improving estimation using data from non-probability samples (for example, Elliott and Haviland (2007)). One way to improve such estimation is to combine a non-probability sample with a probability sample.

The blendR package (available on GitHub) provides four statistically valid estimators of total when combining a non-probability sample with a probability sample. These estimators have applications in many areas, such as: the internet of things, where a non-probability sample could be taken from devices connected to the internet; insurance claims, where claims could be voluntarily reported; and estimation of the death toll due to a natural disaster, where survivors could self-report deaths in a family. In each of these situations, the estimators from blendR can combine the information from the respective non-probability samples with a probability sample to make more accurate estimates. The prevalence of non-probability samples continues to grow in both academia and industry, due in large part to technological advances and the availability of big data. blendR is needed to allow analysts from a variety of disciplines to use non-probability samples to improve estimation.

The estimators are taken from Liu et al. (2017) and Breidt, Opsomer, and Huang (2018). The sampling program considers the non-probability sample as a capture sample and the probability sample as a recapture sample, meaning units selected into the non-probability sample can be again sampled into the probability sample. Capture-recapture methodology provides powerful tools to estimate the total number of units in a population (Le Cren 1965). The goal of the four estimators presented is to make valid estimates of the total of some variable of interest gathered in both samples. The values may disagree for units which are part of both samples (due to measurement error, for example).

The estimators from Liu et al. (2017) are ratio estimators and the one from Breidt, Opsomer, and Huang (2018) is a difference estimator. One ratio estimator uses whether or not the unit was a part of the non-probability sample as auxiliary information, one uses the value of the variable of interest gathered in the non-probability sample as auxiliary information, and the third is a weighted combination of the first two estimators. The difference estimator adds the total value of the variable of interest gathered in the non-probability sample to the estimated difference between the value of the variable in the probability sample and the value of the variable in the non-probability sample. These estimators can be used in any situation of combining samples via a capture-recapture sampling program and have many exciting possible extensions.

The estimators are currently used to estimate the total catch of the fish in several settings, including the fish Red Snapper by Texas Parks and Wildlife (TPWD). TPWD and other entities, including the National Oceanic and Atmospheric Administration (NOAA) estimate the total fish catch in the Gulf of Mexico. The blendR package provides data from a 2016 TPWD capture-recapture sampling program in which the capture sample was a non-probability sample of captains who reported the number of Red Snapper they caught via a smartphone app. The recapture sample was a dockside intercept sample in which boats were boarded and interviewers collected data about the number of Red Snapper caught (a probability sample).

The National Research Council has advised NOAA to continue experiments with electronic reporting to better estimate the total fish caught in marine waters by recreational anglers (National Research Council 2017). Accurate estimation is critical to setting approriate fishing seasons and bag limits. As such, this is an important research field.

This work is part of dissertation research by the author (Benjamin Williams). It is also being used in working papers regarding non-sampling errors and sample size calculations for electronic reporting experiments by a fisheries research team at Southern Methodist University led by Lynne Stokes. Bug reports, contributions, and other useful comments are welcomed as issue tickets on Github and will be attended to in a timely manner.

## References

Breidt, Jay F., Jean D. Opsomer, and Chen-Min Huang. 2018. "Model-Assisted Survey Estimation with Imperfectly Matched Auxiliary Data." In *Predictive Econometrics and Big Data*, 753:21–35. Studies in Computational Intelligence. Springer.

Elliott, Marc N., and Amelia Haviland. 2007. "Use of a Web-Based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33 (2):211–5. http://www.thewitnessbox.com/10498-en.pdf.

Le Cren, E. D. 1965. "A Note on the History of Mark-Recapture Population Estimates." *The Journal of Animal Ecology* 34 (2):453. https://doi.org/10.2307/2661.

Liu, Bingchen, Lynne Stokes, Tara Topping, and Greg Stunz. 2017. "Estimation of a Total from a Population of Unknown Size and Application to Estimating Recreational Red Snapper Catch in Texas." *Journal of Survey Statistics and Methodology* 5 (3):350–71. https://doi.org/10.1093/jssam/smx006.

Lohr, Sharon L. 2010. *Sampling: Design and Analysis*. 2nd ed. Brooks/Cole.

National Research Council. 2017. *Review of the Marine Recreational Information Program*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/24640.