

Stat 472/572 Final Project: As part of the course, your work in this class will culminate in a project involving some theoretical or applied aspect of sampling. You can either work individually or work with other students (at most three students for each group) and submit one project for each team. Final project is **due on Wednesday May 8**. You will need to **submit an electronic version via Canvas**.

Here are some possible topics for the final project:

1. Creating a sampling design for a long-term research project. Collect a pilot sample if possible to help with the design.
2. Analyze data from a large-scale survey such as
The Current Population Survey (CPS),
National Youth Tobacco Survey (NYTS),
The NSF Survey of Doctorate Recipients,
The National Health and Nutrition Examination Survey etc.

You can find a lot of complex survey data from the following websites.

<https://www.nass.usda.gov/> www.cdc.gov

www.fedstats.gov

www.census.gov

<http://www.data.gov>

In your data analysis, be sure to include:

- (a). Description of survey design
- (b). Clearly state your topic. This can be estimating population total, population mean, proportions, discovering the relationship between the variables using survey regression, or other topics engaging your interest.
- (c). Analyses and conclusions. Use what we have covered in class as much as possible and include variance estimation and confidence intervals.
- (d). Nonresponse issues. Discuss how you handle the nonresponses.
- (e). Discuss design features that might affect the results.

3. Study some research problems that you have interest. Some possible topics are

- (a) Comparison of GOF test using bootstrap weights and Rao-Scott first and second order corrected tests.

Rao and Thomas (1988) noted that the need to perform statistical analyses of categorical data is frequently encountered in quantitative sociological research. Methods that take account of survey design features in testing hypotheses have been proposed, including Wald (1943) test, the Rao-Scott first-order and second-order corrected tests (Rao & Scott, 1981), and quasi-score tests (Rao, Scott, & Skinner, 1998) etc. Recently, Kim and Rao (2016) proposed “Hypotheses Testing from categorical survey data using bootstrap weights”. They present a unified approach by constructing bootstrap approximations to weighted likelihood ratio statistics and weighted quasi-score statistics.

Propose to compare the performance of the GOF test by using bootstrap weights to the Rao-Scott first-order and second-order corrected tests.

- (b) Methods for approximating residual variance in nonparametric regression has been discussed by Gasser, Sroka, and Jennen-Steinmetz (1986) etc. Lu (2012) proposed a residual variance estimators for complex survey data by incorporating sampling weights. Try to improve the estimator from Lu.

4. Study some advanced topics, and write a report in your own words with your understanding. Suggested reading material is Dagdou, Goga, and Haziza (2020) “Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison” Dagdou, Goga, and Haziza.

References

- Dagdou, M., Goga, C., & Haziza, D. (2020). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Technical report*, 84, 187–196.
- Gasser, T., Sroka, L., & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73, 625–633.
- Kim, J. kwang, & Rao, J. N. K. (2016). Hypotheses testing from categorical survey data using bootstrap weights. *Proceedings of Statistics Canada Symposium 2016, Growth in Statistical Information: Challenges and Benefits*.

- Lu, Y. (2012). Difference-based variance estimator for nonparametric regression in complex surveys. *Journal of Statistical Computation and Simulation*, 84, 335-343.
- Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J. N. K., Scott, A. J., & Skinner, C. J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.
- Rao, J. N. K., & Thomas, R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.