Stat 9100.3: Analysis of Complex Survey Data

1 Logistics

Instructor: Stas Kolenikov, kolenikovs@missouri.edu

Class period: MWF 1-1:50pm

Office hours: Middlebush 307A, times: TBA

Website: Blackboard http://courses.missouri.edu

Information: This course covers some topics in modern analytical tools developed for complex sample surveys.

Prerequisites: The students will need to have received credit for STAT 4760/7760 or equivalent to be enrolled in this class. In other words, you will have understanding of statistical inference concepts. Having taken STAT 4310/7310 Introduction to Sampling is an advantage.

Other info: Academic integrity is fundamental to the activities and principles of a university. All members of the academic community must be confident that each person's work has been responsibly and honorably acquired, developed, and presented. Any effort to gain an advantage not given to all students is dishonest whether or not the effort is successful. The academic community regards breaches of the academic integrity rules as extremely serious matters. Sanctions for such a breach may include academic sanctions from the instructor, including failing the course for any violation, to disciplinary sanctions ranging from probation to expulsion. When in doubt about plagiarism, paraphrasing, quoting, collaboration, or any other form of cheating, consult the course instructor.

If you have special needs as addressed by the Americans with Disabilities Act (ADA) and need assistance, please notify the Office of Disability Services, A038 Brady Commons, 882-4696 or course instructor immediately. Reasonable efforts will be made to accommodate your special needs.

Grade structure: homeworks (30%) + class presentation (30%) + takehome final (40%)

The homework exercises (about 5–6 throughout the semester) will represent a mix of theoretical questions, and practical examples to be studied with the complex data sets. The class presentation (about 20–25 min) will be one of the additional topics papers, see the list below. Expect the takehome final to be all-inclusive, with theoretical and practical questions, as well as questions based on readings.

Data sets: Students on the biostat track might want to use NHANES data for their homeworks (see links below). Students from social science tracks might want to use GSS or CPS surveys. Students in education might want to use NAEP data. Other data sets might be used from the student's area of interest; those should have sufficiently complex sample design and non-trivial design effects.

Software: Design-based estimation is now incorporated in many software titles. Usability varies from the traditional set of estimators (means, totals, ratios, proportions) to multi-stage designs, and to a variety of analytical tools (linear regression, logistic regression, survival models, and other multivariate techniques). The current leaders appear to be Stata, R and (SAS-callable) SUDAAN. All of them can handle stratified clustered designs with Taylor-series linearization or jackknife and BRR replicate variance estimation, for the linear statistics and a variety of regression estimation procedures, and that is probably as far as most analytic uses of survey would go in the likely applications. A review of the existing software (although it does not seem to have been updated recently) can be found at http://www.hcp.med.harvard.edu/statistics/survey-soft/.

2 Content

The class will consist of several modules, as outlined below.

	Topics	Readings
1.	Basic concepts: SRS, WR, WOR, stratified samples, clustered samples, (Narain-)Horwitz-Thompson estimator. Asymptotic normality	Ch. 1–3 of $\mathcal{TH}$97, Ch. 1 and 2 of $\mathcal{SHS}$89, Dalenius (1994), Ch. 1 of $\mathcal{LP}$04, Ch. 2 of $\mathcal{KG}$99, Ch. 1–2 of $\mathcal{CS}$05; Brewer & Donadio (2003)
2.	Design-based, model-based, model-assisted, predictive approaches to survey inference	Binder & Roberts (2003), Kish & Frankel (1974), Brewer (2002), Särndal, Swensson & Wretman (1992), Ch. 3 of $\mathcal{CS}$05
3.	Survey weights	Ch. 4 of $\mathcal{KG}$99, Sec. 6.2 of $\mathcal{TH}$97, Pfeffermann (1993), Korn & Graubard (1995)
4.	Analysis of subdomains and subpopulations	Skinner (1989) = Ch. 3 of $\mathcal{SHS}$89, Ch. 6 of $\mathcal{LP}$04, Bellhouse & Rao (2002), Hidiroglou & Patak (30)
5.	Nonlinear statistics, regression and estimating equations	Binder (1983), Skinner (1989), Ch. 4 and Sec. 6.4–6.5 of $\mathcal{TH}$97, Fuller (1975), Fuller (2002), Sec. 11.2 of $\mathcal{CS}$05
6.	Missing data	Kalton & Kasprzyk (1986), Little (2003*b*), Ch. 4 of $\mathcal{LP}$04, Ch. 4 of $\mathcal{KG}$99, Ch. 13 of $\mathcal{CS}$05; Little & Vartivarian (2005), Haziza & Rao (2006)
7.	Small area estimation	Rao (2003), Ghosh & Rao (1994), Sec. 10.4 of $\mathcal{CS}$05; Fay & Herriot (1979), Prasad & Rao (1990), Ghosh, Natarajan, Stroud & Carlin (1998), Lehtonen, Särndal & Veijanen (2003), You & Chapman (2006), special issue of Statistics in Transition
8.	Variance estimation and resampling inference	Sec. 4.2 of $\mathcal{TH}$97, Ch. 5 of $\mathcal{KG}$99, Ch. 5 of $\mathcal{LP}$04, Shao (1996), Krewski & Rao (1981), Rao & Wu (1988), Rao, Wu & Yue (1992), Ch. 7 and 9 of $\mathcal{CS}$05
Additional topics		
i.	Empirical likelihood inference	Chen & Qin (1993), Wu (2004), Wu & Rao (2006)
ii.	Multilevel models	Pfefferman, Skinner, Holmes, Goldstein & Rasbash (1998), Rabe-Hesketh & Skrondal (2006)
iii.	Sampling in space and time	Binder & Hidiroglou (1988), Fuller (1990), Ernst (1999), Ch. 7 of $\mathcal{TH}$97
iv.	Bayesian methods	Little (2003*a*) = Ch. 4 of $\mathcal{CS}$03, Ghosh et al. (1998), You & Chapman (2006)
v.	Case-control studies	Scott & Wild (2003) = Ch. 8 of $\mathcal{CS}$03, Ch. 9 of $\mathcal{KG}$99
vi.	Disclosure risk	Skinner & Carter (2003)
vii.	Inverse sampling	Rao, Scott & Benhin (2003)
viii.	Non-sampling error	Lesser & Kalsbeek (1992)
ix.	Post-stratification	Holt & Smith (1979), Valliant (1993)
x.	Survey methodology and cognitive issues	Groves, Couper, Lepkowski, Singer & Tourangeau (2004), Statistics Canada (2003)

3 Readings

The list of topics and readings should not be intimidating. This is the list of "everything-you-need-to-know-about-survey-statistics" (unless you do methodological research in the area). The readings are provided for your reference, so that you could consult your syllabus should the need arise in your practical work to get started with the literature search. The course is divided into the main part that will be delivered by the instructor, with the readings that generally are book chapters, invited papers, or other big reviews of the topic; and the optional part, with the topics to be picked by students for their term presentation, and the readings being the research papers.

There are several great books on the topic of complex survey sampling and data analysis. Some of them, mostly earlier ones, tend to gravitate to the issues of sampling *per se* and mathematical foundations: Kish (1965), Cochran (1977), Wolter (1985), Thompson (1992), Levy & Lemeshow (2003), Chaudhuri & Stenger (2005) (referred to as $\mathcal{CS}05$ above). Other more recent books tend to focus more on the analytical methods developed to address a wide range of practical problems: Skinner, Holt & Smith (1989) [$\mathcal{SHS}89$], Thompson (1997) [$\mathcal{TH}97$], Korn & Graubard (1999) [$\mathcal{KG}99$], Chambers & Skinner (2003) [$\mathcal{CS}03$], Lehtonen & Pahkinen (2004) [$\mathcal{LP}04$]. If you have any of those books in your library, it will cover most of the "first order" topics in the first half of the course, and some of the "second order" selective topics. A summary of historical developments in survey statistics is given in Rao (2005). There are also some highly specialized monographs, such as Särndal et al. (1992), Rao (2003) or Tillé (2006).

There is a broad range of articles published in top journals such as *Annals of Statistics, JASA, JRSSb, Biometrika*, but the leading journal in the field dedicated solely to survey statistics is *Survey Methodology* published by Statistics Canada.

4 Educational objectives

Upon completion of the course, the students will:

- understand the importance of design-based (randomization) inference;
- know the implications of complex sampling designs for point and interval estimation;
- by using the randomization inference paradigm, be able to compute means and variances of simple statistics;
- know and be able to verify the domains of applicability of asymptotic normality, including results for non-linear statistics;
- be aware of the subtleties that arise in variance estimation, and be able to find ways to estimate variances in difficult situations, including those with (adjustments for) non-response;
- specify the major features of complex survey designs in their favorite software;
- perform analysis of (generalized) linear models, including analysis on subdomains, with appropriate design specification;
- be aware of the broad spectrum of research problems in area of survey statistics.

5 Links

	Data sets
NHANES:	http://www.cdc.gov/nchs/nhanes.htm
GSS:	http://www.norc.org/projects/gensoc.asp
CPS:	http://www.census.gov/cps/
NAEP:	http://nces.ed.gov/nationsreportcard/
	Software
Stata:	http://www.stata.com/stata9/svy.html
R:	http://cran.us.r-project.org/src/contrib/Descriptions/survey.html
SUDAAN:	http://www.rti.org/sudaan/

Publications

Survey Methodology journal, open access:	http://www.statcan.ca/bsolc/english/bsolc?catno=12-001-X&CHROPG=1
My personal set of references:	http://www.citeulike.org/user/ctacmo/tag/survey
ASA Survey Research Methods Section:	http://www.amstat.org/sections/SRMS/index.html
Statistics in Transition special issue on SAE:	http://www.stat.gov.pl/english/sit/sit73/index.htm
Statistics Canada MA readings:	http://www.statcan.ca/english/employment/ma/readings.htm
NIH references:	http://archive.nlm.nih.gov/proj/dxpnet/nhanes/docs/doc/sample_survey/references.php

References

Bellhouse, D. R. & Rao, J. N. K. (2002), 'Analysis of domain means in complex surveys', *Journal of Statistical Planning and Inference* **102**, 47–58.

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.

Binder, D. A. & Hidiroglou, M. A. (1988), Sampling in time, *in* P. R. Krishnaiah & C. R. Rao, eds, 'Handbook of Statistics', Vol. 6, North Holland, Amsterdam, pp. 187–211.

Binder, D. A. & Roberts, G. R. (2003), Design-based and model-based methods for estimating model parameters, *in* R. L. Chambers & C. J. Skinner, eds, 'Analysis of Survey Data', John Wiley & Sons, New York, chapter 3.

Brewer, K. (2002), *Combined Survey Sampling Inference*, Arnold/Oxford University Press.

Brewer, K. & Donadio, M. E. (2003), 'The high entropy variance of the Horvitz-Thompson estimator', *Survey Methodology* **29**(2), 189–196.

Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, Wiley series in survey methodology, Wiley, New York.

Chaudhuri, A. & Stenger, H. (2005), *Survey Sampling: Theory and Methods*, Vol. 181 of *Statistics: Textbooks and Monographs*, 2nd edn, Chapman & Hall/CRC, Boca Raton, FL.

Chen, J. & Qin, J. (1993), 'Empirical likelihood estimation for finite populations and the effective usage of auxiliary information', *Biometrika* **80**(1), 107–116.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd edn, John Wiley and Sons, New York.

Dalenius, T. (1994), A first course in survey sampling, *in* P. R. Krishnaiah & C. R. Rao, eds, 'Handbook of Statistics: Sampling', Vol. 6, Elsevier: North Holland, chapter 2.

Ernst, L. R. (1999), The maximization and minimization of sample overlap problems: A half century of results, Technical report, U.S. Bureau of Labor Statistics.

Fay, R. E. & Herriot, R. A. (1979), 'Estimates of income for small places: An application of james-stein procedures to census data', *Journal of the American Statistical Association* **74**(366), 269–277.

Fuller, W. A. (1975), 'Regression analysis for sample survey', *Sankhya Series C* **37**, 117–132.

Fuller, W. A. (1990), 'Analysis of repeated surveys', *Survey Methodology* **16**(2), 167–180.

Fuller, W. A. (2002), 'Regression estimation for survey samples (with discussion)', *Survey Methodology* **28**(1), 5–23.

Ghosh, M., Natarajan, K., Stroud, T. W. F. & Carlin, B. P. (1998), ‘Generalized linear models for small-area estimation’, *Journal of the American Statistical Association* **93**(441), 273–282.

Ghosh, M. & Rao, J. N. K. (1994), ‘Small area estimation: An appraisal’, *Statistical Science* **9**(1), 55–76.

Groves, R. M., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004), *Survey Methodology*, Wiley Series in Survey Methodology, John Wiley and Sons, New York.

Haziza, D. & Rao, J. N. (2006), ‘A nonresponse model approach to inference under imputation for missing survey data’, *Survey Methodology* **32**(1), 53–64.

Hidiroglou, M. A. & Patak, Z. (30), ‘Domain estimation using linear regression’, *Survey Methodology* **1**, 67–78.

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

Kalton, G. & Kasprzyk, D. (1986), ‘The treatment of missing survey data’, *Survey Methodology* **12**(1), 1–16.

Kish, L. (1965), *Survey Sampling*, John Wiley and Sons, New York.

Kish, L. & Frankel, M. R. (1974), ‘Inference from complex samples’, *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

Korn, E. L. & Graubard, B. I. (1995), ‘Analysis of large health surveys: Accounting for the sampling design’, *Journal of the Royal Statistical Society, Series A* **158**(2), 263–295.

Korn, E. L. & Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley and Sons.

Krewski, D. & Rao, J. N. K. (1981), ‘Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods’, *The Annals of Statistics* **9**(5), 1010–1019.

Lehtonen, R. & Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys*, Statistics in Practice, 2nd edn, John Wiley & Sons, New York.

Lehtonen, R., Särndal, C.-E. & Veijanen, A. (2003), ‘The effect of model choice in estimation for domains, including small domains’, *Survey Methodology* **29**(1), 33–44.

Lesser, V. M. & Kalsbeek, W. D. (1992), *Non-sampling Error in Surveys*, John Wiley and Sons, New York.

Levy, P. S. & Lemeshow, S. (2003), *Sampling of Populations: Methods and Applications*, 3rd edn, John Wiley & Sons, New York.

Little, R. J. (2003*a*), The Bayesian approach to sample survey inference, *in* R. Chambers & C. J. Skinner, eds, ‘Analysis of Survey Data’, John Wiley and Sons, chapter 4.

Little, R. J. (2003*b*), Bayesian methods for unit and item nonresponse, *in* R. Chambers & C. J. Skinner, eds, ‘Analysis of Survey Data’, John Wiley and Sons, chapter 18.

Little, R. J. & Vartivarian, S. (2005), ‘Does weighting for nonresponse increase the variance of survey means?’, *Survey Methodology* **31**(2), 161–168.

Pfefferman, D., Skinner, C. J., Holmes, D. J., Goldstein, H. & Rasbash, J. (1998), ‘Weighting for unequal selection probabilities in multilevel models’, *Journal of Royal Statistical Society* **60**(1), 23–40.

Pfeffermann, D. (1993), ‘The role of sampling weights when modeling survey data’, *International Statistical Review* **61**, 317–337.

Prasad, N. G. N. & Rao, J. N. K. (1990), ‘The estimation of the mean squared error of small-area estimators’, *Journal of the American Statistical Association* **85**(409), 163–171.

Rabe-Hesketh, S. & Skrondal, A. (2006), 'Multilevel modelling of complex survey data', *Journal of the Royal Statistical Society, Series A* **169**(4).

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley series in survey methodology, John Wiley and Sons, New York.

Rao, J. N. K. (2005), 'Interplay between sample survey theory and practice: An appraisal', *Survey Methodology* **31**(2), 117–138.

Rao, J. N. K. & Wu, C. F. J. (1988), 'Resampling inference with complex survey data', *Journal of the American Statistical Association* **83**(401), 231–241.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**(2), 209–217.

Rao, J., Scott, A. & Benhin, E. (2003), 'Undoing complex survey data structures: Some theory and applications of inverse sampling (with discussion)', *Survey Methodology* **29**(2), 107–128.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.

Scott, A. & Wild, C. (2003), Fitting logistic regression models in case-control studies with complex sampling, *in* R. Chambers & C. J. Skinner, eds, 'Analysis of Survey Data', John Wiley and Sons, chapter 8.

Shao, J. (1996), 'Resampling methods in sample surveys', *Statistics* **27**, 203–254. with discussion.

Skinner, C. & Carter, R. (2003), 'Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling', *Survey Methodology* **29**(2), 177–180.

Skinner, C. J. (1989), Domain means, regression and multivariate analysis, *in* C. J. Skinner, D. Holt & T. M. Smith, eds, 'Analysis of Complex Surveys', Wiley, New York, chapter 3, pp. 59–88.

Skinner, C. J., Holt, D. & Smith, T. M., eds (1989), *Analysis of Complex Surveys*, Wiley, New York.

Statistics Canada (2003), *Survey Methods and Practices*, Ottawa. Catalogue No. 12-587-XPE.

Thompson, M. E. (1997), *Theory of Sample Surveys*, Vol. 74 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, New York.

Thompson, S. K. (1992), *Sampling*, John Wiley and Sons, New York.

Tillé, Y. (2006), *Sampling Algorithms*, Springer Series in Statistics, Springer, New York.

Valliant, R. (1993), 'Poststratification and conditional variance estimation', *Journal of the American Statistical Association* **88**(421), 89–96.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer, New York.

Wu, C. (2004), 'Some algorithmic aspects of the empirical likelihood method in survey sampling', *Statistica Sinica* **14**, 1057–1067.

Wu, C. & Rao, J. N. K. (2006), 'Pseudo-empirical likelihood ratio confidence intervals for complex surveys', *The Canadian Journal of Statistics/La revue canadienne de statistique* **34**(3), 359–376.

You, Y. & Chapman, B. (2006), 'Small area estimation using area level models and estimated sampling variances', *Survey Methodology* **32**(1), 97–103.