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, Mon 1-2pm, Tue 1-2 pm, Thu 9-10am.

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 (2004)
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), Kim, Michael, Fuller & Kalton (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), 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. (2004), 'Domain estimation using linear regression', *Survey Methodology* **30**, 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.

Kim, J. K., Michael, B. J., Fuller, W. A. & Kalton, G. (2006), 'On the bias of the multiple-imputation variance estimator in survey sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3), 509–521.

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, Series B* **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.

The references are also available at http://www.citeulike.org/user/ctacmo/tag/stat9100svy

Note Title 1/11/2007

STAT 9100.3 COMPLEX SURVEYS – INTRO

- INTRODUCE MYSELF
- HAND OUT SYLLABI
- MECHANICS:
  - CLASS PERIODS
  - STUDENT TALKS
  - HOME ASSIGNMENTS
  - TAKEHOME FINAL
- UNIQUENESS OF THE COURSE!
  - + SHORTAGE OF SURVEY STATISTICIANS
  - + WIDE USE OF SURVEY STATS
    - FEDERAL: BLS, CENSUS; STAT CANADA
    - HEALTH: CDC, NCI, ... – HEALTH DATA
    - PRIVATE: RTI, NORC, WESTAT, ...
      - GOD ONLY KNOWS WHAT ELSE!
  - + MY OWN RESEARCH INTERESTS

- TOPICS IN THE SYLLABUS:
  - BASIC TOPICS IN CLASS BY ME
  - ADVANCED TOPICS: SELECT PAPERS
    - SHOW OF HANDS – SUGGEST TOPICS

- BROAD OBJECTIVES:
  - + OVERALL FAMILIARITY WITH WHAT SVY STAT IS ABOUT
  - + UNDERSTANDING OF THE DESIGN-BASED ESTIMN

- TO THAT EFFECT, A HUGE LIST OF REFS, CLICKABLE LINKS IN SYLLABUS

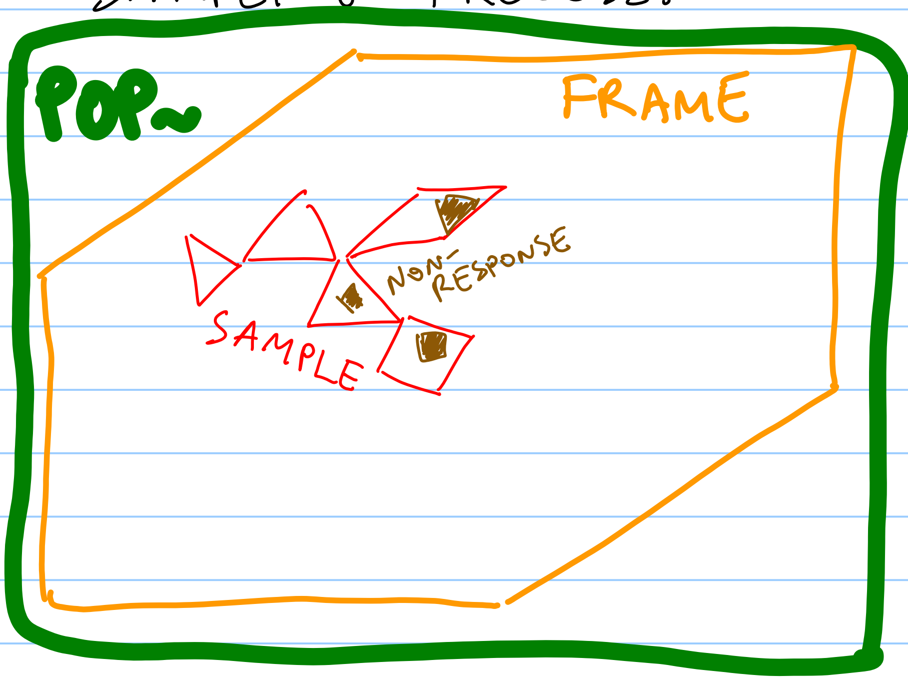
BOOKS ON SAMPLING:

- THOMPSON, M (1997) – THE MAIN TEXT
  INTERMEDIATE TO UPPER LEVEL,
  QUITE RIGHT FOR US
- LEHTONEN & PAHKINEN (2004)
  KORN & GRAUBARD (1999)
  — SOMEWHAT LIGHTER BUT STILL RIGOROUS
- KISH, COCHRAN, THOMPSON (92), LEVY & LEMESHOW
  — GOOD OL' BOOKS, DON'T HAVE
  NEWER STUFF, RATHER DRY
- SHS, CHAMBERS & SKINNER —
  COLLECTIONS OF RESEARCH MONOGRAPHS,
  THE MOST ADVANCED & UP TO DATE
- SÄRNDAL, SWENSSON, WRETMAN —
  MORE DIFFICULT TO DIGEST
- SPECIALIZED MONOGRAPHS: RAO; TILLÉ; ...

FIG. FROM Ch 1 OF LP(04) —
SURVEY PROCESS.
WHERE DO...
- STATISTICIANS
- COGNITIVE /PSYCHOMETRIC PEOPLE
- MANAGERS
- FIELD STAFF
- CLIENTS

... FIT IN?

SAMPLING PROCESS:



NOTE THAT POP~S TO WHICH THE RESEARCHER WANTS TO GENERALIZE MAY NOT BE THOSE FOR WHICH (LARGE SCALE) DATA WERE COLLECTED: TARGET VS. REPRESENTED POP~

OBSERVED SUMMARY Y =

= TRUE Y +

+ (Y [POP~] − Y [FRAME]) + UNDERCOVERAGE

+ RANDOM~ ERROR SAMPLING ERROR

+ INSTRUMENT~ ERROR MEASUREMENT

+ NON-RESPONSE ERROR MEASUREMENT

+ ADJUSTMENTS

DESIGN PARADIGM:

THE RANDOM COMPONENT IS DUE TO RANDOM~, BUT IT IS NOT THE CHARACTERISTIC ITSELF.

DESCRIPTIVE USE: TOTALS, MEANS, RATIOS, PROPORTIONS

ANALYTICAL USE: STAT'L REL~ MODELS (LINEAR, LOGISTIC, SVY REGRESSION; MULTILEVEL & STRUCTURAL EQ~ MODELS)

EXAMPLE OF A REAL COMPLEX SVY:
NHANES III (K & G 99)

RUN INTRO-NHANES.DO

http://www.cdc.gov/nchs/nhanes.htm

NHANES III:

2812 PSUs
↳ 13 wp 1
↳ 2799
I ↳ 34 STRATA, ALLOWS OVERSAMPLING OF NON-WHITES
↳ 2 PSU/STRATUM

II 89 STANDS FOR MOBILE EXAM CENTERS
↳ SEGMENTS

III ↳ LIST HHs IN EACH SEGMENT

IV ↳ INDIVIDUALS ⊆ HH
SCREENING, DEMOGRAPHY
⇒ ≃1 IN 5 SAMPLED HH
CONTRIBUTED OBSERV. UNITS

INSTRUMENTS:

- INDIVIDUAL Q'RE
- BLOOD PRESSURE MEAST: HH
- FAMILY Q'RE
- MEC → AUTOMATED DATA COLLECTION

Note Title — STAT 9100 SVY — SRS & SIMPLE DESIGNS — 1/12/2007

FRAMEWORK & DEFINITIONS

CAPITALS / LOWERCASE
POP^n / SAMPLE

$$\mathcal{U} = \begin{array}{|c|c|c|} 1 & Y_1 & I_{1s} \\ \vdots & Y_2 & \vdots \\ \vdots & \vdots & \\ N & Y_N & I_{Ns} \end{array}$$

UNIVERSE

$\uparrow$ FIXED (column Y) $\quad$ $\uparrow$ RANDOM! (column I)

$\mathcal{S} = 2^{\mathcal{U}}$ — ALL SUBSETS

$T[Y] = \sum_{j=1}^{N} Y_j$ — POP^n TOTAL

$s = (j_1, \ldots, j_n)$ ~ SAMPLE (UNORDERED – SUFF STAT)

PROBABILITY SPACE / SAMPLE DESIGN: $p : \mathcal{S} \to [0,1]$

$\forall s \in \mathcal{S}$ $\quad$ $p(s)$ IS PROBTY OF SELECTING THIS PARTICULAR SAMPLE

EXAMPLES: $\mathcal{U} = \{1, 2, 3\}$ $\quad$ SRS WR, SRSWOR

SAMPLE SIZE: $n(s) = \sum_{j=1}^{N} I_{js}$ = # UNITS IN S

FIXED SIZE DESIGNS: $p(s) = 0 \Leftarrow n(s) \neq n$

INCLUSION PROBTY: $\pi_j = E_{p(s)} I_{js}$

NOTE THAT $\sum_j \pi_j = E\, n(s) > 1$

SELF-WGT DESIGN / EPSEM: $\forall j \quad \pi_j = \text{Const} = n/N$

NOT EVERY EPSEM IS NICE!

JOINT INCLUSION PROBTY: $\pi_{jk} = \mathbb{E}_{p(s)} I_{js} I_{ks}$

IMPORTANT FOR VARIANCE ESTIM~!

SRSWOR:

$\forall s \quad n(s) = n, \quad p(s) = \text{const} = 1/C_N^n$

$\forall j$, # SAMPLES WHERE IT IS USED $= C_{N-1}^{n-1}$

HENCE, $\pi_j = \dfrac{C_{N-1}^{n-1}}{C_N^n} = \dfrac{(N-1)!}{(n-1)!(N-n)!} \dfrac{n!(N-n)!}{N!} = \dfrac{n}{N}$ — EPSEM

JOINT INCLUSION PROBTY:

$\forall$ PAIR j,k # SAMPLES $= C_{N-2}^{n-2}$

HENCE, $\pi_{jk} = \dfrac{C_{N-2}^{n-2}}{C_N^n} = \dfrac{(N-2)!}{(n-2)!(N-n)!} \dfrac{n!(N-n)!}{N!} = \dfrac{n(n-1)}{N(N-1)}$

π_{jk} vs n^2/N^2 ?

$\dfrac{n(n-1)}{N(N-1)} \overset{<}{\text{vs}} \dfrac{n^2}{N^2} \Leftrightarrow \dfrac{n-1}{N-1} \overset{<}{\text{vs}} \dfrac{n}{N} \Leftrightarrow nN - N \overset{<}{\text{vs}} nN - n$

$n < N$

FINITE POPULATION CONSISTENCY:

THE ESTIMATOR = THE POP~ PARR

WHEN $s = \mathcal{U}$

EXPECTATIONS

$$\mathbb{E}_{p(s)} \sum_{j\in s} z_j = \sum_{s\in S} p(s) \sum_{j\in s} z_j =$$

$$= \sum_{s\in S} p(s) \sum_{j=1}^{N} z_j I_{js} = \sum_{j=1}^{N} \left\{ z_j \sum_{s\in S} p(s) I_{js} \right\} = \sum_{j=1}^{N} z_j \pi_j$$

$z_j = 1 \Rightarrow$ SAMPLE SIZE $\quad \mathbb{E}\, n(s) = \sum_{j=1}^{N} \pi_j$

z_j IS SOMETHING REAL & EPSEM $\Bigg\} \Rightarrow \mathbb{E} \sum_{j\in s} y_j = \sum_{j=1}^{N} y_j \frac{n}{N} = \frac{n}{N} T[y]$

$$\bar{y}_n = \frac{1}{n} \sum_{j\in s} y_j \Rightarrow \mathbb{E}\, \bar{y} = T[y]/N \quad \text{— UNBIASED!}$$

EXPANSION ESTIMATOR OF TOTAL:

$$t[y] = N \bar{y}_n \quad \text{FOR EPSEM DESIGNS}$$

FOR ARBITRARY DESIGNS, HORVITZ-THOMPSON (-NARAIN) ESTIMATOR

$$t_{HT}[y] = \sum_{j\in s} y_j / \pi_j$$

$\Rightarrow$ SHOW IT IS UNBIASED — IN CLASS EXERCISE?
— SHORT HWK PROBLEM?

	$p(s)$	π_j	π_{jk}	t_{HT}	$\mathbb{V}\, t_{HT}$
$\{1\}$	0				
$\{2\}$	0				
$\{3\}$	0				
$\{1,2\}$	3/12				
$\{1,3\}$	4/12				
$\{2,3\}$	5/12				
$\{1,2,3\}$	0				

COMPARE TO SRS

DESIGN VARIANCE:

$$\mathbb{V}\Big(\sum_{j\in s} z_j\Big) = \mathbb{V}_{p(s)}\Big(\sum_{j=1}^{N} z_j I_{js}\Big)$$

$$= \sum_{j=1}^{N} z_j^2\, \mathbb{V} I_{js} + \sum_{j\neq k} z_j z_k \operatorname{Cov}(I_{js}, I_{ks}) =$$

$$= \sum_{j=1}^{N} z_j^2 \pi_j(1-\pi_j) + \sum_{j\neq k} z_j z_k (\pi_{jk} - \pi_j\pi_k)$$

$$\operatorname{Cov}(I_{js}, I_{ks}) = \mathbb{E}\, I_{js} I_{ks} - \mathbb{E} I_{js} \mathbb{E} I_{ks} = \pi_{jk} - \pi_j \pi_k$$

FIXED SIZE: $\mathbb{P}\big(\sum I_{js} = n\big) = 1$

$$\sum_{k\neq j}^{N} \operatorname{Cov}(I_{js}, I_{ks}) = \operatorname{Cov}(I_{js}, n - I_{js}) = -\mathbb{V}(I_{js})$$

$$\Rightarrow \quad \mathbb{V}\Big(\sum_{j\in s} z_j\Big) = \frac{1}{2}\sum_{j\neq k} (z_j - z_k)^2 (\pi_j\pi_k - \pi_{jk})$$

HT: $z_j = y_j/\pi_j \Rightarrow$

$$\mathbb{V}\,[t_{HT}[y]] = \sum_{j=1}^{N} y_j^2\Big(\frac{1}{\pi_j} - 1\Big) + \sum_{j\neq k} y_j y_k \Big(\frac{\pi_{jk}}{\pi_j\pi_k} - 1\Big)$$

$$=\{\text{FIXED } n\} = \frac{1}{2}\sum_{j\neq k}\Big(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k}\Big)^2 \Omega_{jk}, \qquad \Omega_{jk} = \pi_j\pi_k - \pi_{jk}$$

? → DERIVE THE VARIANCE FORMULA FOR SRS
$\Rightarrow$ DISCUSS THE FINITE POP~ CORR~
(MENTION U-STAT PRES~ OF $\mathbb{V}$?)

$$\Omega_{jk} = n^2/N^2 - n(n-1)/N(N-1) = \frac{n^2(N-1) - Nn(n-1)}{N^2(N-1)} = \frac{n(N-n)}{N^2(N-1)}$$

DESIGN EFFECT:

$$\operatorname{DEFF}(t) = \frac{\mathbb{V}_{p(s)}[t]}{\underset{\uparrow \text{WOR}}{\mathbb{V}_{SRS}}[t]}$$

ESTIMATORS OF VARIANCE

IF z_{jk} IS A CHARACTERISTIC OF A PAIR OF OBSERVATIONS, THEN AN ESTIMATE OF $T[z_{jk}]$

$$t[z] = \sum_{jk\in s}{}' z_{jk}/\pi_{jk}$$

$\Rightarrow$

$$v(t_{HT}) = \frac{1}{2} \sum\sum_{\substack{j,k\in s \\ j\neq k}} \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k}\right)^2 \frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}$$

YATES-GRUNDY-SEN ESTIMATOR

HT-STYLE ESTIMATOR FOR A U-STAT OF ORDER 2

UNBIASED IF $\forall jk \ \pi_{jk} > 0$

WEIRD IF $\pi_{jk} = 0$!

OTHER TRADITIONAL DESIGNS

SRS WITH REPLACEMENT

↪ n UNITS WITH REPLACEMENT

$$1-\pi_j = \frac{(N-1)^n}{N^n} \sim \frac{N^n - nN^{n-1} + n(n-1)/2\, N^{n-2} - \dots}{N^n}$$

$$\pi_j \approx n/N + o(n/N) \quad \text{as} \quad n/N \to 0$$

$$1-\pi_{ij} = (N-2)^n/N^n = 1 - \frac{2n}{N} + \frac{2n(n-1)}{N^2} - \dots$$

NICE DESIGN: SRSWR + SRSWR = SRSWR!

NOT TRUE FOR WOR

$$\mathbb{E}\, n(s) = \sum_j \pi_j = n - \frac{n(n-1)}{2N} + \dots$$

$$\mathbb{V}[\, t_{HT}[y]\,] = \sum_{j=1}^{N} y_j^2 \left(\frac{1-\pi_j}{\pi_j}\right) + \sum_{j \neq k} y_j y_k \left(\frac{\pi_{jk} - \pi_j \pi_k}{\pi_j \pi_k}\right) = \dots$$

CLUSTERED SAMPLING

$\mathcal{U} = \mathcal{B}_1 \cup \dots \cup \mathcal{B}_L$, ℓ CLUSTERS ARE SELECTED AT RANDOM, EVERY UNIT OBSERVED

$\pi_j = \pi_{jk} = \ell/L$, $j,k \in \mathcal{B}_t$ $\qquad \Omega_{jk} = \ell/L(\ell/L - 1) < 0$

$\pi_{jk} = \dfrac{\ell(\ell-1)}{L(L-1)}$, $j \in \mathcal{B}_t,\ k \in \mathcal{B}_{t'},\ t \neq t'$ $\qquad \Omega_{jk} = \dfrac{\ell(L-\ell)}{L^2(L-1)} > 0$

WHAT IS AN EFFICIENT STRATEGY?

STRATIFIED SAMPLING

$\mathcal{U} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_H$, $|\mathcal{S}_h| = N_h$ KNOWN

WITHIN EACH $\mathcal{S}$, AN SRS IS TAKEN:

$\pi_j = n_h / N_h$ FOR $j \in \mathcal{S}_h$

$$t_{HT}[y] = \sum_h^H N_h \bar{y}_h \; ; \quad \bar{y}_{STR} = \sum_{h=1}^H W_h \bar{y}_h$$

$$\bar{y}_h = \frac{1}{n_h} \sum_{\substack{j \in \mathcal{S}_h \\ j \in S}} y_j \qquad W_h = N_h / \sum_k N_k$$

$V[t_{HT}] = ?$

$\Omega_{jk} = 0$, $j \in \mathcal{S}_h$, $k \in \mathcal{S}_{h'}$, $h \neq h'$

$\Omega_{jk} =$ THAT FOR SRS, $j, k \in \mathcal{S}_h$ (2.46), (2.47)

$\Rightarrow$ WHAT IS THE OPTIMAL STRATIF~ STRATEGY?

HW: Neyman ALLOC~

$\forall j \in \mathcal{S}_h$ COST $= C_h$

TOTAL BUDGET $= C$

OPTIMAL ALLOCATION $= ?$

MY PAPER: BUDGET-OPTIMAL ALLOC~ FOR REPEATED CLUSTERED SVY

OTHER ESTIMATORS OF THE TOTAL

LINEAR ESTIMATORS $\quad e = \sum_{j \in s} d_{js} y_j \,, \quad \sum_{s: j \in s} p(s) d_{js} = 1$

$$= \sum_{j=1}^{N} d_{js} I_{js} y_j \,; \quad \sum_s p(s) I_{js} d_{js} = 1$$

$$V_{p(s)}[e] = \sum_{j=1}^{N} a_j y_j^2 + \sum_{j \neq k} a_{jk} y_j y_k$$

$$a_j = V[d_{js} I_{js}] = \sum_s p(s) d_{js}^2 I_{js} - 1$$

$$a_{jk} = \mathrm{Cov}(d_{js} I_{js}, d_{ks} I_{ks})$$

$$v[e] = \sum_{j \in s} \frac{a_j}{\pi_j} y_j^2 + \sum_{j \neq k} \frac{a_{jk}}{\pi_{jk}} y_j y_k$$

CALIBR~ FOR ARRAY w:

$$\forall \text{ a.e. } s \qquad \sum d_{js} w_j = T[w]$$

$$\Rightarrow \quad MSE(e) = -\frac{1}{2} \sum_{j \neq k} a_{jk} \left(\frac{y_j}{w_j} - \frac{y_k}{w_k}\right)^2 w_j w_k$$

$$a_{jk} = \mathbb{E}\left[(d_{js} I_{js} - 1)(d_{ks} I_{ks} - 1)\right]$$

$$\mathbb{E}_{p(s)}\, e = T[y] \Rightarrow$$

$$v(e) = -\frac{1}{2} \sum_{j \neq k} \left(d_{js} d_{ks} - \frac{1}{\pi_{jk}}\right)\left(\frac{y_j}{w_j} - \frac{y_k}{w_k}\right)^2 w_j w_k$$

PPS SAMPLING:

$$\pi_j \sim y_j \Rightarrow MSE[e] = 0$$

STAT 9100 SVY - MULTISTAGE

Note Title

1/25/2007

1^{ST} STAGE:

RANDOM SELECTION OF PRIMARY SAMPLING UNITS (PSUs)

(TYPICALLY, πps)

2^{ND} STAGE:

RANDOM SELECTION OF SSUs

⋮

OBSERVATION UNIT: THE ONE THAT IS CONTACTED, AND RESPONDS TO THE SVY

BIRTHDAY METHOD FOR WITHIN HH SAMPLING

$$\mathcal{U} = \mathcal{B}_1 \cup \dots \cup \mathcal{B}_L$$

$$|\mathcal{B}_r| = M_r, \quad \sum_{r=1}^{L} M_r = N$$

1^{ST} STAGE:

SAMPLE S_B OF PSU LABELS

2^{ND} STAGE: $\perp\!\!\!\perp$ $\forall r \in S_B$,

SAMPLE S_r IS TAKEN FROM $\mathcal{B}_r$,

$$|S_r| = m(S_r)$$

$$s = \bigcup_{r \in S_B} S_r, \quad n(s) = \sum_{r \in S_B} m(S_r)$$

1^{ST} STAGE INCLUSION PROB'TY:
$$\mathbb{P}[r \in S_B] = \Pi_r$$

2^{ND} STAGE INCLUSION PROB'TY:
$$\mathbb{P}[j \in S_r \mid r \in S_B] = \pi_{j|r}$$

UNIT INCL'N PROB'TY:
$$\pi_j = \Pi_r \, \pi_{j|r}$$

HT ESTIMATOR
$$t[y] = \sum_r \frac{t_r[y]}{\Pi_r}$$
$$\underbrace{t_r[y]}_{\text{HT ESTIMATOR OF THE r-th PSU TOTAL}} = \sum_{j \in S_r} \frac{y_j}{\pi_{j|r}}$$

EXAMPLE:

1) $\Pi_r = M_r \ell / N$ ← πps
2) $\pi_{j|r} = m_r / M_r$ ← SRS

$\Big\}$ $\pi_j = \Pi_r \pi_{j/r} = \frac{m_r \ell}{N}$

EPSEM: $m_r = m \ \forall r$
$$t[y] = \sum_{r \in S_B} \frac{N}{\ell \cancel{M_r}} \sum_{j \in S_r} \frac{y_j \cancel{M_r}}{m} = \frac{N}{n} \sum_r \sum_j y_j = N \bar{\bar{y}} = N \cdot \frac{1}{\ell} \sum_r \bar{y}_r$$

MORE COMPLEX

ESTIMATOR $e = \sum_{r \in S_B} d_r(s_B) \, t_r$

$d_r(s_B)$ ← CALIBR'N COEFTS

t_r ← ESTIMATOR OF TOTAL IN r^{th} PSU

$$\mathbb{E}[e] = \mathbb{E}_{I}\left[\mathbb{E}_{II}(e \mid s_B)\right]$$

$\mathbb{E}_{I}$ ← 1^{ST} STAGE DESIGN

$\mathbb{E}_{II}$ ← 2^{ND} STAGE DESIGN

$\mathbb{E}_{II}(t_r \mid s_B) = T_r$ ← USUALLY WANT THIS TO HOLD

$$\Rightarrow \mathbb{E}[e] = \mathbb{E}_{I} \sum_r d_r(s_B) \, T_r = \sum_{r=1}^{L} T_r = T[y]$$

$$\mathbb{V}[e] = \mathbb{E}_{I}\,\mathbb{V}_{II}(e|s_B) + \mathbb{V}_{I}\,\mathbb{E}_{II}(e|s_B) =$$

$$= \mathbb{E}_{I}\left\{\sum_{r\in s_B} d_r^2(s_B)\,\mathbb{V}_{II}(t_r|s_B)\right\} + \mathbb{V}_{I}\left(\sum_r d_r(s_B)\,T_r\right)$$

E.G. $I = \pi ps$, $II = SRS$:

$$\Rightarrow \quad t[y] = N\bar{\bar{y}},$$

$$\mathbb{V}[N\bar{\bar{y}}] = \frac{N^2}{\ell^2}\sum_r \frac{\pi_r}{m_r}\left(1-\frac{m_r}{M_r}\right)S_r^2 + \frac{1}{2}\sum_{r\neq q}(\pi_r\pi_q-\pi_{rq})\left(\frac{T_r}{\pi_r}-\frac{T_q}{\pi_q}\right)^2$$

$\mathbb{V}$ ESTIM~

NEED TO BUILD UP $j \to r \to S$

HENCE A BACKWARDS DECOMP~ MIGHT BE USEFUL:

$$\mathbb{V}[e] = \mathbb{E}_{II}\,\mathbb{V}_{I}(e|s_r, r=1,\dots,L) + \mathbb{V}_{II}\,\mathbb{E}_{I}(e|s_r, r=1,\dots,L)$$

$\uparrow$ ESSENTIALLY $\mathbb{V}_{II}[t_r]$

$d_r(s_B) = 1/\pi_r$, ℓ = FIXED

$$\Rightarrow \quad 1^{ST} \text{ TERM} = \frac{1}{2}\sum_{r\neq q}^{L}(\pi_r\pi_q-\pi_{rq})\left(\frac{t_r}{\pi_r}-\frac{t_q}{\pi_q}\right)^2$$

HENCE GENERALLY

$$v[e] = v_{I} + \sum_{r\in s_B} d_r^*(s_B)\,v_{II,r}$$

$I = \pi ps$, $II = SRS$ OF SIZE m_r

$\Rightarrow$

$$v_{I} = \frac{1}{2}\sum_{r\neq q\in s_B}\left(\frac{\pi_r\pi_q-\pi_{rq}}{\pi_{rq}}\right)\left(\frac{t_r}{\pi_r}-\frac{t_q}{\pi_q}\right)^2$$

$$d_r^*(s_B) = 1/\pi_r, \quad v_{II,r} = \frac{M_r^2}{m_r}\left(1-\frac{m_r}{M_r}\right)\frac{1}{m_r-1}\sum_{j\in s_r}(y_j-\bar{y}_r)^2$$

* $\ell \ll L$, HIGH ENTROPY WITH SELECTION PROBTY n_r/ℓ

$$\Rightarrow \quad n_{rq} \simeq n_r n_q \,(\ell-1)/\ell$$

$$(n_r n_q - n_{rq})/n_{rq} \sim 1/(\ell-1)$$

$$\Rightarrow \quad v[\,t[y]\,] \simeq \frac{1}{2}\frac{1}{\ell-1}\sum_{r\neq q}\left(\frac{t_r}{n_r}-\frac{t_q}{n_q}\right)^2 \simeq \frac{N^2}{\ell}\left[\frac{1}{\ell-1}\sum_r(\bar{y}_r-\bar{\bar{y}})^2\right]$$

ONLY DEPENDS ON y_r!

* $\mathcal{L}P$ 04:

CLUSTERS OF EQ SIZES $M \to m \ \forall r$

$$V_{CL}[t] = (ML)^2\left[\left(1-\frac{\ell}{L}\right)\frac{S_b^2}{\ell}+\left(1-\frac{m}{M}\right)\frac{S_w^2}{m\ell}\right]$$

$$S_b^2 = \frac{1}{L-1}\sum_{r=1}^{L}(\bar{Y}_r-\bar{Y})^2 \;;\; S_w^2 = \frac{1}{L(M-1)}\sum_r\sum_j(Y_j-\bar{Y}_r)^2$$

$$\uparrow \qquad\qquad \uparrow$$

$$s_b^2 = \frac{1}{\ell-1}\sum_{r=1}^{\ell}(\bar{y}_r-\bar{y})^2 \;;\; s_w^2 = \frac{1}{\ell(m-1)}\sum_r\sum_j(y_j-\bar{y}_r)^2$$

$$v = (ML)^2\left[\left(1-\frac{\ell}{L}\right)\frac{s_b^2}{\ell}+\left(1-\frac{m}{M}\right)\frac{\ell}{L}\frac{s_w^2}{\ell m}\right]$$

$\ell \ll L \Rightarrow 2^{ND}$ TERM IS SMALL

STAT 9100 SVY – PPS SAMPLING / ESTIM~

πPS – PROB PROPORTIONAL TO SIZE

WHY? LOOK AT $\left(\frac{Y_j}{\pi_j} - \frac{Y_k}{\pi_k}\right)^2$ TERMS!

MORE GENERAL: SELECT PSU'S WITH PROB
$\pi_r = \ell \alpha_r$, $\Sigma \alpha_r = 1$, $\alpha_r \leq 1/\ell$ (OTHERWISE INCLUDE wp1)

SAMPLING

OFTEN THE CASE IS THAT $n=2$, SO MOST SAMPLING SCHEMES WORK EASILY WITH $n=2$, WITH POSSIBLE EXTENSIONS TO $n>2$

ASSUME WE CAN SELECT ONE UNIT WITH GIVEN PROBTY.

THE JOINT PROBTIES OF SELECTION, AND HENCE $V[t]$ VARY FROM ONE SCHEME TO ANOTHER!

(*) SUCCESIVE SAMPLING

DRAW ℓ TIMES w/o REPLACEMENT:
1ST DRAW: UNIT r, PROB p_r
2ND DRAW: UNIT q, PROB $p_q/(1-p_r)$

$$\pi_r = p_r\left(1 + \sum_{q \neq r} \frac{p_q}{1-p_q}\right), \quad \pi_{rq} = p_r p_q\left(\frac{1}{1-p_r} + \frac{1}{1-p_q}\right)$$

FIND p_r: $\pi_r = \ell \alpha_r$ ← ITERATIVE PROCEDURE?

(*?)

FELLEGI'S METHOD

(*)

1ST DRAW: $p_r = \alpha_r$

2ND DRAW: $\frac{p_q}{1-p_r}$ WHERE $\forall q \sum_{t \neq q} \frac{\alpha_t p_q}{1-p_t} = \alpha_q$

ROTATING DESIGNS: REPEAT THE 2^{ND} DRAW STRATEGY

(*)

SAMPFORD'S METHOD

1^{ST} DRAW: r, $p_r = \alpha_r$

$2^{ND}, \ldots, \ell^{TH}$ DRAW: WITH REPLACEMENT, $p_r \propto \frac{\alpha_r}{1-\ell\alpha_r}$

REJECT & RESTART IF REPEATED UNITS

$\ell \nearrow \Rightarrow$ REJECT MORE OFTEN, TAKES LONGER - IMPORTANT FOR SIMUL^N!

MADOW'S ORDERED SYSTEMATIC PROCEDURE

$(0,1] = (0, \alpha_1] \cup (\alpha_1, \alpha_1 + \alpha_2] \cup \ldots (1-\alpha_L, 1]$

TAKE $\upsilon \sim U(0, 1/\ell]$

UNIT r IS SELECTED $\Leftarrow \sum_{j=0}^{r-1} \alpha_j < \upsilon + \frac{\xi}{\ell} \leq \sum_{j=0}^{r} \alpha_j, \xi = 0, 1, \ldots, \ell$

AS IS THE CASE WITH SYSTEMATIC SAMPLING, SOME $\pi_{rq} = 0 \Rightarrow$ UNBIASED $\hat{V}$ NOT FEASIBLE

RANDOM SYSTEMATIC PROCEDURE

RANDOMIZE THE ORDER OF UNITS FIRST!

π_{rq} = DIFFICULT :(

SAMPLING WITH REPLACEMENT

SAMPLE WR @ 1^{ST} STAGE

SAMPLE $\perp\!\!\!\perp$ @ SUBSEQ STAGES

RAO-HARTLEY-COCHRAN

$1, \dots, L \to 1, \dots, \ell$ GROUPS

"TOTAL MASS" OF THE jTH GROUP : $P_j = \sum_{r \in j\text{TH GROUP}} \alpha_j$

SELECT 1 PSU $r(j)$ FROM EACH GROUP $\Rightarrow$ α_r / P_j

CONDN'L HT ESTIMATOR: $e = \sum_j t_{r(j)} P_j / \alpha_{r(j)}$

APPROX VARIANCE ESTIMATORS AVAILABLE

BERNOULLI SAMPLING

$\forall r = 1, \dots, L$ TAKE $I_{rs} \sim \text{Bernoulli}(\ell \alpha_r)$

DISADVANTAGE: RANDOM $n(s)$

SIMPLE REJECTIVE SAMPLING

- SELECT ℓ UNITS WR WITH SELECTION PROB p_r,

OR

- TAKE BERNOULLI SAMPLE WITH PROBABILITIES λ_r
  REJECT IF $n(s) \neq \ell$

NICE ENTROPY PROPERTIES.

APPROX$\sim$ OF VARIANCE OF HT ESTIMATOR

$$\pi_{rq} \simeq \pi_r \pi_q \frac{\ell-1}{\ell}$$ FOR (*) DESIGNS

SRSWOR: $\pi_{jk} = \pi_j \pi_k \dfrac{N(n-1)}{n(N-1)} \simeq \pi_j \pi_k$

$$\Rightarrow \Omega_{jk} \simeq \pi_j \pi_k \frac{N-n}{n(N-1)} \sim O(nN^{-2})$$

BREWER & DONADIO (2003)

$$\pi_{jk} \simeq \pi_j \pi_k \frac{c_j + c_k}{2}$$

$c_j \leftarrow$ # OF CHOICES ; SLIGHTLY BETTER c_j: (18) OF B&D (2003)

EFFICIENCY OF PPS SAMPLING (LP04):

$$V_{SRS}[t] - V_{pps}[t] = \frac{N^2}{n} \mathrm{Cov}\left(\alpha, y^2/\alpha\right)$$

STAT 9100 SVY : IMPORTANT STAT RESULTS

Note Title 1/20/2007

CHEBYSHEV'S INEQUALITY

$$MSE_{p(s)}[t(y)] = E_{p(s)}(t(y) - T(y))^2 =$$

$$= \text{For some } c, \quad \sum_{s: |t(y)-T(y)| < c} (\quad)^2 + \sum_{s: |\quad| \geq c} (\quad)^2 \geq$$

$$\geq c^2 \, \mathbb{P}\{ |t(y) - T(y)| \geq c \}$$

$$\Rightarrow \mathbb{P}\{ \qquad \} \leq MSE[t(y)] / c^2$$

EXISTENCE RESULTS

$\int t^*$ is UMVUE: $\mathbb{E}_{p(s)} t^* = T(y), \forall t \quad V_{p(s)}(t^*) \leq V_{p(s)} t$

$\int \Omega$ BE THE RANGE OF Y : $\Omega = \{ \prod_{i=1}^{N} (a_i, b_i) ; a_i < y_i < b_i \}$

$\forall A \in \Omega$:

$$t_A = t^*[y] - t^*[A] + A \qquad , A = \sum_{i=1}^{N} A_i$$

$$\mathbb{E}[t_A] = T[y]$$

$$V[t_A] = E_{p(s)}[t^*[y] - t^*[A] + A - y]^2 = 0$$

WHEN $\mathcal{U} = A$

$$V_p[t^*] \leq \inf_{\Omega} V[t_A] = 0$$

$\Rightarrow$

ONLY DESIGN / ESTIMATOR WITH 0 VARIANCE CAN BE UMVUE

GODAMBE (1955)

LINEAR ESTIMATORS:

$$t = \sum b_{is} y_i I_{is}$$

$\Rightarrow$

$t \equiv t_{HT}$ WITH OTHER DESIGN RESTRICTIONS

MOMENT GENERATING FUNCTIONS

ASSUME $T[Y]=0$

$$M(t) = \mathbb{E}\exp t\sum_{j\in S} y_j = \mathbb{E}_{p(s)}\exp\left[\sum_{j=1}^{N} t I_{js} y_j\right] =$$

$$= \{SRS\} = \frac{1}{C_N^n} \sum_{s:n(s)=n} \exp\left[t \sum_{j\in S} y_j\right]$$

CONDITIONAL OF BERNOULLI SAMPLE:

$$B = \mathbb{P}[n(s)=n] = C_N^n \lambda^n (1-\lambda)^{N-n}, \quad \lambda = n/N$$

↑ IN THE BOOK

JOINT MGF OF SAMPLE SIZE AND SAMPLE SUM

$$M(u,t) = \mathbb{E}\exp\left\{\sum_{j=1}^{N} t I_{js} y_j + I_{js} u\right\} = \prod_{j=1}^{N} \mathbb{E}\exp[I_{js}(ty_j+u)]$$

$$= \prod_{j=1}^{N} \left[\lambda e^{ty_j+u} + (1-\lambda)\right]$$

EXPECTATIONS SUBTRACTED:

$$\left.\begin{aligned} n(s)-n &= \sum_{j=1}^{N} (I_{js}-\lambda) \\ \sum_{j\in S} y_j - \mathbb{E}\sum_{j\in S} y_j &= \sum_{j=1}^{N} (I_{js}-\lambda) y_j \end{aligned}\right\}$$

$$M(u,t) = \mathbb{E}\exp\left[\sum_{j=1}^{N} t(I_{js}-\lambda)y_j + (I_{js}-\lambda)u\right] =$$

$$= \prod_{j=1}^{N} \left[\lambda e^{(1-\lambda)(ty_j+u)} + (1-\lambda)e^{-\lambda(ty_j+u)}\right]$$

COND'L MGF: DIVIDE BY PROB $n(s)=n$

$$M(t) = \frac{\prod_{j=1}^{N} \left[\lambda e^{(1-\lambda)ty_j} + (1-\lambda)e^{-\lambda t y_j}\right]}{B}$$

CENTAL LIMIT THEOREM

SEQ OF ARRAYS

$y_1^{(N)} \dots y_N^{(N)}$, $N \to \infty$

n = FUNCTION OF N

$$\frac{N-1}{N} \mathbb{V}_{p(s)} \sum_{j \in s} y_j = \underbrace{N\lambda}_{n}(1-\lambda) \underbrace{\sum_{j=1}^{N} \overbrace{(y_j - \mu_j)^2}^{=\tilde{y}_j^2} / N}_{S_N'^2} =: V_N$$

TAILS CONDITIONS:

LINDEBERG : $\lim_{N\to\infty} d_N(\varepsilon) = 0 \quad \forall \varepsilon > 0$

$$d_N(\varepsilon) = \frac{1}{S_N^2} \sum_{j : |\tilde{y}_j| > \varepsilon \sqrt{V_N}} \tilde{y}_j^2$$

LYAPUNOV: $\lim_{N\to\infty} \frac{Q}{V_N^{\frac{2+\delta}{2}}} = 0$, $Q = \lambda(1-\lambda) \sum_{j=1}^{N} |\tilde{y}_j|^{2+\delta}$

$$V = N\lambda(1-\lambda) = n(1 - n/N) \to \infty \text{ as } N \to \infty$$

CLT: APPROXIMATE $M(t)$ NEAR ZERO,

IGNORE CONTRIBUTIONS OF "OUTLYING" TERMS

$$\Rightarrow \quad \frac{\sum_{j\in s} - n\mu_y}{V_N^{1/2}} \xrightarrow{d} N(0,1)$$

MORE COMPLEX DESIGNS —

DEPENDS ON THE SAMPLING SCHEME:

- SIMPLE REJECTIVE DESIGN — HAJEK (1964)
- SUCCESIVE SAMPLING — ROSEN (1972)
- STRATIFIED SAMPLES — BICKEL & FREEDMAN (1984)
  ($H \to \infty$ OR $N_h \to \infty$, $2 \le n_h \le N_h - 1$)
- STRATIFIED SAMPLES — KREWSKI & RAO (1981)
  n_h BDD, $H \to \infty$, $\max W_h = O(H^{-1})$
  + LYAPUNOV COND~, + CONV~ OF POP~ COV MX
- HORVITZ-THOMPSON : P.K. SEN (1980, 1988)
- MULTIVARIATE EXTENSIONS

DESIGN-CONSISTENCY

WE KNOW: $\dfrac{e - T[y]}{\sqrt{V[e]}} \xrightarrow{d} N(0,1)$

$T[y]$ AND $V[e]$ (ESP THE LATTER) ARE DESIGN-BASED CONCEPTS.

DOES THE STUDENTIZED VERSION $\dfrac{e - T[y]}{\sqrt{v(e)}}$ ALSO $\xrightarrow{d} N(0,1)$?

$v(e)$ HAS TO BE DESIGN-CONSISTENT!

$$\frac{v(e)}{V(e)} \xrightarrow{P} 1$$

HERE, P IS THE PROBTY OVER SAMPLES $p(s)$

HENCE, NEED TO BE ABLE TO SHOW CONSISTENCY OF $v(e)$

FOR THE BASIC DESIGNS CONSIDERED THUS FAR (SRS, STRATIFIED, HT + YGS, MULTI-STAGE) AND BASIC $v(e)$, THIS CONSISTENCY DOES HOLD

SVY - AUX VARIABLES

Note Title 2/6/2007

SO FAR: FOCUS ON SINGLE VARIABLE ESTIMATORS
SIMPLE EXTENSIONS: AUX INFO ESTIMATORS

RATIO ESTIMATOR

SUPPOSE VARS Y AND X ARE COLLECTED

$$(y_j, x_j) \quad , j \in S$$

INTEREST IS IN RATIO $R = T[Y]/T[X]$

$\Rightarrow$ ESTIMATE BY

$$r = t[y]/t[x] \qquad \text{(COMBINED RATIO ESTIMATOR)}$$

OR

$$r_s = \sum_{h=1}^{H} W_h r_h \quad , r_h = t_h[y]/t_h[x]$$

(SEPARATE RATIO ESTIMATOR)

RATIO ESTIMATOR OF TOTAL:

$$t_r[y] = r\,T[x]$$

↑ ASSUMED KNOWN

REGRESSION ESTIMATOR OF TOTAL:

$$t_L[y] = t[y] + b(T[x] - t[x]),$$

$$b = \frac{S_{xy}}{S_x^2} = \text{SAMPLE REGRESSION COEFT}$$

$$S_{\xi\eta} = \frac{1}{N-1} \sum_{j=1}^{N} (\xi_j - \bar{\xi})(\eta_j - \bar{\eta}), \quad \xi, \eta = X, Y$$

$$s_{\xi\eta} = \frac{1}{n-1} \sum_{j \in s} (\xi_j - \bar{\xi})(\eta_j - \bar{y}), \quad \xi, \eta = x, y$$

SAMPLE AVERAGE

(OR DESIGN-BASED ESTIMATOR, IN GENERAL)

COEFF OF VARIATION: $C_{\xi\eta} = S_{\xi\eta} / \bar{\xi}\bar{\eta}$

$\mathbb{E}\, r \neq R \quad \Leftarrow$ NON-LINEARITY

BIAS $\quad \mathbb{B}[r] = ?$

VARIANCE $\quad \mathbb{V}[r] = ?$

IF THE LATTER IS UNAVAILABLE, MSE $\mathbb{M}[r] = ?$

LINEARIZATION / TAYLOR SERIES EXPANSION APPROACH

SUPPOSE $\theta = f(T[Y])$ IS A POP~ PARR

$\hat{\theta} = f(t[y])$ IS A CONSISTENT ESTIMATOR

$$\hat{\theta} - \theta = \nabla f \cdot (t[y] - T[y]) + \frac{1}{2} (t[y] - T[y])' [\nabla^2 f] (t[y] - T[y])$$

+ SMALLER ORDER TERMS

SUPPOSE $\mathbb{E}_{p(s)}\, t[y] = T[y] \Rightarrow$

$$\mathbb{B}\,\hat{\theta} = \frac{1}{2} \mathbb{E} \, (t[y] - T[Y])' [\nabla^2 f] (t[y] - T[Y]) \simeq O_p(n^{-1}),$$

$$\mathbb{M}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 \approx \nabla f' \, \mathbb{V}(t[Y]) \, \nabla f \qquad (*)$$

ON THE OTHER HAND,

$$\hat{\theta} - \theta \simeq \nabla f' \left(\sum_{j \in s} \frac{y_j}{\pi_j} - T[y] \right)$$

$$= \sum_{j \in s} \{ \nabla f' y_j \} / \pi_j - \nabla f' T[y]$$

= A LINEAR STATISTIC OF

$$u_j = \nabla f' y_j$$

SO $$M[\hat{\theta}] \simeq \sum_{j \neq k} \left(\frac{u_j}{\pi_j} - \frac{u_k}{\pi_k} \right)^2 \Omega_{jk}$$

$$m[\hat{\theta}] \simeq \sum_{\substack{j \neq k \\ \in s}} \left(\frac{u_j}{\pi_j} - \frac{u_k}{\pi_k} \right)^2 \frac{\pi_j \pi_k - \pi_{jk}}{\pi_{jk}}$$

BACK TO THE RATIO ESTIMATOR

DESIGN = SRS

$$\mathbb{B}[t_r[y]] = \frac{1-f}{n} \bar{Y} C_x (C_x - \rho C_y) \quad , C_\xi = \sqrt{C_{\xi\xi}}$$

$$M[t_r(y)] \simeq V[t_r(y)] = \frac{1-f}{n} S_d^2$$

$$S_d^2 = \sum_{j=1}^{N} \frac{(y_j - R x_j)^2}{N-1} = S_{yy} - 2 R S_{xy} + R^2 S_{xx}^2$$

(**)

WITH PLUG-IN ESTIMATOR

$M[t_r] \leq V[t_r]$ FOR REASONABLE DESIGNS

$$| \mathbb{B}(t_r[y]) / V^{1/2}(t_r[y]) | \leq C_{\bar{x}} = \sqrt{(1-f) C_{xx} / n}$$

APPLICATION: SAMPLE MEAN IN CLUSTERED SAMPLES

DESIGN: 1 ST STAGE = SRS OF PSUs, ℓ OUT OF L

2 ND STAGE = SRS OF $n_r = \tau N_r$ UNITS

$$\bar{y} = \frac{t[y]}{t[1]} \underset{\uparrow\, L/\ell}{=} \frac{\sum N_r \bar{y}_r}{\sum N_r} \underset{\uparrow\, \tau}{=} \frac{\sum_r \sum_{j \in B_r} y_j}{\sum_r n_r} \underset{\uparrow\, 1/\ell}{=} \frac{\bar{v}}{\bar{n}}$$

$$v_r = \sum_{j \in B_r} y_j$$

$$v(\bar{v}) = \frac{1}{\ell(\ell-1)} \sum_r (v_r - \bar{v})^2$$

$$v(\bar{n}) = \frac{1}{\ell(\ell-1)} \sum_r (n_r - \bar{n})^2$$

$$\mathrm{cov}(\bar{v}, \bar{n}) = \frac{1}{\ell(\ell-1)} \sum_r (v_r - \bar{v})(n_r - \bar{n})$$

$$\mathrm{var}(\bar{y}) = \left(\frac{1}{\bar{n}^2}\right)\left[v(\bar{v}) - 2\bar{y}\,\mathrm{cov}(\bar{v}, \bar{n}) + \bar{y}^2 v(\bar{n}) \right] =$$

$$= \frac{1}{\bar{n}^2} \frac{1}{\ell(\ell-1)} \sum_r \left[(v_r - \bar{v})^2 - 2\frac{\bar{v}}{\bar{n}}(v_r - \bar{v})(n_r - \bar{n}) + \frac{\bar{v}^2}{\bar{n}^2}(n_r - \bar{n})^2 \right] =$$

$$= \frac{1}{\bar{n}^2 \ell(\ell-1)} \sum_r \left[(v_r - \bar{v}) - \frac{\bar{v}}{\bar{n}}(n_r - \bar{n}) \right]^2 =$$

$$v_r - \cancel{\bar{v}} - \frac{\bar{v}}{\bar{n}} n_r + \cancel{\bar{v}}$$

$$= \frac{1}{\bar{n}^2 \ell(\ell-1)} \sum_r n_r^2 \left[\bar{y}_r - \bar{y} \right]^2$$

$$\bar{y}_r = v_r / n_r$$

A STRANGE LOOKING ESTIMATOR!

HOMEWORK:

(i) DERIVE THE VARIANCE OF THE RATIO (**) USING (*)

(ii) FIND THE LINEARIZED CONTRIBUTIONS u_r FOR THE RATIO ESTIMATOR, AND FIND THEIR VARIANCE AND ITS ESTIMATOR

KG 2.4.9

REGR~ ESTIMATOR

$$t_{reg}[y] = t[y] + b(T[x] - t[x])$$

$$\mathbb{V}[t_{reg}] \simeq \mathbb{M}[t_{reg}] \underset{\substack{\uparrow \\ SRS}}{\simeq} \frac{1-f}{n} \frac{1}{N-1} \sum_{j=1}^{N} D_j^2 = \frac{1-f}{n} S_y^2 (1-\rho^2)$$

$$D_j = (Y_j - \bar{Y}) - B(X_j - \bar{X}) \quad \leftarrow \text{LINEAR~ TERM}$$

$B =$ POPULATION REGR~ COEFT

MOST OF THE TIME,

$$\mathbb{V}[t_{reg}] \leq \mathbb{V}[t_{rat}] \leq \mathbb{V}[t_{HT}]$$

FOR REASONABLY STRONGLY CORRELATED X, Y

OTHER ESTIMATORS OF VARIANCE:

MODEL-BASED / ASSISTED

SVY – REGRESSION W/COMPLEX DATA

Note Title 2/11/2007

RESOURCES:

BINDER (1983)
SKINNER (1989)
FULLER (2002)

I.I.D.:

$$Y_j = X_j \beta + e_j$$

$$\mathbb{E}\, e_j = 0$$

OLS: $\sum_{j=1}^{N} (y_j - x_j B)^2 \to \min$

– IS THIS A SENSIBLE THING?

MAY NOT BE A CORRECT MODEL, BUT AT LEAST THAT'S SOMETHING YOU CAN ALWAYS COMPUTE

NORMAL EQNS:

$$\sum_{j=1}^{N} x_j (y_j - x_j B) = 0$$

– A POPULATION TOTAL OF ZERO $\in \mathbb{R}^p$!

SOMETHING WE COULD TRY TO ESTIMATE

SAMPLE / ESTIMATOR:

$$\sum_{j \in S} x_j (y_j - x_j b) = 0 \qquad \leftarrow \text{OLS}$$

$$\sum_{j \in S} \frac{x_j (y_j - x_j b)}{\pi_j} = 0 \qquad \leftarrow \text{HT}$$

ANALOGY TO RATIO ESTIMN

$$b_w = (X'WX)^{-1}(X'WY)$$

$$W = \text{DIAG}(\pi_j^{-1})$$

UNLESS THE DESIGN IS SELF-WEIGHTING, THE OLS ESTIMATOR MIGHT BE BIASED.

HOWEVER, IF

...

BLAH-BLAH ABOUT $\pi_j^{-1} \in \mathcal{L}(X)$

SEE LATER

THUS TO ENSURE DESIGN CONSISTENCY,

1. USE WEIGHTS
2. USE DESIGN INFO IN THE MODEL

LET US REVISIT OLS $b_1 = b_I$ AGAIN

$$b_1 = \left(\sum_{j\in S} x_j x_j'\right)^{-1}\left(\sum_{j\in S} x_j' y_j\right)$$

SUPPOSE THE DESIGN IS SUCH THAT $b_1 \xrightarrow{p} B$

$$b_1 - B = ?$$

RECALL THE LINEARN FRAMEWORK:

1) BASIC ESTIMATOR t
2) TAYLOR SERIES LINEAR TERM ∇f
3) GET $\mathbb{V}[t]$
4) $M](\hat{\theta}) \approx \nabla f \cdot \mathbb{V}[t] \cdot \nabla f'$

1) BASIC ESTIMATOR:

$$t_b = \sum_{j\in S} x_j (y_j - x_j' b) = 0$$

$$T = \sum_{j=1}^{N} x_j (y_j - x_j' B) = 0$$

2) $$\nabla f(B) = \frac{\partial T}{\partial B} = \sum_{j\in S} x_j x_j', \ \mathbb{E}\,\nabla f(B) = \frac{n}{N}\sum_{j=1}^{N} x_j x_j'$$

3) $$\mathbb{V}[t] = \mathbb{E}_{p(s)}\, t_B t_B' = \mathbb{E}\left(\sum_{j\in S} x_j e_j\right)\left(\sum_{j\in S} x_j e_j\right)'$$

$$= \mathbb{E}\sum_{j\in S} x_j x_j' e_j^2 + \mathbb{E}\sum_{j\neq k} x_j x_k e_j e_k$$

SRS: $\frac{n}{N}\sum_j^N x_j x_j' e_j^2 - \frac{n(n-1)}{N(N-1)} \sum_{j=1}^N x_j x_j' e_j^2$

$\sum_{k \neq j} x_k e_k = -x_j e_j$

$= \frac{n(N-1) - n(n-1)}{N(N-1)} \sum_j x_j x_j' e_j^2$

$= \frac{n(N-n)}{N(N-1)} \sum_j x_j x_j' e_j^2$

4) $V[b] \approx \left(\frac{n}{N}\sum_{j=1}^N x_j x_j'\right)^{-1} \frac{n(N-n)}{N(N-1)} \sum_j x_j x_j' e_j^2 \left(\frac{n}{N}\sum_{j=1} x_j x_j'\right)^{-1}$

$\simeq \frac{1}{n} \left(\frac{1}{N}\sum_{j=1}^N x_j x_j'\right)^{-1} \left[\frac{1}{N}\sum x_j x_j' e_j^2\right] \left(\frac{1}{N}\sum_{j=1}^N x_j x_j'\right)^{-1}$

$v(b) \approx \frac{1}{n}\left(\frac{1}{n}\sum_{j\in S} x_j x_j'\right)^{-1} \left(\frac{1}{n}\sum_{j\in S} x_j x_j' e_j^2\right) \left(\frac{1}{n}\sum_{j\in S} x_j x_j'\right)^{-1}$

AKA SANDWICH/ROBUST ESTIMATOR

COMPLEX DESIGN: $v_L(b) = V_{xx}^{-1} v_{xy} V_{xx}^{-1}$

$V_{xx} = \frac{1}{n}\sum_h \sum_\ell \sum_j x_{h\ell j} x_{h\ell j}^T$

$v_{xy} = \hat{V}[(y - x^T b)x] = \sum_h W_h \frac{(1-f_h)\ell_h}{\ell_h - 1} \sum_\ell (u_{h\ell} - \bar{u}_h)(u_{h\ell} - \bar{u}_h)^T$

$u_{h\ell} = \sum_{j \in B_{h\ell} \cap S_h} u_{h\ell j}, \quad u_{h\ell j} = \frac{1}{n} x_{h\ell j} \left[y_{h\ell j} - b^T x_{h\ell j}\right]$

USE OF DESIGN INFO

$$\bar{Y}_{reg} = \bar{X}_N \hat{\beta}$$

KNOWN POP~ TOTAL

CALIBR~: $\sum_{j \in S} d_{js} \alpha_j = T[\alpha]$

$$\mathbb{E}_{p(s)} \bar{X}_N \hat{\beta} = \mathbb{E}_{p(s)} \bar{X}_N (X'WX)^{-1} X'WY$$

$$(X'WX)^{-1} X'WY = (X'WX)^{-1}(X'WXB + X'WE) = B + (X'WX)^{-1} X'WE$$

FOR IT TO BE UNBIASED, NEED TO HAVE
$\mathbb{E}_{p(s)} (X'WX)^{-1} X'WE = 0$

SUFFICIENT COND~: $\mathbb{E}_{p(s)} X'WE = 0$

LOCATION INVARIANCE WRT Y: $W D_\pi^{-1} \mathbb{1} \in \mathcal{L}(X)$

SO THAT THE DESIGN EXPECT~ CAN BE CARRIED THROUGH $X'W$
(OTHERWISE, Y'S AND E'S MAY "GET LOST" IN THE $X^\perp$ SPACE, AND $\mathbb{E}_{p(s)}$ WILL PRODUCE BIASES)

WAYS TO ENSURE THAT:

1. HAVE $\mathbb{1} \in X$, USE $W = D_\pi$

   $\Rightarrow$ SUPPOSE THIS IS THE SINGLE REGRESSOR, THEN

$$\bar{Y} = \Big(\sum_{j \in S} \pi_j^{-1}\Big)^{-1} \sum_{j \in S} \pi_j^{-1} Y_j = t_{HT}[Y] / t_{HT}[\mathbb{1}]$$

2. HAVE $W = I$, AND USE π_j^{-1} AS ONE OF REGRESSORS

3. HAVE π_j AS ONE OF REGRESSORS, AND USE $W = D_\pi^2$

$$\Rightarrow \tilde{Y} = N^{-1} \sum_{j \in S} y_j / \pi_j \quad \text{— HT ESTIMATOR}$$

4. STRATA WITH FRACTIONS f_h, SRS WITHIN STRATA $\Rightarrow$ CAN USE STRATA INDICATORS TO ENSURE THE LINEAR SPACE RESTRICTION

↳ AN EXAMPLE OF MODELLING DESIGN RATHER THAN PROVIDING DESIGN-BASED INFERENCE

MULTIVARIATE DESIGN EFFECTS (SEC. 2.11 OF SHS)

VECTOR $\hat{\beta} \in \mathbb{R}^p \Rightarrow$ WHAT IS A PROPER WAY TO LOOK AT DEFF?

1D: $\displaystyle DEFF(t, V_0) = \frac{\mathbb{V}_{p(s)}[t]}{V_0[t]}$ ← ESTIMATOR UNDER SOME "NULL" ASSUMPTIONS

$$DEFF(\hat{\theta}, V_0) = \Delta = \mathbb{E}(V_0^{-1})\, \mathbb{V}(\hat{\theta})$$

TYPICALLY, $V_0 = \mathbb{V}_{SRS}[\hat{\theta}]$

GENERALIZED DESIGN EFFECTS:

$$\max_{c\neq 0} \operatorname{deff}(c'\hat{\theta}, c'V_0 c) = \delta_1 \geq \dots \geq \delta_p = \min_{c\neq 0}(c'\hat{\theta}, c'V_0 c) = \operatorname{Spec}\{\Delta\}$$

EFFECT OF GENERALIZED DEFFS:

$$\chi^2 = (\hat{\theta}-\theta)' V_0^{-1} (\hat{\theta}-\theta) \sim \sum_{i=1}^{p} \delta_i \chi^2_1$$

(RAO & SCOTT, 1981)

SATTERTHWAITE (1946) APPROX$^{\sim}$:

$$\frac{\chi^2}{\bar{\delta}(1+a^2)} \sim \chi^2_q \qquad q = \frac{p}{1+a^2}, \quad a = CV\{\delta\}$$

COEFF OF VARIATION

$$p\bar{\delta} = \operatorname{tr}\Delta, \qquad p\bar{\delta}^2(1+a^2) = \operatorname{tr}\Delta^2$$

MISSPECIFICATION EFFECT (SEC 2.2 OF SHS)

- EFFECT OF THE DESIGN ON VARIANCE ESTIMATORS

$$\operatorname{MEFF}(\hat{\theta}, v_0) = \mathbb{V}_{true}[\hat{\theta}] / \mathbb{E}_{true}[v_0]$$

MAY BE DESIGN-BASED OR MODEL-BASED

CLUSTER DEFF $(n=2)$: $1+\rho_{icc}$ – VARIANCE IS 50% MORE THAN WHAT IT WOULD BE UNDER SRS

CLUSTER MEFF : $\dfrac{1+\rho_{icc}}{1-\rho_{icc}}$ – VARIANCE IS 3 TIMES MORE THAN WHAT AN SRS-BASED ESTIMATOR WILL TELL

OFTEN: $\mathbb{E}_{TRUE}\ \sigma_0 \simeq \mathbb{V}_{SRS}[\hat{\theta}]$

BOTH ARE USUALLY ESTIMATED BY SAMPLE-BASED $\hat{\sigma}_0$

DESIGN EFFECTS IN REGRESSION (SHS, SEC. 3.3.4)

WHAT IS THE APPROPRIATE σ_0?

(i) SRS + HOMOSKEDASTICITY : $\sigma_{OLS} = s_e^2 (X'X)^{-1}$

(ii) SRS + HETEROSKEDASTICITY : SANDWICH

(SEE ALSO OTHER CORRECTIONS IN MY MISSPECIFICATION TALK)

$$\text{MEFF}(\hat{\beta}_{OLS}, V_{OLS}) \simeq 1 + \rho C_\sigma C_x$$

$$C_\sigma = CV\ \sigma^2(x_j),\quad C_x = CV\ x_j,\quad \rho = \text{corr}(\sigma^2(x_j), x_j)$$

(iii) NESTED ERRORS / GLS :

$$y_{hj} = x'\beta + u_h + e_{hj}$$

$$E e_{hj} e_{h'j'} = \sigma_u^2 \delta_{hh'} + \sigma_e^2 \delta_{hh'} \delta_{jj'}$$

$$\frac{\mathbb{V}_{\text{NESTED ERRORS}}(\text{SLOPE})}{\mathbb{V}_{OLS}(\text{SLOPE})} \simeq 1 + (\underset{\substack{\uparrow \\ \text{CLUSTER SIZE}}}{M} - 1)\, \rho_{ICC,\varepsilon}\, \rho_{ICC,x}$$

TYPICAL DEFFS:

MEAN	$\to 2$	
SLOPE	$\to 1.2$	
CORR ~	$\to 2$	
R^2	$\to 5$	

RECALL THAT REGRESSING ON DESIGN INFO HELPS CONSISTENCY... REGRESSING ON VARS CORRELATED WITH DESIGN INFO (E.G. X'S THAT ARE CORRELATED WITHIN CLUSTERS IN THE SAME WAY AS Y'S) ACTS AS A SURROGATE

SVY – NONLINEAR MODELS

Note Title 2/17/2007

BINDER (1983)
SKINNER (1989)
THOMPSON (1997, Ch. 4, 6)

SEC. 4.1

POPULATION ESTIM~ EQ:

$$\sum_{j=1}^{N} \varphi_j(y_j, x_j, \theta_N) = 0$$

MEAN: $\varphi_j = y_j - \theta_N$
RATIO: $\varphi_j = y_j - \theta_N x_j$
C.D.F.: $\varphi_j = \mathbb{1}\{y_j \le y\} - \theta_N$

$$\varphi_s(\theta) = \sum_{j \in s} \varphi_j(y_j\, x_j\, \theta)/\pi_j$$

$$\mathbb{E}_{p(s)}\, \varphi_s(\theta) = \sum_{j=1}^{N} \varphi_j(y_j\, x_j\, \theta)$$

$$\mathbb{V}_{p(s)}\, \varphi_s(\theta) = \sum_{j=1}^{N} \left(\varphi_j/\pi_j\right)^2 \pi_j(1-\pi_j) + \sum_{j \ne k}^{N} \frac{\varphi_j}{\pi_j}\frac{\varphi_k}{\pi_k}(\pi_{jk} - \pi_j \pi_k)$$

ESTIMATOR: $v(\varphi_s) = \frac{1}{2} \sum_{j \ne k} w_{jk} \left(\frac{\varphi_j}{\pi_j} - \frac{\varphi_k}{\pi_k} \right)^2$

$$w_{jk} = (\pi_j \pi_k - \pi_{jk})/\pi_{jk}$$

CLT: $\left(\varphi_s(\theta) - \sum_{j=1}^{N} \varphi_j(y_j\, x_j\, \theta) \right) \Big/ \mathbb{V}^{1/2}[\varphi_s(\theta)] \xrightarrow{d} N(0,1)$

WITH APPROPRIATE STUDENTIZED VERSION

TAYLOR SERIES:

$$\frac{1}{N}\varphi_s(\hat\theta_s) - \frac{1}{N}\varphi_s(\theta) = -\frac{1}{N}\varphi_s(\theta) =$$

$$= \frac{\partial\varphi}{\partial\theta}(\hat\theta_s - \theta) + \frac{1}{2}(\hat\theta_s - \theta)' \frac{\partial^2\varphi}{\partial\theta\,\partial\theta'}\Big|_{\bar\theta} (\hat\theta_s - \theta)$$

$\bar\theta$ – INTERMEDIATE POINT

SEC. 4.2 — FUNCTIONS OF TOTALS,
LINEARIZATION TECHNIQUE

$$\theta = g(T), \quad \hat{\theta} = g(t)$$

$$\Rightarrow \quad \varepsilon_L(\hat{\theta};\theta) = \hat{\theta} - \theta = g(t) - g(T) =$$
$$= \sum_\alpha \left.\frac{\partial g}{\partial T_\alpha}\right|_{T_\alpha} (t_\alpha - T_\alpha) + o(\ldots)$$

@ POP~ TOTALS

$$n^{1/2}(t - T) \sim N(0, \ldots), \quad \left\|\frac{\partial g}{\partial T}\right\| > 0$$
$$\Rightarrow \quad \frac{g(t) - g(T)}{V^{1/2}\{\Sigma_\alpha \frac{\partial g}{\partial T_\alpha} t_\alpha\}} = \frac{\hat{\theta} - \theta}{V^{1/2}[\varepsilon]} \xrightarrow{d} N(0,1)$$

ESTIMATOR: $\quad v(\varepsilon_L) = v\left(\sum_{j\in s} u_j/\pi_j\right)\Big|_t$

E.G. Y-G-S ESTIMATOR $\qquad u_j = \sum_\alpha \frac{\partial g}{\partial T_\alpha} y_{\alpha j}$

4.2.3 ~ 4.2.6:
RANDOM GROUPS, BRR, JACKKNIFE, BOOTSTRAP

SEC. 4.3:
U-STATISTICS WITH SRS

SEC. 4.4:
ESTIMATING EQNS WITH
NUISANCE PARAMETERS
MOST IMPORTANT EXAMPLE —
THE COX PROP HAZARD MODEL

BINDER (1983)

DESIGN-BASED SAMPLING DISTR~ FOR PARAMETERS DEFINED AS FUNCTIONS OF DATA VALUES

E.G. REGRESSION: $X^T X B = X^T Y$

GENERALIZED LINEAR MODEL

REVIEW OF GLM:

$$y \sim f(y; \theta, \varphi) =$$
$$= \exp\left[\alpha(\varphi)\{y\theta - g(\theta) + h(y)\} + \delta(\varphi, y) \right]$$

MORE OFTEN GOES INTO DENOMINATOR

$\mathbb{E} Y = g'(\theta) \equiv \mu(\theta)$ — MEAN FN

$\mathbb{V} Y = \mu'(\theta) / \alpha(\varphi)$ — VARIANCE FN

$\theta = f(x'\beta)$ — f IS A KNOWN FN

β IS A VECTOR OF COEFTS

SCORE EQ:

$$\sum [y_k - \mu_k(f(x'\beta))] f'(x'\beta) x = 0$$

REGRESSION: $f =$ IDENTITY LINK, $g = \mu^2/2$, $\alpha = \sigma^{-2}$

LOGIT: $f =$ IDENTITY, $g = \ln p/1-p$, $\alpha = 1$, etc.

IMPLICIT PARAMETERS:

REGRESSION: $X^T X B = X^T Y$

GLM: $\sum_{j=1}^{N} [Y_j - \mu\{f(X_j^T B)\}] f'(X_j^T B) X_j = 0$

GENERALLY: $\sum_{j=1}^{N} \varphi_j (Y_j X_j B) = 0$

ESTIM~ FRAMEWORK: $W_N(\theta) = \sum_{j=1}^{N} \varphi_j(z_j; \theta) - v(\theta)$

POP~ VALUE: $W_N(\theta_0) = 0$

REGULARITY COND$\sim$:

(i) $\theta \in \operatorname{int} \Theta$; NBHOOD $\mathcal{U}(\theta) \subseteq \Theta$

(ii) SEQ OF POP$\sim$ AND DESIGNS: $\{\mathcal{U}_N, P_N(s)\}$:

- SCALED $(t - T) \xrightarrow{d} N(0, \Sigma)$
- $v(t)/V(t) \xrightarrow{P} 1$

(iii) CONTINUITY & LIMITS OF $W_N(\theta)$, $\partial W_N(\theta)$

(iv) CONTINUITY OF $\mathbb{V}[t]$

ESTIM$\sim$ OF θ:

ASSUME $t[u(\theta)]$ IS ASYMPT^{ly} NORMAL:

$$\text{SCALING}\left(t[u(\theta)] - U[\theta]\right) \xrightarrow{d} N(0, \Sigma_u(\theta))$$

$\Sigma_u(\theta)$ ESTIMATED BY $\hat{\Sigma}_u(\theta)$

$\hat{W}(\theta) = t[u(\theta)] - v(\theta)$ - ESTIM EQN FOR θ:

$\hat{\theta}$ IS SUCH THAT $\hat{W}(\hat{\theta}) = 0$

REGRESSION: $u(y_j x_j B) = -(y_j - x_j^T B)x_j$, $v(B) = 0$

$$0 = \hat{W}(\hat{\theta}) \simeq \hat{W}(\theta_0) + \left.\frac{\partial \hat{W}(\theta)}{\partial \theta}\right|_{\theta_0} (\hat{\theta} - \theta_0) + \text{REMAINDER}$$

$$\hat{W}(\theta_0) \simeq -\frac{\partial \hat{W}(\theta_0)}{\partial \theta}(\hat{\theta} - \theta_0)$$

$\mathbb{V}[\quad] \rightarrow$

$$\Sigma_u(\theta_0) \simeq \frac{\partial W_N(\theta_0)}{\partial \theta} \mathbb{V}(\hat{\theta}) \left[\frac{\partial W_N(\theta_0)}{\partial \theta}\right]^T$$

$$\Rightarrow \quad \mathbb{V}[\hat{\theta}] = \left[\frac{\partial W_N(\theta_0)}{\partial \theta}\right]^{-1} \Sigma_u(\theta_0) \left[\frac{\partial W_N(\theta_0)}{\partial \theta}\right]^{-T}$$

WITH ESTIMATOR

$$v[\hat{\theta}] = \left[\frac{\partial W(\hat{\theta})}{\partial \theta}\right]^{-1} \hat{\Sigma}_u(\hat{\theta}) \left[\frac{\partial W(\hat{\theta})}{\partial \theta}\right]^{-T}$$

GLM, CANONICAL LINK $f(\cdot) = \text{IDENTITY}$:

$$\sum_{j=1}^{N} [Y_j - \mu(x_j^T B)] X_j = 0$$

$\Rightarrow$

$$w_N(B) = \sum_{j=1}^{N} [Y_j - \mu(x_j^T B)] x_j$$

$$\frac{\partial w_N}{\partial B} = -\sum_{j=1}^{N} \mu'(x_j^T B) x_j x_j^T = X^T \Lambda(B) X$$

$$\Lambda(B) = \text{diag}[\mu'(x_1^T B), \ldots, \mu'(x_N^T B)]$$

$$\Rightarrow \mathbb{V}[\hat{B}] = [X^T \Lambda(B) X]^{-1} \Sigma(B) [X^T \Lambda(B) X]^{-1}$$

$$\Sigma(B) = \mathbb{V}[t(e_j x_j)]$$

$$e_j = Y_j - \mu(x_j^T B)$$

EACH MATRIX, AND EACH MATRIX ENTRY, IS EITHER A TOTAL OR AN ESTIMATE OF VARIANCE OF A TOTAL

$\Rightarrow$

$$v(B) = \hat{J}^{-1}(\hat{B}) \hat{\Sigma}(\hat{B}) \hat{J}^{-1}(\hat{B})$$

WHERE

$$\hat{J}(\hat{B}) = t[x_j x_j^T \mu'(x_j^T B)]$$

(AKA DERIVATIVES / JACOBIAN IN NEWTON–RAPHSON STEPS)

REGRESSION: $w_N(B) = X^T Y - X^T X B$; $\dfrac{\partial w_N(B)}{\partial B} = -X^T X$

$$\Rightarrow v(\hat{B}) = S_{XX}^{-1} \hat{\Sigma}(\hat{B}) S_{XX}^{-1}$$

$$S_{XX} = t(x_j x_j^T) \qquad \hat{\Sigma}(\hat{B}) = v(\hat{e}_j x_j)$$

EXTENSION:

ESTIMATE SIMULTANEOUSLY $(\bar{Y}, B, R^2)$:

$$W_N(\theta) = \begin{Bmatrix} Y^T \mathbb{1} - N\bar{Y} \\ X^T Y - X^T X B \\ (Y^T Y - N\bar{Y}^2)(R^2 - 1) + Y^T Y - Y^T X B \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}$$

$$\hat{R}^2 = 1 - \frac{S_{YY} - \hat{B}^T S_{XY}}{S_{YY} - N\bar{y}^2}$$

$$\frac{\partial W}{\partial \theta} = \begin{pmatrix} -N & 0 & 0 \\ 0 & -X^T X & 0 \\ 2N\bar{y}(1-R^2) & -Y^T X & Y^T Y - N\bar{y}^2 \end{pmatrix} \rightarrow \left(\frac{\partial W}{\partial \theta}\right)^{-1}$$

$$u_j = \begin{pmatrix} y_j \\ (y_j - x_j^T \hat{B}) x_j \\ (\hat{R}^2 y_j - x_j^T \hat{B}) y_j \end{pmatrix} \rightarrow \sigma[u_j]$$

$$\Rightarrow \quad V[R^2] = (S_{YY} - N\bar{y}^2)\, a^T V[u]\, a$$

$$a = 2\bar{y}(1-R^2) - \hat{B}^T \mathbb{1}$$

LOGISTIC REGRESSION:

$$\mu(X^T B) = \frac{\exp(X^T B)}{1 + \exp(X^T B)}$$

$$W_N(B) = \sum_{j=1}^{N} \left[y_j - \frac{\exp(x_j^T B)}{1 + \exp(x_j^T B)} \right] x_j$$

$$X^T \Lambda(B) X = \sum_{j=1}^{N} x_j x_j^T \frac{\exp(x_j^T B)}{[1 + \exp(x_j^T B)]^2}$$

$$\hat{\Sigma}(\hat{B}) = \sum_{j \in S} x_j x_j^T \left[y_j - \frac{\exp(x_j^T B)}{1 + \exp(x_j^T B)} \right]^2$$

ANOTHER EXAMPLE (SEC. 4.4):

LOG-LINEAR MODELS FOR CATEG DATA

(MULTINOMIAL LOGISTIC MODEL)

MOVE ON TO SKINNER (1989)

= CH 3 OF SHS (1989)

$$\ell(\theta) = \sum_j \ln f(y_j;\theta)$$

$$\hookrightarrow u_j(\theta) = \nabla \ln f(y_j;\theta)$$

POP PARR: $\theta_0 : T[u_j(\theta)] = 0$

PSEUDOMLE $\hat{\theta}_{PML} \leftarrow t[\hat{\theta}_{PML}] = 0$

- NOT THE MLE ESTIMATOR

  $\Rightarrow$ OPTIMALITY DOES NOT HAVE TO HOLD!
- VARIOUS ESTIMATORS t

  $\Rightarrow$ DIFFERENT $\hat{\theta}_{PML}$!

$$As.V[\hat{\theta}_{PML}] = \text{LINEAR} = I(\hat{\theta}_{PML})^{-1} V_L[t(\hat{\theta}_{PML})] I(\hat{\theta}_{PML})^{-1}$$

$$I(\theta) = \frac{\partial}{\partial\theta} t[u(\theta)]$$

$V_L[t(\hat{\theta}_{PML})] \leftarrow$ ESTIMATED IN ANY REASONABLE DESIGN-CONSISTENT WAY

LOGISTIC: $u_j(\theta) = (y_j - p_j[\theta])x_j$

$$I(\theta) = t[p_j(1-p_j)x_j x_j^T]$$

WALD TESTS: χ^2 ; LR TESTS: GEN DEFF

SVY: RESAMPLING ESTIMATORS

Note Title 3/14/2007

RESOURCES:

SHAO (1996) REVIEW
RAO & WU (1988) SCALED BOOTSTRAP
RAO, WU & YUE (1992) BOOTSTRAP IN WEIGHTS
KREWSKI & RAO (1981) CONSISTENCY RESULTS
RAO & WU (1985) COMPARISON OF ESTIMATORS

SUMMARY REVIEW: MY TALK, SEE:

http://web.missouri.edu/~kolenikovs/talks/survey-resampling-by2.pdf

BRR:

LET'S LOOK @ ESTIM~ OF TOTAL

$$\bar{y} = \sum_h W_h \bar{y}_h = \sum_h W_h \frac{y_{h1}+y_{h2}}{2}$$

$$v(\bar{y}) = \sum_h \frac{W_h^2}{2} v(\bar{y}_h) = \sum_h W_h^2 \left[(y_{h1}-\bar{y}_h)^2 + (y_{h2}-\bar{y}_h)^2 \right]$$

$$= \sum_h \frac{W_h^2}{4} (y_{h1}-y_{h2})^2$$

DIVIDE THE SAMPLE INTO HALF-SAMPLES:

$$\bar{y}^{(1)} = \sum_h W_h y_{h1}, \quad \bar{y}^{(2)} = \sum_h W_h y_{h2}$$

$$\bar{y} = \frac{1}{2}\left(\bar{y}^{(1)} + \bar{y}^{(2)}\right)$$

(NEARLY) INDEP ⇒

$$\tilde{v}(\bar{y}) = \frac{1}{4}\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right)^2$$

MORE EFFICIENT:

- TAKE ALL POSSIBLE ½-SAMPLES $\to 2^H$

$$\Rightarrow \frac{1}{2^H} \sum_v (\bar{y}^{(v)} - \bar{y})^2 \equiv v(\bar{y})$$

$$\delta_h^v = \pm 1, \quad \Delta y_h = \frac{1}{2}(y_{h1} - y_{h2})$$

$$\bar{y}^{(v)} = \bar{y} + \sum_h w_h \delta_h^v \Delta y_h$$

$$\bar{y}^{(v)} - \bar{y} = -(\bar{y}^{(-v)} - \bar{y}) \quad \text{, WHERE } (-v) \text{ IS COMPLEM.}$$

$$(\bar{y}^{(v)} - \bar{y})^2 = \frac{1}{4} \sum_h w_h^2 (y_{h1} - y_{h2})^2 + 2 \sum_{h h'} \delta_h^{(v)} \delta_{h'}^{(v)} w_h w_{h'} \Delta y_h \Delta y_{h'}$$

$\sum_v \Rightarrow$ THE SECOND TERM GOES AWAY!

- TAKE A BALANCED SUBSET: $\sum_v \sum_{h h'} \delta_h^{(v)} \delta_{h'}^{(v)} = 0$

$$\Rightarrow v_{BRR}(\bar{y}) = v(\bar{y}) \quad !$$

JACKKNIFE: G GROUPS

$$t^{(v)} = \frac{G}{G-1} \sum \frac{y_j}{\pi_j} \quad ; \quad t[y] = \frac{1}{G} \sum_{v=1}^{G} t^{(v)}$$

$$t^v = \sum_{j \in \text{GROUP } v} \frac{y_j}{\pi_j} \Rightarrow$$

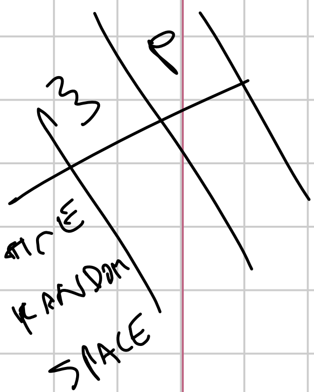
$$G t - (G-1) t^{(v)} = t^v$$

$$-(G-1)[t^{(v)} - t] = t^v - t$$

$$v(t) = \frac{1}{G(G-1)} \sum_v (t_v - t)^2 = \frac{G-1}{G} \sum_v (t^{(v)} - t)^2$$

Note Title 4/1/2007

SVY : DESIGN vs MODEL-BASED



MAIN REF: BINDER & ROBERTS (2003)
SEE ALSO: THOMPSON (1997, CH. 5); HANSON, MADOW & TEPPING (1983); RUBIN-BLEUER & SCHIOPU-KRATINA (2005)

PRO-MODEL ARGUMENT:
IF THE DATA $\sim f(x, \theta)$, AND THE MODEL IS TRUE, THEN THE MODEL-BASED ESTIMATES AND MODEL-BASED VARIANCES ARE OPTIMAL!

STOCHASTIC COMPT:
RESOLVED AT THE TIME OF MEASUREMENT, SPACE = $\mathbb{R}^n$, RANDOM VARS $= (y_j, x_j)$

PRO-DESIGN ARGUMENT:
WE HAVE TO HYPOTHESIZE ABOUT A POTENTIAL ∞ POPULATION TO JUSTIFY MODEL-BASED ARGUMENT, WHILE IN THE DESIGN-BASED APPROACH, THE FINITE POP~ PARAMETERS ARE WELL-DEFINED, AND THERE IS A FINITE POP~ SENSE OF CONSISTENCY.

STOCHASTIC COMPT:
RANDOM~ IN SAMPLING RESOLVED AT TIME OF TAKING THE SAMPLE; THE VALUES OF VARS OF INTEREST ARE CONSIDERED TO BE FIXED; SPACE = $\{\text{SAMPLES}\}, p(s)$
RANDOM VARS $= I_{js}$

SIMPLEST CASE:
ESTIMATING THE MEAN

MODEL:

$Y_1, \dots, Y_n$ ARE SAID TO BE i.i.d.

$\mathbb{E} Y = \beta, \; \mathbb{V} Y = \sigma^2$

$\Rightarrow$ MODEL-BASED ESTIMATOR IS $\bar{}$

$\mathbb{E}_\xi \bar{Y} = \beta; \; \mathbb{V}_\xi \bar{Y} \; \sigma^2/n$

DESIGN IS ASSUMED TO BE IGNORABLE:

$p(s)$ DOES NOT DEPEND ON Y_i's

DESIGN:

$Y_{j1} \dots Y_{jn} \leftarrow$ POP^n $\mathcal{U} = \{Y_1, \dots, Y_N\}$

POP^n QUANTITIES: $b = \bar{y}_{\mathcal{U}} = \frac{1}{N} \sum_{j=1}^{N} y_j$

SAMPLE ESTIMATES: SRS $\Rightarrow$ $\hat{b} = \frac{1}{n} \sum_{j \in s} y_j$

$\mathbb{E}_p \hat{b} = b; \; \mathbb{V}_p \hat{b} = \frac{1}{n} (1-f) S^2$

NON-SRS $\Rightarrow$ $\hat{b} = \hat{b}_{HT} = \frac{1}{N} \sum_{j \in s} y_j / \pi_j$

REVIEW OF o, O, o_p, O_p

COMPARISON STRATEGY:

- TAKE {DESIGN, MODEL}-BASED ESTIMATOR
- ESTABLISH {MODEL, DESIGN} PROPERTIES
- GOOD ESTIMATORS: THOSE THAT SATISFY THE OTHER STRATEGY CONSISTENCY (OR EVEN EFFICIENCY)
- FRAMEWORK: FINITE POP^n ASYMPTOTICS $N \to \infty, \; n \to \infty$: CLT IS APPLICABLE
- DISTINGUISH $o_p(n^q)$ UNDER DESIGN PROBTIES; $o_\xi(n^q)$ UNDER MODEL PROBTIES, $o(n^q)$ IN ASYMPT. FRAMEWORK

PARAMETERS OF INTEREST:

- $\mu_j = \mathbb{E}_\xi Y_j$ @ INDIV LEVEL
- $\beta = \sum_{j=1}^{N} I_{js} \mu_j / n$ @ SAMPLE LEVEL

(MOST OFTEN, $\mu_j = \mu \ \forall j$, ALTHOUGH NOT NECESSARILY)

- $\bar{\mu} = \sum_{j=1}^{N} \mu_j / N$ @ POP~ LEVEL
- RELATION BTW THE TWO: UNDER H_0,

$$\beta = \bar{\mu} + o(1) \quad (o_p(1)\ ?)$$

- $b \equiv \bar{Y}_U = \frac{1}{N} \sum_{j=1}^{N} Y_j$ @ POP~ LEVEL

TARGETS OF INFERENCE:

$\beta \longleftarrow$ MODEL-BASED

$b \longleftarrow$ DESIGN-BASED

MODEL-BASED ESTIMATOR:

$$\hat{\beta} = \sum_{j \in s} I_{js} c_{js} Y_j$$

c_{js} ARE CHOSEN S.T.:

$$\hat{\beta} = \beta + o_\xi(1) \implies \hat{\beta} = \bar{\mu} + o_\xi(1)$$

DESIGN-BASED ESTIMATOR:

$$\hat{b} = \sum_{j \in s} I_{js} d_{js} Y_j$$

d_{js} ARE CHOSEN S.T.:

$$\forall j, \ \mathbb{E}_p (I_{js} d_{js}) = \frac{1}{N} + o(N^{-1}),$$

$$\hat{b} = b + o_p(1)$$

ESTIM OF b

$$\mathbb{E}_p(\hat{b}) = \sum_{j=1}^{N} y_j \, \mathbb{E}_p(I_{js} d_{js}) = \frac{1}{N}\sum_{j=1}^{N} y_j + o_p(1) = b + o_p(1)$$

$$\mathbb{E}_p(\hat{\beta}) = \sum_{j=1}^{N} y_j \, \mathbb{E}_p(I_{js} c_{js}) \overset{?}{=} b + o(1)$$

THE MODEL IS TRUE

$$\Rightarrow \mathbb{E}_\xi b = \frac{1}{N}\sum \mathbb{E}_\xi y_j = \bar{\mu} \, ; \quad b = \bar{\mu} + o_\xi(1)$$

$$\mathbb{E}_\xi(b - \hat{\beta}) = \mathbb{E}_\xi\left[\bar{\mu} + o(1) - (\beta + o_\xi(1))\right] =$$

$$= \mathbb{E}_\xi\left[(\bar{\mu} - \beta) + o(1)\right] = o(1)$$

BOTH $\hat{\beta}$ AND b DIFFER FROM THEIR $\mathbb{E}_\xi$ BY $o(1) \Rightarrow$

$$\begin{aligned}
\mathbb{E}_p(\hat{\beta}) - b &= \mathbb{E}_\xi\left(\mathbb{E}_p(\hat{\beta}) - b\right) + o(1) \\
&= \mathbb{E}_\xi\left(\mathbb{E}_p(\hat{\beta} - b)\right) + o(1) \\
&= \mathbb{E}_p\left(\mathbb{E}_\xi(\hat{\beta} - b)\right) + o(1) \\
&= \mathbb{E}_p\left(o(1)\right) + o(1) = o(1)
\end{aligned}$$

$$\mathbb{E}_\xi \hat{b} = \sum_{j \in s} I_{js} c_{js} \mu_j = \bar{\mu} + o_p(1)$$

$$\mathbb{E}_\xi b = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E} y_j = \bar{\mu}$$

$$\mathbb{E}_p \mathbb{E}_\xi \hat{\beta} = \mathbb{E}_p \mathbb{E}_\xi \hat{b} = \bar{\mu} + o(1)$$

DESIGN-BASED & TOTAL VARIANCES

MODEL $\to$ POP $\sim$ $\to$ SAMPLE

$O_\xi(N^{-1/2})$ $\quad$ $O_p(n^{-1/2})$

$n = o(N)$

$\mathbb{V}_{p\xi}(\hat{b}) = ?$

$$\mathbb{V}_p(\hat{b}) = \mathbb{V}_p\Big(\sum_{j\in s} I_{js}\, d_{js}\, y_j\Big) = \sum_j \sum_{j'} \Delta_{jj'} y_j y_{j'}$$

$$\Delta_{jj'} = \mathrm{Cov}_p(I_{js} d_{js}, I_{j's} d_{j's})$$

$$\mathbb{V}_{\xi p}(\hat{b}) = \mathbb{V}_\xi \mathbb{E}_p(\hat{b}) + \mathbb{E}_\xi \mathbb{V}_p(\hat{b}) =$$

$$= \mathbb{V}_\xi\big[b + O_\xi(N^{-1/2})\big] + \mathbb{E}_\xi \sum_{jj'} \Delta_{jj'} y_j y_{j'} + o(n^{-1})$$

DUE TO FPC

$$= \mathbb{V}_\xi\Big[\frac{1}{N}\sum_j y_j\Big] + \sum_{jj'} \Delta_{jj'} \mathbb{E}_\xi y_j y_{j'} + o(n^{-1}) + O_\xi(N^{-1/2})?$$

$$= \frac{1}{N^2}\sum_{jj'} \sigma_{jj'} + \sum_{jj'} \Delta_{jj'}(\sigma_{jj'} + \mu_j \mu_{j'}) + o(n^{-1})$$

$$\sigma_{jj'} = \mathrm{Cov}_\xi(y_j, y_{j'}) ; \quad \frac{1}{N^2}\sum_{jj'} \sigma_{jj'} = \bar{\sigma}_{..}$$

i.i.d. (AND SOME OTHER WELL MIXING CASES)

$$\Rightarrow \bar{\sigma}_{..} = O(N^{-1}) = \mathbb{V}_\xi \mathbb{E}_p \hat{b}$$

$$\Rightarrow \mathbb{V}_{\xi p}(\hat{b}) = \sum_{jj'} \Delta_{jj'}(\sigma_{jj'} + \mu_j \mu_{j'}) + O(N^{-1}) + o(n^{-1})$$

$$\simeq \mathbb{E}_\xi \mathbb{V}_p[\hat{b}]$$

$$v(\hat{b}) / \mathbb{V}_p(\hat{b}) = 1 + o(1) \Rightarrow v_p(\hat{b}) / \mathbb{V}_{\xi p}(\hat{b}) = 1 + o(1)$$

DESIGN-CONSISTENCY $\Rightarrow$ TOTAL CONSISTENCY

$$MSE_{p\xi}(\hat{\beta}) = \mathbb{E}_\xi \mathbb{E}_p (\hat{\beta} - b)^2 =$$

$$= \mathbb{E}_\xi \{ BIAS_p^2(\hat{\beta}) + V_p(\hat{\beta}) \}^2$$

$$= O(N^{-1}) + O(n^{-1})$$

UNDER MODEL USUALLY USUALLY

SPECIFIC EXAMPLE:

STRATIFIED SAMPLING : 2 STRATA

$$\mathbb{E}_\xi y_j = \mu$$

$$\hat{\beta} = \bar{y} = \frac{1}{n}\sum_{j \in s} y_j \; ; \; \hat{b} = W_1 \bar{y}_{1s} + W_2 \bar{y}_{2s}$$

$$\mathbb{E}_p \hat{\beta} = (n_1 \bar{y}_{1U} + n_2 \bar{y}_{2U})/n \neq \mathbb{E}_p \hat{b} = \frac{N_1 \bar{y}_{1U} + N_2 \bar{y}_{2U}}{N} = b$$

DESIGN-INCONSISTENCY OF $\hat{\beta}$!

$$BIAS_p(\hat{\beta}) = \mathbb{E}_p \hat{\beta} - b = \underbrace{\left(\frac{n_1}{n} - \frac{N_1}{N}\right)}_{\alpha}(\bar{y}_{1U} - \bar{y}_{2U}) = \alpha(\bar{y}_{1U} - \bar{y}_{2U})$$

MODEL IS TRUE $\Rightarrow$

$$\mathbb{E}_\xi \bar{y}_{1U} = \mathbb{E}_\xi \bar{y}_{2U} = \mu \; ; \; \mathbb{E}_\xi \mathbb{E}_p \hat{\beta} = \mu \; ; \; \mathbb{E}_p(\hat{\beta}) = \mu + o(1)$$

ALSO, $\mathbb{E}_\xi b = \mu$, $b = \mu + o(1)$

$\Rightarrow \mathbb{E}_p \hat{\beta} = b + o(1)$ — ASYMPTOTIC DESIGN-UNBIASEDNESS

NOW: IF WE USE $\hat{\beta}$ TO ESTIMATE b,

$$MSE_p(\hat{\beta}) = \mathbb{E}_p(\hat{\beta} - b)^2 =$$

$$= BIAS_p^2(\hat{\beta}) + V_p(\hat{\beta})$$

$$= \alpha^2(\bar{y}_{1U} - \bar{y}_{2U})^2 + (n_1 S_{1U}^2 + n_2 S_{2U}^2)/n^2 + o(n^{-1})$$

$$\mathrm{MSE}_p(\hat{\beta}) \quad \text{vs} \quad \mathbb{V}_p(\hat{b}) \quad ?$$

$$\alpha^2(\bar{y}_{1u}-\bar{y}_{2u})^2 + \frac{n_1 S_{1u}^2 + n_2 S_{2u}^2}{n^2} \quad \text{vs} \quad \frac{N_1^2 S_{1u}^2}{N^2 n_1} + \frac{N_2^2 S_{2u}^2}{N^2 n_2}$$

TAKE $\mathbb{E}_\xi$ OF BOTH SIDES:

$$\mathbb{E}_\xi \mathrm{MSE}_p(\hat{\beta}) = \alpha^2\sigma^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right) + \frac{\sigma^2}{n} = \frac{\sigma^2}{n} + O(N^{-1}) = \frac{\sigma^2}{n} + o(n^{-1})$$

VS

$$\mathbb{E}_\xi \mathbb{V}_p(\hat{b}) = \sigma^2\left(\frac{N_1^2}{N^2 n_1} + \frac{N_2^2}{N^2 n_2}\right) \geq \frac{\sigma^2}{n}$$

WITH STRICT INEQUALITY WHEN $n_1/N_1 \neq n_2/N_2$

NON-LINEAR STATISTICS

MODEL-BASED ANALYSIS:

$$\theta = g(\mu_1, \dots, \mu_m), \quad \hat{\theta} = g(\bar{y}_{1s}, \dots, \bar{y}_{ms})$$

$$\Rightarrow \hat{\theta} - \theta = \sum_{k=1}^m \frac{\partial g}{\partial \mu_k} \underbrace{(\bar{y}_{ks} - \mu_k)}_{O_\xi(n^{-1/2})} + o_\xi(n^{-1/2}) \quad \leftarrow \text{TAYLOR SERIES MODEL EXP}^{n}$$

$$= \sum_{j \in s} u_j / n + o_\xi(n^{-1/2}) = \sum_{j=1}^N I_{js} u_j / n + o_\xi(n^{-1/2})$$

WHERE $u_j = u_j(\mu) = \nabla g^T \cdot (y_{js} - \mu)$, $\mathbb{E}_\xi u_j = 0$

$$\mathbb{E}_p[g(\bar{y})] = g(\mu) + \sum_{j=1}^N \pi_j u_j(\mu)/n + o_\xi(n^{-1/2})$$

$$\mathbb{E}_{\xi p}[g(\bar{y})] = g(\mu) + o(n^{-1/2})$$

DESIGN-BASED ANALYSIS

POP. VALUE $\bar{Y}_{1U}, \dots, \bar{Y}_{mU}$

SAMPLE ESTIMATOR $\bar{y}_{1d} \dots \bar{y}_{md}$

$$g(\bar{y}_d) - g(\bar{Y}_U) = \sum_{j=1}^{N} I_{js} d_{js} u_j(\bar{Y}_U) + o_p(n^{-1/2})$$

$$\mathbb{E}_p [g(\bar{y}_d)] = g(\bar{Y}_U) + \sum \left[\frac{1}{N} + o(N^{-1}) \right] u_j (\bar{Y}_U) + o(n^{-1/2})$$
$$= g(\bar{Y}_U) + o(n^{-1/2})$$

$$\mathbb{E}_{\xi p} [g(\bar{y}_d)] = g(\mu) + o(n^{-1/2})$$

$$\mathbb{V}_p [g(\bar{y}_d)] = \sum_{jj'} \Delta_{jj'} u_j(\bar{Y}_U) u_{j'}(\bar{Y}_U) + o(n^{-1})$$

$$\approx \mathbb{V}_{\xi p} [g(\bar{y}_d)]$$

MODEL-ONLY PROPERTIES

USUALLY, ξ-PROPERTIES OF $\hat{\beta}$ ARE WELL-KNOWN

E.G. MLE $\hat{\beta}$ ⇒ ASYMPT. NORMALITY, ASYMPT. EFFICIENCY

$\mathbb{E}_\xi \hat{b} = ?$

$$\mathbb{E}_\xi \hat{b} = \mathbb{E}_\xi [\sum I_{js} d_{js} Y_j] = \sum I_{js} d_{js} \mu_j$$

$$= \mathbb{E}_p [\mathbb{E}_\xi (\hat{b})] + o(1) \quad \leftarrow LLN_p$$
$$= \mathbb{E}_\xi [\mathbb{E}_p (\hat{b})] + o(1)$$
$$= \mathbb{E}_\xi [\bar{Y}_U] + o(1) = \sum \mu_j / N + o(1)$$

$\mathbb{E}_\xi \hat{b}$ & $(*)$ ⇒ $\mathbb{E}_\xi \hat{b} = \beta + o(1)$ -

ASYMPT. MODEL-UNBIASED!

$\mathbb{V}_\xi \hat{b} = ?$

$= \mathbb{V}_\xi \left(\sum_j I_{js} d_{js} Y_j \right) = \sum_{jj'} I_{js} I_{j's} d_{js} d_{j's} \sigma_{jj'}$

$= \mathbb{E}_p \left(\qquad \right) + o(n^{-1}) \qquad \Leftarrow LLN_p$

$= \sum_{jj'} \Delta_{jj'} \sigma_{jj'} + o(n^{-1})$

$= \mathbb{V}_{\xi p} \hat{b} + o(n^{-1})$ PROVIDED $\sum \mu_j \mu_{j'} \Delta_{jj'} = o(n^{-1})$

$= \mathbb{E}_\xi \underbrace{\mathbb{V}_p \hat{b}}_{} + o(n^{-1})$

THE LEADING TERM OF $\mathbb{V}_{\xi p} \hat{b}$

$\sum \mu_j \mu_{j'} \Delta_{jj'} = O(n^{-1}) \Leftarrow$

- $\mu_j = \mu \;\; \forall j$, $\;\; \mathbb{P}_p[n(s) = n] = 1$ (FIXED SIZE)

  $\Rightarrow \sum \mu_j \mu_{j'} \Delta_{jj'} = 0$
- $\mu_j = 0 \;\; \forall j \Rightarrow \sum \mu_j \mu_{j'} \Delta_{jj'} = 0$

  ESTIMATING EQUATIONS!

MODEL VIOLATION

IS $\hat{\beta} \xrightarrow{\xi} \beta$?

IF IT IS, IS $\sigma_\xi(\hat{\beta}) / \mathbb{V}_\xi[\hat{\beta}] \to 1$?

CRITICAL ASSUMPTIONS:

- SAMPLE DESIGN IS IGNORABLE
- STRUCTURE FOR THE MODEL MEANS/COVARS

MODEL VIOLATIONS:

- INFO SAMPLING: $\mathbb{E}[Y_j \mid I_{js} = 1] \neq \mu_j = \mathbb{E}_\xi[Y_j]$
- WRONG STRUCTURE OF μ_j OR $\sigma_{jj'}$

SAMPLING IS NOT IGNORABLE

$\Rightarrow$ WHAT IS A REASONABLE TARGET FOR MODEL INFERENCE?

LET'S STILL CONSIDER / ASSUME

$$\beta = \mathbb{E}_\xi \bar{y}_U + o(1) = \sum_{j=1}^{N} \mu_j / N + o(1) \quad (*)$$

SEEN EARLIER:

1) $\mathbb{E}_\xi [\hat{b}] = \beta + o(1)$

2) $v_p [\hat{b}] = V_\xi [\hat{b}] + o(1)$ UNDER CERTAIN ALTHOUGH REASONABLE CONDITIONS

- ESTIMATING EQNS APPROACH IS ESPECIALLY BENEFICIAL! THEN $v_p[\hat{b}] \approx V_\xi[\hat{b}]$ EVEN IF THE DESIGN IS NON-IGNORABLE, BUT THE MEAN STRUCTURE IS CAPTURED
- IF THE MEAN STRUCTURE IS NOT MODELLED CORRECTLY, HOWEVER, THEN $\mathbb{E}_{p\xi}\, u_j(\bar{y}_u)$ MAY NOT BE ZERO, AND THEN $V_{p\xi} [\hat{b}] > v_p[\hat{b}]$

PROPERTIES OF $\hat{\beta}$?

$$\hat{\beta} = \sum_j I_{js}\, c_{js}\, Y_j \quad \leftarrow \text{ BLUE OR SOMETHING}$$

- NON-IGNORABLE DESIGN:

$$\mathcal{U} = \{0, \dots, 0, 1, \dots, 1\} \quad ; \quad P_\xi[u_j = 1] = \mu, \ \beta = \mathbb{E}_\xi[\bar{y}_u] = \mu$$

$$\pi_j = \begin{cases} p_0, & y_j = 0 \\ p_1, & y_j = 1 \end{cases} \rightarrow \text{INFORMATIVE}$$

$$\mu_{j1} = \mathbb{E}_\xi[y_j \mid I_{js} = 1] = \mathbb{P}[y_j = 1 \mid I_{js} = 1] = \frac{\mu p_1}{(1-\mu)p_0 + \mu p_1}$$

$$\text{BIAS}[\bar{y}_S] = \mu_{j1} - \mu = \frac{\mu(1-\mu)(p_1 - p_0)}{(1-\mu)p_0 + \mu(p_1)}$$

- WRONG STRUCTURE

ξ: $\begin{matrix} Y_1 \\ Y_2 \end{matrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \frac{\sigma^2}{n} \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \right)$

$\Rightarrow$ BLUE $\hat{\mu} = 0.8 y_1 + 0.2 y_2$

$\mu_1 \neq \mu_2$, $\beta = \mathbb{E}_\xi \bar{Y}_n = \frac{\mu_1 + \mu_2}{2}$

$\Rightarrow$ BIAS $\hat{\mu} = \mathbb{E}_{p\xi}[\hat{\mu} - \beta] = 0.3(\mu_1 - \mu_2)$

	P	ξ
ESTIMATOR	$\hat{b}$	$\hat{\beta}$
CONSISTENCY FOR b?	✓	WHEN THE MODEL HOLDS
CONSISTENCY FOR β?	✓	WHEN THE MODEL HOLDS

VARIANCE ESTIM~ $\mathbb{V}_p[\hat{b}] \simeq \mathbb{V}_{p\xi}[\hat{b}]$!!

$\mathbb{V}_p$ MODEL IS TRUE $\Rightarrow$ OFTEN, $\mathbb{V}_p[\hat{b}] < \mathbb{V}_p[\hat{\beta}]$

$\mathbb{V}_\xi$ CUMBERSOME ... $(\text{INFO MX})^{-1}$

MODEL IS VIOLATED?

$\mathbb{E}_{\xi p} \hat{b} = \beta + o(1)$

$\mathbb{V}_{\xi p}[\hat{b}] \simeq \mathbb{V}_p[\hat{b}]$

UNDER MILD COND~

$\hat{\beta}$ IS BIASED FOR β

MEANS ARE CORRECT

$\Rightarrow \mathbb{V}_p[\hat{\beta}] \simeq \mathbb{V}_{p\xi}[\hat{\beta}]$

$\mathbb{V}_p$'S IN THOSE CASES ARE (ASYMPT.) EQUIVALENT TO $\mathbb{V}_\xi$ IMPLIED BY LITERATURE ON M-ESTIMATES AND MISSPECIFIED MODELS (HUBER 67,74; WHITE 82)

ARGUABLY, MSE IS THE BEST CRITERION?

STAT 9100 SVY – MISSING DATA & WEIGHTS

Note Title 4/12/2007

MISSING DATA

- MISSING BY DESIGN
  - TWO-PHASE SAMPLING:
    X IS COLLECTED ON EVERYBODY AT TIME t_1
    Y IS COLLECTED ON A SUBSAMPLE AT TIME t_2
  - ROTATION SAMPLES:
    DESCRIBE CPS
  - SHORT AND LONG FORMS

- ATTRITION IN LONGITUDINAL SURVEYS
  - ONLY A FRACTION OF THE UNITS IN WAVE t WILL BE SURVEYED IN WAVE $t+1$

- UNIT NON-RESPONSE:
  (PRACTICALLY) NO INFORMATION CAN BE OBTAINED FROM THE UNIT
  - RESPONSE RATES: AAPOR DEFINITIONS
    (INCLUDES CONTACTS, ELIGIBLE UNITS, COMPLETES, INCOMPLETES, ...)

  SHOW OF HANDS:
  WHAT DO YOU THINK A GOOD RESPONSE RATE IS?

http://www.aapor.org/best_practices_for_survey_and_public_opinion_research.asp
http://www.aapor.org/pdfs/standarddefs_4.pdf

- ITEM NON-RESPONSE:
  FOR UNIT j, SOME DATA ARE COLLECTED, WHILE OTHERS ARE MISSING

FRAMEWORK:

UNIT j

VARIABLES X_{ij}, $i = 1, \dots, p$ – OF INTEREST

$f(x, \theta)$ – MODEL OF INTEREST

$z_{ij} = \mathbb{1}\{ X_{ij} \text{ IS OBSERVED} \}$

MISSING DATA MECHANISM

$$p_{ij} = \mathbb{P}[z_{ij} = 1 \mid X_{ij}, X_{-ij}, X_{ij'}, z_{ij}, z_{ij'}, \psi]$$

MISSING COMPLETELY @ RANDOM:

$$\mathbb{P}[z_{ij} \mid \text{STUFF}] = \psi$$

MISSING @ RANDOM

$$\mathbb{P}[z_{ij} \mid \text{STUFF}] = \mathbb{P}[z_{ij} \mid X_{-ij}, X_{ij'}, z_{ij}, z_{ij'}, \psi]$$

NOT MISSING @ RANDOM / INFORMATIVELY MISSING

$$\mathbb{P}[z_{ij} \mid \text{STUFF}] = \mathbb{P}[z_{ij} \mid X_{ij}, \dots]$$

THE MAIN THEOREM OF MISSING DATA:

THE MISSING DATA MECHANISM IS IGNORABLE $\Leftarrow$ MAR & $\psi \cap \theta = \emptyset$

– FACTORIZE LIKELIHOOD!

MISSING BY DESIGN:

$z_j \propto \mathbb{1}$ + RANDOMIZATION

UNIT NON-RESPONSE:

ASSUME $\mathbb{P}[z_j \mid \text{DESIGN VARS}] = \psi$

$\Rightarrow \mathbb{P}[\text{IN SAMPLE}] = \mathbb{P}[\text{SELECTION}]\,\mathbb{P}[\text{RESPONSE}]$

$\Rightarrow$ USE WEIGHT $= (\mathbb{P}[\text{IN SAMPLE}])^{-1}$

ITEM NON-RESPONSE $\Rightarrow$???

- LIKELIHOOD METHODS
- IMPUTATION METHODS
  - HOT DECK
  - MULTIPLE IMPUTATION

WEIGHTS

KORN & GRAUBARD (1999) - CH 4
PFEFFERMANN (1993)

PURPOSES OF WEIGHTS IN STAT PROGRAMS

- FREQUENCY WEIGHTS: HOW MANY IDENTICAL UNITS HAVE BEEN OBSERVED - ONLY MAKES SENSE FOR SIMPLE CATEGORICAL DATA / CONTINGENCY TABLES
  USE: REPRODUCE CONTINGENCY TABLES, HISTOGRAMS, ... % PLOTS
- VARIANCE-ADJUSTING WEIGHTS:
  IF $\mathbb{V} y_j \propto \sigma^2 h_j$, THEN THE LKHD IS
  $$\ln \ell(y_j; \mu_j, \sigma_j^2) \sim \frac{1}{2}\sum_j \frac{(x_j - \mu_j)^2}{\sigma_j^2} \sim \sum_j \frac{(x_j - \mu_j)^2}{h_j}$$
  SO ONE CAN RUN THE ESTIMATION PROCEDURE WITH WEIGHT h_j^{-1}.
  OFTEN, THOSE WEIGHT ARISE WHEN THE DATA ARE AGGREGATED BASED ON GROUPS OF DIFFERENT SIZES:
  $$\mathbb{V}[y_j] = \sigma^2 / n_j, \text{ SO } h_j \propto n_j^{-1}$$
  WHICH MAKES IT SOMEWHAT LIKE FREQ WEIGHTS

- SAMPLING / SURVEY / PROBABILITY WEIGHTS TO CORRECT FOR DIFFERENTIAL PROBTIES OF SELECTION
  FREQ INTERPRETATION:
  EACH UNIT REPRESENTS w_j UNITS OF POPULATION

SAMPLING WEIGHTS $\leftarrow$

(1) PROBABILITY OF SELECTION, HT WAY:

$$w_j^{BASE} = 1/\pi_j$$

(2) NON-RESPONSE ADJUSTMENTS:

IF OUT OF n_j UNITS IN THE DESIGNED SAMPLE, THE DATA WERE COLLECTED ON $\tilde{n}_j$ UNITS ONLY, THEN ASSUMING MAR | CLUSTER INDICATOR,

$$\hat{\mathbb{P}}[\text{RESPONSE}] = \tilde{n}_j / n_j$$

$$w_j^{NR} = 1/\hat{\mathbb{P}}[\text{RESPONSE}] = n_j / \tilde{n}_j$$

(3) FRAME UNDERCOVERAGE/POSTSTRATIF-

IF CERTAIN DESIGN & OUTCOME VARS z_j ARE KNOWN FOR ALL POP-UNITS. SEX-AGE-RACE: AVAILABLE FROM CENSUS, BUT ARE NOT KNOWN BEFORE SAMPLING, SO CANNOT BE USED TO STRATIFY BEFOREHAND

$$\sum_{j \in S} w_j^B w_j^{NR} \underset{\substack{\uparrow \\ \text{USUALLY 0/1} \\ \text{CELL INDICATORS}}}{z_{jk}} = t_w[z_k]$$

$$\Rightarrow w_j^{PS} = \frac{T[z_k]}{t_w[z_k]} \qquad \text{IF } z_{jk} = 1$$

THEN THE RESULTING FINAL WEIGHT WOULD BE

$$w_j = w_j^{BASE} \, w_j^{NR} \, w_j^{PS}$$

THE USE OF WEIGHTS:

$u_j(\theta)$ IS ESTIMATING EQ ~ FOR θ

$$0 = \sum_{j=1}^{N} u_j(\theta_0) \quad \text{FOR POP} \sim$$

$$0 = \sum_{j \in S} w_j \, u_j(\hat{\theta}) \quad \text{FOR SAMPLE}$$

- LINEAR STATISTICS: EXPLICIT SOLUTIONS
- NL (GLM): ITERATIVE ALGORITHMS

INEFFICIENCY OF WEIGHTED ESTIMATION

$$\bar{y}_w = \frac{t_w[y]}{t_w[1]} = \frac{\sum_j w_j y_j}{\sum_j w_j}$$

$$\mathbb{V}_\xi[\bar{y}_w] = \mathbb{V}_\xi\left[\frac{\sum_j w_j y_j}{\sum_k w_k}\right] = \frac{\sum_j w_j^2 \sigma^2}{(\sum_j w_j)^2}$$

$$\text{DEFF}_w = \frac{\mathbb{V}_\xi[\bar{y}_w]}{\mathbb{V}_{SRS}[\bar{y}]} = \frac{\sum_j w_j^2 \sigma^2}{(\sum_j w_j)^2} \Big/ \frac{\sigma^2}{n} =$$

$$= \left(\frac{1}{n}\sum_j w_j^2\right) \Big/ \left(\frac{1}{n}\sum_j w_j\right)^2 = 1 + \frac{\text{"}\mathbb{V}[w]\text{"}}{\text{"}\mathbb{E}[w]^2\text{"}} =$$

$$= 1 + CV_w^2$$

COEFFICIENT OF VARIATION OF WEIGHTS

SEE TABLE 4.4.1 OF K&G/4

$$\text{INEFF}_0 = 1 - \frac{\mathbb{V}_{SRS}[\bar{y}]}{\mathbb{V}_\xi[\bar{y}_w]} = 1 - \frac{1}{\text{DEFF}}; \quad \text{DEFF} = (1 - \text{INEFF}_0)^{-1}$$

RATHER COMMON PRACTICE:
TRUNCATE THE WEIGHTS THAT ARE TOO HIGH

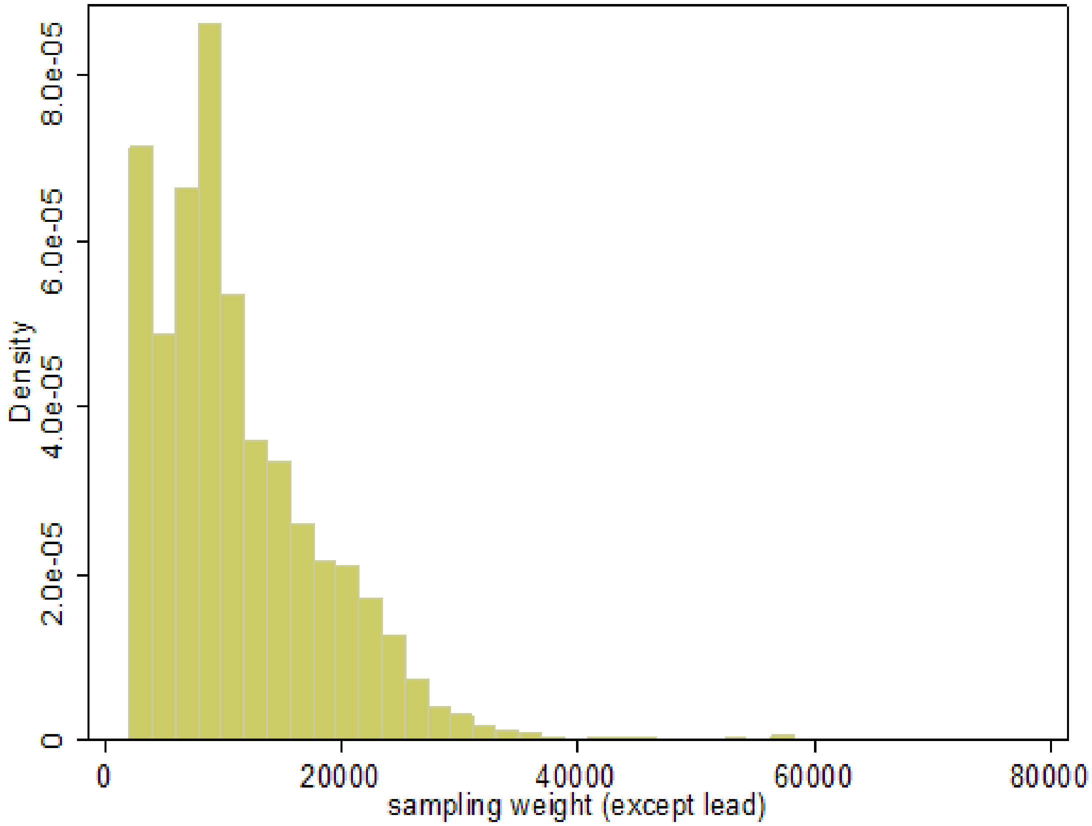
NHANES:

```
. use nhanes2

. sum finalw

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    finalwgt |     10351    11318.47     7304.04       2000      79634

. di 1+r(Var)/( r(mean)*r(mean) )
1.4164382
```



SERIOUS PFEFFERMANN (1993)

THE NEED FOR WEIGHTS ARISES WHEN THE DESIGNS ARE INFORMATIVE, AND MODEL IN THE POPULATION IS DIFFERENT FROM ONE IN THE DATA.

SIMPLE EXAMPLE:

DESIGN VARIABLE z_j : $\mathbb{P}[j \in s] \propto z_j$

VARIABLE OF INTEREST y_j :

$$\mathrm{Cov}[y_j, z_j] > 0 \Rightarrow \mathbb{P}[\,y_j > \mu_j \mid j \in s\,] > 1/2,\ \mathbb{E}_\xi \bar{y}_s > \mu_y$$

↑SAMPLE

UNBIASED ESTIMATOR: REGRESSION

$$\bar{Y}_R = \bar{Y}_s + b\,(\bar{z}_u - \bar{z}_s), \quad b = \sigma_{yz}/\sigma_z^2$$

REQUIRES THE KNOWLEDGE OF ALL VALUES z_j IN POP~ TO MAKE DESIGN IGNORABLE!

NOTATION:

DESIGN VARS : $z_j \sim g(z, \varphi)$

SAMPLE DESIGN : $p(s) = \mathbb{P}[\{I_{js}\} \mid Y, z, \psi]$

MODEL OF INTEREST: $y_j \sim f(y \mid z, \theta)$

$Y = (Y_s, Y_{-s})$ ← SAMPLED & NON-SAMPLED

C.F. MISSING DATA LITERATURE

JOINT DISTR~: $f(Y, z \mid \theta, \varphi) = f(Y \mid z, \theta)\, g(z, \varphi)$

SAMPLE DISTR~: OVER UNOBSERVED VARIABLES

$$f(Y_s, I_s, z \mid \theta, \varphi, \psi) = \int f(Y_s, Y_{-s} \mid z, \theta)\, g(z, \varphi)\, P(I \mid Y_s, Y_{-s}, z, \psi)\, dY_{-s}$$

NON-INFORMATIVE DESIGN:

$$f(Y_s, z \mid \theta, \varphi) = \int f(Y_s, Y_{-s} \mid z, \theta)\, g(z, \varphi)\, dY_{-s}$$

E.G. $\mathbb{P}[\,I_s \mid Y_s, Y_{-s}, z, \psi] = \mathbb{P}[\,I_s \mid z, \psi]$

WHEN z'S ARE KNOWN FOR ALL POP~ UNITS!

EXTENSIONS: IGNORABILITY FOR REGRESSION

$$f(Y_s \mid X_s, z, I) = f(Y_s \mid X_s, z)$$

IF FURTHER $= f(Y_s \mid X_s) \rightarrow$ OLS!

IGNORING INFORMATIVE DESIGNS:

- BIAS OF POINT ESTIMATES
- POOR PERFORMANCE OF INFERENCE PROCEDURES
- NO INTERPRETATION OF $\mathbb{E}_{p\xi}\hat{\beta}_{OLS}$
- POPULATION-BASED CASE-CONTROL STUDIES:

$$Y = \begin{cases} 1, & \text{RARE} \\ 0, & \end{cases}$$

$$\mathbb{P}[I_{js}] = \begin{cases} 1, & Y_j = 1 \\ \pi \ll 1, & Y_j = 0 \end{cases}$$

MODEL: $f(y \mid x, \theta)$ – LOGISTIC REGRESSION

(PSEUDO)MLE VS WEIGHTED?

IF THE MODEL HOLDS IN THE POPULATION, $\theta_{\text{INTERCEPT}}$ IS BIASED, BUT ALL θ_x'S ARE IN FACT MLE'S!

TESTING IGNORABILITY

REGRESSION CONTEXT:

OPTIMAL / OLS / UNWEIGHTED / MLE $\hat{\beta}$
VS.
WEIGHTED / DESIGN-CONSISTENT $\hat{\beta}_w$

H_0: DESIGN IS IGNORABLE $\Rightarrow \operatorname{plim}_{n,N\to\infty} {}_{p\xi}[\hat{\beta} - \hat{\beta}_w] = 0$

$$\lambda = \hat{D}^T [v(\hat{D})]^{-1} \hat{D}, \quad \hat{D} = \hat{\beta} - \hat{\beta}_w$$

IMPLEMENTATION:

$$\left.\begin{array}{l} Y_s = \underbrace{X_s\hat{\beta}}_{OLS} + \underbrace{W_s X_s}_{AUX}\hat{\gamma} + \text{ERROR} \\ \qquad\qquad\qquad\quad \text{REGRESSORS} \end{array}\right\} \begin{array}{l}\text{TEST} \\ H_0: \gamma = 0\end{array}$$

GENERAL HAUSMAN TEST:

$\hat{\theta}_0, \hat{\theta}_1 \xrightarrow{P} \theta$, $\hat{\theta}_0$ IS ASYMPT EFFICIENT, $\sqrt{n}(\hat{\theta}_i - \theta) \xrightarrow{d} N(0, V_i)$

$\Rightarrow \operatorname{AsCov}(\hat{\theta}_0, \hat{\theta}_1 - \hat{\theta}_0) = 0$, $\mathbb{V}(\hat{\theta}_1 - \hat{\theta}_0) = \mathbb{V}(\hat{\theta}_1) - \mathbb{V}(\hat{\theta}_0) \geq 0$,

$$(\hat{\theta}_1 - \hat{\theta}_0)^T (V_1 - V_0)^{-1} (\hat{\theta}_1 - \hat{\theta}_0) \xrightarrow{d} \chi^2_{\dim\theta}$$

SIGNIFICANT COMPTS $\Rightarrow$ IMPORTANT DESIGN VARS/INTERACTIONS!

$\mathbb{V}[\theta_0], \mathbb{V}[\theta_1] \leftarrow$ DESIGN-BASED ESTIMATORS

APPROACHES TO INCORPORATE WEIGHTS

- MODIFICATION OF MODEL-DEPENDENT ESTIMATORS: REPLACE ALL SUMS $\frac{1}{n}\sum_{i=1}^{n} \text{WHATEVER}_i$ BY $\frac{1}{N}\sum_{i=1}^{n} w_i \,\text{WHATEVER}_i$

  $$b_{OLS} = \frac{1}{n}\sum x_i y_i - \left(\frac{1}{n}\sum x_i\right)\left(\frac{1}{n}\sum y_i\right) \Big/ \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)^2$$

  AMBIGUITY: MORE THAN A SINGLE DESIGN-CONSISTENT ESTIMATOR MIGHT BE AVAILABLE (E.G. t_{HT}, t_{RATIO}, t_{REG}, ...)

- MODELS THAT YIELD DESIGN-CONSISTENT ESTIMATORS

  E.G. FIXED STRATUM EFFECT MODELS

  $$Y_{hi} \mid \mu_h, \sigma_h \sim N(\mu_h, \sigma_h^2), \; P(\mu_h, \ln \sigma_h) \propto 1$$

  RANDOM EFFECT MODELS

  $$\mu_h \mid \mu\, \delta^2 \sim N(\mu, \delta^2), \quad P(\mu, \ln \sigma_h, \ln \delta) \sim 1$$

- PSEUDO-LKHD: MODIFY LKHD SCORE EQNS

  $$u(\theta) = \partial \ln f(y, \theta) \Rightarrow$$

  $$\sum_{j \in S} w_j \, u_j(\theta) = 0$$

- ESTIMATING FUNCTIONS

  UNBIASED: $\mathbb{E}_\xi \, g(y, \theta) = 0$

  OPTIMAL: $g^* = \arg\min_g \mathbb{E}_\xi (g^2) \Big/ \left[\mathbb{E}_\xi \frac{dg}{d\theta}\Big|_{\text{TRUE } \theta}\right]^2$

  SAMPLE: DESIGN-UNBIASED

  $$\mathbb{E}_p \, h(y_s, \theta) = g^*(y, \theta) \quad \forall \text{ POPN } y, \theta$$

  OPTIMAL: $h^* = \arg\min \mathbb{E}_{p\xi} h^2(y_s, \theta) \Big/ \left[\mathbb{E}_{p\xi} \frac{dh}{d\theta}\right]^2$

  h^* TURNS OUT TO BE H-T!

- WEIGHTS AS SURROGATE SUMMARIES OF DESIGN VARIABLES

  $a = (a_1, \ldots, a_N)$ IS AN ADEQUATE SUMMARY OF z := $\mathbb{P}(I_S \mid z) = \mathbb{P}(I_S \mid a)$

  $\Rightarrow (\pi_j)_{j=1}^{N}$ IS THE MINIMAL POSSIBLE SUMMARY

  (TREATMENT EFFECT LITERATURE: PROPENSITY SCORE)

  DON'T HAVE TO WORK OUT THAT WAY: (π_j)'s MIGHT BE TOO COARSE!

- MLE FROM WEIGHTED DISTR~:

  MODEL $P(y_i, \alpha) = P(I_{jS} \mid Y_j)$ —

  MODIFY DISTR~ OF Y TO ACCOUNT FOR PROBTY OF OBSERVING Y

  $$f(y;\lambda) = \frac{p(y_j;\alpha)\, f(y_j;\theta)}{\mathbb{P}[I_{jS}=1]} =: f(y_j \mid I_{jS}=1)$$

  $$\mathbb{P}[I_{jS}=1] = \int p(y_j, \alpha)\, f(y_j;\theta)\, dy_j$$

  $$\Rightarrow \text{LKHD}(\theta) \propto \prod f(y_j;\theta) \Big/ \left\{\int P(y_j;\alpha) f(y_j;\theta)\, dy\right\}^n$$

  PROPORTIONALITY INCLUDES α

  EMPIRICAL WAY: ESTIMATE α, PLUG IN

  JOINT WAY: max OVER α, θ

- EMPIRICAL LKHD:

  $$\max \prod p_j \;\Big|\; \sum p_j g(x_j) = 0$$

  WHERE $\sum_{j=1}^{N} g(x_i) = 0$ IN POP~

  $$\theta = \int u(x)\, dF_N(x)$$

  $$\Rightarrow \hat{\theta} = \int u(x)\, dF_W(x), \quad F_W(x) = \sum p_j \delta(x_j)$$

  - CORRECTS FOR KNOWN POP~ QUANTITIES

IMPUTATION:

A PROCEDURE OF COMING UP WITH A VALUE FOR MISSING DATA

- SINGLE IMPUTED VALUE:
  $\{y^*_{ij}$, IMPUTATION INDICATOR $z_{ij}\}$
  USUALLY, UNBIASED ESTIM~ IS FEASIBLE, BUT VARIANCE ESTIM~ IS PROBLEMATIC
- MULTIPLY IMPUTED VALUE:
  SEVERAL PLAUSIBLE VALUES Y^*_{ij} — EXTRA VARIABILITY PROVIDED AS NEEDED

IDEALLY Y^*_{ij} COMES FROM PREDICTIVE DISTR~

$$f(Y_{ij} \mid \underset{\text{OTHER VARS}}{Y_{(-i)j}}, \underset{\text{OTHER OBS}}{Y_{i(-j)}}, \underset{\text{DESIGN}}{z_j} ; \underset{\text{MODEL}}{\theta}, \underset{\text{MISSING DATA}}{\alpha}, \underset{\text{DESIGN}}{\Psi}\}$$

- MEAN IMPUT~: $y^*_{ij} = \bar{Y}_j$ (M.B. OVER IMPUT~ CELL)

- REG~ IMPUT~: $y^*_{ij} = X^T_{(-i)j}\gamma$, $\gamma \leftarrow$ (D.C.) REGRESSION $Y \to X_{(-i)}$, THE VARS AVAILABLE FOR jTH CASE

- STOCH/ REGR~ IMPUT~:
  $$y^*_{ij} = X_j^T\gamma + \varepsilon^*_{ij}, \quad \varepsilon^*_{ij} \sim N(0, s^2) \text{ OR EMPIRICAL DISTR~ OF } e_{ij}$$
  OR $$y^*_{ij} = \begin{cases} 0 \\ 1 \end{cases}, \quad \mathbb{P}[1] = \Lambda(x^T_{(-i)j}\gamma)$$

- HOT DECK IMPUT~:
  $$Y^*_{ij} = Y_{i(-j)}$$, M.B. OVER IMPUT~ CELL (OFTEN STRATA OR PSU)
  POPULAR WITH DATA PROVIDERS, BUT PROPERTIES ARE NOT SO WELL KNOWN

- COLD DECK IMPUT~:

  y^*_{ij} ← VALUE FROM EXTERNAL SOURCE (REGISTER, CENSUS, LAST YEAR VALUE, ...)

RANDOM COMPONENTS:

ξ ← MODEL
p ← DESIGN
$*$ ← IMPUT~

$\}$ $p*$ IF AVAILABLE?

VARIANCE ESTIM~ WITH IMPUTED DATA?

⚠ NON-RESPONSE BIAS ⇒ $MSE[\hat{\theta}] > V[\hat{\theta}]$?

- EXPLICIT FORMULAE?

$$v(t_{HT}(Y_{obs})) = \sum_{i=1}^{\ell} \frac{(\ell t_i/\pi_i - t)^2}{\ell(\ell-1)}$$

⇐ (i) t_i ARE UNBIASED FOR CLUSTER TOTALS
(ii) t_i's ARE $\perp\!\!\!\perp$ ⇐ ADJUSTMENT/IMPUT~ CELLS DO NOT CUT THROUGH CLUSTERS

- RESAMPLING:

BRR/BSTRAP/JKNIFE → IMPUTE FOR THE $^{(r)}$ DATASET → ESTIMATE → RESULTING VARIABILITY SUFFICES!

• IMPUT~ MUST PROVIDE CONSISTENT $\hat{\theta}^{*(r)}$!

* SOMETIMES MODIF~ MIGHT BE REQUIRED - E.G. NO RESPONDENTS IN A PARTICULAR BSTRAP SUBSAMPLE!

- MULTIPLE IMPUT~

d-th IMPUTED DATA SET $\to \hat{\theta}_d, v_d \; ; d=1,\dots,D$

↳ COMBINE

- POINT ESTIMATE $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$

- WITHIN-IMPUT~: $\bar{v}_D = \frac{1}{D}\sum_{d=1}^{D} v_d$

- BETWEEN-IMPUT~: $B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta}_D)(\hat{\theta}_d - \bar{\theta}_D)^T$

- TOTAL VARIABILITY:

$$T_D = \bar{v}_D + \frac{D+1}{D} B_D$$

- NON-RESPONSE INFO:

$$\gamma_D = (1+1/D)\, B_D T_D^{-1}$$

- INFERENCE:

$$(\theta - \bar{\theta}_D)\, T_D^{-1/2} \sim t_\nu$$

$$\nu = (D-1)\left(1 + \frac{1}{D+1}\bar{W}_D B_D^{-1}\right)^2$$

↑ SATTERTHWAITE APPROX~

OR

$$\nu^* = (\nu^{-1} + \hat{\nu}_{obs}^{-1})^{-1}$$

$$\hat{\nu}_{obs} = (1-\hat{\gamma}_D)\left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}$$

ν_{com} = DESIGN D.F.

IMPUTATIONS = ?

- TYPICAL ADVICE: 3÷5
- EMPIRICAL EVIDENCE ON STABILITY?
- DESIGN DEGREES OF FREEDOM?

STAT 9100 SVY : DIFFICULT SITUATIONS

Note Title 4/25/2007

http://www.stata.com/support/faqs/st

- ZERO SURVEY WEIGHTS?
- SUBDOMAIN ESTIMATION
  - ↳ SET WEIGHTS OF IRRELEVANT UNITS TO ZERO
  - ↳ D.F. = #PSUS WITH DOMAIN OBS − #STRATA WITH DOMAIN
- SINGLETON PSUS / STRATA (E.G. SELF-REPRESENTING, $\pi_j = 1$)
  - ↳ COLLAPSE STRATA : INCREASE D.F., OVERACCOUNT B/W STRATA
  - ↳ USE SELF-REPRESENTING PSUS AS PSEUDO-STRATA ← SSU
- LACK OF D.F.S (KORN & GRAUBARD 1995)
- PANEL DATA
  - OTHER EXAMPLES OF MISMATCH BETWEEN THE SURVEY POP~ & THE ANALYSIS POP~: BINDER & ROBERTS (2006)
- MATCHING & NBHOOD STRUCTURE − NEED π_{ij} ?
- SURVEYS OVER TIME

  GOOD TIMES TO SURVEY? OPTIMIZE OVERLAP? SEASONALITY?
  TIME IN SAMPLE EFFECT? ROTATING SURVEYS
- NON-SMOOTH ESTIMATORS (QUANTILES)
- VARIANCE ESTIM~ WITH IMPUTED DATA
- SURVEY BOOTSTRAP

  SCALING, NON-RESPONSE, IMPUT~, POST-STRAT~ ADJ
- OUTLIERS & INFLUENTIAL OBSERVATIONS (TRADITIONAL SENSE + BIG WEIGHTS)
- MULTIVARIATE METHODS
- MULTILEVEL MODELS
- NON-NORMAL ASYMPTOTICS ?
  - χ^2 GOODNESS-OF-FIT → RAO-SCOTT MIXTURES
- NETWORK SAMPLING − EXTREMELY INFO!