Analysis of Complex Health Survey Data

Stas Kolenikov
@StatStas@mastodon.online

NORC

Views and opinions are those of the instructor only,
not those of NORC or ASA

ICHPS 2023

https://skolenik.github.io/ICHPS2023-svy/

NORC

Target audience and objectives

This short course is aimed at the audience with general knowledge of statistics at graduate level. No knowledge of survey statistics is assumed

- but a 2 hour course won't make you an expert in it, either!

The course will cover:

- Role of survey data in today's multi-faceted world of data collection
- Basic features of survey data
- Basic principles of survey inference
- The most important peculiarities of analyzing survey data

By the end of the course, you will be better prepared to read technical reports and papers describing complex health survey data. You will be better equipped to understand the syntax of the complex health survey data analysis modules in R and Stata (self-study).

NORC

What informs health policy?

If we want to improve health outcomes of constituents of health policy, how are the differences in health outcomes related to the differences in what may potentially affect health?

- Biological pathways and experiments (pharma RCT)
- Biological pathways and observational studies
  - leading to universally applicable health policies
  - e.g., smoking and tobacco
- Socioeconomic determinants and observational studies (survey data)
  - leading to allocation of limited resources across population groups
  - e.g., targeted cancer screening or immunization campaigns

NORC

What is a complex survey?

If the methodology description of your data set contains keywords or phrases like

- stratified sampling
- multistage sampling
- random digit dialing
- nonresponse adjustment
- probability weights
- replicate weights

. . . it is a complex survey data set, and you are in the right workshop to learn more about them!

Korn & Graubard (1999), Heeringa et al. (2017), Lumley (2010), Chambers & Skinner (2003), Pfeffermann & Rao (2009), Wu & Thompson (2020); Skinner & Wakefield (2017); West et al. (2016).

NORC

Finite population surveys

- Developed initially (1920–1930s) to study finite populations with fixed characteristics
  - volume of a tree
  - last year revenue of a business
  - total grain harvest by a farm
  - person's sex
- Random variables are 0/1 indicators of being in the sample vs. not
- Inference/statistical distributions are with respect to the sampling mechanism
- 1970s and on: statistical models are incorporated and blended with sampling inference

NORC

Sample surveys vs. anything else

- Clinical trials
  - assume biology is the same, hence sample selection is not an issue
  - inference is wrt randomized assignment
- Administrative records
  - by-products of systems built by somebody else
  - created for reasons other than research (e.g., billing)
  - concepts/definitions/categories may not match the research questions
  - cannot uncover how people adapt to the admin structures
- Big Data
  - terrific at answering the "what is happening" question
  - sweeps rationale/motives for human behavior under the carpet — awful at "why" question
  - ML/AI methods have $\approx$ zero transparency
  - "$N =$ all" is a misleading slogan

NORC

Dimensions to break down

- Human populations vs. establishment populations (hospitals, practices, providers)
- Sampling frames and modes
  - List of patients
  - Population register
  - Area sample of the general population : face-to-face/specimen interview
  - Random digit dialing sample of the general population : phone interview
  - Address-based sample of the general population : mail or face-to-face interview
  - Respondent driven samples : SAQ or face-to-face interview
  - Screening for target population
  - Combinations: multiple frame and/or multiple mode surveys
- Probability panels

NORC

Health surveys in the U.S.

- National Health Interview Survey (NHIS): area sample, face-to-face; CDC/NCHS
- National Health And Nutrition Examination Survey (NHANES): area sample, face-to-face, biological specimen; CDC/NCHS
- Behavioral Risk Factor Surveillance Survey (BRFSS): phone survey; CDC, state health agencies
- National Survey of Drug Use and Health (NSDUH): area sample, face-to-face; SAMHSA
- Consumer Assessment of Healthcare Providers and Systems (CAHPS): list sample of patients, multimode; AHRQ+CMS
- Medical Expenditure Panel Survey (MEPS): three-fold survey of households, their insurers, and their medical providers; AHRQ
- National Inpatient Sample of the Healthcare Cost and Utilization Project (HCUP-NIS): sample of discharges; discharge and hospital data elements; AHRQ

NORC

International health surveys

- Demographic and Health Surveys (DHS) program:
  - reproductive age women
  - area samples (cluster $\approx$ village), face-to-face
  - $\sim$90 countries, every $\sim$3–8 years
- Multiple Indicator Cluster Surveys (UNICEF):
  - women and children
  - area samples (cluster $\approx$ village), face-to-face
  - $\sim$120 countries, irregular $\sim$3–15 years
- International Tobacco Control Policy Evaluation Project: sample design varies by country
  - phone and web in developed countries
  - anything goes in developing countries, face-to-face augmented with other modes

NORC

Language

Population $\mathcal{U} \equiv$ units for which the generalizable knowledge is sought; $|\mathcal{U}| = N$.

Observation unit $\equiv$ the entity on which the survey measurements are taken.

Sample $\mathcal{S} \equiv$ units that are selected for observation; $|\mathcal{S}| = n$.

Frame $\equiv$ a method to identify, and often contact, a unit from the target population; a link between the target and feasible population.

Coverage $\equiv$ relation between frame(s) and population. A lot of the times, no single frame covers the entire population of interest.

Sampling unit $\equiv$ the entity obtained from the sampling frame with a single draw; may match the observation units, or be a group of observation units.

EPSEM design $\equiv$ equal probability of selection method; a sampling design in which all observation units have the same probability of selection. **Does not equate i.i.d. sample.**

NORC

Features of complex surveys

The Big Four features are:

- Stratification
- Clustering
- Unequal probabilities of selection
- Weight adjustments

Other statistical features may include

- multiple phase sampling
  - subsampling respondents for specimen data collection
- multiple frames
  - all phone surveys use both landline and cell frames
- mode effect adjustments (prominent in CAHPS)

NORC

Stratification

Stratification $\equiv$ breaking up the population/frames into mutually exclusive groups before sampling.

- Geographic regions in f2f samples
- Diagnostic groups in patient list samples

Why?

- Oversample subpopulations of interest if they can be identified on the frame(s)
- Oversample areas of higher concentration of the target rare population
- Ensure specific accuracy targets in subpopulations of interest
- Utilize different sampling designs/frames in different strata
- Balance things around/avoid weird outlying samples/spread the sample across the whole population
  - Deeply stratified samples: dozens/hundreds of strata, 2 PSUs per stratum

NORC

Cluster samples

Cluster, or multistage sampling design $\equiv$ sampling groups of units rather than the ultimate observation units.

- Geographic units (e.g., census tracts) in f2f samples
- Entities in natural hierarchies (e.g., hospitals/practices and providers within a practice)

Why?

- Lists of observation units are not available, but survey statistician can obtain lists of higher level units for which residence or health service of the ultimate observation units can be established
- Reduce interviewer travel time/cost in f2f surveys
- Interest in multilevel modeling of hierarchical structures

Terminology: PSU $\equiv$ primary sampling unit $\approx$ cluster

NORC

Unequal probabilities of selection

Why?

- Oversample (smaller) subpopulations of interest (e.g., ethnic/racial minorities)
- Oversample areas of higher concentration of the target rare population
- Result of multiple stage/cluster sampling
  - Most f2f samples are drawn using probability proportional to size (PPS) sampling at the first few stages, fixed sample size at last stage $\Rightarrow$ approximately EPSEM
  - If measures of sizes are not accurate, or differential nonresponse is encountered, no longer EPSEM
- Unintended result of multiple frame sampling
  - dual phone users, i.e., those who have both landline and cell phone service, are more likely to be selected

NORC

Weight adjustments

Why? Corrections for...

- eligibility
- nonresponse (unavoidable in real world)
- frame noncoverage
- frame overlap in multiple frame surveys
- statistical efficiency

Kalton & Flores-Cervantes (2003), Valliant et al. (2013), Valliant & Dever (2017), Kolenikov (2016)

NORC

Sampling is about doing the best job for the $$$

In the end of the day, all of the sampling features are there for 1+ of the following reasons:

- Save money
  - use cluster samples to save on travel costs
  - use stratified samples to realize statistical efficiency gains
- Cannot get the full population listing
  - . . . so have to use area samples to gradually zoom down to individuals
  - . . . so have to use infrastructure created for a different purpose (telecom or postal) to contact people
- Overcome the real world data collection difficulties
  - nonresponse weight adjustments

NORC

Design-based inference

Recall the finite population sampling paradigm:

- observed characteristics are fixed (e.g., person's height and weight, presence of a medical condition)
- random variables are 0/1 indicators of being vs. not being in the sample

Remember $\langle \Omega, \mathcal{F}, P \rangle$ from your Billingsley-based class?

- elementary outcome = particular sample, i.e., subset of the population $\mathcal{U}$
- probability associated with the outcome = probability to draw a particular sample
- Hence Ω is a discrete space (although combinatorially large)
- . . . all events are finite unions of elementary outcomes, and
- . . . all sampling distributions are histograms

NORC

Design-based inference, cont.

"Easy" statistics: those linear in random variables.

We like linear statistics because expectations can be carried through, and we can produce CLT-type results through multiplication of characteristic functions.

- Trick question: is the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ a linear statistic?
- Reminder: survey statisticians think in terms of sample inclusion indicators $\mathbb{I}_i, i \in \mathcal{U}$.
- Recast the sample mean in those terms: $\bar{y} = \frac{\sum_{i\in\mathcal{U}} y_i \mathbb{I}_i}{\sum_{i\in\mathcal{U}} \mathbb{I}_i}$, which is a ratio of two random variables
  - ...unless the sample size n is fixed by design, which rarely happens in practice

NORC

Design-based inference, cont.[2]

So, what *are* the linear functions of random variables? In survey statistics, these are the totals.

- Target of inference (unobserved): population total $T[y] = \sum_{i \in \mathcal{U}} y_i$
- Sample totals (with weights): $t_w[y] = \sum_{i \in \mathcal{S}} w_i y_i \equiv \sum_{i \in \mathcal{U}} \mathbb{I}_i w_i y_i$

How to unbiasedly estimate the population total?

- Define probability of selection $\mathrm{Prob}[i \in \mathcal{S}] = \pi_i$, then
- $\mathbb{E}_p t_w[y] = \sum_{i \in \mathcal{U}} \mathbb{I}_i w_i y_i = \sum_{i \in \mathcal{U}} \pi_i w_i y_i$
- To make this unbiased for $T[y]$, irrespective of values of $\{y_i | i \in \mathcal{U}\}$, it makes sense to set weight so that $\pi_i w_i = 1 \forall i$, i.e. $w_i = 1/\pi_i$.

We just derived the Horvitz & Thompson (1952) estimator!

Note that we made no assumptions regarding y_i $\Rightarrow$

SURVEY INFERENCE IS ESSENTIALLY NONPARAMETRIC.

NORC

Design-based inference, cont.[3]

What can we say about the means of sampling distributions (point estimation)?

- To produce unbiased estimates of totals, we definitely need the Horvitz-Thompson weights.
- All other statistics are nonlinear, e.g., the weighted mean is a ratio

$$\bar{y}_w = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i} \equiv \frac{t_w[y]}{t_w[1]}$$

- If some sort of LLN holds, then functions of consistently estimated quantities will be consistent, so...
- ...it appears to make sense to use weights for all analyses.

NORC

Design-based inference, cont.[4]

Weighted mean again:

$$\bar{y}_w = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i} \equiv \frac{t_w[y]}{t_w[1]}$$

From the expression for $\bar{y}_w$, it can be seen that weights "may not matter" if y_i are "uncorrelated" with weights w_i...

- ...although the sense of "uncorrelated" is very poorly defined

Typically, survey statisticians exploit knowledge of the features of the population of interest in their sampling designs, so expecting that measurements and weights would be "uncorrelated" is unwarranted, if not dangerous, for producing unbiased estimates.

NORC

Design-based inference, cont.[5]

What can we say about the variances of sampling distributions (standard errors)?

- Variances are associated with second moments of random variables
- The random variables in surveys are $\mathbb{1}_i$, so...
  - $\mathbb{E}\mathbb{1}_i^2 = \mathbb{E}\mathbb{1}_i = \pi_i$
  - $\mathbb{E}\mathbb{1}_i\mathbb{1}_j \equiv \pi_{ij} = \text{Prob}[\text{both } i,j \in \mathcal{S}]$
- Thus, sampling variances depend on the *first order* probabilities of selection, and especially on the *joint*, or *second order* probabilities of selection:

$$\mathbb{V}(t[y]) = \sum_{i,j \in \mathcal{U}} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_i y_j$$

NORC

Design-based inference, final

Given what we just learned about means and variances in survey sampling, what are the impacts of the survey features on analysis of survey data?

- Stratification
- Clustering
- Unequal probabilities of selection
- Weight adjustments

Impacts could be on...

- Point estimates
- Variances — there is so much interest in them that survey statisticians coined a special term, *design effect*

$$\text{DEFF} = \frac{\mathbb{V}[\hat{\theta}; \text{actual design}]}{\mathbb{V}[\hat{\theta}; \text{SRS}]}$$

NORC

Stratification

- Does not have impact on π_i and w_i per se, although...
- ...if different strata are sampled at different rates π_i, weights will differ
- Samples are taken independently between strata $\Rightarrow$ different expressions for the second order probabilities for units in the same vs. different strata
  - $\pi_{ij} = \pi_i \pi_j$; $\pi_{ij} - \pi_i \pi_j = 0$ so no cross-strata terms in variance
- In the end of the day, stratification *typically* improves precision of the estimates
  - ...but to actually get this gain in your software output, you need to know what formula to use for variance

NORC

Cluster sampling

- Does not have impact on π_i and w_i per se
- Units within the same cluster have **much** higher probabilities of joint selection than units in different clusters ($\pi_{ij} \gg \pi_i\pi_j; \pi_{ij} - \pi_i\pi_j > 0$)
- Variance is clearly affected, and cluster samples typically result in a loss of precision vis-a-vis "flat" samples

$$\mathrm{DEFF}_{\mathrm{cluster}} = 1 + \rho_{\mathrm{ICC}}(\bar{m} - 1)$$

where

- ρ_{ICC} is the *intraclass correlation* $\approx$ a fraction of total variance that is due to between-cluster variance
  - typically low single digit %
- $\bar{m}$ is the (average) number of observations per cluster in the sample

Extreme cluster sample: systematic samples

NORC

Unequal probabilities of selection

- Probabilities of selection affect weights and hence point estimates
- Second order probabilities are affected $\Rightarrow$ variances are affected, as well

(Model-based) DEFF due to unequal weighting is

$$\mathrm{DEFF}_{\mathrm{uwe}} = 1 + \mathrm{CV}_w^2 = \frac{n \sum_{j \in \mathcal{S}} w_j^2}{\left[\sum_{j \in \mathcal{S}} w_j\right]^2}$$

- The second order probabilities of selection in unequal probability sampling designs are one of the most convoluted problems in sampling statistics!
- Brewer & Hanif (1983) list 50 methods of unequal probability sampling, for a good measure

NORC

Weight adjustments

- Weights affect point estimates
- Weight adjustments are based on the sample data
  - weights are random rather than fully-prespecified probability weights
- Sampling variability in weights is difficult to account for
  - Replicate variance estimation methods offer hope
- Weight adjustments increase dispersion of weights $\Rightarrow \mathrm{DEFF}_{\mathrm{uwe}}$
- On the other hand, weight calibration improves precision of estimates

$$\mathrm{DEFF}_{\mathrm{calib}}[\bar{y}] \approx 1 - R^2_{y:\mathrm{calib}}$$

where $R^2_{y:\mathrm{calib}}$ is from the regression of the outcome being analyzed on the calibration variables

Deville & Särndal (1992), Henry & Valliant (2015), Devaud & Tillé (2019)

NORC

Variance estimation

- Analytical methods: linearization $\Rightarrow$ sandwich estimators
- Computational methods: replicate variance estimation
  - Mimic the sampling design to create subsamples of you data
  - Utilize variability in estimates between subsamples to inform variance estimation
  - Balanced repeated replication (BRR; McCarthy (1969))
  - Jackknife (Krewski & Rao 1981)
  - Bootstrap (Rao & Wu 1988, Sitter 1992)
  - Successive difference (SDR)

Shao (1996), Rust & Rao (1996), Kolenikov (2010); relevant sections of Heeringa et al. (2017) or Lumley (2010).

NORC

Summary statistics

- Starting point: estimates of totals $t_w[\mathbf{y}]$
  - Estimation of proportions: define $y = 0/1$ indicator
- Point estimation: functions $g(t_w[y_1], t_w[y_2], \ldots)$
  - correlation: $r(x,y) = \frac{t_w[1]t_w[xy] - t_w[x]t_w[y]}{\sqrt{(t_w[1]t_w[x^2] - t_w[x]^2)(t_w[1]t_w[y^2] - t_w[y]^2)}}$
- Variance estimation: sandwich

$$v[g(t_w[y_1], t_w[y_2], \ldots)] = \{\nabla g\}' \{v[(t_w[y_1], t_w[y_2], \ldots)]\} \{\nabla g\}$$

NORC

Regression models

Parametric statistical models have peculiar relations with survey statistics (Skinner 1989, Binder & Roberts 2003, 2009, Pfeffermann 2011).
Census regression and normal equations:

$$y_i = Bx_i + e_i, \quad \sum_{i \in \mathcal{U}} x_i(y_i - Bx_i) = 0$$

Normal equations represent population totals (equal to zero). Plug in sample estimators:

$$\sum_{i \in \mathcal{S}} w_j x_j (y_j - b x_j) = 0 \Rightarrow b = (X'WX)^{-1}(X'Wy), W = \mathrm{diag}\{w_j\}$$

Inference: sandwich variance estimates (Fuller 1975, Binder 1983).

NORC

Subpopulations / domains

Domain $\equiv$ *subpopulation* $\equiv$ a nontrivial subset of the population

Variance estimation complexities:

- Sample size (denominator of s^2/n) is random even when the overall sample size is fixed by design
- Pairwise selection probabilities within the domain are not the true pairwise probabilities
- In practice, one has to work with filtering variables $d_i y_i$ instead of y_i where d_i is the domain indicator
- The full data set has to be used; one cannot simply drop cases that are not used for the analysis

West et al. (2008)

NORC

Complex survey analysis software

R — Stata — SAS — SUDAAN — SPSS

The bare minimum:

- survey features: stratification, clustering, unequal weights
- methods: descriptive statistics (means, totals, tabulations)

Better working set:

- survey features:
  - calibrated weights
  - replicate weights
  - multistage samples
  - multiphase samples
- methods:
  - regression models
  - hooks for custom methods to ride on

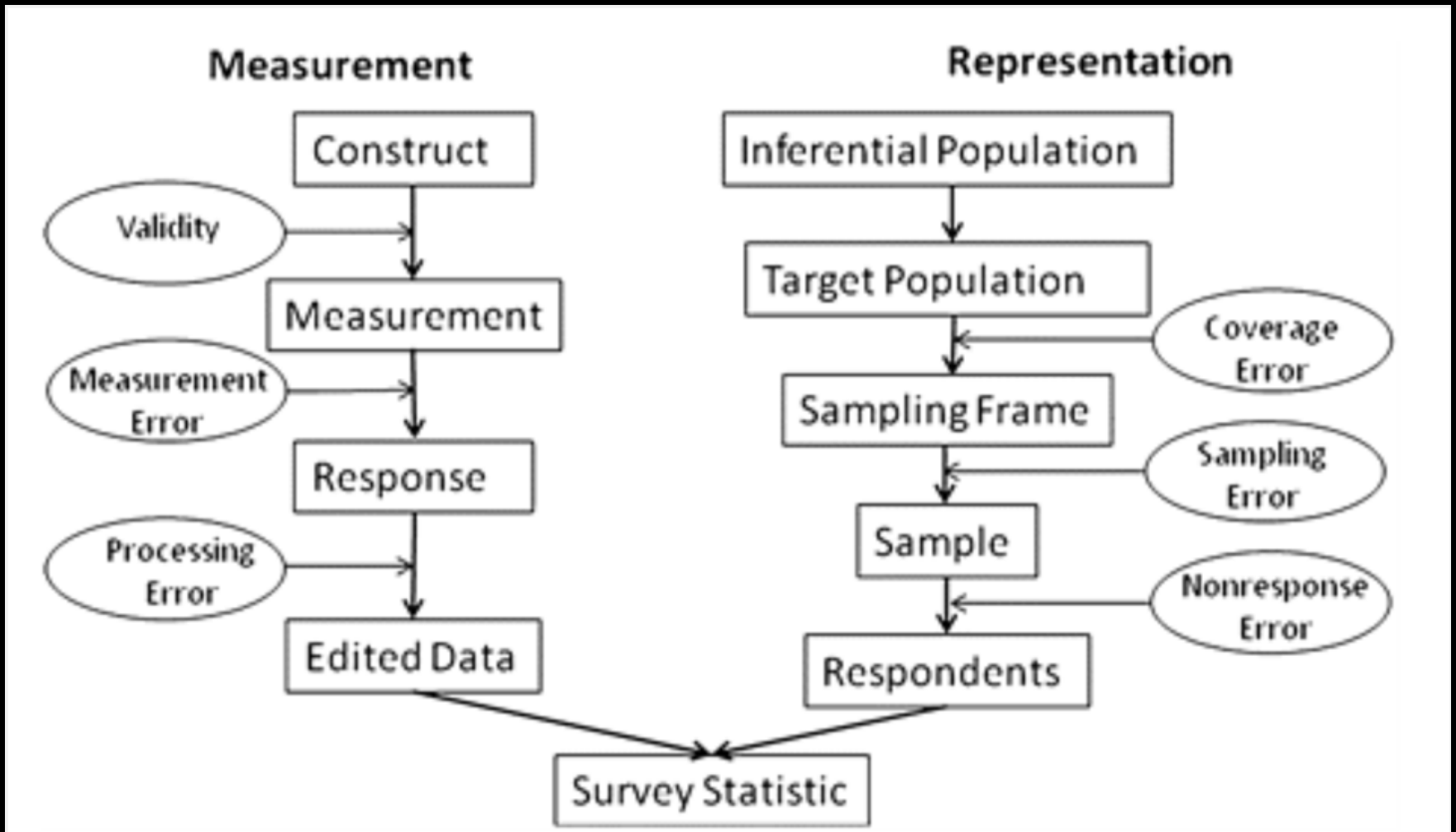
NORC

Advanced topics

- Multilevel models
- Small area estimation (MRP is one of the many methods)
- Missing (and imputed) data
- Case-control studies
- Causal inference
- Survival analysis
- Bayesian methods

NORC

Do you want to collect your own survey data?



From: Total Survey Error: Past, Present, and Future
Public Opin Q. 2010;74(5):849-879. doi:10.1093/poq/nfq065
Public Opin Q | © The Author 2011. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Focus on biases

Statisticians are reasonably good in quantifying the sampling error. The “social science” side survey methodologists concentrate on controling the measurement errors part of TSE (Groves et al. 2009, Groves & Lyberg 2010).

- Coverage bias
- Psychology of survey response
- Nonresponse (noncontact + refusal) bias
- Social desirability bias
- Cognitive shortcuts
- Mode effects
- Interviewer effects
- Multicultural, multilingual

Handbook of Health Survey Methods (Johnson 2015)

NORC

Additional resources

Professional organizations:

- Survey Research Methods Section of the ASA (SRMS)
- American Association for Public Opinion Research (AAPOR)

Online education:

- SRMS list of degree programs: http://community.amstat.org/surveyresearchmethodssection/links
- Joint Program in Survey Methodology online: https://jpsmonline.umd.edu/
- Coursera Specialization: https://www.coursera.org/specializations/data-collection

List of competencies:

- Special issue of AAPOR's *Survey Practice* online journal: https://www.surveypractice.org/issue/594-vol-8-issue-2-2015

NORC

What I covered today

1. Examples of complex survey data
2. Sampling designs and complex survey features
   - Terminology
   - Big Four: stratification, clustering, prob selection, weighting
3. Impact of complex survey features on analysis
   - Biases
   - Design effects
   - Subpopulations / domains
4. Other topics, one slide each
   - Software for complex survey data analysis
   - Advanced uses of survey data
   - Collecting survey data
   - Resources

NORC

Appendix I: Software

Code examples: https://github.com/skolenik/ICHPS2023-svy

NORC

Complex survey analysis software

R — Stata — SAS — SUDAAN — SPSS

The bare minimum:

- survey features: stratification, clustering, unequal weights
- methods: descriptive statistics (means, totals, tabulations)

Better working set:

- survey features:
  - calibrated weights
  - replicate weights
  - multistage samples
  - multiphase samples
- methods:
  - regression models
  - hooks for custom methods to ride on

NORC

R

`library(survey)` — available from CRAN (Lumley 2010)
`library(ReGenesees)` — requires custom installation

1. Declare your survey design (e.g., identifiers of units, strata, clusters; weight variables; etc.)
2. Apply survey-aware functions provided by those packages to obtain design-correct inference

Both packages support nearly any design imaginable, as well as estimation of common statistical models (e.g. GLMs)

Overall impression: very solid

`library(srvyr)` — a tidyverse interface to {`survey`}
`library(sampling)` — a suite of sampling methods (Tillé 2006)

NORC

Stata

Official Stata includes the suite of `svy` routines.

1. `svyset` your data (identifiers of units, strata, clusters; weight variables; calibration variables)
2. prefix nearly any estimation command with `svy` for design-correct inference
3. clear (but complex) ways for third-party modules to comply with `svy` requirements
4. calibrated weights (although poorly documented)

Overall impression: very solid
Lags behind R: sampling; beats R: variety of supported models

Heeringa et al. (2017), Kolenikov & Pitblado (2014).

NORC

SAS

The suite of `SAS PROC SVYWHATEVER` procedures:

- need to declare the survey features within every procedure (copy/paste errors?)
- a limited set of methods coded in `SAS PROC SVYWHATEVER`
- other procedures produce ridiculous results with weights

Overall impression: OK for the general 90% tasks

Lags behind R and Stata: variety of methods.

Beats Stata with `PROC SURVEYSELECT`.

NORC

SUDAAN

SAS-callable custom software (SUrvey DAta ANalysis)

- need to declare the survey features within every procedure (copy/paste errors?)
- a wider set of methods compared to SAS (including Cox regression)
- sophisticated weight adjustment methods

Overall impression: very strong within its limited scope; future development unclear

NORC

SPSS

- weighted procedures produce ridiculous results
- requires a separate purchase of Complex Samples module
- replicate weights are not supported
- future development unclear

Overall impression: meh

NORC

Software and rubrics

	R	Stata	SAS	SUDAAN	SPSS
Integration	`library (survey)`	Official `svy:`	PROC SVY*	SAS-callable	Complex Samples ($$)
Big three	+	+	+	+	+
Replicate weights	+	+	+	+	-
Calib $\mathbb{V}$	+	+	-	+	-
Models	++	++	+	+	?

Appendix II: Statistical models

NORC

Advanced topics

- Multilevel models
- Small area estimation
- Missing (and imputed) data
- Case-control studies
- Causal inference
- Survival analysis
- Bayesian methods

NORC

Multilevel models

$$y_{ij} \sim f(\theta_{ij}) \text{ (exponential family)}, \quad \theta_{ij} = x'_{ij}\beta + Zu_i$$

- Mixed models in biostatistics
- Multilevel models in social sciences
- If clusters and individuals are sampled in an informative way, variance components that are central in the maximum (quasi-)likelihood estimation may work in strange ways
- Open research question since Pfeffermann et al. (1998) as to how scale the weights at the lower levels (to tweak biases in variance estimates, efficiency gains)

NORC

Small area estimation

Try to do more with less – provide estimates for domains with insufficient sample size

- Fay & Herriot (1979) model:

$$g(\text{poverty rate}_i) = x_i'\beta + \text{model error}_i + \text{sampling error}_i$$

- Unit models (Rao & Molina 2015)
  - sample model

$$y_{ij} = \mathbf{x}_{ij}'\beta + u_i + e_{ij}$$

  - combined with known population totals $\Rightarrow$ estimate

$$\bar{Y}_i = f_i\bar{y}_i + (1 - f_i)(\bar{\mathbf{X}}_i'\hat{\beta} + \hat{u}_i)$$

  - GLM-like extensions are feasible
- Rediscovered independently in other literatures (economics: poverty mapping; political science: MRP), invariably with problems (Molina & Rao 2010)

NORC

Imputed data

If casewise deletion, imputation by (mean, mode, etc.), single imputation are bad, is multiple imputation (Rubin 1996) a good answer?

- Kim et al. (2006) — multiple imputation may not play well with the complex designs
- Shao & Sitter (1996) bootstrap:
  1. Draw bootstrap sample reflecting complex design
  2. Impute once
  3. Estimate, store the pseudo-value
  4. Repeat 1–3 the “bootstrap” number of times (hundreds, not single digits)
  5. Combine using the bootstrap rules (although Rubin rules would be about the same)

NORC

Case-control studies

- Cases: sampled at 100% rate (or close)
- Controls: match cases, in some appropriate sense
- Controls are sampled at much lower rate, $\approx$ the condition prevalence

The interest is often in the causes of the condition — tempting to run a logistic regression with the condition as the outcome?

- Regression with weights is consistent wrt design, but likely very inefficient
- Regression without weights sounds scary as the design is **very** informative
- If sampling of the controls is not informative, then only the intercept is biased (Scott & Wild 2003)

NORC

Survival analysis

Nearly all of the expertise is encoded in Jerry Lawless' brain (Lawless 2003).

- Survival analysis == instantaneous hazard rates $\approx$ moderately high frequency of observations (days); surveys $\approx$ one time operations
- Even in longitudinal survey studies, time between data collection waves ($\sim$year) may contain several events
- Severe left and right censoring issues
- Proportional hazards Cox model: Binder (1992)

NCHS linked the National Death Index (NDI) with NHIS, NHANES, some other surveys.

NORC

Bayesian methods

Mostly forced marriage:

- Bayesian methods proceed by tightening the distributions of parameters driven by the likelihood supplied by the data
- Survey inference (nonparametric!): likelihood of an SRS sample = $1/\binom{N}{n}$ and does not depend on y_i
- One can construct models and priors that allow reproducing some of the simple textbook formulae (Little 2012)

Bayesian methods work very well when complementing survey inference conditional on the sample

- Missing data imputation
- Small area estimation
- Adaptive sampling designs

NORC

Interface of svy and te

Looks like we need three probability spaces:

1. Superpopulation model ξ to produce $(y_i^{(0)}, y_i^{(1)}, x_i)$
2. Sampling p
3. Treatment assignment T

- Large scale evaluation: ξ creates the universe, T is a randomized assignment, the survey follows everyone adjusting for nonresponse: $p \succ T \perp \xi$
- Survey experiments (e.g. question wording, mode assignment): ξ creates the universe, p draws the sample, T is conditional on the sample drawn, $T \perp \xi \otimes p$.
- Everything else: the probabilities spaces do not separate.

NORC

Propensity scores and weights

- Dong et al. (2020): literature review of the more commonly used methods to estimate PATE or PATT
  1. ignore everything
  2. throw weights in as PS predictors
  3. ignore weights in PS modeling and subclassification, use weights in outcome regression within subclasses
  4. multiply the weights through for the outcome analysis
  5. that is not the full factorial space!
- Ridgeway et al. (2015): conditions on when unweighted propensity score analysis is problematic
- Lenis et al. (2017), Austin et al. (2018): weights must be used in the outcome analysis, may not matter as much for PS model itself if covariates are balanced in population / weighted sample
  - Dong et al. (2020) walked it back
- Mercer et al. (2017): Rubin's framework applied to probability / non-probability samples

NORC

IPW ain't H-T

Estimator: if the estimator is $\tau_{\text{ATE}} = \frac{yt}{p} - \frac{(1-t)y}{1-p} := \sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i}$, what are the components?

Matching the concepts from the T space to the p space:

- Population: $\{-y^0, y^1\}$
- Sample: $n = 1$
- Probabilities: $\pi_{01} = 0$
- Variance: $(y^1 + y^0)^2$

Do these make sense?

NORC

References I

Austin, P. C., Jembere, N. & Chiu, M. (2018), 'Propensity score matching and complex surveys', *Statistical Methods in Medical Research* **27**(4), 1240–1257. PMID: 27460539.
URL: *https://doi.org/10.1177/0962280216658920*

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

Binder, D. A. (1992), 'Fitting Cox's proportional hazards models from survey data', *Biometrika* **79**(1), 139–147.

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.

NORC

References II

Binder, D. A. & Roberts, G. R. (2009), Design- and model-based inference for model parameters, *in* D. Pfeffermann & C. R. Rao, eds, 'Sample Surveys: Inference and Analysis', Vol. 29B of *Handbook of Statistics*, Elsevier, Oxford, UK, chapter 24.

Brewer, K. & Hanif, M. (1983), *Sampling with Unequal Probabilities*, Springer-Verlag, New York.

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

Devaud, D. & Tillé, Y. (2019), 'Deville and sarndal's calibration: revising a 25-years-old successful optimization problem (with discussion)', *Test* **28**(4), 1033–1065.

Deville, J. C. & Särndal, C. E. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**(418), 376–382.

NORC

References III

Dong, N., Stuart, E. A., Lenis, D. & Nguyen, T. Q. (2020), 'Using propensity score analysis of survey data to estimate population average treatment effects: A case study comparing different methods', *Evaluation Review* **44**(1), 84–108. PMID: 32672113. **URL:** *https://doi.org/10.1177/0193841X20938497*

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.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2009), *Survey Methodology*, Wiley Series in Survey Methodology, 2nd edn, John Wiley and Sons, New York.

Groves, R. M. & Lyberg, L. (2010), 'Total survey error: Past, present, and future', *Public Opinion Quarterly* **74**(5), 849–879.

NORC

References IV

Heeringa, S. G., West, B. T. & Berglund, P. A. (2017), *Applied Survey Data Analysis*, Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, 2nd edn, Chapman and Hall/CRC.

Henry, K. A. & Valliant, R. (2015), 'A design effect measure for calibration weighting in single stage samples', *Survey Methodology Journal* **41**(2), 315–331.

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**(260), 663–685.

Johnson, T. P., ed. (2015), *Handbook of Health Survey Methods*, Wiley Handbooks in Survey Methodology, Wiley, Hoboken, NJ.

Kalton, G. & Flores-Cervantes, I. (2003), 'Weighting methods', *Journal of Official Statistics* **19**(2), 81–97.

NORC

References V

Kim, J. K., Brick, M. 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.

Kolenikov, S. (2010), 'Resampling inference with complex survey data', *The Stata Journal* **10**, 165–199.

Kolenikov, S. (2016), 'Post-stratification or non-response adjustment?', *Survey Practice* **9**(3). available at http://www.surveypractice.org/index.php/SurveyPractice/article/view/3

Kolenikov, S. & Pitblado, J. (2014), Analysis of complex health survey data, *in* T. P. Johnson, ed., 'Handbook of Health Survey Methods', Wiley, Hoboken, NJ, chapter 29.

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

NORC

References VI

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.

Lawless, J. F. (2003), Event history analysis and longitudinal surveys, *in* R. L. Chambers & C. J. Skinner, eds, 'Analysis of Survey Data', Wiley, chapter 15, pp. 221–243.

Lenis, D., Nguyen, T. Q., Dong, N. & Stuart, E. A. (2017), 'It's all about balance: propensity score matching in the context of complex survey data', *Biostatistics* **20**(1), 147–163.
URL: *https://doi.org/10.1093/biostatistics/kxx063*

Little, R. J. (2012), 'Calibrated Bayes, an alternative inferential paradigm for official statistics', *Journal of Official Statistics* **28**(3), 309–334.

Lumley, T. S. (2010), *Complex Surveys: A Guide to Analysis Using R (Wiley Series in Survey Methodology)*, Wiley, New York.

NORC

References VII

McCarthy, P. J. (1969), 'Pseudo-replication: Half samples', *Review of the International Statistical Institute* **37**(3), 239–264.

Mercer, A. W., Kreuter, F., Keeter, S. & Stuart, E. A. (2017), 'Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference', *Public Opinion Quarterly* **81**(S1), 250–271.
URL: *https://doi.org/10.1093/poq/nfw060*

Molina, I. & Rao, J. N. K. (2010), 'Small area estimation of poverty indicators', *Canadian Journal of Statistics* **38**(3), 369–385.

Pfeffermann, D. (2011), 'Modelling of complex survey data: Why model? why is it a problem? how can we approach it?', *Survey Methodology* **37**(2), 115–136.

Pfeffermann, D. & Rao, C. R., eds (2009), *Handbook of Statistics, Volume 29: Sample Surveys*, North Holland.

NORC

References VIII

Pfeffermann, 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.

Rao, J. N. K. & Molina, I. (2015), *Small Area Estimation*, Wiley series in survey methodology, 2nd edn, John Wiley and Sons, Hoboken, NJ.

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.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A. & Kabeto, M. U. (2015), 'Propensity score analysis with survey weighted data', *Journal of Causal Inference* **3**, 237–249.

Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American Statistical Association* **91**(434), 473–489.

NORC

References IX

Rust, K. F. & Rao, J. N. (1996), 'Variance estimation for complex surveys using replication techniques', *Statistical Methods in Medical Research* **5**(3), 283–310.

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 (with discussion)', *Statistics* **27**, 203–254.

Shao, J. & Sitter, R. R. (1996), 'Bootstrap for imputed survey data', *Journal of the American Statistical Association* **91**(435), 1278–1288.

Sitter, R. R. (1992), 'Comparing three bootstrap methods for survey data', *The Canadian Journal of Statistics* **20**(2), 135–154.

NORC

References X

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. & Wakefield, J. (2017), 'Introduction to the design and analysis of complex survey data', *Statistical Science* **32**(2), 165–175.

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

Valliant, R. & Dever, J. (2017), *Survey Weights: A Step-by-step Guide to Calculation*, Stata Press, College Station, TX.

Valliant, R., Dever, J. A. & Kreuter, F. (2013), *Practical Tools for Designing and Weighting Survey Samples*, Springer.

West, B. T., Berglund, P. & Heeringa, S. G. (2008), 'A closer examination of subpopulation analysis of complex-sample survey data', *Stata Journal* **8**(4), 520–531(12).

NORC

References XI

West, B. T., Sakshaug, J. W. & Aurelien, G. A. S. (2016), 'How big of a problem is analytic error in secondary analyses of survey data?', *PloS One* **11**(6), e0158120.

Wu, C. & Thompson, M. E. (2020), *Sampling Theory and Practice*, Springer, New York.

NORC