Analyzing Data Collected with Complex Survey Design

Hsueh-Sheng Wu

CFDR Workshop Series

June 28, 2021

BGSU
Center for Family and Demographic Research

Outline

- What does complex survey design mean?
- Probability sampling designs
- Weight variables
- Differences between sample and population
- Consideration of using weight variables
- Introduction to -*svy*- command in Stata
- Use the -svyset- command for different datasets
- Examples of analyzing data collected with a complex survey design
- Conclusions

What Does Complex Survey Design Mean?

- Social scientists often analyze data collected from a subgroup of individuals (i.e., a sample) to infer the characteristics of the target population where these individuals were selected from
- When a sample is selected from the target population with non-probability sampling methods, such as quota sampling, convenience sampling, and snowball sampling, it is impossible to use sample statistics to infer the statistics of the target population
- When a sample is selected with probability sampling methods, such as simple random sampling, cluster sampling, and stratified sampling, it provides researchers a way of using probability weights, primary sampling units (PSU), strata, and finite population correction (FPC) to infer population statistics using sample statistics.
- To better protect the confidentiality of respondents, some national datasets that were collected with probabilty sampling designs have started to inlcude only replicate weight variables, and not probability weights, PSU, strata, and FPC in the data. For these datasets, replicate weight variables should be used.

Probability Sampling Designs

Simple Random Sampling (SRS):

- Each individual in the sampling frame gets a number
- Random numbers are used to decide which individuals are selected

Systematic Sampling:

- Use the ratio of sample size to the population size to decide the sampling interval
- Arrange the respondents in an order
- Randomly select a respondent within the first sampling interval. This respondent is the first selected respondent
- Use the position of the first selected respondent and the sampling interval to decide the next selected respondent

Probability Sampling Designs (Continued)

Stratified Sampling:

- Divide the sample frame into different strata and then randomly select respondents from each stratum
- Stratified sampling assumes that heterogeneity across strata and homogeneity within each stratum
- Researchers can over- or under-sample some sub-population. For example, over-sample minority group to make sure the sample has enough variation and statistical power.
- Over- or under-sampling can be done because different sampling methods and ratios can be used for different strata

Probability Sampling Designs (Continued)

Cluster Sampling:

- Divide the sampling frame into different clusters and then randomly select respondents from each cluster
- Assume that homogeneity across clusters and heterogeneity within each cluster
- Cluster sampling is used because it saves time, money, and energy (e.g., geographical proximity saves the cost and time of traveling)
- Cluster sampling generally has more variability or less efficiency than Simple Random Sampling

Probability Sampling Designs (Continued)

Multi-stage sampling: combination of SRS, systematic sampling, stratified sampling, or cluster sampling

Example 1 (Add Health, a stratified two-stage sampling)

- The sampling frame is stratified by region, urbanization, school size, school type, and race composition
- 80 high schools and 52 middle schools were selected with unequal probability at the first stage
- 90,000 students were selected to fill out in-school Add Health questionnaire, and 27,000 of them fill out in-home questionnaire
- Add Health data oversampled twins and siblings of twins; non-related adolescents residing together; disabled minority students; Blacks from well-educated families; and minority students who are Chinese, Cubans, and Puerto Ricans

Probability Sampling Designs (Continued)

Example 2 (American Community Survey, two-phrase three-stage sampling)

- Main processing:
  - In August, select housing units from National Master Addressing File (MAF) created by Census Bureau
  - Assign housing units to five existing sub-frames
  - Decide which sub-frame to use
  - Select housing units
  - Select individuals
- Supplement phase:
  - In January, select housing units that are not listed in the National Master Addressing File (MAF) in the previous August.
  - Assign housing units to five existing sub-frames
  - Decide which sub-frame to use
  - Select housing units
  - Select individuals

Weight Variables

The specification of sampling designs usually rely on the following variables

• Weights: There are different types of weight variables. The most common one is the probability weight, calculated as the inverse of the probability of being selected in the sample

• Primary sampling unit (PSU): PSU is the first unit that is sampled in the design, indicating where sampling design started

• Stratum: Stratum is used in stratified sampling. In general, each stratum should have at least two elements in a stratum. If not, you need to specify how to solve this problem by using the -singleunit- option in Stata

Weight Variables (Continued)

- Finite Population Correct (FPC): FPC is used to correct the standard error of the estimate, especially when the sampling fraction is large

- Replicate weights: Replicate weights eliminate the needs of providing PSUs and Strata in the data file, so it can better reserve the confidentiality of respondents. The logic of replicate weights is as follows:
  - Divide the population into subsamples
  - Obtain the estimate of the parameter from the full sample and sub-samples
  - The difference in the estimates is used to calculate the variance of the parameter.

Differences between Sample and Population

- Different weights for sampled respondents
  - The under-coverage of the sampling frame
  - Under- and/or over-sampling of certain demographic groups
- Different survey participation
  - Refusal of participation
  - Lost in the follow-up surveys
  - Item-nonresponse
- Augmented data
  - Replenish sample

Consideration of Using Weight Variables

- What is the research question?
- What data set will be use?
- What statistic models will be used?

- What is the population of the data set?
- What is the sampling design of the data file?
- What are weight variables available?

- Can statistical software incorporate the survey design and run the statistical model simultaneously?

Introduction to -*svy*- Command in Stata

The -*svy*- command allows Stata to analyze survey data while taking into account how the data were collected

Use the -*svyset*- to specify the survey design of the data

One-stage clustered design with stratification:

svyset su1 [pweight=pw], strata(strata), fpc(fpc1)

Multiple-stage designs:

svyset su1 [pweight=pw], fpc(fpc1) strata(strata) || su2, fpc(fpc2) || su3, fpc(fpc3)

Then use the -*svy*- subcommands to analyze the survey data

Introduction to -*svy*- Command in Stata

Table 1. Additaional Analyses for data collected with complex survey data

Comands	Analysis	Comands	Analysis
svy: biprobit	Bivariate probit regression for survey data	svy: nl	Nonlinear least-squares estimation for survey data
svy: clogit	Conditional (fixed-effects) logistic regression for survey data	svy: oprobit	Ordered probit regression for survey data
svy: cloglog	Complementary log-log regression for survey data	svy: poisson	Poisson regression for survey data
svy: cnreg	Censored-normal regression for survey data	svy: probit	Probit regression for survey data
svy: cnsreg	Constrained linear regression for survey data	svy: proportion	Estimate proportions for survey data
svy: glm	Generalized linear models for survey data	svy: ratio	Estimate ratios for survey data
svy: gnbreg	Generalized negative binomial regression for survey data	svy: scobit	Skewed logistic regression for survey data
svy: heckman	Heckman selection model for survey data	svy: slogit	Stereotype logistic regression for survey data
svy: heckprob	Probit model with sample selection for survey data	svy: stcox	Cox proportional hazards model for survey data

Introduction to -*svy*- Command in Stata

Table 1. Additaional Analyses for data collected with complex survey design

Comands	Analysis	Comands	Analysis
svy: hetprob	Heteroskedastic probit regression for survey data	svy: streg	Parametric survival models for survey data
svy: intreg	Interval regression for survey data	svy: tobit	Tobit regression for survey data
svy: ivprobit	Probit model with endogenous regressors for survey data	svy: total	Estimate totals for survey data
svy: ivregress	Single-equation instrumental-variables regression for survey data	svy: treatreg	Treatment-effects regression for survey data
svy: ivtobit	Tobit model with endogenous regressors for survey data	svy: truncreg	Truncated regression for survey data
svy: logistic	Logistic regression for survey data, reporting odds ratios	svy: zinb	Zero-inflated negative binomial regression for survey data
svy: logit	Logistic regression for survey data, reporting coefficients	svy: zip	Zero-inflated Poisson regression for survey data
svy: mean	Estimate means for survey data	svy: ztnb	Zero-truncated negative binomial regression for survey data
svy: mprobit	Multinomial probit regression for survey data	svy: ztp	Zero-truncated Poisson regression for survey data
svy: nbreg	Negative binomial regression for survey data		

Use the -svyset- Command for Different Datasets

American Community Survey (*ACS*):

svyset[pweight=perwt], vce(brr) brrweight(repwtp1-repwtp80) fay(.5) mse

The National Longitudinal Study of Adolescent to Adult Health (*Add Health*):

svyset psuscid [pw = wt_var], strata(region)

Current Populaton Survey (*CPS*):

svyset [iw=wtsupp], sdrweight(repwtp1-repwtp160) vce(sdr)

General Social Survey (*GSS*):

svyset [weight=wtssall], strata(vstrat) psu(vpsu) singleunit(scaled)

Use the -svyset- Command for different Datasets (Continued)

National Longitudinal Survey of Youth 1997 (*NLSY97*):

svyset [pweight=WTVAR] , strata(vstrat) psu(vpsu) singleunit(scaled)

The Survey of Income and Program Participation (*SIPP*):

svyset [pweight=_yourwgt], brrweight(RepWt_1-RepWt_n) fay(.5) vce(brr) mse

US Census 2000:

svyset serialno [pweight = pweight]

Examples of Analyzing Data Collected with a Complex Survey Design

See the supplement command and log files

Conclusions

- The way that data were collected determines what design variables should be used in data analyses and how well sample statistics approximate population statistics. Without considering design variables, researchers are unlikely to get an accurate analysis results.
- Each data set has its own research purpose and is uniquely designed. It is very imporant to consult the user guide of a dataset and decide how to best specify the -svyset- command for the data set.
- While using the same longitudinal data set, the setup of -svy- command may be different, depending on what the target population is, whether a sub-population is examined, and whether strata have only one sampling unit.
- After the -svyset- command is setup, you can analyze data using the --svy- command along with the regular analysis command
- For further question, feel free to contact me at wuh@bgsu.edu or stop by my office (5D Williams Hall)