WELCOME TO THE 2024 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES: BEST PRACTICES IN THE USE OF NDACAN DATA

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

SURVEY DESIGN AND USING WEIGHTS

JULY 31, 2024

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

NDACAN SUMMER TRAINING SERIES SCHEDULE

July 10 — NCANDS: Strengths & Limitations

July 17 — Assessing Reporting Issues in NCANDS & AFCARS

July 24 — AFCARS: Strengths & Limitations

July 31 — Survey Design & Using Weights

August 7 — NSCAW III for Experienced & New Users

August 14 — NYTD: Strengths & Limitations

SESSION AGENDA

- Survey sampling
- Survey weights
- Survey analysis in programming languages

TERMS

- Target population
- Sample population
- Primary sampling unit (PSU)
- Secondary sampling unit (SSU)
- Strata
- Cluster

SURVEY SAMPLING

SAMPLING

- Want to make inference about a *target population* but can't get data on every single unit (e.g. person) in the population so take a *sample*
- Want to minimize sampling error and bias, and survey time and cost, while maximizing coverage and precision
- Sampling is almost always done without replacement
  - Samples are usually "small enough" with respect to the target population such that one unit's probability of inclusion will not affect another unit's
  - If a sample becomes "large enough" with respect to the target population (for example, sample of 90% of the full target population) then would need to consider a *finite population correction* in weighted analysis

PROBABILITY SAMPLING

- Each unit has a calculable probability of being sampled
- The Ideal: Simple random sample
  - Every unit of measure in the target population is selected with equal probability and therefore has unbiased representation
  - Because of this unbiased randomness and Law of Large Numbers, if you take "large enough" sample, the sample estimates will converge to the population values
  - Cons
    - May miss very small populations because of imbalance aka unrepresentative groups across geographic regions of other important domains of interest, e.g. race, sex, age
    - Not usually practical for planning/implementation or costs
    - In reality, usually heterogenous response rates by geography, survey methods, demographics

COMMON PROBABILITY SAMPLING METHODS

- Stratified sampling – divide the population into homogenous, mutually exclusive groups, i.e. *strata*; then independent samples taken within strata
  - Ensure adequate sample size for subgroups of interest (usually imbalanced)
  - Increases precision
- Cluster sampling – population divided into groups/clusters and then randomly select number of clusters, where all units in chosen clusters are included in sample (e.g. the cluster is the sampling unit, contrary to stratified)
  - Mutual homogeneity but internal heterogeneity
  - Increases sampling efficiency
- Multi-stage designs – can use cluster and/or stratified sampling to divide population at multiple levels

MAIN BIASES IN PROBABILITY SAMPLING

- Non-response bias – people who don't respond may be characteristically different from those who are responding
- Selection bias - some units have a differing probability of selection that is unaccounted for by the researcher
- Coverage bias - some population members do not appear in the sample frame (under-coverage), e.g. homelessness, incarceration, out of reach from surveying technique (such as phone)

SURVEY DESIGN

DESIGNING A SURVEY

- Define target population – the largest encompassing group of all units (e.g. individuals) to which inference and conclusions can be made
- Define sampling frame and design
  - Define strata or clusters that the whole target population can be divided into
    - Usually based on geography (e.g. states) or demographics
  - Define any second stage strata or clusters
    - Based on geography or natural organization (e.g. child welfare agencies in a state)
  - Define any additional sampling clusters – e.g. based on "domains" or demographics of interest, or groups to oversample
  - Define primary sampling units that will be randomly sampled (e.g. children, or families)
- A survey design is decided and set before data collection begins and should remain unchanged for multi wave data

EXAMPLE: NSCAW 2

- National Survey of Children and Adolescent Wellbeing – longitudinal survey with 3 waves
- **Target population**:
  - All children in the U.S. who were subjects of child abuse or neglect investigations conducted by child protective services (except those living in 8 states where laws interfered with survey administration, and thus removed from sampling frame)
- **Multi-stage stratified design**:
  - U.S. divided into 9 strata – 8 correspond to largest states, 9th is all remaining states
  - Within strata, PSUs were geographic areas that encompass the population served by a single CPS agency (usually equivalent to counties)
  - All children within PSU were categorized into 5 mutually exclusive domains (e.g. infants receiving services, children 1-17.5 receiving services, etc) and then randomly sampled within the domain

WEIGHTS

"*Survey weighting is a mess*" - Andrew Gelman

WHY WEIGHT

- Using survey weights in analysis ensures conclusions and inference are applicable to the whole target population
- Adjust for survey design error and bias
- Without survey weights, standard error calculations from statistical programming languages will be underestimated, and significant results may be false
- The rule of thumb is to use survey weights if available
  - Survey weights are almost always recommended for descriptive statistics (e.g. means, proportions)
  - There is less consensus about always using survey weights in statistical models (depends on many factors)

WHAT DO WEIGHTS DO

- Survey weights compensate for estimation bias from:
  - Unequal selection probabilities
  - Unit non-response
  - Loss of population coverage
  - Survey administration issues (e.g. data collection pauses, re-sampling or adjusting mid-way)
- Every primary sampling unit who has a valid observation will have a survey weight
- "Final" analysis weights are usually the cumulative product of multiple adjustments for each stage of sampling (e.g. a cumulative probability of selection and response adjustments at each stage)

BASE WEIGHTS

- Adjust for inclusion probability during sampling
  - Inverse probability of being chosen in each stage/strata/domain
  - If you were less likely to be chosen your weight will be higher, i.e. larger representation
- Need estimates of the number of units in the target population (aka reference population) and within each defined strata/domain – needed to calculate probabilities and for calibration later

WEIGHT ADJUSTMENTS

- Additional weight adjustment factors may be calculated to adjust for other bias such as:
  - Non-response
    - Can incorporate probability of non-response based on characteristics, e.g. sociodemographic
  - Survey problems
    - Anything that necessitates revising the original survey design and/or additional resampling – non-compliance, higher than expected non-response
    - Extended duration of survey administration – makes it hard to set a reference population, time induced biases/changes in responses

WEIGHT CALIBRATING (CTD.)

- Sometimes constructed weights can have large variation which can reduce precision
- Calibrating and adjustments may come in any order and may be done multiple times, usually finish weighting construction with trimming and/or calibrating

WEIGHT CALIBRATING

- **Smoothing**
  - Model-based weights that use observed survey quantities (rather than non-random inverse probabilities)
  - Reduces variability in weights
- **Calibration/post-stratification**
  - Adding survey weights across the sampling frame strata and domains should add up to the known/estimated target population totals
  - After weighting adjustments, take the sum of weights and get multiplier such that the sum will equal the total population
  - Decreases bias due to non-response and underrepresented groups
- **Trimming/Winsorization**
  - Extreme value weights can be outlier and heavily influence variance therefore trimming will set the largest value weights equal to some predefined percentile (e.g. 99^{th})
  - Reduces variability but will increases bias – delicate balance and careful consideration of cutoffs

WEIGHTS IN A MULTI-WAVE SURVEY

- There will be at least one weight for each wave (if not more) in multi wave survey data
- Weights in the first wave of any multi-wave survey will be constructed in the same way as discussed (e.g. base weights and adjustments)
- All subsequent waves' weights are usually simply just the first waves' weight cumulatively multiplied with additional adjustments for attrition, e.g. the probability of responding up to that wave

CHOOSING WEIGHTS IN A MULTI-WAVE SURVEY

- The choice of weights in a multi-wave survey will depend on
  - The estimates or model of interest and any corresponding variables used
  - The "path" in which someone can take to arrive at each wave
    - Some surveys are such that people must respond to each wave to be eligible for the next wave.
    - Some surveys are such that if you responded at the first wave, you could respond to any subsequent wave

SURVEY ANALYSIS IN PROGRAMMING LANGUAGES

SURVEY FUNCTIONS IN PROGRAMMING LANGUAGES

- Need to define the sampling frame to your programming language – strata, (primary and secondary) sampling units, probability weights
- Stata
  - *svyset* and *svy* prefix
  - https://www.stata.com/manuals/svy.pdf
- R
  - *survey* package
  - https://stats.oarc.ucla.edu/r/seminars/survey-data-analysis-with-r/
- SPSS
  - *Complex Samples* add-on survey analysis
- SAS
  - *PROC SURVEY*
  - https://stats.oarc.ucla.edu/sas/seminars/sas-survey/

WORKING WITH SUBSAMPLE

- Often want to do analysis on a subpopulation of the survey sample (e.g. only certain race or sex, or age constraints)
- You should always leave all observations in memory (don't drop anything) even when doing subsample analysis - this affects variance calculation and the underlying sampling frame
- Define your subpopulation in a binary variable where 1 indicates your subsample, and 0 not
- Then use this variable in your programming language to tell it to do subsample analysis

STATA EXAMPLE WITH NSCAW II

- **First, define survey design**

Sample code for defining the survey design:

```
. svyset nscawpsu [pweight= nanalwt], strata(stratum)

Sampling weights:   nanalwt
              VCE:  linearized
     Single unit:   missing
        Strata 1:   stratum
  Sampling unit 1:  nscawpsu
            FPC 1:  <zero>
```

STATA EXAMPLE WITH NSCAW II (PT. 2)

- Use `svy` prefix in order to incorporate svy design to any analysis
- Example, get proportion of children in Wave 1 who are male/female

```
svy: prop chdGendr
```

```
. svy: prop chdGendr
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =  8                  Number of obs    =     5,872
Number of PSUs   = 82                  Population size = 2,474,846
                                       Design df       =        74
```

	Proportion	Linearized std. err.	Logit [95% conf.	interval]
chdGendr				
Male	.5086575	.0124677	.4838141	.5334581
Female	.4913425	.0124677	.4665419	.5161859

STATA EXAMPLE WITH NSCAW II (PT. 3)

- Compare proportion with svy prefix and without

svy: prop chdGendr

```
. svy: prop chdGendr
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =  8                    Number of obs   =     5,872
Number of PSUs   = 82                    Population size = 2,474,846
                                         Design df       =        74
```

	Proportion	Linearized std. err.	Logit [95% conf. interval]	
chdGendr				
Male	.5086575	.0124677	.4838141	.5334581
Female	.4913425	.0124677	.4665419	.5161859

prop chdGendr

```
. prop chdGendr

Proportion estimation                    Number of obs = 5,872
```

	Proportion	Std. err.	Logit [95% conf. interval]	
chdGendr				
Male	.5137943	.0065225	.5010016	.5265689
Female	.4862057	.0065225	.4734311	.4989984

STATA EXAMPLE WITH NSCAW II (PT. 4)

- Specify subpopulation of just females
  - `gen subsamp_gender = .`
  - `replace subsamp_gender = 1 if chdGendr == 2`
  - `replace subsamp_gender = 0 if chdGendr == 1`
- Get distribution of poverty level for females
  - `svy, subpop(subsamp_gender): tab cgdpovrt`

STATA EXAMPLE WITH NSCAW II (PT. 5)

```
svy, subpop(subsamp_gender): tab cgdpovrt
```

```
. svy, subpop(subsamp_gender): tab cgdpovrt
(running tabulate on estimation sample)

Number of strata =  8                    Number of obs    =     5,872
Number of PSUs   = 82                    Population size = 2,474,846
                                         Subpop. no. obs =     2,855
                                         Subpop. size    = 1,215,997
                                         Design df       =        74

------------------------------
% Federal |
Poverty   |
Level     |  proportion
----------+-------------------
  Missing |       .0718
    < 50% |       .2433
 50% - <1 |       .2912
 100%-200 |       .2354
    >200% |       .1583
          |
    Total |          1
------------------------------
Key: proportion = Cell proportion
```

```
tab cgdpovrt if chdGender == 2
```

```
. tab cgdpovrt if chdGendr == 2

  % Federal |
    Poverty |
      Level |      Freq.     Percent        Cum.
------------+-----------------------------------
    Missing |        274        9.60        9.60
      < 50% |        584       20.46       30.05
50% - <100% |        690       24.17       54.22
  100%-200% |        695       24.34       78.56
      >200% |        612       21.44      100.00
------------+-----------------------------------
      Total |      2,855      100.00
```

REFERENCES

- Lumley, Thomas. *Complex surveys: a guide to analysis using R*. John Wiley & Sons, 2011.
- Lohr, Sharon L. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.
- STATA SURVEY DATA REFERENCE MANUAL (PDF)
  - https://www.stata.com/manuals/svy.pdf
- Gelman, Andrew. "Struggles with survey weighting and regression modeling." (2007): 153-164.
- Bollen, Kenneth A., et al. "Are survey weights needed? A review of diagnostic tests in regression analysis." *Annual Review of Statistics and Its Application* 3 (2016): 375-392.

QUESTIONS?

SARAH SERNAKER
STATISTICIAN

SARAH.SERNAKER@GMAIL.COM

NEXT WEEK…

August 7, 2024
at 12pm (Eastern)

Presenter:
Marianne Kluckman, MPH
RTI, International

Topic:
Approaching NSCAW III for Experienced and New Users