A Methodology Template for Constructing Variables and Applying Survey Weights in Complex Survey Design Studies Using R

Andrew Jergel, MPH[1]; Scott Gillespie, MS, MPH[1]; Lilian Zapata, MD[2]; Kiesha Fraser Doh, MD[3,4]

[1]Pediatric Biostatistics Core, Department of Pediatrics, Emory University, Atlanta, GA, USA.; [2]Emory University School of Medicine, Atlanta, GA, USA.; [3]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA.; [4]Children's Healthcare of Atlanta, Atlanta, GA, USA.

Background

Purpose: Literature and general information relating to complex survey design (CSD) studies are not readily available. The purpose of this is to provide investigators and statisticians with a starting place in working with these databases.

Study Aim: assess the relationship between Healthcare Utilization (HU), Adverse Childhood Experience (ACE) scores, and Gun Violence Exposure (GVE) in Adolescents.

Database: The Future of Families and Child Wellbeing Study (FFCWS) database.

FFCWS Quick Facts

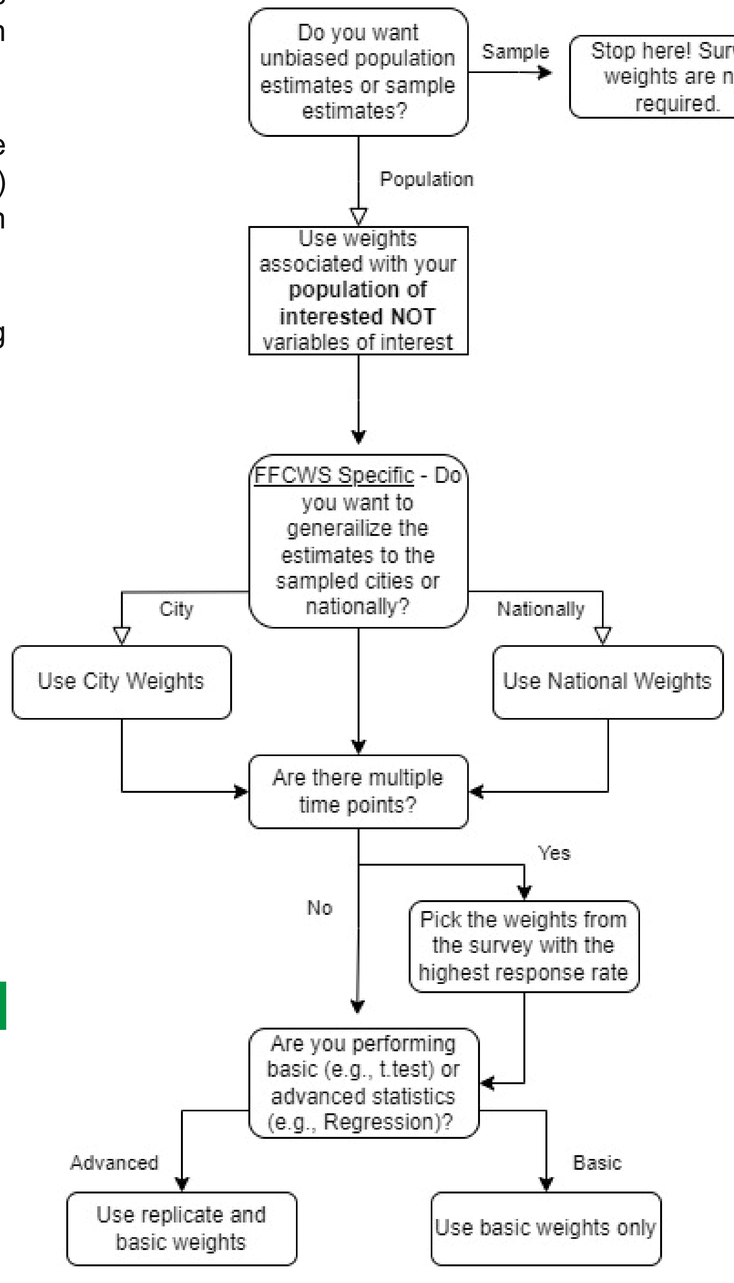
- **Complex longitudinal survey design.**
- **Follows 4,898 teens and their families.**
- **Data currently available for 6 different time points in the focal child's life (Birth and ages 1, 3, 5, 9, 15).**
- **The focal child, both parents (if applicable), primary caregiver, and teachers are surveyed but not at every time point.**
- **National and City Survey Weights for each survey and years are available.**
- **Provide data for GVE from the Gun Violence Archive.**

Tools Utilized and Methods

Overview of Tools and Methods Used:

- R Studio
- Essential R Packages: survey and gtsummary
- Essential Functions: svydesign(), tbl_svysummary() and svyby()
- Statistics: Means with standard deviations (SD), counts with percentages, and 95% Confidence Intervals.

Methods – Survey Weighting



Variable Selection and Construction

Selecting and constructing variables in CSD studies is the same as any other study. The difference is you might be using variables across multiple survey types and time points.

1. **It is OKAY to select variables across multiple survey types and time points if it fits your research question.**
2. **It is OKAY to construct variables using variables obtained from multiple survey types and time points.**

Results

Final Dataset

- 7 final variables constructed from 200+ variables across nine surveys from parents and focal teen from birth, year 5, 9, and 15.
- Used basic national-level survey weights for teens in year 9.
- No subjects are removed yet.

Survey Object

- The "dataset" is converted to a "survey object" using the svydesign() function in the *survey* package. Use svyrepdesign() with replicate weights.
- Subjects who should be excluded are done so in this step using the subset() function as the survey object will adjust survey weights appropriately.

Analysis

- Survey weighted summary and inferential statistics like means (SD) and n (%) can be produced using tbl_svysummary() from *gtsummary* package.
- However, if replicate weights were used other functions like svyby() from *survey* package would need to be used.

Additional Information

Sample Code and Documentation

SCAN ME

FFCWS Information

SCAN ME