Methods

Methods

Simulation studies are widely used in survey statistics to evaluate the performance of different sampling methods and designs under various scenarios. Testing survey weight diagnostic tests requires simulating data under different conditions continuously. The few simulation studies to test the performance of several diagnostic tests either used purely simulated data or existing survey data then alter the selection probabilities by using the survey data as the population to survey from.

The data generating process for this simulation study will utilize existing survey data from the Bureau of Labor Statistics' Consumer Expenditure Survey public use interview data file. The dataset for 2015 contains consumer unit characteristics, assets, and expenditure data for consumers in the United States collected by the Census Bureau for the Bureau of Labor Statistics by interview and diary surveys. For more information, please visit [insert citation for https://www.bls.gov/cex/].

For selecting the endogenous and exogenous variables to simulate, choosing intuitive and strong empirical relationships is key for running the survey weight diagnostic tests. WHYY? Many diagnostic tests assume the unweighted exogenous variables $\vec{X}$ are a significant predictor of the endogenous variable Y.

talk about data talk about tests themselves

Constructing survey weights begin with first determining the selection probability of an unit being selected from the population. Let $\pi_{Si} = \mathbb{P}(i \in S)$ be the selection probability of surveying observation i from sample S. The classic Horvitz-Thompson (HT) estimator of the population total Y is defined as $\hat{Y} = \sum_{i \in S} w_{Si} y_i$ where weight $w_{Si} = \pi_{Si}^{-1}$. In case of non-response, the HT estimator can be further generalized by replacing w_{Si} with $w_i = \pi_i^{-1}$ and $\pi_i = \mathbb{P}(i \in S, i \in F, i \in R)$ with F being the units in the target population within the sampling frame and R being the units in S that respond to the survey.

Since survey weights are used when there is some sort of sampling bias, design-based inference is primarily used in diagnostic tests to accommodate for departures from some model assumptions