ITC 4CV Wave 1 Core Data Subset

Subset the Wave 1 Data to Conduct an Appropriate Statistical Analysis Using the Weighted Data

P. Driezen, M. E. Thompson, G. Li & C. Boudreau

2017-11-14

Introduction

The purpose of this document is two-fold. First, it defines a "core subset" of the ITC 4 Country Smoking and Vaping Survey Wave 1 data set that will serve the purpose of a majority of analyses. This core subset is the full Wave 1 data set with the following exclusions:

1. Respondents recruited as part of the Australian CCV vaper sample are excluded.
2. Long-term quitters are excluded (those who quit smoking more than 2 years before their survey date).
3. Respondents who report having never smoked on a daily basis are excluded.

Authors whose analyses will use only the core subset will be able to use syntax, available on request, to create the core subset. In their papers, they can then refer to this document for the definition of core subset.

Note: the "core subset" should not be confused with the data set "itc4v_core", which is the core linking data set containing identifiers and wave membership variables.

The second purpose is to document how the ITC 4CV data from Wave 1 can be used to conduct an analysis with the existing released data, either by subsetting the data or by conducting a subpopulation (or domain) analysis. Sample code is available on request for SAS, Stata, SUDAAN, and SPSS to create a subset (namely the core subset) of the Wave 1 data. Analysis can then be conducted in one of two ways:

1. Including only those respondents meeting the inclusion criteria in the subset of data and
2. By conducting a subpopulation (or domain) analysis that relies on all respondents sampled, so that complete information is used to estimate variances and standard errors.

Although the second approach is the more correct way to analyze survey data involving subpopulations, the differences between the two approaches tend to be negligible when **all** sampling strata are represented and **most** respondents are included in both data sets. However, with smaller subpopulations, it is recommended that approach #2 is employed to obtain correct variance estimates for estimated means, proportions and regression models. It should also be noted that if the subpopulation is one for which a separate weight is already provided, e.g., cigarette smokers as a whole, for which kWTS201v is the weight, the subpopulation specification is unnecessary.

Applying the Exclusion Criteria and Creating the Data Sets

In the sample code, applying the exclusion criteria and setting up the data proceed as follows:

- Only respondents included in Wave 1 are extracted from the core linking data set ("itc4v_core" where ineM1 = 1)
- Indicator variables are defined to flag respondents meeting the various exclusion criteria, which are:
    1. Respondents recruited as part of the Australian CCV vaper sample are excluded
    2. Long-term quitters are excluded (those who quit smoking more than 2 years ago)
    3. Respondents who report having never smoked on a daily basis are excluded
- A binary indicator variable is then created to flag any respondent meeting one or more of these exclusion criteria. Respondents meeting none of these criteria are included in the analysis (include = 1) while those having at least one exclusion criteria are excluded (include = 0).
- Two separate data sets are created. The first subsets the data by removing those respondents having a value of 0 for the "include" variable. The second subset includes all respondents, so that a subpopulation analysis can be conducted using any of the major statistical software packages (SAS, Stata, SUDAAN, or SPSS).
- For the analysis, the sampling weight is based on a combination of kWTS101v (used for Canada, the US, and England) and kWTS103v (for the Australian **non-CCV** sample)