Guide to weighting harmonized NYTS data for the 1999 and 2000 survey years

# Overview

The National Youth Tobacco Survey (NYTS) is a survey of students in 6th through 12th grade. In the fall of 1999, the first year of NYTS data was fielded as a baseline; the 2000 NYTS was conducted in the spring of the same academic year. Since the 2000 survey, data have been collected approximately every other year. Because the survey uses a stratified cluster sampling design, survey weights and other design information are provided that allow for estimates that are representative of the national population of students.

The NYTS weighting protocol changed between the 2000 and 2002 samples. Analyses must be adjusted in order to combine the data from 1999 or 2000 with later years, or to calculate *total* population size for these individual survey years. This document provides context and guidance for researchers using the harmonized NYTS survey.

In the 1999 and 2000 NYTS samples, the weights were scaled so that the sum of all weights included would equal the sample size of the survey, while surveys fielded in 2002 and later were scaled to the total population targeted by the survey instead. Though this change does not affect calculated proportions using only a single year of data, it does affect analyses that combine multiple years of data across this change in scaling and calculations of weighted totals in these years. The provided weights in 1999 and 2000 can be made comparable to later years of the survey if multiplied by a constant to scale them to the population size.

# Scaling sample weights

Survey weights are commonly scaled to either the sample size or the population size. Scaling the weights to the sample size, as was done in the 1999 and 2000 NYTS, allows for easier comparisons of totals to the original sample size. This may be helpful if, for example, an analyst wants to ensure that an estimate is based on a large enough number of responses. Scaling the weights to the population size, on the other hand, allows for easier statements about the population at large. Rescaling the 1999 and 2000 NYTS requires knowing the target population size of the post-stratified dataset.

# Sources of population data for the NYTS

The NYTS has used two different sources for population data on students since first being fielded in 1999. This population data serves two purposes: first, to create the sample frame used to select students; second, to post-stratify the sample to be representative of all students. From 1999 to 2006, the NYTS used data from the Quality Education Data (QED) database. Beginning in 2009, the data come from the National Center for Educational Statistics (NCES). Because the QED data are not publicly available, we estimate the target population size for the 1999 and 2000 NYTS samples using the NCES dataset.

# Calculating the NYTS total population from NCES Data

The 2011 NYTS methodology documentation includes detailed information on raw counts of schools and students used in post-stratification adjustments, making it a good test case for replication. Using the NCES data, we first estimate the total population targeted by the 2011 NYTS. As outlined by the 2011 NYTS methodology report, we use NCES data from the most recent academic year that precedes the 2011 NYTS: the 2009-2010 Common Core Data Public Elementary/Secondary School Universe data, and the 2007-2008 Private School Universe Survey. After restricting the data to only include "regular"[[1]](#footnote-1) schools, we sum the total of 6th through 12th grade students. Table 1 compares the population total from the 2011 NYTS methodology report with both the MPC calculation and the sum of the weights in the public use 2011 NYTS data. The calculated estimate from the MPC team matches that from the 2011 NYTS methodology report exactly, and differs only slightly from the sum of the weights in the public use 2011 NYTS data. This difference is likely due to sample weight trimming by the NYTS[[2]](#footnote-2).

**Table 1: Population estimates for NYTS 2011**

|  |  |
| --- | --- |
| Population calculated by MPC[[3]](#footnote-3) | 27,395,817 |
| Population from 2011 Methodology | 27,395,817 |
| Sum of NYTS 2011 weights | 27,394,239 |

After confirming that our total count of students exactly match those in the 2011 NYTS methodology report[[4]](#footnote-4), we may employ the same approach for the 1999 and 2000 NYTS samples. While this is likely not the precise methodology employed originally when NYTS was using the QED data, we are confident that our estimates are comparable with the approach and data sources used in 2009-forward NYTS samples. Because the 1999 and 2000 NYTS were fielded in the same academic year, the total population size will be the same for both samples. We use public school data from 1998-1999 academic year, and private school data from the 1997-1998 academic year as these are the most recent NCES data preceding the 1999 and 2000 NYTS samples. We estimate the population size of the NYTS in the 1999-2000 school year to be 25,336,606.

# Validity of Population Estimates

Because the MPC calculation is based on the NCES data, which was not used to establish the total population in the 1999-2006 iterations of the NYTS, it may bias the population size relative to the population calculated by the QED data in these early years. Table 2 compares the sum of weights from the public use NYTS data (based on the private QED dataset) with population estimates calculated by the MPC team using NCES data as outlined above. The differences in calculations for the 2002, 2004, and 2006 samples provide context for the scale of differences between our estimated total population in the 1999 and 2000 NYTS, and the target population actually used by NYTS that we cannot measure.

**Table 2: Comparison of Population Estimates for 2002-2006**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NYTS Sample | Academic Year | Most Recent NCES Public School Data[[5]](#footnote-5) | Most Recent NCES Private School Data[[6]](#footnote-6) | Sum of NYTS Weights | Population estimate from NCES data |
| 2002 | 2001-2002 | 2000-2001 | 1999-2000 | 25,568,633 | 25,941,263 |
| 2004 | 2003-2004 | 2002-2003 | 2001-2002 | 25,531,273 | 27,095,834 |
| 2006 | 2005-2006 | 2004-2005 | 2003-2004 | 26,345,581 | 27,586,677 |

The numbers from the NCES data are consistently larger, but we do not expect this to affect analyses dramatically. Possible sources for these differences include: 1) Differences in the sampling and frequency of coverage between the NCES and the QED data sources and 2) The NYTS practice of weight trimming which reduces the size of large weights. Analysts who desire a consistent total population estimate across all years may choose to rescale the 2002-2006 NYTS samples to the population specified in Table 2 or calculate the total population from the NCES data sources for all years[[7]](#footnote-7).

# Sample Code

We are providing analysts with sample code on how to modify the 1999 and 2000 data for analyses requiring the weights that are rescaled to the population total. The sample code below demonstrates how this rescaling can be done in STATA and R, though the logic is consistent regardless of statistical package. Sample code for SAS and/or SPSS can be provided upon request.

## R

**library(survey)**

# Load dataset

**nyts <- read.csv("nytsall\_integrated.csv")**

# When treating each sample as a separate sample in time

# multiply the weight by population size divided by the

# sample size so the sum of the weights is equal to the

# population size.

**nyts$wtscaled <- nyts$perwt**

**nyts$wtscaled <- ifelse(nyts$year == 1999,**

**nyts$wtscaled \* 25336606 / 15058,**

**nyts$wtscaled)**

**nyts$wtscaled <- ifelse(nyts$year == 2000,**

**nyts$wtscaled \* 25336606 / 35828,**

**nyts$wtscaled)**

**nyts$cigev\_i <- ifelse(nyts$cigev == 2, 1,**

**ifelse(nyts$cigev == 9, NA, 0))**

**nyts\_svy <- svydesign(ids = ~psunum, weight = ~wtscaled,**

**strata = ~stratanum, data = nyts)**

# survey.lonely.psu command allows for variance estimates when

# there is only one psu in a stratum. May not be correct

# for all analyses.

**options(survey.lonely.psu = "adjust")**

# Calculate the total number of students who have ever smoked

# cigarettes over all years of the NYTS.

**svyby(~cigev\_i, ~year, nyts\_svy, svytotal)**

## Stata

// Load dataset

**use "nytsall\_integrated.dta", clear**

// When treating each sample as a separate sample in time

// multiply the weight by population size divided by the

// sample size so the sum of the weights is equal to the

// population size.

**gen wtscaled = perwt /**

**replace wtscaled = wtscaled \* 25336606 / 15058 if year == 1999**

**replace wtscaled = wtscaled \* 25336606 / 35828 if year == 2000**

// singleunit command allows for variance estimates when

// there is only one psu in a stratum. May not be correct

// for all analyses.

**svyset psunum [pweight = wtscaled], strata(stratanum) ///**

**singleunit(centered)**

// Calculate the total number of students who have ever smoked

// cigarettes over all years of the NYTS.

**gen cigev\_i = cigev == 2**

**replace cigev\_i = . if cigev == 9**

**svy: total cigev, over(year)**

1. The "regular" definition excludes special education, vocational, Montessori, private schools with a special emphasis, early childhood programs/child care center and other/alternative schools. This follows guidance from the 2011 NYTS methodology. [↑](#footnote-ref-1)
2. Please see the 2011 NYTS methodology report for more details on NYTS weight trimming [↑](#footnote-ref-2)
3. Sources: Common Core of Data’s Public Elementary/Secondary School Universe Survey Data (2009-2010 v.1a) and Private School Universe Survey (2007-2008) [↑](#footnote-ref-3)
4. Both the total population and the number of students by grade and school-type (public or private) match Table 4-4 of the 2011 NYTS methodology report exactly. [↑](#footnote-ref-4)
5. NCES public school data is available every year so the NYTS used the previous year's survey to form the NYTS sample in 2011. Our estimates also use the previous year's data to try to be comparable to how weights were originally calculated. [↑](#footnote-ref-5)
6. NCES private school data is available every other year. [↑](#footnote-ref-6)
7. By using NCES data directly, analysts may choose to use the NCES data that corresponds to the year of the NYTS, rather than the preceding year. Detailed information on calculating the population from the NCES datasets is available in the 2011 NYTS methodology report. [↑](#footnote-ref-7)