**Sample codes for exploratory data analysis**

**1. Survey design**

**1.1 Population Assessment of Tobacco and Health (PATH) Study**

Samples for the study in Wave 1 were selected using a four-stage, stratified probability sampling design, with replenishment samples used in later waves (Wave 4 and Wave 7) to address attrition. Replicate weights are used in the PATH Study to accurately estimate the variance of survey statistics. For more details on the survey design and replicate weights, refer to the [Population Assessment of Tobacco and Health (PATH) Study Public-Use Files.](https://www.icpsr.umich.edu/web/NAHDAP/studies/36231/datadocumentation)

Create a survey design object using PATH data. For illustration, we used the Wave 1 data. Configure R to calculate the mean square error (MSE) of replicate weights. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use svrepdesign function

library(survey)

# Set option to calculate the mean square error (MSE) of replicate weights

options(survey.replicates.mse = TRUE)

svy\_design <- svrepdesign(

# Specifies the unique identifier for each respondent, ensuring that the survey design accounts for the individual sampling units

id = ~PERSONID,

# Defines the sampling weights for the respondents

weights = ~R01\_A\_PWGT,

# Indicates the replicate weights used for variance estimation

repweights = "R01\_A\_PWGT[1-100]+",

# Specifies the method used for creating the replicate weights. In PATH Study, Fay’s method is used, which is a variant of balanced repeated replication (BRR)

type = "Fay",

# A parameter for Fay's method, which determines the amount of perturbation applied to the replicate weights

rho = 0.3,

# The dataset that contains the survey data and the relevant weight variables

data = PATH\_wave1

)

**1.2 National Health Interview Survey (NHIS)**

NHIS is a cross-sectional household interview survey. For more details on the survey design and weights, refer to the official guide of [National Health Interview Survey](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2023/srvydesc-508.pdf).

Create a survey design object using NHIS survey data. For illustration, we used the 2023 data. Configure R to adjust for lonely primary sampling units (PSUs) and ensure accurate variance estimation. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use svydesign function

library(survey)

# Set option to adjust for lonely primary sampling units (PSUs)

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

# Create a survey design object using the svydesign function

svy\_design <- svydesign(

id = ~fpx, # Primary sampling unit (PSU) variable

strata = ~pstrat, # Stratification variable

weights = ~wtfa\_sa, # Sampling weights

PSU = ~ppsu, # Primary sampling unit (PSU) for complex designs

data = NHIS\_2023, # Your data set

nest = TRUE # The strata and PSUs are nested

)

**1.3 National Survey on Drug Use and Health (NSDUH)**

The sample for the NSDUH was selected using a multistage, deeply stratified sample design. For more details on the survey design and weights, refer to the official guide of [National Survey on Drug Use and Health](https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2022/NSDUH-2022-datasets/NSDUH-2022-DS0001/NSDUH-2022-DS0001-info/NSDUH-2022-DS0001-info-codebook.pdf).

Create a survey design object using NSDUH data. For illustration, we used the 2022 data. Configure R to adjust for lonely primary sampling units (PSUs) and ensure accurate variance estimation. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use svydesign function

library(survey)

# Set options to adjust for lonely primary sampling units (PSUs)

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

# Create a survey design object using the specified survey design parameters

svy\_design <- svydesign(

id = ~verep, # Primary sampling unit (PSU) variable

strata = ~VESTR\_C, # Variance stratum variable

weights = ~ANALWT2\_C, # Sampling weights

data = NSDUH\_2022, # Your data set

nest = TRUE # The strata and PSUs are nested

)

**1.4 Monitoring The Future (MTF)**

For MTF study, a multistage random sampling procedure is used to secure the nationwide sample of 12th graders each year. For more details on the survey design and weights, refer to the official guide of [Monitor The Future](https://monitoringthefuture.org/wp-content/uploads/2022/12/mtf2022.pdf).

Create a survey design object using MTF data. For illustration, we used the 2022 data. Configure R to adjust for lonely primary sampling units (PSUs) and ensure accurate variance estimation. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use the svydesign function

library(survey)

# Set options to adjust for lonely primary sampling units (PSUs)

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

# Create a survey design object using the specified survey design parameters

svy\_design <- svydesign(

id = ~psu, # Primary sampling unit (PSU) variable

strata = ~strata, # Stratification variable

weights = ~weight, # Sampling weights

data = MTF\_2022, # Your dataset

nest = TRUE # Indicates that the strata and PSUs are nested

)

**1.5 National Youth Tobacco Survey (NYTS)**

The National Youth Tobacco Survey (NYTS) is an annual cross-sectional survey of students in

grades 6–12 in U.S. public and private schools. For more details on the survey design and weights, refer to the official guide of [National Youth Tobacco Survey](https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/pdfs/2023_NYTS_Methodology_508.pdf).

Create a survey design object using NYTS data. For illustration, we used the 2022 data. Configure R to adjust for lonely primary sampling units (PSUs) and ensure accurate variance estimation. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use the svydesign function

library(survey)

# Set options to adjust for lonely primary sampling units (PSUs)

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

# Create a survey design object using the specified survey design parameters

svy\_design <- svydesign(

id = ~id, # Primary sampling unit (PSU) variable

strata = ~strata, # Stratification variable

weights = ~weights, # Sampling weights

data = NYTS\_2023, # Your dataset

nest = TRUE # Indicates that the strata and PSUs are nested

)

**1.6 The Tobacco Use Supplement to the Current Population Survey (TUS-CPS)**

The Current Population Survey sample is a multistage stratified sample of approximately 72,000 assigned housing units from 852 sample areas. For more details on the survey design and weights, refer to the official guide of [Current Population Survey](https://www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf).

Create a survey design object using TUS-CPS data. For illustration, we used the 2022 data. Configure R to calculate the mean square error (MSE) of replicate weights. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use svrepdesign function

library(survey)

# Set option to calculate the mean square error (MSE) of replicate weights

options(survey.replicates.mse = TRUE)

# create a survey design object using replicate weights

repweight <- TUSCPS\_2022[, paste0("pwsrwgt", 1:160)]

svy\_design = svrepdesign(

weights = ~PWSRWGT, # sampling weights

repweights = repweight, # replicate weights for TUS-CPS

type = "BRR", # replication method

fay.rho = 0.5, # set the rho parameter for Fay's method

data = TUSCPS\_2022 # your data set

)

**1.7 Youth Risk Behavior Surveillance System (YRBSS)**

The national Youth Risk Behavior Survey (YRBS) uses a three-stage cluster sample design to produce a representative sample of 9th through 12th grade students. For more details on the survey design and weights, refer to the official guide of [Youth Risk Behavior Survey](https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2021/2021_YRBS_Data_Users_Guide_508.pdf).

Create a survey design object using YRBSS data. For illustration, we used the 2021 data. Configure R to adjust for lonely primary sampling units (PSUs) and ensure accurate variance estimation. Modify the code by replacing the placeholders (highlighted in red) with the appropriate variable names from your dataset, as used in our sample code.

# Load the required ‘survey’ package to use the svydesign function

library(survey)

# Set options to adjust for lonely primary sampling units (PSUs)

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

# Create a survey design object using the specified survey design parameters

svy\_design <- svydesign(

id = ~ PSU, # Primary sampling unit (PSU) variable

strata = ~ STRATUM, # Stratification variable

weights = ~ WEIGHT, # Sampling weights

data = YRBS\_2021, # Your dataset

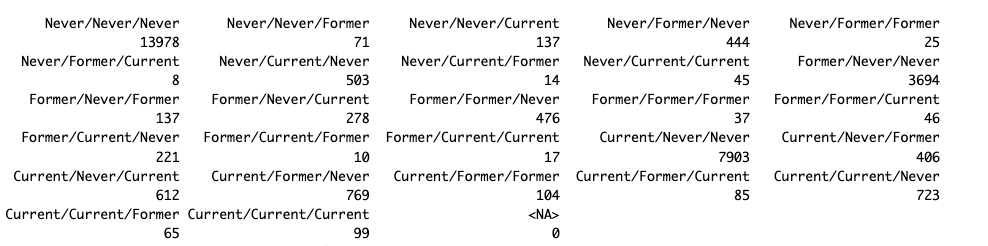
nest = TRUE # Indicates that the strata and PSUs are nested

)

**2. Survey metrics**

Assuming the options and svrepdesign function above have been executed, the following code calculates various survey metrics, including weighted count, weighted percent, standard error of the weighted percent, 95% CI for weighted percent (lower bound of the weighted percent, and upper bound of the weighted percent).

For illustration, *tobacco\_status* is a variable representing a combination of tobacco product use. For cigarettes, cigars, and e-cigarettes, the output will be 27 statuses, each representing a combination of use status for cigarettes, cigars, and e-cigarettes: "Never/Never/Never", "Never/Never/Former", "Never/Never/Current"… For example, “Current/Former/Never” indicates that a person is a current smoker, a former cigar smoker, and has never used e-cigarettes. The following figure displays all 27 combinations of product use statuses.



*var1* and *var2* are placeholders representing categories that you may want to use for stratifying your survey data. For example, var1 is age and var2 sex. These variables will be included as columns in the long format dataset prepared for the analysis.

2.1 Overall survey metrics

Use group\_by() without any additional variables and handle var1 and var2 as "Overall".

library(srvyr)

# Calculate overall survey metrics by grouping the data by tobacco use status combinations

# Compute weighted counts, percentages with 95% confidence intervals, and unweighted counts

# Add overall indicators for sex and age

# Convert the survey design object to a survey design object

metrics\_all <- as\_survey(svy\_design) %>%

# Group by tobacco use status combinations, and ensures that all levels are included

group\_by(tobacco\_status, .drop = FALSE) %>%

# Calculate the outputs for each group

summarize(

# Calculate the weighted count

"weighted count" = survey\_total(),

# Calculate the weighted percent, including the standard error and 95% confidence interval

"weighted percent" = survey\_mean(vartype = c("se", "ci"), level = 0.95, proportion = TRUE),

# Calculate the unweighted count

"unweighted\_count" = unweighted(n())

) %>%

# indicate that the metrics are for which group, here for the overall population

mutate(var1 = "Overall", var2 = "Overall") %>%

ungroup()

2.2 Survey metrics by *var1*

# Calculation is similar to the overall survey metrics

# But it additionally groups the data by age categories, providing outputs for each age group within each tobacco use status combination

metrics\_by\_var1 <- as\_survey(svy\_design) %>%

group\_by(var1, tobacco\_status, .drop = FALSE) %>%

summarize(

"weighted count" = survey\_total(),

"weighted percent" = survey\_mean(vartype = c("se", "ci"), level = 0.95, proportion = TRUE),

"unweighted\_count" = unweighted(n())

) %>%

mutate(var2 = "Overall") %>%

ungroup() %>%

rename(var1\_name = var1)

2.3 Survey metrics by *var1* and *var2*

# Calculation is similar to the overall survey metrics

# But it additionally groups the data by age and sex, providing outputs for each combination of sex, age, and tobacco use status within each tobacco use status combination

metrics\_by\_var1\_var2 <- as\_survey(svy\_design) %>%

group\_by(var1, var2, tobacco\_status, .drop = FALSE) %>%

summarize(

"weighted count" = survey\_total(),

"weighted percent" = survey\_mean(vartype = c("se", "ci"), level = 0.95, proportion = TRUE),

"unweighted\_count" = unweighted(n())

) %>%

ungroup() %>%

rename(var1\_name = var1, var2\_name = var2)

**3. Export data**

Combine the survey metrics and recode for readability.

library(dplyr)

library(tidyr)

# Combine survey metrics into a data frame

data <- rbind.data.frame(metrics\_all, metrics\_by\_var1, metrics\_by\_var1\_var2)

# Select relevant columns and recode for readability

overall\_metrics <- data %>%

dplyr::select(tobacco\_status, var1, var2,

"weighted count", "unweighted\_count", "weighted percent",

"weighted percent\_se", "weighted percent\_low", "weighted percent\_upp") %>%

# Filters out rows with zero unweighted counts

filter(unweighted\_count > 0)