

Complex Sampling Weights
and
Preparing 2020 BRFSS Module Data for Analysis

July 2021



Overview

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based telephone survey that collects data on a number of health outcomes, health-related risk behaviors, use of preventive services, and chronic conditions from noninstitutionalized adults who reside in each of the states and participating US territories. In 2011, BRFSS changed its data collection procedures, structure, and weighting methodology so the traditional landline telephone-based data set could include data from participants using cell phones. The BRFSS uses a core set of questions and gives states the option to include multiple modules that focus on specific health issues. Not all optional modules are collected by all of the states, and states may opt to collect module data from just subsets of their survey participants. Core questions are used during all interviews. Researchers using BRFSS data should conduct analyses with complex sampling procedures; they also should appropriately stratify and weight the data in their work. Potential bias resulting from selection probabilities and noncoverage among segments of the population can be reduced through weighting.

Researchers conducting analysis of variables from the core-only section should use the variable `_LLCPWT` for weighting.

BRFSS has created this document to guide users analyzing variables from 2020 BRFSS optional modules and/or combinations of module and core variables. The 2020 BRFSS data sets include data that respondents provided by landline telephone or cell phone. Data users should note that newer weighting procedures are likely to affect trend lines when comparing BRFSS data collected up to 2010 with data from 2011 and afterward; because of these changes, users are advised NOT to make direct comparisons with pre-2011 data, and instead, begin new trend lines with that year.

Data users should become familiar with the information presented in this document prior to performing analyses. Information about the changes to 2020 BRFSS data collection is available in the [2020 Data Comparability Report](#).

Weighting BRFSS Core Data

BRFSS has grouped the 2020 data into four data sets:

- 2020 BRFSS Questionnaire data (Combined Landline Telephone and Cellular Telephone)

And three versions

- 2020 BRFSS Combined Landline Telephone and Cellular Telephone Version 1
- 2020 BRFSS Combined Landline Telephone and Cellular Telephone Version 2
- 2020 BRFSS Combined Landline Telephone and Cellular Telephone Version 3

Researchers should understand that BRFSS data were collected by asking core and/or module questions on the annual questionnaire. Users, therefore, may need to use data from up to 4 different data sets, described below.

- Use the Combined Landline Telephone and Cell phone data if questions are exclusively from the core section *or* if questions are in common modules that were asked on both the landline survey and cellphone survey (*Note: See the section below, Using BRFSS Data from Multiple Data Sets, for more about common modules.*)
- Use Version 1, Version 2, and/or Version 3 Combined Landline Telephone and Cell Phone data sets if states collected multiple version questionnaires.

In all cases, researchers conducting complex sampling analyses from the *core-only* section should use the variable **_LLCPWT** for weighting, **_STSTR** for stratification, and the variable **_PSU** for primary sampling unit, or clustering for short. The following table displays the description of the data, names of the data sets, and the variable names of the final weight.

Data description	Data set name	Final weight variable name
Combined Landline Telephone and Cellular Telephone	LLCP2020	_LLCPWT
Combined Landline Telephone and Cellular Telephone Version 1	LLCP20V1	_LCPWTV1
Combined Landline Telephone and Cellular Telephone Version 2	LLCP20V2	_LCPWTV2
Combined Landline Telephone and Cellular Telephone Version 3	LLCP20V3	_LCPWTV3

The examples below demonstrate how to use **_LLCPWT**, **_STSTR** & **_PSU** when analyzing variable **HAVARTH4** (*Ever told you had some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia*) and **PHYSHLTH** (*how many days during the past 30 days was your physical health not good?*) in the core section using the 2020 Louisiana data set. Examples are given in both SAS and R.

Example in SAS:

```
LIBNAME IN '\BRFSS\';           /* Use LIBNAME to set the location of the BRFSS data set */

PROC SURVEYMEANS                /* Calculate average number of physical healthy days */
  DATA = IN.LLCP2020 (WHERE=(_STATE= 22)); /* FIPS code for Louisiana is 22 */
  WEIGHT _LLCPWT;               /* Use _LLCPWT because question is from the core section */
  STRATA _STSTR;
  CLUSTER _PSU;
  VAR PHYSHLTH;
RUN;

PROC SURVEYFREQ;               /* Calculate percentage of population with arthritis */
  DATA = IN.LLCP2020(WHERE= (_STATE = 22));
  WEIGHT _LLCPWT;               /* Use common weight because question is from the core section */
  STRATA _STSTR;
  CLUSTER _PSU;
  TABLE HAVARTH4;
RUN;
```

Example in R:

```

# Install the survey package if it is not already installed
install.packages("survey")

# Call the library for the current R session
library(survey)

# Read in BRFSS data
load("\\BRFSS\\BRFSS.rdata")

# Subset the data for Louisiana
BRFSS <- BRFSS[BRFSS$state == 22, ]

# Set options for allowing a single observation per stratum
options(survey.lonely.psu = "adjust")

# Create survey design
brfssdsgr <- svydesign(
  id=~1,
  strata = ~ststr,
  weights = ~llcpwt,
  data = BRFSS)

# calculate average number of physical healthy days
# Variable to analyze
svymean(~physhlth,
  brfssdsgr,
  na.rm = TRUE)

# calculate percent in each arthritis category
svymean(~factor(havarth4),
  brfssdsgr,
  na.rm = TRUE)

```

Using BRFSS Data from Multiple Data Sets

Due to the complex nature of state-based data collection processes, users may have to create a data set that fits their research needs. The following two examples illustrate how to prepare module data for analysis when states have collected module data in a variety of ways.

Prior to conducting analyses using optional module data, users should always review all relevant documents for the given data year, such as the [Overview](#), [Codebook](#), [Comparability of Data](#), and [Modules by State](#) in particular to identify states that collected data of interest and determine which questionnaire version states used. Remember to check FIPS state codes in each data set to avoid duplication.

Individual states may have chosen to use a number of optional modules, depending on each state's needs. Individual states may also have chosen to divide their samples and use different modules in the

subsamples that were distinguished by version of the surveys. **Modules that appeared in each version of a state's questionnaire are called common modules.** Each state that collected common modules did so on all its versions of combined landline telephone and cell phone.

Example 1: Cancer Survivorship: Type of Cancer

The example below uses the module **Cancer Survivorship: Type of Cancer** to demonstrate how to combine and reweight data from multiple data sets for analysis. This module was selected because states collected the data either as a common module in the combined landline telephone and cell phone or by splitting the sample and offering the module on one or more versions of the data collection. For ease in using the 2020 optional modules, please refer to the two documents published with the annual data set: [Module weights for 2020](#) and [States by Module](#). In addition to showing which states used which modules, these tables list names and descriptions of the data sets, versions, and weight variables from the corresponding data sets. If users want to analyze variables from certain modules, these two documents are a good place to start. The following information has been copied from the two module documents to show segments of the tables:

Module weights for 2020

State	Description	Dataset	Data weight	Module(s)
Maryland	Combined Land Line and Cell Phone data	LLCP2020	_LLCPWT	E-Cigarettes, Marijuana Use, Pre-Diabetes
	Combined Land Line and Cell Phone data, version 2	LLCP20V2	_LCPWTV2	Cancer Survivorship: Course of Treatment, Cancer Survivorship: Pain Management, Cancer Survivorship: Type of Cancer
	Combined Land Line and Cell Phone data, version 3	LLCP20V3	_LCPWTV3	Adverse Childhood Experience

States by Module

Module	Description	Dataset	Data weight	State(s)
Cancer Survivorship: Type of Cancer	Combined Land Line and Cell Phone data	LLCP2020	_LLCPWT	Arizona, Connecticut, Delaware, Georgia, Guam, Hawaii, Indiana, Louisiana, Massachusetts, Michigan, Mississippi, Missouri, Montana, New Jersey, New Mexico, North Carolina, Rhode Island, South Dakota, Utah, Vermont, Virginia, Wisconsin
	Combined Land Line and Cell Phone data, version 1	LLCP20V1	_LCPWTV1	Kansas, Maine, Nebraska, New York, Ohio, Oklahoma
	Combined Land Line and Cell Phone data, version 2	LLCP20V2	_LCPWTV2	Maryland
	Combined Land Line and Cell Phone data, version 3	LLCP20V3	_LCPWTV3	New York

- 1) The module **Cancer Survivorship: Type of Cancer** listed in the 2020 document [States by Module](#) shows that there are 29 states that collected data on this module, 22 of them used the

common version of the module, 6 states used version 1, 1 state used version 2 and 1 state used version 3 of the module. See detailed information below:

- a. Twenty-two states collected this data as a common module; therefore, we need to extract data for these states from LLC2020. They are:
Arizona, Connecticut, Delaware, Georgia, Guam, Hawaii, Indiana, Louisiana, Massachusetts, Michigan, Mississippi, Missouri, Montana, New Jersey, New Mexico, North Carolina, Rhode Island, South Dakota, Utah, Vermont, Virginia, and Wisconsin.
- b. 6 states, *Kansas, Maine, Nebraska, New York, Ohio, and Oklahoma* used version 1 of the module. We therefore need to extract data from LLC20V1 for these 6 states.
- c. *Maryland* used version 2 of the module, we will extract data from LLC20V2 for Maryland.
- d. *New York*, in addition to using version 1, also used version 3 of the module; therefore, we need to extract data from LLC20V3 for New York.

2) Obtain data sets:

- To download the Combined Landline Telephone and Cellular Telephone data set (LLC2020):
 - Go to the data [webpage](#), and choose the 2020 data set. From there, read all of the 2020 documents including the [overview](#), [codebook](#), and the [module list](#). The data set is available in ASCII and SAS transport formats. Record layout, formats, and SAS code are included to read-in the data.
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 1 data set:
 - Go to the [webpage](#) with Combined Landline and Cellular Telephone Survey Multiple Questionnaire Version Data. There are 3 separate SAS data sets corresponding to Questionnaire Version 1, 2, and 3. Choose Version 1: [LLC20V1](#).
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 2 data set:
 - Follow the same step above for Version 1, but instead, choose Version 2: [LLC20V2](#).
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 3 data set:
 - Follow the same step above for Version 1, but instead, choose Version 3: [LLC20V3](#)

3) Generate a working data set:

Create a new and uniform final weight variable from each version of the data sets that were downloaded in steps above, combine data into one working data set. This step will require users to do the following:

- keep all the states that collected the module data in each of the data sets,
- rename the corresponding weight variable to a new and consistent weight variable,
- combine all the data sets into one that contains the renamed weight variable. Details are shown in the SAS syntax below.

Note: Since New York used Version 1 and Version 3 of the module, the weights from each version of the data set need to be adjusted by checking the sample size for each version.

Example in SAS:

```
LIBNAME IN 'BRFSS\';
```

**Extract data from landline and cell phone combined dataset, begin with the main dataset, where the module Cancer Survivorship: Type of Cancer was used as a common module in the landline and cellular telephone samples;*

```
DATA LLCPP; /*Select 22 states that used Cancer Survivorship: Type of Cancer as a common module; */ SET
IN.LLCP2020 (WHERE=( _STATE IN (4,9,10,13,15,18,22,25,26,28,29,30,34,35,37,44,46,49,50,51,55,66)));
RENAME _LLCPWT=_FINALWT;
RUN;
```

```
DATA LLCPPV1; /*Kansas, Maine, Nebraska, New York, Ohio, Oklahoma used version 1;
SET IN.LLCP20V1 (WHERE=( _STATE IN (20,23,31,36,39,40)));
RENAME _LCPWTV1=_FINALWT;
RUN;
```

```
DATA LLCPPV1 LLCPPV1_NY; *Subset V1 data for NY from LLCPPV1 from the step above in order to reweight V1 & V3
NY data;
SET LLCPPV1;
IF _STATE=36 THEN OUTPUT LLCPPV1_NY;
ELSE OUTPUT LLCPPV1;
RUN;
```

```
DATA LLCPPV2;
SET IN.LLCP20V2 (WHERE=( _STATE IN (24))); *Maryland used version 2;
RENAME _LCPWTV2=_FINALWT;
RUN;
```

```
DATA LLCPPV3_NY;
SET IN.LLCP20V3 (WHERE=( _STATE IN (36))); *New York used version 3;
RENAME _LCPWTV3=_FINALWT;
RUN;
```

Since New York used version 1 and version 3 of the module, we need to manually adjust the Weight for New York's data follow these steps:

- 1) Find the sample size (N1) for LLCPPV1_NY (LandlineV1 + CellphoneV1, $N1 = (1921+2634) = 4555$)
- 2) Find the sample size (N3) for LLCPPV3_NY (LandlineV3 + CellphoneV3, $N3 = (2059+2677) = 4736$)
- 3) Combine sample sizes for LLCPPV1_NY and LLCPPV3_NY, $(N1+N3 = (4555+4736) = 9291)$

- 4) Use the sample size from each version divided by the combined sample size to get proportions, therefore, the proportion for version 1 of NY data is **P1** = $N1/(N1+N3) = 4555/9291 = 0.49$
the proportion for version 3 of NY data is **P3** = $N3/(N1+N3) = 4736/9291 = 0.51$

Since the proportions are so close, we can either use the calculated P1 & P3 or divide the _FINALWT by 2.

The SAS codes for the weight adjustment is:

```
DATA LLCPNY(RENAME=(NYWT=_FINALWT)); *Rename the adjusted weight var NYWT to _FINALWT;
SET LLCPV1_NY(IN=A)
    LLCPV3_NY(IN=B);
IF A THEN NYWT= (_FINALWT * 0.49);
ELSE IF B THEN NYWT= (_FINALWT * 0.51);
DROP _FINALWT; * drop the unadjusted _FINALWT;
RUN;

DATA CANTERTYPE; *Combine all datasets which will be used to analyze the data;
SET LLCPV1 LLCPV2 LLCPNY; *LLCPNY contains V1 & V3 of NY DATASETS with adjusted WEIGHT;
RUN;
```

Example in R:

Begin with the main data set for states that used the Cancer Survivorship: Type of Cancer as a common module.

use FIPS state codes for the 22 states that used common version of the module

```
llcp <- llcp2020[llcp2020$state %in% c(4,9,10,13,15,18,22,25,26,28,29,30,34,35,37,44,46,49,50,51,55,66), ]
```

Rename the final weight variable for the main data set

```
llcp$finalwt <- llcp$llcpwt
```

Choose the states that used version 1 of the module – Kansas, Maine, Nebraska, New York, Ohio, Oklahoma

```
llcpv1 <- llcp20v1[llcp20v1$state %in% c(20,23,31,39,40), ]
```

Rename the final weight variable for the version 1 data set

```
llcpv1$finalwt <- llcpv1$llcpwtv1
```

Extract data for NY version 1 data set

```
llcpv1ny <- llcp20v1[llcp20v1$state %in% c(36), ]
```

Rename & adjust the final weight variable for the version 1 NY data set

```
llcpv1n$finalwt <- llcpv1$llcpwtv1 * (0.49)
```

Choose the states that used version 2 of the module – Maryland (Fips code=24)

```
llcpv2 <- llcp20v2[llcp20v2$state %in% c(24), ]
```

Rename the final weight variable for the version 2 data set

```
llcpv2$finalwt <- llcpv2$llcpwtv2
```

Choose the states that used version 3 of the module – New York

```
llcpv3ny <- llcp20v3[llcp20v3$state %in% c(36), ]
```



```
# Rename and adjust the final weight variable for NY the version 3 data set
llcpv3ny$finalwt <- llcpv3$llcpwtv3* (0.51)
```

```
# Combine all data sets
cancertype <- rbind(llcp, llcpv1, llcpv1ny, llcpv2, llcpv3ny)
```

Example 2: Pre-Diabetes

The example below uses the Pre-Diabetes Module to demonstrate how to combine and reweight data from multiple data sets.

- 1) The module **Pre-Diabetes** listed in the document [States by Module](#) shows that there are total of 43 states used the Pre-Diabetes optional module in 2020.

- 37 states used the common version of this module; therefore, we need to extract data for these 37 states from the dataset LLC2020. These states are:

Alabama, Alaska, Arizona, California, Connecticut, Delaware, Florida, Georgia, Guam, Hawaii, Idaho, Illinois, Indiana, Iowa, Kentucky, Maine, Maryland, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, North Dakota, Oregon, Pennsylvania, Puerto Rico, Utah, Virgin Islands, Virginia, West Virginia, Wisconsin, Wyoming

- 4 states used version 1 of the module; therefore, we need to extract data from LLC20V1 for these 4 states *Kansas, Massachusetts, Ohio, Oklahoma*.
- 2 states used version 2 of the Pre-Diabetes module and therefore, we need to extract data from LLC20V2 for these 2 states, *Colorado & New York*.
- *New York* also used version 3 of the Pre-Diabetes module and therefore, we need to extract data for *New York* from LLC20V3.

Note: New York used both version 2 & version 3 of the Pre-Diabetes module for 2020. The weight variable will need to be adjusted for New York before we combine and analyze the data. We shall follow the same steps for weight adjustment as we did in example 1.

2) Obtain data sets:

- To download the Combined Landline Telephone and Cellular Telephone data set (LLC2020):
 - Go to the data [webpage](#), and choose the 2020 data set. From there, read all of the 2020 documents including the [Overview](#), [Codebook](#), and the [States by Module](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read-in the raw data.

- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 1 dataset:
 - Go to the webpage with Combined Landline and Cellular Telephone Survey Multiple Questionnaire Version Data. There are 3 separate SAS data sets corresponding to Questionnaire Version 1, 2, and 3. Choose Version 1: [LLCP20V1](#).
 - To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 2 data set:
 - Follow the same step above for Version 1, but instead, choose Version 2: [LLCP20V2](#).
 - To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 3 data set:
 - Follow the same step above for Version 1, but instead, choose Version 3: [LLCP20V3](#).
- 3) **Generate a working data set:** Create a new and uniform final weight variable from each of the data sets, combine data into a working data set. This step will require users to do the following:
- a. Keep all of the states that collected the module data in each of the data sets
 - b. Rename the corresponding weight variable to a consistent weight variable and adjust the weight variable if any state uses more than 1 version of the module;
 - c. Combine all the data sets into one that contains the renamed weight variable. Details are shown in the code below

Example code in SAS:

```
LIBNAME IN '\BRFSS\';
```

Begin with the main dataset, where **Pre-Diabetes was used as a common module in the combined landline and cellular telephone samples for 37 states: Alabama, Alaska, Arizona, California, Connecticut, Delaware, Florida, Georgia, Guam, Hawaii, Idaho, Illinois, Indiana, Iowa, Kentucky, Maine, Maryland, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, North Dakota, Oregon, Pennsylvania, Puerto Rico, Utah, Virgin Islands, Virginia, West Virginia, Wisconsin, Wyoming;*

```
DATA LLCP;
SET IN.LLCP2020 (WHERE=(_STATE in
(1,2,4,6,9,10,12,13,15,16,17,18,19,21,23,24,26,27,28,29,30,31,32,33,34,35,38,41,42,49,51,52,54,55,56,66,72)));
RENAME _LLCPWT=_FINALWT;      *Rename weight variable to be consistent across new data sets;
run;
```

**Extract data for states that used Questionnaire V1;*

```
DATA LLCPV1;
SET IN.LLCP20V1 (WHERE=(_state IN (20,25,39,40)); * Kansas, Massachusetts, Ohio, Oklahoma;
RENAME _LCPWTV1 = _FINALWT;      *Rename weight variable to be consistent across new data sets;
RUN;
```

**Extract data for states that used Questionnaire V2;*

```
DATA LLCPV2;
SET IN.LLCP20V2 (WHERE=(STATE IN (8,36)); * Colorado, New York;
RENAME _LCPWTV2 = _FINALWT; *Rename weight variable to be consistent across new data sets.
RUN;
```

**Subset NY data from V2 datasets & prepare for weight readjustment, LLCPV2 contains only Colorado data;*

```
DATA LLCPV2_NY LLCPV2;
SET LLCPV2;
IF STATE=36 THEN OUTPUT LLCPV2_NY; * New York;
ELSE OUTPUT LLCPV2; * Colorado;
RUN;
```

**Extract data for states that used Questionnaire V3;*

```
DATA LLCPV3_NY;
SET IN.LLCP20V3 (WHERE=(STATE IN (36)); *New York;
RENAME _LCPWTV3 = _FINALWT; *Rename weight variable to be consistent across new data sets;
RUN;
```

To manually adjust the Weight for V2 & V3 Pre-Diabetes module for New York, follow these steps.

- 1) Find the sample size (N2) for LLCPV2_NY (LandlineV2 + CellphoneV2, $N2 = (1981 + 2752) = 4733$)
- 2) Find the sample size (N3) for LLCPV3_NY (LandlineV3 + CellphoneV3, $N3 = (2059 + 2677) = 4736$)
- 3) Combine sample sizes for LLCPV2_NY and LLCPV3_NY, $(N2 + N3 = (4733 + 4736) = 9469)$
- 4) Use the sample size from each version divided by the combined sample size to get proportions, therefore, the proportion for version 2 of NY data is $P2 = N2 / (N2 + N3) = 4733 / 9469 = 0.50$
the proportion for version 3 of NY data is $P3 = N3 / (N2 + N3) = 4736 / 9469 = 0.50$

Sample sizes from step 1 & 2 are so similar for version 2 & version 3, we can simply adjust the _FINALWT by dividing it by 2.

The SAS codes for the weight adjustment is:

**LLCPNY contains v2 & v3 of Pre-Diabetes data for NY with the adjusted weight;*

```
DATA LLCPPNY(RENAME=(NYWT=_FINALWT)); *Rename the adjusted weight var NYWT to _FINALWT;
SET LLCPV2_NY
    LLCPV3_NY;
NYWT= (_FINALWT/2 );
DROP _FINALWT; * drop the unadjusted _FINALWT;
RUN;
```

**Combine all datasets together;*

```
data PREDIAB;
SET LLCP LLCPV1 LLCPV2 LLCPPNY;
RUN;
```

Dataset PREDIAB is ready to be analyzed.

Example in R:

Begin with the main data set for 37 states that used the Pre-Diabetes module as a common module

```
llcp <- llcp2020[llcp2020$state %in% c(1,2,4,6,9,10,12,13, 15,16,17,18,19,21,23,24,26,
27,28,29,30,31,32,33,34,35,38,41,42,49,51,52,54,55,56,66,72), ]
```

```
# Extract Questionnaire V1 data for Kansas, Massachusetts, Ohio, Oklahoma
nrow(llcp20v1[llcp20v1$state %in% c(20,25,39,40), ])
```

```
# Extract Questionnaire V2 data for Colorado
nrow(llcp20v2[llcp20v2$state %in% c(8), ])
```

```
# Extract Questionnaire V2 data for New York
nrow(llcp20v2ny[llcp20v2$state %in% c(36), ])
```

```
# Extract Questionnaire V3 data for New York;
nrow(llcp20v3ny[llcp20v3$state %in% c(36), ])
```

```
# Rename the final weight variable for the main data set
llcp$finalwt <- llcp$llcpwt
```

```
# Rename the final weight variable for the version 1 data set
llcpv1$finalwt <- llcpv1$lcpwtv1
```

```
# Rename the final weight variable for the version 2 data set
llcpv2$finalwt <- llcpv2$lcpwtv2
```

```
# Rename the final weight variable for the version 2 NY data set
llcpv2ny$finalwt <- llcpv2ny$lcpwtv2 * (.50)
```

```
# Rename the final weight variable for the version 3 data set
llcpv3ny$finalwt <- llcpv3ny$lcpwtv3 * (.50)
```

```
# Combine all data sets and it's ready for analysis
fampln <- rbind(llcp, llcpv1, llcpv2, llcpv2ny, llcpv3ny)
```

Combining Multiple Years of BRFSS Data

There are times we need to analyze data from multiple years, such as calculating the average prevalence of a variable in core sections for 2018 and 2020. In this case, we need to combine data from 2018 and 2020. The weight variable for 2018 and 2020 needs to be adjusted proportionally for the combined data set. There are two ways to adjust the weight variable. The sample size for each year helps to determine how we should adjust them. If the sample sizes are very similar, we can simply divide the original weight variable by 2. If they are different, we need to adjust the weight variable proportionally by following these steps:

- 1) Find out the sample size for 2018 and 2020
- 2) Sum up the total sample size for 2018 and 2020
- 3) Find the proportion for 2018 (use the sample size for 2018 divided by the combined sample size for 2018 & 2020)
- 4) Find the proportion for 2020 (Use the sample size for 2020 divided by the combined sample size for 2018 & 2020)
- 5) The adjusted weight for 2018 is ($_LLCPWT(2018) * \text{proportion for 2018}$)

- 6) The adjusted weight for 2020 is $(_LLCPWT(2020) * \text{proportion for 2020})$

The example below uses the 2018 and 2020 data from New York to demonstrate how to combine and reweight data on the core section from multiple years.

For 2018, the sample size for New York is 35,767. For 2020, the sample size is 14,769. The sample size for 2018 is much larger than that for 2020, therefore the weights need to be adjusted proportionally. Follow the steps listed above; we can find adjusted weights for New York for 2018 & 2020.

- 1) The sample size is 35,767 for 2018, and 14,769 for 2020
- 2) The combined sample size is 50,536
- 3) The proportion for 2018 is $35,767/50,536 = .71$
- 4) The proportion for 2020 is $14,769/50,536 = .29$;
- 5) The adjusted weight for 2018 is $(_LLCPWT(2018) * .71)$
- 6) The adjusted weight for 2020 is $(_LLCPWT(2020) * .29)$

Obtain data sets:

- To download the 2018 Combined Landline Telephone and Cellular Telephone Questionnaire data set (LLCP2018):
 - Go to the data [Webpage](#), and choose the 2018 data set. From there, read all of the 2018 documents including the [Overview](#), [Codebook](#), and the [Modules By State](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read-in the raw data.
- To download the 2020 Combined Landline Telephone and Cellular Telephone Questionnaire data set (LLCP2020):
 - Go to the data [Webpage](#), and choose the 2020 data set. From there, read all of the 2020 documents including the [Overview](#), [Codebook](#), and the [Modules By State](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read-in the raw data.
- Generate a new and uniform final weight variable (the adjusted weight variables) from each of the data sets and combine the data into one working data set. This step will require users to:
 - Rename the corresponding weight variable to a consistent weight variable; and
 - Combine all the data sets into one data set that contains the consistent weight variable.

Again, users should note that new weighting procedures are likely to affect trend lines when comparing BRFSS data collected before and after 2011; because of the changes, users are advised NOT to make direct comparisons with pre-2011 data, and instead, to begin new trend lines with that year.

Reweighting Example in SAS:

```
Libname D18 "\BRFSS2018";
Libname D20 "\BRFSS2020";
```

```
/* Check sample size for each 2018 & 2020 */
```

```
DATA NY18;
  SET D18.LLCP2018(WHERE=(_STATE=36));
  RUN;
  /* NOTE: There were 35767 observations read from the data set D18.LLCP2018  WHERE _STATE=36*/
```

```
DATA NY20;
  SET D20.LLCP2020(WHERE=(_STATE=36));
  RUN;
  /* NOTE: There were 14769 observations read from the data set D20.LLCP2020  WHERE _STATE=36*/
```

```
  /* Create weighted dataset with the adjusted weight variable (FINALWT) for each year */
```

```
DATA COMBINEYRS;
  SET D18.LLCP2018 (IN=A)
      D20.LLCP2020 (IN=B);
  WHERE _STATE=36;
  IF A THEN FINALWT = LLCPPWT * .71;
  IF B THEN FINALWT = LLCPPWT * .29;
  RUN;
```

Reweighting Example in R:

```
# Subset 2018 and 2020 data for New York
NY18 <- llcp2018[llcp2018$state == 36, ]
NY20 <- llcp2020[llcp2020$state == 36, ]

# Count number of observations in each year
nrow(NY18)
nrow(NY20)

# Create a final weight based on proportion of observations in each year
NY18$finalwt <- NY18$llcpwt * (.71)
NY20$finalwt <- NY20$llcpwt * (.29)

# Combine the data sets and reweight
NY1820 <- rbind(NY18, NY20)
```

~~~ END~~~