An Introduction to Analyzing the NAWS Public Access Data

March 2018

Release 1.1

Contents

Overview on Conducting Analysis with the NAWSPAD

This document provides an introduction to the NAWS Public Access Data (NAWSPAD), alerting the user to statistical issues related to the National Agricultural Workers Survey's (NAWS) complex sampling design and outlining available options for addressing them. The NAWS uses a stratified multi-stage sampling design to account for seasonal and regional fluctuations in the level of employment in crop agriculture. The NAWS has seven levels of sampling. The first two levels are to create 36 strata determined by three interviewing cycles and 12 agricultural regions. The next level consists of the primary sampling unit (PSU) within each stratum. The PSU is the county cluster (Farm Labor Area). The next four levels of sampling occur within the county cluster. Within each county cluster, the NAWS draws random samples of counties, ZIP Code regions, and employers. At the final level of sampling, workers are randomly sampled within employers. Further information on the NAWS sampling design is available in the *Statistical Methods of the National Agricultural Workers Survey* available on the NAWS website (https://www.doleta.gov/naws/).

The NAWS's multi-level sampling design adds complexity to data analysis in two ways:

1. To produce accurate statistics, users need to apply sampling and post-sampling weights to account for differences in sampling probabilities and to correct for non-response at the region and cycle levels.
2. Users who wish to calculate the standard error of point estimates need to account for the complex multi-stage sampling design. The NAWSPAD supports using the balanced repeated replication (BRR) method. This method does not require information on the strata or PSU, which are suppressed from the NAWSPAD for privacy reasons.

Software Options for Analyzing the NAWS Data

The NAWSPAD files are available in SAS, Microsoft Excel, and CSV (comma separated values) formats. These files can be read into other data analysis programs, such as SPSS and Stata. NAWS users who do not have access to analysis software such as SAS, SPSS, R, or Stata, or users who only need to quickly calculate the mean or proportion of specific variables, can perform these calculations using the NAWS Excel files. In addition, there are several documents to support users, including codebooks, questionnaires, and explanation of the NAWS methodology. A list of these documents is provided below.

Calculate Weighted Means and Proportions

Users with access to statistical software such as SAS, SPSS, R, or Stata should consult the respective manuals for how to calculate weighted means and proportions using the weighting variable PWTYCRD. MS Excel users can calculate weighted means and proportions using the instructions in Chapter 1 “How to Calculate Weighted Means and Weighted Proportions from NAWS Public Access Data Using Excel.” This chapter provides step-by-step instructions on performing the calculations. The process can be tedious if analyzing many variables but does provide the same means and proportions produced by statistical software.

Calculate Design-Corrected Standard Errors

The NAWS provides a BRR method to estimate design-corrected standard errors using common statistical software programs. This method involves generating replicates, then calculating the point estimates for each replicate and the variance of the replicate estimates. This variance is the estimated sampling variance of the statistic of interest. Chapter 2 contains additional explanation of the BRR method.

This document is organized into two chapters; each chapter provides specific guidance and detailed explanations:

- Chapter 1: How to Calculate Weighted Means and Weighted Proportions using Excel
- Chapter 2: Using Replicate Weights and Calculating Design-Corrected Standard Errors

Additional Documentation to Support NAWSPAD Users

The following additional documentation about the NAWS and the NAWSPAD may be useful. These documents are available on the NAWS website (https://www.doleta.gov/naws/).

- NAWSPAD Variables and Labels
- The NAWSPAD codebook
- Supporting statement [Part A of the Paperwork Reduction Act (PRA) Information Collection Request (ICR)]
- Statistical Methods of the National Agricultural Workers Survey (Part B of the PRA ICR)
- Field sampling protocols
- Questionnaires in English and Spanish
- For users interested in regional analysis:
    - Sampling regions

- Correspondence between NAWS and USDA Farm Labor Survey sampling regions
- Analysis regions in the NAWSPAD file

Chapter 1

How to Calculate Weighted Means and Weighted Proportions using Excel

This chapter provides step-by-step instruction on calculating weighted means and proportions using Excel.

Step 1

Download the NAWS Excel files

From the “Public Data Files” link on the NAWS website (https://www.doleta.gov/naws/), download the two files that are available in Excel format: NAWS_A2E185 (which contains the first half of the variables) and NAWS_F2Y185 (which contains the second half of the variables).

Step 2

Apply Sampling Weights

The NAWS uses a complex sampling design that includes both stratification and clustering; for this reason, users must make use of sampling weights to adjust the relative value of each farmworker so that population estimates may be obtained from the sample. The NAWS sampling weight variable PWTYCRD includes a factor which correctly proportions the data for analysis. The PWTYCRD variable is used for almost all NAWS analysis and allows merging several years of data together. At least two consecutive years of data should be combined to obtain robust. The PWTYCRD variable can be found in both the NAWS_A2E185 and NAWS_F2Y185 files. Further information on how the NAWS sampling weight is calculated is provided in *Statistical Methods of the National Agricultural Workers Survey*, available on the NAWS website (http://www.doleta.gov/naws/).

Step 3

Calculate Weighted Means and Proportions

There are two types of variables in the NAWSPAD: continuous and categorical variables. Users can calculate weighted means for continuous variables using the instructions in Step 3a, and weighted proportions for categorical variables using the instructions in Step 3b. Examples of continuous variables include age of farmworker, wages, number of work days, and numbers of years in the United States. Categorical variables include those for which each farmworker respondent is assigned a value indicating his/her membership in one of several possible categories. Examples of categorical variables include gender, employment status, and legal status. Users should consult the codebook for more information on the variables contained in the NAWSPAD.

Step 3a

Calculate Weighted Means

The mean, or average, of a variable is the sum of all values across all the observations divided by the number of observations. This sample mean assumes that each observation in the sample has an equal weight. However, when calculating the mean of a variable in the NAWSPAD, users need to apply the sampling weight variable PWTYCRD to adjust for the relative value of each farmworker in the sample, because each farmworker contributes differently to the final mean depending on his/her sampling weight.

The weighted mean can be calculated in Excel with a formula that uses both the SUMPRODUCT and SUM functions. For example, if users want to calculate the mean age of farmworkers then the SUMPRODUCT function is used to calculate the numerator in the formula, which is the sum of the products of each farmworker's age and his/her sampling weight (PWTYCRD). The SUM function is used to calculate the denominator in the formula, which is the sum of all farmworker sampling weights (PWTYCRD).

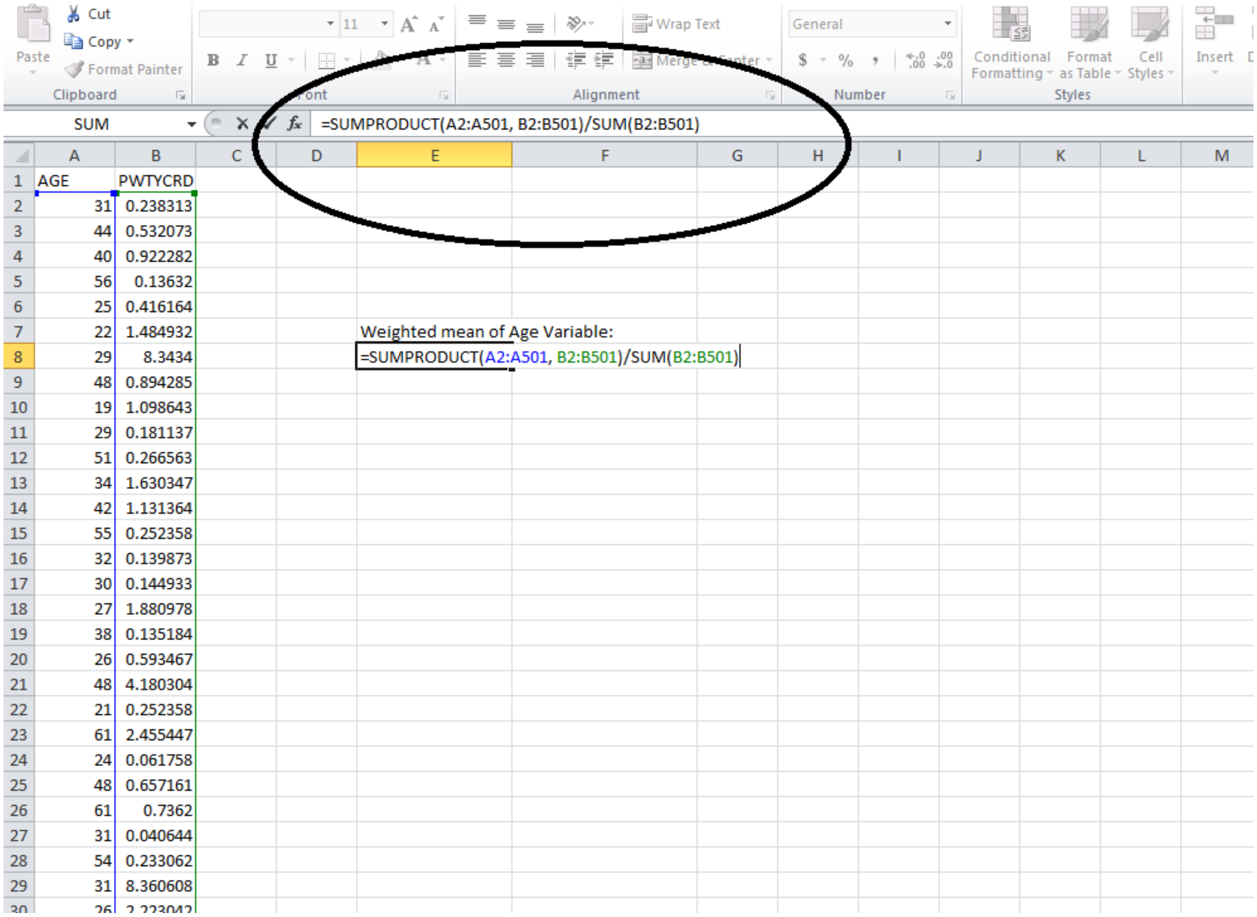
Calculate Weighted Means for All Fiscal Years

The following instructions show how to calculate the weighted mean of a continuous variable for all fiscal years in the dataset combined in Excel. For the purpose of illustration, we will use AGE as the variable of interest and we assume that the Excel data set contains 500 farm workers interviewed in all fiscal years.

1. Determine the variable(s) for which you would like to calculate the weighted mean(s). In our example, we want to calculate mean AGE.
2. Locate the variable of interest and the PWTYCRD variable in the NAWS Excel file. For our example, the variable AGE is located in column A and the variable PWTYCRD is located in column B (see Figure 1). There are 500 rows in our example dataset, one for each farmworker interviewed. The first row contains the variable names.
3. The Excel formula for calculating a weighted mean is SUMPRODUCT(array1,array2,array3,...)/SUM(number1, [number2], [number3], ...). Using our example, the formula to calculate farmworkers' mean age is SUMPRODUCT(A2:A501, B2:B501)/SUM(B2:B501).
    a. SUMPRODUCT(A2:A501, B2:B501) calculates the numerator, which sums the products of each farmworker's age (AGE) and his/her sampling weight (PWTYCRD).

b. SUM(B2:B501) calculates the denominator, which is the sum of all 500 farmworker sampling weights.
c. The full formula SUMPRODUCT(A2:A501, B2:B501)/SUM(B2:B501) calculates the weighted mean age for the 500 farmworkers in the data set. Figure 1 shows an illustration of the Excel data set with the formula written in it.

Figure 1: Formula for Weighted Mean



=SUMPRODUCT(A2:A501, B2:B501)/SUM(B2:B501)

	A	B
1	AGE	PWTYCRD
2	31	0.238313
3	44	0.532073
4	40	0.922282
5	56	0.13632
6	25	0.416164
7	22	1.484932
8	29	8.3434
9	48	0.894285
10	19	1.098643
11	29	0.181137
12	51	0.266563
13	34	1.630347
14	42	1.131364
15	55	0.252358
16	32	0.139873
17	30	0.144933
18	27	1.880978
19	38	0.135184
20	26	0.593467
21	48	4.180304
22	21	0.252358
23	61	2.455447
24	24	0.061758
25	48	0.657161
26	61	0.7362
27	31	0.040644
28	54	0.233062
29	31	8.360608
30	26	2.223042

Weighted mean of Age Variable:
=SUMPRODUCT(A2:A501, B2:B501)/SUM(B2:B501)

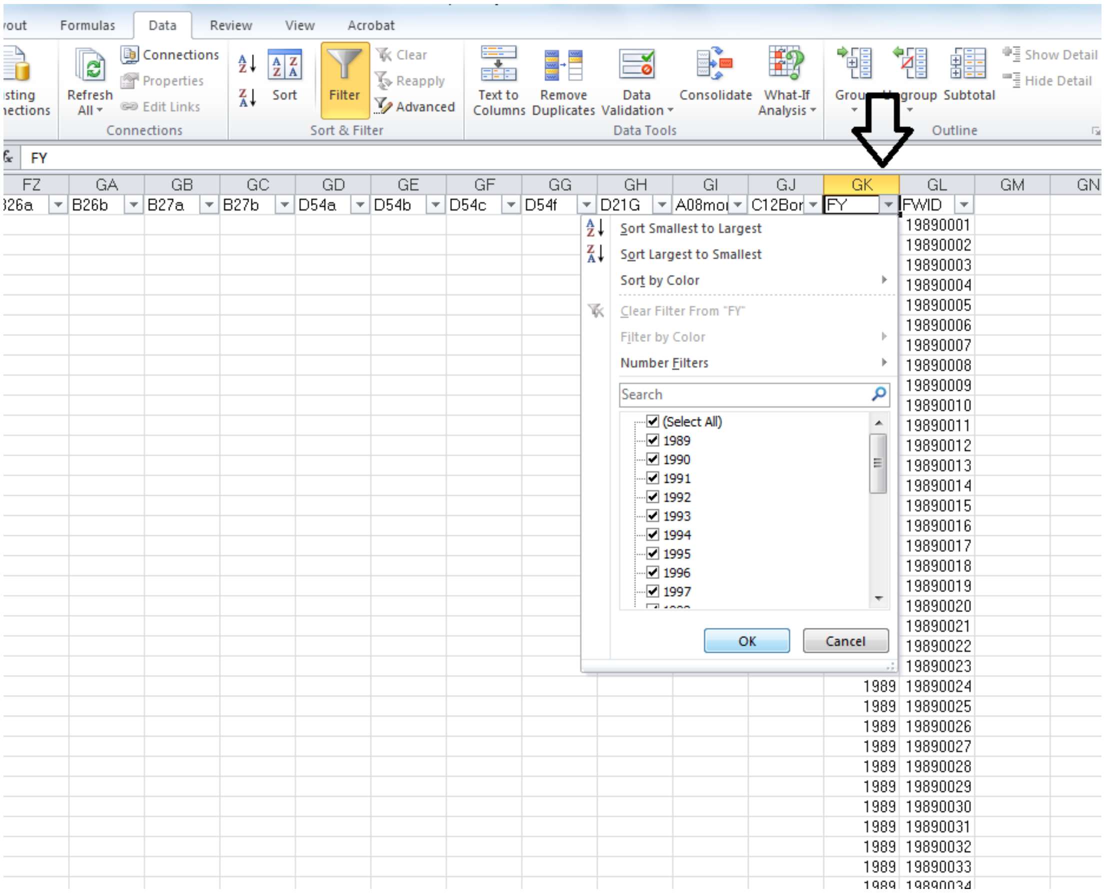
Calculate Weighted Means for Specific Fiscal Years

The previous step described how to calculate weighted means for a continuous variable using all fiscal years in the dataset. It is also possible to calculate weighted means for specific years using the Fiscal Year (FY) variable. To ensure that your results are as accurate as possible, you need to combine at least two consecutive fiscal years of data (e.g., 2015 and 2016). For more information about combining fiscal years, please consult the *Statistical Methods of the National Agricultural Workers Survey*, available on the NAWS website (https://www.doleta.gov/naws/), which describes the statistical methods of the NAWS.

There are two options to create a worksheet containing data for specific fiscal years.

Option 1: Use the Filter option in Excel Before entering any formulas into your worksheet, select all cells in the worksheet, go to the “Data” tab on the menu bar at the top of the worksheet (see Figure 2), and click the “Filter” button. A small drop-down arrow will appear next to each variable name. Click on the drop-down arrow next to the “FY” variable name (see Figure 2). This will open a dialogue box and at the bottom you will see a list of all the fiscal years available for selection (see Figure 2). Select the fiscal years you are interested in and click “OK”. You can pick as many fiscal years as you want, but you should pick at least two consecutive years. Now that you have applied the filter, you can calculate weighted means using the formula described in the “Calculate Weighted Means for All Fiscal Years” section above, steps 1 through 3.

Figure 2: Apply a Filter to "FY"



Option 2: Copy and Paste Before you do any calculations, copy the column headings for all the variables you need (i.e., the variable of interest, PWTYCRD, and FY) and the rows corresponding to the fiscal years you are interested in. Paste them into a new worksheet. Make sure you copy all the rows included in the fiscal years you have chosen; missing even one row can change your results. In the new worksheet, calculate

weighted means using the formula described in the “Calculate Weighted Means for All Fiscal Years” section above, steps 1 through 3.

For more help with calculating weighted means in Excel, please consult the “How to calculate weighted averages in Excel” page on the Microsoft website (http://support.microsoft.com/kb/214049).

Step 3b

Calculate Weighted Proportions

When calculating weighted proportions for categorical variables, users must apply the sampling weight variable PWTYCRD to adjust for the relative value of each farmworker in the sample (see Step 2 Apply Sampling Weights at the beginning of this document). The most efficient way to calculate weighted proportions in Excel is to first organize the data using PivotTables. The following instructions use PivotTables to help you easily calculate weighted proportions for all fiscal years combined or for specific fiscal years.

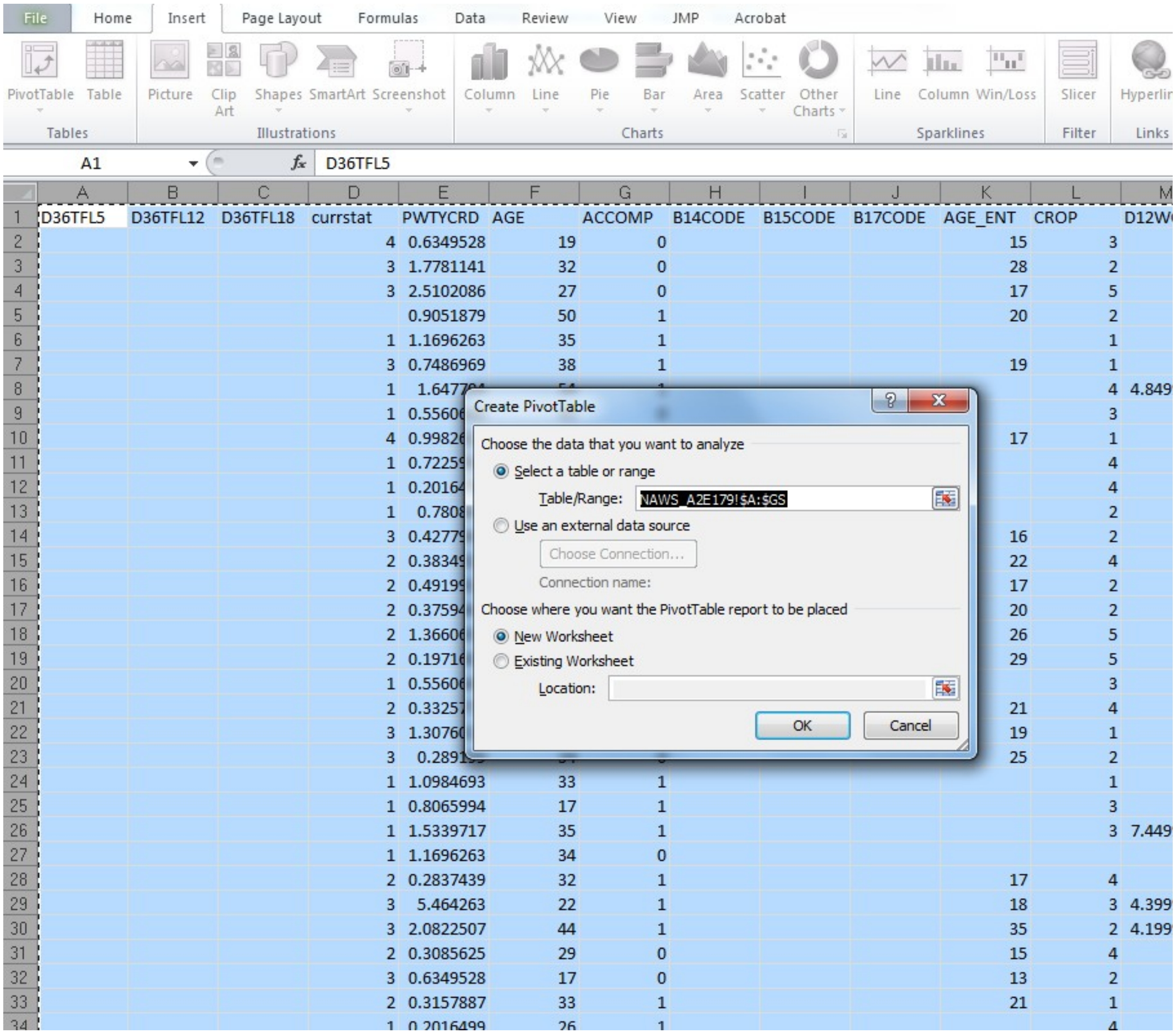
Calculate Weighted Proportions for All Fiscal Years

The instructions below describe the method for calculating weighted proportions for all fiscal years in the dataset. For the purpose of illustration, we will use the variable CURRSTAT (the farmworker’s legal status at the time of interview), which has four categories coded with the numeric values 1 through 4. The calculation described here will yield the weighted proportion for each of the four categories of CURRSTAT.

To calculate weighted proportions for all fiscal years combined:

1. Determine the variable(s) for which you would like to calculate the weighted proportions. In our example, we want to calculate the weighted proportions of CURRSTAT.
2. Highlight all the cells in your spreadsheet. Go to the “Insert” tab on the menu bar at the top of the worksheet and select “PivotTable” (see Figure 3). The “Create PivotTable” dialogue box will appear, and two options will automatically be selected: “Select a table or range” (which shows the range of all the cells selected in our spreadsheet) and “New Worksheet” (see Figure 3). Click “OK”.

Figure 3: Create a PivotTable



A1 — D36TFL5

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	D36TFL5	D36TFL12	D36TFL18	currstat	PWTYCRD	AGE	ACCOMP	B14CODE	B15CODE	B17CODE	AGE_ENT	CROP	D12W
2				4	0.6349528	19	0				15	3	
3				3	1.7781141	32	0				28	2	
4				3	2.5102086	27	0				17	5	
5					0.9051879	50	1				20	2	
6				1	1.1696263	35	1					1	
7				3	0.7486969	38	1				19	1	
8				1	1.6477							4	4.849
9				1	0.55606							3	
10				4	0.99826						17	1	
11				1	0.72259							4	
12				1	0.20164							4	
13				1	0.7808							2	
14				3	0.42779						16	2	
15				2	0.38349						22	4	
16				2	0.49199						17	2	
17				2	0.37594						20	2	
18				2	1.36606						26	5	
19				2	0.19716						29	5	
20				1	0.55606							3	
21				2	0.33257						21	4	
22				3	1.30760						19	1	
23				3	0.2891						25	2	
24				1	1.0984693	33	1					1	
25				1	0.8065994	17	1					3	
26				1	1.5339717	35	1					3	7.449
27				1	1.1696263	34	0						
28				2	0.2837439	32	1				17	4	
29				3	5.464263	22	1				18	3	4.399
30				3	2.0822507	44	1				35	2	4.199
31				2	0.3085625	29	0				15	4	
32				3	0.6349528	17	0				13	2	
33				2	0.3157887	33	1				21	1	
34				1	0.2016499	26	1					4	

Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: NAWS_A2E179!$A:$GS

○ Use an external data source

Choose Connection...

Connection name:

Choose where you want the PivotTable report to be placed

● New Worksheet

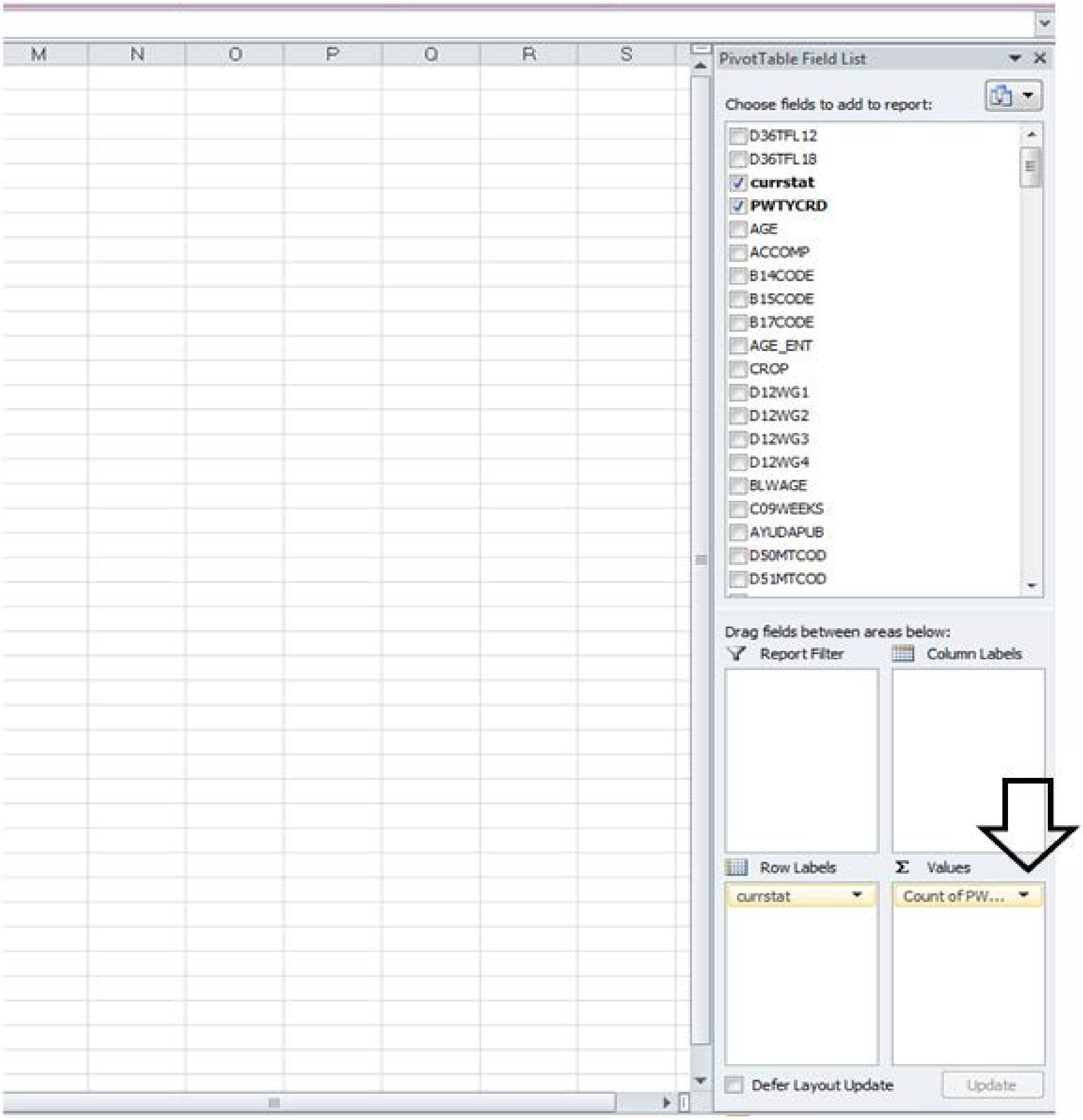
○ Existing Worksheet

Location:

OK Cancel

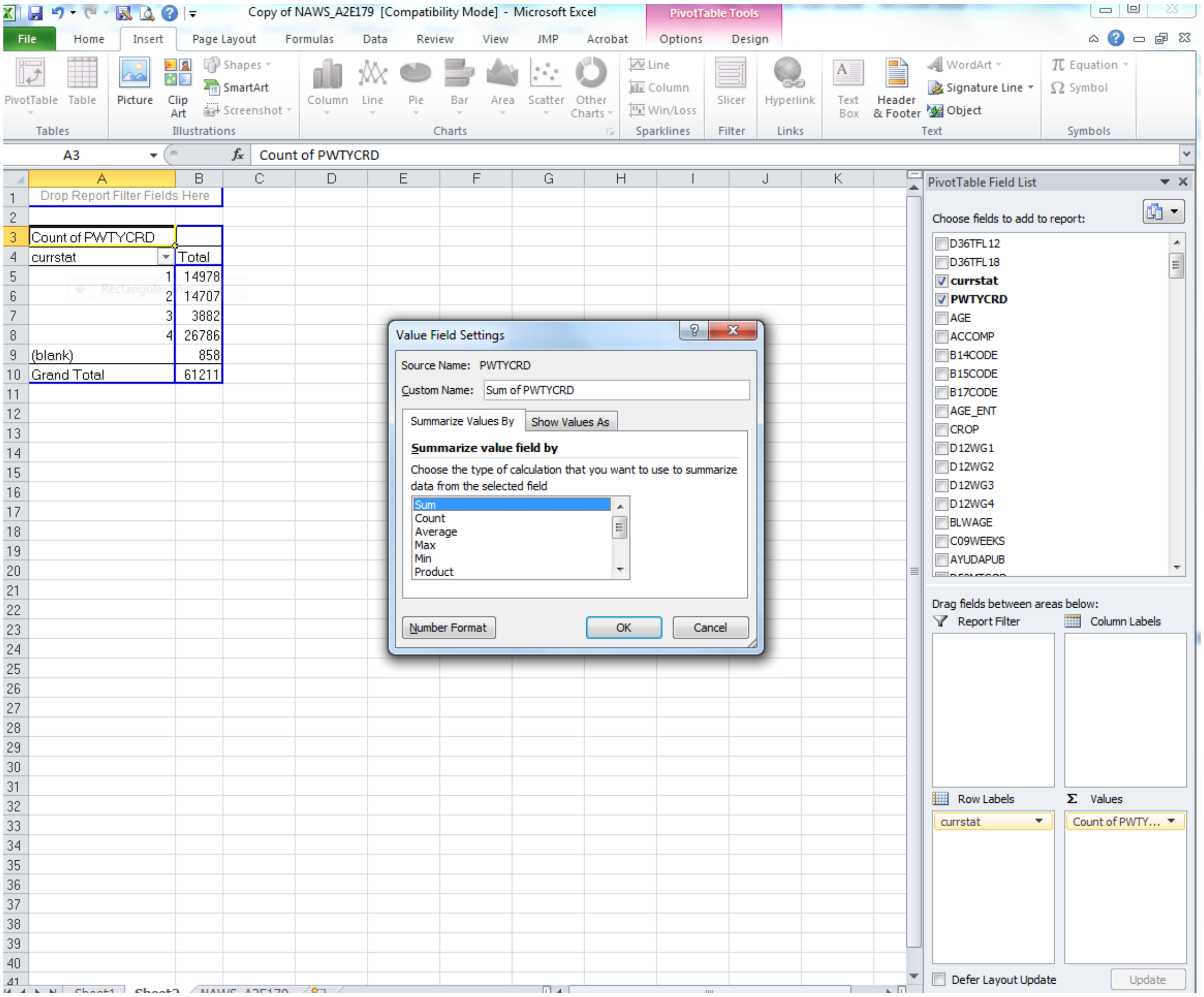
3. After clicking “OK”, a new worksheet will open. On the left side of the new worksheet you will see an empty PivotTable. On the right side you will see a PivotTable Field List from which you can select the variables that you want to add to your PivotTable (see Figure 4). Using our example of CURRSTAT as the variable of interest:
    a. Drag CURRSTAT and drop it into the “Row Labels” area.
    b. Drag PWTYCRD and drop it into the “Values” area; you will see “Count of PWTYCRD” with a drop-down arrow next to it (see Figure 4).

Figure 4: PivotTable Field List



c. Click on the drop-down arrow next to “Count of PWTYCRD” and click “Value Field Settings”. In the Value Field Settings dialogue box, select the “Sum” option under “Summarize value field by” (see Figure 5).

Figure 5: Value Field Settings for PWTYCRD



Copy of NAWS_A2E179 [Compatibility Mode] - Microsoft Excel

A3 f_x Count of PWTYCRD

Drop Report Filter Fields Here

Count of PWTYCRD	
currstat	Total
1	14978
2	14707
3	3882
4	26786
(blank)	858
Grand Total	61211

Value Field Settings

Source Name: PWTYCRD

Custom Name: Sum of PWTYCRD

Summarize Values By | Show Values As

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

Sum
Count
Average
Max
Min
Product

Number Format | OK | Cancel

PivotTable Field List

Choose fields to add to report:

□ D36TFL12
□ D36TFL18
☑ **currstat**
☑ **PWTYCRD**
□ AGE
□ ACCOMP
□ B14CODE
□ B15CODE
□ B17CODE
□ AGE_ENT
□ CROP
□ D12WG1
□ D12WG2
□ D12WG3
□ D12WG4
□ BLWAGE
□ C09WEEKS
□ AYUDAPUB

Drag fields between areas below:

Report Filter | Column Labels

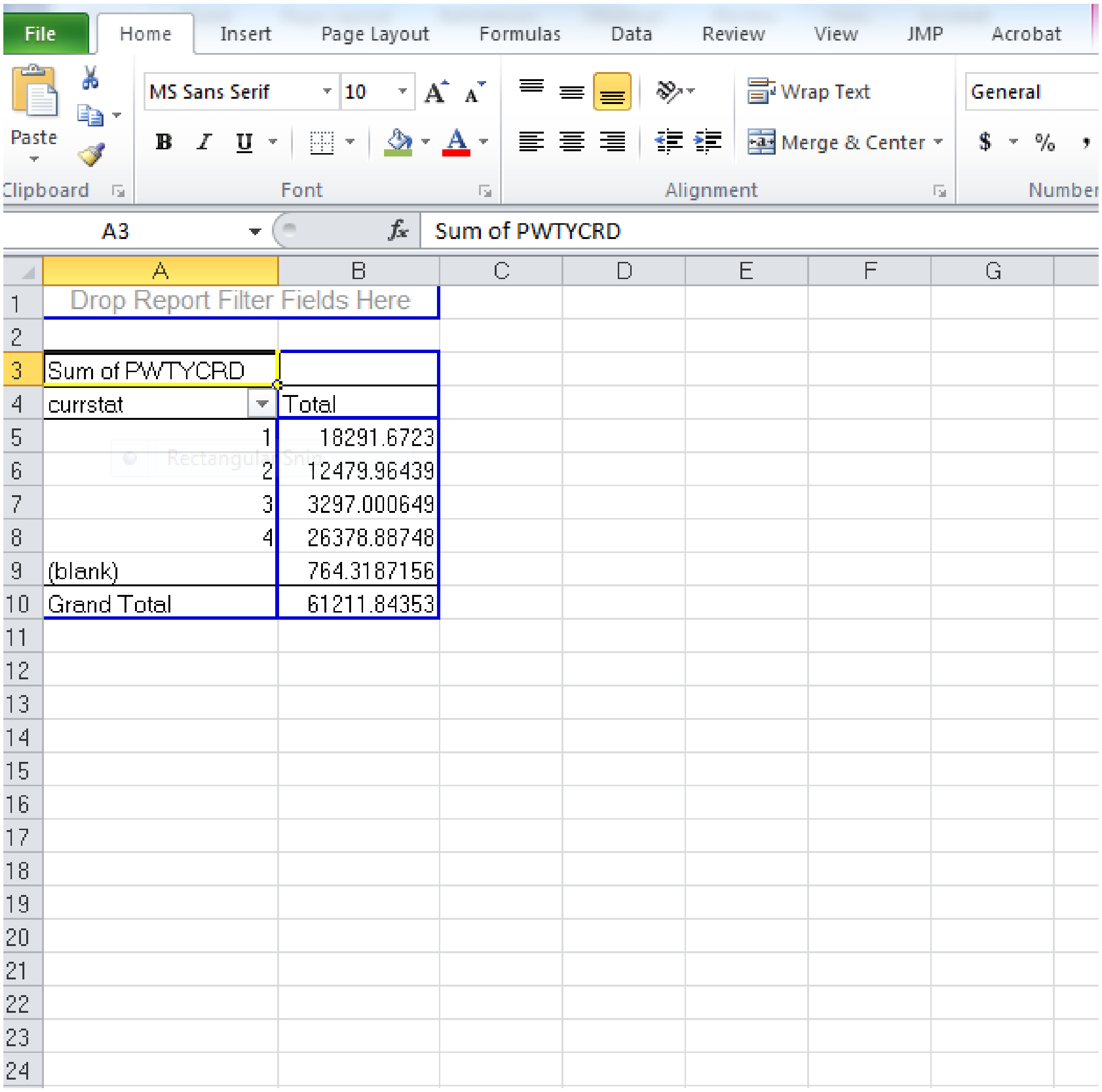
Row Labels: currstat

Values: Count of PWTY...

□ Defer Layout Update | Update

d. Click “OK” to create your PivotTable. The left column in the table contains the four categories of CURRSTAT. The right column contains the sum of the PWTYCRD values for each CURRSTAT category (see Figure 6); this is the column that you will use to calculate your weighted proportions.

Figure 6: Created PivotTable



A3 | Sum of PWTYCRD

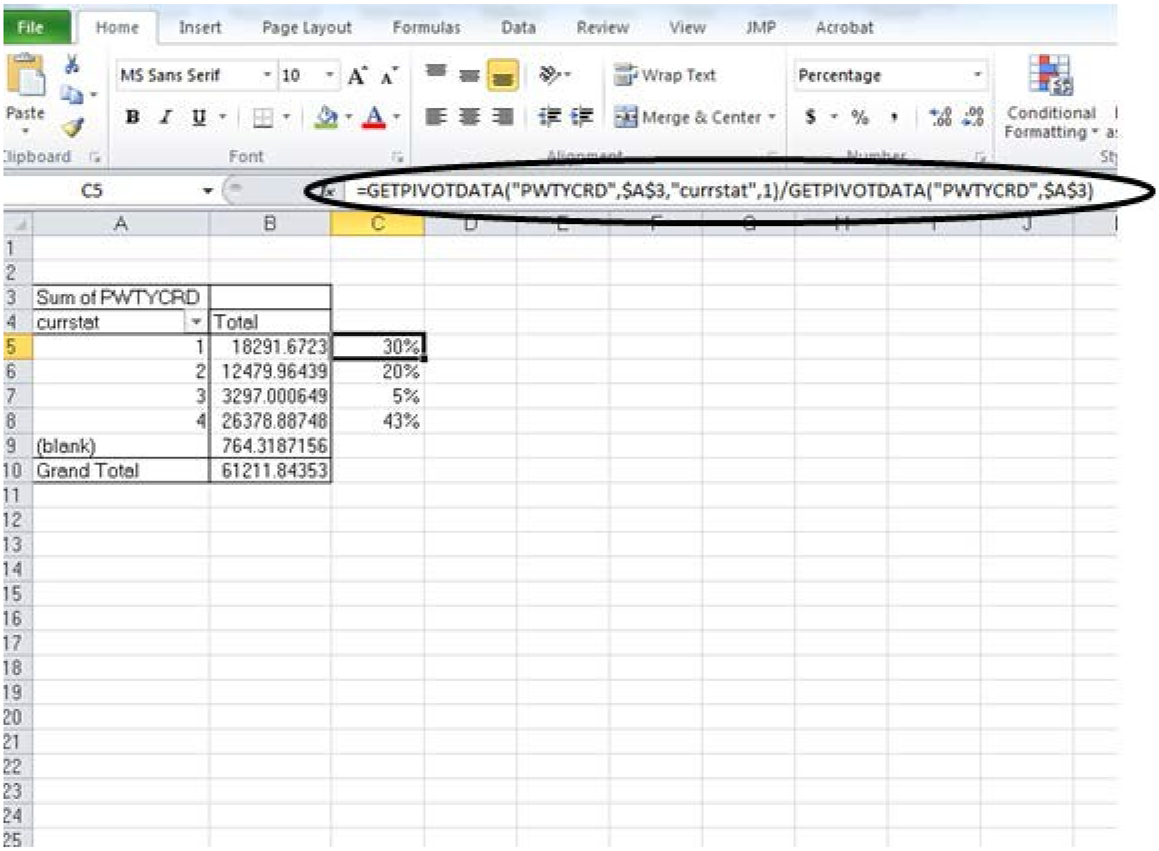
	A	B
1	Drop Report Filter Fields Here	
2		
3	Sum of PWTYCRD	
4	currstat	Total
5	1	18291.6723
6	2	12479.96439
7	3	3297.000649
8	4	26378.88748
9	(blank)	764.3187156
10	Grand Total	61211.84353

Note: The columns in your PivotTable may have headings different from those pictured based on the version of Excel you are using. The values in the right column will likely differ from those pictured, depending on the fiscal years of data you are analyzing.

4. Now you are ready to calculate the weighted proportions of the variable (i.e., how often each category of the variable occurs). For each category of the variable (i.e., row in the PivotTable):
    a. Create a formula that divides the value in the right column of the PivotTable by the “Grand Total” (see Figure 7).
    b. Convert the resulting values from decimals to percentages (see Figure 7). Note: Because you are calculating weighted proportions based on cells that were created using a PivotTable, you cannot use the fill handle to apply the formula for the first category of the variable to the other

categories of the variable. If you want to be able to use the fill handle option, copy the PivotTable and paste it elsewhere in your worksheet using “Paste Special” with the “Values” option. This will remove all embedded formulas and provide you with a static table that you can use for your calculations.

Figure 7: Formula for Weighted Proportion



C5

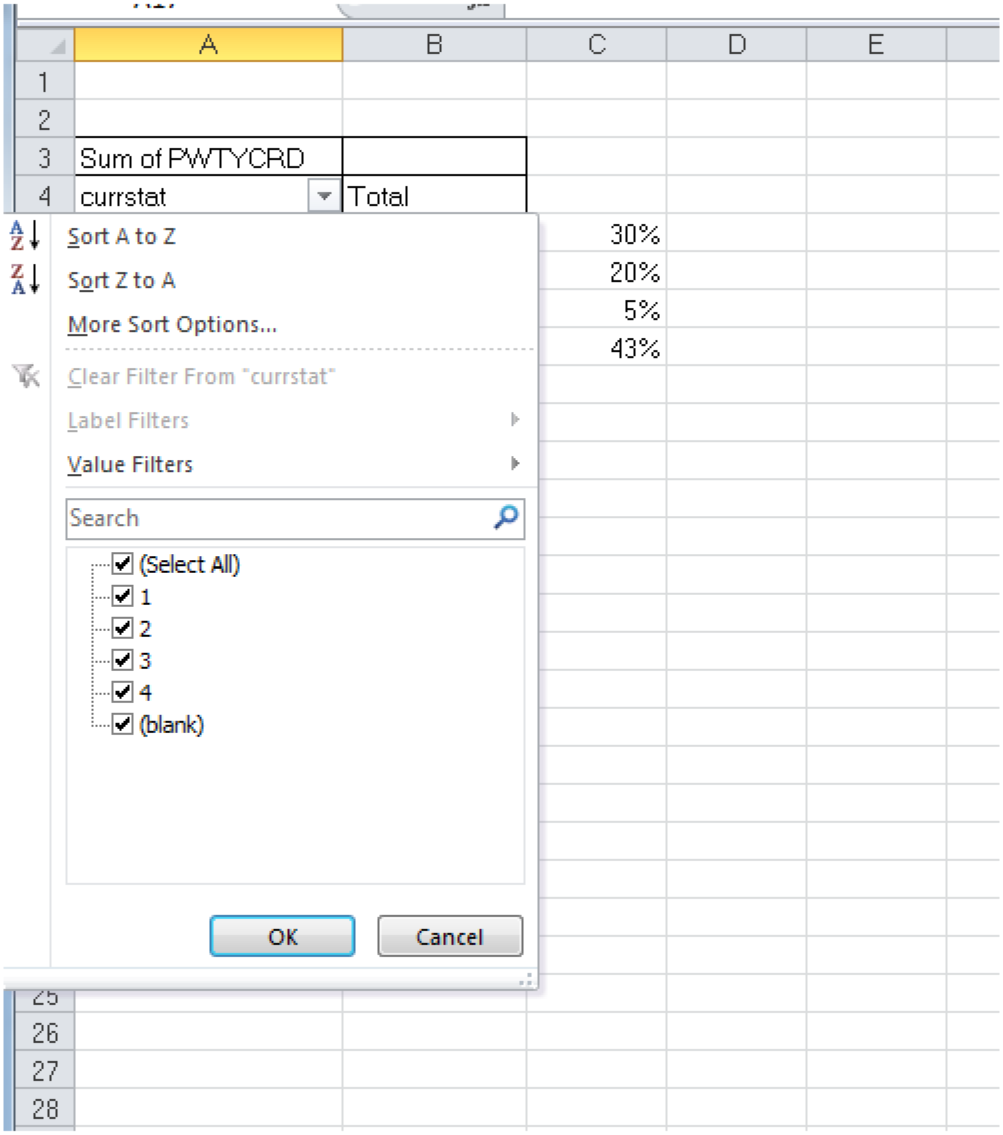
```
=GETPIVOTDATA("PWTYCRD",$A$3,"currstat",1)/GETPIVOTDATA("PWTYCRD",$A$3)
```

Sum of PWTYCRD		
currstat	Total	
1	18291.6723	30%
2	12479.96439	20%
3	3297.000649	5%
4	26378.88748	43%
(blank)	764.3187156	
Grand Total	61211.84353	

Note that because your formula for calculating weighted proportion uses cells in a PivotTable, you cannot use the fill handle to apply the formula for the first category of the variable to the other categories of the variable. If you want to be able to use the fill handle option, copy the PivotTable and paste it elsewhere in your worksheet using “Paste Special” with the “Values” option. This will remove all embedded formulas and provide you with a static table that you can use for your calculations.

The category labeled “(blank)” represents the respondents who are missing a value for the variable. If left in the PivotTable, the category of missing values will comprise a share of respondents. If you want to exclude missing values from your analysis, you can filter them out by clicking the down arrow next to CURRSTAT, unchecking the box next to “(blank)” in the filter menu, then clicking “OK” (see Figure 8).

Figure 8: Filter “(blank)” from Created PivotTable



For additional help creating PivotTables in Excel, please consult the “Create or delete a PivotTable or PivotChart report” page on the Microsoft website (https://support.office.com/en-us/article/Create-or-delete-a-PivotTable-or-PivotChart-report-d09e4d07-8cd6-4b60-afad-8fb67418800f?CorrelationId=213dd195-0deb-4f22-8ec8-e80eca1c2fd0&ui=en-US&rs=en-US&ad=US).

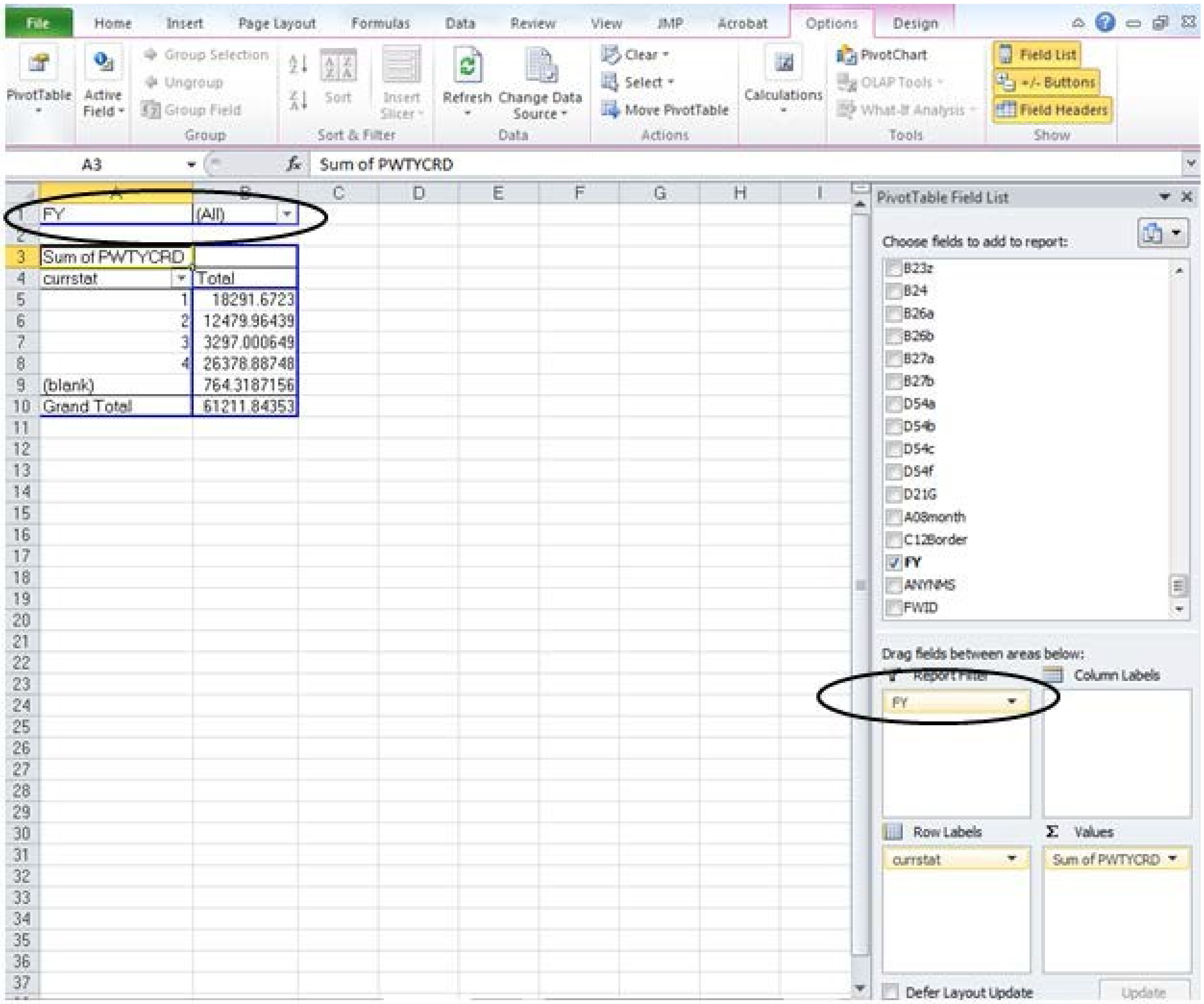
Calculate Weighted Proportions for Specific Fiscal Years

The process for calculating weighted proportions for specific fiscal years is very similar to the process for calculating weighted proportions for all fiscal years. To ensure that the data results are as accurate as possible, you need to combine at least two consecutive fiscal years of data (e.g., 2015 and 2016). For more information about combining fiscal years, please consult the *Statistical Methods of the National Agricultural Workers Survey* which describes the statistical methods of the NAWS.

There are two options for calculating weighted proportions for specific fiscal years.

Option 1: Use the Report Filter option in a PivotTable Create a PivotTable by following steps 1 through 3 in the “Calculate Weighted Proportions for All Fiscal Years” section above. After you have changed the “Value Field Settings” for PWTYCRD to “Sum of PWTYCRD” as described in step 3, return to the PivotTable Field List on the right side of the screen, locate the variable “FY”, and drag and drop it into the “Report Filter” area below the list. Once you have done this, the variable “FY” will appear in the “Report Filter” section above your created PivotTable (see Figure 9).

Figure 9: Apply “FY” to the Report Filter



A3 Sum of PWTYCRD

FY	(All)
Sum of PWTYCRD	
currstat	Total
1	18291.6723
2	12479.96439
3	3297.000649
4	26378.88748
(blank)	764.3187156
Grand Total	61211.84353

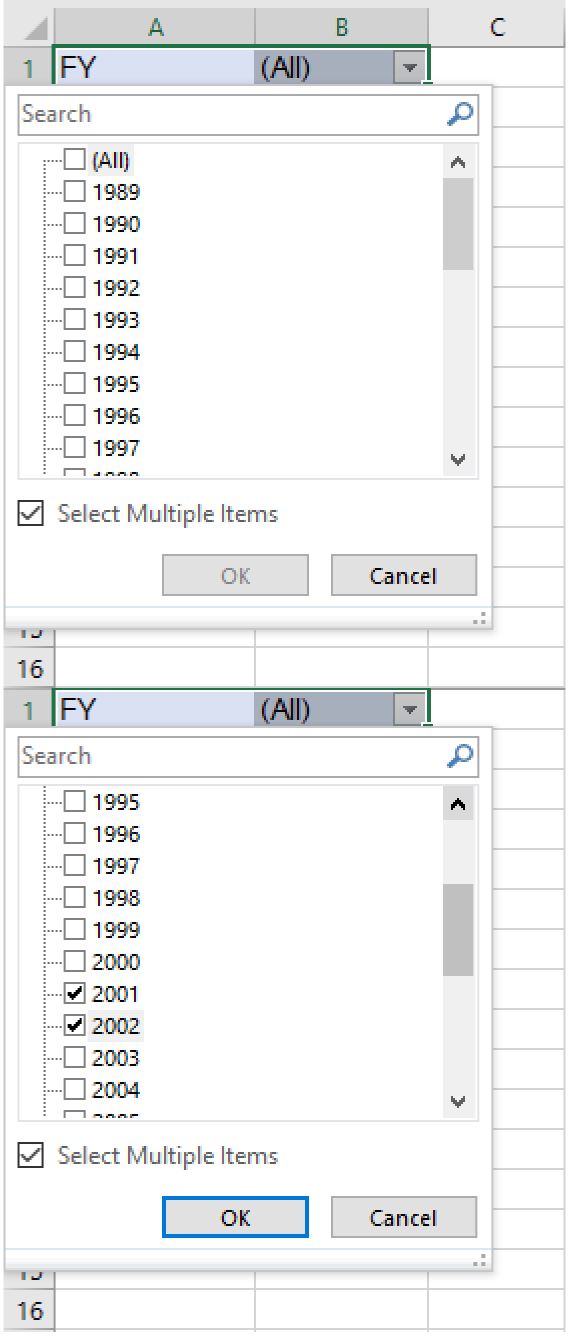
Report Filter: FY

Row Labels: currstat

Values: Sum of PWTYCRD

The next step is to select the fiscal years you are interested in. Click the down arrow next to (All) and check the “Select Multiple Items” box. A check box will appear to the left of each Fiscal Year in the list; click in the check box to the left of (All) to deselect all of them. Locate the years that you are interested in, check the box for each (we use 2001-2002, for the purpose of illustration), then click “OK” (see Figure 10).

Figure 10: Select Specific Fiscal Years



Once the report filter is applied, the sum of the PWTYCRD values for each category of CURRSTAT and the Grand Total for the table will recalculate to reflect the weighted frequencies for only those fiscal years which you have selected. The final step is to calculate the weighted proportions according to the procedure described in step 4 of the “Calculate Weighted Proportions for All Fiscal Years” section above.

Option 2: Copy and Paste Before you do any of the steps involved in creating a PivotTable, copy the column headings for all the variables you need (i.e., the variable of interest, PWTYCRD, and FY) and the rows corresponding to the fiscal years you are interested in. Paste them into a new worksheet. Make sure you copy all the rows included in the fiscal years you have chosen; missing even one row can change your results. In the new worksheet, create a PivotTable and calculate weighted proportions according to the process described in steps 1 through 4 of the “Calculate Weighted Proportions for All Fiscal Years” section above.

For additional help creating PivotTables in Excel, please consult the “Create or delete a PivotTable or PivotChart report” page on the Microsoft website (https://support.office.com/en-us/article/Create-or-delete-a-PivotTable-or-PivotChart-report-d09e4d07-8cd6-4b60-afad-8fb67418800f?CorrelationId=213dd195-0deb-4f22-8ec8-e80eca1c2fd0&ui=en-US&rs=en-US&ad=US).

Chapter 2

Using Replicate Weights and Calculating Design-Corrected Standard Errors

The current NAWSPAD dataset includes BRR weights that can be used to estimate design-corrected standard errors using common statistical software programs. This chapter provides an overview of the reasoning behind the selection of the BRR method, describes how to use the replicate weights to generate design-corrected standard errors, and provides a general overview of the procedures NAWS used to create the replicate weights.

Selection of the BRR Method

In complex survey analysis, there are two general methods to calculate standard error: linearization methods (also known as Taylor Series Linearization) and replicate methods. In replicate methods, a series of smaller subsamples, or replicates, are generated from the full sample. Once the replicates are formed, the standard error is calculated for each subsample and the full sample's standard error is estimated using the variation among the replicate estimates.

Replicate-based estimators are extremely flexible, allowing for standard error calculation for a wide range of point estimates, from regression coefficients and ratios to quantiles, all based on the same set of replicate weights. On the other hand, standard errors for discontinuous functions such as quantiles are not possible using linearization methods. Another advantage of using replicate-based estimators is that, unlike the linearization methods, the user does not need the primary sampling unit (PSU) identifiers to run standard error calculations.

There are several methods (e.g., jackknife, bootstrap, BRR) that can be used to create replicate weights. NAWS utilizes Fay's method of BRR, as this particular form of BRR offers several advantages in terms of ease of implementation, stability, and efficiency over jackknife and bootstrap replicate methods. BRR variance estimators also produce actual confidence interval coverage that reflects nominal confidence levels better than the other estimators. With respect to jackknife methods, BRR gives consistent estimation for discontinuous functions (such as quantiles) under certain sampling designs and for analyzing population subsets for which sample size within PSUs is small.

The NAWSPAD contains survey responses, their associated final sampling weights, 80 replicate weights, and the variable FAYWTYRS. The procedures and syntax needed to carry out standard error calculation in SAS, Stata, and R are presented below. An explanation of how the replicate weights were created can be found at the end of this chapter.

The NAWS BRR weights are not valid for single years or for any combination of years that does not appear as a value of the NAWSPAD variable FAYWTYRS value. *Users must always condition their analyses on values of the FAYWTYRS variable*, which contains an identifier for the two-year survey block applicable to each observation. When using FAYWTYRS to analyze a subset of years (e.g., only the FAYWTYRS value representing 2015–2016) the software environment may issue a warning to the effect that selecting a subset in this manner does not yield a valid subdomain analysis, but this warning can be safely ignored. Each two-year survey block constitutes a single survey episode for separate analysis and should not be considered a subdomain.

Using the Replicate Weights

Using BRR weights, design-corrected standard errors can be calculated using standard statistical software packages, including SAS, Stata, and R. The statistical procedure involves generating replicates, then calculating the point estimates for each replicate and the variance of the replicate estimates. This variance is the estimated sampling variance of the statistic of interest.

To understand the ways in which various software packages estimate the design corrected errors, it is useful to look at the equations. Let $\hat{\theta}$ be the point estimate based on the full sample, **y**, corresponding to the population parameterθ. Let $\hat{\theta}_r$ be the point estimate based on $\mathbf{y}_{brr}^{(r)}$, the re-weighted rth half-sample replicate, where r = 1, 2,...,R.

Then

$$\hat{\theta} = \bar{\bar{\theta}}_R = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r \qquad (1)$$

and

$$\hat{V}[\hat{\theta}] = \frac{1}{R(1-\rho)^2}\sum_{r=1}^{R}\left(\hat{\theta}_r - \bar{\bar{\theta}}_R\right)^2 \qquad (2)$$

where $\bar{\bar{\theta}}_R$ is the average value of $\hat{\theta}_r$ over the R replicates and $\hat{V}[\hat{\theta}]$ is the estimated variance of the statistic of interest. An alternative form for the estimated variance replaces $\bar{\bar{\theta}}_R$ by the full sample point estimate $\hat{\theta}$:

$$\hat{V}[\hat{\theta}] = \frac{1}{R(1-\rho)^2}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2 \qquad (3)$$

The alternative form in equation (3) is implemented in SAS, whereas both R and Stata use equation (2) by default, with options to switch to the former, if desired (through the *mse* option). Presently there is no option to switch formulas in SAS. Although the two methods will produce slightly different estimates, the estimators are asymptotically identical and, in practice, differences in estimates are very small.

This section shows how to use the replicate weights in SAS, Stata, and R. The example assumes that the current NAWS datafile (*naws_all.sas7bdat*) has been downloaded from the official NAWS website [(https://www.doleta.gov/naws/)](https://www.doleta.gov/naws/) and resides in the directory "./naws/data" under the user's active profile. The example shows the correct calculation of means, proportions, and subdomain mean analyses for: WAGET1 (worker compensation expressed as an hourly rate), MIGTYPE2 (worker migrant status), and AGE (worker calendar age) by GENDER (coded 0=male and 1=female). The example assumes the analysis is being conducted using fiscal years (FY) 2015 and 2016.

SAS Users

```
libname naws './naws/data';

data naws;

set naws.naws_all;

run;     /* this database contains the full set of analysis and weight variables */

proc surveymeans data=naws varmethod=BRR(fay=.5);

repweights    fay01 fay02 fay03 fay04 fay05 fay06 fay07 fay08 fay09 fay10 fay11 fay12
fay13 fay14 fay15 fay16 fay17 fay18 fay19 fay20 fay21 fay22 fay23 fay24 fay25 fay26
fay27 fay28 fay29 fay30 fay31 fay32 fay33 fay34 fay35 fay36 fay37 fay38 fay39 fay40
fay41 fay42 fay43 fay44 fay45 fay46 fay47 fay48 fay49 fay50 fay51 fay52 fay53 fay54
fay55 fay56 fay57 fay58 fay59 fay60 fay61 fay62 fay63 fay64 fay65 fay66 fay67 fay68
fay69 fay70 fay71 fay72 fay73 fay74 fay75 fay76 fay77 fay78 fay79 fay80;

weight pwtycrd;

var waget1;

where faywtyrs=20152016;

run;
```

```
proc surveymeans data=naws varmethod=BRR(fay=.5) missing;

repweights    fay01 fay02 fay03 fay04 fay05 fay06 fay07 fay08 fay09 fay10 fay11 fay12
fay13 fay14 fay15 fay16 fay17 fay18 fay19 fay20 fay21 fay22 fay23 fay24 fay25 fay26
fay27 fay28 fay29 fay30 fay31 fay32 fay33 fay34 fay35 fay36 fay37 fay38 fay39 fay40
fay41 fay42 fay43 fay44 fay45 fay46 fay47 fay48 fay49 fay50 fay51 fay52 fay53 fay54
fay55 fay56 fay57 fay58 fay59 fay60 fay61 fay62 fay63 fay64 fay65 fay66 fay67 fay68
fay69 fay70 fay71 fay72 fay73 fay74 fay75 fay76 fay77 fay78 fay79 fay80;

weight pwtycrd;

var migtype2;

class migtype2;

where faywtyrs=20152016;

run;

proc surveymeans data=naws varmethod=BRR(fay=.5);

repweights    fay01 fay02 fay03 fay04 fay05 fay06 fay07 fay08 fay09 fay10 fay11 fay12
fay13 fay14 fay15 fay16 fay17 fay18 fay19 fay20 fay21 fay22 fay23 fay24 fay25 fay26
fay27 fay28 fay29 fay30 fay31 fay32 fay33 fay34 fay35 fay36 fay37 fay38 fay39 fay40
fay41 fay42 fay43 fay44 fay45 fay46 fay47 fay48 fay49 fay50 fay51 fay52 fay53 fay54
fay55 fay56 fay57 fay58 fay59 fay60 fay61 fay62 fay63 fay64 fay65 fay66 fay67 fay68
fay69 fay70 fay71 fay72 fay73 fay74 fay75 fay76 fay77 fay78 fay79 fay80;

weight pwtycrd;

var age;

domain gender;

where faywtyrs=20152016;

run;
```

Stata Users

(You must first convert the datafiles to Stata format. Several methods exist, but if you have a working SAS installation, this can be most easily done as follows:

```
libname naws './naws/data'; proc

export data=naws.naws_all

outfile="./naws/data/naws_all.dta" dbms=stata replace;
```

```
run;
```

In the instructions below we assume that these Stata files exist in the current working directory.)

```
use naws_all
svyset _n [pweight=pwtycrd], brrweight(fay01-fay80) fay(.5) vce(brr)

svy: mean waget1 if faywtyrs==20152016 estat effects

svy: tab migtype2 if faywtyrs==20152016, se deff missing svy: mean age if
faywtyrs==20152016, over(gender)

estat effects, srssubpop
```

Note the need to specify the srssubpop option with estat effects following an estimation that makes use of the over() option (for tabulation over factor levels); this asks Stata to compute design effects separately with respect to the subdomains defined by levels of the factor variable instead of with respect to the entire sample; the latter is usually the desired quantity in tabulations.

R Users

(The foreign library in R can read Stata files of older format but not SAS files, so we recommend that you first convert the files to Stata format; see the Stata instructions above for a method using SAS tools. If your Stata installation is version 13 or later, use the saveold command instead of save in order to facilitate compatibility. We assume that the current working directory contains Stata files named *naws_all.dta*

```
library(foreign) library(survey)

naws_all=read.dta("naws_all.dta")

dim(naws_all.dta) # confirm observation and variable counts

naws.svr=svrepdesign(repweights="fay[0-9]",weights=~pwtycrd,

type="Fay",rho=.5,data= naws_all.dta)

summary(naws.svr) # check that 80 replicates are specified

svymean(~waget1,subset(naws.svr,faywtyrs==20152016), deff="replace",na.rm=TRUE)

svymean(~factor(migtype2),subset(naws.svr,faywtyrs==20152016),
deff="replace",na.rm=TRUE)

svyby(~age,~gender,subset(naws.svr,faywtyrs==20152016),svymean,
deff="replace",na.rm=TRUE)
```

Note the need to use the na.rm=TRUE argument in order to direct R to use a casewise-deletion policy (which is applied by default in both SAS and Stata). Also, the most appropriate design effect for NAWS analysis uses the with-replacement form of the denominator (because simple random sampling from the total population of eligible farm workers in the U.S. is effectively with-replacement sampling, and because the survey weights used in NAWS do not sum to population totals); this is the default behavior in Stata when a finite population correction factor is not specified, but in R it needs to be explicitly requested using the "replace" option on the deff argument.

How the Replicate Weights Were Created

This section provides an overview of the procedure used to create the NAWS replicate weights. The BRR method requires that each stratum contains exactly two PSUs. Although this can be a restrictive requirement for some surveys, the NAWS survey was designed to obtain at least two PSUs within each stratum. In cases of deviations, strata can be merged or split to form the two PSUs. These modifications were implemented in NAWS as suggested in Wolter's *Introduction to Variance Estimation*.[1,2]

Once each stratum contains exactly two PSUs, and designated PSU 1 or 2, the replicates are formed. In the case of NAWS, 80 replicate weights were chosen as a conservative number to accommodate any necessary splitting of strata. Each of the 80 columns represents one replicate and its elements determine whether PSU 1 is selected (+1) or not (-1). As a result of orthogonality, each PSU within each stratum is incorporated into half of the replicates, and each pair of PSUs from different strata appears in half the samples. These “balanced” half-samples ensure that the sampling variance of the replicate estimates is an asymptotically (i.e., for large sample sizes) unbiased estimator for the true sampling variance of the point estimate. In simpler terms, the procedure ensures that the replicates remain relatively representative of the full sample so that the expected value of the variance estimated using the replicates equals the sampling variance of the full sample.

The NAWSPAD BRR weights used Fay’s method, a variant of BRR, which invokes a weighted selection of PSUs (specified by the parameter rho, $0 < \rho \leq 1$) that incorporates information from the full sample into each replicate. The selected PSU is given ($2-\rho$) times its survey weight while the unselected PSU is given ρ times its survey weight in a particular replicate. For example, if a rho of 0.5 is chosen, then the weight of the selected PSU is given 2-0.5=1.5 times its sampling weight while the unselected PSU is

[1] Wolter, K. (2007). *Introduction to Variance Estimation*. Springer Science & Business Media.

[2] It should be noted that positive bias is introduced when effective stratifications are modified. In other words, reduction in variance attained through stratification is potentially lost upon collapsing with another stratum. However, it is desirable to keep standard error estimation on the conservative side (i.e. positive bias) as it is not possible to account for all sources of error in the NAWS. Similar standard errors were obtained in tests using other estimation methods.

given 0.5 times its sampling weight in the replicate. A rho of 1 results in the unmodified BRR method. An important consequence of a weighted selection procedure is the avoidance of patterns of zeros in the replicate weights generated by unmodified BRR methods, which may enable users to infer PSU membership. For NAWS, a rho of ρ = 0.5 was employed based on advantages in efficiency, conservativeness, and stability, in addition to greater confidentiality of participants.