```
* Analysis of American Community Survey PUMS Data using Stata
* Dataset used is the 2013 1-year PUMS file for California
*
* Created by:
* Tim Bates
* Phillip R. Lee Institute for Health Policy Studies
* University of California, San Francisco
* Timothy.Bates@ucsf.edu
*
* Created on: March 25, 2015
*
********************************************************************************

* TELL STATA THAT YOU ARE USING COMPLEX SURVEY DATA
* AND INSTRUCT STATA TO USE "PWGTP" AS THE PERSON WEIGHT VARIABLE

* This will be necessary to generate weighted estimates

svyset [pw=pwgtp]

* If you wanted to use the replicate weights to get the most accurate esimates of variance
* you would use the following command instead (just remove the asterisk (*) that precedes it).

* svyset [pw=pwgtp], sdrweight(pwgtp1-pwgtp80) vce(sdr)
```

* IDENTIFY SAMPLE CASES WHERE OCCUPATION IS REPORTED AS REGISTERED NURSE

* OCCP is the variable identifying occupation and there are 3 occupation codes associated with RNs:

* 3255 (Registered Nurses)

* 3256 (Nurse Anesthetists)

* 3258 (Nurse Practitioners & Nurse Midwives)

* We're going to collapse all of these codes into a single binary variable that takes a value of 1 if

* occupation is "registered nurse" and 0 if not.

generate rn=0

replace rn=1 if occp==3255

replace rn=1 if occp==3256

replace rn=1 if occp==3258

* Now we'll exclude sample cases where educational attainment is reported as less than a high school diploma.

* SCHL is the educational attainment variable; a value of SCHL less than 16 indicates less than a HS diploma.

```
replace rn=0 if schl<16

* Give the new variable a name.

label variable rn "id occupation is registered nurse"



* IDENTIFYING CURRENTLY EMPLOYED REGISTERED NURSES

* Not everyone whose occupation is reported is actually employed.

* We'll create another binary variable to identify currently employed registered nurses.

* First, set values of the new variable identical to the RN id variable created above.

generate rn_employed=rn

* Now exclude sample cases where employment status is reported as either "unemployed" or "not in labor force".

* ESR is the employment status variable; a value of 3 is "unemployed", a value of 6 is "not in labor force"

replace rn_employed=0 if esr==3

replace rn_employed=0 if esr==6

* Give the new variable a name.
```

label variable rn_employed "id currently employed registered nurse"

* COMPARE THE NUMBER OF EMPLOYED VS NOT-EMPLOYED RNs

tab rn rn_employed, row

* The output will tell you what share of RNs in sample were not currently employed in nursing at the time of the survey.

* COMPARE AGE OF EMPLOYED VS NOT-EMPLOYED RNs

* When we created the binary variable to identify RNs currently employed in nursing, most of the RNs who got excluded had an

* employment status of "not in labor force", as opposed to "unemployed".

* It's likely that these RNs not in the labor force have a different age profile compared to currently employed RNs.

* We can easily test this assumption.

* We'll create a variable that takes a value of 1 for all RNs, regardless of employment status, and a value of 2 for RNs

* currently employed in nursing.

generate rn_compare=.

replace rn_compare=1 if rn==1

replace rn_compare=2 if rn_employed==1

* Give the new variable a name.

label variable rn_compare "employed vs. not employed RNs"

* Label the values of the new variable to make the output easier to read

label define rn_compare_lbl 1 "Not currently employed RNs" 2 "Currently employed RNs"

* Attach these value labels to the variable

label values rn_compare rn_compare_lbl

* Now we'll estimate the median age for each group to see if it's different

table rn_compare [pw=pwgtp], c(median age)

* Note that we had to specify the use of the variable "PWGTP"; this is always the case when using Stata's "table" command,

* even if we've already instructed Stata to use "PWGTP" as our weight variable.

* The output will most likely show that RNs in sample who are not currently employed in nursing are much older than those that

* are currently employed in nursing.

* TABULATE SHARE OF EMPLOYED RNs WHO HOLD A BACHELOR'S IN NURSING DEGREE

* FOD1P & FOD2P are the variables identifying the field of study for bachelor's degree

* We want to identify all sample cases where field of bachelor's degree is reported as "nursing"

* FOD1P==6107 or FOD2P==6107

generate bsn=0

replace bsn=1 if fod1p==6107

replace bsn=1 if fod2p==6107

lab variable bsn "id field of bachelors degree is nursing"

* Now we can simply specify that we want only want to include employed RNs in our tabulation

svy, subpop(rn_employed): tab bsn, count cell format(%9.3g)

* TABULATE SHARE OF EMPLOYED RNs WHO HOLD A BACHELOR'S IN NURSING OR HIGHER DEGREE IN ANY FIELD

* First identify all sample cases where educational attainment is reported as master's or higher degree

* SCHL is the educational attainment variable

* Master's or higher degree includes values of SCHL>=22

generate bsn_higher=0

replace bsn_higher=1 if schl>=22

* Next recode your new variable to also include all of the sample cases where field of bachelor's degree is reported as "nursing"

replace bsn_higher=1 if bsn==1

label variable bsn_higher "id bsn or higher degree"

* Now we can simply specify that we want only want to include employed RNs in our tabulation

svy, subpop(rn_employed): tab bsn_higher, count cell format(%9.3g)

* DETERMINE THE RACIAL & ETHNIC COMPOSITION OF EMPLOYED RNs

* Race and ethnicity are treated as distinct concepts in the American Community Survey. Hispanic ethnicity is treated as a

* binary condition; either you are Hispanic or you aren't. However, people who identify with any race group can also be Hispanic.

* When describing the racial & ethnic composition of RNs (or any group), the convention is to distinguish race from Hispanic ethnicity.

* For example, the share of RNs who are White, non-Hispanic; or Asian, non-Hispanic. In contrast, we allow RNs who identify as Hispanic

* to be from any race group.

* The variable describing Hispanic ethnicity in the ACS is HISP.

* The variable describing race is RAC1P; the categories are identical to those used in Census data.

* Since we want to distinguish race and ethnicity when describing the racial & ethnic composition of RNs, we need to create a new variable.

* However, for some race groups, when we distinguish race from ethnicity, the number of RNs in sample is going to be too small

* to generate meaningful estimates. For example, RNs who are Native Hawaiian or Pacific Islander, but not Hispanic will be very few in number.

* These smaller groups should be combined to form a larger group with an adequate number of sample observations.

* Before creating our new variable that distinguishes race from ethnicity, we can identify which race groups are going to have too few sample

* observations by cross-tabulating race and Hispanic ethnicity

* To make this easier, we'll first create a binary variable that collapses the 23 different categories of Hispanic ethnicity into one.

* This variable will take a value of 1 if the person in sample is Hispanic, and 0 if not.

```
generate hisp_id=0

replace hisp_id=1 if hisp!=1

label variable hisp_id "id hispanic ethnicity"

label define hisp_id_lbl 0 "Non-Hisp" 1 "Hisp"
```

```
label values hisp_id hisp_id_lbl


* Next we'll assign labels to the different values of the race variable, to make it easier to read the output.


#delimit ;


label define rac1p_lbl

1 "White"

2 "Blk/Afr Am"

3 "Amer Ind"

4 "AK Native"

5 "Amer Ind & AK Native"

6 "Asian"

7 "HI/Pac Island"

8 "Some other race"

9 "Two or more race" ;


label values rac1p rac1p_lbl ;


#delimit cr


* Now we cross-tabulate race with our new binary variable for Hispanic ethnicity, for the population of currently employed RNs.


tab rac1p hisp_id if rn_employed==1
```

* The output will show you which race groups need to be combined, focus on the column showing race, by non-Hispanic ethnicity.

* In the California data, the number of sample observations describing either Native American RNs, or Native Hawaiian/Pacific Islander RNs

* is too small to use for generating estimates. In your state, you may need to combine even more groups due to small sample counts.

* It's also worth pointing out that one of the race categories is "Some other race", and race is cross-tabulated with Hispanic ethnicity

* most of these cases are going to be also coded as "Hispanic". But what about sample cases where race is coded

* as "some other race" and ethnicity is coded as "not Hispanic"? For all intents and purposes, these could be considered "unknown", or "unreported".

* When creating our new variable to distinguish race & ethnicity, treat these sample cases as missing information.

* Next we'll create our variable that distinguishes race & ethnicity, taking into account that some race groups need to be combined,

* and that sample cases where race is coded as "some other race" and Hispanic ethnicity is coded as "not Hispanic" are going to

* be treated as missing information. In other words, we're going to describe the racial & ethnic composition only for RNs whose

* race & ethnicity is known.

* I'm going to identify the following groups in this new variable:

* White, not Hispanic

* Black or African American, not Hispanic

* Asian, not Hispanic

* Other race, not Hispanic (this will include all Native American groups, Native Hawaiian/Pacific Islander, and two or more races)

* Hispanic, any race

* Based on the data for your state, you may have to recombine the race & ethnicity variables somewhat differently.

generate race_eth=.

* White, not Hispanic

replace race_eth=1 if rac1p==1 & hisp_id==0

* Black or African American, not Hispanic

replace race_eth=2 if rac1p==2 & hisp_id==0

* Asian, not Hispanic

replace race_eth=3 if rac1p==6 & hisp_id==0

* Other race, not Hispanic (this will include all Native American groups, Native Hawaiian/Pacific Islander, and two or more races)

replace race_eth=4 if rac1p==3 & hisp_id==0

replace race_eth=4 if rac1p==4 & hisp_id==0

replace race_eth=4 if rac1p==5 & hisp_id==0

replace race_eth=4 if rac1p==7 & hisp_id==0

```
replace race_eth=4 if rac1p==9 & hisp_id==0

* Hispanic, any race

replace race_eth=5 if hisp_id==1

* I haven't assigned any code for "some other race", so Stata will treat it was missing information.

* Now we'll assign labels to the values of the new race/ethnicity variable just created to make the output easier to read.

#delimit ;

label define race_eth_lbl

1 "White NH"

2 "Blk/Afr Amer NH"

3 "Asian NH"

4 "Other race NH"

5 "Hispanic" ;

#delimit cr

* Now we'll attach these value labels to the values of tthe new race/ethnicity variable so that they display in our output.

label values race_eth race_eth_lbl
```

* Finally, we'll tabulate the racial & ethnic composition of employed RNs.

svy, subpop(rn_employed): tab race_eth, count cell format(%9.3g) obs