```
> # analyze survey data for free (http://asdfree.com) with the r language
> # panel study of income dynamics
> # replication of umich statistics
>
> # note that the folks at the panel study of income dynamics claim there's no "recommended"
variance calculation
> # so i asked one of the authors of this paper -
> # http://psidonline.isr.umich.edu/Publications/Papers/tsp/2011-05_Heeringa_Berglung_Khan.pdf
> # - to provide me with example sas output that i could precisely match.  they did.  here it is:
> #
https://raw.github.com/ajdamico/usgsd/master/Panel%20Study%20of%20Income%20Dynamics/umich%20outpu
t.pdf?raw=TRUE
> # and this script matches those results exactly.  but it only works exactly on the 1968-2009
individual cross-year file.
> # the 1968-2009 individual cross-year file is already obsolete, so if you run this script,
> # you'll get every-so-slightly different numbers.  to assuage your concerns, i've run the
whole script on the 1968-2009 file
> # and saved the output here:
> #
https://raw.github.com/ajdamico/usgsd/master/Panel%20Study%20of%20Income%20Dynamics/replication%2
0output.pdf?raw=TRUE
> # hope that works for you.  if it doesn't, e-mail psidhelp@umich.edu and ask for the
1968-2009 file yourself ;)
>
> # if you have never used the r language before,
> # watch this two minute video i made outlining
> # how to run this script from start to finish
> # http://www.screenr.com/Zpd8
>
> # anthony joseph damico
> # ajdamico@gmail.com
>
> # if you use this script for a project, please send me a note
> # it's always nice to hear about how people are using this stuff
>
> # for further reading on cross-package comparisons, see:
> # http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Damico.pdf
>
>
> # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # #
>
################################################################################################
####################
> # prior to running this analysis script, the umich individual cross-year file must be
downloaded to your local disk #
> # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # #
> #
https://raw.github.com/ajdamico/usgsd/master/Panel%20Study%20of%20Income%20Dynamics/download%20al
l%20microdata.R  #
> # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # #
> # that script will place all necessary psid files whever you specified, probably the "C:/My
Directory/PSID/" folder #
>
```

```
################################################################################################
####################
> # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # #
>
>
>
################################################################################################
####
> # replicate example output provided by the authors of the psid's design-based sampling error
report #
>
>
>
> # set your working directory.
> # the R data file (.rda) should have been stored within this folder
> # use forward slashes instead of back slashes
>
> # uncomment this line by removing the `#` at the front..
> setwd( "C:/My Directory/PSID/" )
> # ..in order to set your current working directory
>
>
> # remove the # in order to run this install.packages line only once
> # install.packages( "survey" )
>
>
>
> # no need to edit anything below this line #
>
>
> # # # # # # # # #
> # program start #
> # # # # # # # # #
>
> require(survey)# load survey package (analyzes complex design surveys)
Loading required package: survey

Attaching package: ‘survey’

The following object is masked from ‘package:graphics’:

    dotchart

>
>
> # load the individual cross-year file
> load( "ind.rda" )
>
>
> # limit the file to only the variables needed
> KeepVars <-
+ c(
+ "er32000" ,# sex
+ "er34020" , # education level
+ "er31997" ,# primary sampling unit
```

```
+ "er31996" ,# strata
+ "er34046"# weights
+ )
>
>
> # create a "skinny"data.frame object that only contains the
> # columns you need for this analysis,
> # specified in character vector 'KeepVars'
> x <- ind[ , KeepVars ]
>
>
> # to free up memory, remove the full r data frame
> rm( ind )
>
> # clear up RAM
> gc()
         used (Mb) gc trigger  (Mb)  max used  (Mb)
Ncells 201021  5.4     407500  10.9    350000   9.4
Vcells 539882  4.2   85318147 651.0 105952641 808.4
>
>
> # perform all recodes on the `individual` table #
>
> # create a `completed_ed` column that simply blanks out
> # `er34020` values of 98 or 99, but otherwise uses
> # whatever's already in `er34020`
> x <-
+ transform(
+ x ,
+ completed_ed = ifelse( er34020 %in% 98:99 , NA , er34020 )
+ )
>
> # end of all recodes #
>
> # create survey design object with PSID design information
> y <-
+ svydesign(
+ ~er31997 ,
+ strata = ~er31996 ,
+ data = x ,
+ weights = ~er34046 ,
+ nest = TRUE
+ )
>
> # extract the unweighted available number of observations
> unwtd.count( ~completed_ed , y )
       counts SE
counts  23461  0
>
> # calculate the mean of the `completed_ed` column created above..
> c_ed <- svymean( ~completed_ed , y , na.rm = TRUE )
> # ..to create a `svystat` object `c_ed`
> class( c_ed )
[1] "svystat"
>
> # extract the actual statistic..
```

```
> coef( c_ed )
completed_ed
    9.797017
>
> # ..the standard error..
> SE( c_ed )
             completed_ed
completed_ed   0.08946531
>
> # ..and confidence intervals, both default..
> confint( c_ed )
                2.5 %   97.5 %
completed_ed 9.621668 9.972365
>
> # ..and matching sas.
> confint( c_ed , df = degf( y ) )
                2.5 %   97.5 %
completed_ed 9.618235 9.975799
>
> # run the same query, broken down by the sex variable
> c_ed_by_sex <-
+ svyby(
+ ~completed_ed ,
+ ~er32000 ,
+ y ,
+ svymean ,
+ na.rm = TRUE
+ )
>
>
> # extract the actual statistic..
> coef( c_ed_by_sex )
       1        2
9.647527 9.943074
>
> # ..the standard error..
> SE( c_ed_by_sex )
[1] 0.10723929 0.08272307
>
> # ..and confidence intervals, both default..
> confint( c_ed_by_sex )
     2.5 %    97.5 %
1 9.437342  9.857712
2 9.780940 10.105208
>
> # ..and matching sas.
> confint( c_ed_by_sex , df = degf( y ) )
     2.5 %    97.5 %
1 9.433227  9.861828
2 9.777765 10.108383
>
> # sum up the counts by gender
> svytotal( ~factor( er32000 ) , y )
                     total      SE
factor(er32000)1 148094393 6225232
factor(er32000)2 153388460 6221975
```

```
factor(er32000)9         0       0
>
> # calculate the proportion of each gender,
> # also printing the results to the screen
> ( c_sex <- svymean( ~factor( er32000 ) , y ) )
                    mean     SE
factor(er32000)1 0.49122 0.0033
factor(er32000)2 0.50878 0.0033
factor(er32000)9 0.00000 0.0000
>
> # then, all by its lonesome, print the standard error too.
> SE( c_sex )
factor(er32000)1 factor(er32000)2 factor(er32000)9
      0.00330169       0.00330169       0.00000000
>
>
> # for more details on how to work with data in r
> # check out my two minute tutorial video site
> # http://www.twotorials.com/
>
> # dear everyone: please contribute your script.
> # have you written syntax that precisely matches an official publication?
> message( "if others might benefit, send your code to ajdamico@gmail.com" )
if others might benefit, send your code to ajdamico@gmail.com
> # http://asdfree.com needs more user contributions
>
> # let's play the which one of these things doesn't belong game:
> # "only you can prevent forest fires" -smokey bear
> # "take a bite out of crime" -mcgruff the crime pooch
> # "plz gimme your statistical programming" -anthony damico
```