Module 3 examples - R code

September 2019

Examples illustrating the over-sampling of some demographic groups and demonstrating the importance of using weights in analyses

# Load required packages

library(survey)  
library(dplyr)

# Display Version Information   
cat("R package versions:\n")

## R package versions:

for (p in c("base", "survey","dplyr")) {   
 cat(p, ": ", as.character(packageVersion(p)), "\n")  
}

## base : 3.5.2   
## survey : 3.36   
## dplyr : 0.8.0.1

# Data preparation

## Download & Read SAS Transport Files

# function to download the required survey cycles for a component file   
downloadNHANES <- function(fileprefix){  
 print (fileprefix)  
 outdf <- data.frame(NULL)  
 for (j in 1:length(letters)){  
 urlstring <- paste('https://wwwn.cdc.gov/nchs/nhanes/',yrs[j],'/',fileprefix,letters[j],'.XPT', sep='')  
 download.file(urlstring, tf <- tempfile(), mode="wb")  
 tmpframe <- foreign::read.xport(tf)  
 outdf <- bind\_rows(outdf, tmpframe)  
 }  
 return(outdf)  
}  
  
# Specify the survey cycles required, with corresponding file suffixes  
yrs <- c('2015-2016')  
letters <- c('\_i')  
  
# Download data for each component  
# Demographic (DEMO)  
DEMO <- downloadNHANES('DEMO')

## [1] "DEMO"

# BPX blood pressure exam   
BPX <- downloadNHANES('BPX')

## [1] "BPX"

# BPQ blood pressure questionnaire   
BPQ <- downloadNHANES('BPQ')

## [1] "BPQ"

# Merge component files  
# Keep all records in DEMO, even if that SEQN does not match to BPQ or BPX files  
one\_tmp <- left\_join(DEMO, select(BPX, "SEQN", starts\_with("BPXSY"), starts\_with("BPXDI")), by="SEQN") %>%  
 left\_join(., select(BPQ, "SEQN", "BPQ050A") , by="SEQN")

## Create derived analysis variables (using dplyr functions)

df <- mutate(one\_tmp,   
 # create indicator for overall summary  
 one = 1,  
 # Hypertension prevalence  
 # Count Number of Nonmissing SBPs & DBPs  
 n\_sbp=rowSums(!is.na(select(one\_tmp, starts\_with("BPXSY")))),  
 n\_dbp=rowSums(!is.na(select(one\_tmp, starts\_with("BPXDI"))))) %>%  
 # Set DBP Values Of 0 To Missing For Calculating Average  
 mutate\_at(vars(starts\_with("BPXDI")), list(~na\_if(., 0))) %>%  
 mutate(  
 # Calculate Mean Systolic and Diastolic (over non-missing values)   
 mean\_sbp=rowMeans(select(., starts\_with("BPXSY")), na.rm=TRUE),  
 mean\_dbp=rowMeans(select(., starts\_with("BPXDI")), na.rm=TRUE),  
 # Create 0/1 indicator for hypertension  
 # "Old" Hypertensive Category variable: taking medication or measured BP > 140/90   
 # as used in NCHS Data Brief No. 289  
 # Variable bpq050a: now taking prescribed medicine for hypertension, 1 = yes   
 htn\_old=case\_when( mean\_sbp>=140 | mean\_dbp>=90 | BPQ050A ==1 ~ 1,  
 n\_sbp > 0 & n\_dbp > 0 ~ 0),  
 # for reference: "new" definition of hypertension prevalence, based on taking medication or measured BP > 130/80   
 # From 2017 ACC/AHA hypertension guidelines   
 # Not used in Data Brief No. 289  
 htn\_new=case\_when( mean\_sbp>=130 | mean\_dbp>=80 | BPQ050A ==1 ~ 1,  
 n\_sbp > 0 & n\_dbp > 0 ~ 0),  
 # Create race and Hispanic ethnicity categories for oversampling analysis   
 # combined Non-Hispanic white and Non-Hispanic other and multiple races, to approximate the sampling domains  
 race1 = factor(c(3, 3, 4, 1, NA, 2, 4)[RIDRETH3],  
 labels = c('NH Black','NH Asian', 'Hispanic', 'NH White and Other')),  
 # Create race and Hispanic ethnicity categories for hypertension analysis   
 raceEthCat= factor(c(4, 4, 1, 2, NA, 3, 5)[RIDRETH3],  
 labels = c('NH White', 'NH Black', 'NH Asian', 'Hispanic', 'NH Other/Multiple')),  
 # Create age categories for adults aged 18 and over: ages 18-39, 40-59, 60 and over  
 ageCat\_18 = cut(RIDAGEYR,  
 breaks = c(17, 39, 59, Inf),  
 labels = c('18-39','40-59','60+')),   
 # Define subpopulation of interest: non-pregnant adults aged 18 and over who have at least 1 valid systolic OR diastolic BP measure   
 inAnalysis= (RIDAGEYR >=18 & ifelse(is.na(RIDEXPRG), 0, RIDEXPRG) != 1 & (n\_sbp > 0 | n\_dbp > 0))   
 )

# Estimates for graph - Distribution of race and Hispanic origin, NHANES 2015-2016

Module 3, Examples Demonstrating the Importance of Using Weights in Your Analyses  
Section “Adjusting for oversampling”  
Proportion of unweighted interview sample

df %>% count(race1) %>%   
 mutate(prop= round(n / sum(n)\*100, digits=1))

## # A tibble: 4 x 3  
## race1 n prop  
## <fct> <int> <dbl>  
## 1 NH Black 2129 21.4  
## 2 NH Asian 1042 10.5  
## 3 Hispanic 3229 32.4  
## 4 NH White and Other 3571 35.8

Proportion of weighted interview sample

df %>% count(race1, wt=WTINT2YR) %>%  
 mutate(prop= round(n / sum(n)\*100, digits=1))

## # A tibble: 4 x 3  
## race1 n prop  
## <fct> <dbl> <dbl>  
## 1 NH Black 37789477. 11.9  
## 2 NH Asian 17701706. 5.6  
## 3 Hispanic 55750392. 17.6  
## 4 NH White and Other 205239470. 64.9

Proportion of US population

# Input population totals from the American Community Survey, 2015-2016  
# available on the NHANES website: https://wwwn.cdc.gov/nchs/nhanes/responserates.aspx#population-totals  
# counts from tab "Both" (for both genders), total row (for all ages)  
  
data.frame(group=c('Non-Hispanic White and Other', 'Non-Hispanic Black', 'Non-Hispanic Asian', 'Hispanic'), n=c(194849491+10444206, 38418696, 17018259, 55750392 )) %>%  
 mutate(prop = round(n / sum(n)\* 100, digits=1))

## group n prop  
## 1 Non-Hispanic White and Other 205293697 64.9  
## 2 Non-Hispanic Black 38418696 12.1  
## 3 Non-Hispanic Asian 17018259 5.4  
## 4 Hispanic 55750392 17.6

# Comparison of weighted and unweighed estimates for hypertension, NHANES 2015-2016

Module 3, Examples Demonstrating the Importance of Using Weights in Your Analyses Section “Why weight?”

## Prevalence of hypertension among adults aged 18 and over, overall and by race and Hispanic origin group

### Unweighted estimates

Unweighted estimate - for adults aged 18 and over

df %>% filter(inAnalysis==1) %>% summarize(mean=round(mean(htn\_old)\*100, digits=1))

## mean  
## 1 35.9

Unweighted estimate - for adults aged 18 and over, by race and Hispanic origin

df %>% filter(inAnalysis==1 ) %>% group\_by(raceEthCat) %>% summarize(mean=mean(htn\_old)\*100, n=n())

## # A tibble: 5 x 3  
## raceEthCat mean n  
## <fct> <dbl> <int>  
## 1 NH White 37.0 1790  
## 2 NH Black 44.5 1173  
## 3 NH Asian 24.5 633  
## 4 Hispanic 33.4 1702  
## 5 NH Other/Multiple 34.0 206

### Weighted estimates

## WARNING

The following commands are intended to demonstrate the importance of using the sample weight in your analyses. The weighted estimate produces the correct **point estimates** for the prevalence of hypertension. However, your analysis must account for the complex survey design of NHANES (e.g. stratification and clustering), in order to produce correct **standard errors** (and confidence intervals, statistical tests, etc.). Do not use this step as a model for producing your own analyses!  
See the Continuous NHANES tutorial Module 4: Variance Estimation for a complete explanation of how to properly account for the complex survey design using commands in the “survey” package

Weighted estimates - for adults aged 18 and over

df %>% filter(inAnalysis==1) %>% summarize(mean=round(weighted.mean(htn\_old, WTMEC2YR)\*100, digits=1))

## mean  
## 1 32.1

Weighted estimate - for adults aged 18 and over, by race and Hispanic origin

df %>% filter(inAnalysis==1 ) %>% group\_by(raceEthCat) %>% summarize(mean=weighted.mean(htn\_old, WTMEC2YR)\*100)

## # A tibble: 5 x 2  
## raceEthCat mean  
## <fct> <dbl>  
## 1 NH White 33.4  
## 2 NH Black 40.1  
## 3 NH Asian 24.6  
## 4 Hispanic 23.1  
## 5 NH Other/Multiple 34.8

## Example of how to use the survey package to estimate the prevalence of hypertension, with correct standard errors

See Module 4: Variance Estimation for details

# Define survey design for overall dataset   
NHANES\_all <- svydesign(data=df, id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTMEC2YR, nest=TRUE)  
  
# Create a survey design object for the subset of interest   
# Subsetting the original survey design object ensures we keep the design information about the number of clusters and strata  
NHANES <- subset(NHANES\_all, inAnalysis==1)

Proportion and standard error, for adults aged 18 and over

svyby(~htn\_old, ~one, NHANES, svymean) %>% mutate(htn\_old = round(htn\_old\*100, digits=1), se=round(se\*100, digits=1))

## one htn\_old se  
## 1 1 32.1 1.2

Proportion and standard error, for adults aged 18 and over by race and Hispanic origin

svyby(~htn\_old, ~raceEthCat, NHANES, svymean) %>% mutate(htn\_old = round(htn\_old\*100, digits=1), se=round(se\*100, digits=1))

## raceEthCat htn\_old se  
## 1 NH White 33.4 1.6  
## 2 NH Black 40.1 2.1  
## 3 NH Asian 24.6 2.7  
## 4 Hispanic 23.1 2.3  
## 5 NH Other/Multiple 34.8 5.1

## Age distribution among Hispanic adults, weighted and unweighted

To support the statement in the tutorial text that the unweighted estimate over-represents Hispanic adults aged 60 and over, compared with their actual share of the Hispanic adult population.

### Unweighted age distribution among Hispanic adults in the analysis

df %>% filter(inAnalysis==1 & raceEthCat=='Hispanic') %>%   
 count(ageCat\_18) %>%   
 mutate(prop= round(n / sum(n)\*100, digits=1))

## # A tibble: 3 x 3  
## ageCat\_18 n prop  
## <fct> <int> <dbl>  
## 1 18-39 628 36.9  
## 2 40-59 517 30.4  
## 3 60+ 557 32.7

Unweighted, Hispanic adults aged 60 and over comprise 33% of Hispanic adults in the analysis sample.

### Weighted age distribution among Hispanic adults in the analysis population

df %>% filter(inAnalysis==1 & raceEthCat=='Hispanic') %>% count(ageCat\_18, wt=WTMEC2YR) %>%  
 mutate(prop= round(n / sum(n)\*100, digits=1))

## # A tibble: 3 x 3  
## ageCat\_18 n prop  
## <fct> <dbl> <dbl>  
## 1 18-39 17906014. 50.1  
## 2 40-59 12403266. 34.7  
## 3 60+ 5439742. 15.2

When properly weighted, Hispanic adults aged 60 and over comprise 15% of Hispanic adults in the US non-institutionalized civilian population