432 Class 03

https://thomaselove.github.io/432-2023/

2023-01-24

Today's Agenda

Incorporating Survey Weights …

1. In estimating means and confidence intervals
2. In building linear regression models
3. Into a more detailed t-test approach using NHANES

Primary Source:
https://bookdown.org/rwnahhas/RMPH/survey-design.html

Today's R Setup

```
knitr::opts_chunk$set(comment = NA)

library(broom)
library(janitor)
library(gtsummary)
library(knitr)
library(mosaic)

library(nhanesA) # data source
library(survey) # survey-specific tools

library(tidyverse)

theme_set(theme_bw())
```

Section 1

Incorporating survey weights (an introduction)

What are survey weights?

In many surveys, each sampled subject is assigned a weight that is equivalent to the reciprocal of his/her probability of selection into the sample.

$$\text{Sample Subject's Weight} = \frac{1}{Prob(selection)}$$

but more sophisticated sampling designs require more complex weighting schemes. Usually these are published as part of the survey data.

There are several packages available to help incorporate survey weights in R, but I will demonstrate part of the `survey` package today.

An NHANES Example

Let's use the NHANES 2013-14 data and pull in both the demographics (DEMO_H) and total cholesterol (TCHOL_H) databases.

```
demo_raw <- nhanes('DEMO_H')
tchol_raw <- nhanes('TCHOL_H')
```

Detailed descriptions available at

- https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DEMO_H.htm
- https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/TCHOL_H.htm

Weighting in NHANES

Weights are created in NHANES to account for the complex survey design. A sample weight is assigned to each sample person. It is a measure of the number of people in the population represented by that sample person.

The sample weight is created in three steps:

1. the base weight is computed, which accounts for the unequal probabilities of selection given that some demographic groups were over-sampled;
2. adjustments are made for non-response; and
3. post-stratification adjustments are made to match estimates of the U.S. civilian non-institutionalized population available from the Census Bureau.

Source: https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module3.aspx

Weights in our NHANES data

The `DEMO` file contains two kinds of sampling weights:

- the interview weight (`WTINT2yr`), and
- the MEC exam weight (`WTMEC2yr`)

NHANES also provides several weights for subsamples. A good rule for NHANES is to identify the variable of interest that was collected on the smallest number of respondents. The sample weight that applies to that variable is the appropriate one to use in your analysis.

In our first case, we will study total cholesterol and use the weights from the MEC exam.

What Variables Do We Need?

- SEQN = subject identifying code
- RIAGENDR = sex (1 = M, 2 = F)
- RIDAGEYR = age (in years at screening, topcode at 80)
- DMQMILIZ = served active duty in US Armed Forces (1 = yes, 2 = no)
- RIDSTATR = 2 if subject took both interview and MEC exam
- WTMEC2YR - Full sample 2 year MEC exam weight
- LBXTC = Total Cholesterol (mg/dl) - this is our outcome

The first five of these came from the DEMO_H file, and the first and last comes from TCHOL_H.

Merge the DEMO and TCHOL files

```
dim(demo_raw)
```

```
[1] 10175    47
```

```
dim(tchol_raw)
```

```
[1] 8291    3
```

```
joined_df <- inner_join(demo_raw, tchol_raw, by = c("SEQN"))

dim(joined_df)
```

```
[1] 8291   49
```

Create a small analytic tibble

```
nh1314 <- joined_df |> # has n = 8291
    tibble() |>
    filter(complete.cases(LBXTC)) |> # now n = 7624
    filter(RIDSTATR == 2) |> # still 7624
    filter(RIDAGEYR > 19 & RIDAGEYR < 40) |> # now n = 1802
    filter(DMQMILIZ < 3) |> # drop 7 = refused, n = 1801
    mutate(FEMALE = RIAGENDR - 1,
           AGE = RIDAGEYR,
           US_MIL = ifelse(DMQMILIZ == 1, 1, 0),
           WT_EX = WTMEC2YR,
           TOTCHOL = LBXTC) |>
    select(SEQN, FEMALE, AGE, TOTCHOL, US_MIL, WT_EX)
```

Our nh1314 analytic sample: Variables

```
nh1314 |> tabyl(FEMALE, US_MIL) |>
  adorn_totals(where = c("row", "col")) |> adorn_title()
```

```
       US_MIL
 FEMALE      0  1 Total
      0    829 45   874
      1    921  6   927
  Total   1750 51  1801
```

```
df_stats(~ AGE + TOTCHOL, data = nh1314) |>
  rename(med = median, na = missing) |>
  kable(digits = 1)
```

response	min	Q1	med	Q3	max	mean	sd	n	na
AGE	20	24	30	34	39	29.5	5.8	1801	0
TOTCHOL	69	156	178	203	417	181.0	37.4	1801	0

Our nh1314 analytic sample: Weights

Each weight represents the number of people exemplified by that subject.

```
favstats(~ WT_EX, data = nh1314) |>
  rename(na = missing) |>
  kable(digits = 1)
```

min	Q1	median	Q3	max	mean	sd	n	na
8430.5	24694	34642.1	59560.7	125680.3	44528.7	26027.4	1801	0

Create nh_design survey design

```r
nh_design <-
    svydesign(
        id = ~ SEQN,
        weights = ~ WT_EX,
        data = nh1314)

nh_design <- update( nh_design, one = 1)

## this one = 1 business will help us count
```

Unweighted counts, overall and by sex

```
sum(weights(nh_design, "sampling") != 0)
```

```
[1] 1801
```

```
svyby( ~ one, ~ FEMALE, nh_design, unwtd.count)
```

```
  FEMALE counts se
0      0    874  0
1      1    927  0
```

```
svyby( ~ one, ~ FEMALE + US_MIL, nh_design, unwtd.count)
```

```
    FEMALE US_MIL counts se
0.0      0      0    829  0
1.0      1      0    921  0
0.1      0      1     45  0
1.1      1      1      6  0
```

Weighted counts, overall and by groups

Weighted size of the generalizable population, overall and by groups.

```
svytotal( ~ one, nh_design )
```

```
       total      SE
one 80196108 1104558
```

```
svyby( ~ one, ~ FEMALE * US_MIL, nh_design, svytotal)
```

```
    FEMALE US_MIL        one        se
0.0      0      0 37185326.4 1225990.7
1.0      1      0 40151728.1 1192408.4
0.1      0      1  2509429.8  419477.5
1.1      1      1   349624.1  157476.1
```

Use the survey design to get weighted means

What is the mean of total cholesterol, overall and in groups?

```
svymean( ~ TOTCHOL, nh_design, na.rm = TRUE)
```

```
           mean     SE
TOTCHOL 181.25 1.0172
```

```
svyby(~ TOTCHOL, ~ FEMALE + US_MIL, nh_design,
      svymean, na.rm = TRUE)
```

```
    FEMALE US_MIL  TOTCHOL       se
0.0      0      0 182.3569 1.575994
1.0      1      0 180.0248 1.368408
0.1      0      1 186.6966 5.354835
1.1      1      1 164.1984 6.535223
```

Unweighted Mean of TOTCHOL

```
nh1314 |>
  summarise(n = n(), mean(TOTCHOL)) |>
  kable(digits = 2)
```

n	mean(TOTCHOL)
1801	181.01

Note that we're using `summarise` to ensure that we get the **dplyr** package's version of `summarize`.

Unweighted Group Means of TOTCHOL

```
nh1314 |> group_by(FEMALE, US_MIL) |>
  summarise(n = n(), mean(TOTCHOL)) |>
  kable(digits = 2)
```

FEMALE	US_MIL	n	mean(TOTCHOL)
0	0	829	182.22
0	1	45	187.11
1	0	921	179.71
1	1	6	169.50

Measures of uncertainty (Survey-Weighted)

Mean of total cholesterol within groups with 90% CI?

```
grouped_result <- svyby(~ TOTCHOL, ~ FEMALE + US_MIL,
                        nh_design, svymean, na.rm = TRUE)
coef(grouped_result)
```

```
     0.0      1.0      0.1      1.1
182.3569 180.0248 186.6966 164.1984
```

```
confint(grouped_result, level = 0.90)
```

```
         5 %     95 %
0.0 179.7646 184.9492
1.0 177.7739 182.2756
0.1 177.8887 195.5045
1.1 153.4489 174.9478
```

- Get standard errors with `se(grouped_result)`, too.

Placing estimated means in `res`

```
res <- tibble(
  type = rep(c("Unweighted", "Survey-Weighted"),4),
  female = c(rep("Female", 4), rep("Male", 4)),
  us_mil = rep(c("Military", "Military",
                 "Not Military", "Not Military"), 2),
  MEAN = c(169.5, 164.1984, 179.71, 180.0248,
           187.11, 186.6966, 182.22, 182.3569) )
```

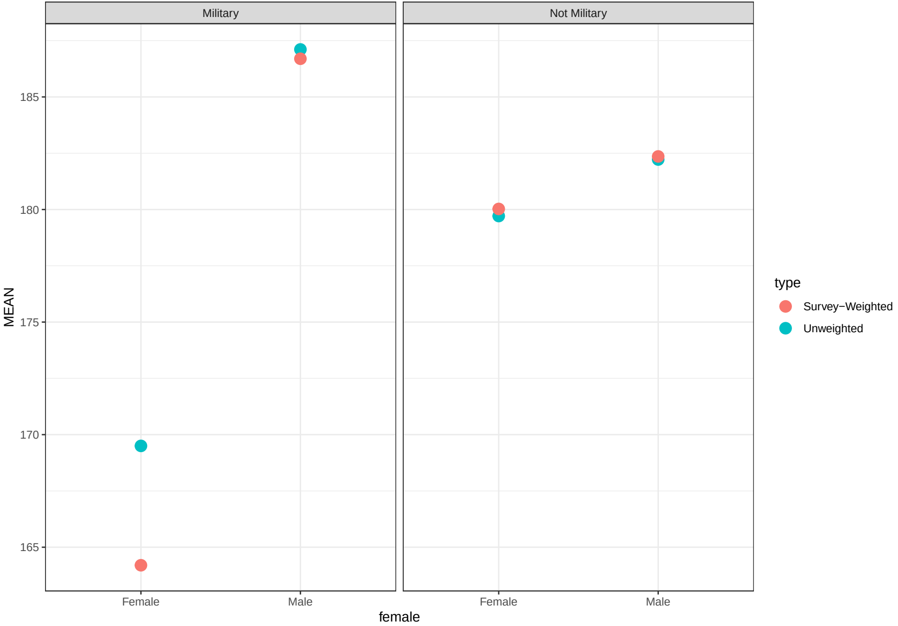
The Estimated Means

```
res |> kable(digits = 1)
```

type	female	us_mil	MEAN
Unweighted	Female	Military	169.5
Survey-Weighted	Female	Military	164.2
Unweighted	Female	Not Military	179.7
Survey-Weighted	Female	Not Military	180.0
Unweighted	Male	Military	187.1
Survey-Weighted	Male	Military	186.7
Unweighted	Male	Not Military	182.2
Survey-Weighted	Male	Not Military	182.4

```
ggplot(res, aes(x = female, y = MEAN, col = type)) +
  geom_point(size = 4) +
  facet_wrap(~ us_mil) ## plot shown on next slide
```

Plotting the Estimated Means



Section 2

Building Models

Models for TOTCHOL in our nh1314 data

First, we'll ignore the weighting, and fit one model with main effects of all three predictors (model mod1) and then a second model which incorporates an interaction of FEMALE and US_MIL.

```
mod1 <- lm(TOTCHOL ~ AGE + FEMALE + US_MIL, data = nh1314)

mod2 <- lm(TOTCHOL ~ AGE + FEMALE * US_MIL, data = nh1314)
```

The interaction term means that the effect of FEMALE on TOTCHOL depends on the US_MIL status.

Unweighted Model mod1 (no interaction)

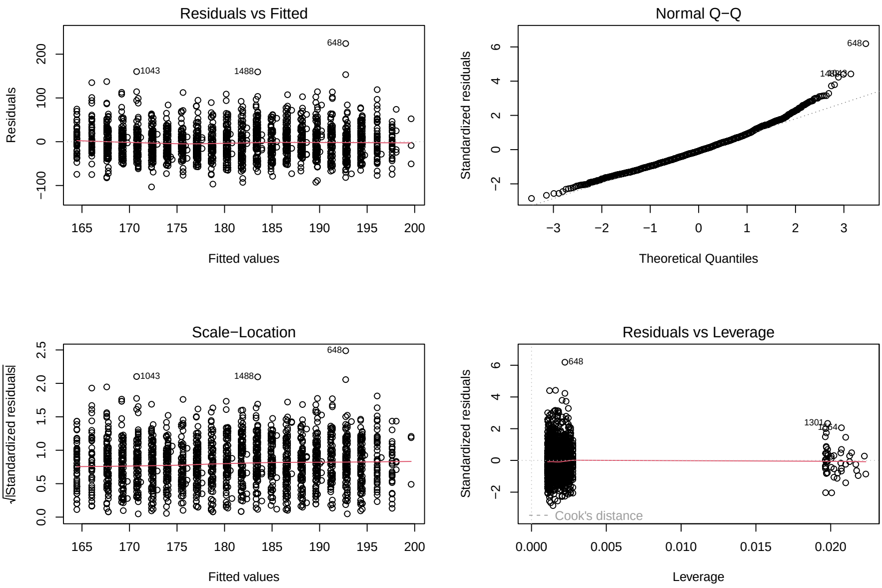
```
tidy(mod1, conf.int = TRUE, conf.level = 0.90) |>
  select(-statistic) |> kable(digits = 2)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	136.35	4.49	0.00	128.95	143.74
AGE	1.57	0.15	0.00	1.33	1.81
FEMALE	-3.31	1.73	0.06	-6.16	-0.47
US_MIL	2.00	5.20	0.70	-6.56	10.57

```
glance(mod1) |> select(r2 = r.squared, adjr2 = adj.r.squared,
    AIC, BIC, sigma, nobs, df) |> kable(dig = c(4,4,1,1,3,0,0)
```

r2	adjr2	AIC	BIC	sigma	nobs	df
0.061	0.0594	18052.7	18080.2	36.28	1801	3

Residuals for Model `mod1`

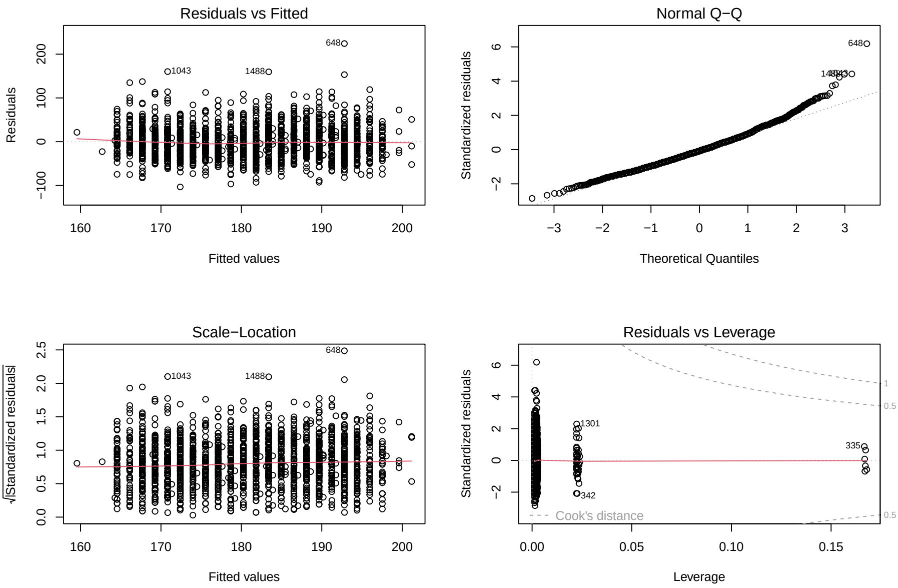


Unweighted Model `mod2` (with interaction)

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	136.30	4.49	0.00	128.91	143.69
AGE	1.57	0.15	0.00	1.33	1.81
FEMALE	-3.15	1.74	0.07	-6.01	-0.29
US_MIL	3.64	5.55	0.51	-5.50	12.78
FEMALE:US_MIL	-13.34	15.87	0.40	-39.45	12.77

r2	adjr2	AIC	BIC	sigma	nobs	df
0.0613	0.0593	18054	18087	36.282	1801	4

Residuals for Model `mod2`



Perform a survey-weighted generalized linear model

Again, we'll run two models, first without and second with an interaction term between FEMALE and US_MIL.

```
glm1_res <- svyglm(
    TOTCHOL ~ AGE + FEMALE + US_MIL,
    nh_design, family = gaussian())
```

```
glm2_res <- svyglm(
    TOTCHOL ~ AGE + FEMALE * US_MIL,
    nh_design, family = gaussian())
```

Gaussian family used to generate linear regressions here.

Model 1 Results

```
tidy(glm1_res, conf.int = TRUE, conf.level = 0.90) |>
  select(-statistic) |> kable(digits = 2)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	137.13	5.00	0.00	128.89	145.36
AGE	1.56	0.17	0.00	1.29	1.84
FEMALE	-3.21	2.01	0.11	-6.52	0.09
US_MIL	0.59	5.04	0.91	-7.70	8.89

```
glance(glm1_res) |>
  select(nobs, AIC, BIC, everything()) |> kable(digits = 1)
```

nobs	AIC	BIC	null.deviance	df.null	deviance	df.residual
1801	21.6	2344965	2498023	1800	2344935	1797

Model 2 Results

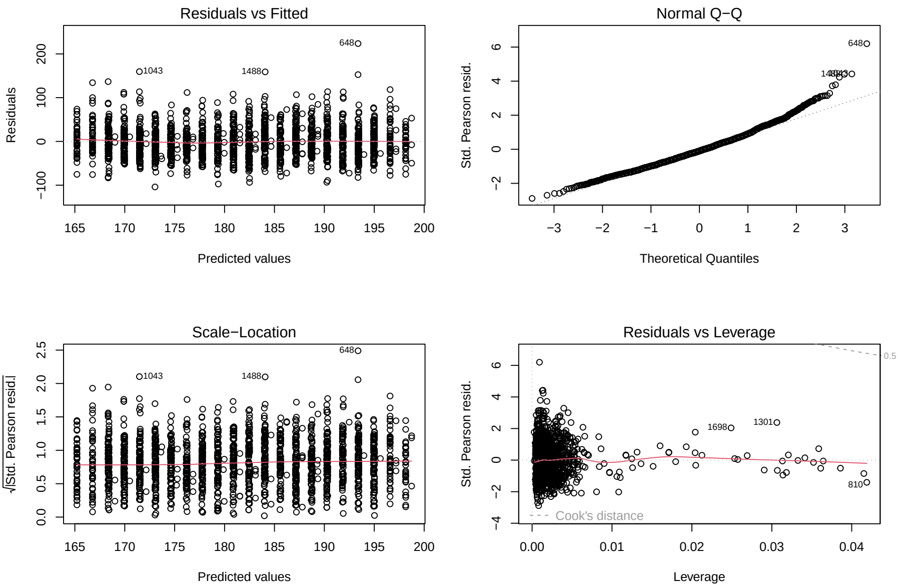
```
tidy(glm2_res, conf.int = TRUE, conf.level = 0.90) |>
  select(-statistic) |> kable(digits = 2)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	136.86	5.01	0.00	128.63	145.10
AGE	1.57	0.17	0.00	1.29	1.85
FEMALE	-2.87	2.03	0.16	-6.21	0.47
US_MIL	3.43	5.47	0.53	-5.58	12.43
FEMALE:US_MIL	-22.07	8.55	0.01	-36.14	-7.99

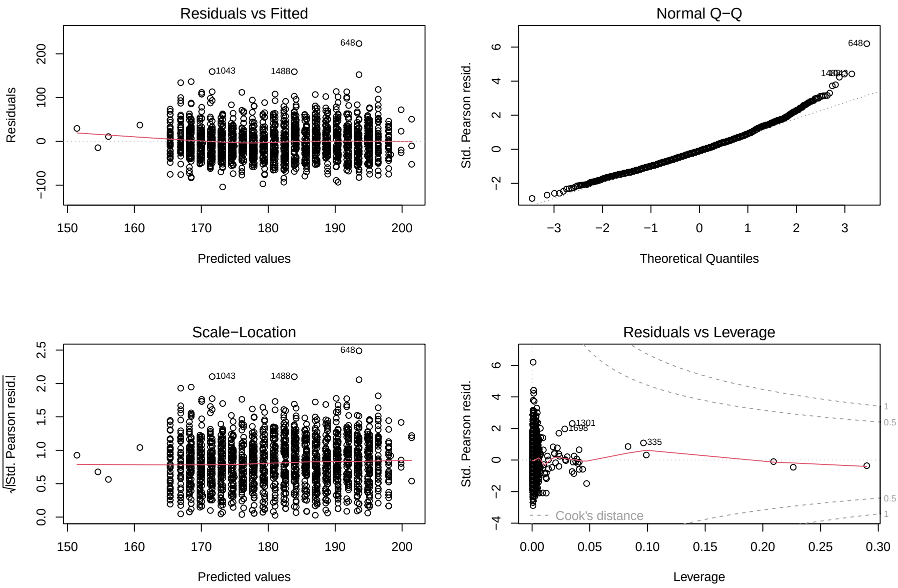
```
glance(glm2_res) |>
  select(nobs, AIC, BIC, everything()) |> kable(digits = 1)
```

nobs	AIC	BIC	null.deviance	df.null	deviance	df.residual
1801	22.2	2341671	2498023	1800	2341633	1796

Residuals for Model `glm1_res`



Residuals for Model `glm2_res`



Section 3

A More Complete Weighted NHANES Analysis (See https://wwwn.cdc.gov/nchs/data/tutorials/DB303_Fig1_R.R)

New Question, New Data

We are now interested in looking at the percentage of persons aged 20 and over with depression, by age and sex, in the US in 2013-2016. Start by pulling in the relevant data using `nhanesA`...

```
DEMO_H <- nhanes('DEMO_H') |> select(SEQN, RIAGENDR,
                    RIDAGEYR, SDMVSTRA, SDMVPSU, WTMEC2YR)
DEMO_I <- nhanes('DEMO_I') |> select(SEQN, RIAGENDR,
                    RIDAGEYR, SDMVSTRA, SDMVPSU, WTMEC2YR)
DEMO <- bind_rows(DEMO_H, DEMO_I)
DPQ_H <- nhanes('DPQ_H')
DPQ_I <- nhanes('DPQ_I')
DPQ <- bind_rows(DPQ_H, DPQ_I)
```

Merge DEMO and DPQ files and create derived variables

```
dat2 <- left_join(DEMO, DPQ, by = "SEQN") |> tibble() |>
  # Set 7=Refused and 9=Don't Know To NA
  mutate(across(.cols = DPQ010:DPQ090,
                ~ ifelse(. >=7, NA, .))) %>%
  mutate(one = 1,
         PHQ9_score = rowSums(select(. , DPQ010:DPQ090)),
         Depression = ifelse(PHQ9_score >= 10, 100, 0),
         Sex = factor(RIAGENDR, labels = c("M", "F")),
         Age_group = cut(RIDAGEYR,
            breaks = c(-Inf, 19, 39, 59, Inf),
            labels = c("Under 20", "20-39", "40-59", "60+")),
         WTMEC4YR = WTMEC2YR/2,
         inAnalysis = (RIDAGEYR >= 20 & !is.na(PHQ9_score))) |
  select(-starts_with("DPQ"))
```

Define Survey Design

Here's the survey design for the overall data set:

```
NH_des_all <- svydesign(data = dat2, id = ~ SDMVPSU,
  strata = ~ SDMVSTRA, weights = ~ WTMEC4YR, nest = TRUE)

dim(NH_des_all)
```

```
[1] 20146    13
```

Here's the survey design object for the subset of interest: adults aged 20 and over with a valid PHQ-9 depression score:

```
NH_des_dat2 <- NH_des_all |> subset(inAnalysis)

dim(NH_des_dat2)
```

```
[1] 9942   13
```

Define a function to call svymean and unweighted count

```
ourSummary <- function(varformula, byformula, design){
  # Get mean, stderr, and unweighted sample size
  c <- svyby(varformula, byformula, design, unwtd.count )
  p <- svyby(varformula, byformula, design, svymean )
  outSum <- left_join(select(c,-se), p)
  outSum
}
```

Estimate overall prevalence of depression

```
ourSummary(~ Depression, ~ one, NH_des_dat2)
```

```
  one counts Depression        se
1   1   9942   8.056844 0.3599894
```

Estimate prevalence of depression in various strata

```r
## By sex
ourSummary(~ Depression, ~ Sex, NH_des_dat2)
```

```
  Sex counts Depression        se
1   M   4821   5.549344 0.4293217
2   F   5121  10.427654 0.5658239
```

```r
## By age
ourSummary(~ Depression, ~ Age_group, NH_des_dat2)
```

```
  Age_group counts Depression        se
1     20-39   3328   7.744613 0.5236944
2     40-59   3307   8.429826 0.6164284
3       60+   3307   7.971216 0.7797954
```

Estimate prevalence of depression by Age and Sex

```
## By sex and age
ourSummary(~ Depression, ~ Sex + Age_group, NH_des_dat2)
```

```
  Sex Age_group counts Depression        se
1   M     20-39   1654   5.513778 0.6461045
2   F     20-39   1674  10.050321 0.8036891
3   M     40-59   1556   5.222060 0.7699895
4   F     40-59   1751  11.477238 1.2011361
5   M       60+   1611   6.052782 0.8295114
6   F       60+   1696   9.579923 1.0534115
```

Compare Prevalence between Male and Female

Across all age groups:

```
svyttest(Depression ~ Sex, NH_des_dat2)
```

```

    Design-based t-test

data:  Depression ~ Sex
t = 6.8246, df = 29, p-value = 1.706e-07
alternative hypothesis: true difference in mean is not equal t
95 percent confidence interval:
 3.416354 6.340267
sample estimates:
difference in mean
           4.87831
```

Compare Prevalence between Male and Female

In people ages 40-59:

```
svyttest(Depression ~ Sex, subset(NH_des_dat2, Age_group == "4
```

```
	Design-based t-test

data:  Depression ~ Sex
t = 3.8688, df = 29, p-value = 0.0005706
alternative hypothesis: true difference in mean is not equal t
95 percent confidence interval:
 2.948407 9.561949
sample estimates:
difference in mean
          6.255178
```

Differences by Age Group, among Adults

```
svyttest(Depression ~ Age_group, subset(NH_des_dat2,
                  Age_group=="20-39" | Age_group=="40-59"))
```

```
    Design-based t-test

data:  Depression ~ Age_group
t = 0.79398, df = 29, p-value = 0.4337
alternative hypothesis: true difference in mean is not equal t
95 percent confidence interval:
 -1.079836  2.450262
sample estimates:
difference in mean
         0.6852129
```

Next Time?

Linear Regression and ANOVA / ANCOVA models

Reminders

1. Please complete the **Minute Paper after Class 3** by noon tomorrow (Wednesday 2023-01-25)
2. Get started on **Lab 2**, due next Monday 2023-01-30 at 9 PM
3. Continue reading **How To Be A Modern Scientist**