`survey` package: stratified designs

Week 3

Stat 260, St. Clair

Design object: Stratified sampling

```
> library(survey)
> my_design <- svydesign(id, fpc, weights, strata, mydata)
```

- `id` defines the sampling units
- `fpc` gives N_h or n_n/N_n for fpc correction → pop. stratum sizes
- `weights` sampling weights $\underline{N_h/n_h}$
- `strata` gives the stratification variable(s)
    - if more than one variable defines strata use `strata = ~var1 + var2`

Lohr Examples 3.2 and 3.6

The file ~~agsrs.cs~~v agstrat contains farm data collected from a SRS of $n = 300$ counties from $N = 3078$ in the US.

```
      > library(SDaA)
→     > str(agstrat)    # looks at the ``structure'' of the data frame's va
      'data.frame':    300 obs. of  17 variables:
       $ county  : Factor w/ 261 levels "ALEXANDER COUNTY",..: 180 115 254
       $ state   : Factor w/ 46 levels "AL","AR","AZ",..: 27 13 32 20 44 2
       $ acres92 : int  297326 124694 246938 206781 78772 210897 507101 33
       $ acres87 : int  332862 131481 263457 190251 85201 229537 552844 33
       $ acres82 : int  319619 139111 268434 197055 89331 213105 541015 35
       $ farms92 : int  725 658 1582 1164 448 583 321 986 1249 488 ...
       $ farms87 : int  857 671 1734 1278 483 699 371 1065 1251 518 ...
       $ farms82 : int  865 751 1866 1464 527 693 341 1208 1320 571 ...
       $ largef92: int  54 14 20 23 6 34 163 56 86 216 ...
       $ largef87: int  54 13 19 17 5 32 180 36 78 204 ...
       $ largef82: int  42 14 16 9 5 23 176 42 69 193 ...
       $ smallf92: int  58 42 175 56 56 8 10 90 42 16 ...
       $ smallf87: int  67 36 186 66 49 19 24 115 38 37 ...
strata $ smallf82: int  48 38 184 55 48 13 16 132 28 24 ...
     → $ region  : Factor w/ 4 levels "NC","NE","S",..: 1 1 1 1 1 1 1 1 1
       $ rn      : int  805 241 913 478 1028 496 969 42 676 383 ...
Nh/nh→ $ weight  : num  10.2 10.2 10.2 10.2 10.2 ...
```

Design

N_h

We need to add **stratum** population sizes to the data frame:

```
> # recode maps pop sizes to the right regions
> library(dplyr)
> agstrat$N <- recode(agstrat$region,
+                     NC = 1054,
+                     NE = 220,
+                     S = 1382,
+                     W = 422)
> # check if recoding worked:
> agstrat %>%
+   group_by(region) %>%
+   summarize(min(N), max(N))
# A tibble: 4 x 3
  region `min(N)` `max(N)`
  <fct>     <dbl>    <dbl>
1 NC         1054     1054
2 NE          220      220
3 S          1382     1382
4 W           422      422
```

to pop. size

old value = new value

→ N_h

Design

We need to add **stratum** sampling weights N_h/n_h to the data frame:

```
> # sample sizes:
> table(agstrat$region)

 NC   NE    S    W
103   21  135   41
> # recode maps sample sizes to the right regions
> agstrat <- agstrat %>%
+   group_by(region) %>%
+   mutate(n = n())  %>% ungroup()
> agstrat$n[1:180]
  [1] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103
 [19] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103
 [37] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103
 [55] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103
 [73] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103
 [91] 103 103 103 103 103 103 103 103 103 103 103 103 103  21  21  21
[109]  21  21  21  21  21  21  21  21  21  21  21  21  21  21  21  21
[127] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
[145] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
[163] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
```

→ n_h

n_h

↓

$n()$ ⇒ count # rows in each region

NC

NE

S

Design

We need to add **stratum** sampling weights N_h/n_h to the data frame:

```
> # add weights
> agstrat$wts <- agstrat$N/agstrat$n
> # check work:
> agstrat %>%
+   group_by(region) %>%
+   summarize(min(wts), max(wts))
# A tibble: 4 x 3
  region `min(wts)` `max(wts)`
  <fct>       <dbl>      <dbl>
1 NC           10.2       10.2
2 NE           10.5       10.5
3 S            10.2       10.2
4 W            10.3       10.3
```

$\longrightarrow N_{NC}/n_{NC}$

Design

```
> library(survey)
> design_strat <- svydesign(id= ~1,
+                           fpc= ~N,          Nh
+                           weights= ~wts,       Nh/nh
+                         * strata = ~region,
+                           data= agstrat)
> summary(design_strat)
Stratified Independent Sampling design
svydesign(id = ~1, fpc = ~N, weights = ~wts, strata = ~region,
    data = agstrat)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09545 0.09768 0.09768 0.09747 0.09772 0.09772
Stratum Sizes:
            NC NE   S  W
obs        103 21 135 41
design.PSU 103 21 135 41
actual.PSU 103 21 135 41
Population stratum sizes (PSUs):
  NC   NE    S    W
1054  220 1382  422
Data variables:
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92'
```

Lohr Examples 3.2 and 3.6

The variable `acres92` records the number of farming acres in a county in 1992.

```
> # SRS estimate/SE of total farm acres
> svytotal(~acres92, design_strat)
            total       SE
acres92 909736035 50417248
```

→ strat!

```
> mean_obj <- svymean(~acres92, design_strat)
> mean_obj
          mean    SE
acres92 295561 16380
> confint(mean_obj, df = degf(design_strat))
         2.5 %   97.5 %
acres92 263325 327796.5
```

Lohr Examples 3.2 and 3.6

What proportion of counties in the US have fewer than 200,000 farming acres?

```
> agstrat$lt200k92<- ifelse(agsrs$acres92 < 200000,
+                          "less than 200k", "greater than 200k")
> design_strat <- update(design_strat, lt200k92 = agstrat$lt200k92)
> svymean(~lt200k92, design_strat)
                             mean     SE
lt200k92greater than 200k 0.48973 0.0273
lt200k92less than 200k    0.51027 0.0273
```

Lohr Examples 3.2 and 3.6: estimating within strata

svyby

```
> # estimated region means
> region_mean <- svyby(~acres92, # variable
+                      ~region,  # strata
+                      design_strat, # design *
+                      svymean)  # gets mean estimates
> region_mean    # SRS estimates for each region
   region   acres92       se
NC     NC 300504.16 16107.59
NE     NE  97629.81 18149.49
S       S 211315.04 18925.35
W       W 662295.51 93403.65
> confint(region_mean,df=degf(design_strat))
       2.5 %   97.5 %
NC 268804.25 332204.1
NE  61911.41 133348.2
S  174069.74 248560.3
W  478476.12 846114.9
```

} SRS est/SE

→ FPC $\frac{n_h}{N_h}$

Lohr Examples 3.2 and 3.6: stratified vs SRS

The **design effect** of a design compares it's SE^2 to what you'd expect from a SRS:

for a specific estimate
variable

$$DEff = \frac{V(\hat{t}_{str})}{V(\hat{t}_{SRS})} \approx 0.7945$$

```
> # Design effect estimated from the stratified sample:
> svytotal(~acres92, design_strat, deff=T)
              total          SE   DEff
acres92 909736035  50417248 0.7945
```

The variance for estimating total is about 20% lower under a stratified design compared to an *equal sized* SRS.

$$DEff = \frac{SE^2 \text{ complex}}{SE^2 \text{ SRS}}$$

⇒ DEff will change depending on the estimate / variable used