Intro to the `survey` package

Week 2

Stat 260, St. Clair

`survey` package

- use to analyze survey data by defining the sampling design
- **use for all estimation problems that involve a data set!**
    - don't do a "by hand" calculation when you can use `survey`

Basic elements

1. Define a **design object**
2. Estimate mean/total/proportion using the design object
3. (more) can also graph and model using a design's sampling weights

Design object

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

- `id` defines the sampling units
- `fpc` gives N or n/N for fpc correction
- `weights` sampling weights, not needed if self-weighting (all weights equal)

Estimation

Estimates/SE:

```
> svytotal(~vars, my_design)  # pop total estimate
> svymean(~vars, my_design)   # pop mean estimate
```

Confidence interval:

```
> mean_obj <- svymean(~vars, my_design)   # pop mean estimate
> confint(mean_obj)  # CI using N(0,1)
> confint(mean_obj, df = degf(my_design))  # CI using t
```

Lohr Examples 2.5 and 2.10

The file `agsrs.csv` contains farm data collected from a SRS of $n = 300$ counties from $N = 3078$ in the US.

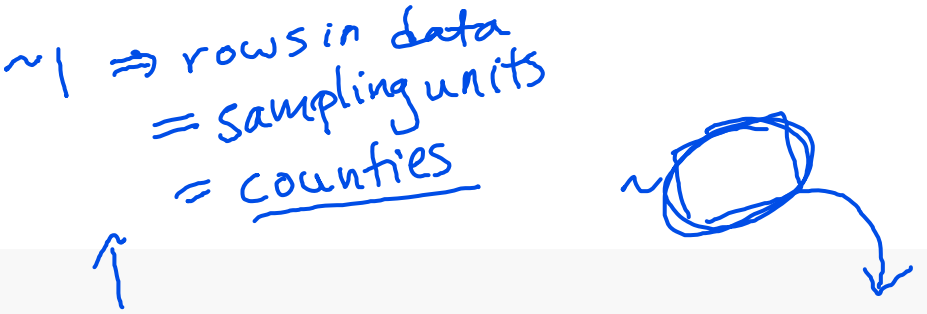
```
> library(SDaA)   # data package for the textbook
> str(agsrs)
'data.frame':    300 obs. of  14 variables:
 $ county  : Factor w/ 256 levels "ADAMS COUNTY",..: 45 46 127 139 14
 $ state   : Factor w/ 42 levels "AL","AR","CA",..: 1 1 1 1 1 1 2 2 2
 $ acres92 : int  175209 138135 56102 199117 89228 96194 57253 210692
 $ acres87 : int  179311 145104 59861 220526 105586 120542 66305 2235
 $ acres82 : int  194509 161360 72334 231207 113618 134616 80909 2275
 $ farms92 : int  760 488 299 434 566 436 320 1051 419 278 ...
 $ farms87 : int  842 563 362 471 658 521 411 1103 526 306 ...
 $ farms82 : int  944 686 447 622 748 650 477 1169 512 369 ...
 $ largef92: int  29 37 4 48 7 20 6 23 7 87 ...
 $ largef87: int  28 41 4 66 9 17 4 32 5 86 ...
 $ largef82: int  21 42 3 62 9 23 9 27 5 72 ...
 $ smallf92: int  57 12 16 14 11 18 17 42 15 8 ...
 $ smallf87: int  47 44 20 11 23 32 12 41 35 6 ...
 $ smallf82: int  66 47 30 28 27 29 27 49 18 4 ...
```

Design

We need to add population size and sampling weights to the data frame:

```
> agsrs$N <- 3078  # population size
> nrow(agsrs)   # sample size
[1] 300
> agsrs$wts <- agsrs$N/nrow(agsrs)  # sampling weights for SRS (N/n)
> head(agsrs$wts)
[1] 10.26 10.26 10.26 10.26 10.26 10.26
```

Design




```
> library(survey)
> design_srs <- svydesign(id= ~1, fpc= ~N, weights= ~wts, data= agsrs
> summary(design_srs)
Independent Sampling design
svydesign(id = ~1, fpc = ~N, weights = ~wts, data = agsrs)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09747 0.09747 0.09747 0.09747 0.09747 0.09747
Population size (PSUs): 3078
Data variables:
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
 [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92
[13] "smallf87" "smallf82" "N"        "wts"
```

Lohr Examples 2.5 and 2.10

~ var1 + var2 + var3

The variable `acres92` and `acres87` record the number of farming acres in a county in 1992 and 1987, respectively.

```
> # SRS estimate/SE of total farm acres
> svytotal(~acres92 + acres87, design_srs)
             total       SE
acres92 916927110 58169381
acres87 929413560 58216264
```

mean per county

```
> mean_obj <- svymean(~acres92, design_srs)
> confint(mean_obj, df = degf(design_srs))
            2.5 %   97.5 %
acres92 260706.3 335087.8
```

299

Lohr Examples 2.5 and 2.10

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

```
> agsrs$lt200k92<- ifelse(agsrs$acres92 < 200000,
+                         "less than 200k", "greater than 200k")
> table(agsrs$lt200k92)

greater than 200k    less than 200k
              147               153
```

```
> design_srs <- update(design_srs, lt200k92 = agsrs$lt200k92)
> svymean(~lt200k92, design_srs)
                              mean     SE
lt200k92greater than 200k 0.49 0.0275
lt200k92less than 200k    0.51 0.0275
```

add to the existing design