Survey Methods II: Statistical Analysis of Surveys in R

CSDE Workshop

Jessica Godwin

January 27, 2026

Resources and Materials

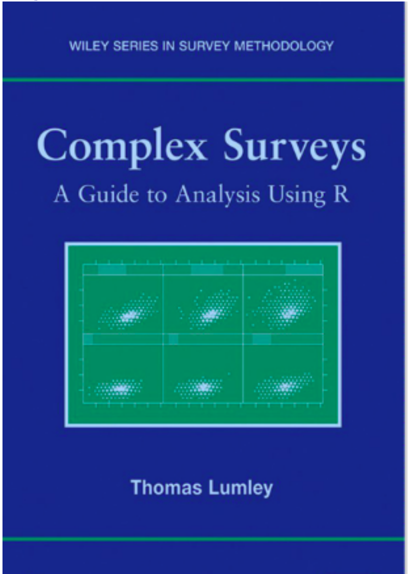
Install survey

- Survey Package Documentation

```
install.packages("survey", dependencies = TRUE)
```

Complex Surveys in R

- Lumley, Thomas (2004). Analysis of Complex Survey Samples. Journal of Statistical Software, 9(8), 1–19. https://doi.org/10.18637/jss.v009.i08
- Lumley, Thomas (2011). Complex surveys: a guide to analysis using R (Vol. 565). John Wiley & Sons.



Survey Statistics

- Lohr, Sharon L (1999). Sampling: Design and analysis. Pacific Grove, CA: Duxbury Press. https://doi.org/10.1201/9780429298899
- Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. Springer Science & Business Media. https://link.springer.com/book/9780387406206
- CSSS/STAT 529 (Spring, taught by Elena Erosheva or Jon Wakefield)

Why survey statistics?

Why survey statistics?

- If our outcome is binary:

Why survey statistics?

- If our outcome is binary:
  - Did our data arise from flipping a coin? or
  - Did our data arise from drawing from an urn?

Why survey statistics?

- If our outcome is binary:
  - Did our data arise from flipping a coin? or
  - Did our data arise from drawing from an urn?
- What does flipping a coin or drawing from an urn have to do with surveys with human respondents?

Finite vs. superpopulations

For observations $i = 1, \ldots, n$, let

$$y_i = \begin{cases} 1, & \text{success}, \\ 0, & \text{failure}. \end{cases}$$

- Superpopulation: If $y \sim \text{Bernoulli}(p)$,
  - $E[y] = p \;\; Var(y) = p(1-p)$.
- Finite population: If $y \sim \text{Hypergeometric}(N, K, n)$,
  - $E[y] = \frac{K}{N} \;\; Var(y) = \frac{K}{N}\left(1 - \frac{K}{N}\right)\left(1 - \frac{n}{N}\right)$. How do we say what $\widehat{p}$ means in either case? Is it the same?

Finite vs. superpopulations, cont'd

If $y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$,

$$\widehat{p} = \frac{\sum_{i=1}^{n} y_i}{n},$$
$$\widehat{Var}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n}.$$

If $y_i \overset{\text{iid}}{\sim} \text{Hypergeometric}(N, K, n)$,

$$\widehat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{k}{n}$$
$$\widehat{Var}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n} \times (1 - \frac{n}{N}).$$

How do we say what $\widehat{p}$ means in either case? Is it the same?

Survey Designs

Simple random sampling(SRS)

- Under an **SRS** of n observations

$$\Pr(\text{subject } k \in \text{sample}, S) = \pi_k = \frac{1}{N}$$

$$\Pr(\text{subjects } k, k' \in \text{sample}, S) = \pi_{k,k'} = \frac{1}{N} \times \frac{1}{N}.$$

- Under an **SRSWOR** of n observation

$$\Pr(\text{subject } k \in \text{sample}, S) = \pi_k = \frac{n}{N}$$

$$\Pr(\text{subjects } k, k' \in \text{sample}, S) = \pi_{k,k'} = \frac{n}{N} \times \frac{n-1}{N-1}.$$

Specifying the Design: `svydesign`

- `svydesign()` needs to know what columns in your data (if any) represent
  - sampling weights (`weights`)
  - strata (`strata`)
  - clusters (`ids`)
  - units
  - finite population correct (`fpc`)

```
library(survey)
data(api)
?api
?svydesign
```

Specifying the Design: `svydesign`

- `apisrs$pw`: sampling weight
- `apisrs$fpc`: finite population correction, i.e. *N* for an SRS
- `apisrs$stype`: school type (elementary, middle, high)

```
table(apisrs$pw,
      apisrs$fpc)
```

```
##
##          6194
##   30.97   200
```

```
6194/200
```

```
## [1] 30.97
```

Specifying the Design: `svydesign`

- `apisrs$pw`: sampling weight
- `apisrs$fpc`: finite population correction, i.e. *N* for an SRS
- `apisrs$stype`: school type (elementary, middle, high)

```
table(apisrs$pw,
      apisrs$stype)
```

```
##
##           E   H   M
##   30.97 142  25  33
```

Specifying the Design: svydesign

```r
srs_des <- svydesign(ids = ~1, weights = ~pw,
                     fpc = ~fpc, data = apisrs)
srs_des
```

```
## Independent Sampling design
## svydesign(ids = ~1, weights = ~pw, fpc = ~fpc, data = apisrs)
```

Systematic sampling

- Select every r^{th} sampling unit from the sampling frame of length N: $r \times n \leq N < r \times (n+1)$
  - What is π_k for individual $k = r$? $k = r + 1$?

Systematic sampling

- Select every r^{th} sampling unit from the sampling frame of length N: $r \times n \leq N < r \times (n+1)$
  - What is π_k for individual $k = r$? $k = r + 1$?
  - Can a systematic sample be implemented so that it is the equivalent of an SRS?

Systematic sampling

- Select every r^{th} sampling unit from the sampling frame of length N: $r \times n \leq N < r \times (n+1)$
  - What is π_k for individual $k = r$? $k = r + 1$?
  - Can a systematic sample be implemented so that it is the equivalent of an SRS?
  - What is $\pi_{r,r+1}$?

Systematic sampling

- Select every r^{th} sampling unit from the sampling frame of length N: $r \times n \leq N < r \times (n+1)$
  - What is π_k for individual $k = r$? $k = r + 1$?
  - Can a systematic sample be implemented so that it is the equivalent of an SRS?
  - What is $\pi_{r,r+1}$?
- Random single start $\rightarrow$ what changes?

Systematic sampling

- Select every r^{th} sampling unit from the sampling frame of length N: $r \times n \leq N < r \times (n+1)$
  - What is π_k for individual $k = r$? $k = r + 1$?
  - Can a systematic sample be implemented so that it is the equivalent of an SRS?
  - What is $\pi_{r,r+1}$?
- Random single start $\rightarrow$ what changes?
- Multiple starts
  - No individual sampling probabilities are 0 or 1
  - Joint sampling probabilities defined

Stratified simple random sampling (strSRS)

- Consider $h = 1, \dots, H$ strata from each of which you want to sample n_h individuals.

$$
\begin{aligned}
\Pr(\text{subject } k \in S_h) =& \pi_k = \frac{n_h}{N_h} \\
\Pr(\text{subjects } k, k' \in S_h) =& \pi_{k,k'} = \frac{n_h}{N_h} \times \frac{n_h - 1}{N_h - 1} \\
\Pr(\text{subjects } k \in S_h, k' \in S_{h'}) =& \pi_{k,k'} = \frac{n_h}{N_h} \times \frac{n_{h'}}{N_{h'}}.
\end{aligned}
$$

strSRS, cont'd

- Why stratify? Why not an SRS or SRSWOR?

strSRS, cont'd

- Why stratify? Why not an SRS or SRSWOR?
  - Availability of **sampling frame**
  - Cost, convenience, speed
  - $N_1, \ldots, N_h$ vary widely
  - Rare outcomes within certain strata
  - We know strata are related to outcome of interest $\rightarrow$ precision gains!
- What happens if we ignore the stratification?
  - Waste a lot of folks' money!!
  - Implicit assumption that outcome of interest doesn't differ by strata
  - $\rightarrow$ obscure differences in outcomes by strata
  - $\rightarrow$ OVERESTIMATE variance/standard errors
  - $\rightarrow$ worsens variability in outcomes between strata grows and within strata shrinks
  - $\rightarrow$ worsens as variability in $\pi_{k \in S_h}$ between strata grows

Specifying the Design: svydesign

- apisrs$pw: sampling weight
- apisrs$fpc: finite population correction, i.e. N_h for an strSRS
- apisrs$stype: strata; chool type (elementary, middle, high)

```
table(apistrat$pw, apistrat$fpc)
```

```
##
##                       755 1018 4421
##   15.1000003814697     50    0    0
##   20.3600006103516      0   50    0
##   44.2099990844727      0    0  100
```

```
755/50
```

```
## [1] 15.1
```

Specifying the Design: svydesign

- apisrs$pw: sampling weight
- apisrs$fpc: finite population correction, i.e. N_h for an strSRS
- apisrs$stype: strata; chool type (elementary, middle, high)

```
table(apistrat$pw,
      apistrat$stype)
```

```
##
##                       E   H   M
##   15.1000003814697    0  50   0
##   20.3600006103516    0   0  50
##   44.2099990844727 100   0   0
```

Specifying the Design: svydesign

```r
strsrs_des <- svydesign(ids = ~1,
                        strata = ~stype,
                        weights = ~pw,
                        fpc = ~fpc,
                        data = apistrat)
strsrs_des
```

```
## Stratified Independent Sampling design
## svydesign(ids = ~1, strata = ~stype, weights = ~pw, fpc = ~fpc,
##     data = apistrat)
```

Cluster sampling

Consider sampling $c = 1, \ldots, C$ clusters or **primary sampling units (PSU)** from your population of N_C clusters and N **units**.

Indivuals k are the **observation units** contained within clusters on which we will make measurements.

One-stage cluster sampling

$$\Pr(\text{PSU } c \in S) = \frac{C}{N_c}$$

$$\pi_{k \in Sc} = \begin{cases} 1, & \text{PSU } c \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Two-stage cluster sampling

Sample m_c from M_c units in cluster c.

$$\Pr(\text{PSU } c \in S) = \frac{C}{N_c}$$

$$\pi_{k \in S_c} = \begin{cases} \frac{m_c}{M_c}, & \text{PSU } c \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Cluster sampling, cont'd

- Probability proportional to size (PPS) sampling
  - $\pi_c \propto M_c$
  - When does this make sense?
- Why implement a cluster sample?
  - The only sampling frame we have is a list of groups of observation units
  - Cost and convenience
- What happens if we ignore clustering in our sample?
  - The m_c observation units sampled in cluster c are **not** independent samples
  - $\rightarrow$ we have LESS information than m_c observations from an SRS
  - $\rightarrow$ we will UNDERESTIMATE variances and standard errors if we ignore this dependence
  - $\rightarrow$ this underestimation worsens as the correlation between outcomes from individuals in a cluster increases

Specifying the Design: svydesign

```
clus_des <- svydesign(ids = ~dnum,
                      weights = ~pw,
                      fpc = ~fpc,
                      data = apiclus1)
clus_des
```

```
## 1 - level Cluster Sampling design
## With (15) clusters.
## svydesign(ids = ~dnum, weights = ~pw, fpc = ~fpc, data = apiclus1)
```

Specifying the Design: svydesign

```
twoclus_des <- svydesign(ids = ~dnum + snum,
                         weights = ~pw,
                         fpc = ~fpc1 + fpc2,
                         data = apiclus2)
twoclus_des
```

```
## 2 - level Cluster Sampling design
## With (40, 126) clusters.
## svydesign(ids = ~dnum + snum, weights = ~pw, fpc = ~fpc1 + fpc2,
##     data = apiclus2)
```

Complex surveys

Multi-stage sampling

- **Example:** DHS (among others) stratify clusters by administrative divisions× urban/rural → select women withihn households within clusters within strata
- **Stratified two-stage cluster sampling**
- PSUs → **secondary sampling units** (SSUs) → observation units
- One could stratify within clusters if a sampling frame necessitates (never encountered this yet)

Multi-phase sampling

- Fancy term for trying again to reach non-respondents!!
- Sub-sample (perhaps fully) your nonrespondents in attempts to get a response.

Designs of Common Surveys

- Demographic and Health Surveys (DHS)
  - https://dhsprogram.com/
  - Kenya DHS 2014 Final Report https://dhsprogram.com/pubs/pdf/FR308/FR308.pdf
- Youth Risk Behavior Survey (YRBS)
  - https://www.cdc.gov/healthyyouth/data/yrbs/index.htm
  - 2019 National YRBS Data User's Guide https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2019/2019_National_YRBS_Data_Users_Guide.pdf
- American Communiy Survey (ACS)
  - https://www.census.gov/programs-surveys/acs
  - Design & Methodology https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_ch04_2014.pdf

Estimation with Survey Data

Horvitz-Thompson estimators

- Each individual k has their responses weighted by their **sampling weight** $w_k = \dfrac{1}{\pi_k}$
  - i.e. an individual with low chance of being sampled $\rightarrow \pi_k$ small $\rightarrow w_k$ big
  - w_k can be interpreted as number of individuals in the finite population that individual k's response represents
  - **Caveat:** nonresponse
- **Average** or **arithmetic mean**

$$\frac{\sum_{k=1}^n y_k}{n} \stackrel{?}{=} \frac{\sum_{k=1}^n w_k y_k}{\sum_{k=1}^n w_k} = \frac{\sum_{k=1}^n \frac{N}{n} y_k}{\sum_{k=1}^n \frac{N}{n}}$$

$$= \frac{\frac{N}{n}\sum_{k=1}^n y_k}{\frac{N}{n}\sum_{k=1}^n 1} = \frac{N}{n}\left(\frac{\sum_{k=1}^n y_k}{\frac{N}{n} \times n}\right)$$

$$= \frac{N}{n}\left(\frac{\sum_{k=1}^n y_k}{N}\right) = \frac{\sum_{k=1}^n y_k}{n}$$

Horvitz-Thompson estimators

- Each individual k has their responses weighted by their **sampling weight** $w_k = \dfrac{1}{\pi_k}$
    - i.e. an individual with low chance of being sampled $\rightarrow \pi_k$ small $\rightarrow w_k$ big
    - w_k can be interpreted as number of individuals in the finite population that individual k's response represents
    - **Caveat:** nonresponse
- **Weighted average**

$$\sum_{k=1}^{n} w_k y_k \text{ such that } w_k \in [0,1] \text{ and } \sum_{k}^{n} w_k = 1$$

Totals

- Consider a population of size N, a sample of size n, where each individual has outcome Y_k
- Y_k is **not** random, but Z_k is

$$Z_k = \begin{cases} 1, & k \in S \\ 0, & \text{otherwise.} \end{cases}$$

- Once sample taken $y_k = Y_k \times Z_k$ denotes an individual's observed response (may contain measurement error)
  - $E[y_k] = E[Y_k \times Z_k] = Y_k E[Z_k] = Y_k \times \pi_k$

Totals

- The population total of outcomes Y is

$$T = \sum_{k=1}^{N} Y_k$$

$$\widehat{T} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

$$\widehat{Var}(\widehat{T}) = \sum_{k,k'} \frac{y_k y_{k'}}{\pi_k \pi_{k'}} - \frac{y_k y_{k'}}{\pi_{kk'}}$$

Totals: Stratified sampling

$$\widehat{T} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} \sum_{k=1}^{n_h} w_{hk} y_{hk},$$

$$\widehat{Var}(\widehat{T}) = \sum_{h=1}^{H} \widehat{Var}(\widehat{T}_h) = \sum_{h=1}^{H} \sum_{k,k'} \frac{y_{hk} y_{hk'}}{\pi_{hk} \pi_{hk'}} - \frac{y_{hk} y_{hk'}}{\pi_{hkk'}},$$

- Calculate variance in terms of each individual's difference from their respective strata total.

Totals: Cluster sampling

$$\widehat{T} = \sum_{c=1}^{C} T_c = \sum_{c=1}^{C} \sum_{k=1}^{N_c} w_{ck} y_{ck} = \sum_{c=1}^{C} w_c \sum_{k=1}^{N_c} y_{ck},$$

- Calculate the variance in terms of each cluster total's difference from the overall population total

Totals: Stratified two-stage cluster sampling

$$\widehat{T} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} \sum_{c_1=1}^{C_{1h}} \widehat{T}_{h[c_1]}$$

$$= \sum_{h=1}^{H} \sum_{c_1=1}^{C_{1h}} \sum_{c_2=1}^{C_{2h}} \widehat{T}_{h[c_1:c_2]} = \sum_{h=1}^{H} \sum_{c_1=1}^{C_{1h}} \sum_{c_2=1}^{C_{2h}} \sum_{k=1}^{n_{c_2}} w_{h[c_1:c_2]k} y_{h[c_1:c_2]k}$$

$$\widehat{Var}(\widehat{T}) = \sum_{h=1}^{H} \widehat{Var}(\widehat{T}_h).$$

- Apply methods from previous two in appropriate summation order

Totals: svytotal

```
?svytotal
svytotal(~enroll, design = srs_des)
```

```
##           total     SE
## enroll 3621074 169520
```

```
svytotal(~enroll, design = strsrs_des)
```

```
##           total     SE
## enroll 3687178 114642
```

```
svytotal(~enroll, design = twoclus_des, na.rm = TRUE)
```

```
##           total     SE
## enroll 2639273 799638
```

Means

- The population mean of outcomes Y is

$$\overline{Y} = \frac{\sum_{k=1}^{N} Y_k}{N}$$

$$\widehat{\overline{Y}} = \frac{\sum_{k=1}^{n} w_k y_k}{N} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

$$\widehat{Var}(\widehat{\overline{Y}}) = \frac{\widehat{Var}(\widehat{T})}{N^2}$$

$$\overset{=}{SRS} \left(1 - \frac{n}{N}\right) \times \frac{1}{n} \sum_{k=1}^{n} (y_k - \bar{y})$$

Means: svymean

```
?svymean
svymean(~enroll, design = srs_des)
```

```
##           mean     SE
## enroll 584.61 27.368
```

```
svymean(~enroll, design = strsrs_des)
```

```
##           mean     SE
## enroll 595.28 18.509
```

```
svymean(~enroll, design = twoclus_des, na.rm = TRUE)
```

```
##           mean     SE
## enroll 526.26 80.341
```

Proportions

- The population mean of binary outcomes Y or **prevalence** is

$$P = \frac{\sum_{k=1}^{N} Y_k}{N}$$

$$\widehat{P} = \frac{\sum_{k=1}^{n} w_k y_k}{N} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

$$\widehat{Var}(\widehat{P}) = \frac{(\widehat{P}(1 - \widehat{P}))}{N}$$

Proportions: svyciprop

```
confint(svymean(~sch.wide, design = srs_des))
```

```
##                    2.5 %    97.5 %
## sch.wideNo  0.1319288 0.2380712
## sch.wideYes 0.7619288 0.8680712
```

```
svyciprop(~sch.wide, design = srs_des)
```

```
##                    2.5% 97.5%
## sch.wide 0.815 0.756 0.863
```

Proportions: svyciprop

```
confint(svymean(~sch.wide, design = strsrs_des))
```

```
##                    2.5 %    97.5 %
## sch.wideNo  0.1243371 0.2197669
## sch.wideYes 0.7802331 0.8756629
```

```
svyciprop(~sch.wide, design = strsrs_des)
```

```
##                    2.5% 97.5%
## sch.wide 0.828 0.775 0.871
```

Ratio Estimation

- What if we don't know N or don't have the finite population corrections?

$$\widehat{\bar{Y}} = \frac{\widehat{T}}{\widehat{N}} = \frac{\sum_k w_k y_k}{\sum_k w_k}$$

  - Now what is $\widehat{Var}(\widehat{\bar{Y}})$? `survey` uses Taylor linearization.

- What if we have some other variable X that we measured in our survey and know population totals for?

$$\widehat{T_Y} = \widehat{T_Y}\frac{T_X}{\widehat{T_X}} = \sum_k w_k y_k \times \frac{T_X}{\sum_k w_k x_k}$$

  - If we're over(under)estimating T_X, then maybe we're over(under)estimating T_Y

Ratio Estimation: svyratio

```
srs_des_nofpc <- svydesign(ids = ~1, weights = ~pw,
                           data = apisrs)

svymean(~enroll, design = srs_des)
```

```
##          mean     SE
## enroll 584.61 27.368
```

```
svymean(~enroll, design = srs_des_nofpc)
```

```
##          mean     SE
## enroll 584.61 27.821
```

```
strsrs_des_nofpc <- svydesign(ids = ~1,
                              strata = ~stype,
                              weights = ~pw,
                              data = apistrat)
svymean(~enroll, design = strsrs_des)
```

```
##           mean     SE
## enroll 595.28 18.509
```

```
svymean(~enroll, design = strsrs_des_nofpc)
```

```
##           mean     SE
## enroll 595.28 18.941
```

```r
srs_des_nofpc <- update(srs_des_nofpc, counter = 1)
svyratio(numerator = ~enroll, denominator = ~counter, design = srs_des_nof
```

```
## Ratio estimator: svyratio.survey.design2(numerator = ~enroll, denominat
##     design = srs_des_nofpc)
## Ratios=
##         counter
## enroll   584.61
## SEs=
##          counter
## enroll 27.82121
```

```r
svymean(~enroll, design = srs_des_nofpc)
```

```
##           mean     SE
## enroll 584.61 27.821
```

Small Area Estimation: svyby

```r
svyby(~enroll, by = ~stype, design = srs_des, svytotal)
```

```
##   stype    enroll        se
## E     E 1849900.0  99738.62
## H     H  890666.2 187717.67
## M     M  880508.1 151805.23
```

```r
svyby(~enroll, by = ~stype, design = strsrs_des, svytotal)
```

```
##   stype    enroll       se
## E     E 1842584.3 72581.33
## H     H  997128.5 69239.40
## M     M  847464.7 55502.96
```

General Linear Models: svyglm

```
srs_mod <- svyglm(sch.wide ~ ell + meals + mobility,
                  design = srs_des, family = quasibinomial())
strsrs_mod <- svyglm(sch.wide ~ ell + meals + mobility,
                  design = strsrs_des, family = quasibinomial())
```

General Linear Models: svyglm

```
coefs_table <- data.frame(SRS_Coef = coef(srs_mod),
                          SRS_SE = SE(srs_mod),
                          StrSRS_Coef = coef(strsrs_mod),
                          StrSRS_SE = SE(strsrs_mod))
round(coefs_table, digits = 3)
```

```
##             SRS_Coef SRS_SE StrSRS_Coef StrSRS_SE
## (Intercept)    1.744  0.456       0.836     0.456
## ell           -0.022  0.011      -0.002     0.013
## meals          0.011  0.009      -0.003     0.009
## mobility      -0.015  0.022       0.061     0.032
```

Post-stratification

What if we don't have a **probability survey**? OR What if our ideal stratification scheme was not possible to implement given our sampling frame?

```
pop.types <- apipop %>%
  group_by(stype) %>%
  summarize(Freq = n())

srs_post <- postStratify(srs_des, ~stype, pop.types)
```

Data Visualization

Functions in the `survey` package

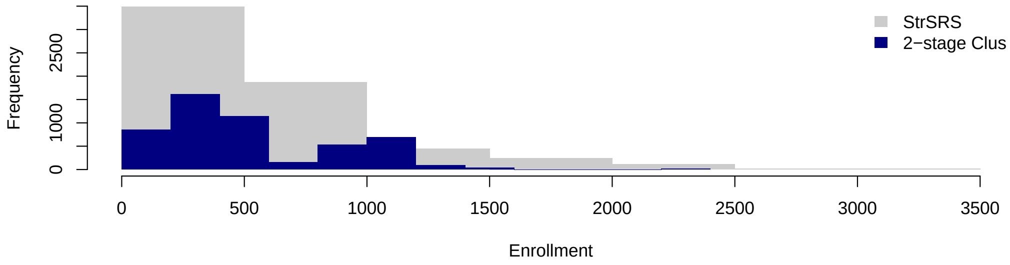
Instead of plotting the data in your sample, the following visualizations give you a sense of the data in your sample AND their relative contribution to the population.

- Histograms (`svyhist`) of weighted outcomes
- Boxplots (`svyboxplot`) of weighted outcomes (by group, if desired)
- Scatterplots (`svyplot`) of two variables showing relative weight of observations (e.g. with transparency or character size)
- Scatterplots by group (`svycoplot`) of two variables conditional on the value of other variables (e.g. binary measure of exposure) using the hexagonal binning method available in `svyplot`

Histograms: svyhist

```
svyhist(~enroll, design = strsrs_des,
        main = "", xlab = "Enrollment",
        probability = FALSE, col = 'grey80', border = FALSE)
svyhist(~enroll, design = twoclus_des,
        main = "", xlab = "Enrollment",
        probability = FALSE, col = 'navy', border = FALSE, add = TRUE)
legend('topright', bty = 'n',
       fill = c("grey80", "navy"), border = FALSE,
       legend = c("StrSRS", "2-stage Clus"))
```

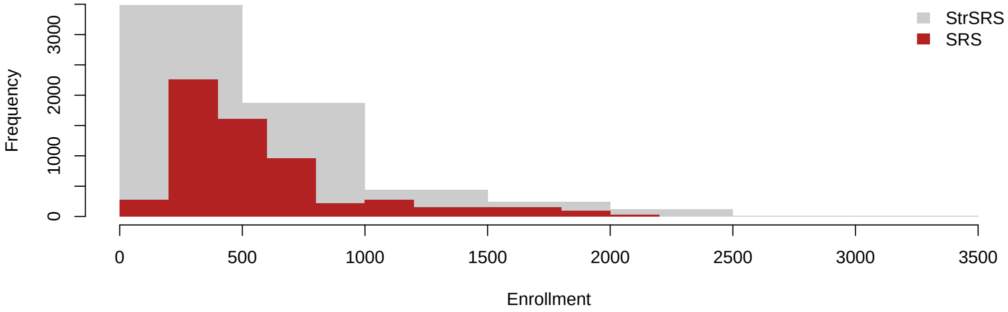
Histograms: `svyhist`



Histograms: svyhist

```
svyhist(~enroll, design = strsrs_des,
        main = "", xlab = "Enrollment",
        probability = FALSE, col = 'grey80', border = FALSE)
svyhist(~enroll, design = srs_des,
        main = "", xlab = "Enrollment",
        probability = FALSE, col = 'firebrick', border = FALSE,
        add = TRUE)
legend('topright', bty = 'n',
       fill = c("grey80", "firebrick"), border = FALSE,
       legend = c("StrSRS", "SRS"))
```

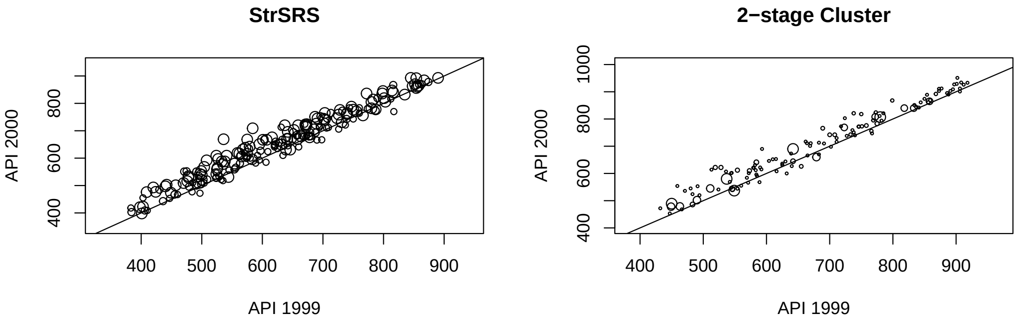
Histograms: `svyhist`



Scatterplots: svyplot

```
par(mfrow = c(1,2))
svyplot(api00~api99, design=strsrs_des, style="bubble",
        col = "firebrick",
        xlab = "API 1999", ylab = "API 2000", main = "StrSRS")
abline(0,1)
svyplot(api00~api99, design=twoclus_des, style="bubble",
        col = "firebrick",
        xlab = "API 1999", ylab = "API 2000", main = "2-stage Cluster")
abline(0,1)
```

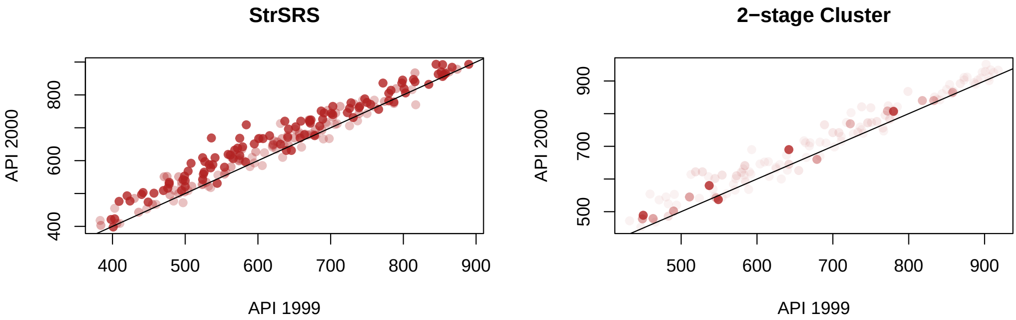
Scatterplots: `svyplot`



Scatterplots: svyplot

```
par(mfrow = c(1,2))
svyplot(api00~api99, design=strsrs_des, style="transparent",
        basecol = "firebrick", pch=19,
        xlab = "API 1999", ylab = "API 2000", main = "StrSRS")
abline(0,1)
svyplot(api00~api99, design=twoclus_des, style="transparent",
        basecol = "firebrick", pch=19,
        xlab = "API 1999", ylab = "API 2000", main = "2-stage Cluster")
abline(0,1)
```

Scatterplots: `svyplot`



Etc.

What didn't we cover?

- Post-stratification and raking (`survey::postStratify`)
- Replicate weights (`survey::svrepdesign`)
- Non-response
- Multi-phase sampling
- Model-based estimation

Specific Question: Contrasts

- Constrasts are just linear combinations of random variables where the weights add up to 0, e.g. averages and differences.

$$
\begin{aligned}
E[aX + bY] =& aE[X] + bE[Y] \\
E[X - Y] =& 1 \times E[X] + (-1) \times E[Y] \\
Var(aX + bY) =& a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y) \\
Var(X - Y) =& 1^2 \times Var(X) + (-1)^2 \times Var(Y) + 2(1)(-1) \times Cov(X, Y)
\end{aligned}
$$

Specific Question: Contrasts

```
cont_total <- svytotal(~api00+api99, strsrs_des)
svycontrast(cont_total, list(diff=c(1,-1)))
```

```
##       contrast    SE
## diff    203736 12705
```

```
vcov(cont_total)
```

```
##              api00      api99
## api00 3396439386 3521991247
## api99 3521991247 3808949720
```

```
sqrt(vcov(cont_total)[1,1] + vcov(cont_total)[2,2] +
      2*1*(-1)*vcov(cont_total)[1,2])
```

```
## [1] 12704.59
```