Analyzing Data from Complex Surveys

Thomas Lumley

11(12) July 2023

Introductions

Thomas Lumley

Developer of R survey package

Applied research in cardiovascular epi/genetics

Introduce yourselves (briefly). What do you want to get from the course

Why complex surveys?

The case-control design can be understood without considering the sampling in detail, but other modern designs cannot.

Weighted survey ideas are helpful in thinking about designs

Weighted estimation is a simple unifying approach; maximum likelihood may not be available and need not be much more efficient

When more efficient estimates exist, they are more sensitive to model assumptions (if non-response is not an issue).

Outline

- **Session 1**:
  - *Basic ideas of complex survey sampling*
  - *Telling the computer about your survey*
  - *Estimating population totals*
- **Session 2**:
  - *Estimating everything else*
- **Session 3**
  - *Data analysis with survey data*
- **Session 4**:
  - *Two-phase case-control and case-cohort designs*
  - *Raking of weights*

Important Notice

Please ask questions.

That's the benefit of having me here rather than just a pile of papers.

Survey sampling

Three basic concepts

- stratification
- oversampling
- clustering

Stratification

A sample of 6000 people from the US would on average contain

- 140 people from Washington (State)
- 11 people from Washington, DC

but the actual number from each region would vary.

Part of the variance of any statistic computed from the sample comes from the variation in the number of people from each region.

The mean of a sample with 7 people from DC will be different from a sample with 15

Stratification

If we can fix the number of people sampled in each region, we can eliminate between-region differences from the variance of our statistics, increasing precision. This is called a **stratified sample**, the regions are **strata**

Taking a stratified sample is possible only if we have a population list that includes the stratum for each person. (*sampling frame*)

The extra precision comes from using the extra information in this population list.

Stratification **always** decreases variance, perhaps not by very much.

Oversampling

Individuals need not be sampled with the same probabilities

- some may be more informative than others (cases vs controls)
- we might want to report statistics for subpopulations and need large enough sample size for them

The former decreases variance; the latter increases it for population summaries but decreases it for subpopulations

Clustering

If a survey involves a physical visit to each participant, it is less expensive to sample people who are physically close together

- homes in the same neighbourhood
- students in the same classroom
- workers in the same factory
- medical records in the same hospital

We often sample a small number of clusters and then sample people from each cluster.

Cluster sampling **increases** variance for the **same sample size**, but may reduce variance for the **same cost**

Clustering may lead to variation in sampling probabilities

Multistage sampling

Take a sample, then take a subsample from it

- Sample schools, then sample classrooms within schools
- Sample counties, then sample neighbourhoods within counties
- Sample universities, then sample academics stratified by department within each university

$$\pi_i = \Pr(\text{chosen at stage 1}) \times \Pr(\text{chosen at stage 2}|\text{in stage 1})$$

Notation

- N population size
- n sample size
- π_i probability that unit i would be sampled
- π_{ij} probability that both units i and j would be sampled
- w_i weights, usually (adjusted versions of) π_i^{-1}
- R_i sampling indicator: $E[R_i] = \pi_i$

Estimating population totals

Population total T_X of X is

$$T_X = \sum_{i=1}^{N} X_i$$

Horvitz-Thompson estimator is

$$\hat{T} = \sum_{i=1}^{N} \frac{R_i}{\pi_i} X_i$$

Since $E[R_i/\pi_i] = 1, \hat{T}$ is unbiased as long as $\pi_i > 0$ for all units in the population.

Estimating variances

$$\mathrm{var}\left[\sum_{i=1}^{N} \frac{R_i}{\pi_i} X_i\right] = \sum_{i,j=1}^{N} \frac{X_i X_j}{\pi_i \pi_j} \mathrm{cov}[R_i, R_j]$$

Estimate this using observed pairs $(i,\ j)$

$$\widehat{\mathrm{var}}\left[\sum_{i=1}^{N} \frac{R_i}{\pi_i} X_i\right] = \sum_{i,j=1}^{N} \frac{R_i R_j}{\pi_{ij}} \frac{X_i X_j}{\pi_i \pi_j} \mathrm{cov}[R_i, R_j]$$

Telling the computer

The Horvitz-Thompson formula needs π_{ij}. The computer can work these out from π_i and the strata and clusters.

If you designed the survey, you know all this information.

With public-use data you typically know only the sampling weights $(1/\pi_i)$ and the first-stage strata and clusters.

Describing (multistage) surveys to R

- Identifiers for sampling units (at each stage, optionally)
- Identifiers for strata (at each stage, optionally)
- Weights (or sampling probabilities at each stage, or population sizes at each stage)
- Population sizes at each stage (optionally)

`svydesign()` returns a survey design object containing data and design information.

Describing (multistage) surveys to Stata

- at the beginning: `svyset` describes the clusters, strata, sampling weights, population sizes (optionally)
- then use `svy:` prefix before any analysis command

What data to include

You need to describe the design for **the whole sample**

- even if you only want to analyse for avocado farmers, use `svydesign()` or `svyset` on the whole data set - (it's ok subset on strata, eg state for ACS)
- but there could be records in your data file that aren't part of the sample - eg, for NHANES if you want the clinical examination sample you need to drop records for people who aren't part of it (`WTMEC2yr` missing)

Example: California schools

Academic Performance Index: standardised test in schools

Population: 6194 schools in California, in 757 districts.

- a cluster sample of all schools in 15 districts
- a stratified unequal sample of 100 elementary schools, 50 middle schools, 50 high schools
- a two-stage cluster sample of 40 districts and up to 5 schools from each

Cluster sample

Using $w_i = M/m$

```
dclus1<-svydesign(id=~dnum, fpc=~fpc, data=apiclus1)
svytotal(~enroll, dclus1)
```

```
##           total      SE
## enroll 5076846 1389984
```

Cluster sample

Rescaling w_i to sum to known $N = 6194$

```
dclus1r<-svydesign(id=~dnum, weights=~pw, data=apiclus1,fpc=~fpc)
svytotal(~enroll, dclus1r)
```

```
##           total     SE
## enroll 3404940 932235
```

Estimate is improved: true $T_{\text{enroll}} = 3811472$

Stratified sample

```
dstrat<-svydesign(id=~snum, strata=~stype, fpc=~fpc, data=apistrat)
svytotal(~enroll, dstrat)
```

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

```
dstrat<-svydesign(id=~1, strata=~stype, fpc=~fpc, data=apistrat)
svytotal(~enroll, dstrat)
```

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

Two-stage cluster sample

Using $w = \pi_i^{-1} = \frac{N_i}{n_i}\frac{M}{m}$

```
dclus2<-svydesign(id=~dnum+snum, data=apiclus2,fpc=~fpc1+fpc2)
svytotal(~enroll, dclus2, na.rm=TRUE)
```

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

Rescaling w_i to sum to known $N = 6194$

```
dclus2<-svydesign(id=~dnum+snum, weights=~scaledw,
                  data=apiclus2,fpc=~fpc1+fpc2)
svytotal(~enroll, dclus2, na.rm=TRUE)
```

```
##           total     SE
## enroll 3187501 965738
```

More typical public-use data

```
data(nhanes)
design <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA,
              weights=~WTMEC2YR, nest=TRUE,data=nhanes)
design
```

```
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (31) clusters.
## svydesign(id = ~SDMVPSU, strata = ~SDMVSTRA, weights = ~WTMEC2YR,
##     nest = TRUE, data = nhanes)
```

Stata

```
use apiclus1
svyset dnum [pw=pw]
svyset dnum [pw=pw], fpc(fpc)
use apiclus2
svyset dnum [pw=pw], fpc(fpc1) || _n, fpc(fpc2)
svy: total enroll
use nhanes
svyset SDMVPSU [pw=WTMEC2YR], strata(SDMVSTRA)
```

Resampling

Analogs of the bootstrap and jackknife:

- *JK1* jackknife leaving out one cluster at a time
- *JKn* stratified version of JK1
- *BRR* split data into half, lots of times
- *bootstrap* resample clusters (several variants exist)

Usually implemented by including **replicate weights** in the data file (set a weight to 0 to omit a cluster)

Used by, eg, American Community Survey, California Health Interview Survey

Replicates

$$\widehat{var}\left[\hat{T}\right] = k\sum_{r=1}^{R} a_r (T_r^* - \hat{T})^2$$

(or use $\bar{T}^*$ instead of $\hat{T}$)

Jackknife has $k \times a_r \sim 1$, bootstrap and BRR have $k \times a_r \sim 1/R$.

Public-use data will tell you what multiplier to use: if they don't, use 1.0. They tend to just use the one multiplier, not different ones per replicate

- ACS: $k = 4/80, a_r = 1$
- CHIS: $k = 1, a_r = 1$

CHIS

```
chis<-read_dta("~/r-survey/www/svybook/Adult.dta")
```

There are sampling weights in `rakedw0` and replicate weights in `rakedw1`-`rakedw80` and the documentation describes them as jackknife weights with multiplier 1.

You could specify the columns (420-499) or specify the pattern of the weights,

```
chis_design<-svrepdesign(weights=~rakedw0, repweights=chis[,420:499],data=chis,
  scale=1, rscales=1,type="other")
```

or

```
chis_design <- svrepdesign(weights=~rakedw0,repweights="rakedw[1-9]+",data=chis,
  scale=1, rscales=1,type="JKn")
```

Either way

```
> chis_design
Call: svrepdesign.default(weights = ~rakedw0, repweights = chis[, 420:499],
    data = chis, scale = 1, rscales = 1, type = "other")
with 80 replicates.
> svymean(~bmi_p, chis_design, na.rm=TRUE)
        mean     SE
bmi_p 26.536 0.0418
> dim(chis_design)
[1] 43020   500
```

Further reading

Antony Damico's site `asdfree.com` has instructions on downloading and using a wide range of survey datasets in R