



EcoHealth  
Alliance

# Complex Survey Sampling and Analysis

**Ernest Guevarra**

**19 August 2021**

# Outline

- What are complex surveys?
- Concepts and principles of complex surveys
  - Survey design
  - Survey analysis
- Using R for complex survey design and analysis
  - `{survey}` package
  - `{smartr}` package
  - `{bbw}` package
- Useful resources

# What are complex surveys?

- A population is specified; data values unknown but are fixed (not random);
- A random sample is drawn from the population. This is what we call the **sample design**;
- Given **sample design**, probabilities can be known/calculated;
- Analysis of complex survey samples aims to estimate features of the fixed population hence is a **design-based inference**.

# Complex survey design

- **Probability samples or random samples** - procedure for taking samples from a population and not just the data we happen to end up with.
- **Probability samples or random samples** is a fundamental concept in design-based inference;
- Involves a sampling procedure that draws any subset of X number from the population with an equal likelihood of being selected.

# Complex survey design: simple random sample

- Drawing a sample of X number from a full list of members of a fixed population.

Lottery



Random number table

Number	Household
1	Aweno
2	Serrano
3	Coopsman
4	Franco
5	Oka
6	Parvanta
7	Roquefort
8	Moki
9	Stirling

Random number table				
7648	2352	6959	1937	
2554	6804	9098	4316	
4318	2346	7276	1880	
7136	9603	0163	3152	
7000	2865	8357	4475	
9804	0042	1106	7949	
2932	9958	9582	2235	
1140	1164	7841	1688	

# Complex survey design: systematic sample

- Determine a **sampling interval**

$$\text{sampling interval} = \left\lfloor \frac{\text{number of samples needed}}{\text{total number of population}} \right\rfloor$$

- Select a random number from the series of numbers between 1 and the sampling interval. This will be the **random starting point** for the systematic selection.
- Using the **random starting point** and the **sampling interval**, select the samples from a list of all members of the population starting from the random starting point position and then for every successive sampling interval position.

# Complex survey design: systematic sample

# Complex survey design: stratified random sample/clustered random sample

- Draw a simple random sample of Y number from each sub-grouping or strata or cluster of the population to get X number of sample from the total population; or,
- Draw a simple random sample of Z number from a simple random sample of Y number of sub-grouping or strata or cluster of the population to get X number of sample from the total population.

# Complex survey design: sample size

- Sample size estimation will be dictated by the complex survey design;
- Rule of thumb: A sample drawn via **simple random sample** will require smaller sample sizes compared to a sample drawn with stratification or clustering;
- Stratified and/or cluster samples often require some form of sample size inflation to account for loss of sampling variance due to increase homogeneity of samples from within a cluster.

# Complex survey analysis

- Complex survey analysis is shaped by the survey design;
- Analysis must factor in how the sample was drawn and the probabilities of a sample being selected;
- The probability of a sample being selected is called its **sampling weight**;
- The **sampling weight** is used to calibrate the contribution of sample to the overall outcome measure/indicator.
- Analysis that doesn't take into account survey design and **sampling weight** can potentially produce biased (inaccurate) results.

# Using R for complex survey analysis: {survey} package

- {survey} package authored and maintained by Thomas Lumley is the main R package that provides functions for implementing complex survey analysis;
- It can be installed from CRAN and loaded to R as follows:

```
install.packages("survey")
library(survey)
```

# Using R for complex survey analysis: `{survey}` package

- The typical workflow for complex survey analysis using the `{survey}` package is:
  - Describe the sample design to R and keep as an R object
  - Apply required analysis to the design-specified R object

# Example of an analysis on a simple random sample

```
library(survey)      # Load survey package
data(api)           # Load api dataset

## Describe simple random sample design
srs_design <- svydesign(ids = ~1, data = apisrs)

## Estimate the population mean of enrollees across the schools
svymean(~enroll, srs_design)

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



# Example of an analysis on a simple random sample

```
## Estimate the population total of enrollees across the schools  
svytotal(~enroll, srs_design)
```

```
##          total      SE  
## enroll 116922 5564.2
```

```
## Estimate the total number of different types of schools in California  
svytotal(~stype, srs_design)
```

```
##          total      SE  
## stypeE    142 6.4333  
## stypeH     25 4.6888  
## stypeM     33 5.2625
```



# Examples of an analysis on a stratified sample

```
## Describe stratified sample design
strat_design <- svydesign(ids = ~1, strata = ~stype, fpc = ~fpc, data = apistrat)

## Estimate the population mean of enrollees
svymean(~enroll, strat_design)

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



# Examples of an analysis on a stratified sample

```
## Estimate the population total of enrollees  
svytotal(~enroll, strat_design)
```

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

```
## Estimate the total number of enrollees by school type  
svytotal(~stype, strat_design)
```

```
##          total SE  
## stypeE   4421  0  
## stypeH   755   0  
## stypeM  1018   0
```



# Useful resources

- Thomas Lumley's book *Complex Surveys: a guide to analysis using R*, published by Wiley.
- Thomas Lumley's accompaniment website for the Complex Surveys book -  
<http://r-survey.r-forge.r-project.org/svybook/index.html>

# Thank you!

Slides can be viewed at <https://ecohealthalliance.github.io/complexsurveys> or PDF version downloaded at

<https://ecohealthalliance.github.io/complexsurveys/complexsurveys.pdf>

R scripts for slides available at <https://github.com/ecohealthalliance/complexsurveys>