Short Course on Statistics and Data Analysis

Brian Fannin, Mary Meyer, Jean Opsomer

May 5–9, 2014

Introduction of Us

- Brian Fannin:
- Mary Meyer:
- Jean Opsomer:

Introduction of You

Course Overview

- Day 1
  1. Review of R and R-Studio
  2. Survey: recap of key characteristics
  3. Survey: introduction to `R survey` package
- Day 2
  4. Survey: descriptive statistics
  5. Data analysis: plotting and linear regression

Course Overview (2)

- Day 3
  6. Data analysis: logistic regression
  7. Survey: plotting
- Day 4
  8. Survey: linear and logistic regression
  9. Survey: plotting and mapping your data
- Day 5
  10. Apply what you've learned: data analysis/survey

1. Review of R and R-Studio

2. Survey: Recap of Key Characteristics

- Terminology: finite population, elements, sampling frame, sampling units
- Probability sampling
- Stratification, clustering, weights
- Survey estimation and inference

Survey Terminology

- In surveys, the target of estimation is a specific finite population, e.g.
  - adult women in Kigali (of specific age, on specific date)
  - children attending elementary school in Rwanda
  - coffee farms in Rwanda
  - gorilla families in Volcanoes National Park
  - ...
- Goal of survey: describe characteristics of finite population

Survey Terminology (2)

- Population is composed of elements
  - a woman (living in Kigali, of specific age, on specific date)
  - a child
  - a farm
  - a gorilla family
  - ...
- Survey data are collected on sample of elements

Survey Terminology (3)

- In practice, often difficult to directly sample elements from finite population
- Instead, sampling frame is used, e.g.
  - list of dwellings in Kigali (arranged in districts and villages)
  - list of elementary schools in Rwanda
  - list of farms in Rwanda
  - map of Volcanoes National Park, divided into geographic units
  - ...

Survey Terminology (4)

- Sampling frame is collection of sampling units
  - sampling units can be elements
  - sampling units can contain smaller sampling units (PSU: primary sampling units, SSU: secondary sampling units...)
- E.g.
  - a dwelling, containing multiple individuals
  - a school, containing children
  - a farm, which might/might not grow coffee
  - a geographic unit, which might/might not have gorillas present at time of survey
  - ...

Survey Terminology (5)

- When conducting a survey, important to have access to good sampling frame
  - undercoverage: if sampling frame is “too small,” this causes bias
  - overcoverage: if sampling frame is “too large,” this causes inefficiency
- E.g.
  - undercoverage: some farms are not on list, because list is out of date
  - overcoverage: some farms on list do not grow coffee

- For now, ignore differences between population and frame, sampling units and elements
- Probability sampling: sample is selected from population using random selection process
  - “random” $\neq$ “equal probability”
  - “random” $\neq$ ”unknown”
- When sample is obtained from population by probability sampling, resulting estimates are statistically valid and have known statistical properties

A Toy Example

- District with 4 farms

Farm	Size	Coffee
1	4	1
2	6	3
3	6	5
4	20	15
	36	?

- Known: total number of farms, farm sizes
- Unknown: coffee crop harvested
- Goal: estimate total coffee planted in district using probability sampling

Toy Example: Scenario 1

- Sampling design: draw sample of size 2 out of 4 farms with equal probability
- Randomness comes from sampling design (=how sample is selected), while population is fixed
- We can enumerate all possible samples from this (toy) population

$\rightarrow$ Let us do this here...

  * Estimator
  * Sampling distribution, expectation, variance
  * Variance estimation

Toy Example: Scenario 2

- General survey principle: when information is available for population, it can be used to improve the precision of survey estimators
  - here: farm 4 is major contributor to total crop
- Improved sampling design: always select farm 4, select 1 of remaining farms with equal probability

$\rightarrow$ example of stratification: population is divided into non-overlapping strata, which are each surveyed independently

Toy Example: Scenario 3

- General survey principle: when information is available for population, it can be used to improve cost efficiency of survey
  - here: suppose farms 1 and 2 are located very close to each other
- More cost-effective sampling design: select sample of size 2 with equal probability, but farms 1-2 are treated as single sampling unit

$\rightarrow$ Example of clustering: population is divided into non-overlapping clusters, which are randomly selected

Survey Estimation and Inference

- General principles
  1. sampling design can be complicated, but it is known
  2. unbiased (approximately) estimators use weighting with inclusion probabilities (available for each design)
  3. variance estimation: exact formulas available for each design, but often replaced by simpler approximations in practice

Survey Estimation and Inference (2)

- Variance estimation principles
  1. always account for weighting
  2. always account for stratification
  3. always account for first level of clustering (PSUs), but can ignore subsequent levels (SSUs, etc)
  4. can account for sampling fractions

3. Survey: Introduction to `R` `survey` Package

- Estimation and inference for surveys requires specialized software
  - weighted estimates
  - variance estimation includes stratification, clustering, sampling fractions
- In R, `survey` package provides modern and convenient environment for analyzing survey data

3. Survey: Introduction to R survey Package (2)

- Analyzing data using `survey` involves two steps
  1. create design object
  2. perform analysis
- Consider toy example again, but implementing sampling and estimation with `survey`
- Reference: Thomas Lumley (2010), "Complex Surveys: A Guild to Analysis Using R," Wiley.

3. Survey: Introduction to `R` `survey` Package (3)

- Main R commands we will use:
  - `sample`: draw a random sample from a list
  - `svydesign`: create a design object
  - `svytotal`: estimate totals using survey data and a design object

Syntax

- sample([list], [sample size], probs=[probabilities])
- svydesign(~[PSU variable], strata=~[stratum variable], weights=~[weight variable], fpc=~[stratum size variable], data=[dataset name])
- svytotal(~[survey variable1]+[survey variable2], [svydesign name])

4. Survey: Descriptive Statistics

- Now ready to use a “full-sized” dataset and perform survey analysis
- Topics: means, totals, multiple variables, new variables, quantiles, domains/subpopulations, ratios

A Realistic Example

- Sampling Frame: file containing 14837 records with variables
  - Prov.Name (5 provinces)
  - Province, District, Secteur, Cellule
  - Village (unique name within province)
  - HH (number of households in village)

A Realistic Example (2)

- Village-level variables created:
  - Dist.Water: average distance to improved water source
  - Vaccinated: fraction of children who are vaccinated
  - HH.Size: average household size
  - HH.Size.Adult: average number of adults per household
  - Prim.School.M: fraction of adult males who completed primary school
  - Prim.School.F: fraction of adult females who completed primary school
  - Birth.Weight: average birth weight

A Realistic Example (3)

- Sample is drawn by Stratified PPS sampling
  - stratified by province
  - sample size of 100 villages per stratum (500 total)
  - PPS: Probability of selection Proportional to village Size (=number of households)

Preparing for Analysis

1. Add missing design variables to sample data file: stratum, cluster, weights, stratum sizes
2. create design object using svydesign()

Analysis (1)

1. Population means and totals

   $\rightarrow$ svymean(), svytotal()

2. Contrasts and creating new variables

   $\rightarrow$ svycontrast(), update()

3. Population quantiles

   $\rightarrow$ svyquantile()

4. Population ratios

   $\rightarrow$ svyratio()

Analysis (2)

5. Subpopulations

$\longrightarrow$ subset(), svyby()

Analysis Exercises

1. Estimate the average village size and the average number of adults per village in Rwanda, and give standard errors
2. Estimate the total of number of children in Rwanda, and give standard errors
3. Estimate the total number of unvaccinated children in Rwanda, and give standard errors
4. Estimate the fraction of unvaccinated children in Rwanda, and give standard errors
5. Estimate the 10%, 25%, 50%, 75%, 90% percentiles of the village fractions of males and females who completed primary school in Rwanda, and give confidence intervals

Analysis Exercises (2)

6. Estimate the average birthweight overall and by province, and give standard errors
7. Estimate the average distance to an improved water source overall and by province, and give standard errors
8. Estimate the average birthweight for villages above and below the national median distance to an improved water source, and give standard errors

Two-stage (Cluster) Sampling

- In two-stage sampling, PSUs (clusters) are selected according to specified sampling design A, and SSUs are selected in each PSU according to specified sampling design B
- Common two-stage design:
  1. design A is (stratified) PPS with respect to PSU size
  2. design B is simple random sampling in PSU
- E.g. PSUs are villages, SSUs are households

A Realistic Example 2

- Starting from Rwanda village frame, 10 villages are selected from each province using PPS
- In each village, 10 households are selected using simple random sampling
- Two types of variables
  - PSU-level variables
  - SSU-level variables

PSU-level Variables

- As before:
  - Prov.Name (5 provinces)
  - Province, District, Secteur, Cellule
  - Village (unique name within province)
  - HH (number of households in village)
  - (other variables inherited from PSU survey)

SSU-level Variables

- Household-level variables created:
  - HH.ID: unique household identifier
  - Dist.Water: distance to improved water source
  - Vaccinated: whether the children in the family are vaccinated (yes/no)
  - HH.Size: household size (count)
  - HH.Size.Adult: number of adults in household (count)

Preparing for Analysis for Two-Stage Sampling

- Principles:
  - sampling design A (for PSUs) needs to be completely specified
  - sampling weights are product of weights for designs A and B
  - adding information for design B (for SSUs) improves precision of variance estimates

Preparing for Analysis for Two-Stage Sampling (2)

1. Add missing design variables to sample data file:
   - weights
   - strata for sampling design A, PSU identifiers, stratum sizes (in number of PSUs) for sampling design A
   - if available: strata for sampling design B (in each PSU), SSU identifiers, stratum sizes (in number of SSUs for each PSU) for sampling design B
2. create design object using svydesign()

Analysis for Two-Stage Sampling

$\Rightarrow$ Once design object is created, analysis is identical to what we did for single-stage designs earlier!

Analysis Exercises

1. Estimate total number of adults and total number of children in Rwanda and give standard errors
2. For your previous question, compare the results when giving only 1st-stage design information vs. giving complete 2nd-stage design information
3. Estimate the fraction of children who are vaccinated in Rwanda, and give standard errors
4. Estimate the total number of unvaccinated children in Rwanda and by province, and give standard errors

Analysis Exercises (2)

5. Estimate the 25%, 50%, 75% percentiles of the household sizes in Rwanda, and give confidence intervals
6. Estimate the 25%, 50%, 75% percentiles of the household sizes by province