Jon Wakefield

Departments of Biostatistics and Statistics
University of Washington
Course Details

Course Timetable
Course Details
Logistics

- **Instructor:** Jon Wakefield, jonno@uw.edu
- **Time:** Tuesday/Thursday 10:00–11.20, via zoom.
- **JW Office Hour:** Tuesday 2:00–3:00, https://washington.zoom.us/j/92020700746, or by appointment.
- **Teaching Assistant:** Peter Gao petergao@uw.edu
- **Office Hour:** To be determined.
- **Text:** There is no required text; recommended texts are given below.
- **Materials:** Available at the canvas course website.
THE SYLLABUS AS ACCORDING TO THE CATALOG DESCRIPTION

- Design and implementation of selection and estimation procedures.
- Emphasis on human populations.
- Simple, stratified, and cluster sampling.
- Multistage and two-phase procedures.
- Optimal allocation of resources.
- Estimation theory.
- Replicated designs,
- Variance estimation.
- National samples and census materials.
What will not be covered

Design of survey instruments:

- How to collect data accurately: misunderstanding questions, refusal to participate, recall bias, etc.
- Cognitive aspects of survey designs: question placement and order effect, interview fatigue, etc.
- Piloting questionnaires.

Practical aspects of implementation:

- Finding lists to sample from.
- Costs and benefits of different interview/questionnaire types (telephone-based, web-based, in-person, etc).
- Interviewer training, data entry.

We will look at the implications for statistical analysis of some of these issues, however.
Main texts

  Written around the R survey package.

  Very well-written and clear mix of theory and practice.

  Specific to health data.

  Excellent on the theory though steep learning curve and hard to dip into if not familiar with the notation. Also, anti- model-based approaches.
All computing will be carried out in R, using the survey package.

Most of the methods we use are implemented in Stata also, but here we use R only, and will make extensive use of Thomas Lumley’s survey package.

Many other choices, e.g. SUDAAN (SAS), VPLX (US Census Bureau, free but limited).

For more choices, see http://www.hcp.med.harvard.edu/statistics/survey-soft/
The course grade will be given on the basis of:

- Homworks (around 5).
- A project of each student’s choice, to be handed in at the end of the quarter.

More details on the project may be found on the class website.
When possible, lectures will be split between:

- Describing techniques and methodology.
- Implementation via demonstrations of analyses with real survey data, using the `survey` package in R.

Very little of the course evaluation will involve mathematical derivations, but for completeness I will include proofs of important results. Slides of this nature will have titles in red.
Course Timetable
WEEK 1
- L1: Course overview, motivation, introduction to survey package, examples, need for specialized methods of analysis; design-based inference.
- L2: Simple random sampling (SRS); SRS with survey package.

WEEK 2
- L3: Stratified random sampling, systematic sampling; NMIHS data.
- L4: Optimal stratified random sampling and choice of strata.

WEEK 3
- L5: R programming.
- L6: Cluster sampling, multistage sampling.

WEEK 4
- L7: Ratio and quantile estimation.
- L8: Scottish household survey case study.

WEEK 5
- L9: Design considerations.
- L10: Graphics and visualization.
WEEK 6
  - L11: Linear regression.
  - L12: Resampling and replicates.

WEEK 7
  - L13: Logistic regression.
  - L14: Post-stratification and raking.

WEEK 8
  - L15: Calibration.
  - L16: Two-phase studies.

WEEK 9
  - L17: Missing data and non-response.
  - L18: Small area estimation.

WEEK 10
  - L19: Model-based estimation.
  - L20: Model-assisted estimation.