# BST 5420

# Sampling Theory and Survey Design in Public Health

Steven E. Rigdon

Spring 2019

# Elements of the Sampling Problem

- "The objective of sample surveys is to make inferences about a population from information contained in a sample selected from that population." [1]

[1] from *Elementary Survey Sampling, 7th ed.*, by Mendenhall, Schaeffer, Ott, and Gerow"

# §1.2 Requirements of a Good Sample: Terminology

- **Observation unit** – An element is an object on which a measurement is taken (often, but not necessarily, a person)
- **Inference** – saying something about a quantity that we do not know, e.g., point estimate, confidence interval, result of hypothesis test
- **Target Population** – The complete collection of elements about which we wish to make an inference
- **Sample** – A subset of the population selected to represent the population.

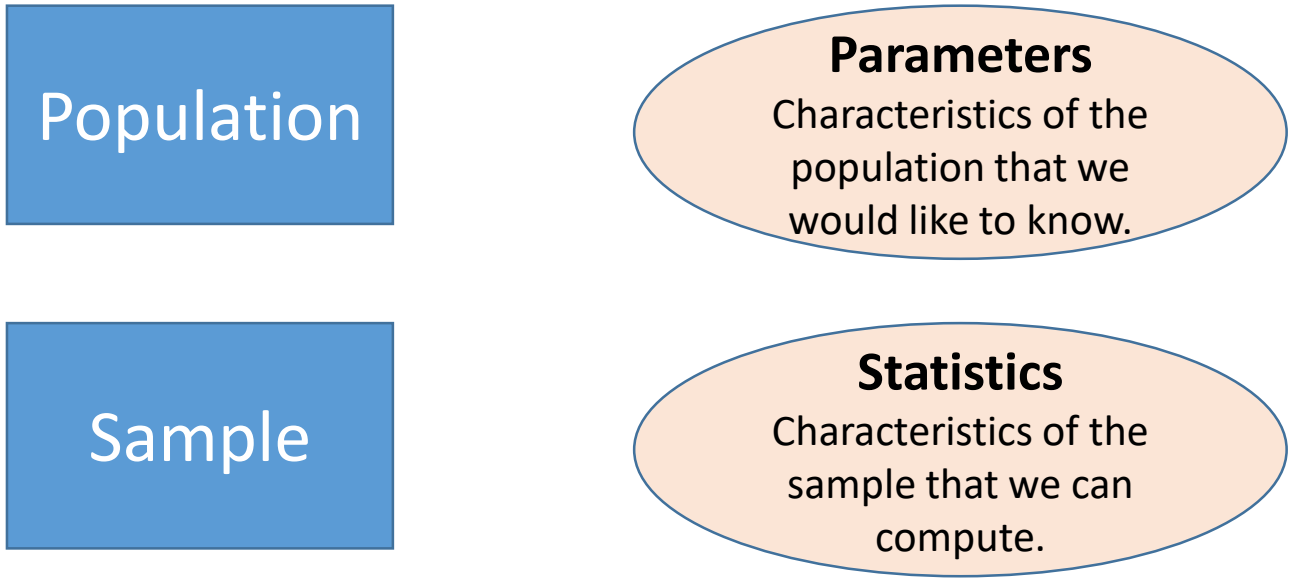
# Terminology (continued)

- **Sampled population –** The collection of all possible observation units that might have been chosen. (The population from which the sample was chosen.)  Ideally

  Sampled Population = Target Population

- **Sampling unit –** A unit that can be selected for a sample.  Could differ from observation unit, e.g. households are selected, but persons are interviewed.
- **Sampling Frame –** A list (or equivalent) of sampling units in the population from which a sample may be selected, e.g., list of citizens and their phone numbers.

# Population – Parameter; Sample - Statistic



# Experiment vs. Survey

- In an ***experiment***, some treatment is imposed on the subjects who are involved in the study.
- In a ***survey*** or ***observational study***, no treatments are imposed, and subjects are just “observed”.

# Design of Studies: Examples

- Investigators want to see the effect of various jobs on the risk of getting carpel tunnel syndrome. They take a sample of workers, identify their job category, and see whether they contract CTS.

  Survey or observational study

- Investigators want to see the effect of Vitamin E on the progression of Alzheimer's disease. They take a sample of patients with Alzheimer's disease, study their diet to estimate their Vitamin E intake, and study their progression of Alzheimer's.

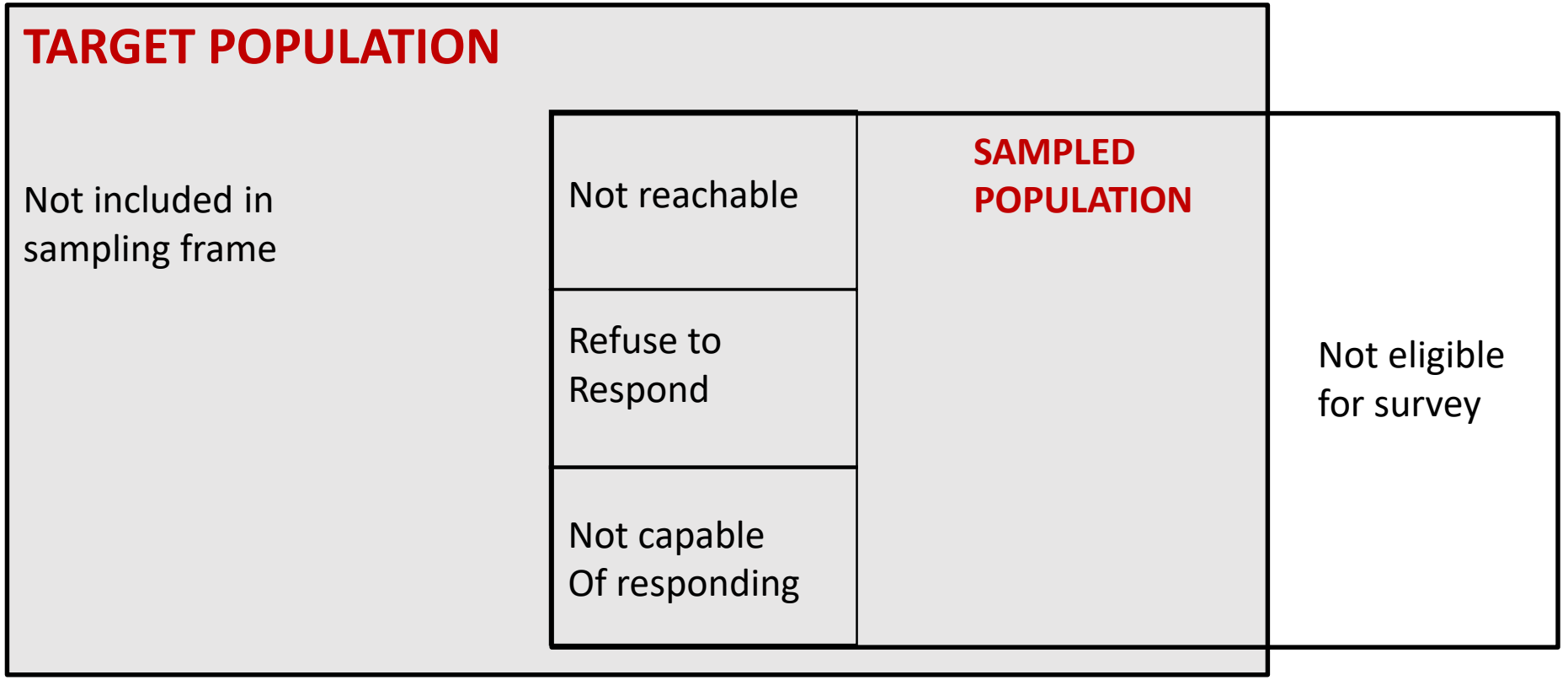
  Survey or observational study

# Design of Studies: Examples

- Investigators want to see the effect of Vitamin E on the progression of Alzheimer's disease. They take a sample of patients with Alzheimer's disease. Half are given a dose of Vitamin E daily, and half are given a placebo (a pill identical to the observational drug, but with no active ingredients)

  Experiment

# Populations and Samples



# §1.3 Selection Bias

Selection bias occurs when some observation units are sampled at different rates, i.e., some are more likely than others to be in sample

**Convenience Sample** – sample taken because the units are readily available

**Self selection** – People choose to whether to participate.

# Examples of Bias

- 1936 Presidential Election (see next slide)
- Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987). E.g. 87% of women are "not satisfied emotionally with their relationships."

| | | |
|---|---|---|
| Self selected | Mailed to professional women's groups | Very long (127 questions) Fatigue factor |
| Loaded questions | Vague wording | |

- CPH Exam – More than 90% of SLU students pass
- Internet Polls

# Gallup Polls 1936 – 1968

| Year | Gallup Final Result (Winner) | Election Result (Winner) | Error of Estimation |
|---|---|---|---|
| 1936 | 55.7% FDR | 62.5% FDR | – 6.8% |
| 1940 | 52.0% FDR | 55.0% FDR | – 3.0% |
| 1944 | 51.5% FDR | 52.3% FDR | – 0.8% |
| 1948 | 44.5% Truman | 49.9% Truman | – 5.4% |
| 1952 | 51.0% Eisenhower | 55.4% Eisenhower | – 4.4% |
| 1956 | 59.5% Eisenhower | 57.8% Eisenhower | 1.7% |
| 1960 | 51.0% Kennedy | 50.1% Kennedy | 0.9% |
| 1964 | 64.0% Johnson | 61.3% Johnson | 2.7% |
| 1968 | 43.0% Nixon | 43.5% Nixon | – 0.5% |

# Quota Sampling

- Collect data from subjects until a quota for each of several demographic categories is met.
- For example: by gender: 50% men, 50% women by race: 66% white, 16% Hispanic, 13% African American, 5% Asian
- Or, by both categories together

| | Men | Women |
|---|---|---|
| White | 33.0% | 33.0% |
| Hispanic | 8.0% | 8.0% |
| African Am | 6.5% | 6.5% |
| Asian | 2.5% | 2.5% |

# Probability Sample

Each unit in the population has a known probability of selection, and a random mechanism is used to choose specific units.

- Most applications in textbook are probability samples.
- Probability sample allow us to make inference to the population.

# Quota Sampling vs. Probability Sampling

- Quota sampling – Find 33 white men, 33 white women, etc. and survey them.
- Quota sampling – Subjectivity on who to choose … based on availability, approachability, …
- Probability sampling – Choose people from sampling frame and contact ***them***, not anybody else.
- Problem with quota sampling … there are many more demographic categories than you might pre-specify.

# Gallup Polls 1936 – 1968

| Year | Gallup Final Result (Winner) | Election Result (Winner) | Error of Estimation |
|---|---|---|---|
| 1936 | 55.7% FDR | 62.5% FDR | – 6.8% |
| 1940 | 52.0% FDR | 55.0% FDR | – 3.0% |
| 1944 | 51.5% FDR | 52.3% FDR | – 0.8% |
| 1948 | 44.5% Truman | 49.9% Truman | – 5.4% |
| 1952 | 51.0% Eisenhower | 55.4% Eisenhower | – 4.4% |
| 1956 | 59.5% Eisenhower | 57.8% Eisenhower | 1.7% |
| 1960 | 51.0% Kennedy | 50.1% Kennedy | 0.9% |
| 1964 | 64.0% Johnson | 61.3% Johnson | 2.7% |
| 1968 | 43.0% Nixon | 43.5% Nixon | – 0.5% |

Quota Samples Used

Probability Samples Used

# Sources of Bias

- **Undercoverage** – failing to include all of the target in the sampling frame
- **Self selection**
- **Convenience Sampling**
- **Judgement sampling** – Interviewers (oftentimes visually) choose who to select
- Using any selection method where the probability of inclusion is associated or correlated with the variable being measured (even if the association is unknown to the investigator)

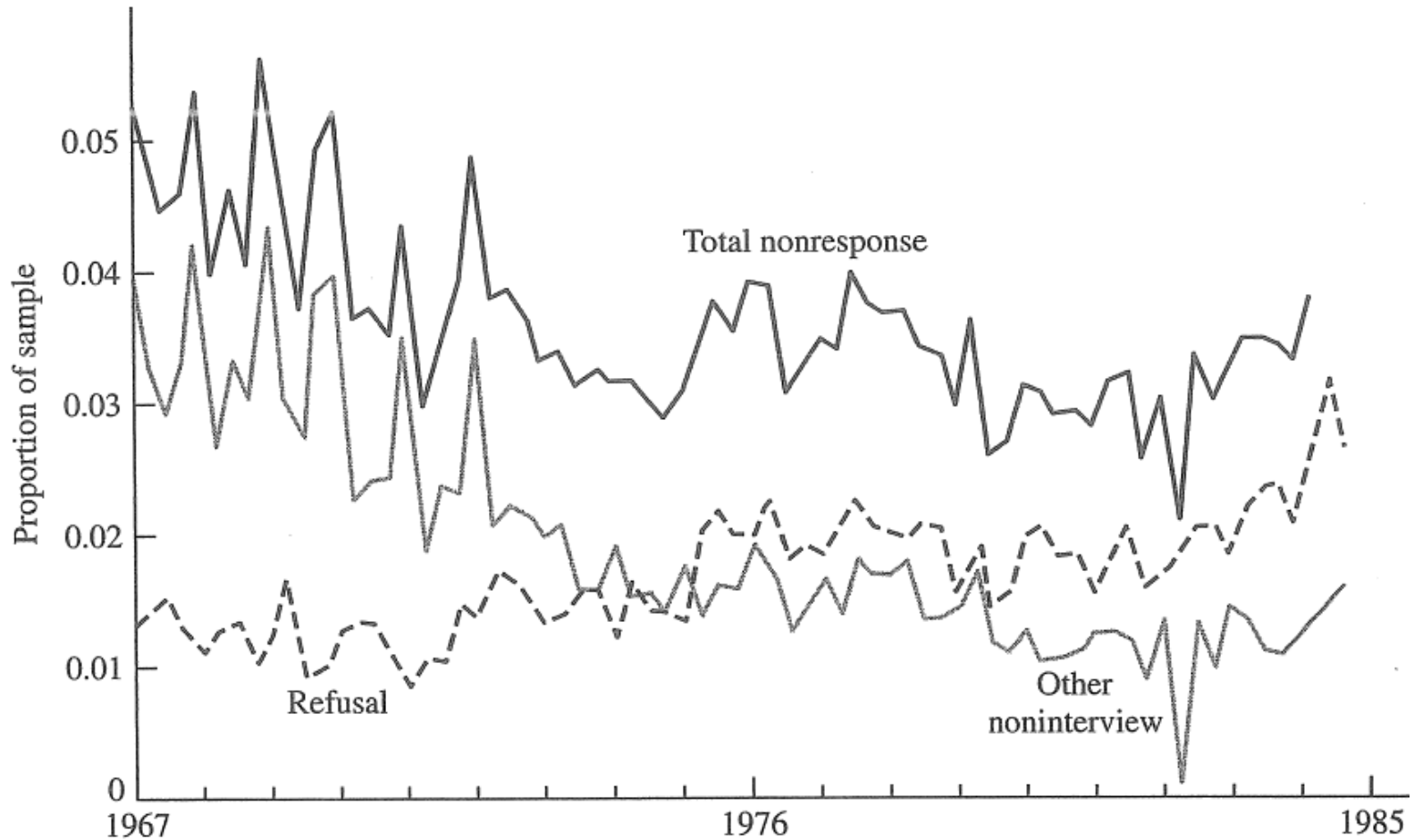
# More Sources of Bias

- **Overcoverage** – Units not in target population appear in sampling frame. (Often occurs when sampling frame is not screened.)
- **Multiple listings**
- **Substituting a convenient unit instead of the one selected**
- **Nonresponse**

# Nonresponse

FIGURE 2.1

Nonresponse rates for the National Health Interview Survey, 1967–1985



# §1.4 Measurement Error

**Measurement error** occurs when response has a tendency to differ from the true

**Sources:**

| | | |
|---|---|---|
| Lying | Don’t understand question | Forgetting (recall bias) |
| Give different answers to different interviewers | Give answers the interviewer wants to hear | Vague words (e.g. Do you own a car?) |

# §1.5 Questionnaire Design

- Always test the questions before “going live”.
- Keep it simple, clear, and short
- Use specific rather than general questions.  For example, rather than “Have you been attacked or threatened recently?” ask “Has anyone attacked or threatened you in any of these ways? (a) … (b) … (c) …”
- Stick to the topic of interest.
- Decide on open or closed questions

# §1.5 More on Questionnaire Design

- Report the exact question asked.
- Avoid loaded or leading questions, that is, questions that will motivate the respondent to answer the way you would like.
- Avoid double negatives.
- Write questions that elicit honest responses.
- Avoid **double-barreled** questions (i.e., questions that ask two things at once). For example “Do you agree with Bill Clinton’s $50 billion bailout of Mexico?” Two questions: Mexico bailout and Clinton’s opinion

# §1.5 Still More on Questionnaire Design

- Use forced-choice, rather than agree/disagree.

  Q1: Do you agree or disagree with this statement: Most men are better suited emotionally for politics than are most women?

  Q2: Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?

Actual responses Vs. schooling

| | **Years of Schooling** | | |
|---|---|---|---|
| | 0 – 11 | 12 | 13+ |
| Q1: Percent “Agree” | 57 | 44 | 39 |
| Q2: Percent “men better suited” | 33 | 38 | 28 |

# §1.6 Sampling and Nonsampling Errors

- **Census** – an attempt to choose the entire population as the sample
- **Sampling Error** – the error made because we took a sample, not a census.
- **Nonsampling Error** – errors that would be present even if we were able to take a census.  Examples: most of the biases described previously

| | |
|---|---|
| Vague terms used in questions | Lying |
| Nonresponse | Double barreled questions |

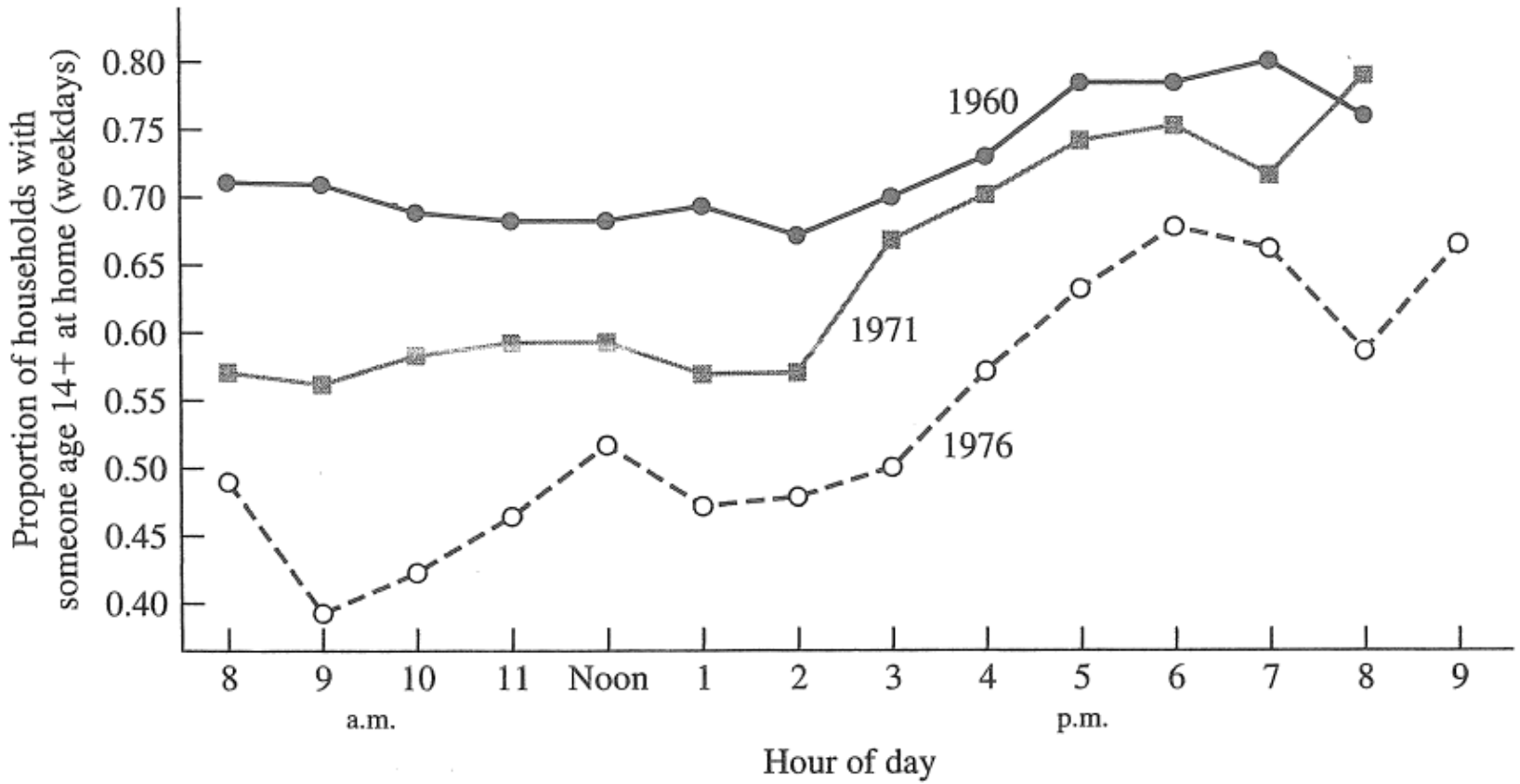
# Reducing Errors in Surveys

- Call when people are home
- Callbacks
- Rewards and incentives ($, raffle, access to web site)
- Training
- Data checks
- Careful questionnaire construction

# When to call

FIGURE 2.3

Proportion of households in which at least one person aged 14 or older was at home (weekdays)



# Methods of Data Collection

- Personal interviews

  (+) most people respond, interviewer can note specific reactions or clarify questions)

  (–) cost, interviewers must be uniformly trained, may deviate from protocol

# Methods of Data Collection

- Telephone interviews

  (+) lower cost

  (–) frame may not equal population, may not reach household (only 20% of numbers are households, rest are unused, belong to businesses, etc.)

# Methods of Data Collection

- RDD (Random Digit Dialing)  Select an exchange (area code + first 3 digits) from list of 42,000 exchanges used mostly for residences, then randomly select the last 4 digits.
- RDD reaches residence 23% of the time
- Modified RDD – if the above reaches a residence, then next call, which would come from same exchange, uses the same number, but with the last 2 digits changed randomly.  Reaches residence 57%.

# Methods of Data Collection

- Self administered questionnaire – usually by mail, but could be web survey

  (+) much lower cost

  (–) very low response rate (resulting in selection bias)

# Direct Observation

- Not all surveys involve people as elements
- Vehicles on a street, medical records, etc.

# Types of Probability Samples We Will Study

- **simple random sample** – A simple random sample (SRS) of size $n$ is one with the property that *any* set of size $n$ has the same chance of being selected.
- **stratified random sample** – First, divide the population into groups that are homogeneous in some sense; each of these is called a **stratum**. Then, take a SRS from each stratum.

# Types of Probability Samples We Will Study

- **cluster sample** – The population is broken into many pieces that are naturally close together; these are called **clusters**. First, take a SRS of clusters. Then survey each element of the selected clusters.
- **two-stage cluster sample** – Take a SRS of clusters. Then within each cluster, take a SRS of elements.
- **systematic sample** – From the sampling frame, select a random starting point and then select every $k^{th}$ element.

# Outline for much of BST 5420