Data collection principles

Sumanta Basu

Reading

- Reading: Textbook Sections 1.1, 1.2.1, 1.3 - 1.5
- Recommended Exercise: 1.1, 1.9, 1.13, 1.17, 1.27, 1.29
- OpenIntro Lab 0: Introduction to R (under blackboard folder ‘Lab1’)

Data Sampling

main focus of course will be on

- data analysis
- statistical inference
- careful interpretation of results

but **how you collect data** is essential (e.g., Agent Orange example)

many contexts for data collection

- surveys
- observational studies
- scientific experiments

Sampling: a crucial statistical concept

historically, statistics concerned gathering census-style data on entire population of interest (“state”)

> “You don’t have to eat the whole ox to know the meat is tough.” – Samuel Johnson

why not examine entire population?

- expensive, slow
- may be infeasible (e.g., if observation destroys what’s being measured)

modern notion of statistics: learn about the population by examining a small fraction of it

Statistical Sampling

Terminology

Population

- the collection of all units of interest
- implied by the research question being studied
- e.g., all Cornell graduate students

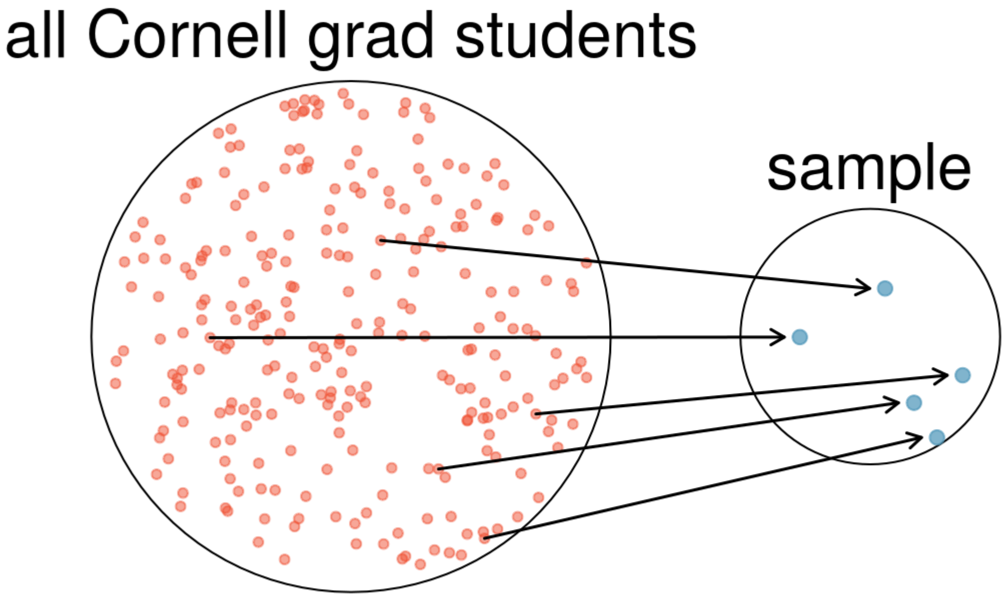
Sample

- any subset of the units from the population
- e.g., 20 Cornell graduate students

Unit or **subject** or **element**

- individual entity from the population
- e.g., a Cornell graduate student

Example



Some distinctions

Randomness

- Population is **fixed / non-random / deterministic**
- Samples could be (often are) **random**

Observability

- Population can be (often is) unobservable
- Sample is observable (can measure and do analysis with it)

What is the **source of randomness** in a sample?

- Sampling Scheme / Sampling Strategy !!

How do we choose the sample?

- Critically important question
- A bad sampling strategy can invalidate all conclusions
- A good sampling method ensures that samples are (with high probability) **representative of the population**
- Choice of sampling method depends on objective of study and feasibility

Sampling design process

Define target population

- What collection of units would you like to describe?
- Implied by the research question being studied
- Population exclusions? (e.g., patients too ill for study)

Example

Research question: What proportion of Cornell graduate students entered graduate studies immediately after graduating college?

- Population?
- The unit?
- The sample?
- Exclusions?

Determine sampling frame

enumerate the population - i.e., a “list” of the population which will help you reach the sample

sampling units may be households, streets, telephone numbers, fields, etc.

example: if population is all Cornell students, could use

- registrar
- university directory

common problems with lists: access, omissions, out-of-date, duplicates

Selecting a sampling design

Probability sampling

- all units in population have a known (non-zero) probability of being included in sample
- simple random sampling
- stratified sampling
- cluster sampling

Non-probability sampling

- some units have zero chance of being included or are included with unknown probability
- convenience sampling, judgment sampling, snowball sampling, quota sampling

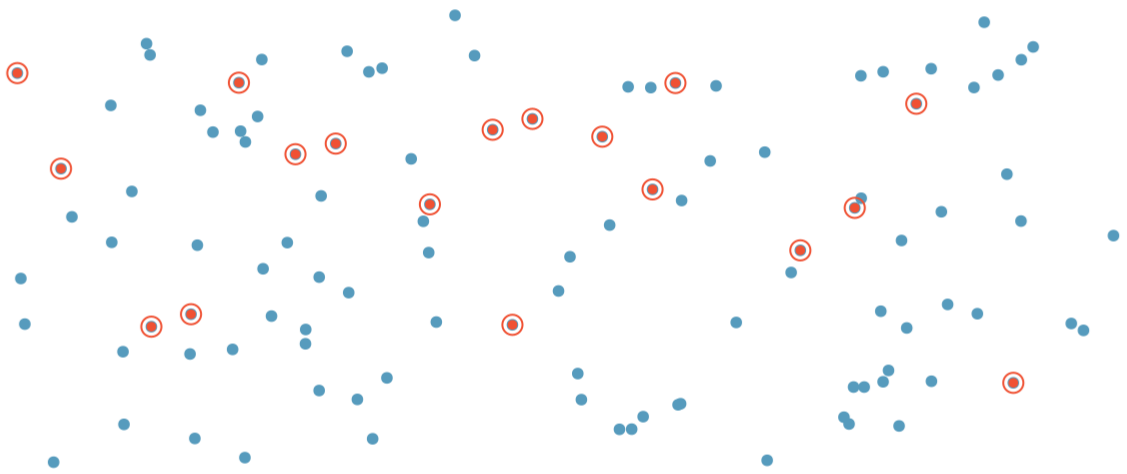
Probability sampling

Probability sampling

- aka random sampling
- objective procedure
- probability of selecting each unit is nonzero and known in advance
- ensures representative sample of the population
- sampling error can be assessed
- results can be projected to population
- more expensive than non-probability samples

Simple random sampling (SRS)

- like “picking names out of a hat”
- an SRS of size n is such that **all possible samples of size n are equally likely**



Is this an SRS of 6010 students?

Include in sample if last name starts with the letter “M”.

SRS of 6010 students

- number students from 1 to 150
- now, choose 10 at random (without replacement)

```
sample(1:150, size=10, replace=FALSE)
```

```
##  [1]  88 134  86  47  71  87  10 106  92 144
```

```
sample(1:150, size=10, replace=FALSE)
```

```
##  [1] 105  53  47 145 110 103  33  37  93  88
```

SRS

- simplest method that gives valid inferences
- intuitive, yet sophisticated, idea
- simplest calculations
- basis of most good sampling methods
- workhorse of the sampling world

SRS

- but is it safe to assume that SRS is representative of whole population?

> Theorem: When a sample is selected by SRS, it differs very little from the entire population.

(We’ll make this precise later.)

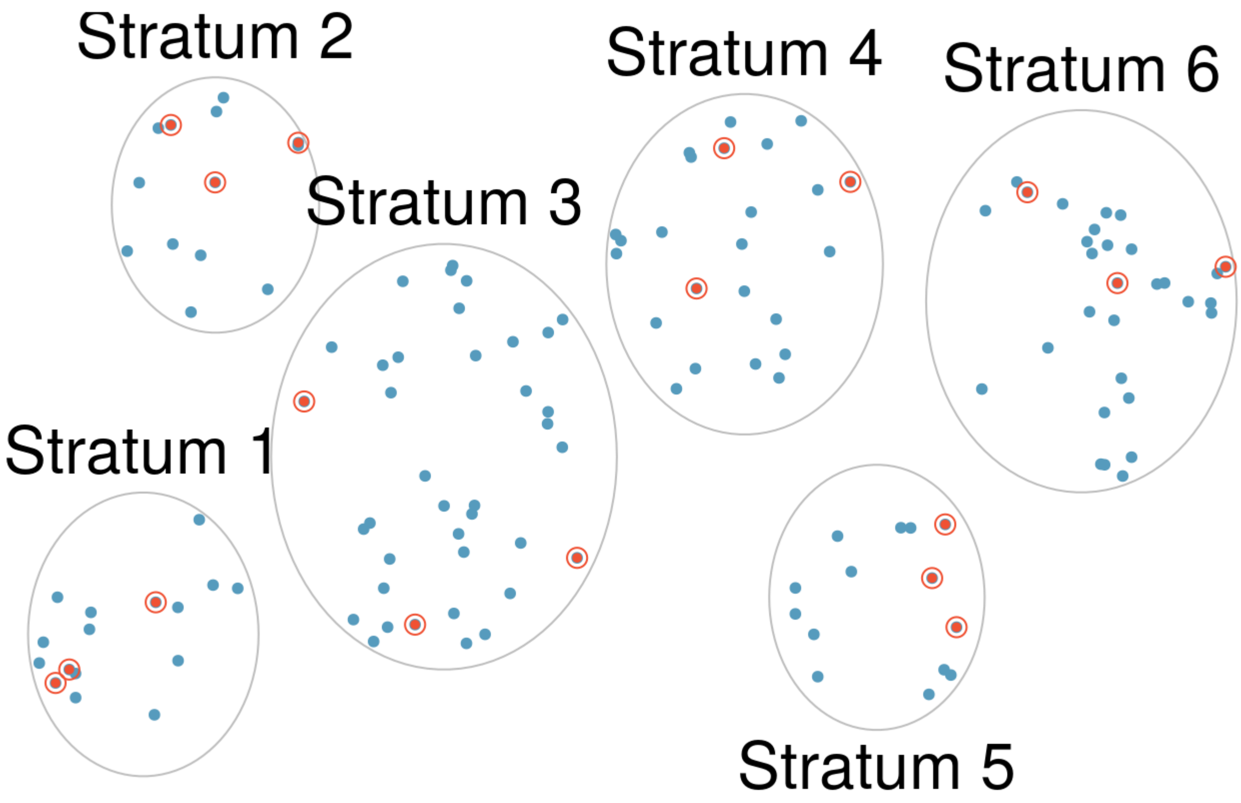
- This course primarily focused on inference based on data obtained from SRS

Stratified random sampling

chosen sample forced to contain units from each “statum” of population

- goal: to equalize “important” variables (e.g., gender, race, geographical area, etc.)
- procedure:
    - partition population (strata are mutually exclusive and exhaustive) based on a characteristic
    - draw simple random samples within each stratum

Stratified random sampling



Stratified random sampling

- smaller sampling error than SRS (removes a source of variation)
- representativeness when **proportional sampling** used
- e.g., 6010 students:

Proportional Sampling

	Social Science	Science
Population	50	100
Sample	5	10

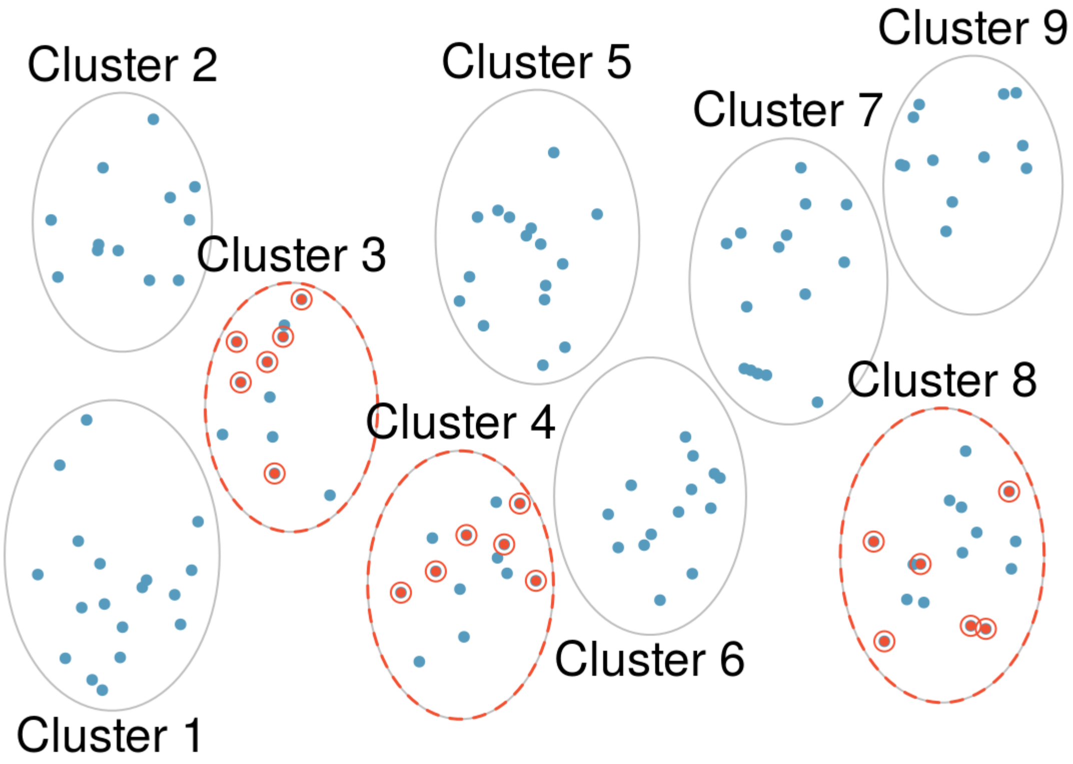
- **disproportional stratified sampling** used when some strata are too small OR more important OR more diverse

Cluster sampling

- partition population into clusters (ideally, each adequately represents population)
- SRS of a few clusters
- within each chosen cluster do SRS

Motivation: more economical

Cluster sampling



Example

- cluster sampling by geographic region
- divide up map, sample regions
- within each region sample units
- can be more economical

Selecting a sampling design

use stratified sampling

- when primary objective is to compare groups
- when it may reduce sampling error (homogeneous strata)

use cluster sampling

- when there are fixed costs associated with each data collection location
- when list of clusters is available but no list of units themselves

Non-probability sampling

Non-probability sampling

- subjective procedure in which **probability of selection for some units is zero or unknown** before drawing the sample

- sampling error cannot be computed

- non-representative sample so results *cannot* be projected to the population

- cheaper and faster, but allows for limited inference

Convenience sampling

- send survey to your friends on facebook
- stop people on the street
- student volunteers (e.g., undergrad psych students!)

Example

Example

Slate (2013):

> “Sixty-seven percent of American psychology studies use college students, for example. This means that many or even most of the subjects are teenagers.”

Judgment sampling

- researcher tries to select a sample that seems to be most appropriate for study
- e.g., sampling from a particular shopping mall that researchers “think” is representative of their target market

Snowball sampling

- initial subjects recruit additional ones (“referral sampling”)
- used to sample difficult to reach populations (e.g., drug users, sex workers, etc.)
- biased (e.g., over-represents people who have a lot of friends)

Quota sampling

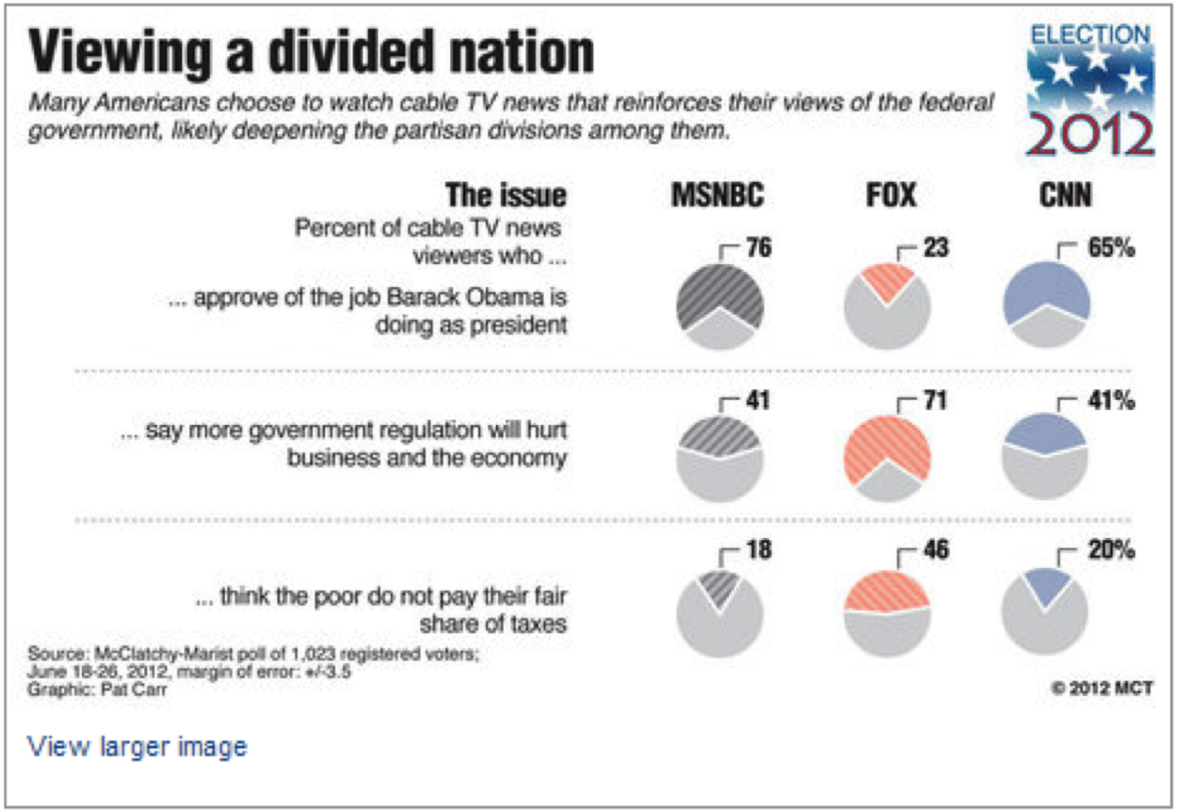
Procedure:

- partition population based on a characteristic (as in stratified random sampling)
- establish how many from each group must be sampled
- non-randomly choose required number from each group

Non-probability sampling

- may be less time consuming or less expensive
- if certain elements have zero probability, then sampling does not fully represent population
- if probabilities of being in sample are not known, cannot project estimates from sample to population
- drawing inferences with a non-representative sample is dangerous (e.g., TV viewer polls)

TV viewer polls



Viewing a divided nation

Many Americans choose to watch cable TV news that reinforces their views of the federal government, likely deepening the partisan divisions among them.

The issue Percent of cable TV news viewers who ...	**MSNBC**	**FOX**	**CNN**
... approve of the job Barack Obama is doing as president	76	23	65%
... say more government regulation will hurt business and the economy	41	71	41%
... think the poor do not pay their fair share of taxes	18	46	20%

Source: McClatchy-Marist poll of 1,023 registered voters; June 18-26, 2012, margin of error: +/-3.5
Graphic: Pat Carr

© 2012 MCT

View larger image

http://www.mcclatchydc.com/2012/07/26/157899/fox-mnsbc-viewers-see-world-differently.html

Errors in sampling

Random sampling error

- sample selected is not representative of population due to chance
- controlled by sample size (larger sample means less likely to have non-representative sample)

Non-sampling error

- systematic error
- **not controlled by sample size**

Types of non-sampling errors

Non-response error

- problematic if non-responders and responders are different

Response or data error

- systematic bias during data collection, analysis, or interpretation
- respondent error (lying, forgetting, etc.)
- interviewer bias
- recording errors
- poorly designed questionnaires

Example: 1948 presidential election

Three major polls predicted Dewey would win.

They were wrong…

What went wrong?

- stopped polling too soon (two weeks before election)
- telephone polls oversampled wealthy
- in person polls used stratified sampling to obtain samples representative of overall US population in terms of race/gender. Not representative of US **voting** population.
- Pollsters used quota sampling

Example: Hormone replacement therapy

1991: Every woman should get on HRT immediately because it prevents heart attacks! (50% reduction)

2002: Every woman should get off HRT immediately because it increases risk of heart attacks! (30% increase)

What happened?

- 1991 data from the *Nurses' Healthy Study*, which could only answer the question "What are the health characteristics of nurses who choose to undergo HRT therapy versus nurses who don't?"
- 2002 data came from randomized clinical trial
- selection bias (healthy volunteer problem)

Experiments

- in the best case, one can randomly select which units are assigned a treatment
- one then compares units in the treatment group to those in the control group

Four fundamental principles in designing experiments

1. **Controlling**: attempting to ensure that units are treated identically except for treatment

2. **Randomization**: randomly assigning units to treatment/control "evens out" any uncontrollable differences and prevents accidental bias

3. **Replication**: increased sample size reduces statistical fluctuations

4. **Blocking**: stratify based on a likely-relevant variable (e.g., gender) and then randomly assign treatment/control within each block.