2. Collecting data

Two fundamentally different ways of collecting data:

Observational study: Passive - record data without interfering with the process or system being observed.

Example: Gallup poll

Experiment: Active - make carefully controlled changes to the system and record response.

Example: Salk vaccine field trial

Observational study and experiment differ fundamentally in the kind and strength of conclusions that can be drawn.

In particular, it is impossible to conclusively establish causality from an observational study - a weakness of epidemiology.

Distinguish between *providing evidence for decision making under uncertainty* and *establishing scientific truth*.

References:

D.S. Moore, *Statistics: Concepts and Controversies*, Freeman, 1979, Ch. 1, 2.

D. Freedman, R. Pisani, and R. Purves, *Statistics*, Norton, 1998, Ch. 1.

D. Huff, *How to Lie with Statistics*, Norton, 1954.

J. Devore and N. Farnum, *Applied Statistics for Engineers and Scientists*, Duxbury, 1999, Ch. 4.

2.1 Observational studies

Population: Entire group of objects about which information is wanted.

Unit: Member of the population.

Variable: Characteristic of interest.

Census: Observe value(s) of variables(s) for all units in the population.

Sample survey: Observe value(s) of variable(s) for subset of the population.

Sample: Subset of population used to gain information about the whole.

Sampling frame: List of units from which sample is chosen.

Why sampling?

- Less costly
- Census might not be possible (destructive testing)
- Sampling might give more accurate results (observer fatigue; need for follow-up)

Examples for use of sampling:

- Gallup poll
- Nielsen ratings
- Accounting
- Assignment of radio stations' royalties to composers

Sampling design vs *convenience sampling*

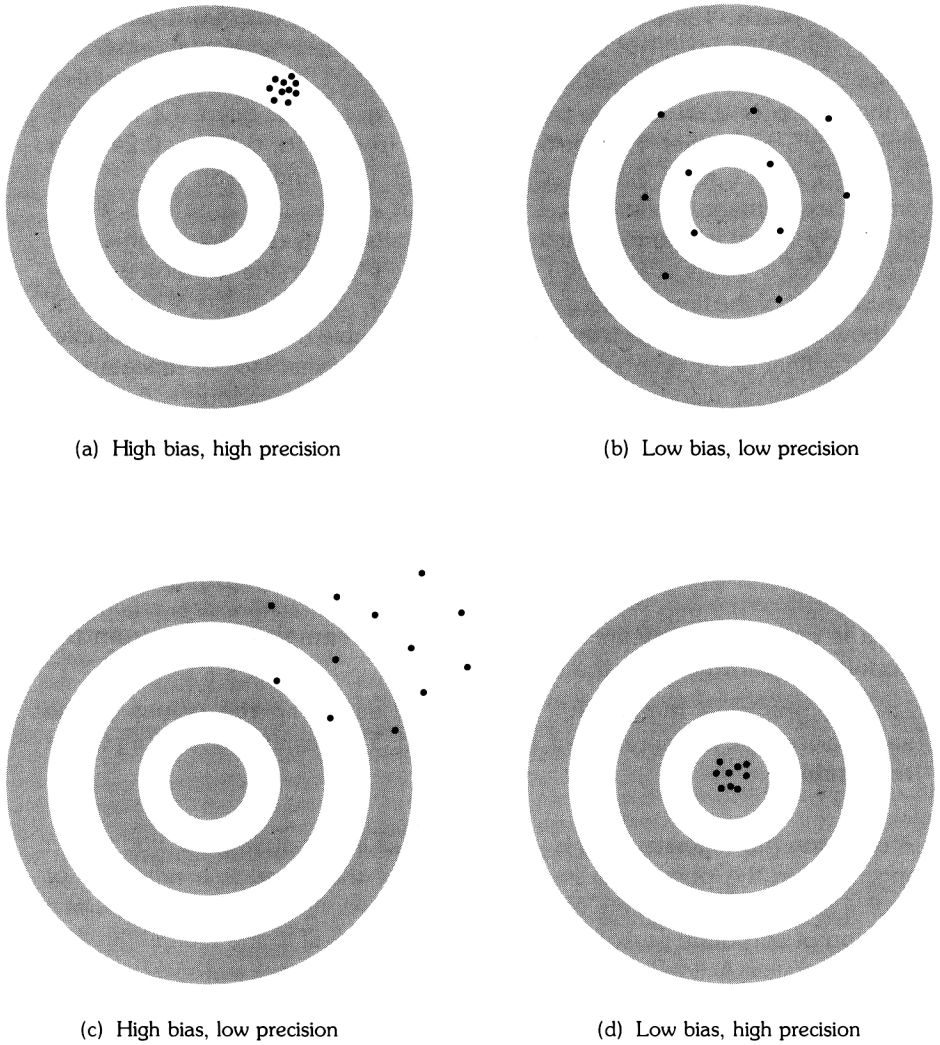
Parameter: Numerical characteristic of the population (such as population mean)

Statistic: Numerical characteristic of a sample (such as sample mean)

We typically us the value of a statistic as an estimate (guess) for the population parameter. For example, we estimate the population mean by the sample mean.

Bias: Systematic error - deviation of sample statistic from population parameter that is consistent accross samples.

Lack of precision: (variance) Variation in value of the statistic accross samples due to randomness in choosing samples.



(a) High bias, high precision

(b) Low bias, low precision

(c) High bias, low precision

(d) Low bias, high precision

Figure 2. Bias and lack of precision in sample results.

Simple random sampling (SRS)

Suppose we draw a sample of size n from a population of size N.

If we sample without replacement, the number of possible samples is

$$\frac{N(N-1)\cdots(N-n+1)}{n!}$$

SRS: All samples have the same chance of being chosen.

Properties of SRS:

- When sampling frame lists entire population, SRS gives unbiased estimates.
- Precision depends on sample size and can be made arbitrarily high by chosing large enough sample.
- As long as sample size is small fraction of population size, *precision depends on sample size, not on population size.*

Recall sampling experiment; will be proved later in class.

Probability sampling

One can often increase precision by giving different units in the sampling frame different probabilities of being included in the sample.

Probability sample: Every unit in the sampling frame has a known nonzero probability of being chosen.

Example: Stratified sampling

- Divide sampling frame into groups (or strata) of units that are of special interest, or that resemble each other.
- Take a separate SRS in each stratum and combine them.

Note: Have to take probabilities into account when computing estimate of population parameter from sample.

Estimating a population parameter by sampling can go wrong

Sampling errors: Errors caused by act of taking sample instead of a census.

Random sampling error (unavoidable)

Non-random sampling error

- Convenience sampling
- Volunteer subjects
- Sampling frame systematically different from population (telephone sampling). Literary Digest prediction for 1936 Roosevelt vs Landon presidential election.

The following 15 slides were created by Alan Polansky, Northern Illinois University.

Presidential Election of 1936

- F. D. Roosevelt vs. Alfred Landon
- United States was just making progress with the great depression.
- Roosevelt: Continue to help the unemployed.
- Landon: Balance the budget.
- Many thought Roosevelt would win easily.

Literary Digest Poll

- Literary Digest had called the winner of every presidential election since 1916.
- Polled 2.4 million people – one of the largest polls ever taken.
- Digest Prediction: Landon will win with 57% of the popular vote.
- Election Result: Roosevelt wins with 62% of the popular vote! (Landon 38%)

The Gallup Poll

- At the same time George Gallup was just setting up his survey organization.
- Polled people before the Literary Digest poll was complete.
- Predicted election result and what the outcome of the Literary Digest poll would be.
- Used a sample of 50,000 people.

The Results

	Roosevelt's Percentage
The Election Result	62%
Literary Digest Prediction	43%
Gallup Prediction of Digest Prediction	44%
Gallup Prediction of the Election Result	56%

What Happened?

The Literary Digest poll was based on a much larger sample – but made a much bigger error – How?

Literary Digest Sampling

- A list of 10 million names and addresses of individuals was taken from phone books and club membership lists.
- A survey was sent to each of the 10 million people.
- This was their first mistake: selection bias

Digest Selection Bias

- The poor tended not to have telephones or belong to clubs.
- The digest selection procedure was biased in that it tended to exclude poor voters from the sample.
- The rich and poor voted very differently in 1936.
- The very large sample did not help with this problem.

Selection Bias

- A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way.
- This guarantees that you get a representative cross section of the public.
- A systematic tendency on the part of the sampling procedure to exclude on kind of person or another is called selection bias.
- Selection bias can create major errors in polls.
- When a selection procedure is biased, taking a large sample does not help. This just repeats the mistake on a larger scale.

Literary Digest Sampling (Continued)

- Only 2.4 million people (24%) replied to the Literary Digest poll.
- Non-respondents can be very different from respondents. When there is a high non-response rate, look out for response bias.
- Special sampling procedures are carried out to try to eliminate non-response bias.

The Election of 1948

- Dewey vs. Truman
- Three surveys predicted Dewey was going to win.
- Results:

Candidates	Predictions			Results
	Crossley	Gallup	Roper	
Truman	45	44	38	50
Dewey	50	50	53	45
Thurmond	2	2	5	3
Wallace	3	4	4	2

What Happened?

- All three polls used quota sampling.
- In quota sampling each interviewer was assigned a fixed quota of subjects to interview; the numbers falling into certain categories (residence, sex, age, economic status, etc.)
- Otherwise the interviewer may pick who they wish.

Problems with Quota Sampling

- Generally quotas are based on census data to try to make the sample look like the population.
- It seems reasonable, but it does not work well in practice.
- There is usually unintentional bias on the part of the interviewers.
  - The 1948 poll had too many republicans
  - Republicans were generally easier to get in touch with:
    - Had telephones and permanent addresses.
    - Lived in nicer areas.

Chance Methods in Sampling

- To guarantee a good sample, most surveys now use chance methods to select individuals for polls.
- Simple random sampling:
  - Drawing at random without replacement.
  - Everyone in the population has an equal chance of being sampled.
- Eliminates interviewer and selection bias.

Multistage Cluster Sampling

- Used when a simple random sample is not practical.
- Method:
  - Divide up the population into geographic regions
  - Randomly select some of the regions
  - Divide up the selected regions into sub-regions
  - Randomly select some of the sub-regions
  - Repeat until individuals are selected

How Well Do Chance Methods Work?

Year	Sample Size	Winning Candidate	Gallup Prediction	Election Result	Error
1952	5385	Eisenhower	51.0	55.4	4.4
1956	8144	Eisenhower	59.5	57.8	1.7
1960	8015	Kennedy	51.0	50.1	0.9
1964	6625	Johnson	64.0	61.3	2.7
1968	4414	Nixon	43.0	43.5	0.5
1972	3618	Nixon	62.0	61.8	0.2
1976	3439	Carter	49.5	51.1	1.6
1980	3500	Reagan	51.6	55.3	3.7
1984	3456	Reagan	59.0	59.2	0.2
1988	4089	Bush	56.0	53.9	2.1
1992	2019	Clinton	49.0	43.2	5.8

Estimating a population parameter by sampling can go wrong

Nonsampling errors: Errors that would be present even if we were to take a census.

- Non-response - missing data
- Response errors - subject cannot remember or is lying. (Mention biased coin idea).
- Processing errors
- Effects of data collection method - phone survey vs. face-to-face; race of interviewer.
- Loaded questions - *Do you favor banning private ownership of handguns in order to reduce the rate of violent crime?*

A word of caution

Whenever you see a result of a survey. ask:

- What was the population?
- How was the sample selected?
- What was the sample size?
- How were the subjects contacted? What was the response rate?
- When was the survey conducted?
- What exatly was measured; what questions were asked?