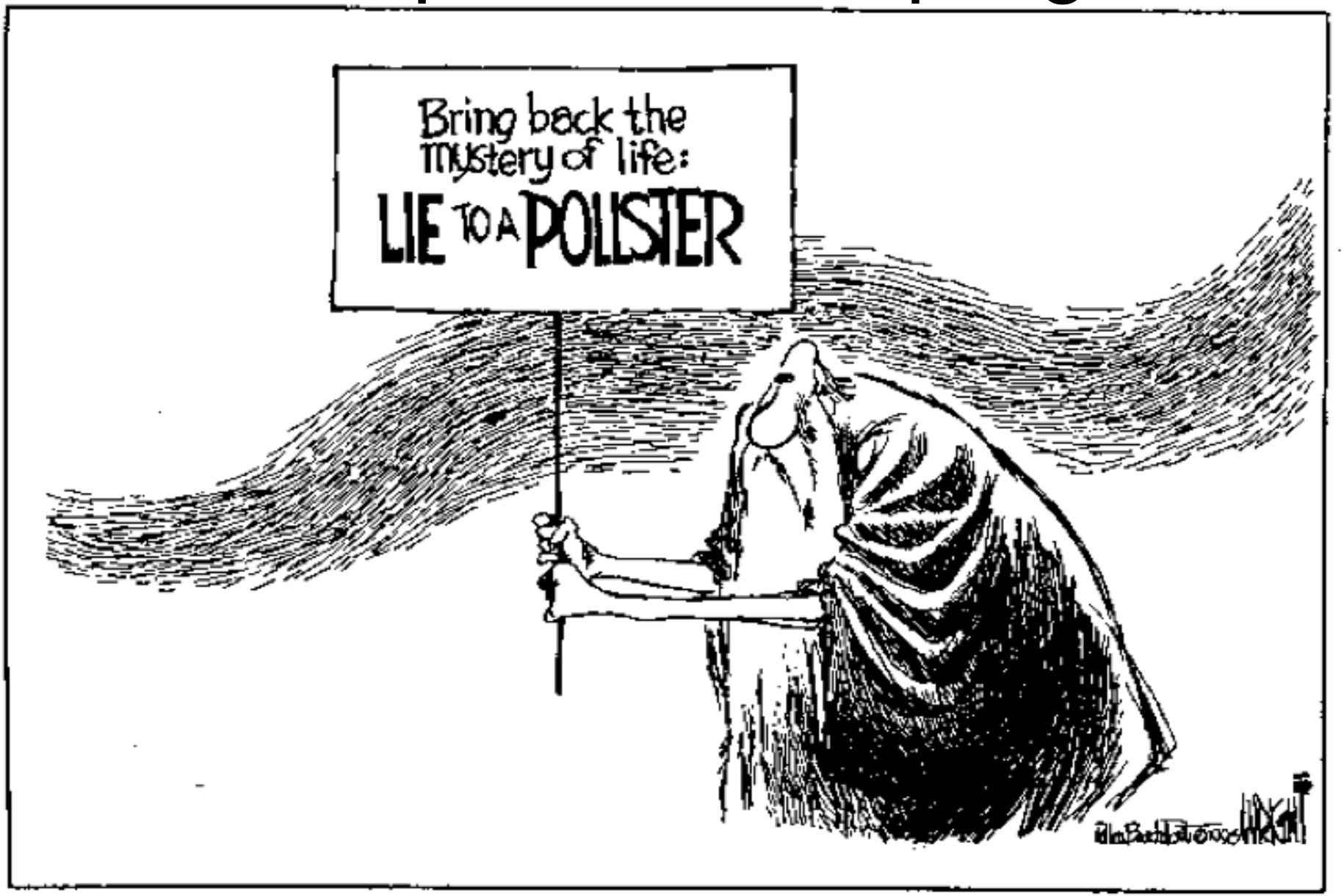
Chapter 19: Sampling



Terminology

- Population: A group of individuals (or objects)
- Sample: Part of a population
- Parameter: Numerical value associated with the population
- Statistic: Numerical value computed using sample
- Inference: Generalizations from sample to population

1936 Presidential Election

Alf Landon Campaign Poster, 1936

Gallup vs Literary Digest

LD vs Gallup

- Sent out 10 million questionnaires (2.4 responses)
- Gallup predicted LD's prediction (3000)
- Gallup's own prediction: FDR with 56%
- LD's prediction: FDR losing with 43%

- What went wrong?

LD debacle

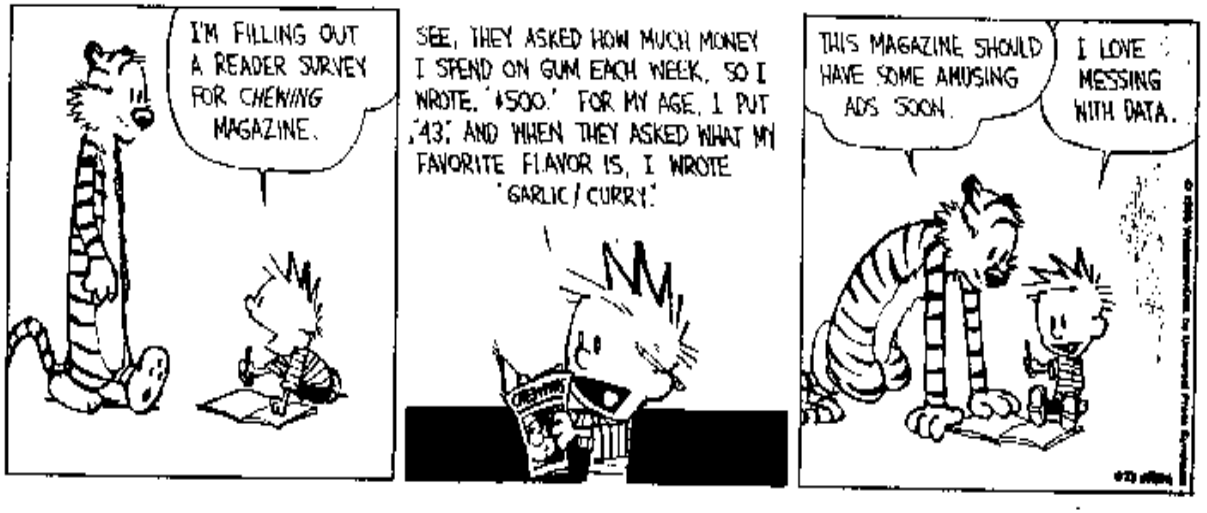
- Source of addresses (“Selection bias”)
- Only 2.4 million (out of 10 million) responded (“Non-response bias”)
- LD went bankrupt soon after.
- Gallup used different methods to get a “random sample”

Why sample?

Why sample?

- Want to know something about the population (who will it elect?) but cannot afford to poll every citizen.
- Quality testing: cannot test every product.
- Market research: will a product succeed?
- Accounting: Expensive to verify a large number of invoices. IRS samples a fraction of tax returns
- Less expensive and time-consuming

Methods of sampling

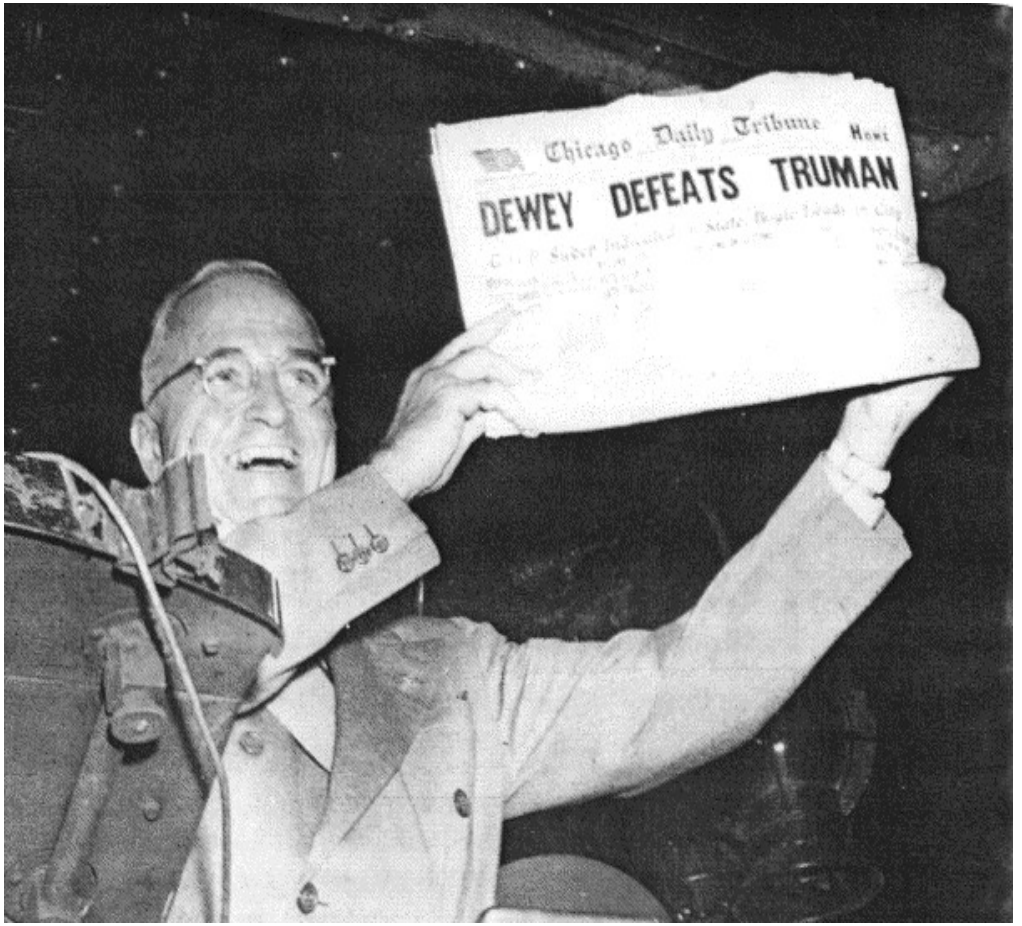


Methods of Sampling

- Quota sampling
- Convenience sampling
- Simple Random Sampling: Draw from the population without replacement.

Quota sampling

- Decide the categories, and their numbers
- Let the interviewers pick whoever they like, within these categories.
- Human choice leads to error



Sources of bias

- Selection bias
- Nonresponse bias (Ex: Ann Landers & parents)
- Timing bias (Ex: NFL poll & football fav. sport)
- Wording bias (Ex: Bill Clinton)

"Now thinking of Bill Clinton as a person do you have a favorable or unfavorable opinion of him?"

40% favorable

"Now I'd like to get your opinion about some people in the news. As I read the name, please say if you have a favorable or unfavorable opinion of this person."

55% favorable

Bias

- Note: bias is **not** prejudice.
- It just means that the error is not due to chance.
- Chance error is called SAMPLING error.
- Bias is NON-SAMPLING error.

Random Sampling

- Simple Random Sampling: Draw from a population without replacement (box model)
- Multistage sampling: Take a SRS of counties, within each take a SRS of townships, and within each of those a SRS of blocks. Sample the entire block.
- Stratified sampling: Divide population into groups (“strata”) and take SRS from each group.
- Systematic sampling: Survey every nth person on a list. How to randomize?

Ex. We want to survey a random sample of 300 passengers on a flight from LA to Mumbai.

1) Randomly generate a list of 30 passengers.

2) Randomly pick 5 first class and 25 economy.

3) Pick one out of first 30 and every tenth after.

4) Randomly select seat position (W, C, A) and then randomly select 30 in those seats.

From xckd.com