

# Sampling

<b>General</b>	<b>2</b>
Population vs sample . . . . .	3
Key ideas . . . . .	4
Population parameter . . . . .	5
Sample and statistic . . . . .	6
<b>US elections</b>	<b>7</b>
The 1936 US election . . . . .	8
The Digest's sampling method . . . . .	9
Problems . . . . .	10
But... . . . . .	11
Gallup . . . . .	12
The 1948 US elections . . . . .	13
<b>Quota sampling</b>	<b>14</b>
Quota sampling . . . . .	15
Example . . . . .	16
Pros and cons. . . . .	17
Back to 1948 election . . . . .	18
Republican bias. . . . .	19
<b>Probability methods</b>	<b>20</b>
Probability methods. . . . .	21
Simple random sampling . . . . .	22
Problem. . . . .	23
Solution: multistage cluster sampling . . . . .	24
Election predictions . . . . .	25
Iraq. . . . .	26
<b>How to evaluate a sample?</b>	<b>27</b>
How to find out if a sample is good? . . . . .	28

**Population vs sample**

- Usually we want to know something about a *population*. Examples
  - ◆ All adolescents with severe acne in the world
  - ◆ All adolescents with acne in the US
- It is often infeasible to look at the entire population.
- Instead, we look at a smaller group, a *sample*.
- The sample should be *representative* of the population, i.e. similar to the population.
- Why? Otherwise the conclusions about the sample do not generalize well to the population.

3 / 28

**Key ideas**

- The method of choosing the sample matters a lot
- The best methods involve the planned use of chance, and leave no room for personal choice
- An ideal sampling method:
  - ◆ Identify population
  - ◆ List all individuals in the population
  - ◆ Draw random sample with a probability method
  - ◆ The results of the sample are generalizable to the population
- Not always possible in praxis...

4 / 28

**Population parameter**

- We want to know a *parameter* (=numerical fact) about a *population*.
- Example: What is the percentage of Republican voters in the next presidential election?
  - ◆ population: voters in the next presidential election
  - ◆ parameter: percentage of Republican voters
- Example: What is the number of deaths in Iraq due to the war?
  - ◆ population: people of Iraq
  - ◆ parameter: the number of deaths due to the war

5 / 28

## Sample and statistic

- It is infeasible to look at the entire population. *Hence, we'll never know the population parameter exactly!*
- We only examine part of the population, a *sample*.
- We compute a *statistic* from the sample:
  - ◆ Percentage of Republican voters in the sample
  - ◆ Number of deaths due to the war in sampled families
- The *statistic* is used to *estimate* the *population parameter*. Statistics are what we know (for a given sample), parameters are what we want to know.
- Statistics are random: when we take a different sample, we get a different value.
- We then make *inference*: we generalize the results of the sample to the population.
- Good inference is only possible if the sample resembles the population. We need a good sample!

6 / 28

## US elections

7 / 28

### The 1936 US election

- Roosevelt (Democrat)  $\Leftrightarrow$  Landon (Republican)
- The Digest's prediction:  
Landon wins, and Roosevelt gets 43% of the votes.
- Election result: Roosevelt wins, with 62% of the votes.
- The Digest predicted the wrong winner, and the predicted percentage was off by almost 20 percentage points!
- What went wrong?

8 / 28

### The Digest's sampling method

- The Digest made a list of people by combining phone books and lists of club membership.
- They randomly picked 10 million people from this list, and mailed them a questionnaire.
- 2.4 million people returned the questionnaire.
- 43% of these people planned to vote for Roosevelt.
- That was their prediction.

9 / 28

## Problems

- Two main problems with the Digest sampling methods:
  - ◆ *Selection bias*: Only 1 in 4 households had a phone at the time. The Digest's list of eligible voters tended to screen out the poor, because they had no phone or club membership.
  - ◆ *Non-response bias*: Only about 1 in 4 households that received a questionnaire returned it. The people who did not respond may have different voting habits than the people who responded.
- So... the 2.4 million people who returned the questionnaire did not represent the 10 million people who were polled, let alone the population of US voters. The sample was *biased*.

10 / 28

## But...

- We had a big sample: 2.4 million... Doesn't that fix the problems?
  - ◆ No! When a procedure is biased, taking a larger sample does not help. This just repeats the basic mistake on a larger scale.
- Why did we first see this problem in the 1936 elections?
  - ◆ Before 1936 the rich and poor tended to vote similarly. In 1936, the poor voted overwhelmingly for Roosevelt, and the rich for Landon.

11 / 28

## Gallup

- In the same year, Gallup predicted:
  - ◆ the Digest's prediction: 44% (truth: 43%)
  - ◆ election results: Roosevelt wins with 56% (truth: 62%)
- How did he predict the Digest prediction so well?
  - ◆ He took a random sample of 3000 people (much smaller than 10 million!) from the same list that the Digest used, and mailed those a questionnaire
- How did he predict the election results?
  - ◆ He used a method called *quota sampling*. His sample size was 50,000.
  - ◆ His method worked better than the Digest's method. He predicted the correct winner, but was still off by 6 percentage points.

12 / 28

## The 1948 US elections

- Truman (Democrat) ⇔ Dewey (Republican)
- Three major polls predicted Dewey as winner:
  - ◆ Crossley: Dewey 50% and Truman 45%
  - ◆ Gallup: Dewey 50% and Truman 44%
  - ◆ Roper: Dewey 53% and Truman 38%
- Election results:
  - ◆ Dewey 45% and Truman 50%
  - ◆ Truman won against the prediction of all three polls!
- What went wrong with the polls?
  - ◆ All polls used quota sampling. This did not give good samples.

13 / 28

## Quota sampling

14 / 28

### Quota sampling

- Each interviewer is assigned a fixed number of people to interview, with also the numbers falling into certain categories (sex, age, residence, economic status, etc) fixed.
- The interviewers are free to interview anybody they like as long as they keep to these quotas.

15 / 28

### Example

- A Gallup poll interviewer in St. Louis had to interview 13 people, of whom:
  - ◆ 6 were to live in the suburbs, and 7 in the central city
  - ◆ 7 were to be men, and 6 women
  - ◆ Of the 7 men (with similar quotas for women):
    - 3 were to be under 40 years old, and 4 over 40
    - 1 was to be black, and 6 white
    - Monthly rentals for the 6 white men:
      - ◆ 1 was to pay \$44.01 or more
      - ◆ 3 were to pay \$18.01 to \$44.00
      - ◆ 2 were to pay \$18.00 or less

16 / 28

## Pros and cons

- The quotas ensure that the sample looks like the population w.r.t. some key characteristics.
- This is why Gallup had a better prediction for the 1936 elections than the Digest. (Recall that the Digest's sample contained mostly rich people.)
- But... there are problems with quota sampling:
  - ◆ We may forget to set quotas for some important characteristics.
  - ◆ The method still leaves a lot of freedom to the interviewers to pick people. This can lead to unintentional selection bias.

17 / 28

## Back to 1948 election

- Why did the polls predict the wrong winner?
  - ◆ Republicans were wealthier than Democrats.
  - ◆ Hence, Republicans were more likely to have phones, nicer houses, permanent addresses.
  - ◆ Within each demographic group, the Republicans were a bit easier to interview.
  - ◆ That's why the samples included too many Republicans, and predicted the Republican candidate to win.

18 / 28

## Republican bias

Year	Gallup's prediction of Republican vote	Actual Republican vote	Error in favor of the Republicans
1936	44	38	6
1940	48	45	3
1944	48	46	2
1948	50	45	5

19 / 28

## Probability methods

- A probability method has the following two properties:
  - ◆ it incorporates planned use of chance
  - ◆ it leaves no room for personal choice of the investigators/interviewers
- Examples:
  - ◆ simple random sampling
  - ◆ (multistage) cluster sampling
  - ◆ stratified sampling (won't go into this)

21 / 28

## Simple random sampling

- Method for simple random sampling:
  - ◆ Write the name of each person in the population on a ticket
  - ◆ Put all the tickets in a box
  - ◆ Shake the box and draw a ticket
  - ◆ Shake the box again, and draw another ticket
  - ◆ Continue until we have the sample size we want
- This is called simple random sampling: drawing at random without replacement.
- Nowadays, it is usually done with computers.
- Each person has the same chance to get into the sample. There is no selection bias. Hence, the sample is likely to be a good representation of the population.

22 / 28

## Problem

- Sometimes simple random sampling is not possible:
  - ◆ Elections:
    - In 1930s there was no list of all eligible voters, there were no computers to draw a random sample from these voters.
    - Moreover, there were many people without phones. Sending interviewers to people all over the US would be very expensive.
  - ◆ Estimating casualties in Iraq:
    - Investigators wanted to interview a sample of people and ask them about family members who died.
    - There is no accurate list of all people in Iraq.
    - Interviewing people all over the country would involve a lot of travel, and that is dangerous.

23 / 28

### Solution: multistage cluster sampling

- Divide population into clusters (usually geographically)
- Randomly select a number of clusters
- For each cluster, interview a random sample of people in the cluster
- Applications:
  - ◆ Election polls
  - ◆ Estimating casualties in Iraq
- Advantage: interviewers only have to be stationed in the selected clusters.

24 / 28

### Election predictions

year	sample size	winner	Gallup prediction	election result	error
1964	6600	Johnson	64%	61%	3%
1968	4400	Nixon	43%	44%	1%
1972	3700	Nixon	62%	62%	0%
1976	3500	Carter	50%	51%	1%
1980	3500	Reagan	52%	55%	3%
1984	3500	Reagan	59%	59%	0%
1988	4000	Bush	56%	54%	2%
1992	2000	Clinton	49%	43%	6%
1996	2900	Clinton	52%	49%	3%
2000	3600	Bush	48%	48%	0%
2004	2000	Bush	49%	51%	2%

- The methods work very well, and with small sample size!
- Note the larger error in 1992. This had to do with undecided voters.

25 / 28

### Iraq

- Article in the Lancet (see website, under the tab 'Links')
- Cluster sampling, 33 clusters in each of which they interview 30 households
- Estimated 100,000 more deaths than expected in the first 1.5 years of the war (confidence interval 8000 - 194000)

26 / 28

### How to find out if a sample is good?

- You can often not see this by looking at the data
- So you should find out how the data were gathered!
- Questions to ask:
  - ◆ What is the population?
  - ◆ What is the parameter?
  - ◆ How was the sample chosen? Was there room for personal choice? Did it involve the planned use of chance?
  - ◆ What was the response rate?
  - ◆ How were the questions phrased?
- Be aware that medical studies often use *convenience samples*, for example all patients of a certain hospital/doctor.