Samples, Good and Bad

1 Sample

- In the last lecture we discussed about the differences between population and samples, and different sampling methods.
- In this lecture, we will look what makes a sample a good sample or a bad sample.

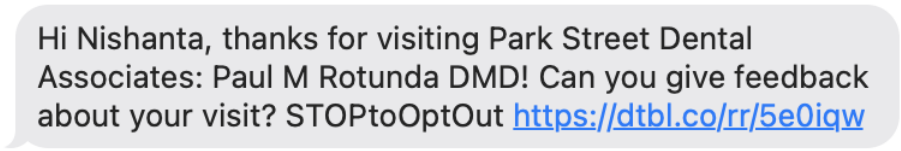
2 How to sample badly?

- The design of a statistical study is biased if it systematically favors certain outcomes. Biases creep in mostly in a voluntary response sample and convenience sampling.
- A voluntary response sample chooses itself by responding to a general appeal. Write-in or call-in opinion polls are examples of voluntary response samples.

  Example:Ann Landers was a popular advice columnist, who wrote a daily column from 1955–2002. She once asked the readers of her advice column, "If you had it to do over again, would you have children?" She received nearly $10,000$ responses, almost 70% saying, "NO!" Can it be true that 70% of parents regret having children? Not at all. This is a voluntary response sample. People who feel strongly about an issue, particularly people with strong negative feelings, are more likely to take the trouble to respond.
- Selection of whichever individuals are easiest to reach is called convenience sampling. The sample is composed of individuals who are easily accessible to the researcher and are willing to be a part of the study.

  Example: Suppose a researcher wants to conduct a survey on the opinions of students regarding the sleep habits of college students. They choose to distribute the survey to students in the cafeteria during lunchtime who are present in the cafeteria at that time and willing to participate. This is an example of convenience sampling as the researcher selects the sample based on convenience and accessibility, as students are readily available during lunchtime.
- Any sampling method that allows some bias is not an ideal sampling method. Biased sampling methods produce data that can be misleading, resulting in incorrect conclusions.
- A sample chosen by chance allows neither favoritism by the sampler nor self-selection by respondents. Choosing a sample by chance attacks bias by giving all individuals an equal chance to be chosen. The simplest way to use chance to select a sample is to place names in a hat (the population) and draw out a handful (the sample). This is the idea of simple random sampling.

Now it's your turn: I recently visited my dentist for usual check-up. After every visit, the dental office sends a text message as attached below. On a scale of 1 to 10, 1 being the lowest and 10 being the highest, how much would you trust the customer satisfaction survey collected by the dental office? Explain your reasoning.



This is a classic example of voluntary response sample. The dental office is collecting the customer satisfaction survey by relying on the voluntary response of the patients. The patients who feel the dentist did a great job, or a terrible job are likely to respond to the survey. Also, the response rate needs to be taken into account to examine non-response bias. I would rate it a 4 out of 10.

Call-in versus random sample polls. A national survey of TV network news viewers found that 48% said they would believe a phone-in poll of 300,000 persons rather than a random sample of 1000 persons. Of the viewers, 42% said they would believe the random sample poll. Explain to someone who knows no statistics why the opinions of only 1000 randomly chosen respondents are a better guide to what all people think than the opinions of 300,000 callers.

Call-in polls and voluntary response polls in general, tend to attract responses from those who have strong opinions on the subject, and therefore, they are often not representative of the population as a whole. A random sample of size 1000 will ideally be representative of the population as a whole. While the 300,000 callers might be an impressive number, the voluntary response of these callers is not trustworthy.

3 Two case studies

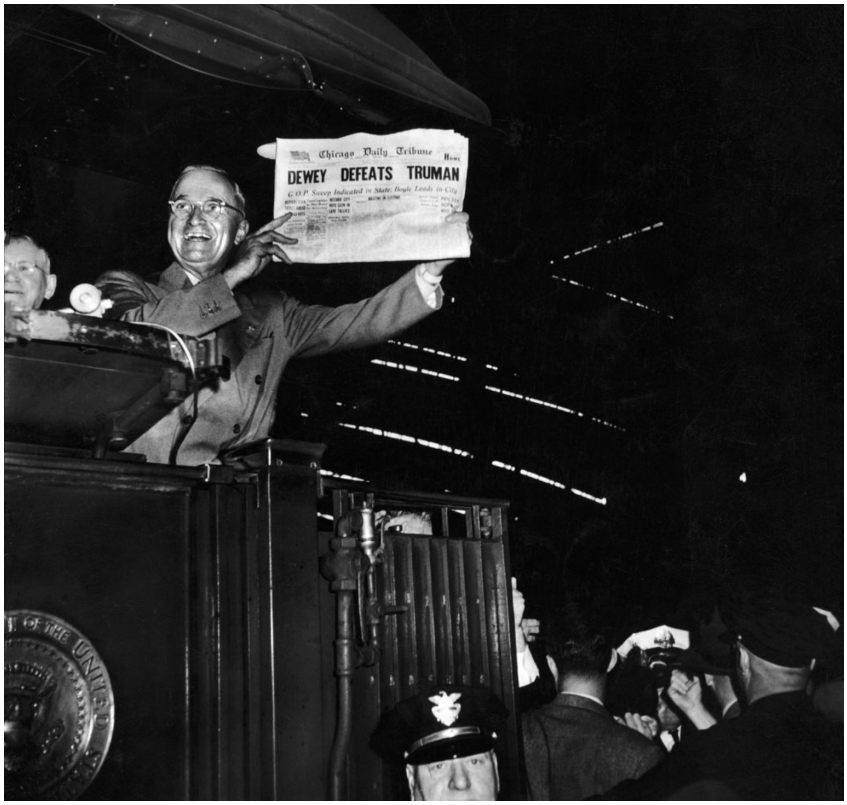
- "Dewey beats Truman"

  The presidential election in 1948 is generally regarded as the greatest upset in American political history. The two candidates were incumbent president Harry Truman and the heavily-favored Republican governor of New York, Thomas Dewey. Dewey was widely tipped by pollsters to win the election easily.

  So the night of November 2, the editors of the Chicago Tribune, based on the advice of their Washington editor and a need for an early first edition, approved a gigantic headline that read, "Dewey Beats Truman."

  But when the dust settled, Truman easily defeated Dewey in the Electoral College, by a 303 to 189 margin. Truman also won by a 50-45 percent popular vote.

  Two days after the election, the president was on his way back to Washington when his train stopped in St. Louis. There, someone handed Truman a two-day-old copy of the Tribune and he posed for a picture with the Chicago Tribune paper. This picture later became one of the most iconic pictures in US political history, and is a classic case study in statistics.



The reason behind this humongous mistake was because the polling companies stopped a few weeks too soon. They had been lulled into thinking that nothing much changes in the last few weeks of the campaign. The polling pioneers admitted their mistakes, reexamined their methods and plunged back to work. They moved gradually away from quota sampling, which questioned a set number of people from different ethnic and age groups, and moved toward random sampling. They extended polling deadlines up until election day and developed their ability to predict those likely to come out and vote. And a big lesson for pollsters from 1948 still holds true today.

- 1936 presidential election

  In the presidential election held in the United States in 1936, the candidates were the incumbent President Franklin Roosevelt (Democrat) and Alf Landon (Republican). *The Literary Digest* was a popular and widely read weekly magazine that ran a poll to predict the winner of the presidential race, and had done so correctly from 1920 to 1932.

  In 1936, *The Literary Digest* mailed a questionnaire to 10 million people to ask about their voting intentions. This extraordinary number of people included readers of *The Literary Digest*, registered car owners and people listed in the phone book. In one of the largest surveys ever, 2.4 million voters replied. The response rate was 24%. The prediction that *The Literary Digest* made based on their survey was that Franklin Roosevelt would receive only 43% of the vote; Landon was predicted to win in a landslide.

  As you may know, Roosevelt won — he obtained 62% of the vote of around 40 million voters.The failure of*The Literary Digest*'s poll was an embarrassment, and *The Literary Digest* subsequently went out of business; eventually its subscriber list was bought by Time magazine.

  Historically, there are two reasons cited for this erroneous prediction by *The Literary Digest*. (1) Because the Digest selected its sample primarily from telephone books and

car registration lists and since these contained, at the time, mostly well-to-do folks who would vote Republican, it is no wonder the magazine mistakenly predicted a Republican win. (2) When there is a poor response rate, there is a potential for non-response bias. It was those who failed to participate in the poll (overwhelmingly supporters of Roosevelt) who were mainly responsible for the faulty prediction. If a large proportion of a sample fails to respond, having a large sample will not help: the results should be regarded as unreliable.