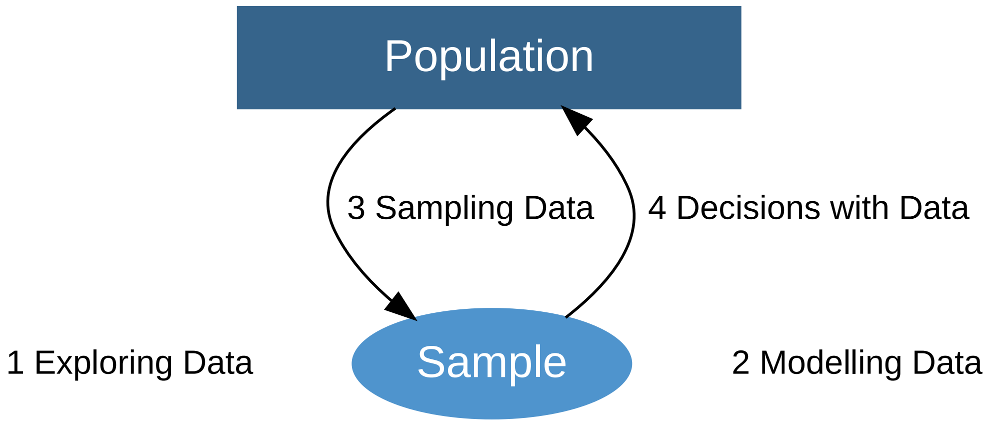
Sample Surveys

Sampling Data | Sample Surveys

Unit Overview



Module3 Sampling Data

Understanding Chance

What is chance?

Chance Variability

How can we model chance variability by a box model?

Sample Surveys

How can we model the chance variability in sample surveys?

Sample Surveys

Data Story | Why didn’t we predict Trump would win?

Population vs Sample

Selecting a sample

Why many 2016 US election polls failed

Summary

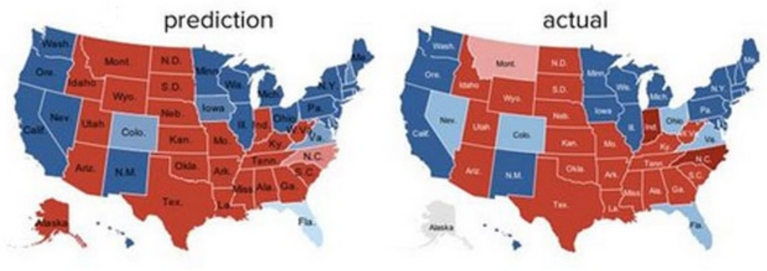
Data Story

Why didn’t we predict Trump would win?

Nate Silver and FiveThirtyEight.com

- In March 2008, Nate Silver established FiveThirtyEight.com, where he developed a system for tracking polls and forecasting the outcome of the 2008 US general election.

- Silver's final 2008 presidential election forecast accurately predicted the winner of 49/50 states. 🔗NYTimes



- They also used their methodology to predict NBA and NFL results, and ecomonic outcomes. They maintained high predictive success winning awards for “data journalism”. 🔗Sports
- It was widely anticipated FiveThirtyEight.com would perform well at the 2016 election where they predicted that Hillary Clinton had as low as a 64.5% chance of winning the election.
- However….
- What went wrong?

Statistical Thinking

With the person next to you discuss:

- Did you think Clinton or Trump would win? Why or why not?
- What could have gone wrong with Silver’s prediction?

Population vs Sample

Populations & Samples

Population & Sample

- The **population** is the full amount of information being studied, collected though a **census**.
- A **sample** is part of the population.

Why we do need samples? What are the issues concerned with analysing samples?

Limitations of a census

Collecting **every** unit of a population

- is hard
- take lots of time
- costs a lot of money
- requires lots of resources.

Parameter & Estimate

Parameter & Estimate

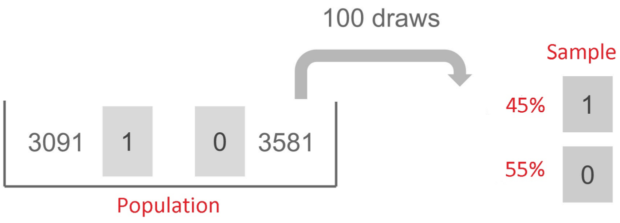
- A **parameter** is a numerical fact about the population which we are interested in. For example the population mean μ.
- An **estimate** (or **statistic**) is a calculation of sample values which best predicts the parameter. For example the sample mean $\bar{x}$.

“The estimate is what the investigator knows. The parameter is what the investigator wants to know.”

In the 2016 US elections, what was the parameter of interest? How was it estimated?

Parameter & Estimate

- Suppose we have a population of 6672 units in the population.
- Suppose 3091 would “vote” 1 and 3581 would “vote” 0.
- Due to constraints, we can only sample 100 people from this finite population and find 45 people vote 1 and 55 vote 0.
- Note: we normally sample without replacement, so we don’t interview the same person twice.



- We can estimate the proportion who vote 1 in the population using the sample.

Finding the best estimate of the parameter

- Much Statistical Theory is concerned with how to find the best estimate of a parameter.
- 2 critical issues are:
  - How was the sample chosen? Is it representative of the population?
  - What estimate is closest to the parameter?

Selecting a sample

4 Common Types of Bias

1. **Selection bias** A systematic tendency to exclude or include one type of person from the sample.

2. **Non-response bias** Caused by participants who fail to complete surveys.
   - What was the response rate?
   - Non-respondents can be very different to respondents.

3. **Interviewer’s bias** When the interviewer has to make a choice of participants in the survey, or when characteristics of the interviewer have an effect on the answer given by participants.

4. **Measurement bias** When the form of the question in the survey affects the response to the question.

Statistical Thinking

With the person next to you propose 4 possible sources of bias in the US Presidential election.

- Selection
- Non-response
- Interviewer
- Measurement

Statistical Thinking

Suggest how the following could occur:

- Selection bias in a phone survey.
- Non-response bias in the Australian postal poll on same sex marriage.
- Interviewer bias, when an employer in a polling company is given a quota.

Examples of Measurement Bias

1. Bias in question wording and order, which impacts responses

Statistical Thinking

How is this wording biased? Suggest a rewording.

- Should a doctor be allowed to murder unborn children who can’t defend themselves?
- How dumb is Trump when it comes to foreign policy!
- Should concerned dog owners vaccinate their pets?

Statistical Thinking

Suggestions:

- Should a doctor be allowed to do abortions?
- What is your view of Donald Trump’s foreign policy?
- Do you think dogs should be required to be vaccinated?

2. Recall bias: People forget details.

3. Sensitive questions: People may not tell the truth.

- Do you use illegal drugs?
- How often do you bathe?
- Do you pay tax?
- What is your age?

4. Lack of clarity in question

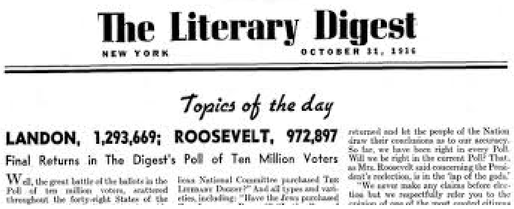
- People may misinterpret the question.
- Certain words may be ambiguous and so mean different things to different people.

5. Attributes of the interview process may cause bias.

Warning about bias and sample size

- When a section process is biased, taking a **larger sample** does not reduce bias, rather it can **amplify** the bias. It repeats the mistake on a larger scale!

- In the famous 1936 US elections, the *Literary Digest* magazine predicted an overwhelming victory for Alfred Landon over Franklin Roosevelt, based on a poll of 2.4 million people! However, Roosevelt won 62% to 38%!! The Digest went bankrupt soon after.
- The problem was that their sampling procedure involved mailing questionnaries to 10 million people, with names and addresses from sources which was biased against the poor.



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

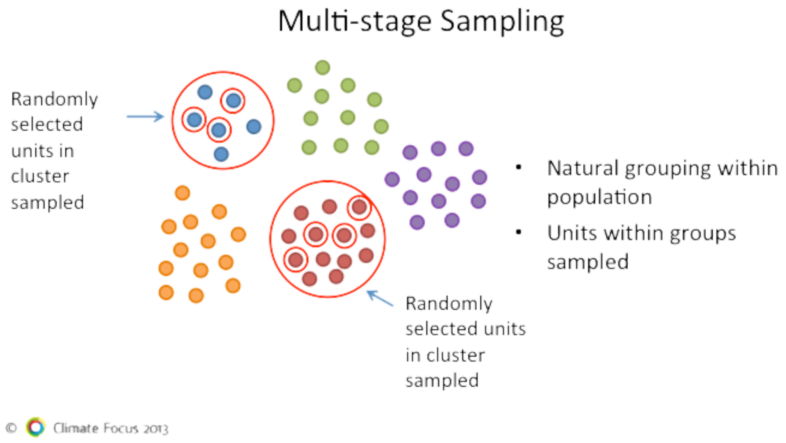
"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens

How to pick a good sample?

- A sampling procedure should be fair - ie it gives a representative cross section of the population.
- We use a **probability method** to pick the sample, so that
  - the interviewer is not involved in the selection. The method of selection is impartial.
  - the interviewer can compute the chance of any particular individuals being chosen. ie There is a defined procedure for selecting the sample, which uses chance. It is objective.
- For example, **Simple random sampling** involves drawing at random without replacement.

1. Multi-stage cluster sampling

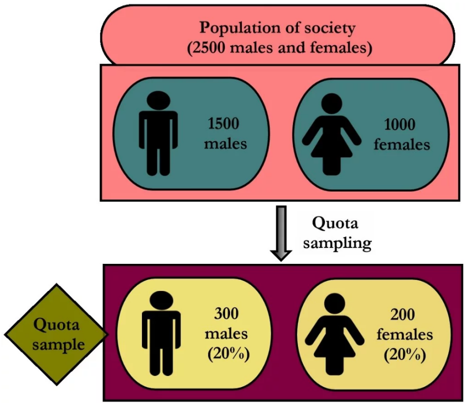
As simple random sampling is often not practical, organisations may use multi-stage cluster sampling. This is a probability sampling technique which takes samples in stages, and individuals or clusters are chosen at random at each stage.


Multi-stage Sampling
Randomly selected units in cluster sampled
Natural grouping within population
Units within groups sampled
Randomly selected units in cluster sampled
© Climate Focus 2013

Copyright for image

2. Quota sampling

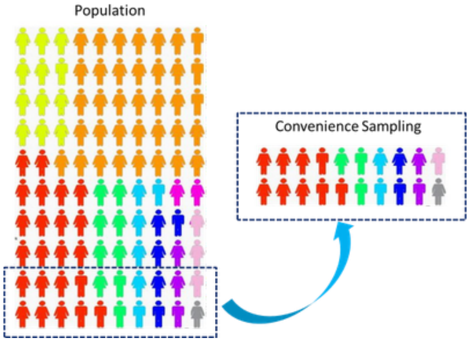
A non-probability sampling technique where the assembled sample has the same proportions of individuals as the entire population with respect to known characteristics, traits or focused phenomenon. This results in unintentional bias from the interviewers when they choose subjects to survey.



Copyright for image

3. Convenience sampling (or “grab sampling”)

A non-probability sampling technique where subjects are selected because of their convenient accessibility. It is definitely not recommended, except possibly to test a survey (pilot).



Copyright for image

Unavoidable Bias

- Even with a probability method determining the sample, bias (eg non-response bias) can easily come in.
- In addition, because the sample is only part of the population, we always have chance error.

Samping & Non-Sampling Error

Estimate = Parameter + Bias + Chance Error
OR:
Estimate = Parameter + Non-sampling Error + Sampling Error

Common methods of surveys

- Face-Face Interviews
- Phone Interviews
- Self-administered Surveys
- Mail
- Online

Why many 2016 US election polls failed

A surprising failure?

- The fact that so many forecasts were off-target was particularly notable given the increasingly wide variety of methodologies being tested and reported via the mainstream media and other channels.
- The traditional telephone polls of recent decades are now joined by increasing numbers of high profile, online probability and non-probability sample surveys, as well as prediction markets, all of which showed similar errors.

However, we note 4 things.

1. The national polls weren't that off

They did predict more people would vote for Clinton. And that's what happened.

2. Some people just don’t answer the phone.

- Many pollsters who do phone polling conduct it via random digit dialing.
- One likely problem is **non-response bias**. This occurs when certain kinds of people systematically do not respond to surveys despite equal opportunity outreach to all parts of the electorate.
- Some groups – including the less educated voters who were a key demographic for Trump on Election Day – are consistently hard for pollsters to reach. It is possible that the frustration and anti-institutional feelings that drove the Trump campaign may also have aligned with an unwillingness to respond to polls. The result would be a strongly pro-Trump segment of the population that simply did not show up in the polls in proportion to their actual share of the population.

3. Some people may lie to pollsters

- Many of those who were polled simply were not honest about whom they intended to vote for, given the highly controversial campaign trail.
- The idea of so-called ``secret Trumpers’’ suggests that support for Trump was socially undesirable, and that his supporters were unwilling to admit their support to pollsters.
- The ``secret Trumper’’ hypothesis has received a fair amount of attention this year. If this were the case, we would expect to see Trump perform systematically better in online surveys, as research has found that people are less likely to report socially undesirable behavior when they are talking to a live interviewer. - This is **measurement bias** which can be improved by changing the mode of data collection when dealing with sensitive issues.

4. It’s hard to capture the real voter

- Because we can’t know in advance who is actually going to vote, pollsters develop models predicting who is going to vote and what the electorate will look like on Election Day.
- This is a notoriously difficult task, and small differences in assumptions can produce sizable differences in election predictions.
- What may have happened is that the usual models of predicting simply didn’t work that year as lots of other things about the election were unusual: high levels of anger and two candidates with high unfavorability ratings. This is an inherent problem in opinion surveys, and cannot be improved by the design of the opinion poll.

Read More …

Summary

Unless a census is possible, information about a population comes from an estimate from a sample. The reliability of such an estimate depends on how the sample was chosen, including consideration of bias and chance error.

Key Words

population, sample, parameter, estimate, representative, selection bias, non-response bias, quota sampling, interviewer's bias, sample random sampling, non-sampling error, sampling error, self-administered

Extension: Non-response bias

- Suppose that we have a survey where we wish to estimate a parameter, $p \in [0,1]$, the proportion to answer “yes” to a question.

- Next suppose that the survey is voluntary where $x \in [0,1]$ and $y \in [0,1]$ are the proportions of survey responders who would vote “yes” and “no” respectively.

- What is the probability of responding “yes” given a response? The calculation required is

$$\frac{p \times x}{p \times x + (1-p) \times y}.$$

- How do we get there? Conditional probability (Topic 19)!

- Using the conditional probability formula (Bayes theorem):

$$P(\text{yes}|\text{respose}) = \frac{P(\text{yes and respond})}{P(\text{respond})}$$

- The numerator can be expanded using the multiplication principal:

$$P(\text{yes and respond}) = P(\text{respond}|\text{yes})P(\text{yes}).$$

- The denominator can be calculated using the law of total probability (assuming one can only vote “yes” or “no”)

$$P(\text{respond}) = P(\text{yes and respond}) + P(\text{no and respond})$$

- The terms on the right hand side can themselves be expanded using the multiplication principal:

$$P(\text{respond}) = P(\text{respond}|\text{yes})P(\text{yes}) + P(\text{respond}|\text{no})P(\text{no})$$

- Putting these together

$$\begin{aligned} P(\text{yes}|\text{respose}) &= \frac{P(\text{respond}|\text{yes})P(\text{yes})}{P(\text{respond}|\text{yes})P(\text{yes}) + P(\text{respond}|\text{no})P(\text{no})} \\ &= \frac{x \times p}{x \times p + y \times (1-p)} \end{aligned}$$

- Notice that if $x = y$ then $P(\text{yes}|\text{respose}) = p$. This means if the non-reponse rate is the same for “yes” and “no” voters then the bias disappears!