Sampling

Communication Research Methods

Jennifer Pan

Assistant Professor
Department of Communication
Stanford University

January 29, 2016

Announcements

Announcements

- No laptop in lecture UNLESS I'm not working in R

Announcements

- No laptop in lecture UNLESS I'm not working in R
- Midterm Feb. 5 (traveling - Megan by Feb 3 before class)

Announcements

- No laptop in lecture UNLESS I'm not working in R
- Midterm Feb. 5 (traveling - Megan by Feb 3 before class)
- Midterm content: today's lecture (Monday Feb. 1 section)

Announcements

- No laptop in lecture UNLESS I'm not working in R
- Midterm Feb. 5 (traveling - Megan by Feb 3 before class)
- Midterm content: today's lecture (Monday Feb. 1 section)
- Midterm review Feb 4, Feb 5 section (no section Feb 8)

Overview

Overview

- Up to now

Overview

- Up to now
    - What is scientific research: theories, concepts, measurements

Overview

- Up to now
  - What is scientific research: theories, concepts, measurements
  - Types of question we want to answer: causation

Overview

- Up to now
  - What is scientific research: theories, concepts, measurements
  - Types of question we want to answer: causation
  - Summarizing data: descriptive statistics

Overview

- Up to now
  - What is scientific research: theories, concepts, measurements
  - Types of question we want to answer: causation
  - Summarizing data: descriptive statistics
  - Basics of R: tool for working with quantitative data

Overview

- Up to now
  - What is scientific research: theories, concepts, measurements
  - Types of question we want to answer: causation
  - Summarizing data: descriptive statistics
  - Basics of R: tool for working with quantitative data
- Today

Overview

- Up to now
  - What is scientific research: theories, concepts, measurements
  - Types of question we want to answer: causation
  - Summarizing data: descriptive statistics
  - Basics of R: tool for working with quantitative data
- Today
  - Getting representative data: random sampling and pitfalls

Terminology

Terminology

- Population

Terminology

- Population
- Sample

Terminology

- Population
- Sample
- Random Sampling

Population

Population: the universe of cases we want to describe

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"
  - Population: all males (of all ages, in the world)

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"
  - Population: all males (of all ages, in the world)
- "What is the average price increase for goods and services?"

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"
  - Population: all males (of all ages, in the world)
- "What is the average price increase for goods and services?"
  - Population: every good and service sold, every store

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"
  - Population: all males (of all ages, in the world)
- "What is the average price increase for goods and services?"
  - Population: every good and service sold, every store
- "How will American vote in the 2016 presidential election?

Population

Population: the universe of cases we want to describe

- "What is the average income of adults in the US?"
  - Population: every person in the US, over 18 years of age
- "Does poor diet increase the risk of prostate cancer?"
  - Population: all males (of all ages, in the world)
- "What is the average price increase for goods and services?"
  - Population: every good and service sold, every store
- "How will American vote in the 2016 presidential election?
  - Population: all US citizens...who intend to vote? who are eligible to vote?

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"
  - Population parameter: risk

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"
  - Population parameter: risk
- "What is the average price increase for goods and services?"

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"
  - Population parameter: risk
- "What is the average price increase for goods and services?"
  - Population parameter: average price

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"
  - Population parameter: risk
- "What is the average price increase for goods and services?"
  - Population parameter: average price
- "How will American vote in the 2016 presidential election?

Population

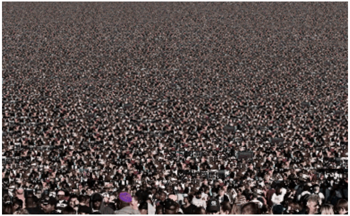
Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- "What is the average income of adults in the US?"
  - Population parameter: average income
- "Does poor diet increase the risk of prostate cancer?"
  - Population parameter: risk
- "What is the average price increase for goods and services?"
  - Population parameter: average price
- "How will American vote in the 2016 presidential election?
  - Population parameter: vote intention

Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

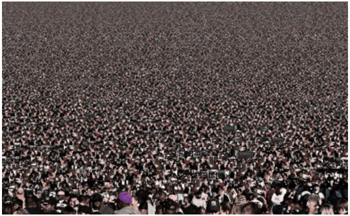
- Census: when we have every single observation from the population



Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

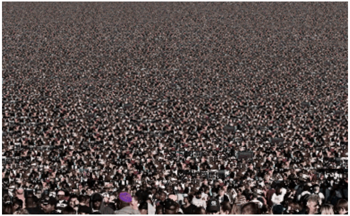
- Census: when we have every single observation from the population
- Rare, usually:



Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

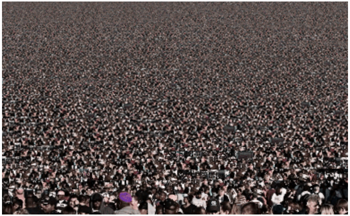
- Census: when we have every single observation from the population
- Rare, usually:
  - Too expensive (2010 census cost $13 billion)



Population

Population: the universe of cases we want to describe Population parameter: the characteristic of the population we care about

- Census: when we have every single observation from the population
- Rare, usually:
  - Too expensive (2010 census cost $13 billion)
  - Too time consuming



Sample

Population: the universe of cases we want to describe

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
    - A sample of US adults

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters
- Calculate a sample statistic

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters
- Calculate a sample statistic
  - Average income of a sample of US adults

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters
- Calculate a sample statistic
  - Average income of a sample of US adults
  - Average risk of prostate cancer of the sample of men

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters
- Calculate a sample statistic
  - Average income of a sample of US adults
  - Average risk of prostate cancer of the sample of men
  - Average price of goods and services in the sample of stores

Sample

Population: the universe of cases we want to describe

- Take a sample of n cases from the population
  - A sample of US adults
  - A sample of men
  - A sample of stores selling goods and services
  - A sample of eligible voters
- Calculate a sample statistic
  - Average income of a sample of US adults
  - Average risk of prostate cancer of the sample of men
  - Average price of goods and services in the sample of stores
  - Vote intention of sample of eligible voters

Sample vs. Population

- Population: the universe of cases we want to describe

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

- Examples:

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

- Examples:
  - We don't care about how individuals in Gallup poll (sample) are going to vote: do care about what that tells us about US voters in general (population)

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

- Examples:
  - We don't care about how individuals in Gallup poll (sample) are going to vote: do care about what that tells us about US voters in general (population)
  - Don't care about whether price of a pizza from Avanti (sample) has increased: do care about what this tells us about inflation in general (population

Sample vs. Population

- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

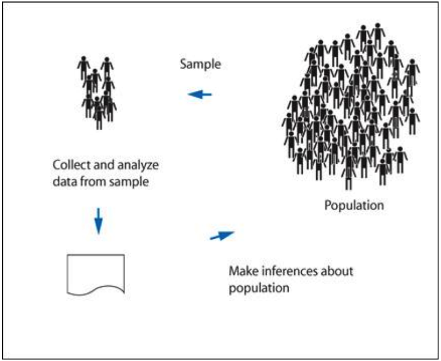
- Examples:
  - We don't care about how individuals in Gallup poll (sample) are going to vote: do care about what that tells us about US voters in general (population)
  - Don't care about whether price of a pizza from Avanti (sample) has increased: do care about what this tells us about inflation in general (population
- Saying something about the population from the sample is called statistical inference

Sample vs. Population

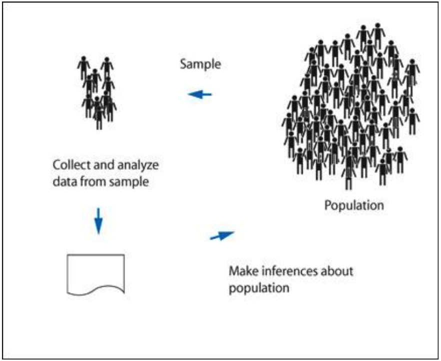
- Population: the universe of cases we want to describe
- Population parameter: the characteristic of the population we care about
- Sample: n cases from the population
- Sample statistic: what we use to estimate the population parameter

- Examples:
  - We don't care about how individuals in Gallup poll (sample) are going to vote: do care about what that tells us about US voters in general (population)
  - Don't care about whether price of a pizza from Avanti (sample) has increased: do care about what this tells us about inflation in general (population
- Saying something about the population from the sample is called statistical inference
- Recall: Inference is a part of what makes research scientific: infer something about the world beyond what we observe

Sample vs. Population

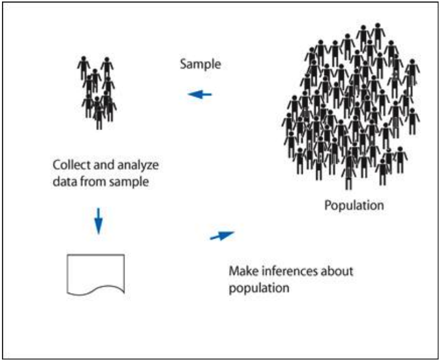

Sample
Collect and analyze data from sample
Population
Make inferences about population

Sample vs. Population


Sample
Collect and analyze data from sample
Population
Make inferences about population

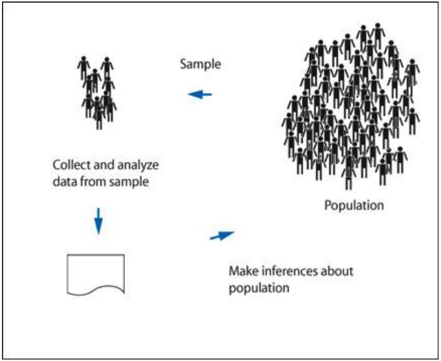
- Unless we have a census, we can never know for sure what the population parameter is...

Sample vs. Population


Sample
Collect and analyze data from sample
Population
Make inferences about population

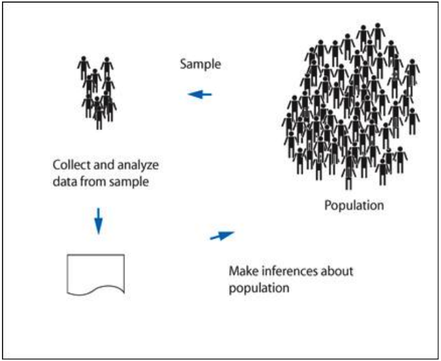
- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)

Sample vs. Population


Sample
Population
Collect and analyze data from sample
Make inferences about population

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)
- Recall Capturing uncertainty is a part of what makes research scientific

Sample vs. Population


Sample
Population
Collect and analyze data from sample
Make inferences about population

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)
- Recall Capturing uncertainty is a part of what makes research scientific
- Afghanistan data: n = 2754 respondents was used to infer experiences of N = 15 million civilians

How uncertain?

- How uncertain (or accurate) will our estimate of the population parameter be?

How uncertain?

- How uncertain (or accurate) will our estimate of the population parameter be?
- Depends on:

How uncertain?

- How uncertain (or accurate) will our estimate of the population parameter be?
- Depends on:
    1. How sample was chosen: must be random sampling

How uncertain?

- How uncertain (or accurate) will our estimate of the population parameter be?
- Depends on:
  1. How sample was chosen: must be random sampling
  2. How large sample is: bigger sample size is better

How uncertain?

- How uncertain (or accurate) will our estimate of the population parameter be?
- Depends on:
  1. How sample was chosen: must be random sampling
  2. How large sample is: bigger sample size is better
  3. Characteristics of population parameter itself: how much variation the population shows

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)
- If there are N units in the population, chance of any particular unit being in the sample must be is:

$$\frac{n}{N}$$

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)
- If there are N units in the population, chance of any particular unit being in the sample must be is:

$$\frac{n}{N}$$

  - N = 230 million possible voters, sample of n = 1000: everyone must have $\frac{1,000}{230,000,000}$ chance of being picked for the sample

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)
- If there are N units in the population, chance of any particular unit being in the sample must be is:

$$\frac{n}{N}$$

  - N = 230 million possible voters, sample of n = 1000: everyone must have $\frac{1,000}{230,000,000}$ chance of being picked for the sample
  - N = 6999 undergrads at Stanford, sample of n = 100 for a undergrad survey: everyone must have $\frac{100}{6999}$ chance of being picked for the survey sample

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)
- If there are N units in the population, chance of any particular unit being in the sample must be is:

$$\frac{n}{N}$$

  - N = 230 million possible voters, sample of n = 1000: everyone must have $\frac{1,000}{230,000,000}$ chance of being picked for the sample
  - N = 6999 undergrads at Stanford, sample of n = 100 for a undergrad survey: everyone must have $\frac{100}{6999}$ chance of being picked for the survey sample
- Produces sample that is representative of the population: if we repeat the same random sampling procedure many times, features of each sample are not identical to those of population, but on average they are identical

Random Sampling

- Simple random sampling: every member of the population must have an equal chance of being chosen: random selection (there are other methods of random sampling)
- If there are N units in the population, chance of any particular unit being in the sample must be is:

$$\frac{n}{N}$$

  - N = 230 million possible voters, sample of n = 1000: everyone must have $\frac{1,000}{230,000,000}$ chance of being picked for the sample
  - N = 6999 undergrads at Stanford, sample of n = 100 for a undergrad survey: everyone must have $\frac{100}{6999}$ chance of being picked for the survey sample
- Produces sample that is representative of the population: if we repeat the same random sampling procedure many times, features of each sample are not identical to those of population, but on average they are identical

Random Sampling

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!

- Random sampling eliminates biased differences between population and sample

Random Sampling

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!

- Random sampling eliminates biased differences between population and sample
- Get it wrong

Random Sampling

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!

- Random sampling eliminates biased differences between population and sample
- Get it wrong
  - Allow people to self-select into answering

Random Sampling

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!

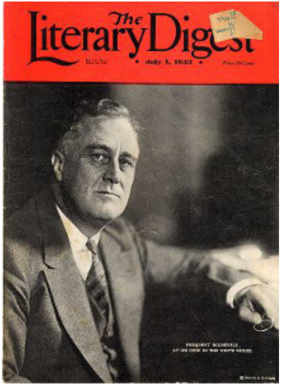
- Random sampling eliminates biased differences between population and sample
- Get it wrong
  - Allow people to self-select into answering
  - Only include certain types of people in sample

Random Sampling

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!

- Random sampling eliminates biased differences between population and sample
- Get it wrong
  - Allow people to self-select into answering
  - Only include certain types of people in sample
- Answers are meaningless!

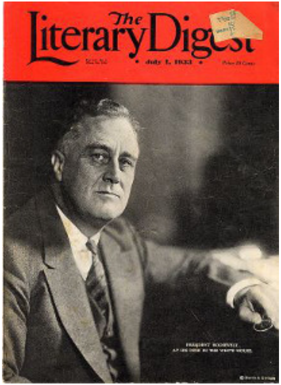
Random Sampling Fail



Example of getting it wrong:

- Literary Digest (LD) poll 1936 to predict Roosevelt vs. Landon election

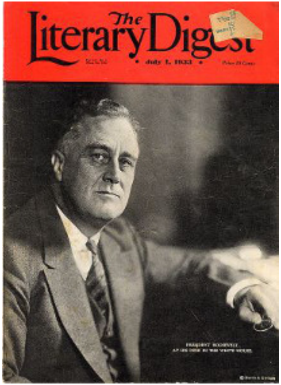
Random Sampling Fail



Example of getting it wrong:

- Literary Digest (LD) poll 1936 to predict Roosevelt vs. Landon election
- LD gets names and address from phone records, automobile clubs, own subscribers

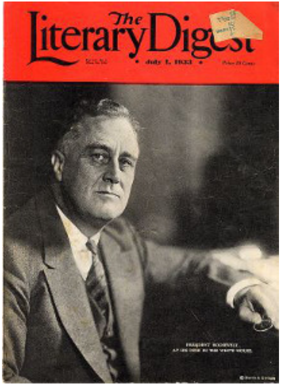
Random Sampling Fail



Example of getting it wrong:

- Literary Digest (LD) poll 1936 to predict Roosevelt vs. Landon election
- LD gets names and address from phone records, automobile clubs, own subscribers
- LD send out 10 million sample ballots, get 2.4 million back (HUGE)

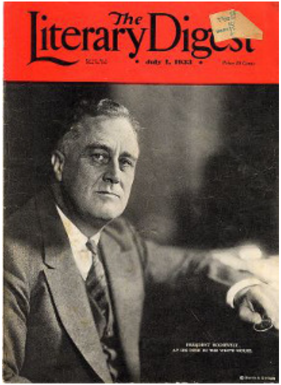
Random Sampling Fail



Example of getting it wrong:

- Literary Digest (LD) poll 1936 to predict Roosevelt vs. Landon election
- LD gets names and address from phone records, automobile clubs, own subscribers
- LD send out 10 million sample ballots, get 2.4 million back (HUGE)
- Prediction:
  Roosevelt 43%, Landon 57%

Random Sampling Fail

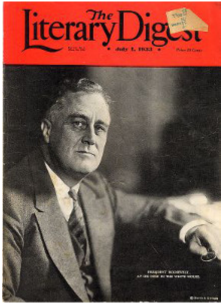


Example of getting it wrong:

- Literary Digest (LD) poll 1936 to predict Roosevelt vs. Landon election
- LD gets names and address from phone records, automobile clubs, own subscribers
- LD send out 10 million sample ballots, get 2.4 million back (HUGE)
- Prediction:
  Roosevelt 43%, Landon 57%
- Actual:
  Roosevelt 60.8%, Landon 36.5%

Random Sampling Fail

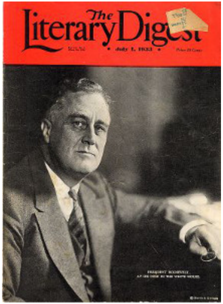
What went wrong?



- Were people in the sample reflective of population of voters?

Random Sampling Fail

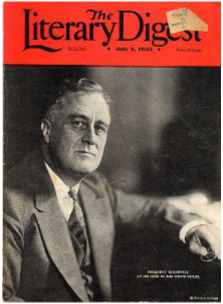
What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study

Random Sampling Fail

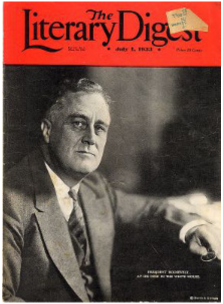
What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study
- Assumed it was everyone in phone directory + those in auto clubs + subscribers

Random Sampling Fail

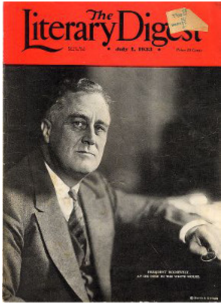
What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study
- Assumed it was everyone in phone directory + those in auto clubs + subscribers
- If you were poor, rural, semi-literate, what is your probability of being in that sample?

Random Sampling Fail

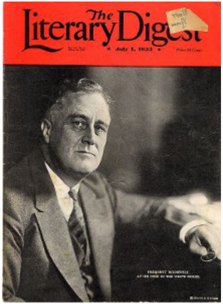
What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study
- Assumed it was everyone in phone directory + those in auto clubs + subscribers
- If you were poor, rural, semi-literate, what is your probability of being in that sample?
  - Was the probability $\frac{n}{N}$ where n = 2.4 million, N = 46 million voters?

Random Sampling Fail

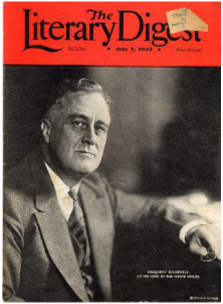
What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study
- Assumed it was everyone in phone directory + those in auto clubs + subscribers
- If you were poor, rural, semi-literate, what is your probability of being in that sample?
  - Was the probability $\frac{n}{N}$ where n = 2.4 million, N = 46 million voters?
  - NO! It was 0

Random Sampling Fail

What went wrong?



- Were people in the sample reflective of population of voters?
- Wrong sampling frame: wrong method for defining population LD wanted to study
- Assumed it was everyone in phone directory + those in auto clubs + subscribers
- If you were poor, rural, semi-literate, what is your probability of being in that sample?
  - Was the probability $\frac{n}{N}$ where $n = 2.4$ million, $N = 46$ million voters?
  - NO! It was 0
- If you were a rich, city-dwelling lawyer, what is your probability of being in thatsample?

Random Sampling Fail

Another example: Republican Primary 2007

- CNBC post-debate online poll: "who won the debate?"
- 7,000 respondents: Ron Paul won with 75% of the vote

Random Sampling: Afghanistan Data

- Altitude and population in surveyed and non-surveyed villages
- Some outliers
- Distribution of these two variables is largely similar between the sampled and non-sampled villages

Sampling Bias

- Sampling bias: wrong sampling frame (some types of people, car owners, rich, more likely to be included in survey than others)
    - How successful is Alcoholics Anonymous? (But who selects into AA? what are their motivations?)
    - If certain 'types' of units are systematically more or less likely to appear in your sample, you probably have a selection effect

Uncertainty

Uncertainty

- Unless we have a census, we can never know for sure what the population parameter is...

Uncertainty

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)

Uncertainty

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)

Depends on:

1. How sample was chosen: random sampling to eliminate bias btw sample and population ✓

Uncertainty

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)

Depends on:

1. How sample was chosen: random sampling to eliminate bias btw sample and population ✓
2. How large sample is: bigger sample size is better

Uncertainty

- Unless we have a census, we can never know for sure what the population parameter is...
- ...but we can estimate it (with varying degrees of uncertainty)

Depends on:

1. How sample was chosen: random sampling to eliminate bias btw sample and population ✓
2. How large sample is: bigger sample size is better
3. Characteristics of population parameter itself: how much variation the population shows

Random Sampling Error

Random Sampling Error

- We can avoid bias, but we get some random sampling error

Random Sampling Error

- We can avoid bias, but we get some random sampling error
  - How sample statistic differs by chance from population parameter

Random Sampling Error

- We can avoid bias, but we get some random sampling error
    - How sample statistic differs by chance from population parameter
    - We know exactly how this error is affecting our estimate (with bias, we never know for sure)

Random Sampling Error

- We can avoid bias, but we get some random sampling error
  - How sample statistic differs by chance from population parameter
  - We know exactly how this error is affecting our estimate (with bias, we never know for sure)

Without sampling bias:

- Random sampling error for average height of students on campus?

Random Sampling Error

- We can avoid bias, but we get some random sampling error
  - How sample statistic differs by chance from population parameter
  - We know exactly how this error is affecting our estimate (with bias, we never know for sure)

Without sampling bias:

Population parameter = sample statistic + random sampling error

- Random sampling error for average height of students on campus?
  - Get random sample of 100 (no bias!)

Random Sampling Error

- We can avoid bias, but we get some random sampling error
  - How sample statistic differs by chance from population parameter
  - We know exactly how this error is affecting our estimate (with bias, we never know for sure)

Without sampling bias:

Population parameter = sample statistic + random sampling error

- Random sampling error for average height of students on campus?
  - Get random sample of 100 (no bias!)
  - Measure heights

Random Sampling Error

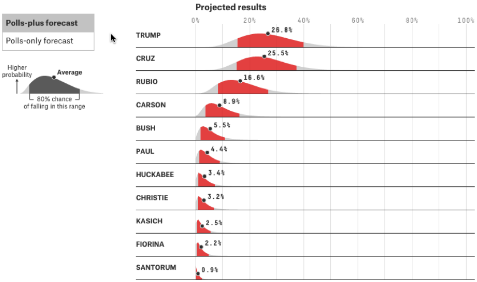
- We can avoid bias, but we get some random sampling error
    - How sample statistic differs by chance from population parameter
    - We know exactly how this error is affecting our estimate (with bias, we never know for sure)

Without sampling bias:

Population parameter = sample statistic + random sampling error

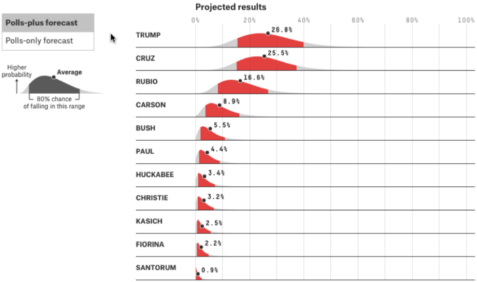
- Random sampling error for average height of students on campus?
    - Get random sample of 100 (no bias!)
    - Measure heights
    - By chance 38% of our sample is less than 5'9", but 36% of true population is less than 5'9"

Random Sampling Error: Primaries



http://projects.fivethirtyeight.com/election-2016/primary-forecast/iowa-republican/

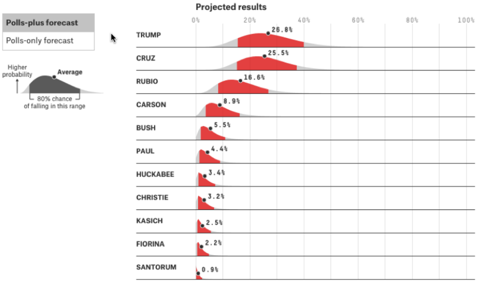
Random Sampling Error: Primaries



http://projects.fivethirtyeight.com/election-2016/primary-forecast/iowa-republican/

- Random sampling error can lead to an overestimate or underestimate relative to true parameter value

Random Sampling Error: Primaries



http://projects.fivethirtyeight.com/election-2016/primary-forecast/iowa-republican/

- Random sampling error can lead to an overestimate or underestimate relative to true parameter value
- Trump 26.8% is somewhere between 15% and 40%

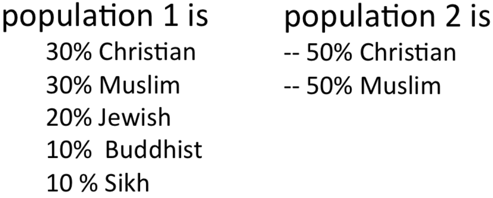
Random Sampling Error: Population Variation, Sample Size

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error



population 1 is

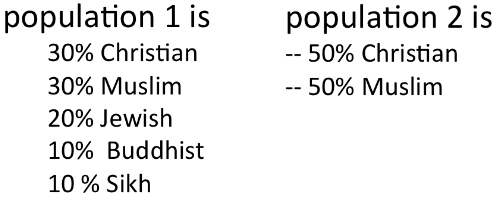
30% Christian
30% Muslim
20% Jewish
10% Buddhist
10 % Sikh

population 2 is

-- 50% Christian
-- 50% Muslim

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error



population 1 is

30% Christian
30% Muslim
20% Jewish
10% Buddhist
10 % Sikh

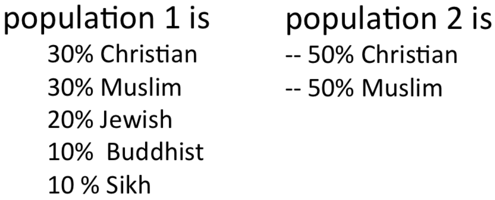
population 2 is

-- 50% Christian
-- 50% Muslim

Which population is likely to be better represented with a sample of 10 people?

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error
- Bigger sample → less random sampling error



population 1 is

30% Christian

30% Muslim

20% Jewish

10% Buddhist

10 % Sikh

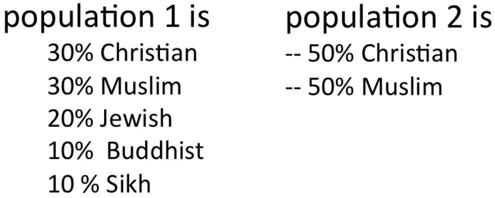
population 2 is

-- 50% Christian

-- 50% Muslim

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error
- Bigger sample → less random sampling error



population 1 is

30% Christian
30% Muslim
20% Jewish
10% Buddhist
10 % Sikh

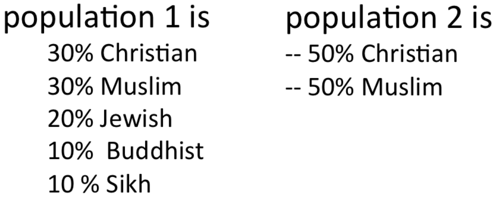
population 2 is

-- 50% Christian
-- 50% Muslim

Population 2 better represented with a sample of 1000 people than 10 people?

Random Sampling Error: Population Variation, Sample Size

- More variation in population → more random sampling error
- Bigger sample → less random sampling error



population 1 is
30% Christian
30% Muslim
20% Jewish
10% Buddhist
10 % Sikh

population 2 is
-- 50% Christian
-- 50% Muslim

Benefit to size not linear: doubling the sample size does not make the estimate twice as good

Surveys and Sources of bias

Surveys and Sources of bias

- Sampling bias and random sampling error apply when doing surveys

Surveys and Sources of bias

- Sampling bias and random sampling error apply when doing surveys
- Other sources of bias in surveys:

Surveys and Sources of bias

- Sampling bias and random sampling error apply when doing surveys
- Other sources of bias in surveys:
  - Unit non-response: failure to reach selected units (in Afghanistan survey, 2754 out of 3097 sampled respondents agreed to participate → 11% refusal rate)

Surveys and Sources of bias

- Sampling bias and random sampling error apply when doing surveys
- Other sources of bias in surveys:
  - Unit non-response: failure to reach selected units (in Afghanistan survey, 2754 out of 3097 sampled respondents agreed to participate → 11% refusal rate)
  - Item non-response: respondents refuse to answer certain survey questions (in Afghanistan survey, the income variable had a non-response rate of approximately 5%)

Surveys and Sources of bias

- Sampling bias and random sampling error apply when doing surveys
- Other sources of bias in surveys:
  - Unit non-response: failure to reach selected units (in Afghanistan survey, 2754 out of 3097 sampled respondents agreed to participate → 11% refusal rate)
  - Item non-response: respondents refuse to answer certain survey questions (in Afghanistan survey, the income variable had a non-response rate of approximately 5%)
  - Mis-reporting

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
  - Asking question about support for foreign forces in Afghanistan sensitive

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
  - Asking question about support for foreign forces in Afghanistan sensitive
  - Will you vote in the next election?

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
  - Asking question about support for foreign forces in Afghanistan sensitive
  - Will you vote in the next election?
  - How often do you lie to people close to you?

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
  - Asking question about support for foreign forces in Afghanistan sensitive
  - Will you vote in the next election?
  - How often do you lie to people close to you?
  - How many sexual partners have you had?

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
    - Asking question about support for foreign forces in Afghanistan sensitive
    - Will you vote in the next election?
    - How often do you lie to people close to you?
    - How many sexual partners have you had?
    - Would you be upset if a black family moved next door?

Surveys and Sources of bias: Misreporting

Different types (reasons) for mis-reporting

- Social desirability bias: respondents choose an answer that is seen as socially desirable regardless of what they really think
  - Asking question about support for foreign forces in Afghanistan sensitive
  - Will you vote in the next election?
  - How often do you lie to people close to you?
  - How many sexual partners have you had?
  - Would you be upset if a black family moved next door?
- Observer effects: people being surveyed are affected (act differently) by being observed

Summing Up

Summing Up

- We estimate population parameters from a sample

Summing Up

- We estimate population parameters from a sample
- Make sure your sample is random to avoid bias

Summing Up

- We estimate population parameters from a sample
- Make sure your sample is random to avoid bias
- Random sampling error is a fact of life

Summing Up

- We estimate population parameters from a sample
- Make sure your sample is random to avoid bias
- Random sampling error is a fact of life
- More variation increases error, larger sample size reduces it

Summing Up

- We estimate population parameters from a sample
- Make sure your sample is random to avoid bias
- Random sampling error is a fact of life
- More variation increases error, larger sample size reduces it
- Problems occur with surveys