Introduction to Data Science

Sampling and Scientific Studies

Components of Statistics

- A general process of investigation:
  - 1. Identify a question or problem.
  - 2. Collect relevant data on the topic.
  - 3. Analyze the data.
  - 4. Form a conclusion.

- Statistics is the study of how best to collect, analyze, and draw conclusions from data (stages 2-4).
  - How best can we collect data?
  - How should it be analyzed?
  - What can we infer from the analysis?

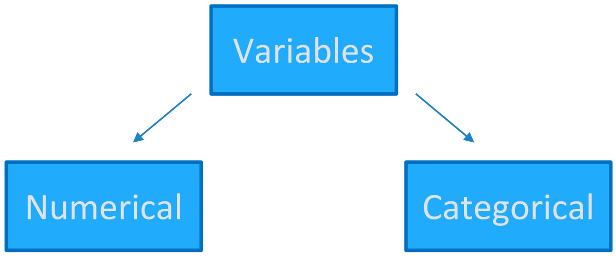
Data Matrix

Variable
↓

email	spam	num_char	line_breaks	format	number
1	No	21705	551	html	small
2	No	7011	183	html	big
3	Yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	No	15829	242	html	small

← Observation
(case)

Types of Variables



Variables

Numerical

(quantitative)

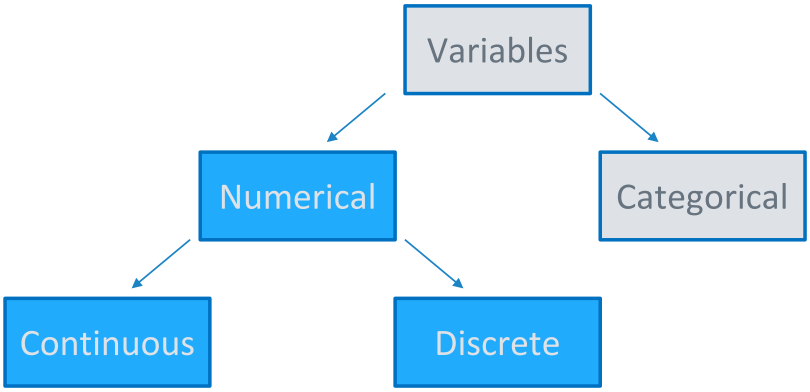
- take on numerical values
- sensible to add, subtract, take averages, etc. with these values

Categorical

(qualitative)

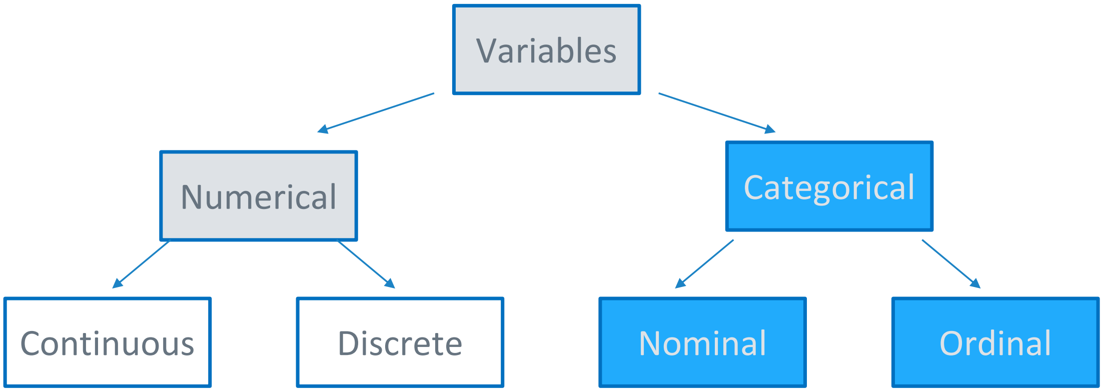
- take on a limited number of distinct categories.
- categories can be identified with numbers, but not sensible to do arithmetic operations

Numerical Variables



- Take on any of an infinite number of values within a given range
- Take on one of a specific set of numeric values

Categorical Variable



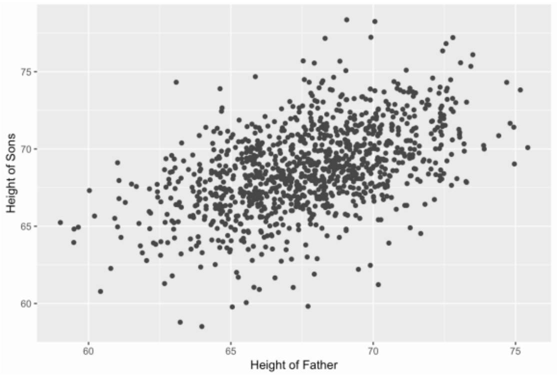
- Levels have an inherent ordering

Example

email	spam	num_char	line_breaks	format	number
1	No	21705	551	html	small
2	No	7011	183	html	big
3	Yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	No	15829	242	html	small
↓ Identity	↓ Nominal Categorical	↓ Discrete Numerical	↓ Discrete Numerical	↓ Nominal Categorical	↓ Ordinal Categorical

Relationships between variables

- Two variables that show some connection with one another are called associated.
- Association can be further described as positive or negative.
- If two variables are not associated, they are said to be independent.



Height of fathers and sons

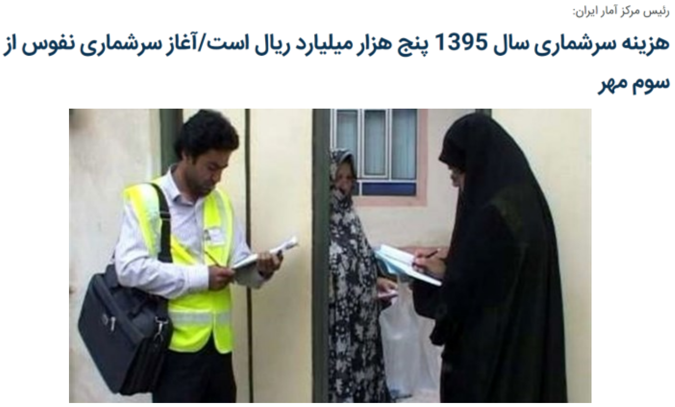
Population and Sample

Population

- Each research question refers to a target population.
- Example:
  - *Research question:* Can adult men become better, more efficient runners on their own, merely by running?
  - *Population of interest: All men over 18*
- Often it is too expensive to collect data for every case in a population.

Census

- Census: collect data from *everyone* in the population.



رئیس مرکز آمار ایران:

هزینه سرشماری سال 1395 پنج هزار میلیارد ریال است/آغاز سرشماری نفوس از سوم مهر

Sampling

- A sample represents a subset of the cases and is often a small fraction of the population.
- Think about sampling something you are cooking: you taste a small part of what you're cooking to get an idea about the dish as a whole.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.

Anecdotal Evidence

- Consider the following statements:
  - My uncle smokes three packs a day and he's in perfectly good health, so smoking doesn't affect your health.
- The conclusion is based on data, but there are two problems:
  - First, the data only represent one or two cases.
  - Second, it is unclear whether these cases are actually representative of the population.
- Data collected in this haphazard fashion are called anecdotal evidence.

Sampling Bias

DO YOU PREFER CATS OR DOGS?

Some Sources of Sampling Bias

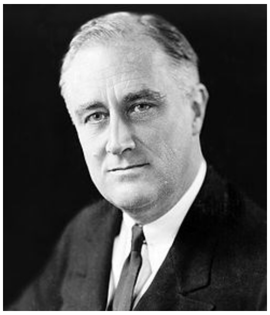
- *Non-response:* If only a *non-random* fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

Sampling Bias Example

- A historical example of a biased sample yielding misleading results:

Alf Landon

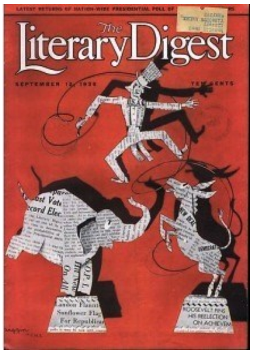
- In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



Franklin D. Roosevelt

The Literary Digest Poll

- The Literary Digest polled 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the winner and FDR would get 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.



What went wrong?

- The magazine had surveyed:
  - its own readers
  - registered automobile owners, and registered telephone users
- These groups had incomes well above the national average of the day which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time.
- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.

Type of Studies

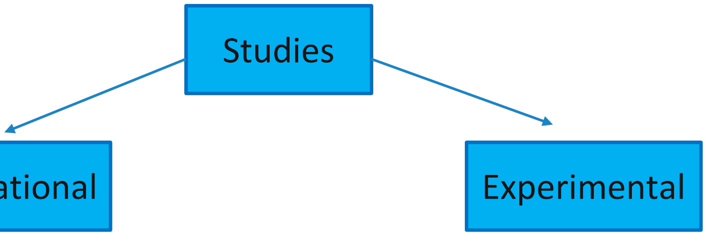
Explanatory and Response Variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{Explanatory variable} \xrightarrow{\text{might affect}} \text{Response variable}$$

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is a high correlation between the two variables.
- We use these labels only to keep track of which variable we suspect affects the other.

Observational Studies & Experiments



Studies

Observational

Experimental

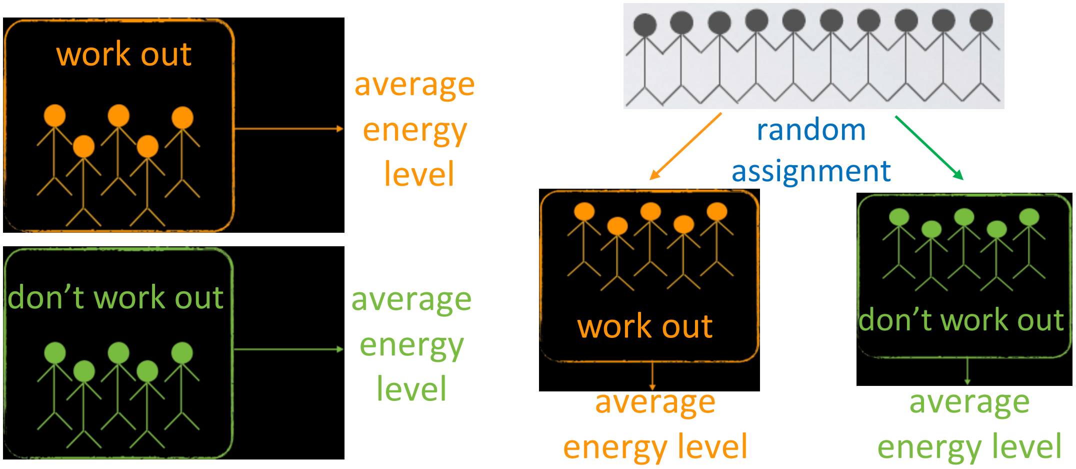
- collect data in a way that does not directly interfere with how the data arise ("observe")
- only establish an association
- retrospective: uses past data
- prospective: data are collected throughout the study

- randomly assign subjects to treatments
- establish causal connections between explanatory and response variables.

Observational vs. Experimental Studies

Observational Study

Experiment



work out

average energy level

don't work out

average energy level

random assignment

work out

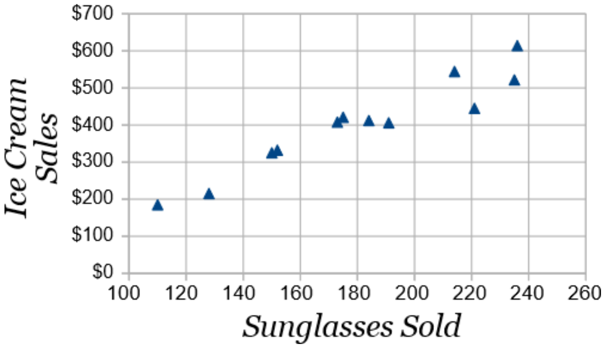
average energy level

don't work out

average energy level

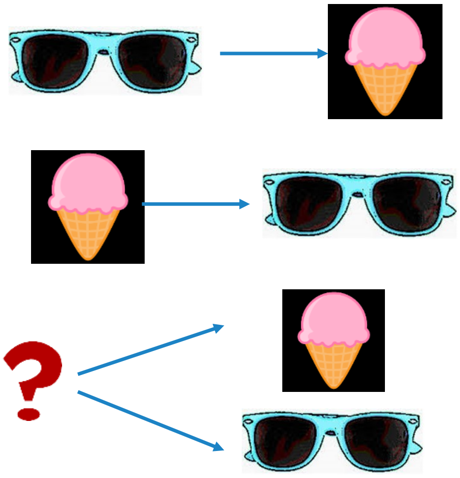
Correlation does not imply causation

- The local ice cream shop keeps track of how much ice cream they sell.
- The ice cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales.



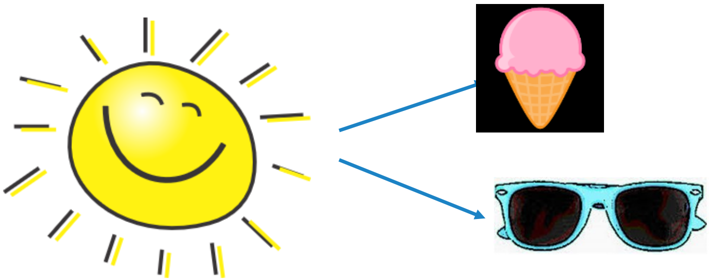
Three possible explanations

1. Sunglasses make people want ice cream!
2. Eating ice cream makes people buy sunglasses!
3. A third variable is responsible for both.

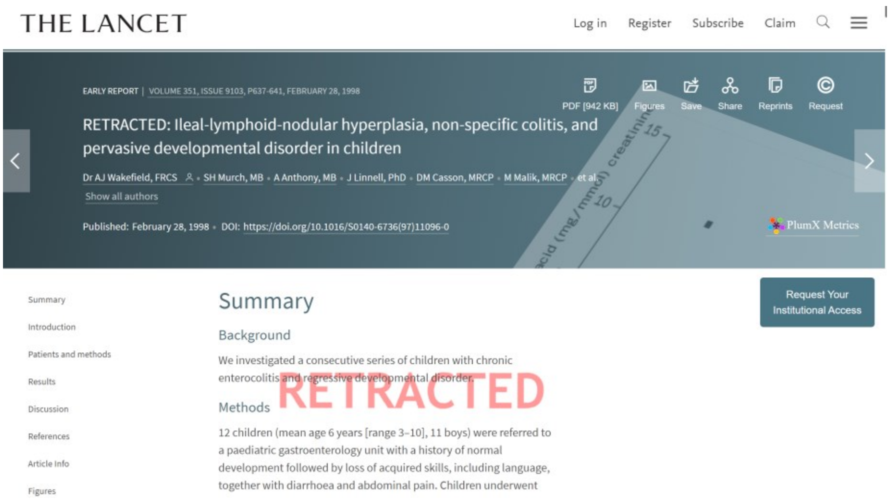


Confounding Variable

- An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called confounders or confounding variables.



MMR Vaccination and Autism



EARLY REPORT | VOLUME 351, ISSUE 9103, P637-641, FEBRUARY 28, 1998

RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Dr AJ Wakefield, FRCS • SH Murch, MB • A Anthony, MB • J Linnell, PhD • DM Casson, MRCP • M Malik, MRCP • et al.

Published: February 28, 1998 • DOI: https://doi.org/10.1016/S0140-6736(97)11096-0

Summary

Background

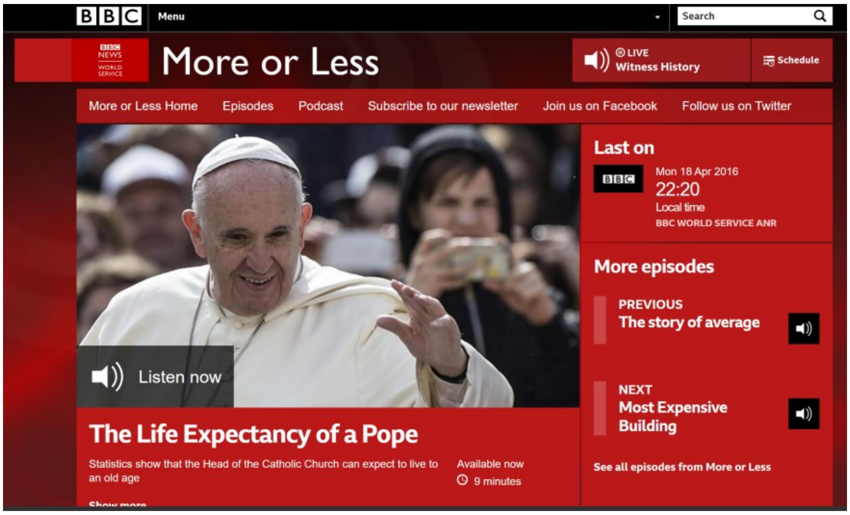
We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

RETRACTED

Methods

12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent

Do popes live longer?



More or Less

Last on

Mon 18 Apr 2016
22:20
Local time
BBC WORLD SERVICE ANR

More episodes

PREVIOUS
The story of average

NEXT
Most Expensive Building

See all episodes from More or Less

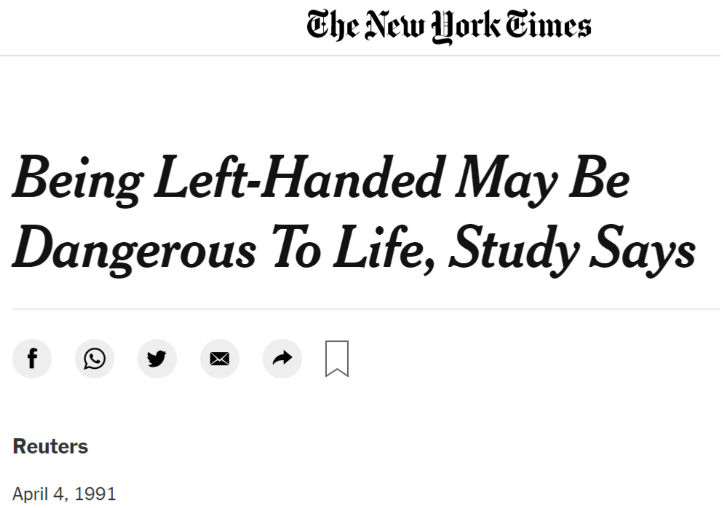
The Life Expectancy of a Pope

Statistics show that the Head of the Catholic Church can expect to live to an old age

Available now
9 minutes

Show more

Left-handedness and Life Expectancy



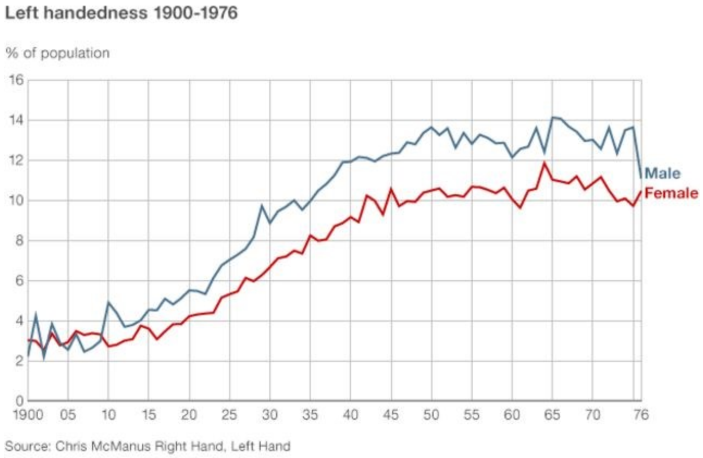
The New York Times

Being Left-Handed May Be Dangerous To Life, Study Says

Reuters

April 4, 1991

Left-handedness and Life Expectancy

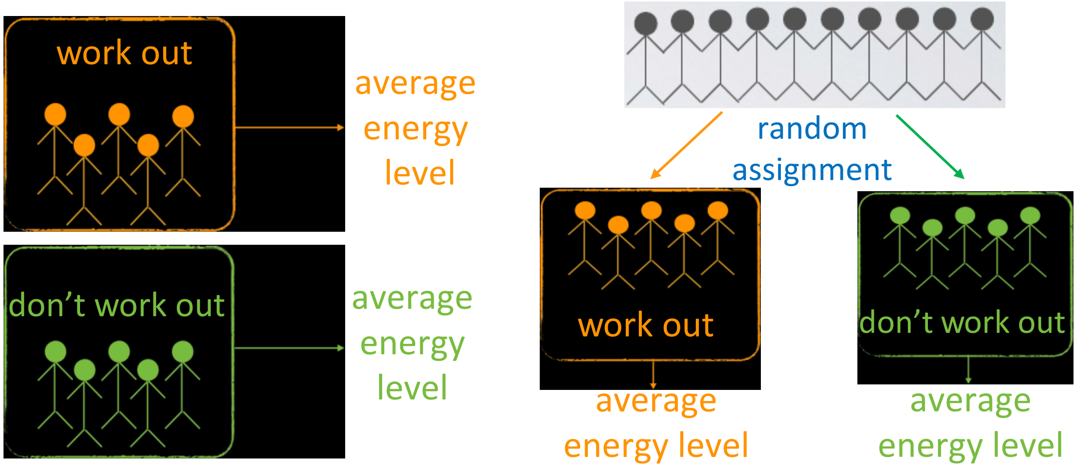


Source: Chris McManus Right Hand, Left Hand

Observational vs. Experimental Studies

Observational Study

Experiment


work out

average energy level

don't work out

average energy level

random assignment

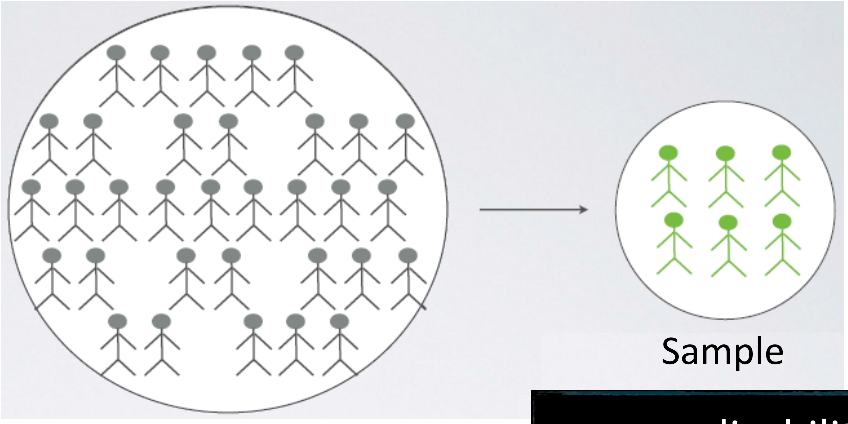
work out

average energy level

don't work out

average energy level

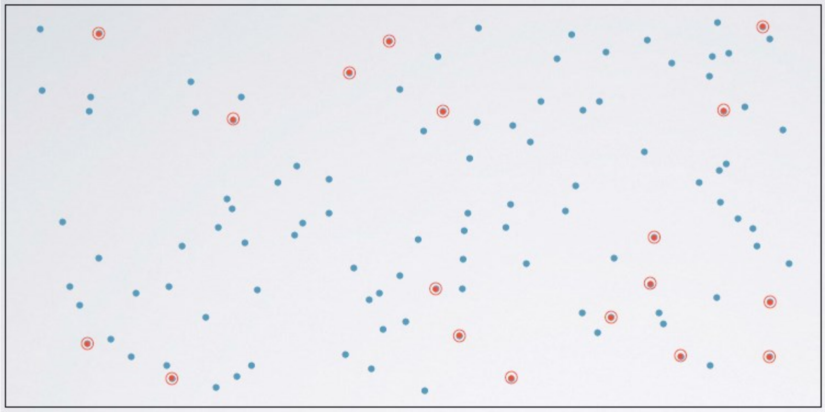
Random Sampling



Population

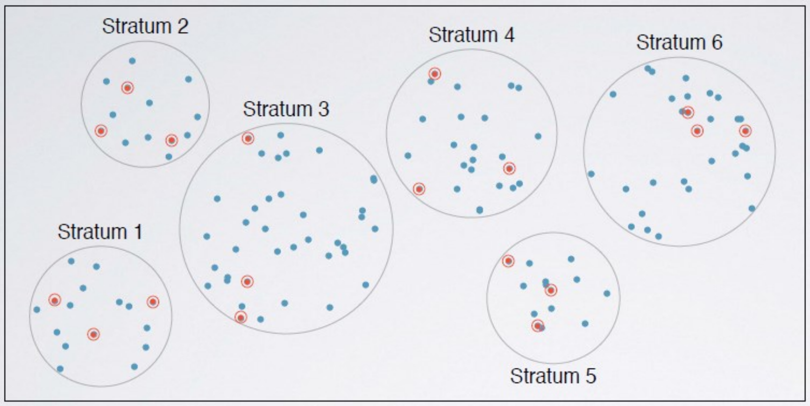
generalizability

Simple Random Sampling (SRS)



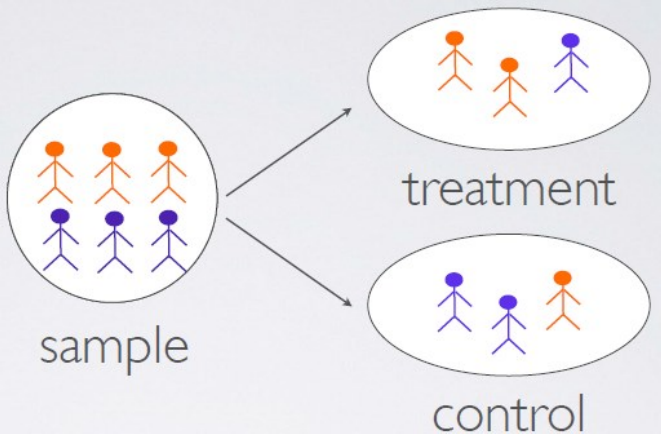
- Each case is equally likely to be selected.

Stratified Sampling



- Divide the population into homogenous strata, then randomly sample from within each stratum.

Random Assignment

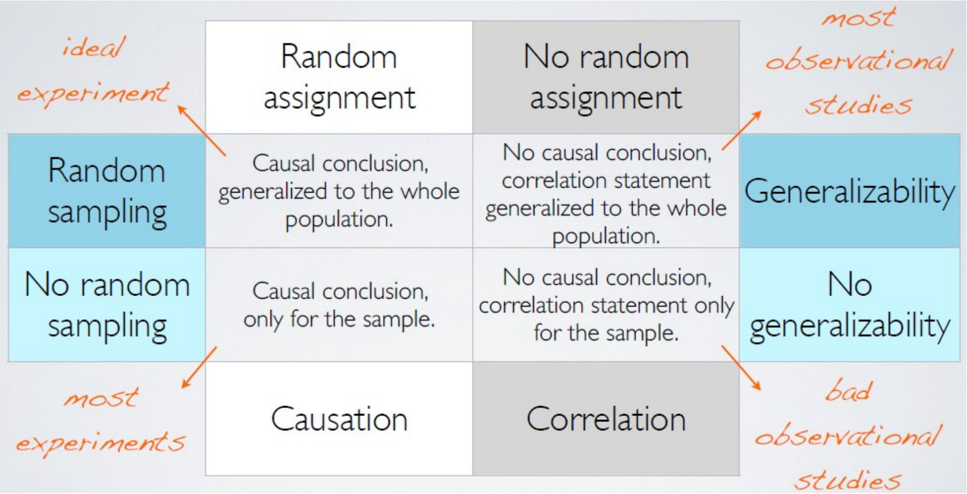


Causality

Principles of Experimental Design

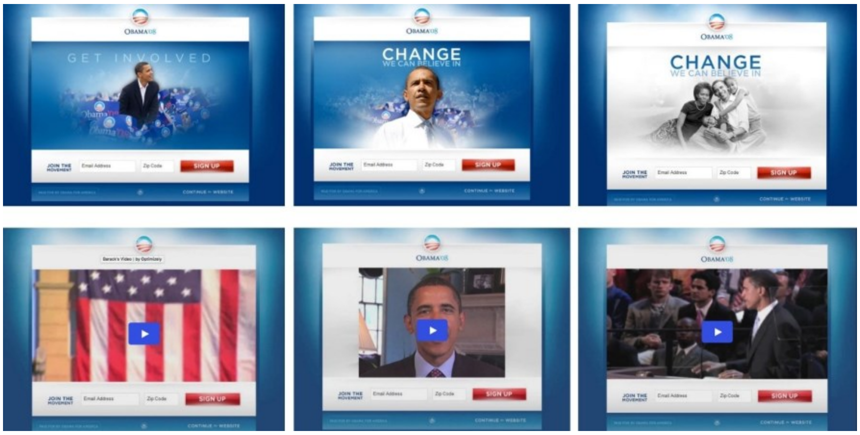
- *Control:* Compare treatment of interest to a control group.
- *Randomize:* Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
- *Replicate:* Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
- *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

Random Assignment vs. Random Sampling



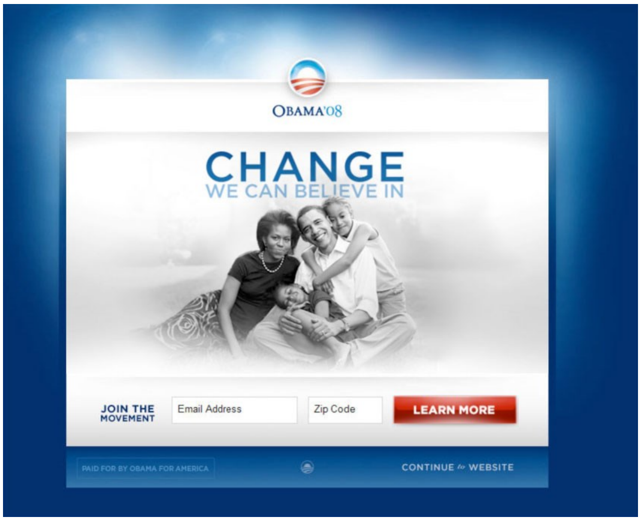
	Random assignment	No random assignment	
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
	Causation	Correlation	

A/B Testing for US Presidential Campaign



JOIN US NOW
LEARN MORE
SIGN UP
SIGN UP NOW

The Winner


OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE