Data: basics

Lecture 01

Objectives

- Contrast population vs. sample
- Understand Data collection/sampling challenges
- Contrast observational study vs. statistical experiment

Data

- *Data* is a collection of things known or assumed as facts
- Data is empirical—it refers to something that is observed, in contrast to unobservable or theoretical concepts
- When you think *data*, think *tables*—each column is a different attribute, each row represents the measurements of each attribute for a specific entity

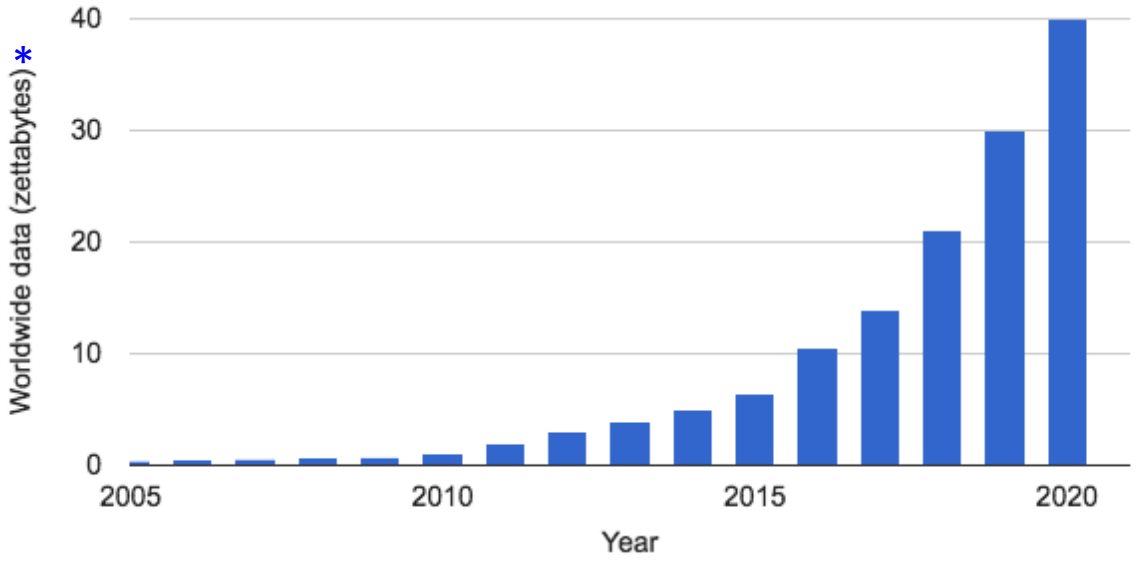
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data sources: examples

User-generated	Business	System monitoring

• Social Networks: Facebook, Twitter…
• Blogs, comments
• Personal documents
• Pictures: Instagram, Snapchat, etc.
• Videos: YouTube etc.
• Internet searches
• Mobile text messages
• User-generated maps
• E-mail
	• Public Agencies
	• Medical records
	• Commercial transactions
	• Banking/stock records
	• E-commerce
	• Credit cards
		• Fixed sensors
		  • Home automation
		  • Weather/pollution sensors
		  • Traffic sensors/webcam
		  • Security/surveillance videos/images
		• Mobile sensors (tracking)
		  • Mobile phone location
		  • Cars
		  • Satellite images
		• Data from computer system Logs
		• Web logs

Digital data: volume



*A zettabyte is one sextillion or 10^{21} bytes.

Big data refers to very large data sets that cannot be processed by traditional methods, and is characterized by high volume, rapid velocity of collection, and variety in type and quality.

Two types of data sets

- Population
- Sample

Population vs. sample

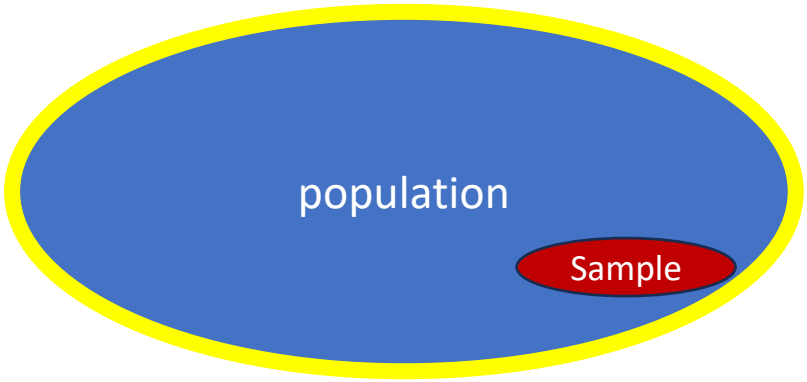
- *Population*: the entire set of all individuals, items, or events of interest
  - Examples: “grades of students in this class”, “all five-card poker hands”, “all car models made by Chevrolet”
  - An *observational unit* is an individual, item, or event of the population for which we create a single data record
    - Example: one specific poker hand in the population of all five-card hands
- A *sample* is a small **subset** of observational units from the population
  - Example: “50 randomly generated five-card poker hands”

The word “statistics” has two distinct meanings

- *Statistics* is the field of study that deals with proper collection, analysis, and interpretation of data.
- *Statistics* can also mean numerical characteristics of data: *mean*, *median*, *variance* etc.

Population and population *sample*

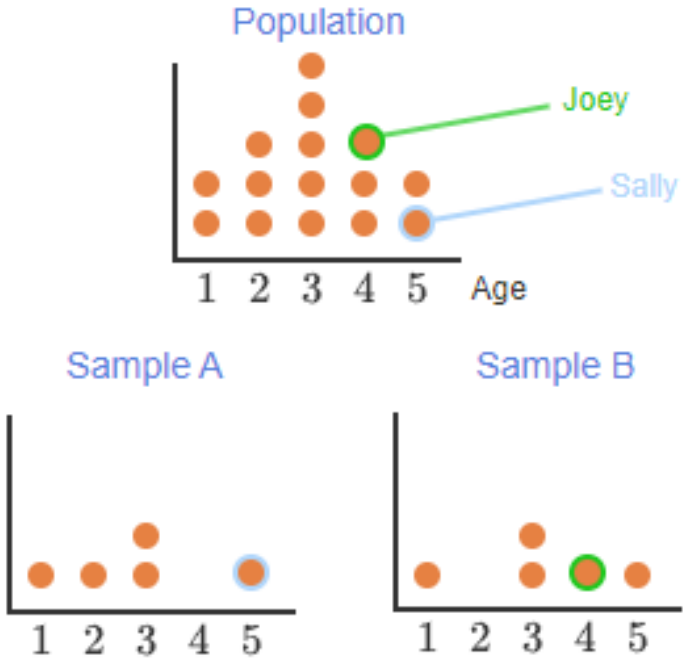
- **S**ample – subset of the population for which the data of interest is collected. We compute the **s**tatistics (as metrics) for a **S**ample.
- **P**opulation - a group of interest: everybody inside the yellow boundary. The goal of statistics (as a discipline) is to learn **P**opulation **p**arameters.



Our goal: learn **P**opulation **P**arameters

Example: parameter vs statistics

Consider two different samples of a population of preschool children in this kindergarten



Population mean: 3
Parameter

Sample A's mean: 2.8
Sample B's mean: 3.2
Statistics

- The population mean is a *parameter*.
- The mean of each sample is a *statistic*.

We want Population Parameters but mostly have Sample Statistics

- In most situations, it is not possible to collect data about the entire population related to some subject one wishes to study.
- Statistics and ML are tools for making inferences about the population parameters based on a sample statistics.

Statistical inference

- *Statistical inference* includes methods that produce estimates about the population based on a sample (or multiple samples)
- True population parameters in most cases cannot be measured directly, and can only be *inferred* from samples

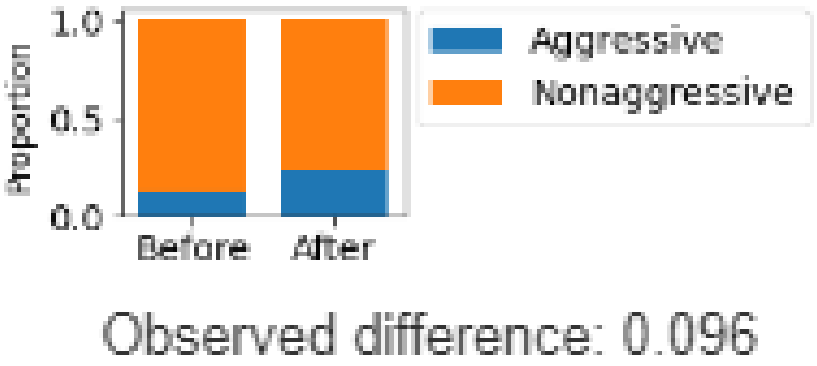
Example: statistical inference

We want to study behavior changes in wild African painted dogs when a new dog is introduced into the pack.

A study is designed to collect behavior data on a random sample of dogs in the zoo before and after the new dog was introduced.

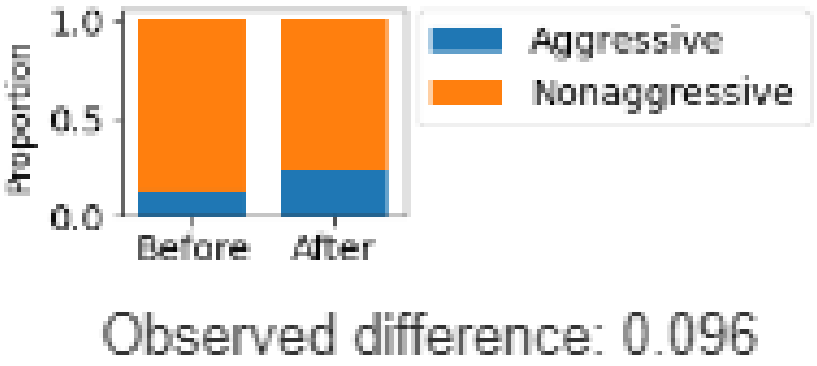
Dog ID	Date	Behavior	New dog	Aggressive
54672	3/2	grooming	no	no
43284	3/2	aggression	no	yes
12114	3/6	sleeping	no	no
...	...	...	...	...
43284	5/18	aggression	yes	yes
87401	5/19	aggression	yes	yes
12114	5/19	eating	yes	no

Example: statistical inference



After the new dog was introduced → the proportion of aggressive behaviors increased by 0.096 (0.128 to 0.224)

Example: statistical inference



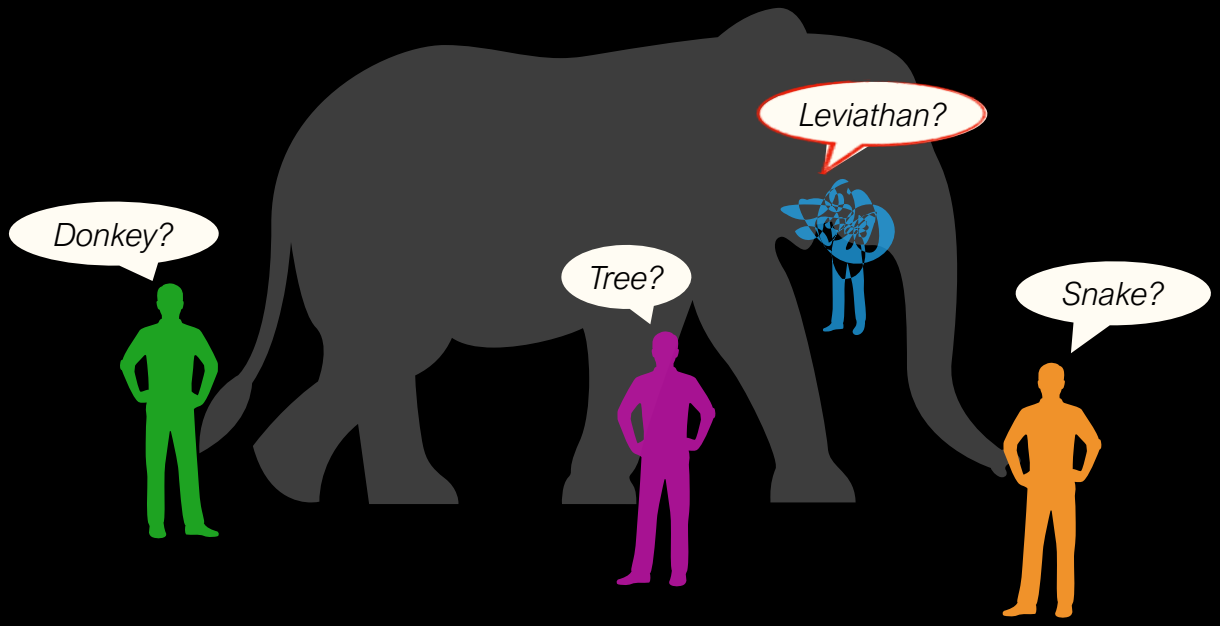
After the new dog was introduced → the proportion of aggressive behaviors increased by 0.096 (0.128 to 0.224)

We want to know:

- Is this change in behavior statistically significant?
- Does it reflect the behavior of the entire population in the zoo?
- Can we extrapolate to the entire species of African pained dogs?

Methods of inferential statistics help us answer these questions

Potential problems with sampling



Each blind person “samples” the object by feeling with their fingers.
Based on that sample, they estimate what object is there.
We can aggregate data from multiple samples to estimate the hidden object (population)

Sampling

- A ***sampling method*** is a process by which observational units are selected from the population to be included in the sample.
- The goal is to have a sample that is a good representative of the population. That is, you want the sample statistics to align with the population parameters.

- The way the data is collected impacts reliability of the results

Sampling methods

- *Random* sampling: select units at random from the entire population
- *Stratified* sampling: divide population into groups **by important attribute**, select random representatives from each *strata*
- *Cluster* sampling: divide population into groups **by non-important attribute**, sample the entire content of some randomly selected clusters
- *Systematic* sampling: select every k-th observational unit:

  where $k = \frac{population\ size}{sample\ size}$

- *Convenience* sampling: select units that are easier to access

Sampling methods: examples

- *Random* sampling: randomly generate 100 10-digit phone numbers with Pittsburgh area code and call them to do survey their love to Pittsburgh
- *Stratified* sampling: in a college with 1800 male and 200 female students a sample of 100 would consist of 90 randomly chosen males and 10 randomly chosen females (but only if gender is believed to be an important attribute!).
- *Cluster* sampling: divide students by their dorm, and survey all students in Tower B and Nordenberg dorms (two dorms randomly selected from all the dorms)
- *Systematic* sampling: 100th person who signs up for a service is asked to complete a customer survey – until we reach the desired sample size
- *Convenience* sampling: a psychology professor uses his students for the study offering them extra-credit

Case study: train passengers satisfaction study in Australia

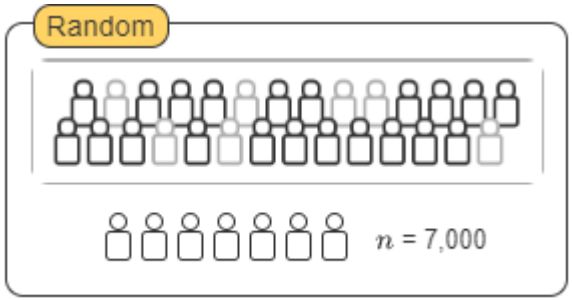
- There are 5 largest cities connected by trains
- There are approximately 2 million adult train passengers per day
- We cannot survey the entire population: we will survey a sample of 7,000 passengers

Sampling example: 1/3

- *Random* sampling:

Sampling example: 1/3

- *Random* sampling:
  - Observational units are selected at random from the population.
  - Each subset is equally likely.



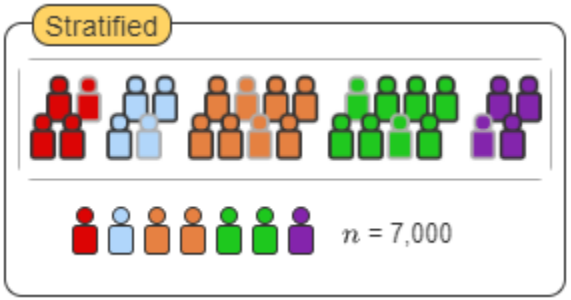
In random sampling, passengers are selected at random from a list of all passengers in the five cities.

Sampling example: 2/3

- *Stratified* sampling:

Sampling example: 2/3

- *Stratified* sampling:
  - The population is first divided into groups, called strata, based on a meaningful feature (of interest).
  - Then observational units are selected from each stratum in proportion to the total strata size.



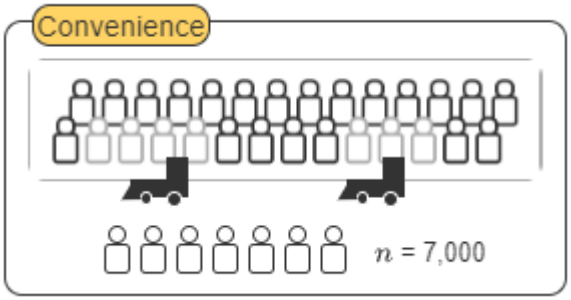
Stratified sampling ensures adequate representation **from each city**.
Passengers are divided into groups, or strata, based on city.
Then from each strata, passengers are selected at random.

Sampling example: 3/3

- *Convenience* sampling:

Sampling example: 3/3

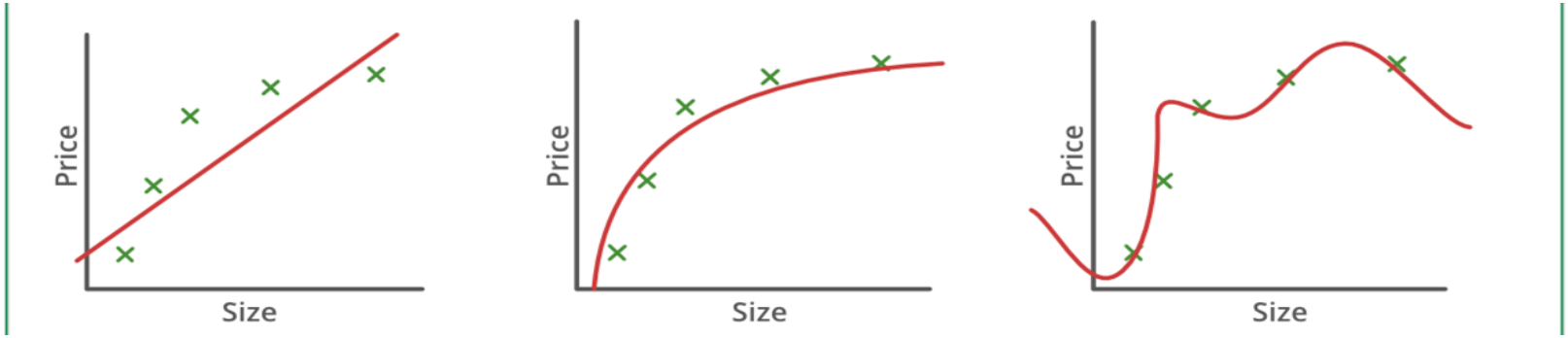
- *Convenience* sampling:
  - Select observational units that are easier to access



Selecting passengers waiting in the train stations uses convenience sampling. This method is easy and quick, but the sample is not likely to be representative of all train passengers.

Problems with sampling: bias

- In data collection, a *sampling bias* is a difference between the parameter inferred from a sample and the true value of the parameter in the population.
  - **Not to be confused** with the statistical model bias: when the model ignores parameters and relationships in the data that actually exist. The model with high bias is insensitive to training data.



High bias model

Low bias model

Some examples of sampling bias

- *Selection* bias: bad selection methodology
- *Leading question (response)* bias: questions elicit particular response
- *Social desirability* bias: respondents answer to please others
- *Self-selection (volunteer)* bias: respondents are a self-selected group
- *Nonresponse* bias: some prefer not to share their opinions

Examples of sampling bias

- *Selection* bias: recruiting volunteers to study effectiveness of a new diet/exercise routine would oversample people that are already interested in healthy lifestyle
- *Leading question (response)* bias: asking questions like "How enjoyable was your recent shopping experience with us?“ assumes the enjoyment
- *Social desirability* bias: people inflate how much they donate to charity
- *Self-selection (volunteer)* bias: conducting poll on X or Twitter
- *Nonresponse* bias: students without strong opinions are unlikely to participate in the evaluation survey

What is the problem with this data sampling?

- A survey posted on a website

What is the problem with this data sampling?

- A survey posted on a website

Voluntary response bias?

What is the problem with this data sampling?

- A survey about the frequency of alcohol consumption

What is the problem with this data sampling?

- A survey about the frequency of alcohol consumption

Social desirability bias?

What is the problem with this data sampling?

- In-person survey about economic policies conducted at a mall in an affluent neighborhood

What is the problem with this data sampling?

- In-person survey about economic policies conducted at a mall in an affluent neighborhood

Selection bias?

What is the problem with this data sampling?

- A survey conducted online by a political news site

What is the problem with this data sampling?

- A survey conducted online by a political news site

Self-selection bias?

Two ways of generating data:

- Observational study: data is collected by recording the responses and measuring features as they occur without any direct influence on the observed data
  - Example: Recording data about birds in my backyard

- Statistical experiment: treatments are first assigned to observational units and then responses are recorded
  - Example: An A/B test is conducted by randomly assigning participants to view one of two possible web page layouts then collecting data on webpage clicks for each layout.
  - With random assignment of treatments to observational units, we can investigate whether the treatment is the cause of the observed response

Observational studies vs experiments