Statistics and Estimation for Computer Science

İstanbul Teknik Üniversitesi

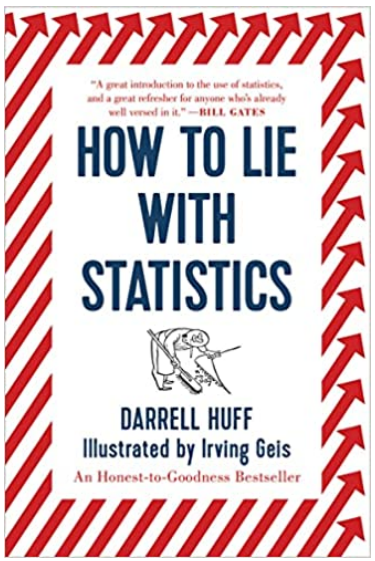
Mustafa Kamasak, PhD

Version: 2022.2.22

Introduction and Concepts

Why Statistics and Estimation

- Data analysis & data science
  "Without data you're just a person with an opinion." – W. Edwards Deming
- "Data" in this sentence has a broader meaning eg.
- It is very easy to collect data
- It is very easy to lie/manipulate with statistics
- "Statistics is scientific data literacy"
- "Probability and Statistics is the difference between gambling and taking risks"



Why Statistics and Estimation

Data literacy

- How to summarize (describe) and visualize data
- How to collect data
- How to design an experiment
- How to determine required data amount
- How reliable are the results
- How to draw conclusions, error probabilities in decision
- How to compare results of different algorithms, classifiers etc.

Why Statistics and Estimation

Human interpretation and reasoning has serious limitations and prone to cognitive biases

- Cannot correctly generalize facts
- Cannot detect patterns from noisy data especialy non-linear patterns and multivariate data
- Frequently detect artificial (non-existing) patterns (Pareidolia)
- Cognitive biases [1]
  - Availability heuristic: The tendency to overestimate the likelihood of events with greater "availability" in memory
  - Confirmation bias: The tendency to search for, interpret, focus on and remember information in a way that confirms one's own view
  - Dunning–Kruger effect: The tendency for unskilled individuals to overestimate their own ability and the tendency for experts to underestimate their own ability
  - ...

[1] https://en.wikipedia.org/wiki/List_of_cognitive_biases

Basic Questions in Statistics

- How to collect data. (Critically important but out of scope in this course!)
- How to describe and summarize the collected data.
- How to draw conclusions from the data.

I check out attendance in the first lecture of my statistics course (data collection)

↓

There are 5 female and 20 male students in the first lecture (description)

↓

Male students are 4 times more interested in statistics compared to female students (inference)

Sample News

Bir Çocuğun Okul Başarısı Evdeki Kitap Sayısı İle Doğru Orantılı

Yapılan bir araştırmada uzmanlar çocuğun okul başarısı ile evdeki kitap sayısının doğru orantılı olduğunu bulmuşlardır. Detayları...

HABER 06/09/2018 15:39

İş ve Yönetimde Kadınlar

Liderlik konumundaki kadınlar işletmelerin daha iyi performans göstermesini sağlıyor

Yeni rapora göre, toplumsal cinsiyet çeşitliliği işletmelerde daha iyi sonuçlar getiriyor; yeteneklerin cezbedilmesini kolaylaştırıyor. Rapor ayrıca, toplumsal cinsiyet çeşitliliğinde süregiden eksikliğin ardındaki nedenleri inceliyor ve döngünün kırılması için öneriler sunuyor.

Why Statistics and Estimation

- Statistical techniques are required for decision-making, interpretation of results etc. in presence of **variability and uncertainty**.
- Sources of variability
  - Inherent to the system
  - Environmental
  - Personal (emotions, mood, experience etc.)
  - Limited precision of measurement
  - Known/unknown sources of noise (thermal, white, pink etc.)
  - ...
- It is not possible to avoid variations and uncertainty
- **Statistical Thinking** is a process
  - to describe, understand the variability which we cannot avoid,
  - to incorporate the variability into decision-making procedure

Descriptive (Betimsel) Statistics

- Interpretation and drawing conclusions from big&raw data is not easy.
- Data has to be
  - Summarized
  - Described
  - Organized
  - Visualized

  for easier interpretation and analysis.
- This branch of statistics is called “descriptive statistics”

Inferential (Çıkarımsal) Statistics

- Typically, it is not possible to reach, access, collect all possible data of interest
- Only a subset can be collected
- Make generalizations about the whole data only from a subset
- Assess the reliability of the generalizations
- This branch of statistics is called "inferential statistics"

Population

- **Population:** Entire collection of individuals, measurements etc. of interest.
- Term comes from human-related studies as the statistics is commonly used
- Typically defined by certain characteristics. For example:
  - Individuals eligible to vote
  - Women between 30-40 years old
- Populations can be defined using "inclusion criteria" or "exclusion criteria"
- In some cases population
  - Cannot be defined/known
  - Cannot be accessed/reached
- Sometimes multiple populations are considered, eg. obesity ratio in Europe vs America.

Population Parameter

- We are interested in some parameter(s) of the population, for example
  - Ratio of obesity
  - Mean height of new-born babies
  - Average reduction in collestrol due to a certain drug
- Population parameter(s) cannot be observed or measured directly due to limitations of cost, time etc.
- Population parameter(s) are estimated from a sample.

Sample

Especially in observational studies, it is not possible to collect data from whole population (due to time, cost etc.)

- **Sample:** A subset of population[2]
- **Sample size:** Number of cases in a sample $\rightarrow$ "Sample of size n"
- **Sampling:** Selection technique of a sample from the population
- **Sampling bias:** Wrong sampling strategies that result in incorrect generalizations from the sample about the population
  - Population is not well defined/listed (no accessible registry)
  - Various cost to reach each case (instance)
  - Not everyone agree to participate

[2] Notice that we typically use sample to denote a single data instance in machine learning.

Other Terms

- Variable: any characteristic or attribute that can take different values
- Observation/Instance/Measurement: A record containing many variables
- Instance/Datum: A single observation
- Data: A collection of observations
- Case: A person within a sample

Sampling

There are various sampling strategies

- Probabilistic methods
  - Simple random
  - Stratified random
  - Cluster random
  - Systematic
- Deterministic methods
  - Convenience ← Used to evaluate courses at ITU
  - Quota
  - Snowball
  - Chunk
  - Extreme cases
  - Judgement

These methods are used for different purposes

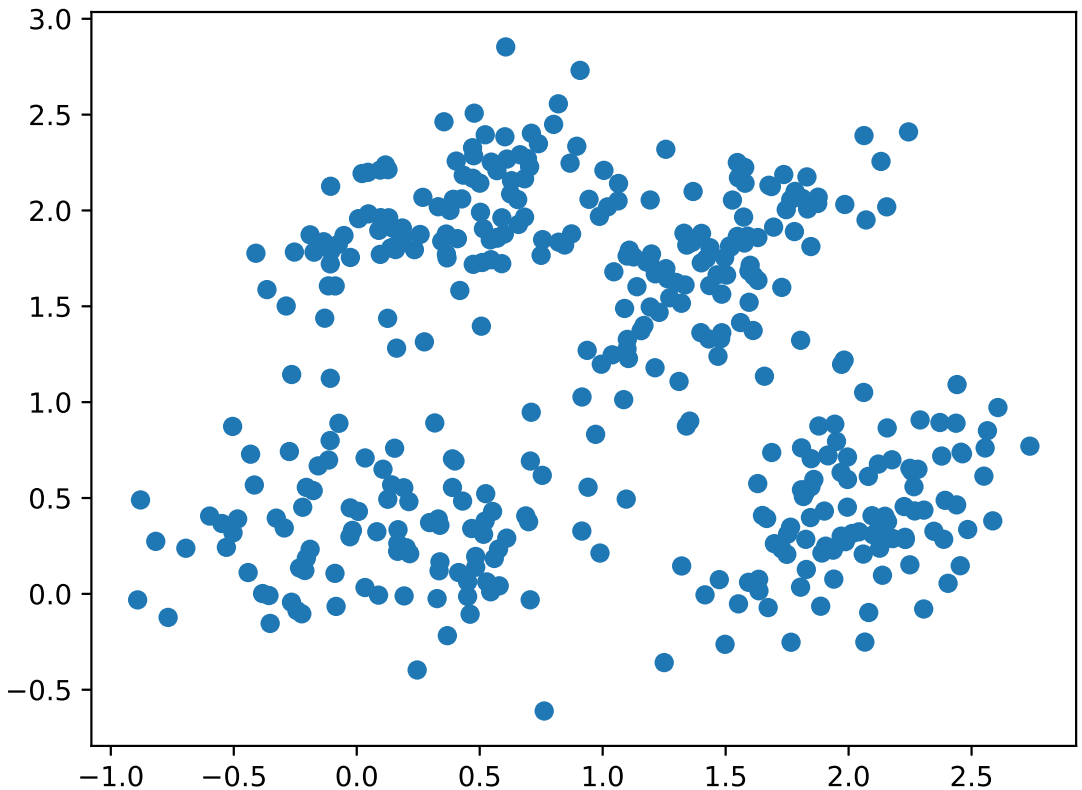
- Deterministic methods are used to build understanding, build a model/theory
- Probabilistic methods are used to describe population, test model/theory

Population

Following population is known and listed in a registry.

- For each instance there are two features in vertical/horizontal axis.
- There are 4 clusters in the population.

A sample of size (10%) will be formed.



Simple Random Sampling

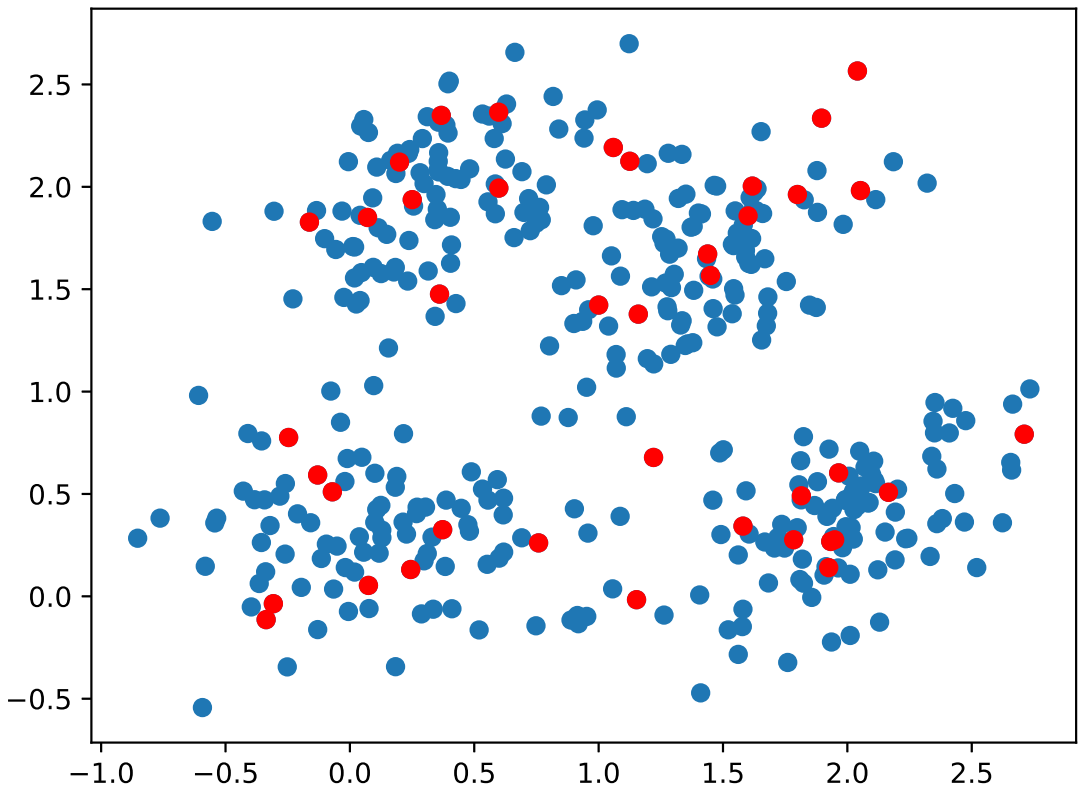
A sample is selected from a population in a way that ensures that every possible sample of size n has the same chance of begin selected.

- Advantages:
  - Very simple and representative
  - Sampling error is easy to calculate
- Disadvantages:
  - Need a list of whole population
  - May require high cost, time, logistic arrangements

Works quite well when the population is homogeneous.

Simple Random Sampling

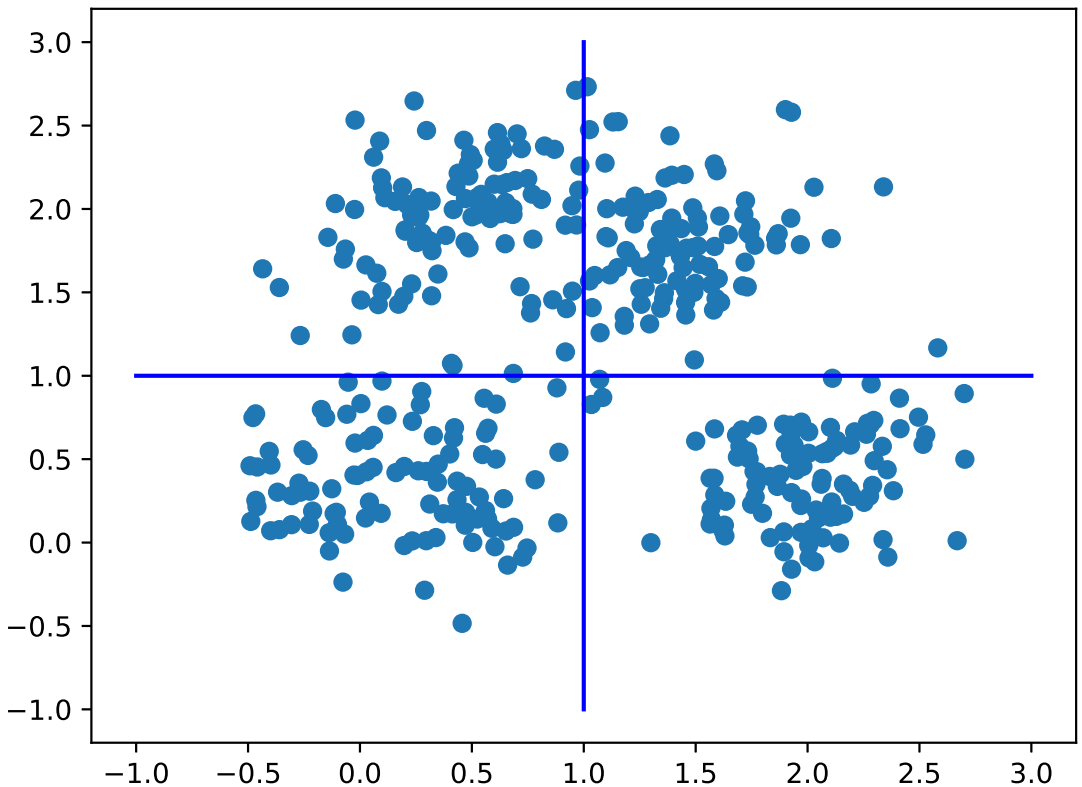
Red dots form our sample.



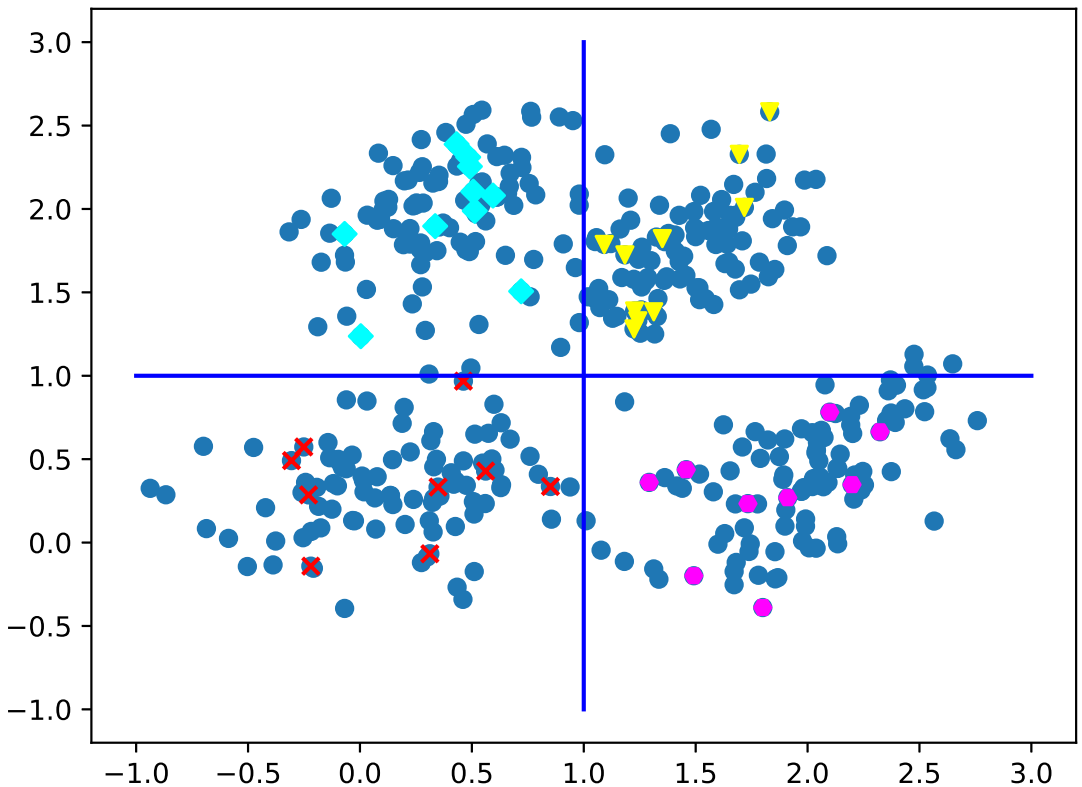
Stratified Random Sampling

- Population is divided into exclusive (non-overlapping) subgroups according to some characteristic such as gender, ethnicity, age etc.
- For example, customer behaviour requires strata according to their income. Income distribution is typically very skewed. Hence, wealthy customers are very few and may not be included into sample with simple random sampling.
- Each subgroup is called stratum (plural: strata).
- Simple random sample within each stratum
  - Advantages:
    - Better precision compared to simple random sampling
    - Better representation of stratifying variable (characteristics)
  - Disadvantages:
    - Need information about stratifying variable
    - Same disadvantages of simple random sampling within each stratum

Stratified Random Sampling

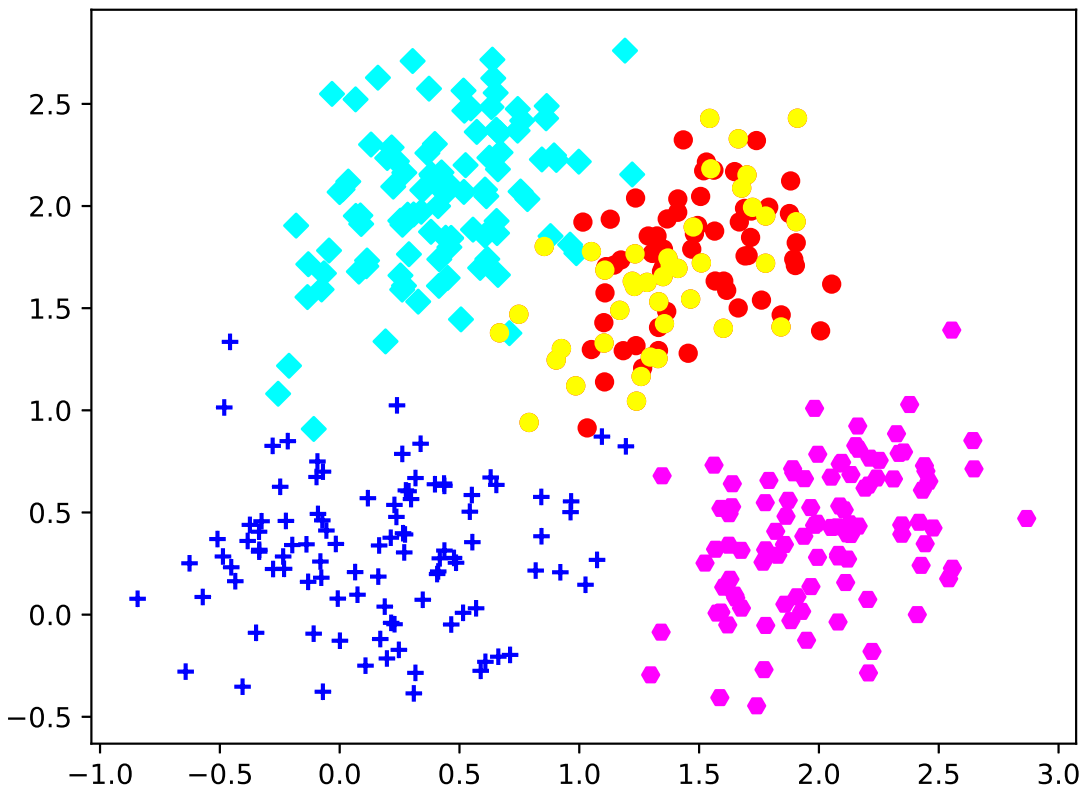


Stratified Random Sampling



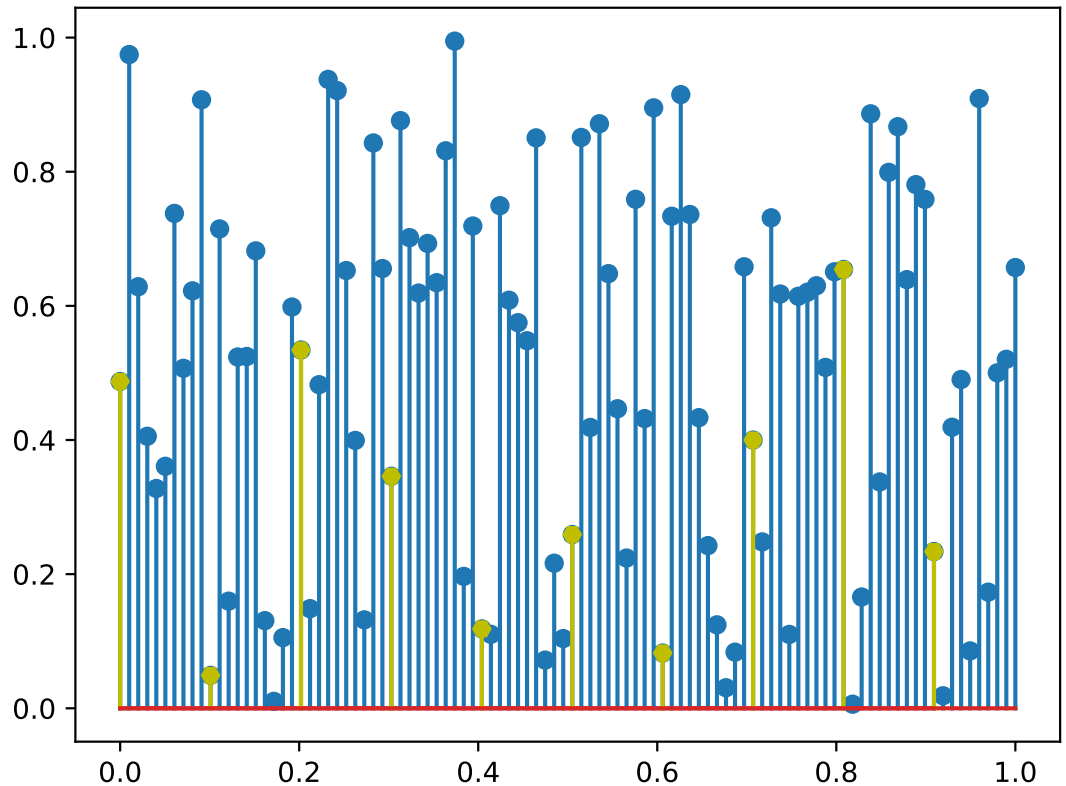
Cluster Random Sampling

- Cluster data
- Choose a random cluster
- Use simple random sampling from the selected cluster



Systematic Sampling

- Choose every n^{th} instance into to the sample
- Especially preferred for streaming data
- Dangerous if there is a pattern in the data!



Type of Studies – Observational vs. Experimental

- **Observational studies:** There is no interference, just observation.
  Typically used in medicine, social sciences, manufacturing etc.
  Used to detect associations between variables. **Association does not mean causality** due to confounding variable(s).
- **Experimental studies:** There are controlled variables (factors).
  Typically used in engineering, medicine [3]
  Used to detect cause-effect between variables and outputs.

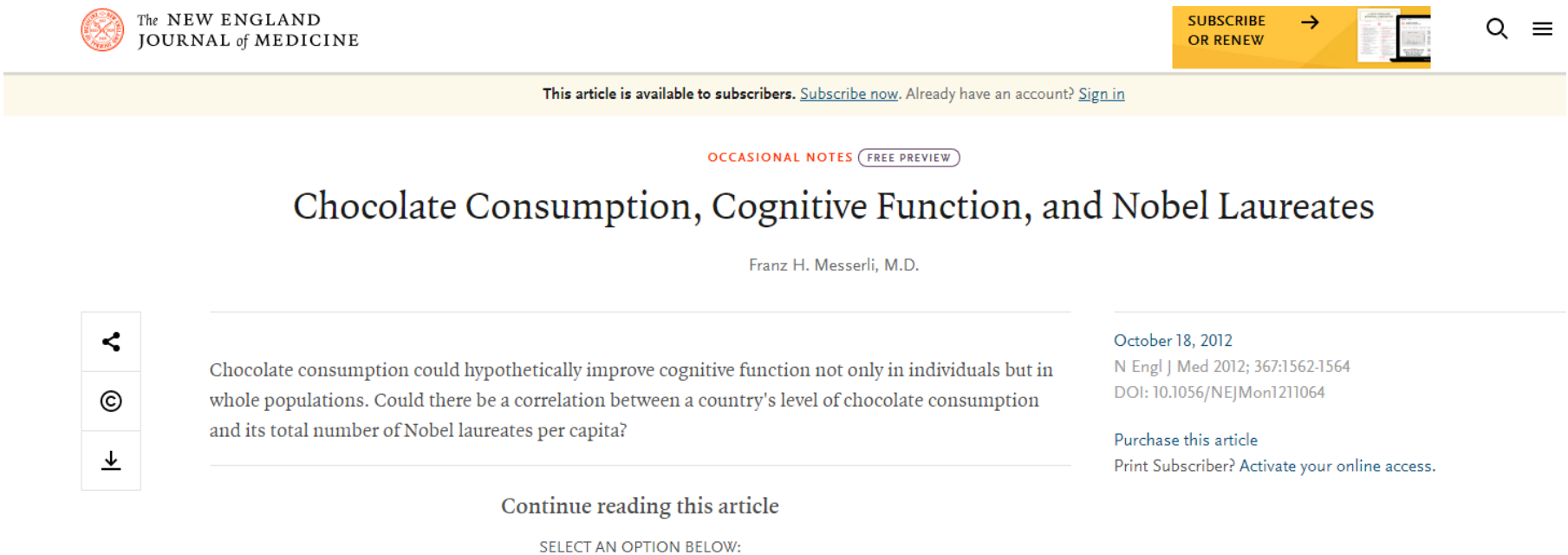
[3] term "trial" is used as it feels unethical to experiment on human health.

Association is NOT causality

Association does not mean causality due to confounding variable(s).
Confounding are typically hidden (unconsidered) variables that are related to all associated variables.

- High correlation between ice-cream consumption and crime rate.
- Weather is a confounding variable that affect both ice-cream consumption and crime-rate.
- Cold weather decrease ice-cream consumption & people typically stay at home.

Confounding variable



The NEW ENGLAND JOURNAL of MEDICINE

SUBSCRIBE OR RENEW

This article is available to subscribers. Subscribe now. Already have an account? Sign in

OCCASIONAL NOTES FREE PREVIEW

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

Chocolate consumption could hypothetically improve cognitive function not only in individuals but in whole populations. Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?

Continue reading this article

SELECT AN OPTION BELOW:

October 18, 2012
N Engl J Med 2012; 367:1562-1564
DOI: 10.1056/NEJMon1211064

Purchase this article
Print Subscriber? Activate your online access.

Confounding variable

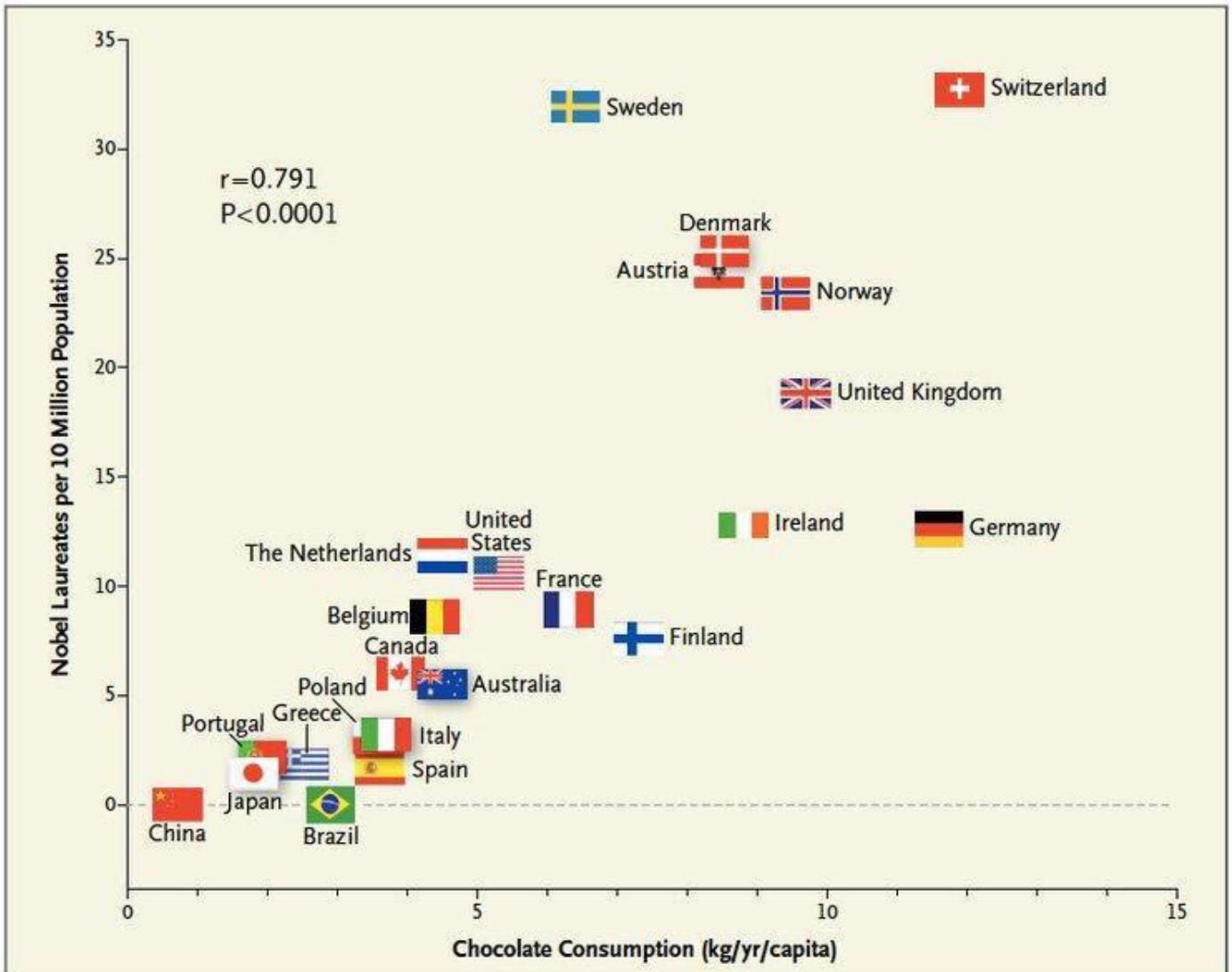
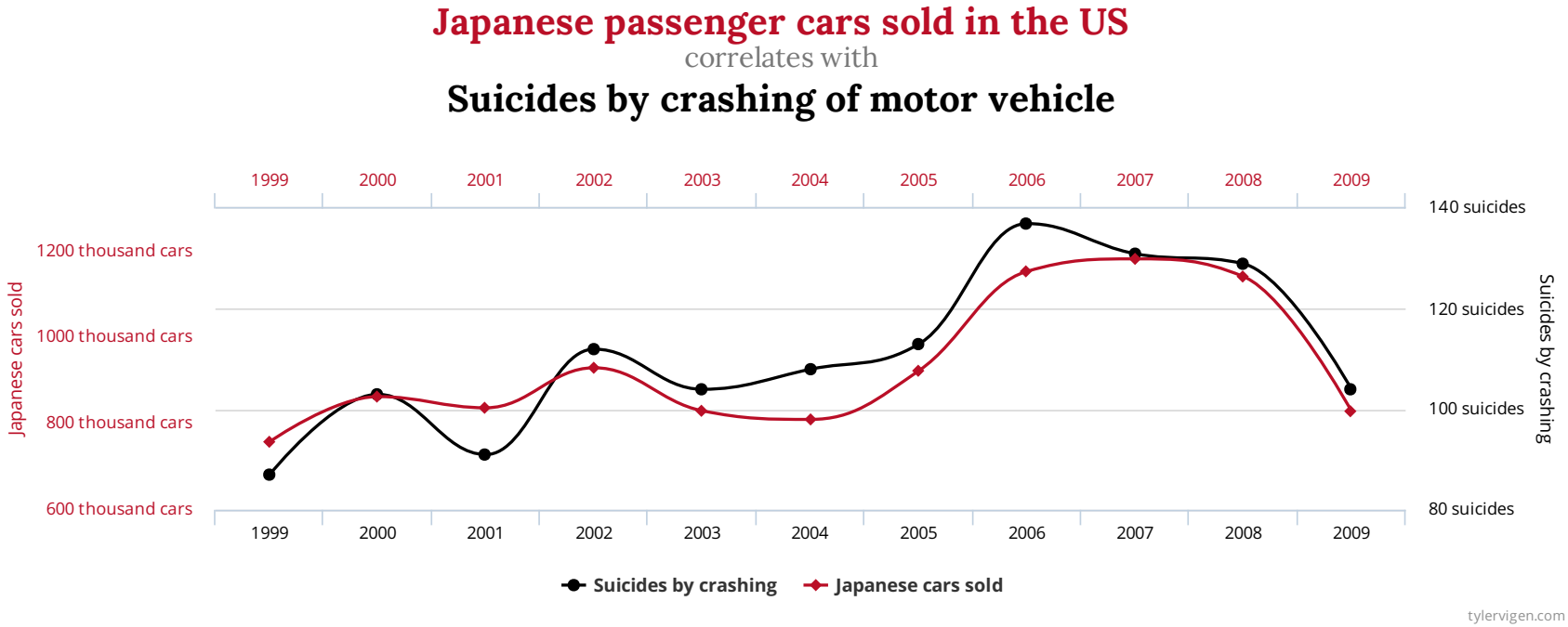


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

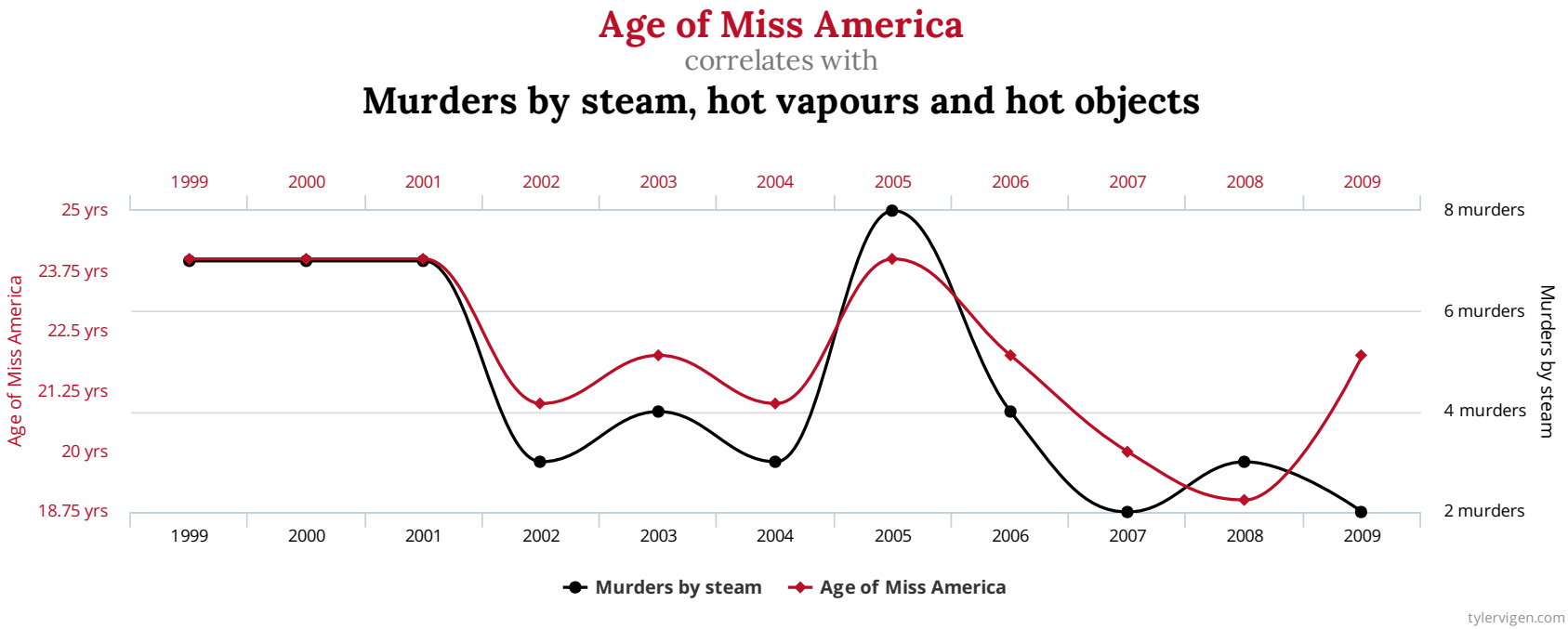
Do NOT push data

- Remember: any parameter obtained from sample is a random variable
- If you try to obtain too many parameters from the same data, you will definitely find spurious associations between unrelated data.



taken from `http://tylervigen.com/`

Do NOT push data



more at `http://tylervigen.com/`

Parameter Estimation

- Parameter estimate (sometimes called sample parameter) is computed from a sample using various methods.
- Two major issues in parameter estimation
  - How accurate is the estimate $\rightarrow$ bias
  - How precise is the estimate $\rightarrow$ variation

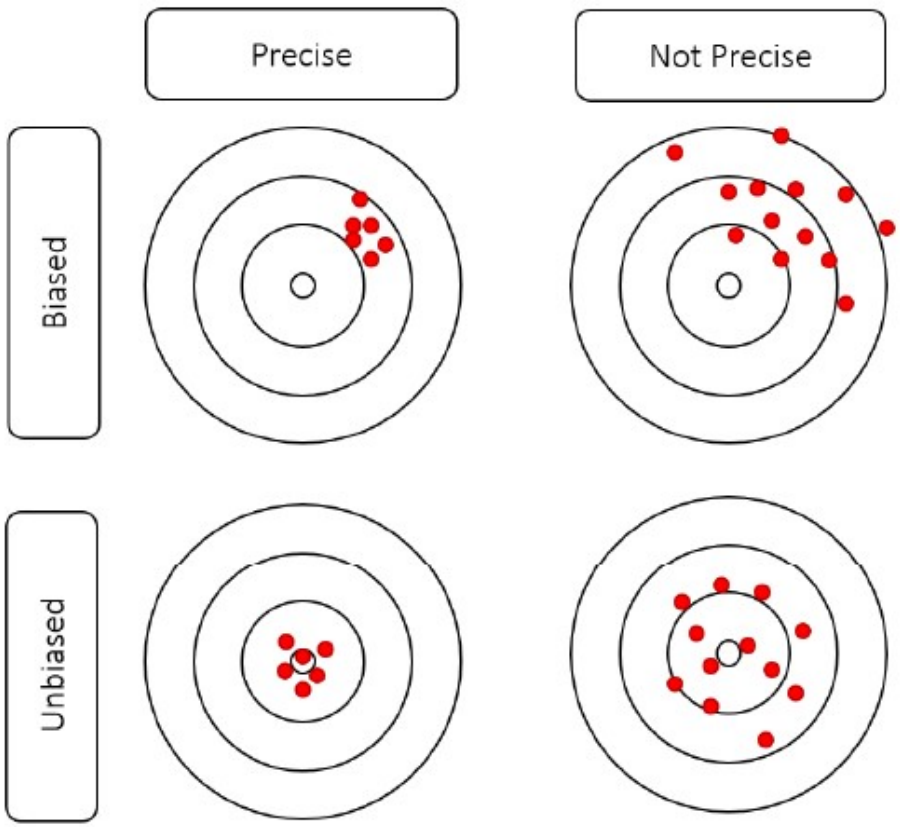


Figure taken from https://www.skuscience.com/blog/how-to-improve-your-sop-thanks-to-forecast-kpi/

Parameter Estimation

- Accuracy of the estimation is typically related to good sampling and good sample representation for population.
- Poor generalization of population $\rightarrow$ large bias / low accuracy.

- Precision of the estimation is typically related to sample size.
- Small sample size $\rightarrow$ low precision / high variability.

Parameter Estimation

What do we want to estimate about a parameter of interest?

- Population distribution (not easy, needs too much data) Instead investigate the parameters of distribution

- Central tendency $\rightarrow$ Mean value
- Dispersion $\rightarrow$ Variation
- Symmetry $\rightarrow$ Skewness
- Flatness/peakedness $\rightarrow$ Kurtosis
- ...

Overview

Population

- Probability distribution
- Greek letters for parameters $(\mu, \sigma, \rho, \cdots)$
- Parameters are not random variable
- Parameters are unknown
- Impossible to measure/-collect/access/define

Sample

- Latin letters for parameters $(\overline{x}, s, r, \cdots)$
- Parameters are computed from data
- Parameters are random variables they have pdf (sampling distribution)
- With correct methods, can be good estimators of population parameters $\rightarrow$ Inference

Population versus Sample Parameters