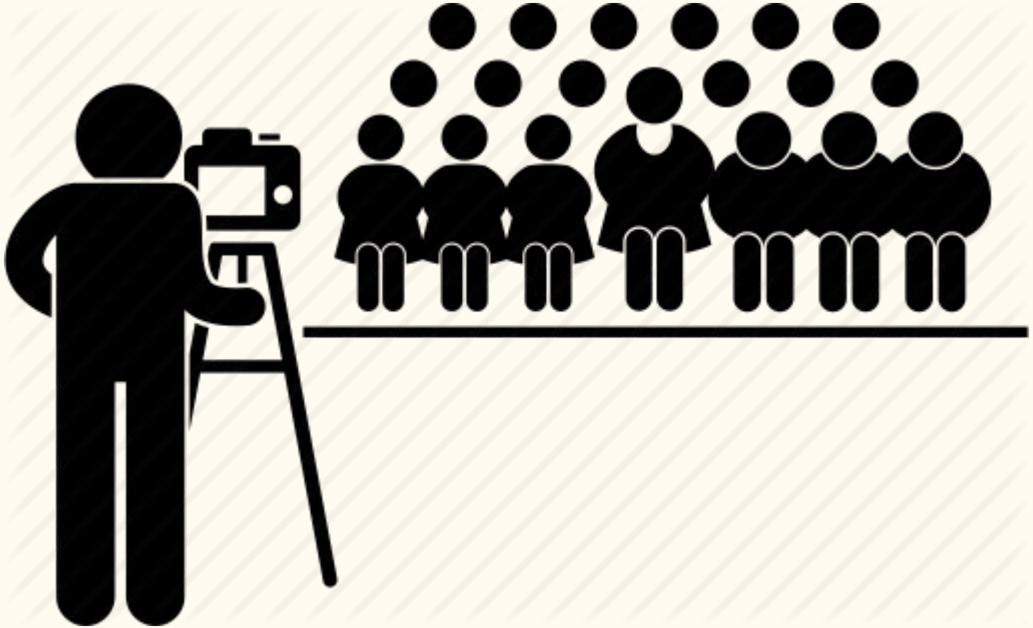
Applied Statistics 1st Lecture: Introduction to Statistics

By: Erika Siregar
Date: September 29, 2020
Venue: PKN STAN (online via zoom)

Rules of Conduct

1. Buka kelas
2. Absensi & Administrasi (mengisi portal)
   a. Hidupkan kamera + unmute + say ‘hadir’.
   b. Foto bersama → upload bukti
3. Materi
   a. Jika ada pertanyaan, gunakan fitur raise hand di zoom.
   b. Akan ada pertanyaan dadakan for random student.
   c. Quiz
   d. Weekly assignment.
   e. Self-study
4. Tutup kelas
5. Kelas akan di-record

ATTENDANCE + GROUP PICTURE



Today's Agenda

1. Kenalan
2. Discuss preliminary survey
3. Discuss main material: Intro to Statistics
4. No quiz or assignment for this week :)

Hello, I am Erika

- **Education:**
  - Bachelor of Applied Science from STIS
  - Master in Computer Science from Old Dominion University, US
- **Work:**
  - BPS: Data scientist, big data engineer and analyst
- **Communities:**
  - R-Ladies Jakarta : @rladiesjkt (IG)
  - Jakarta Machine Learning: @jkt.machinelearning (IG)
- **Connect with me:**
  - Email: erika.mukhlisina@gmail.com, erika@bps.go.id
  - GitHub: https://github.com/erikaris
  - Twitter: @erikaris

What is Statistics

- The science of **planning** studies and experiments, **obtaining data**, and then **organizing**, **summarizing**, **presenting**, **analyzing**, **interpreting**, and drawing **conclusions** based on the data.

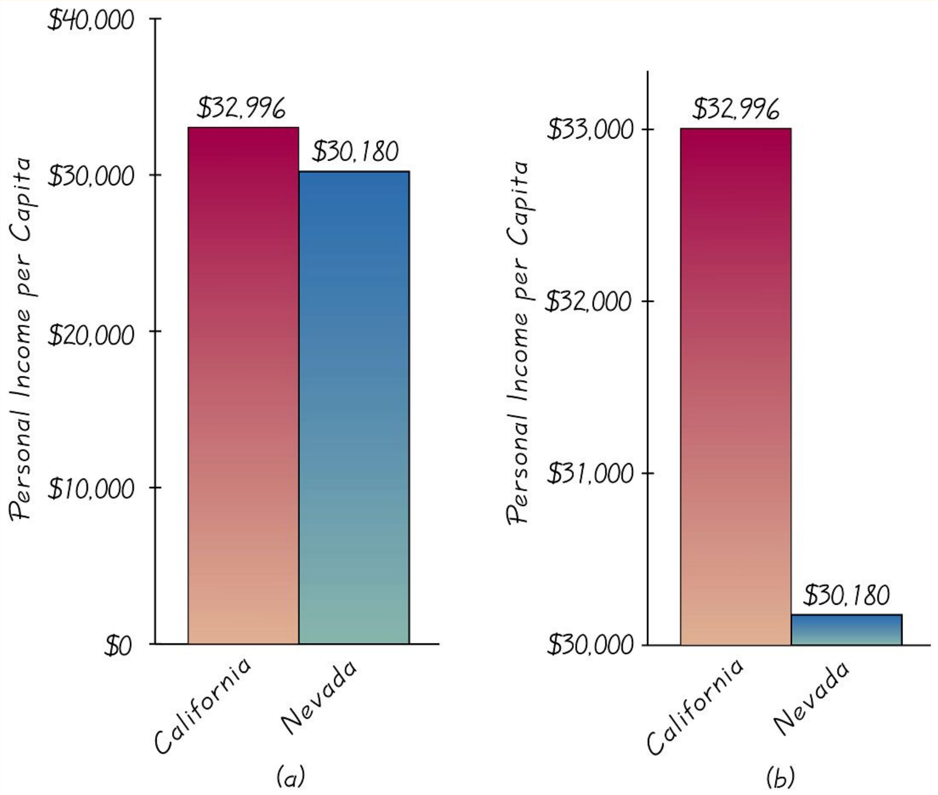
- The facts and figures

Let’s get the sense

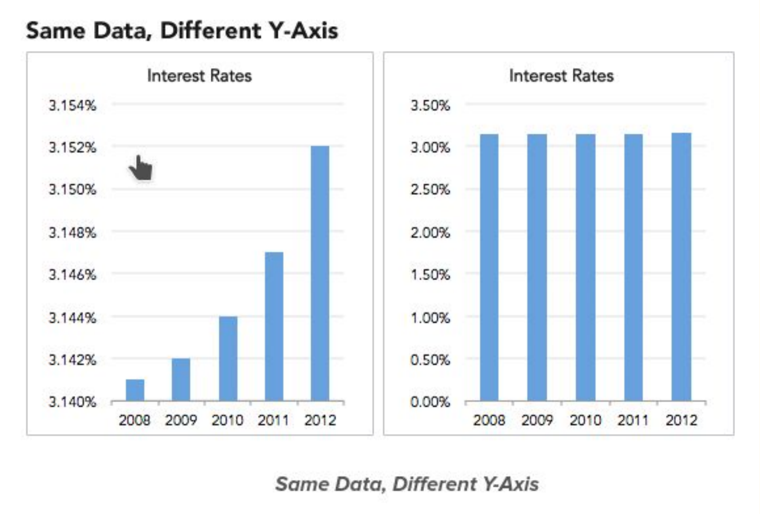
1. Melihat fenomena dan ingin tahu lebih banyak
2. Memutuskan apa yang ingin dicari tahu
3. Data Collection → **Scope: seberapa banyak? Semuanya? Sebagian saja? Kalau sebagian, bagaimana cara memilihnya?**
4. **Preprocessing** (Missing data, outliers, non response) dan **Processing** → (SPSS, R, excel)
5. Analisis dan sajikan (text + figure) → contoh analisis bisa refer to publikasi bps (https://bps.go.id/).

Key Concepts:

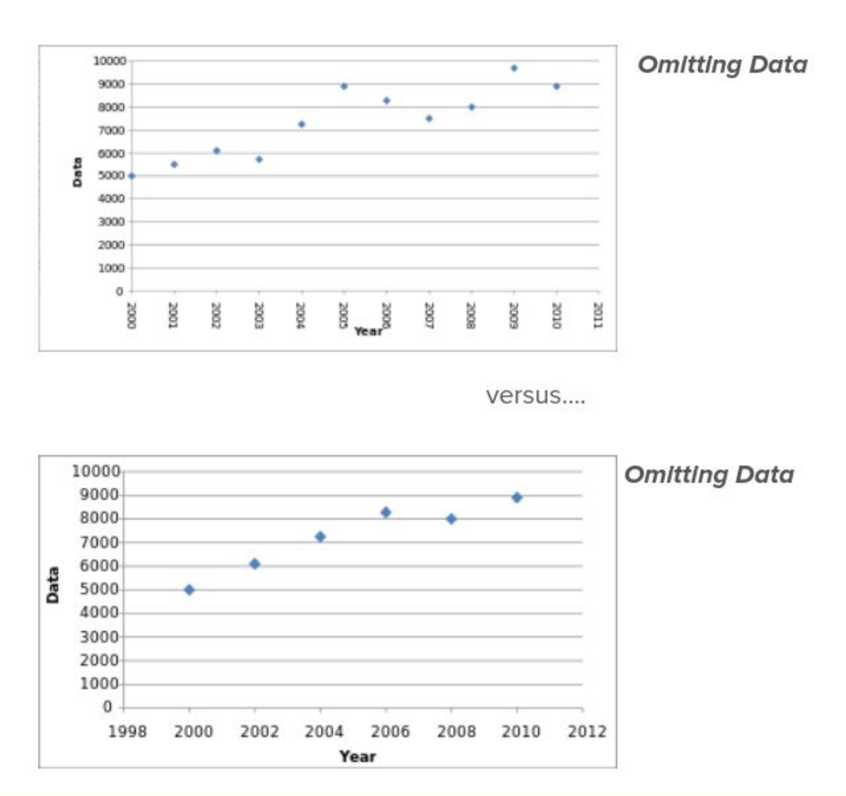
- Statistics requires more **common sense** than mathematical expertise.
- Skills in interpreting information based on data → **kemampuan membaca graph**



Misleading Figures

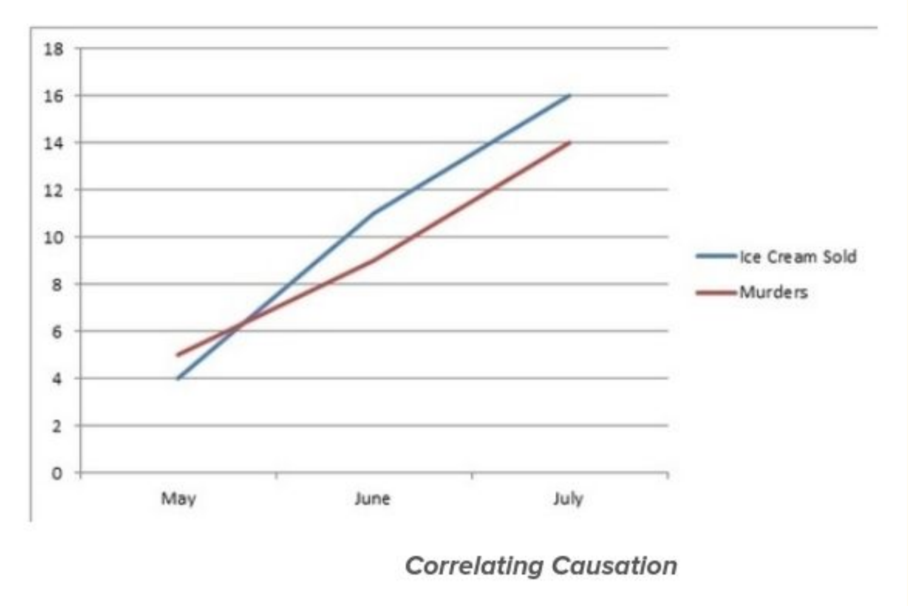


Same Data, Different Y-Axis

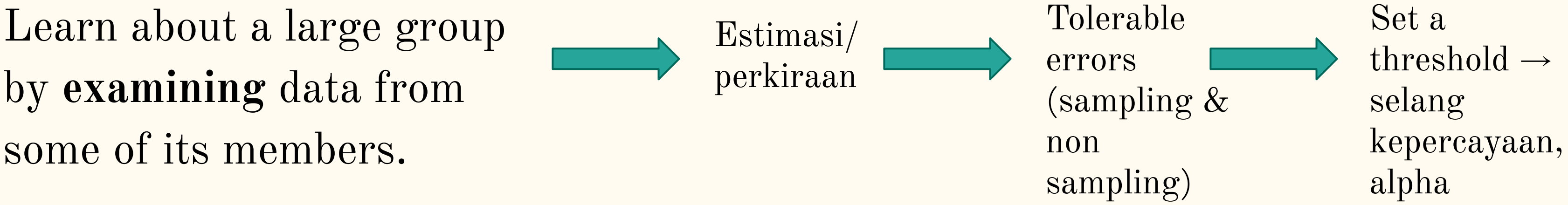


Create a non-exist trend

Misleading Figures



The main idea of Statistics:



Learn about a large group by **examining** data from some of its members.

Estimasi/ perkiraan

Tolerable errors (sampling & non sampling)

Set a threshold → selang kepercayaan, alpha

Data

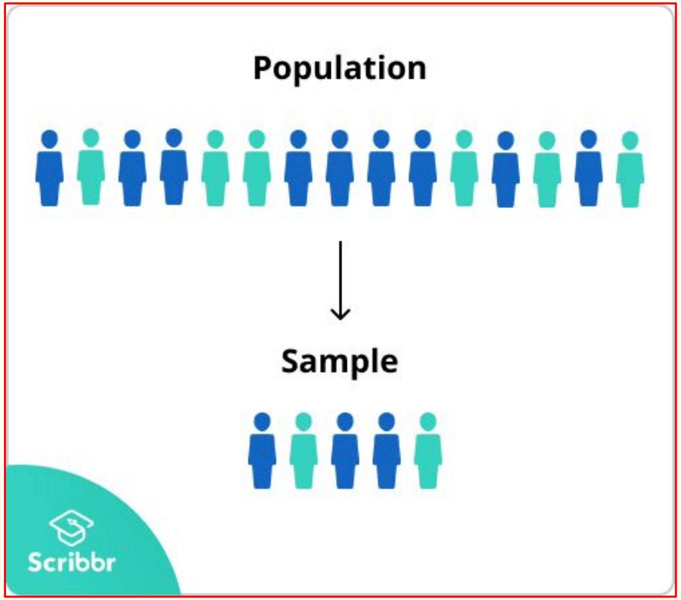
- collections of observations (such as measurements, genders, survey responses)
- Konvensional: tabel, dll
- Big data: social media data, data yang di-scrape dari website (e-commerce, transportasi)
- API → pintu masuk ke data twitter

Statistics Involves Data Collection

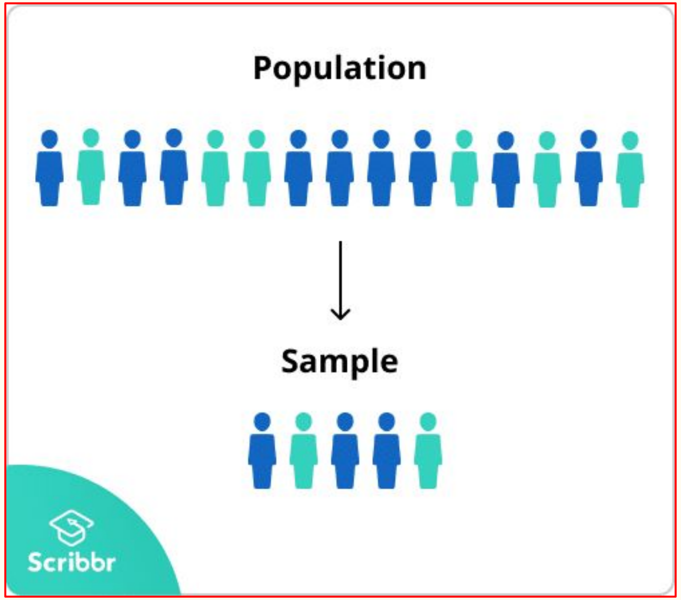
Collection: **kumpulkan data (Scope: seberapa banyak? Semuanya? Sebagian saja? Kalau sebagian, bgmn cara memilihnya)**

→ Ini melahirkan terminologi populasi

dan sampel



REPRESENTATIF → OTHERWISE, BIAS



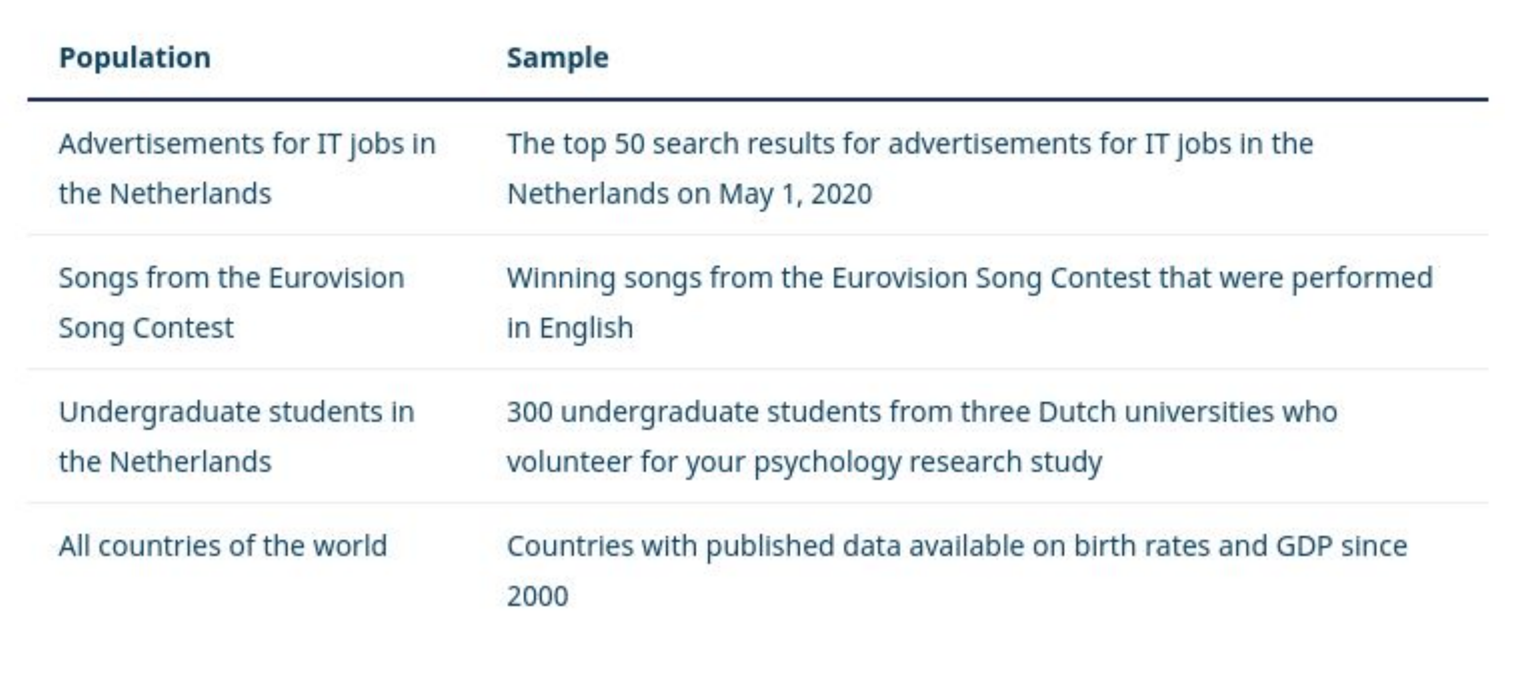
Population = complete collection of all individuals → **TAKE ALL**

Sampel = take some as a representation of the population→ **TAKE SOME**

How to take some? → sampling method

Notes:

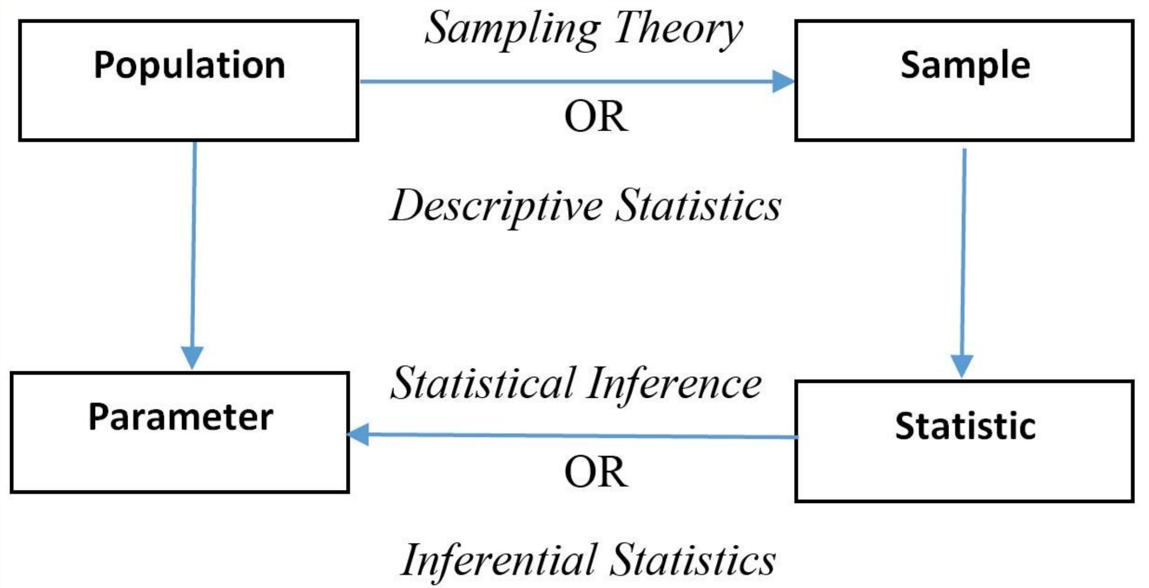
Sample data **must be collected in an appropriate way**, such as through a process of random selection.



Population	Sample
Advertisements for IT jobs in the Netherlands	The top 50 search results for advertisements for IT jobs in the Netherlands on May 1, 2020
Songs from the Eurovision Song Contest	Winning songs from the Eurovision Song Contest that were performed in English
Undergraduate students in the Netherlands	300 undergraduate students from three Dutch universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Populasi vs Sampel

No	Subject	Semua	Sebagian
1	Object observasi	Populasi (ALL)	Sampel (some)
2	Kegiatan	Sensus	Survei → polling, studies.
3.	Measurement	Parameter (mean, median, modus, variance, standar deviasi)	Statistic



Population

Sampling Theory

OR

Descriptive Statistics

Sample

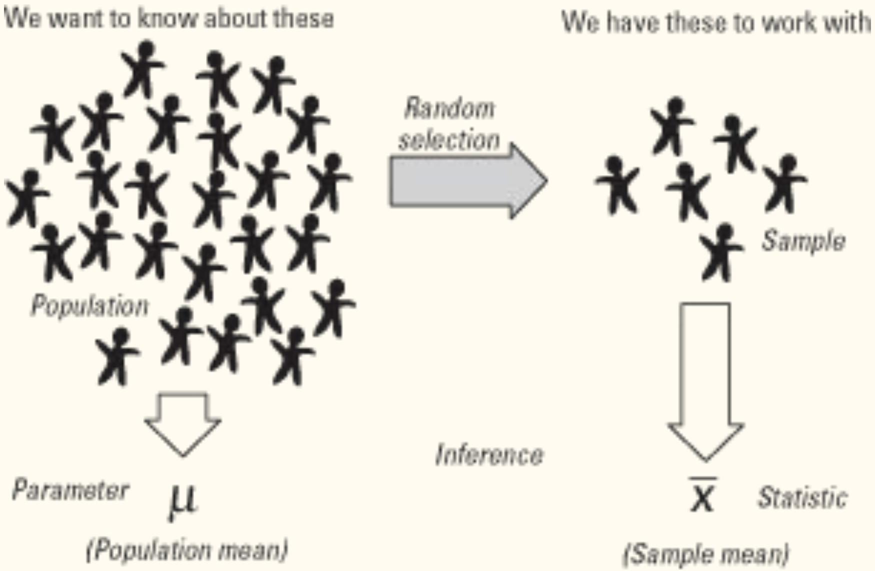
Statistic

Statistical Inference

OR

Inferential Statistics

Parameter



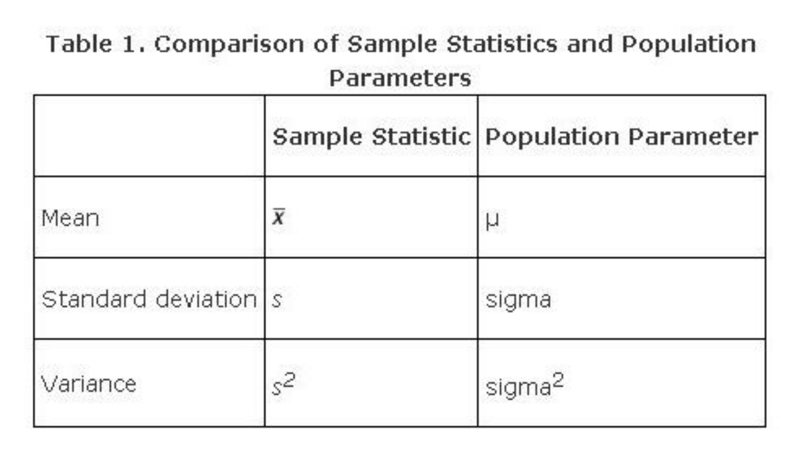
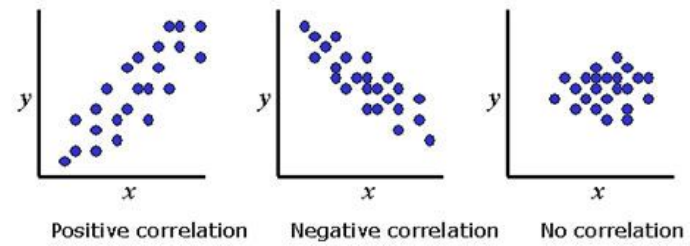


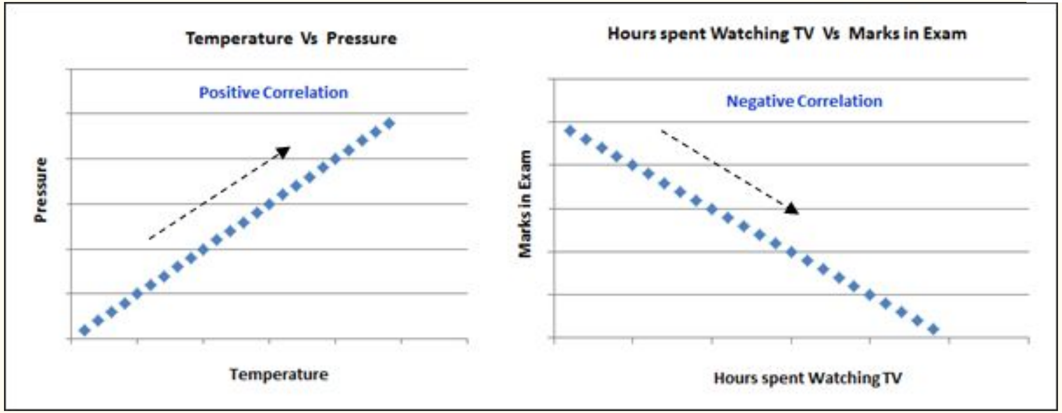
Table 1. Comparison of Sample Statistics and Population Parameters

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	*s*	sigma
Variance	s^2	sigma^2

Correlation

- Adanya hubungan/relasi/asosiasi antara 2 variabel atau lebih
  - Positif: searah → r = 1
  - Negatif: berlawanan arah -> r= -1
  - No correlation → r = 0





Understanding correlation is important in decision making

- Pertumbuhan ekonomi vs inflasi → korelasi positif
- Tingkat pendapatan vs konsumsi
- Jumlah kasus positif covid19 vs tingkat mobilitas penduduk

Data Types

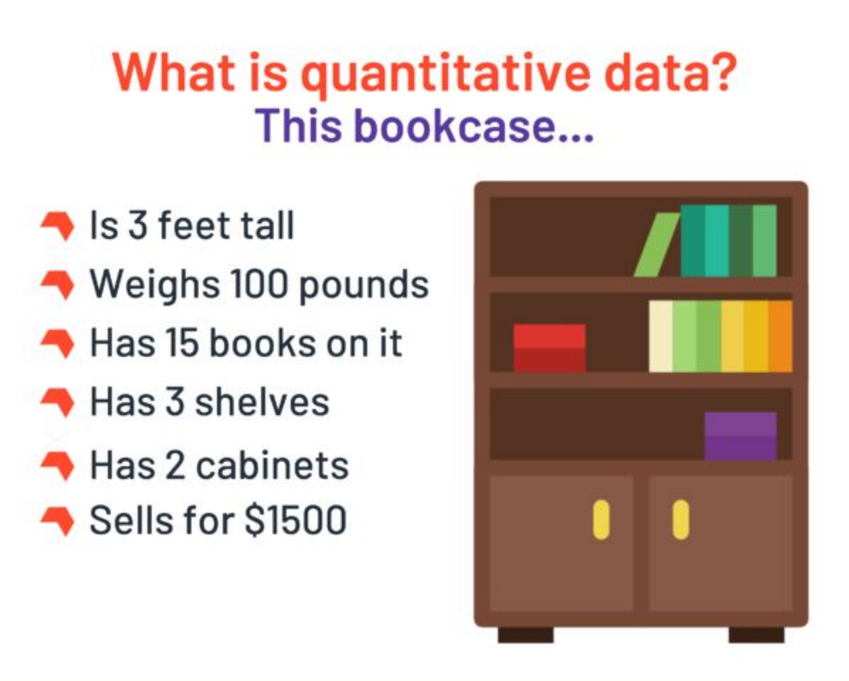
Quantitative

1. Hasil **penghitungan atau pengukuran.**
2. Dibedakan lagi menjadi
   a. **Discrete** → countable: 1, 2, 3, …
   b. **Continuous** → the opposite
3. Contoh:
   a. Nilai UTS Statistik Terapan kelas 5-37 dan 5-38 (D/C)?
   b. Jumlah pengeluaran rumah tangga untuk makan dalam satu minggu. (D/C)?
   c. Jumlah pasien positif covid19. (D/C)?

Qualitative/Categorical

1. Representasi dari kategori
2. Contoh:
   a. Jenis kelamin students di kelas 5-37 dan 5-38
   b. Kondisi koneksi internet students di kelas 5-37 dan 5-38

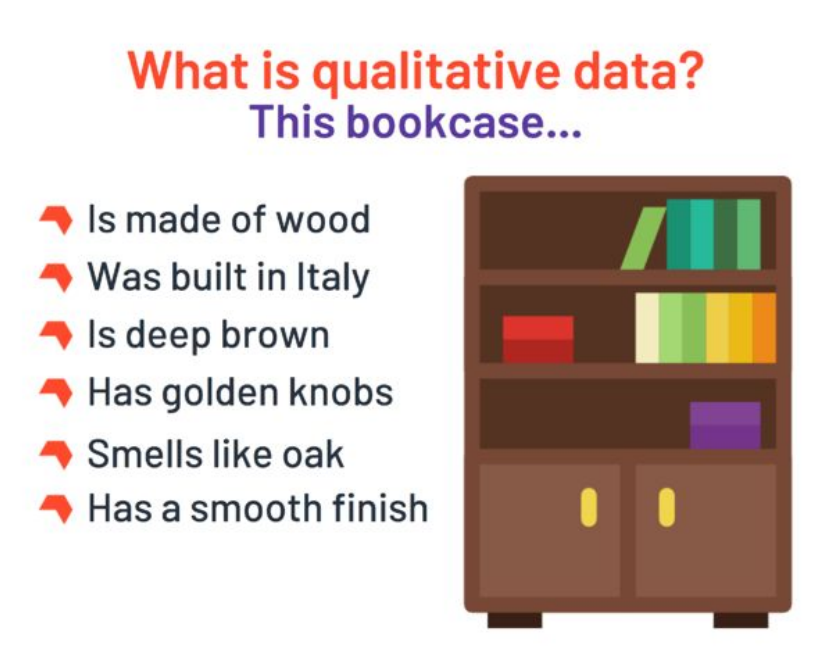
Illustration



What is quantitative data?

This bookcase...

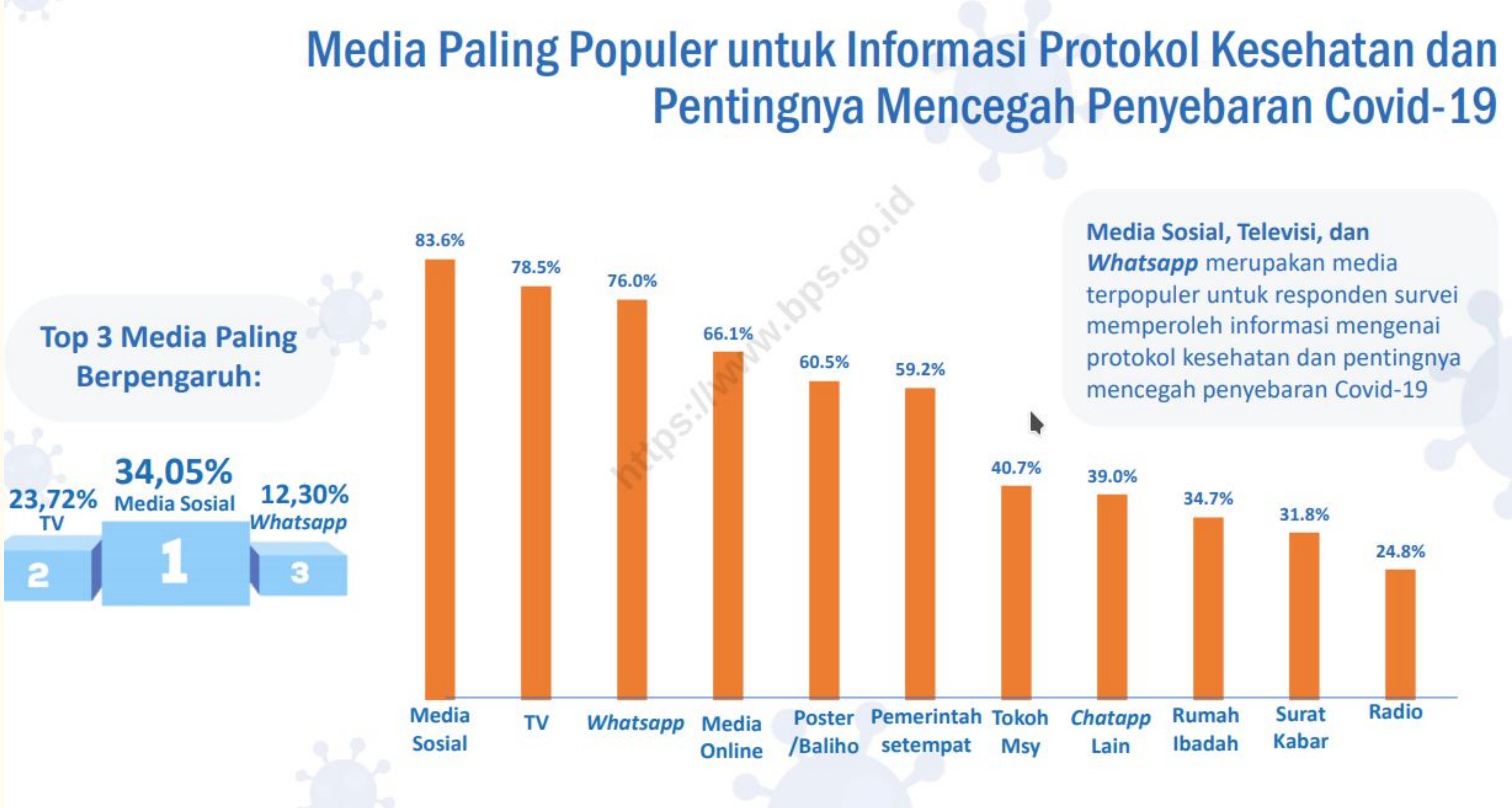
- Is 3 feet tall
- Weighs 100 pounds
- Has 15 books on it
- Has 3 shelves
- Has 2 cabinets
- Sells for $1500



What is qualitative data?

This bookcase...

- Is made of wood
- Was built in Italy
- Is deep brown
- Has golden knobs
- Smells like oak
- Has a smooth finish



Media Paling Populer untuk Informasi Protokol Kesehatan dan Pentingnya Mencegah Penyebaran Covid-19

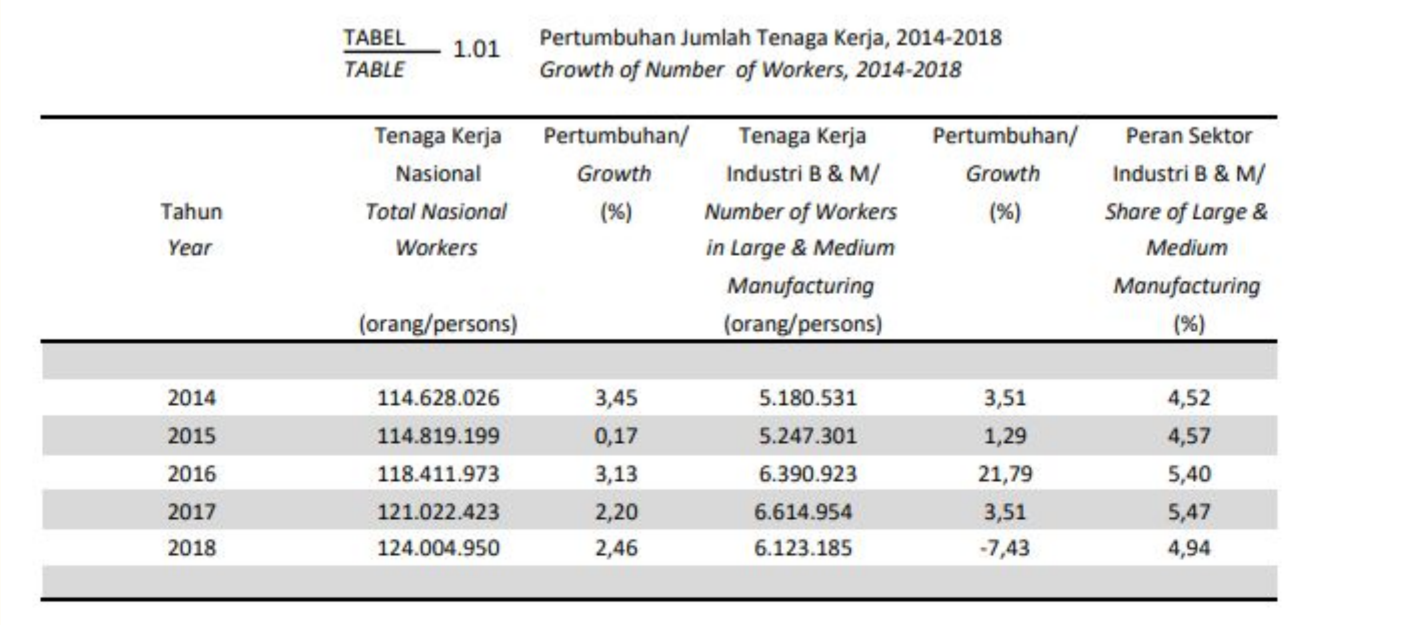
Top 3 Media Paling Berpengaruh:

34,05%
Media Sosial

23,72%
TV

12,30%
Whatsapp

Media Sosial, Televisi, dan *Whatsapp* merupakan media terpopuler untuk responden survei memperoleh informasi mengenai protokol kesehatan dan pentingnya mencegah penyebaran Covid-19



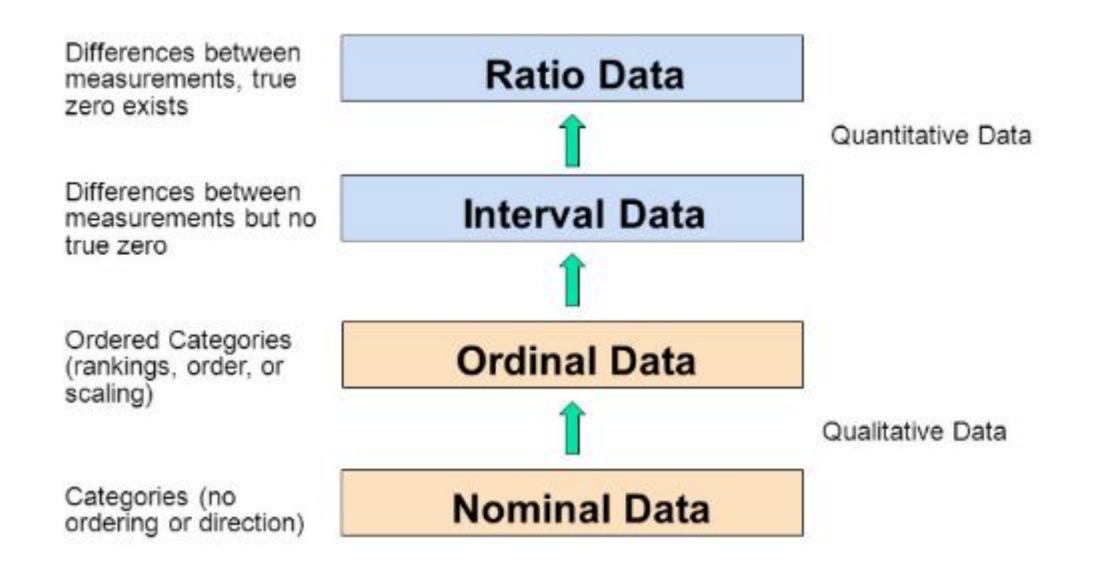
TABEL 1.01 Pertumbuhan Jumlah Tenaga Kerja, 2014-2018

TABLE *Growth of Number of Workers, 2014-2018*

Tahun *Year*	Tenaga Kerja Nasional *Total Nasional Workers* (orang/persons)	Pertumbuhan/ *Growth* (%)	Tenaga Kerja Industri B & M/ *Number of Workers in Large & Medium Manufacturing* (orang/persons)	Pertumbuhan/ *Growth* (%)	Peran Sektor Industri B & M/ *Share of Large & Medium Manufacturing* (%)
2014	114.628.026	3,45	5.180.531	3,51	4,52
2015	114.819.199	0,17	5.247.301	1,29	4,57
2016	118.411.973	3,13	6.390.923	21,79	5,40
2017	121.022.423	2,20	6.614.954	3,51	5,47
2018	124.004.950	2,46	6.123.185	-7,43	4,94

Sumber: BPS

Data Measurement Level



Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

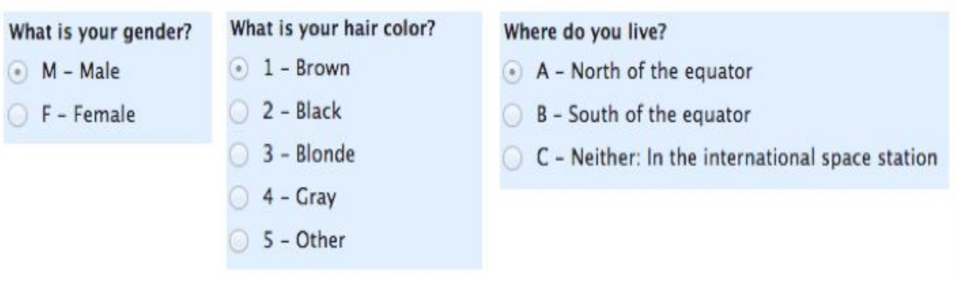
Qualitative Data

Categories (no ordering or direction)

Nominal Data

Types of data measurement

1. Nominal
   a. Categories only, no order
   b. Example:



What is your gender?

● M – Male

○ F – Female

What is your hair color?

● 1 - Brown

○ 2 - Black

○ 3 - Blonde

○ 4 - Gray

○ 5 - Other

Where do you live?

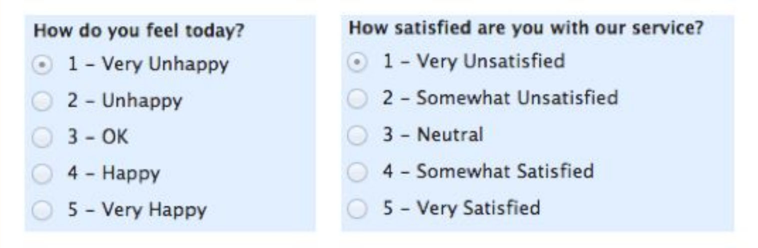
● A – North of the equator

○ B – South of the equator

○ C – Neither: In the international space station

2. Ordinal = nominal + order

   a. Ordered category



How do you feel today?

● 1 – Very Unhappy

○ 2 – Unhappy

○ 3 – OK

○ 4 – Happy

○ 5 – Very Happy

How satisfied are you with our service?

● 1 – Very Unsatisfied

○ 2 – Somewhat Unsatisfied

○ 3 – Neutral

○ 4 – Somewhat Satisfied

○ 5 – Very Satisfied

Types of data measurement

3. Interval = ordinal + ada selisih

a. **similar to ordinal** but the differences or **intervals between values** are equal/constant.
⇒ **kita bisa menyatakan dengan 'terang' berapa selisih antara 2 data.**

b. **Does not have true 0 point.**
   i. addition/substraction = yes
   ii. multiplication/division = no

Example:

1. **Temperature** → berapa selisih antara 20 C dengan 10 C? → 58 F vs 38 F
   There is no true zero because temperature can go into the negatives. Zero is just another point of measurement.
2. **Tahun** → 1000, 2000, 2000 BC

Negative → 0 → positif

Types of data measurement

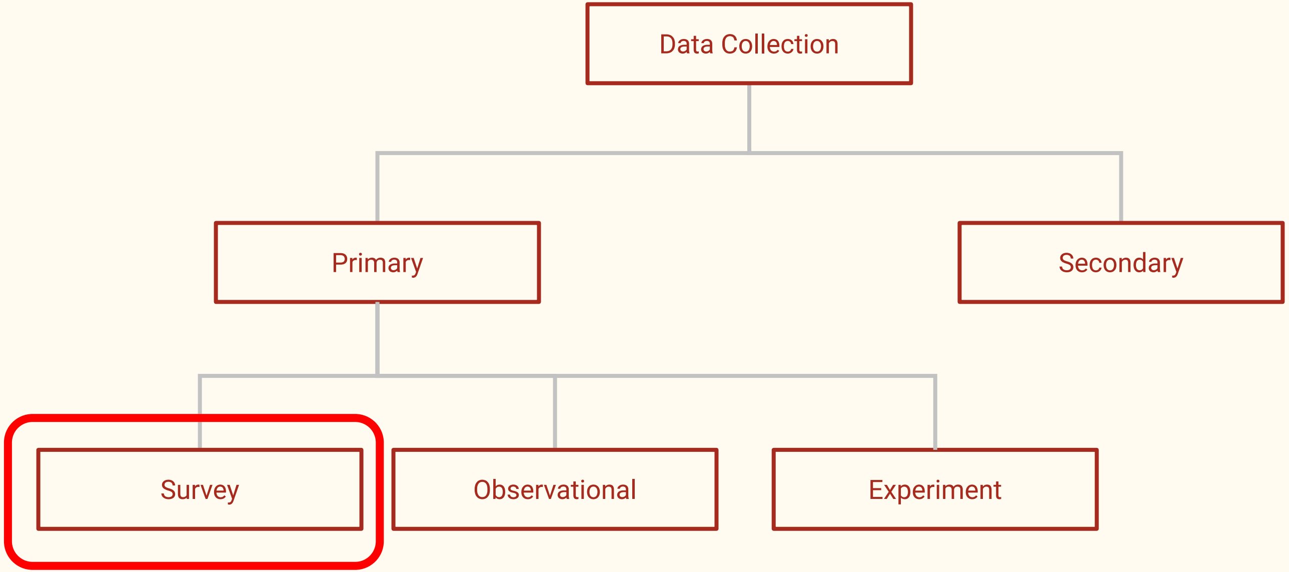
4. Ratio = interval + true 0 point

a. Similar to interval, but
b. Has true zero
c. So we can compute the ratio
d. Example:
   i. Prices of college textbooks ($0 represents no cost, a $100 book costs twice as much as a $50 book)
   ii. Number of unemployed people

0 → positif

True 0 = memang benar2 tidak ada

Data Collection



How to Choose Sample?

1. Simple Random Sampling

Memilih sejumlah n sampel, dimana setiap unit/individu **memiliki peluang yang sama untuk terpilih.**

Simple Random Sampling

How to Choose Sample?

2. Systematic Sampling

Select some starting point and then select **every kth** element in the population

.

How to Choose Sample?

3. Convenience Sampling

Take samples from a group of people easy to contact or to reach.

It’s **prompt, uncomplicated, and economical.**

Example:

- Asking people in the street about who will win the election
- Asking people in the mall about the best coffee in the world

https://youtu.be/aomNbRO5Zac

How to Choose Sample?

Konsumsi RT

1. RT kaya → sampel
2. RT sedang → sampel
3. RT miskin → sampel

4. Stratified Sampling

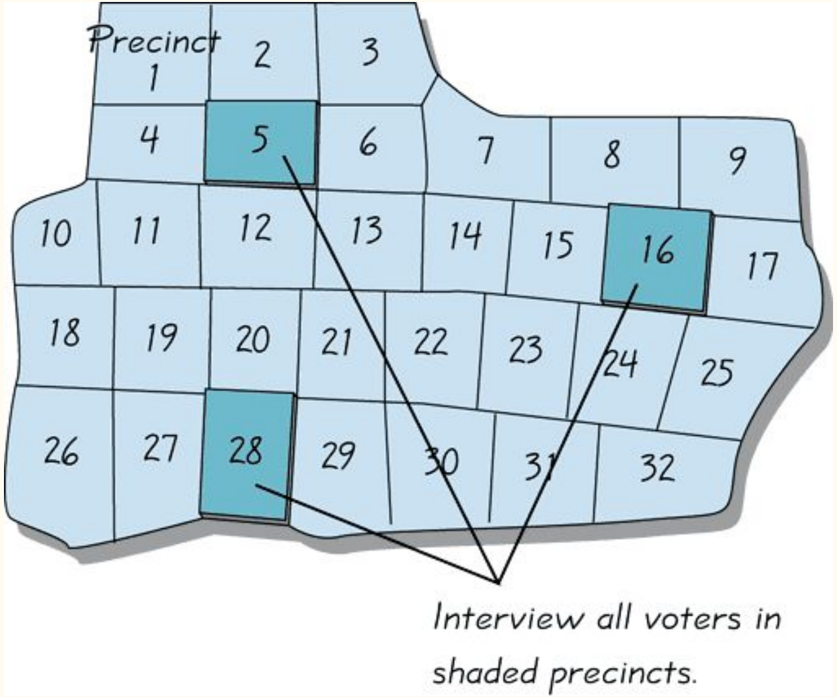
Divide the population into homogenous subgroups (strata). Then, from each strata take n samples.

How to Choose Sample?

4. Clustered sample

Divide the population into **heterogenous naturally-formed** subgroups (cluster). Then, take n clusters.

All individuals in the selected clusters will be used as samples.



How to Choose Sample?

4. Multistage sampling

Pemilihan sample yang dilakukan dalam lebih dari 1 tahap.

Misal:

1. Tahap pertama: memilih sampel kabupaten
2. Tahap kedua: memilih rumah tangga yang akan disampel.

Fun quiz:

Open kahoot.it on your browser

Terima kasih