

Exploratory Data Analysis

Apa Itu Exploratory Data Analysis

Definisi

- Secara definitif, Exploratory Data Analysis mengacu pada proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, untuk menemukan anomali, untuk menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis

Manfaat EDA

- Dapat lebih memahami kondisi dataset yang kita miliki
- dapat memulai pembentukan model Machine Learning dengan lebih baik kedepannya
- Memahami kondisi dataset

Mengapa EDA penting

- Tanpa melakukan Exploratory Data Analysis, kita bisa saja kehilangan banyak informasi penting yang terdapat di dalam dataset
- EDA akan menghemat waktu pengerjaan proyek Data Science

Sumber Data

Internal sources

Spreadsheets (Excel, CSV, JSON, etc.)

Databases: can be queried via SQL, etc.

Text documents

Multimedia documents (audio, video)

External sources

Open data repositories

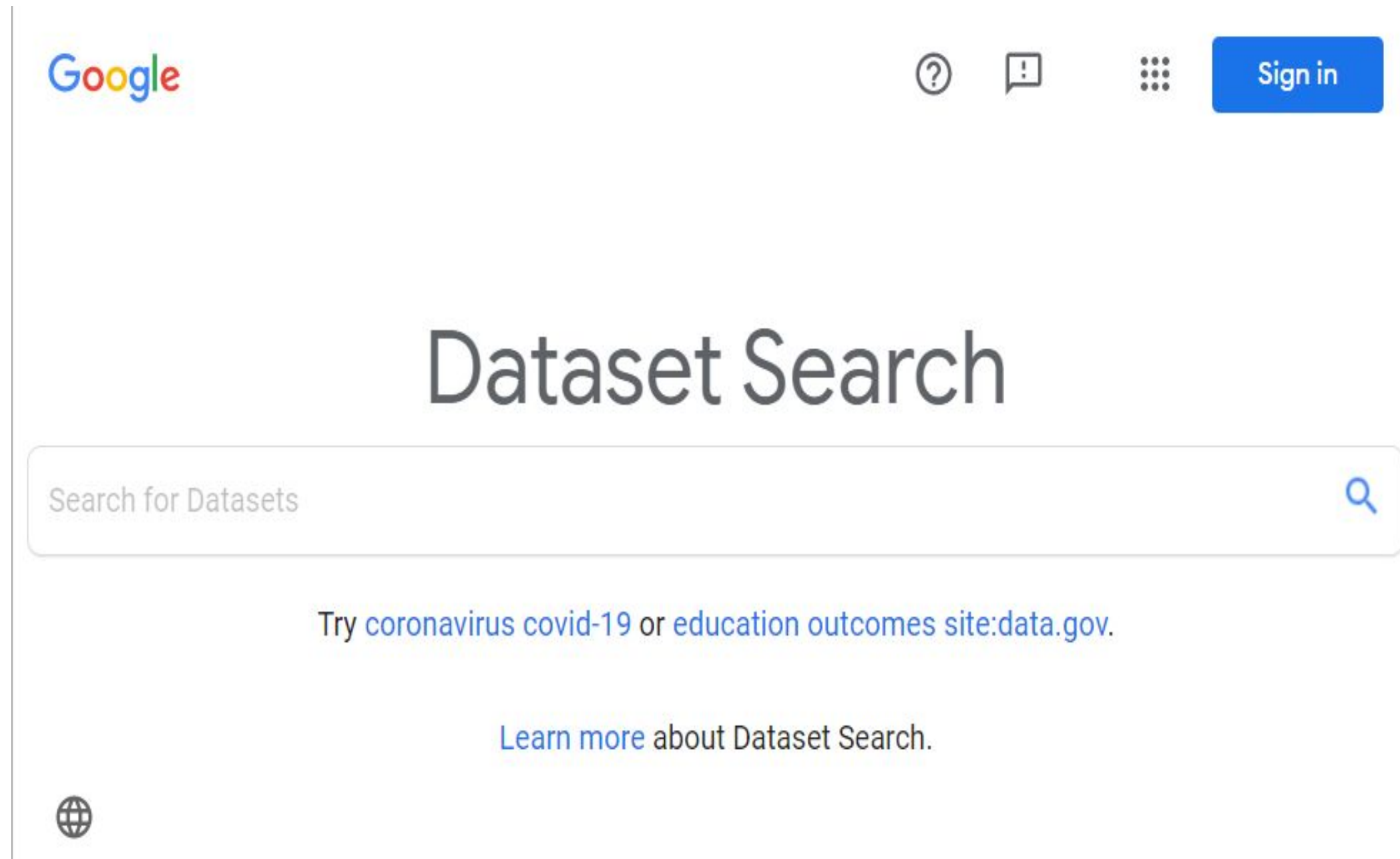
Public domain web pages

Sumber Data Daring

- Portal Satu Data Indonesia (<https://data.go.id>)
- Portal Data Jakarta (<https://data.jakarta.go.id>)
- Portal Data Bandung (<http://data.bandung.go.id>)
- Badan Pusat Statistik (<https://www.bps.go.id>)
- Badan Informasi Geospasial (<https://tanahair.indonesia.go.id/>)
- UCI Machine Learning repository (<https://archive.ics.uci.edu/ml/index.php>)
- Kaggle (<https://www.kaggle.com/datasets>)
- World Bank Open Data (<https://data.worldbank.org>)
- UNICEF Data (<https://data.unicef.org>)
- WHO Open Data (<https://www.who.int/data>)
- IBM Data Asset eXchange (<https://developer.ibm.com/exchanges/data/>)
- DBPedia (<https://www.dbpedia.org/resources/>)
- Wikidata (<https://www.wikidata.org/>) .

Sumber Data Daring

- Cari via Google Dataset Search: <https://datasetsearch.research.google.com>



Susunan Data

Butir data (*datum*): satuan terkecil data; satu nilai untuk satu variable tertentu

Data: kumpulan butir data yang membawa satu kesatuan makna (mendeskripsikan satu objek) tertentu.

Himpunan data (*dataset*): kumpulan data.

Metadata: data yang menjelaskan data yang lain.

symboling	normalized-losses	make	fuel-type
3 ?		alfa-romero	gas
3 ?		alfa-romero	gas
1 ?		alfa-romero	gas
2	164	audi	gas
2	164	audi	gas

"make":

- tipe: string,
- deskripsi: nama pabrikan merek kendaraan

Tipe Data berdasarkan susunannya

	Data terstruktur (structured data)	Data takterstruktur (unstructured data)
Sifat	<ul style="list-style-type: none">• Model data terdefinisikan sebelumnya• Format butir data (biasanya) teks.• Antar butir data terbedakan dengan jelas.• Ekstraksi/kueri langsung cukup mudah.	<ul style="list-style-type: none">• Model data tidak terdefinisikan sebelumnya• Format butir data (biasanya) teks, citra, suara, video, dan format lainnya.• Antar butir data tidak cukup jelas terbedakan karena ketidakteraturan dan ambiguitas.• Ekstraksi/kueri langsung cukup sulit.
Contoh	Data tabular, data berorientasi objek, <i>time series</i>	Data teks dalam dokumen teks bebas, data audio, data video.

Data semi-terstruktur (*semi-structured data*): Data terstruktur yang tidak mengikuti model struktur tabular yang seperti pada basis data relasional, namun tetap mengandung *tags* atau penanda lainnya yang dapat memisahkan elemen-elemen semantik pada data serta mengatur hierarki antara butir-butir datanya.

Tipe butir data (1)

	Nominal/kategori kal	Ordinal	Interval	Rasio
Sifat himpunan asal	Diskret, tidak terurut	Diskret, terurut	Kontinu/numerik, terurut, perbedaan menunjukkan selisih	Kontinu/numerik, terurut, nilai menunjukkan rasio terhadap kuantitas satuan/unit di jenis yang sama
Contoh	Warna (merah, hijau, biru)	Nilai huruf mahasiswa (A, B, C, D, E)	Suhu dalam Celcius, tanggal dalam kalender tertentu	Panjang jalan, suhu dalam Kelvin
Ukuran data menyatakan ...	Membership	Membership, comparison	Membership, comparison, difference	Membership, comparison, difference, magnitude
Operasi matematika	=, ≠	=, ≠, <, >	=, ≠, <, >, +, -	=, ≠, <, >, +, -, ×, ÷

Tipe butir data (2)

	Nominal/kategorikal	Ordinal	Interval	Rasio
Representasi nilai tipikal	Modus	Modus, median	Modus, median, rerata aritmetis	Modus, median, rerata aritmetik, rerata geometrik, rerata harmonik
Representasi sebaran	Grouping	Grouping, rentang (<i>range</i>), rentang antarkuartil	Grouping, rentang (<i>range</i>), rentang antarkuartil, varians, simpangan baku	Grouping, rentang (<i>range</i>), rentang antarkuartil, varians, simpangan baku, koefisien variasi
Memiliki nol sejati yang menyatakan nilai mutlak terbawah.	Tidak	Tidak	Tidak	Ya

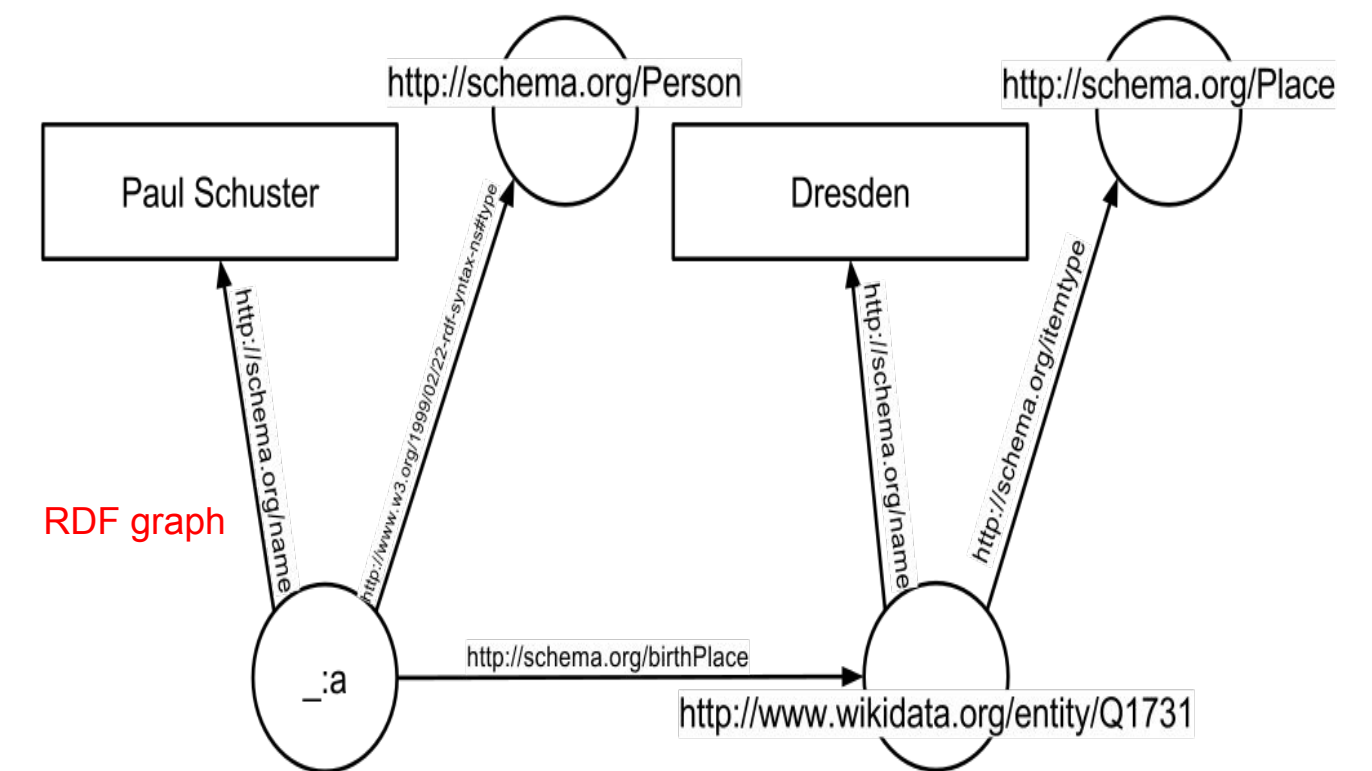
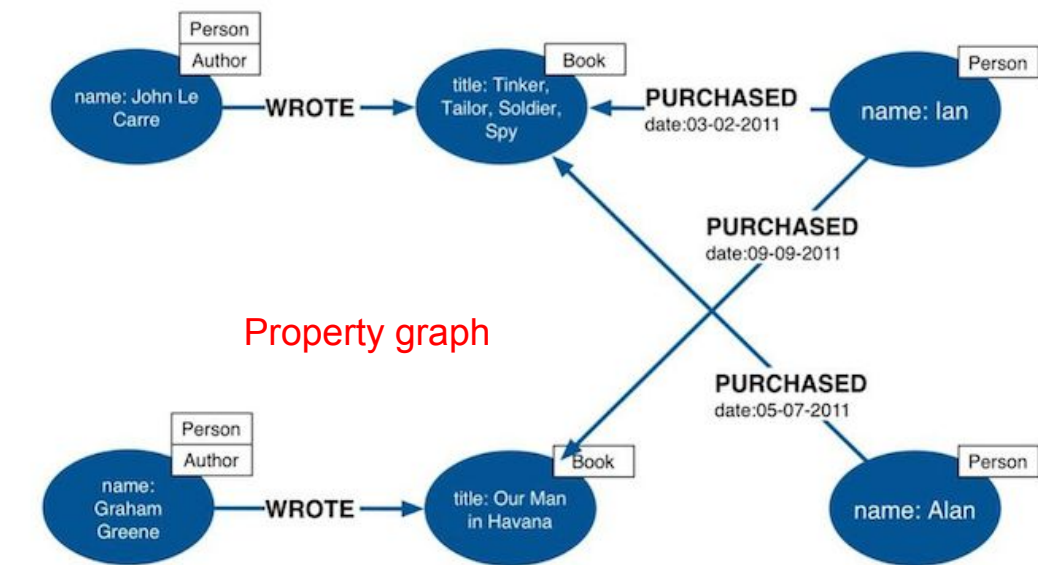
Contoh Model Data Tabular

- Terdiri dari N buah rekord (*record*)
- Masing-masing rekord mengandung D buah atribut
- Rekord = baris, *data point*, instans, *example*, transaksi, tipe entitas, objek, vector fitur.
- Atribut = kolom, *field*, dimensi, fitur.
- Atribut yang sama untuk setiap rekord biasanya diasumsikan memiliki tipe butir data yang sama.
- Struktur dapat bersifat ketat/strict (contoh: basis data relasional) atau longgar/loose (contoh: Excel *spreadsheet*).
- Tergantung keketatan strukturnya, bisa ada bahasa kueri formal untuk mengakses butir-butir data di dalamnya (contoh: SQL).

symboling	normalized-losses	make
3 ?		alfa-romero
3 ?		alfa-romero
1 ?		alfa-romero
2	164	audi
2	164	audi

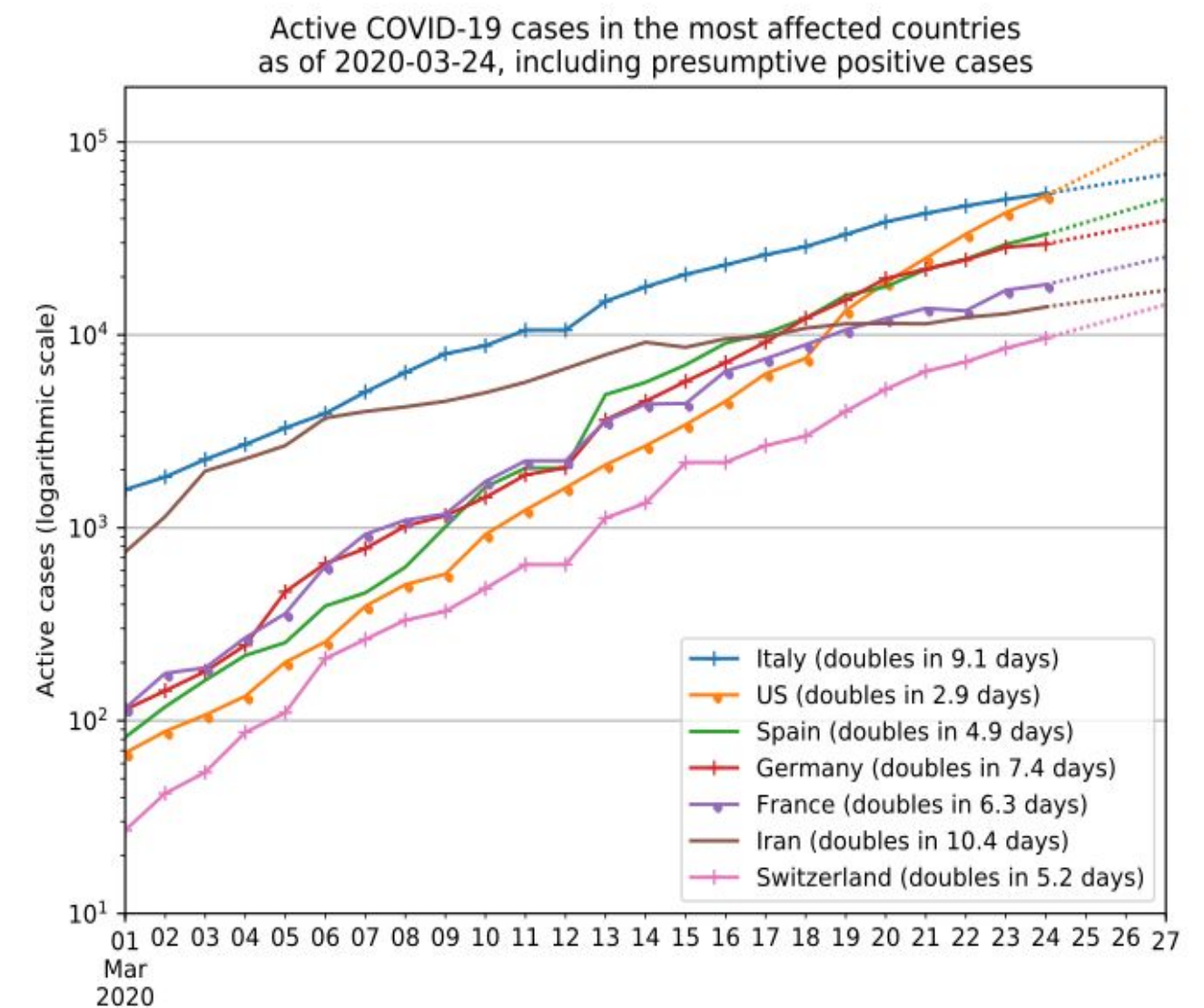
Contoh Model Data Graf/Jejaring

- Tersusun dari simpul-simpul (*nodes*) dan sisi/koneksi antar simpul (*edges*)
- Satu node (biasanya) mewakili satu record
- Dapat mengekspresikan relasi antar record secara eksplisit.
- Termasuk model data graf adalah model data hierarkis/pohon, model data berorientasi objek (*object-oriented data model*).
- Model data graf modern:
 - *Property graph*
 - *Resource description framework (RDF)*



Contoh Model Data Sekuens

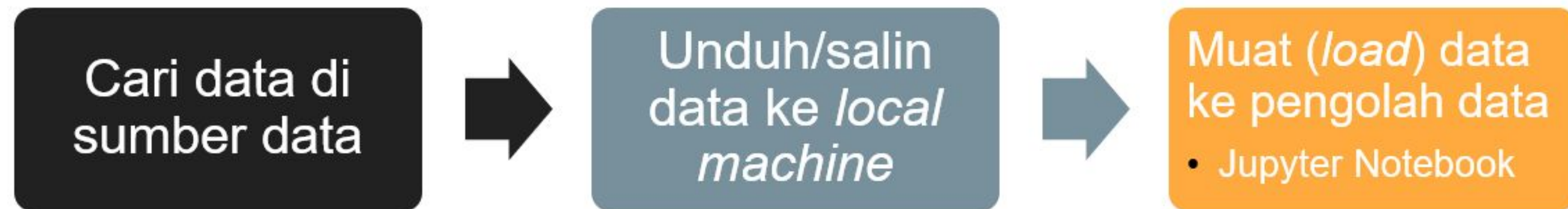
- Tersusun dari rekord-rekord yang terhubung secara sekuensial.
- Contoh: data dari sensor suhu selama suatu rentang waktu.
- Struktur tersirat dari urutan kemunculan rekord
- Rekaman audio dan video dapat dipandang sebagai data sekuens, namun setiap rekordnya sendiri bersifat tidak terstruktur.
- Atribut kontekstual mendefinisikan basis dependensi tersirat. (Contoh: time stamp pada sensor suhu)
- Atribut behavioral: butir-butir data yang nilainya diperoleh dalam suatu konteks tertentu (Contoh: besarnya suhu).
- Jika atribut kontekstualnya adalah waktu/time stamp, maka data sekuens disebut *time series*.



Pengambilan Data

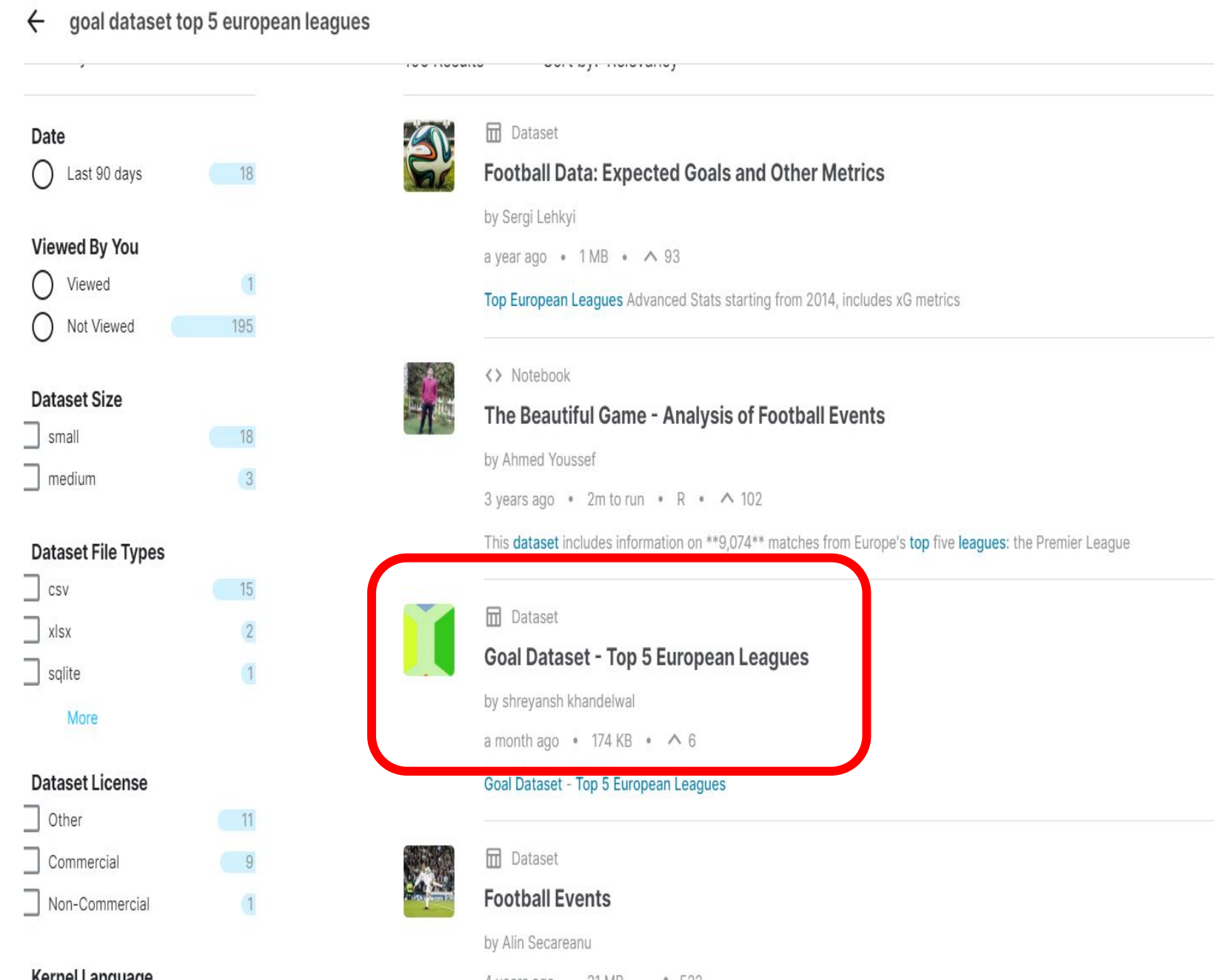
- Pengambilan data secara manual.
- Pengambilan data melalui API
 - Contoh melalui API Kaggle
 - Contoh melalui API Portal Data Bandung
- Pengambilan data melalui *web scraping*
- Pengambilan data melalui akses langsung ke basis data relasional yang ada

Pengambilan data secara manual

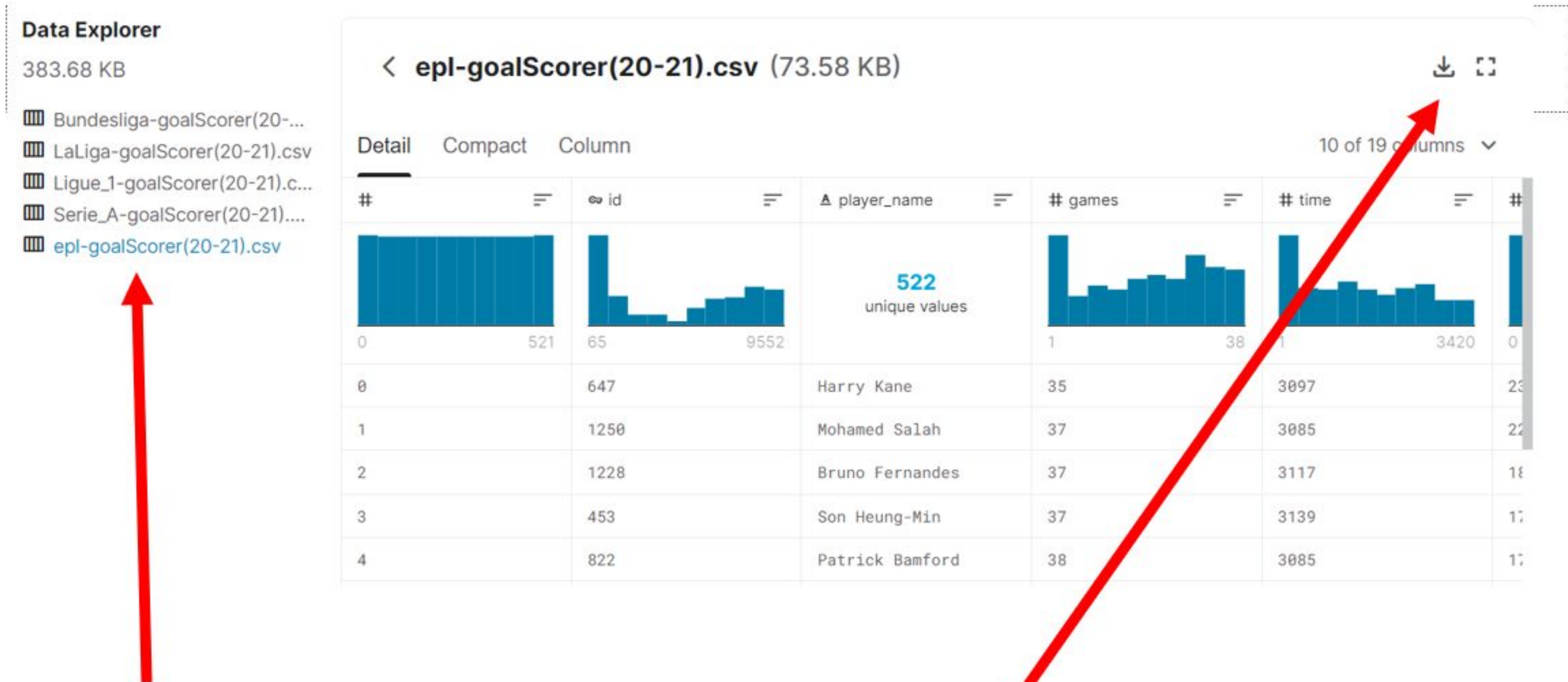


Mengambil data (secara manual) dari Kaggle

- Kita akan mengakses data dari "Goal Dataset – Top 5 European Leagues" dari Kaggle.
- Kunjungi Kaggle.com dan login (buat akun jika perlu)
- Lakukan pencarian "goal dataset top 5 European leagues"
- Klik "Goal Dataset – Top 5 European Leagues"



Mengambil data (secara manual) dari Kaggle



- Di halaman data explorer, pilih "epl-goalScorer (20-21).csv"
- Unduh data dengan mengklik tombol unduh di bagian kanan dan simpan di folder kerja Anda.

Memuat data ke Pandas (1)

- Nyalakan Jupyter Notebook di folder kerja Anda.
- Buka atau buat baru satu skrip ipynb (Python 3)
- Import pandas dan numpy. (Pastikan sudah terinstal sebelumnya).
- Load file CSV yang sudah diunduh sebelumnya (pada contoh "Mengambil Data secara Manual") ke dalam sebuah DataFrame
 - Gunakan perintah `read_csv(...)`

```
In [1]: import pandas as pd  
import numpy as np
```

```
In [2]: path = "epl-goalScorer(20-21).csv"  
df = pd.read_csv(path)
```


Memuat data ke Pandas (2)

- Method `head()` dan `tail()` pada DataFrame membantu kita menampilkan beberapa baris pertama/terakhir dari data yang kita muat.

```
df.head(3)
```

	Unnamed: 0	id	player_name	games	time	goals	xG	assists
0	0	647	Harry Kane	35	3097	23	22.174859	14
1	1	1250	Mohamed Salah	37	3085	22	20.250847	5
2	2	1228	Bruno Fernandes	37	3117	18	16.019454	12

```
df.head()
```

	Unnamed: 0	id	player_name	games	time	goals	xG	assists
0	0	647	Harry Kane	35	3097	23	22.174859	14
1	1	1250	Mohamed Salah	37	3085	22	20.250847	5
2	2	1228	Bruno Fernandes	37	3117	18	16.019454	12
3	3	453	Son Heung-Min	37	3139	17	11.023287	10
4	4	822	Patrick Bamford	38	3085	17	18.401863	7

Mengungkap tipe-tipe data dari setiap kolom

- Atribut `dtypes` pada `DataFrame` berisi tipe data dari setiap kolom.
- Lihat Pandas User Guide untuk detail setiap tipe.
- `dtype: object` di akhir output `dtypes` mewakili `Series` yang merupakan objek Python yang dikembalikan oleh `dtypes` itu sendiri (bukan bagian dari tipe kolom manapun).

```
print(df.dtypes)
```

```
Unnamed: 0      int64  
id              int64  
player_name     object  
games           int64  
time            int64  
goals           int64  
xG              float64  
assists         int64  
xA              float64  
shots           int64  
key_passes      int64  
yellow_cards    int64  
red_cards       int64  
position        object  
team_title      object  
npg             int64  
npxG            float64  
xGChain         float64  
xGBuildup       float64  
dtype: object
```

Mengungkap tipe-tipe data dari setiap kolom

- Dua kolom pertama hanyalah ID numerik yang biasanya tidak memiliki makna riil
- Jadi, dari DataFrame `df`, cukup diambil mulai dari kolom "player_name" (untuk *zero-based index*, kita pakai kolom ke-2 dst).

```
df_noid = df.iloc[:,2:]  
df_noid
```

	player_name	games	time	goals	xG	assists	xA
0	Harry Kane	35	3097	23	22.174859	14	7.577094
1	Mohamed Salah	37	3085	22	20.250847	5	6.528526
2	Bruno Fernandes	37	3117	18	16.019454	12	11.474996
3	Son Heung-Min	37	3139	17	11.023287	10	9.512992
4	Patrick Bamford	38	3085	17	18.401863	7	3.782247
...
517	Jaden Philogene-Bidace	1	1	0	0.000000	0	0.000000
518	Gaetano Berardi	2	113	0	0.074761	0	0.000000
519	Anthony Elanga	1	67	0	0.000000	0	0.000000
520	Femi Seriki	1	1	0	0.000000	0	0.000000

Deskripsi statistik data

DataFrame method `describe()` menampilkan statistik dasar setiap kolom data yang bertipe numerik, mencakup banyaknya data (**count**), rerata aritmetik (**mean**), simpangan baku (**std**), nilai terkecil (**min**), kuartil pertama (**25%**), kuartil kedua/median (**50%**), kuartil ketiga (**75%**), dan nilai terbesar (**max**).

```
df_noid.describe()
```

	games	time	goals	xG	assists	xA	shots	key_passes	yellow_cards	red_cards	npg
count	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000
mean	19.643678	1420.068966	1.862069	2.000806	1.289272	1.376029	17.379310	12.963602	2.061303	0.091954	1.668582
std	11.619836	1031.604819	3.338851	3.317946	2.083350	1.886510	21.572664	16.164361	2.203661	0.295800	2.909929
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	10.000000	470.250000	0.000000	0.074668	0.000000	0.049245	2.000000	1.000000	0.000000	0.000000	0.000000
50%	21.000000	1342.000000	1.000000	0.737295	0.000000	0.691122	10.000000	7.000000	2.000000	0.000000	0.500000
75%	30.000000	2319.000000	2.000000	2.053378	2.000000	2.050509	23.750000	19.000000	3.000000	0.000000	2.000000
max	38.000000	3420.000000	23.000000	22.174859	14.000000	11.474996	138.000000	95.000000	12.000000	2.000000	19.000000

<

Basic Statistic

Central Tendency:

Suatu pengukuran untuk menghitung posisi “Central” dalam sebuah data.

Mean, Median, dan Modus

1

Mean (Arithmetic)

Jumlah semua nilai dibagi
banyaknya

$$\bar{x} = \frac{\sum x}{n}$$

2

Median

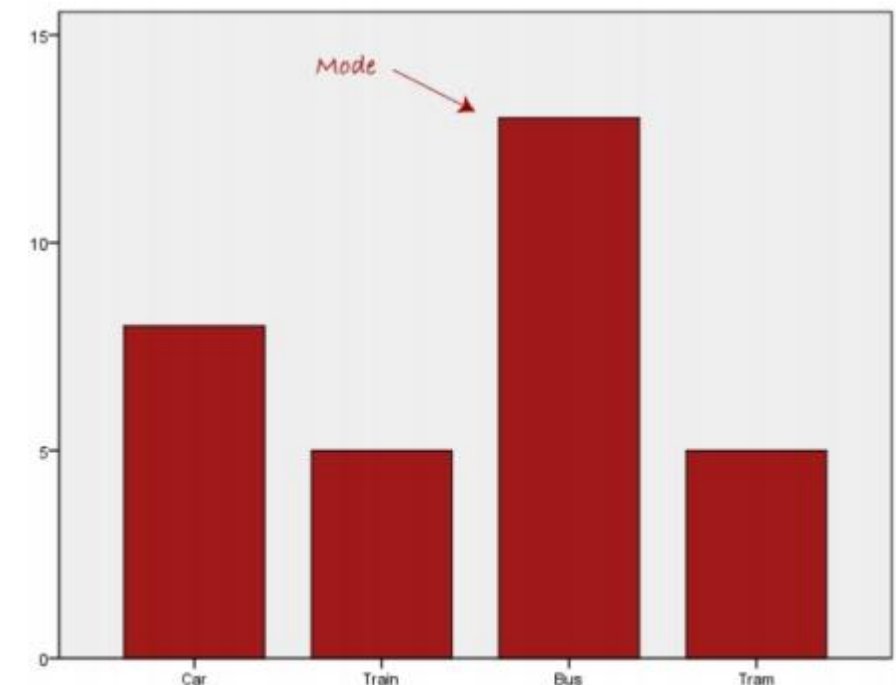
Nilai tengah dari data

14	35	45	55	55	56	56	65	87	89	92
----	----	----	----	----	-----------	----	----	----	----	----

3

Modus

elemen yang paling sering muncul



Kapan 3M digunakan?



Case 1

Staff	1	2	3	4	5	6	7	8
Salary	15k	18k	16k	14k	15k	15k	12k	17k

mean : 15.25

median : 15

- Mean cocok digunakan sebagai pemusatan data



Case 2

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

mean : 30.7

median : 15.5

- Mean kurang cocok digunakan sebagai pemusatan data, karena terdapat pencilan
- Median lebih cocok karen “Robust ” terhadap pencilan



Case 3

Jakarta	Jakarta	Medan	Surabaya
Bali	Bandung	Jakarta	Bandung

modus

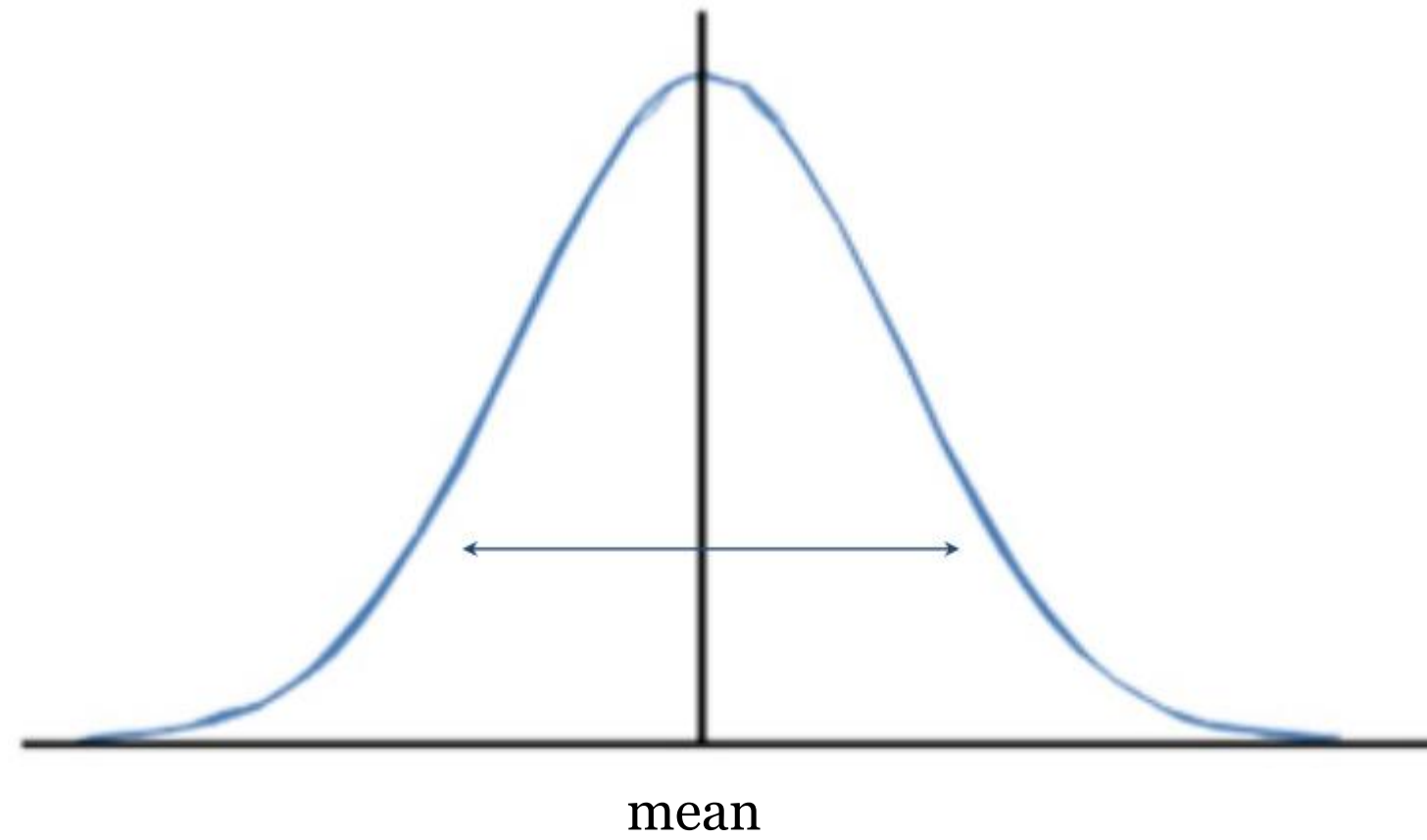
- Modus cocok untuk data ordinal



Variansi & Kovariansi

Suatu pengukuran untuk menghitung seberapa menyebar suatu data.

Variansi



Data
17
15
20
23
16
29

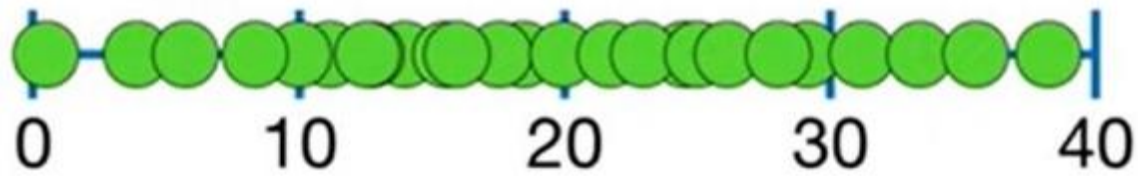
$$\begin{aligned}\text{Population Variance} &= \frac{\sum(x - \mu)^2}{n} \\ &= \frac{(17-20)^2 + (15-20)^2 + (20-20)^2 + (23-20)^2 + (16-20)^2 + (29-20)^2}{6} \\ &= 23.3 \\ \text{std} &= \sqrt{23.3}\end{aligned}$$

Populasi & Sampel



Populasi

Keseluruhan data Bagian dari data

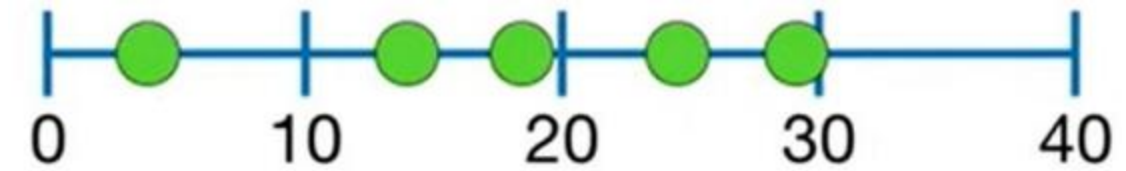


$$\text{Population Variance} = \frac{\sum (x - \mu)^2}{n}$$

$$\text{Population Standard Deviation} = \sqrt{\text{Population Variance}}$$



Sampel



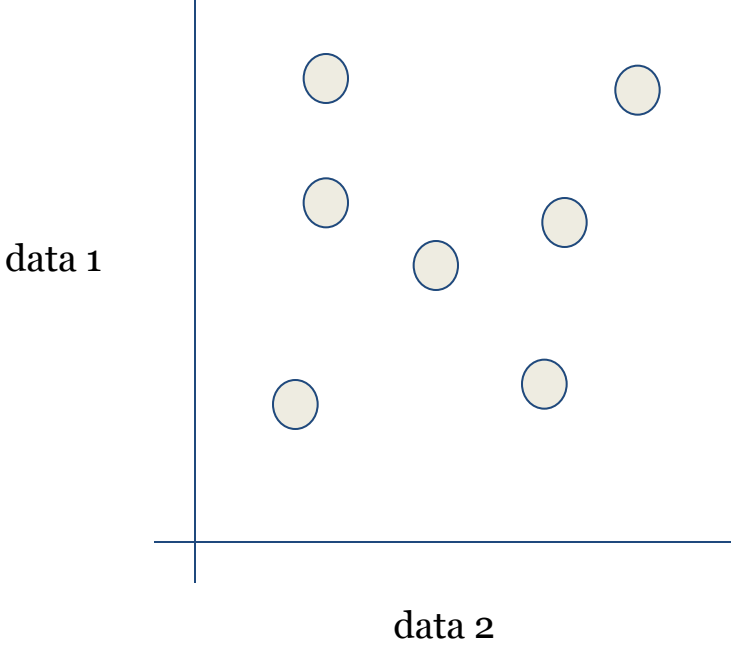
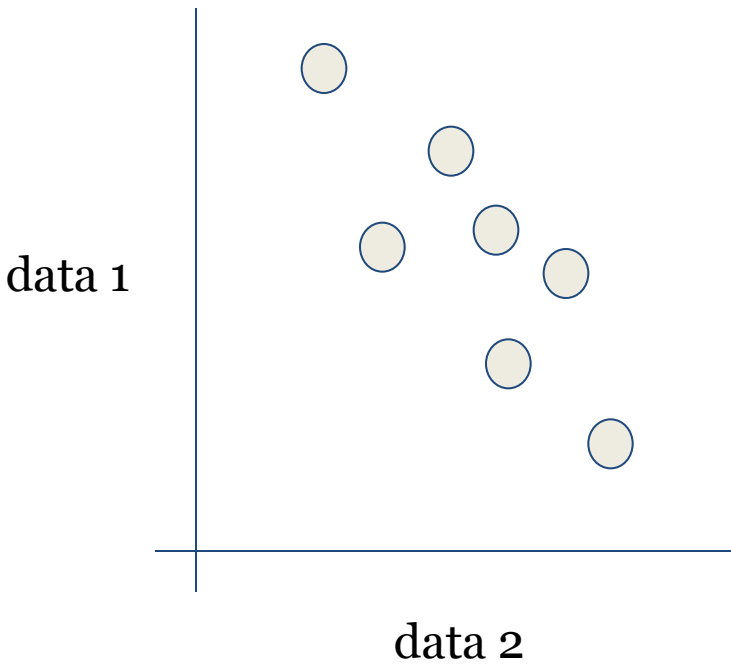
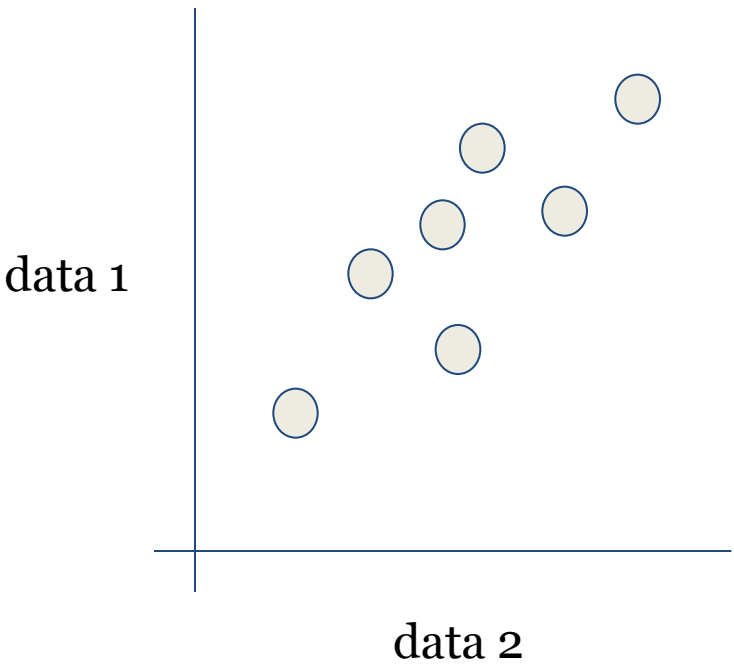
$$\text{Estimated Population Variance} = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$\text{Estimated Population Standard Deviation} = \sqrt{\text{Estimated Population Variance}}$$

Kovariansi

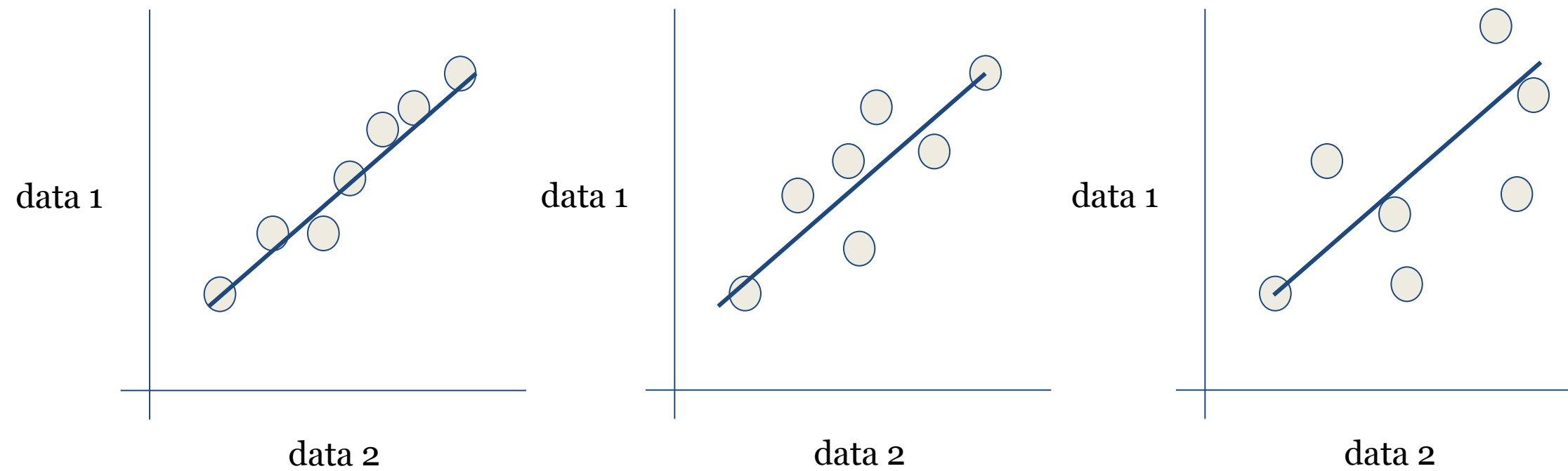
Data1	Data2
17	23
15	34
20	12
23	19
16	24
29	13

Kovariansi = $\frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$



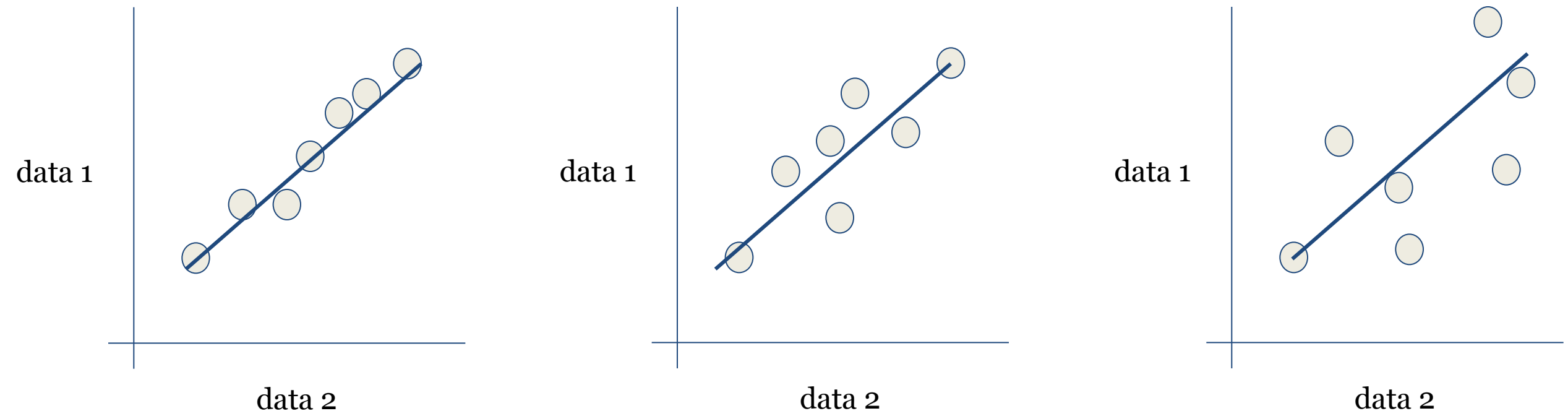
Kovariansi positif Kovariansi Negatif Kovariansi ~ 0

Masalah dengan Kovariansi



Kovariansi positif Kovariansi positif Kovariansi positif

Korelasi



Korelasi 0.9 Korelasi 0.7 Korelasi 0.4

$$\text{Correlation} = \frac{\text{Cov}(x, y)}{\sigma x * \sigma y}$$

standar deviasi variabel x

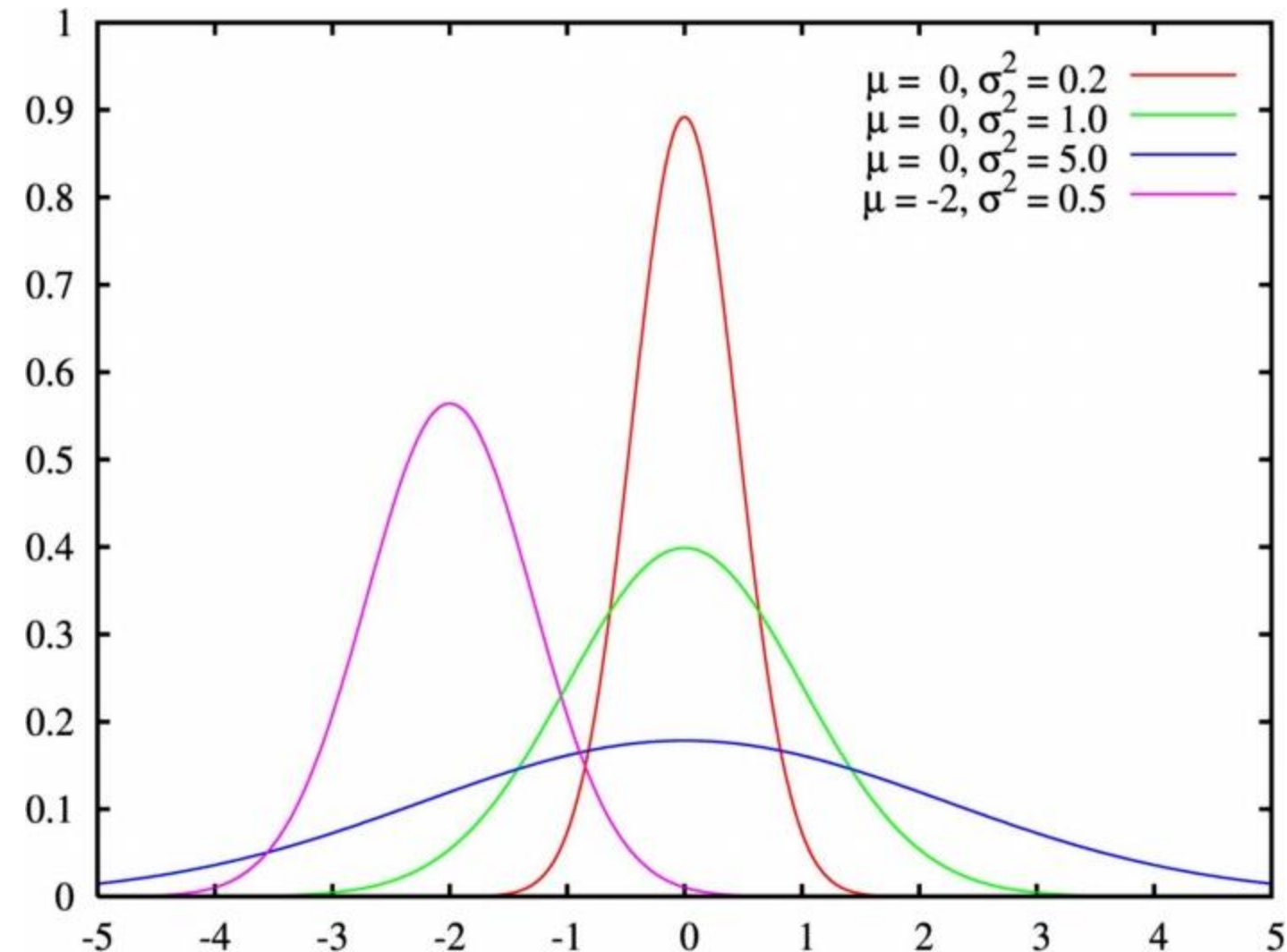
note : Korelasi tidak menyebabkan sebab akibat



Distribusi Normal

Distribusi Normal

Gaussian distribution

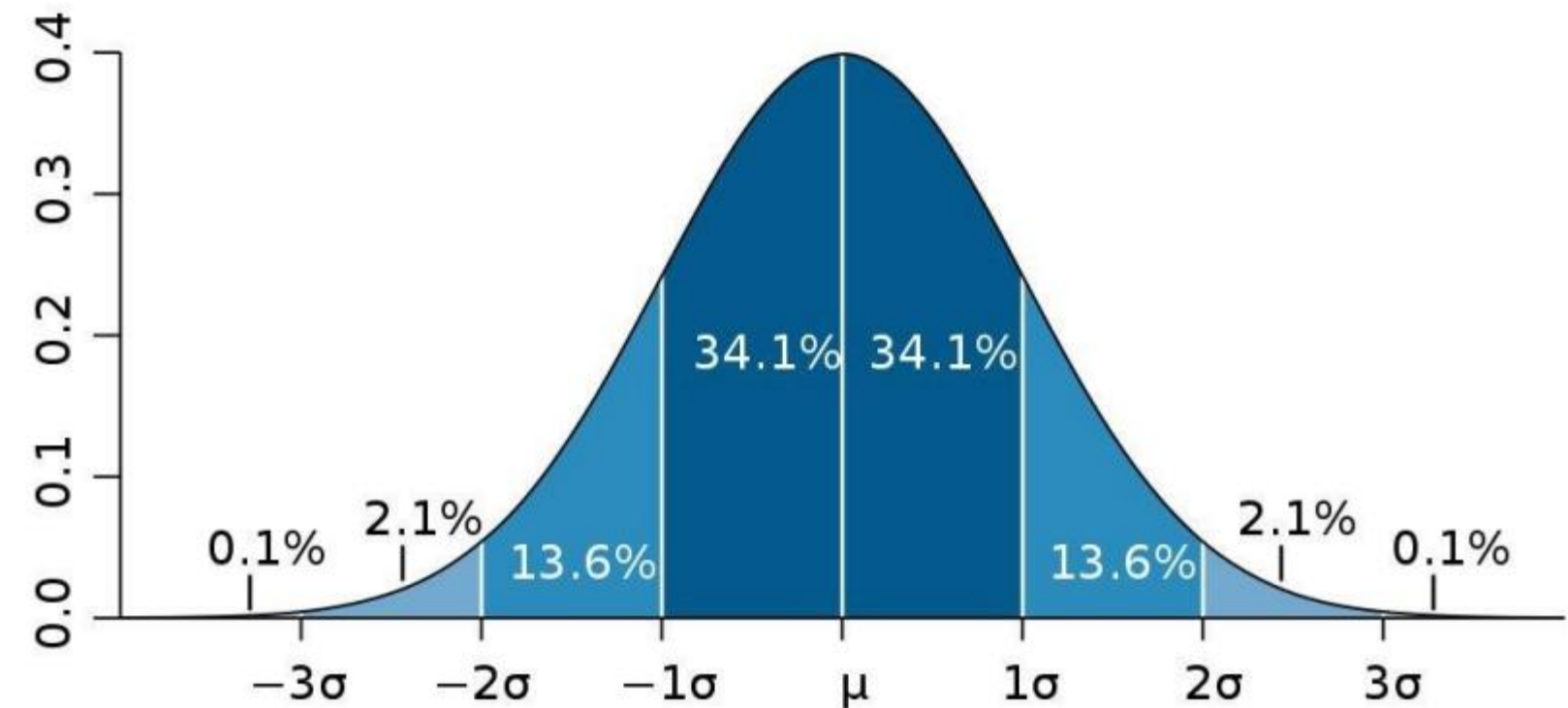


Normal Baku (standar)

rata-rata = 0

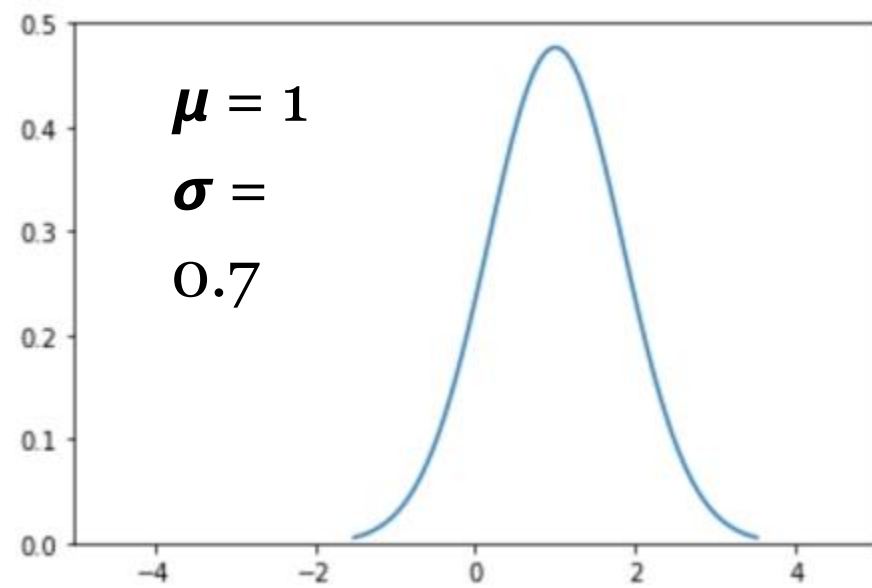
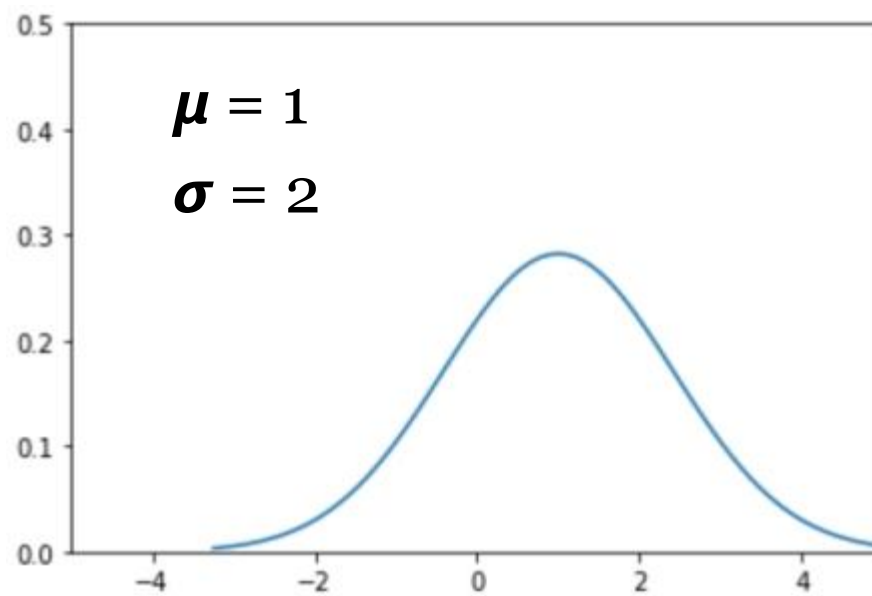
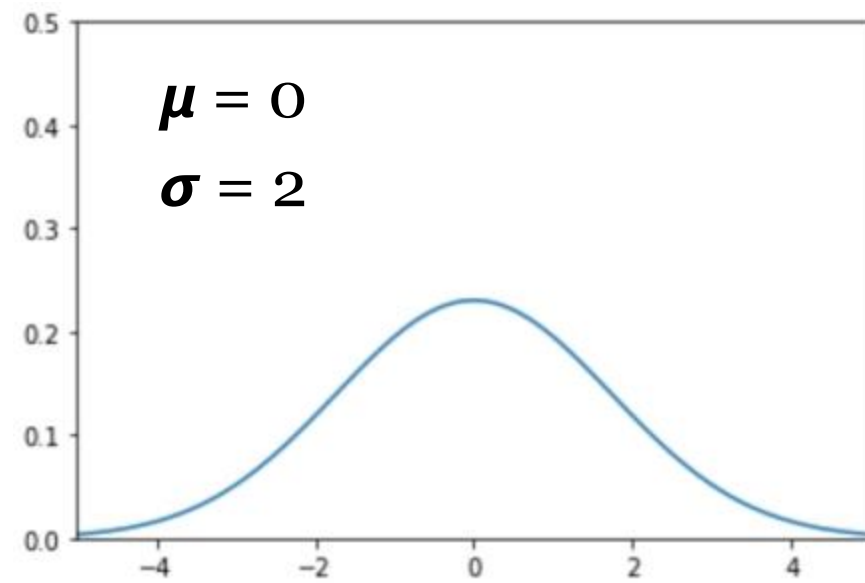
standar deviasi = 1

3 sigma rule

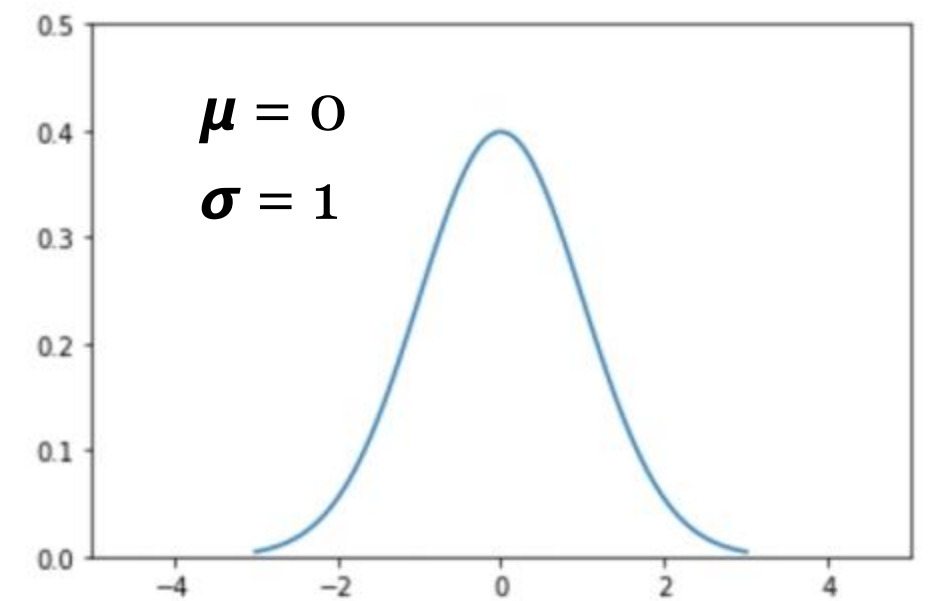


peluang random number diantara $(\mu - \sigma, \mu + \sigma)$ adalah 68.2%

peluang random number diantara $(\mu - 2\sigma, \mu + 2\sigma)$ adalah 95.4%



$$(x - \mu) / \sigma$$



Central Limit Theorem

distribusi dari rata-rata sampel untuk variabel acak akan mendekati distribusi normal jika sampel semakin besar.

(apapun distribusi populasinya)

Contoh Eksperimen

Populasi Sample

Data
1
2
3
4
5
6
7
8
9
10



Data
1
7
8
4
5

Data
3
4
8
2
1

Data
6
2
4
3
9

...

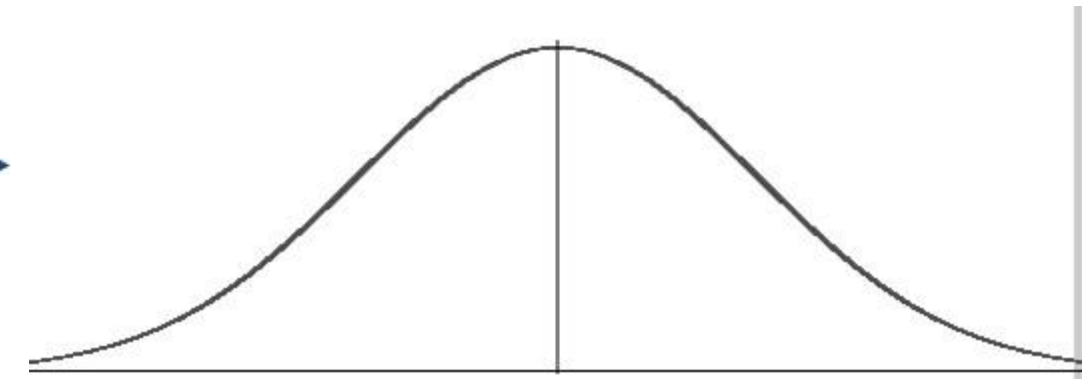
Data
5
3
8
8
10

Data
5
3
9
5
1

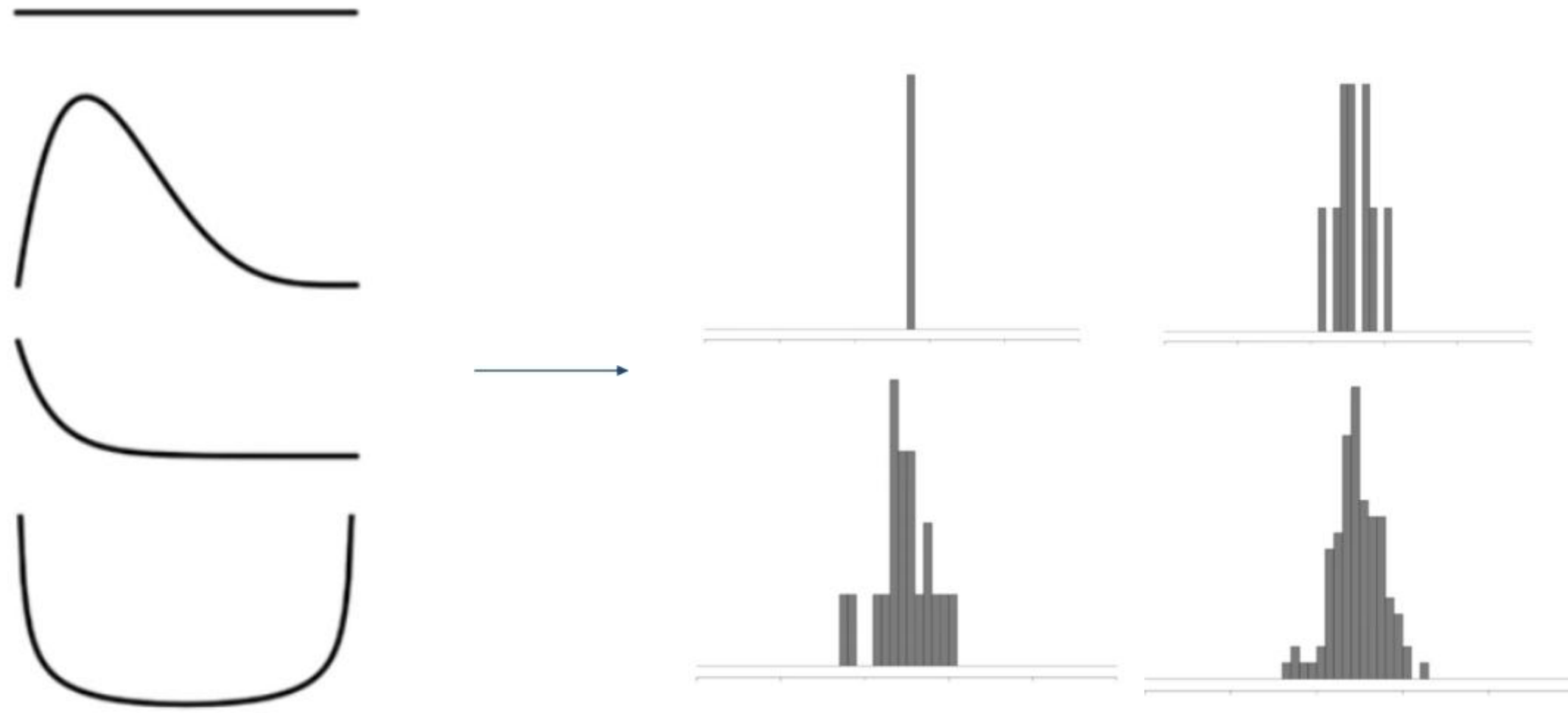
Data
9
1
2
3
5



Average
sample



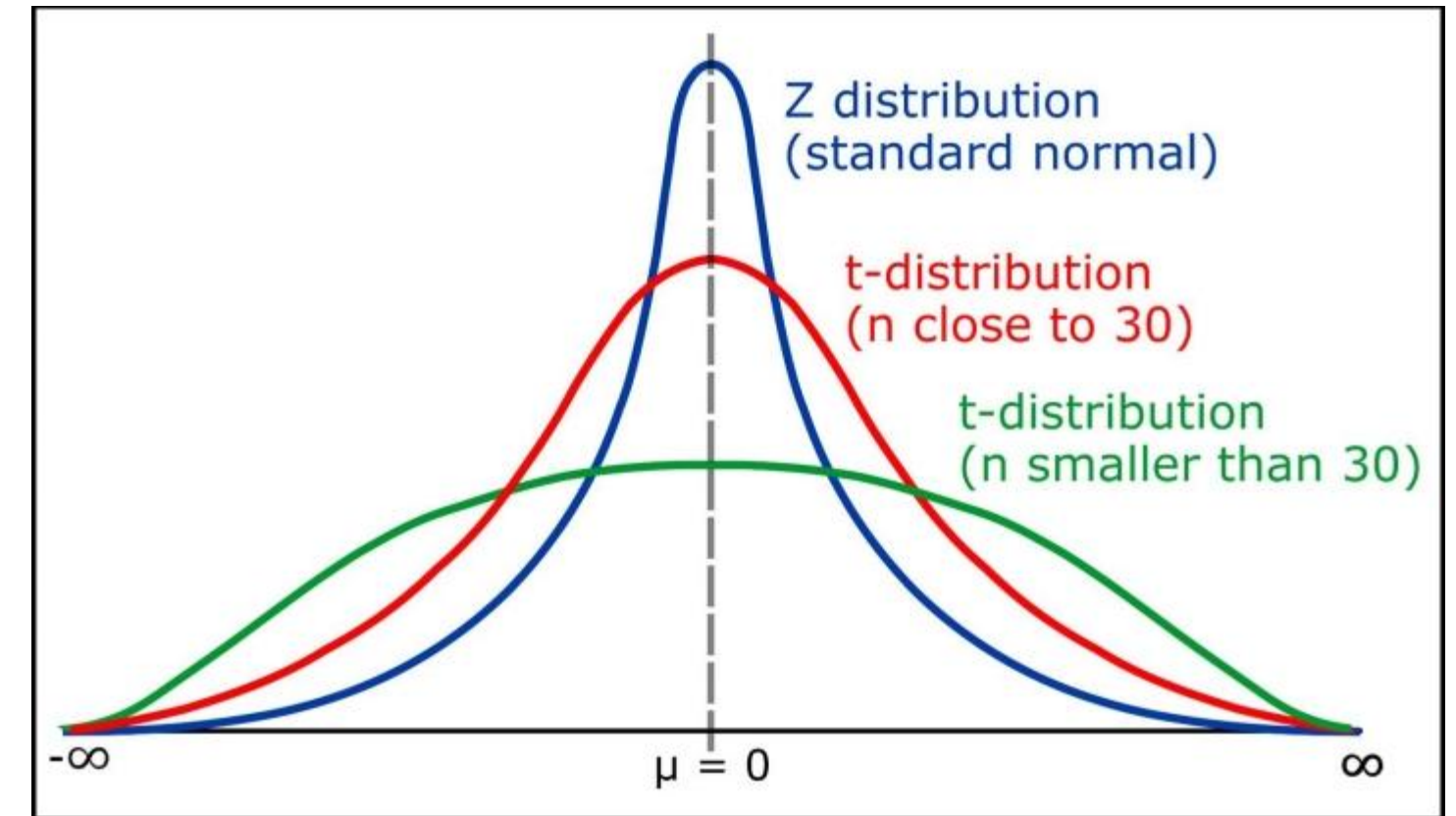
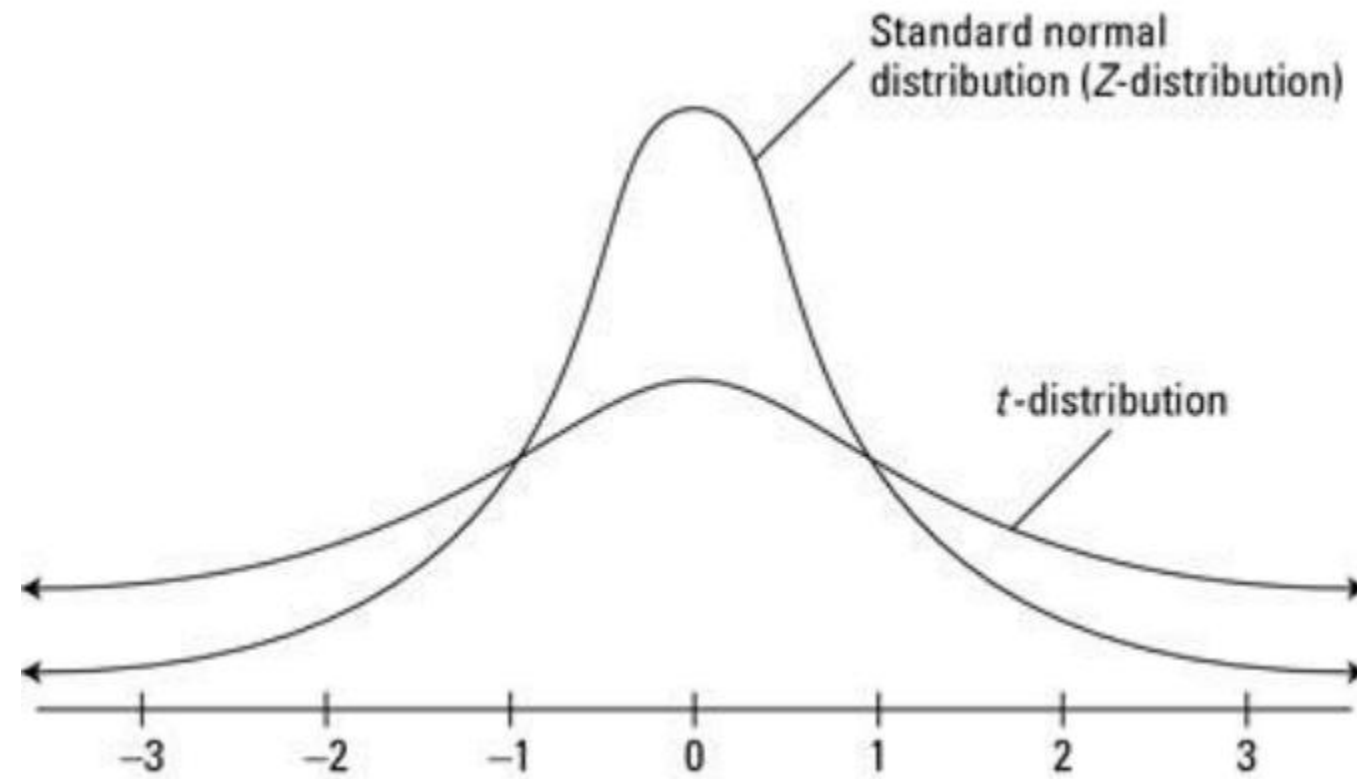
Central Limit Theorem



Apapun distribusi populasi, average sample akan normal, Jika sampel cukup.

rule of thumb : ukuran sampel harus minimal 30 agar CLT terpenuhi

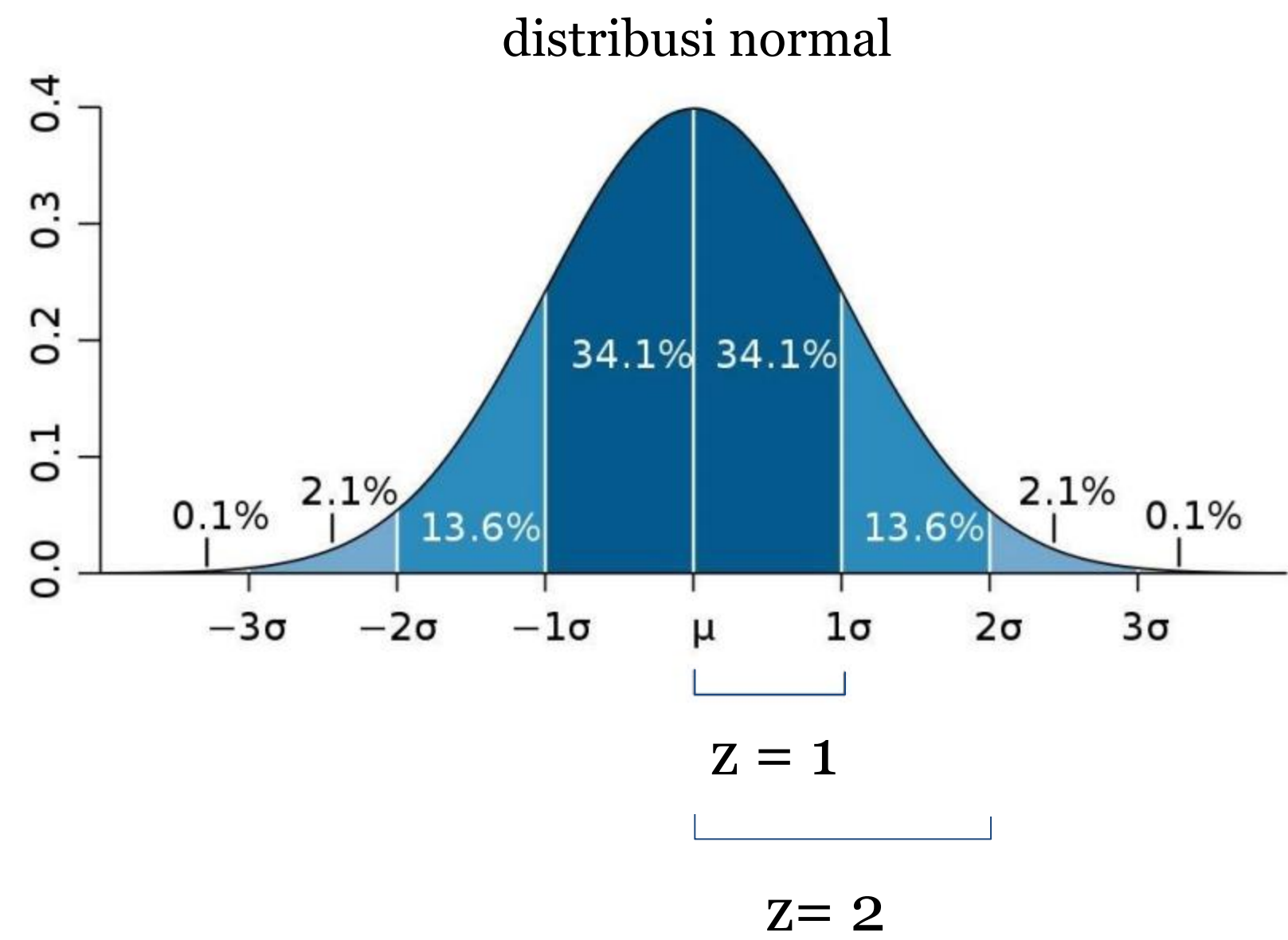
Z-score



Sebelum sampel berjumlah 30, rata-rata sampel berdistribusi t

Ke-"landai"-an t-distribution tergantung jumlah sampel. Jadi,
semakin
banyak sampel semakin mendekati distribusi normal

Z-score

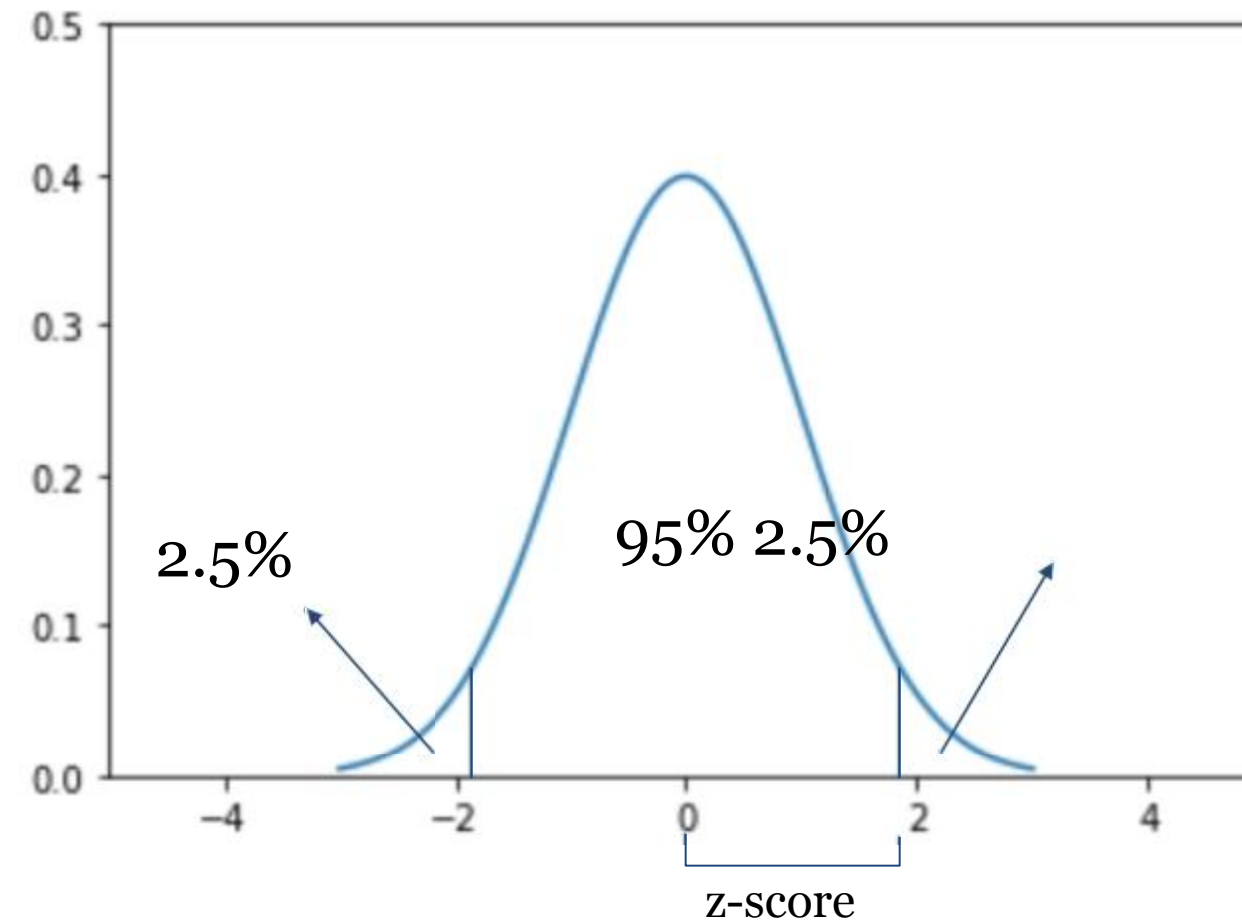


z-score menyatakan seberapa jauh suatu nilai jauh dari rataannya (dalam satuan sigma)

z-score	Peluang
1	0.841
2	0.977
-1	0.158
0.5	0.691

Z-score

jika mau interval dengan peluang 95%, berapa intervalnya?

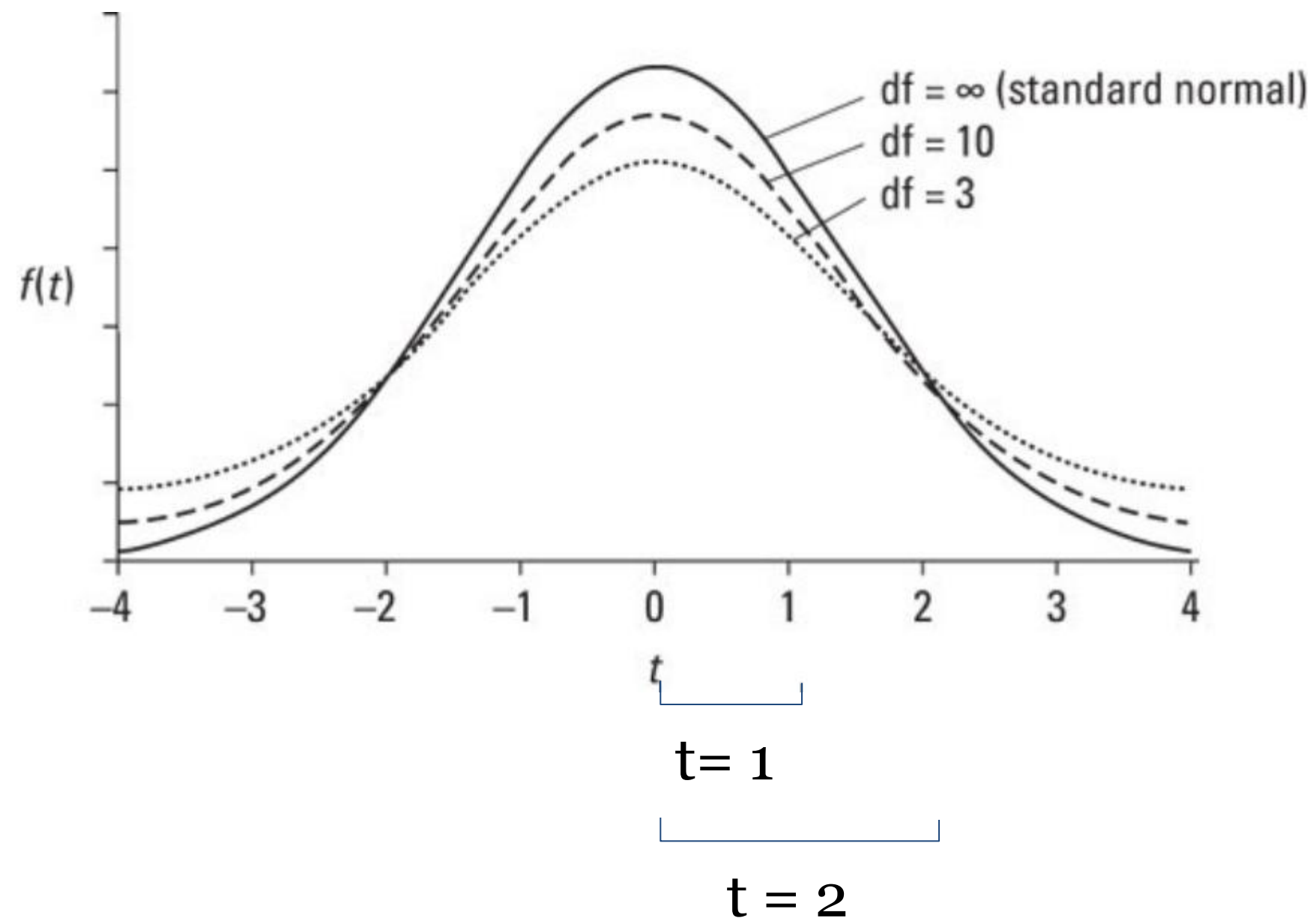


yang digunakan adalah
Z-score dengan
diperoleh z pada peluang 97.5%

interval agar 95% adalah $(\mu - 1.96\sigma, \mu + 1.96\sigma)$

t-score

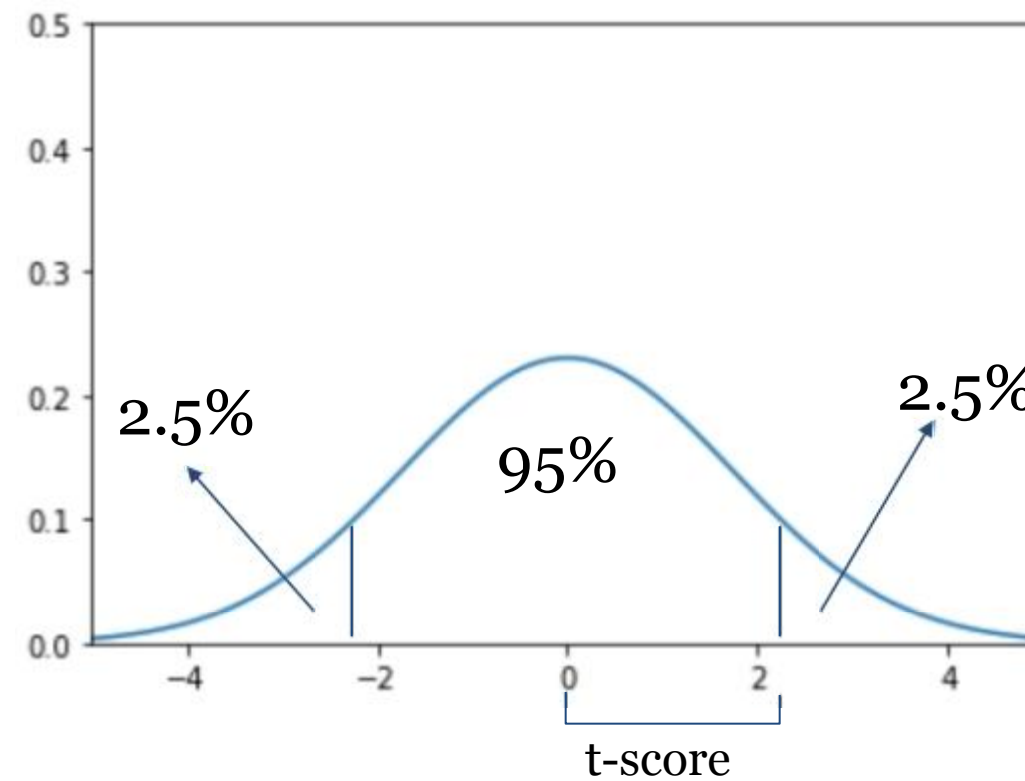
t-score menyatakan seberapa jauh suatu nilai jauh dari rataannya (dalam satuan sigma)



z-score	Peluang (df 3)	Peluang (df 10)	Peluang (df 100)
1	0.841	0.829	0.840
2	0.930	0.963	0.975
-1	0.195	0.170	0.159
0.5	0.674	0.686	0.690

t-score


jika mau interval dengan peluang 95% dengan sampel 20 berapa intervalnya?



yang digunakan adalah t-score dengan peluang 97.5%

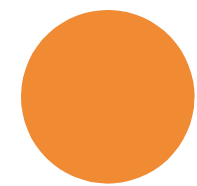
diperoleh $t = 2.1$

interval agar 95% adalah $(\mu - 2.1\sigma, \mu + 2.1\sigma)$



Probability Mass Function & Probability Density Function

Probability Mass Function



fungsi untuk menyatakan peluang dari suatu variabel diskrit

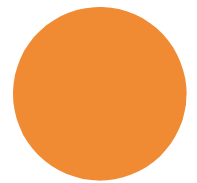
$$P_X(x) = \begin{cases} P(X = x) & \text{untuk } x \text{ di support} \\ 0 & \text{lainnya} \end{cases}$$

contoh :

Misalkan X adalah variabel acak mata dadu normal yang muncul

$$P_X(x) = \begin{cases} \frac{1}{6} & \text{untuk } x = 1, 2, \dots, 6 \\ 0 & \text{lainnya} \end{cases}$$

Probability Mass Function



contoh :

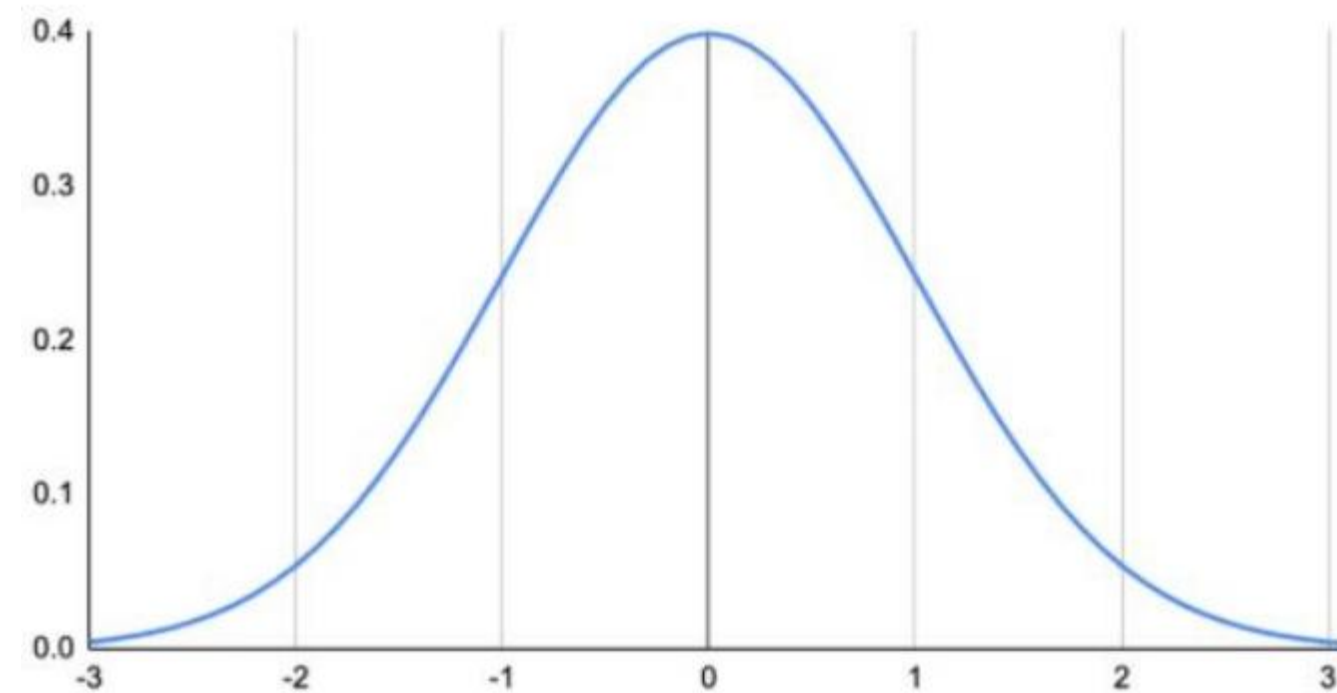
Misalkan X adalah variabel acak jumlah dari dua dadu normal

		Outcome of First Die					
		1	2	3	4	5	6
Outcome of Second Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

$$P_X(x) = \begin{cases} \frac{1}{36} & , x = 2 \\ \frac{2}{36} & , x = 3 \\ \frac{3}{36} & , x = 4 \\ \dots & \\ 0 & , x \text{ lain} \end{cases}$$

Probability Mass Function

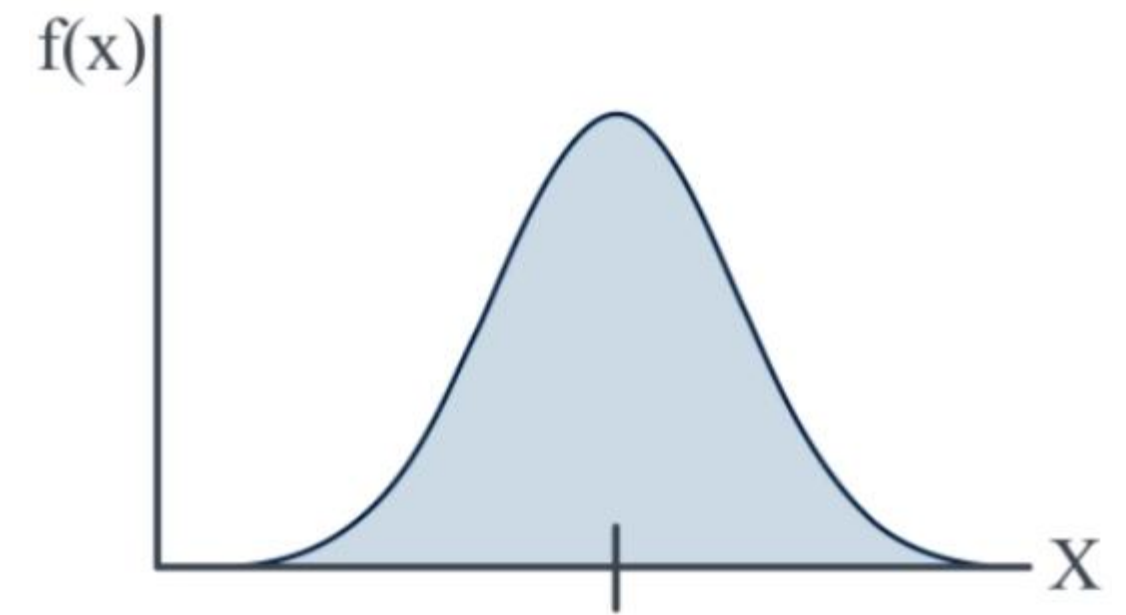
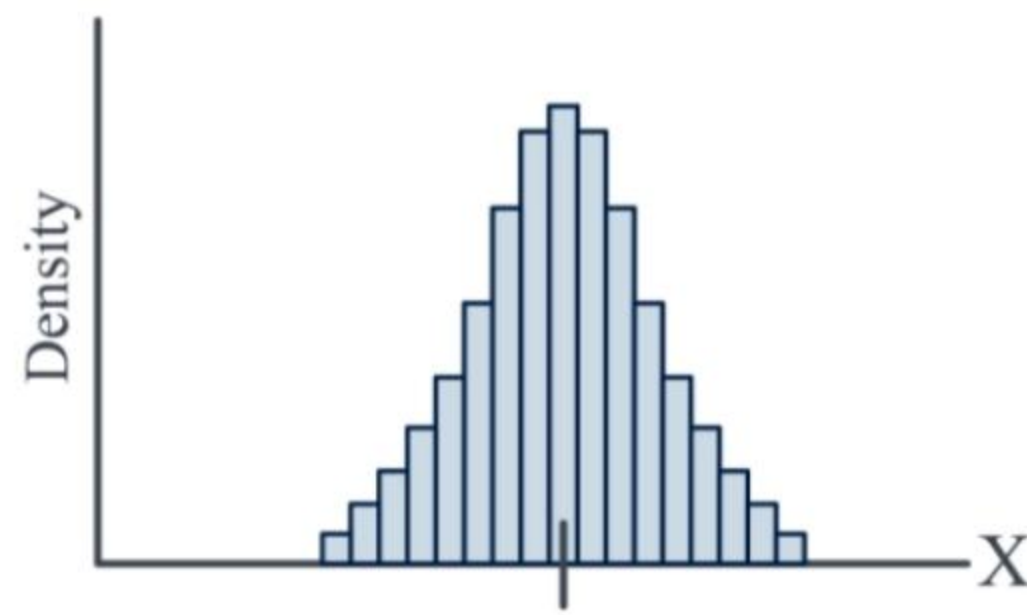
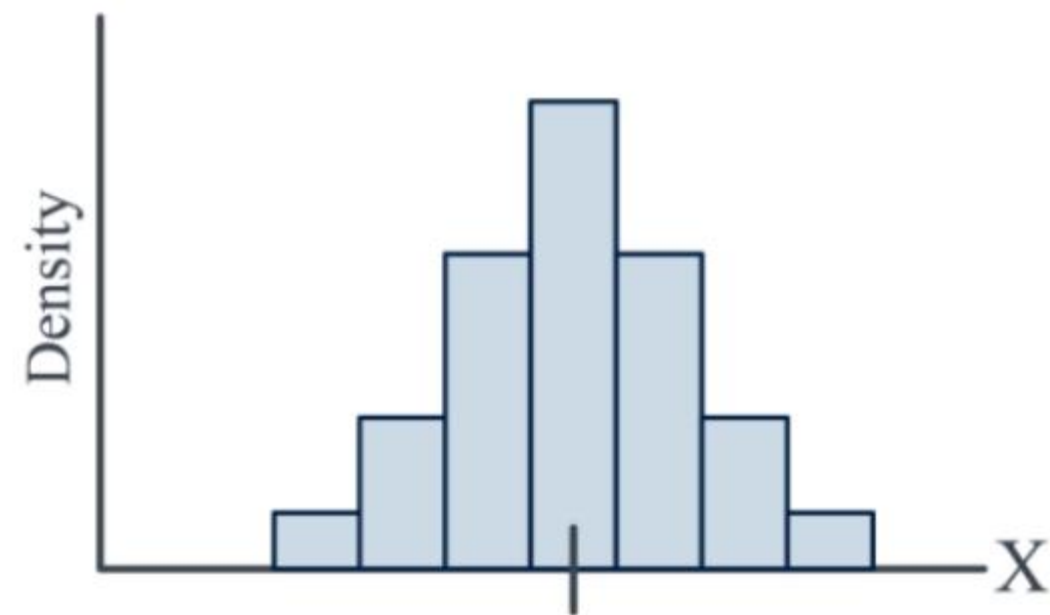
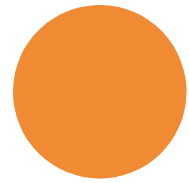
● bagaimana jika di variabel yang kontinu?



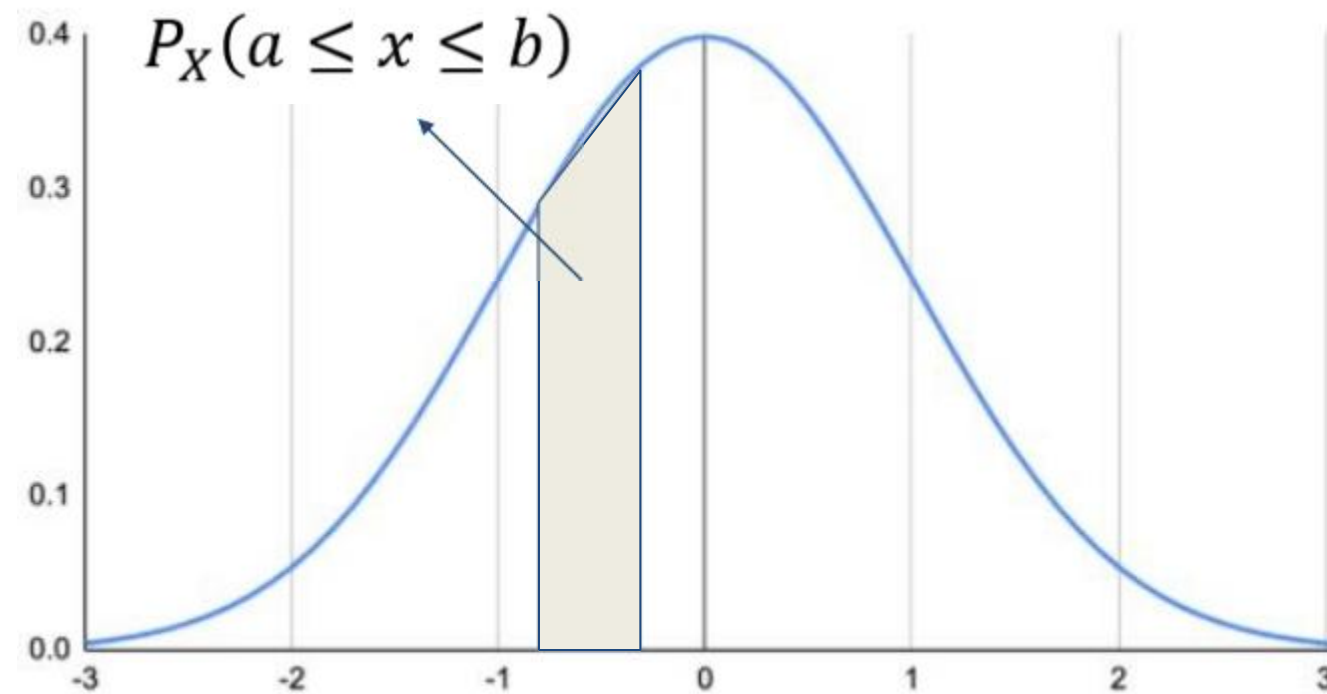
Berapa peluang dari distribusi normal standar diperoleh angka 0?

Probability Density Function

Fungsi yang menyatakan densitas dari distribusi kontinu



Probability Density Function



fungsi densitas distribusi normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

peluang dengan interval sama dengan luas daerah dibawah kurva

$$P_X(a \leq x \leq b) = \int_a^b f(x)dx$$



Ekspektasi

Berapa besar harapan yang kita
terima



Permainan

seandainya sedang bermain dadu. kalau angkanya 6 dikasih 80rb kalau enggak ngasih 10rb.
Profitable atau tidak?

Dadu	Hasil
1	-10rb
2	-10rb
3	-10rb
4	-10rb
5	-10rb
6	80rb

Ekspektasi



Perhitungan ekspektasi

Dadu	Hasil
1	-10rb
2	-10rb
3	-10rb
4	-10rb
5	-10rb
6	80rb

$$Ekspektasi(Bet) = \sum_{x=1}^n output.peluang$$

$$= -10 \cdot \frac{1}{6} - 10 \cdot \frac{1}{6} - 10 \cdot \frac{1}{6} - 10 \cdot \frac{1}{6} - 10 \cdot \frac{1}{6} + 80 \cdot \frac{1}{6}$$

$$= 5$$

$$Ekspektasi(100 \text{ Bet}) = 500$$

$$E[output] = \int output \cdot p(x) dx$$



Premium Game

jika angka dibuat lebih tinggi seperti tabel. melanjutkan permainan atau tidak?

Dadu	Hasil
1	-100rb
2	-200rb
3	-300rb
4	-400rb
5	-500rb
6	1000rb



P-value

Peluang suatu kejadian atau kejadian lain dengan peluang yang sama/ lebih rendah/ ekstrem

P-value

Peluang mendapat kejadian paling tidak sama jarangnya dengan kejadian

Misalkan percobaan lempar koin 1x



p-value dari head

peluang kejadian + peluang kejadian lain dengan peluang yg sama + lebih ekstrem

$$0.5 + 0.5 + 0$$

P-value

Misalkan percobaan lempar koin 2x

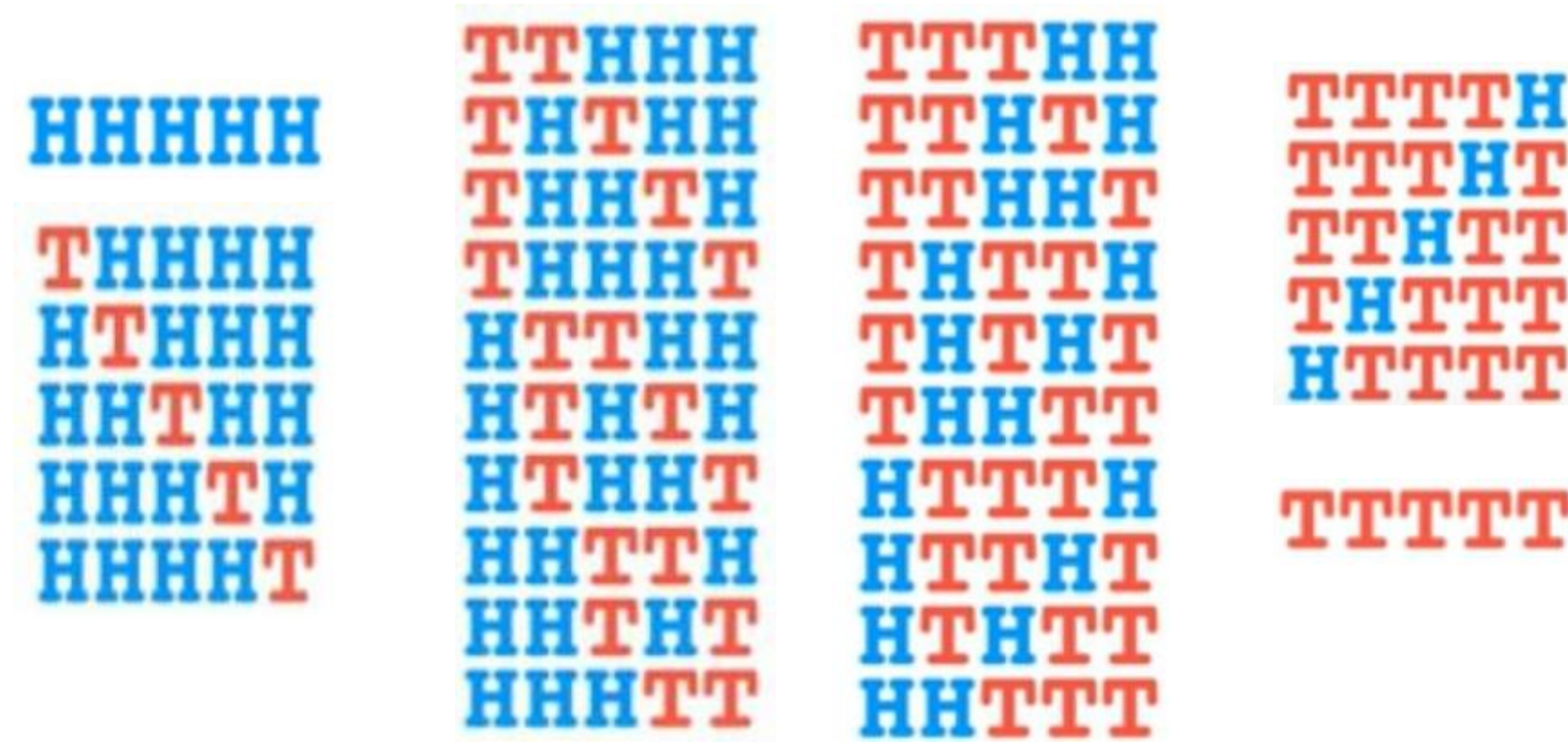


$$\text{p-value dari head 2x} = 0.25[\text{HH}] + 0.25[\text{TT}]$$

$$\text{p-value dari head Tail} = 0.5[\text{HT/TH}] + 0.25[\text{TT}] + 0.25[\text{HH}]$$

P-value

Misalkan percobaan lempar koin 5x



p-value dari head 4 dan tail 1 = $5/32 + 5/32 + 1/32 + 1/32$

peluang
kejadian

kejadian
lain
dengan
peluang
serupa

Kejadian
lebih
ekstrem

p-value dari head 5 ???

P-value

Misalkan percobaan lempar dadu 2x

		Outcome of First Die					
		1	2	3	4	5	6
Outcome of Second Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

p-value dari jumlahnya sama dengan 3

$$p(x = 2) = \frac{1}{36}$$
$$p(x = 3) = \frac{2}{36}$$

$$p(x = 4) = \frac{3}{36}$$

$$p(x = 5) = \frac{4}{36}$$

$$p(x = 6) = \frac{5}{36}$$

$$p(x = 7) = \frac{6}{36}$$

$$p(x = 8) = \frac{5}{36}$$

$$p(x = 9) = \frac{4}{36}$$

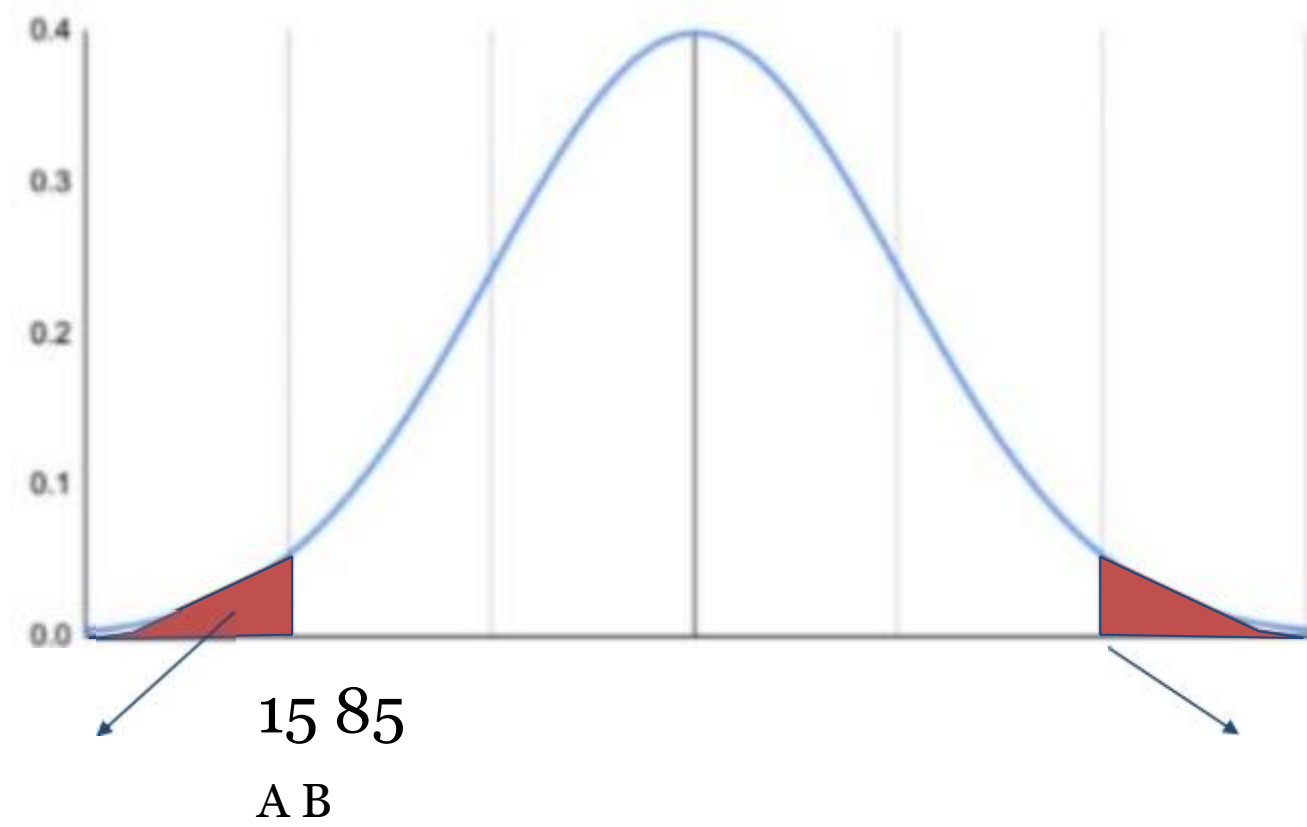
$$p(x = 10) = \frac{3}{36}$$

$$p(x = 11) = \frac{2}{36}$$

$$p(x = 12) = \frac{1}{36}$$

P-value

Misalkan distribusi nilai suatu kelas



p-value dari seseorang bernilai 15 = $A + B$

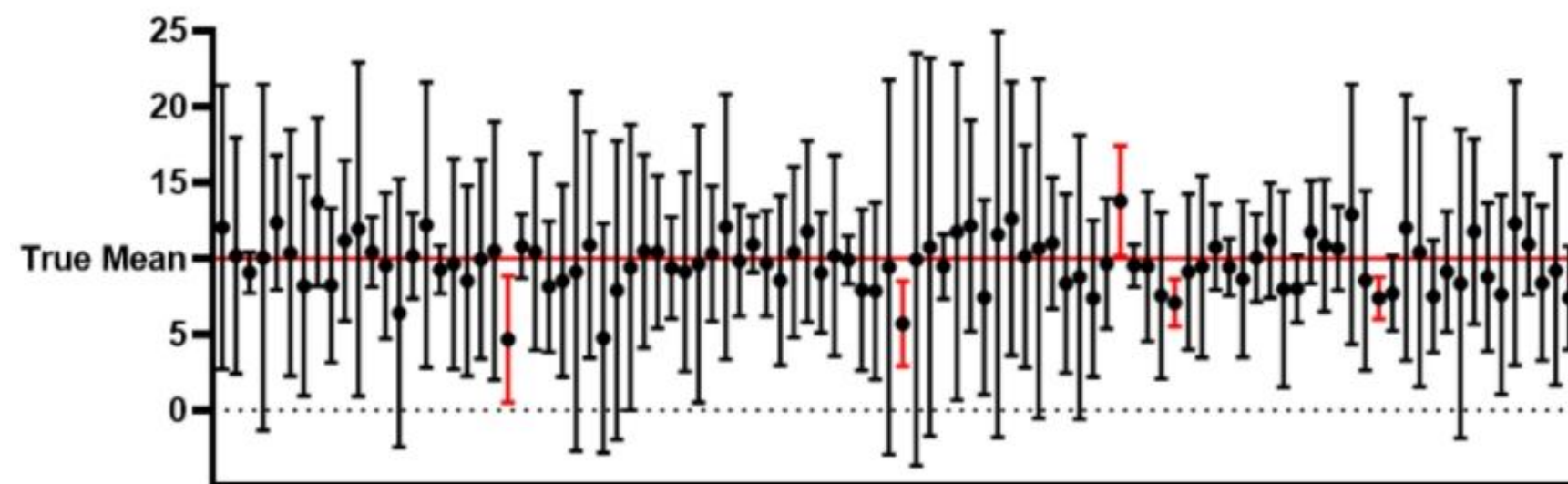


Hypothesis Testing

Confidence Interval

Confidence Interval digunakan untuk menentukan interval suatu parameter (rataan)

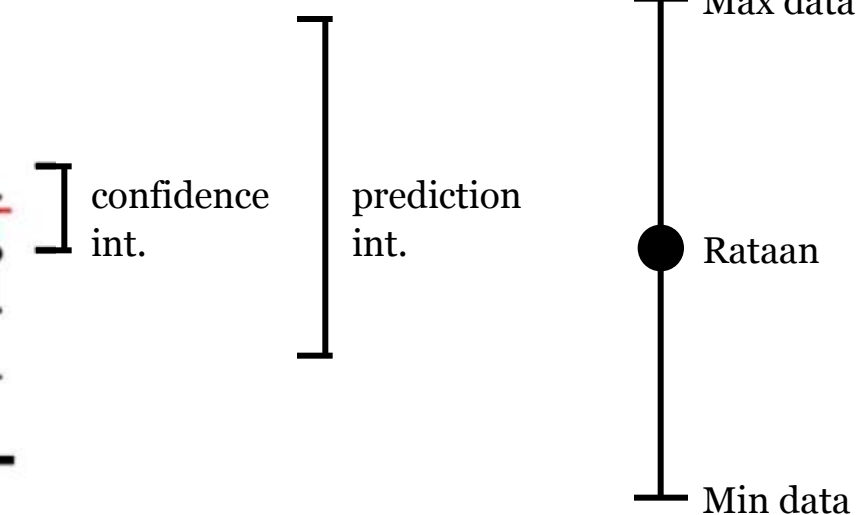
rata-rata tinggi warga Indonesia berada di interval X



Prediction Interval

Prediction Interval digunakan untuk menentukan interval dari data

Tinggi warga Indonesia yang muncul di IG story



Uji Hipotesis in a nutshell

1. Ada orang nge-claim sesuatu (rataan)
2. Kita sampel dan cari rataan dan variansi
3. Ingat Central Limit Theorem
4. Kita cari peluang nya
5. kesimpulan

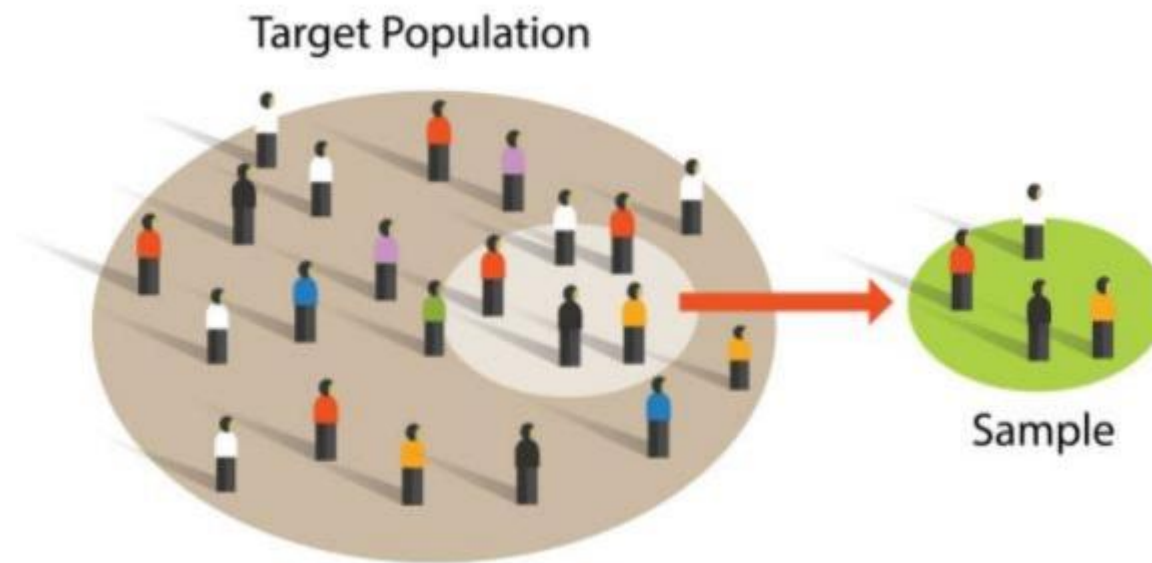
A little flashback

1

distribusi populasi

mean = μ

Var = σ^2

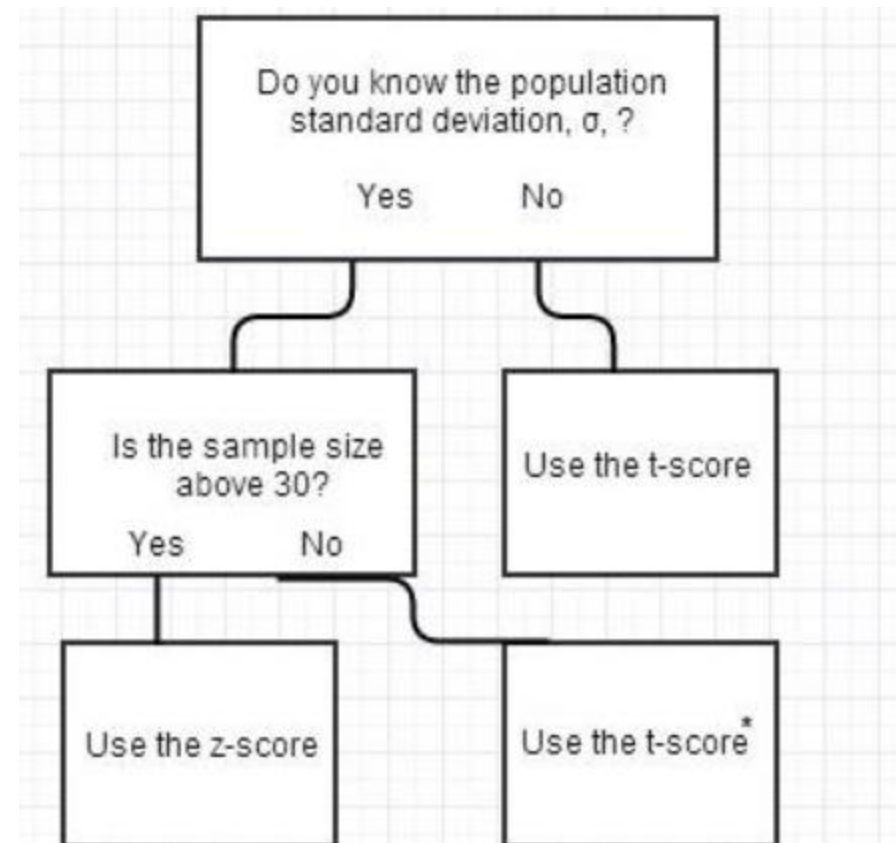
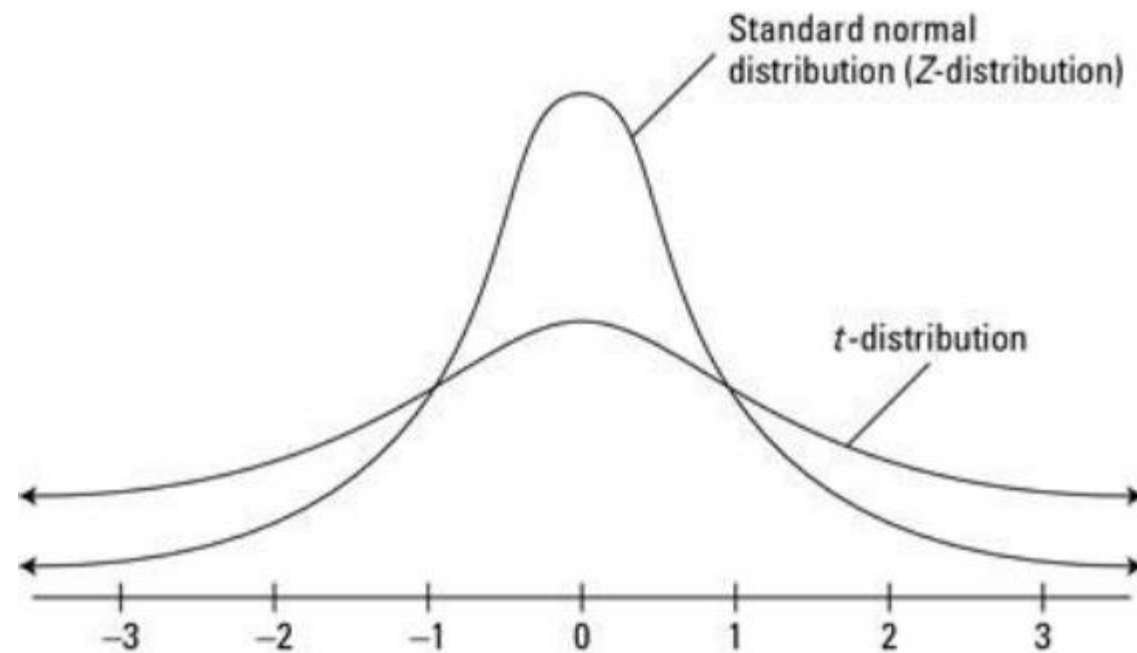


distribusi rata-rata
sample

$$N\left(\mu, \frac{\sigma^2}{n}\right)$$

Jika sampel cukup

2

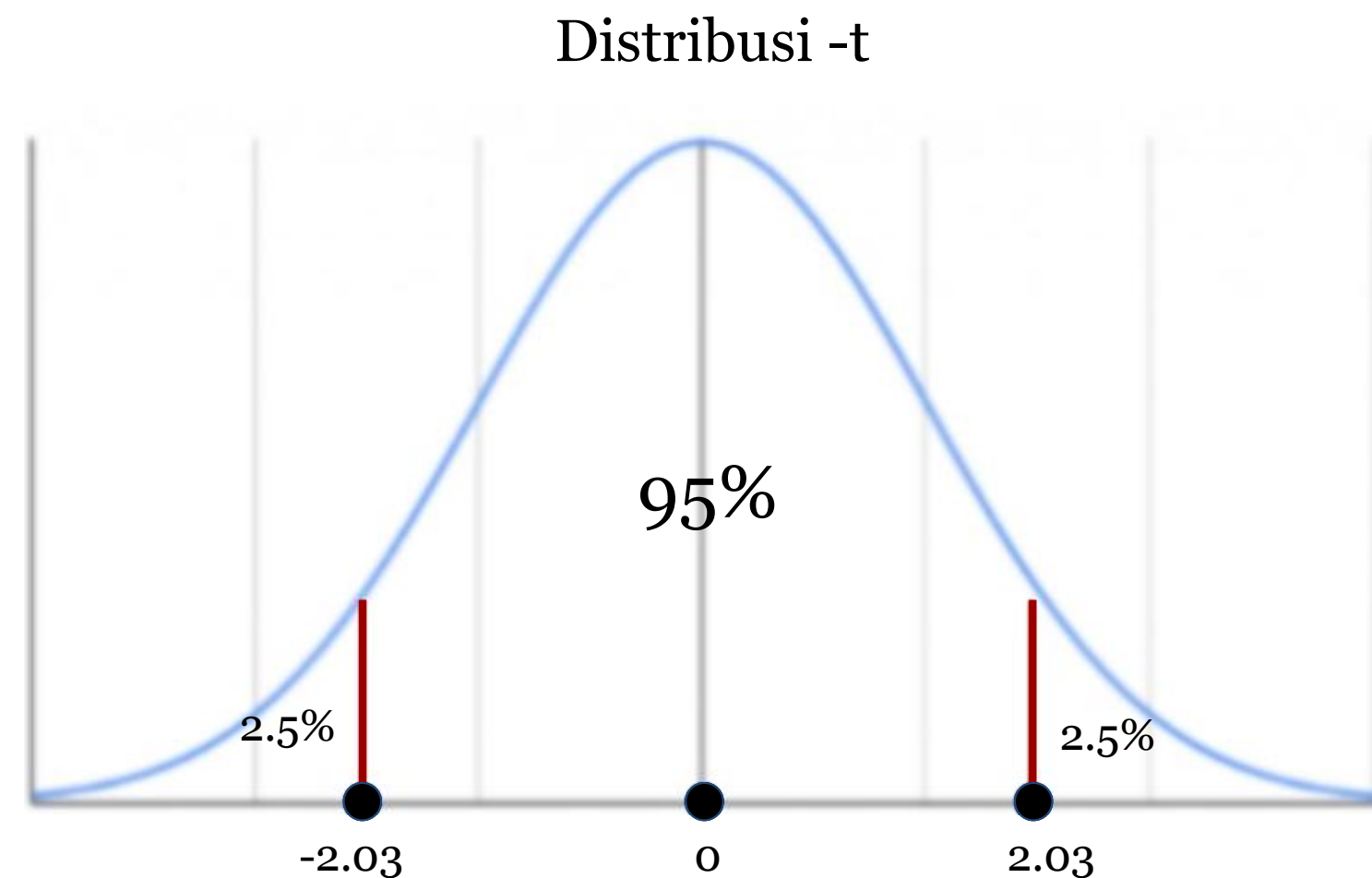


* Replace s in the t-score formula with σ

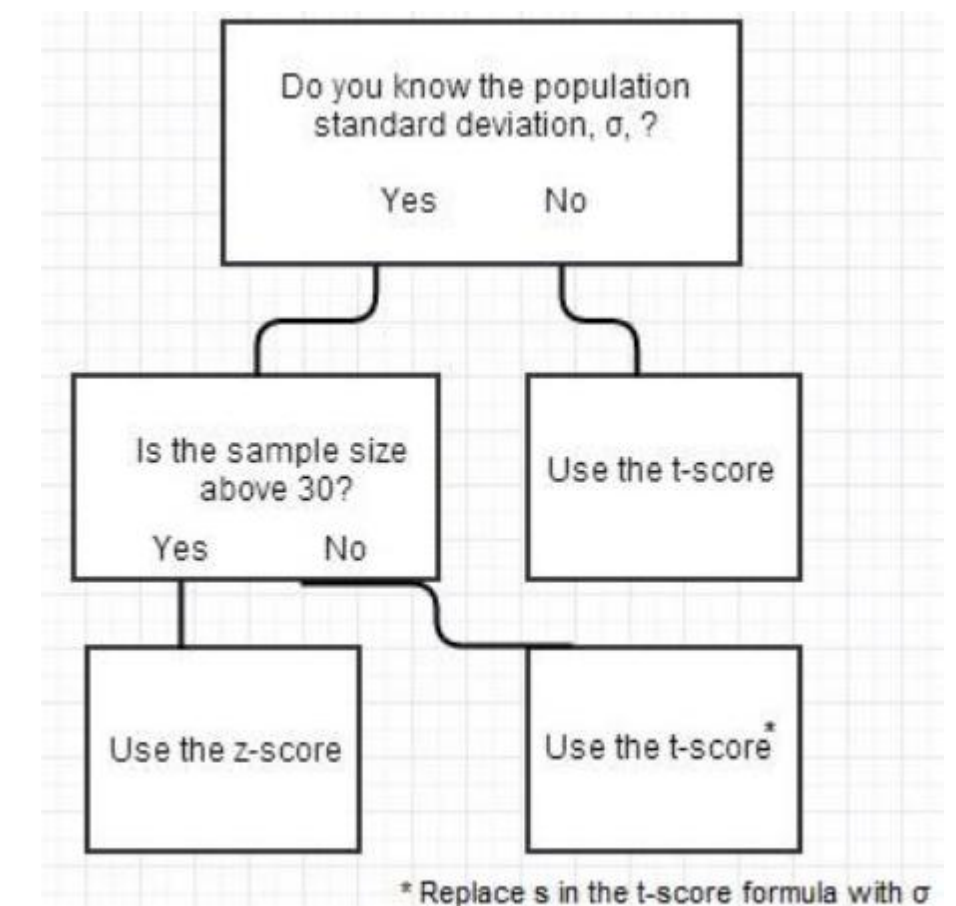
Case 1

Rata-rata berat badan orang-orang di kota A adalah 168 lbs. Seseorang tidak percaya. Dia mengambil sampel 36 orang dan diperoleh rata-rata beratnya 169,5 lbs dengan std 3,9. Dengan 95% interval kepercayaan, Apakah hipotesis benar? berdasarkan hasil sampel

- 2 tail case
- $H_0: \mu_0 = 168$
- $H_a: \mu \neq 168$
- menggunakan distribusi t karena sampel > 30 dan



variansi populasi tidak diketahui



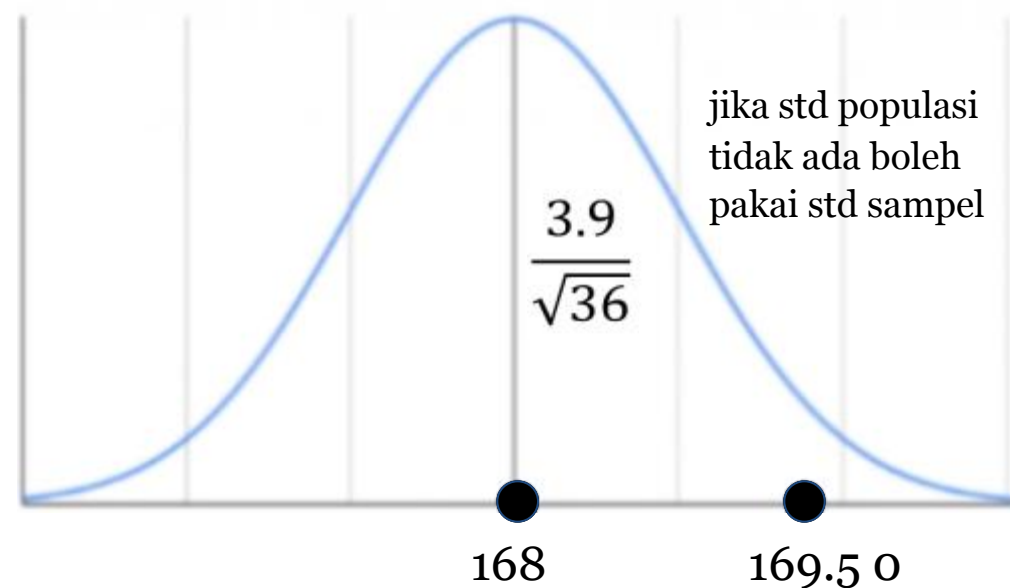
Case 1

Rata-rata berat badan orang-orang di kota A adalah 168 lbs. Seseorang tidak percaya. Dia mengambil sampel 36 orang dan diperoleh rata-rata beratnya 169,5 lbs dengan std 3,9. Dengan 95% interval kepercayaan, Apakah hipotesis benar? berdasarkan hasil sampel

$$H_0: \mu_0 = 168$$

$$H_1: \mu_0 \neq 168$$

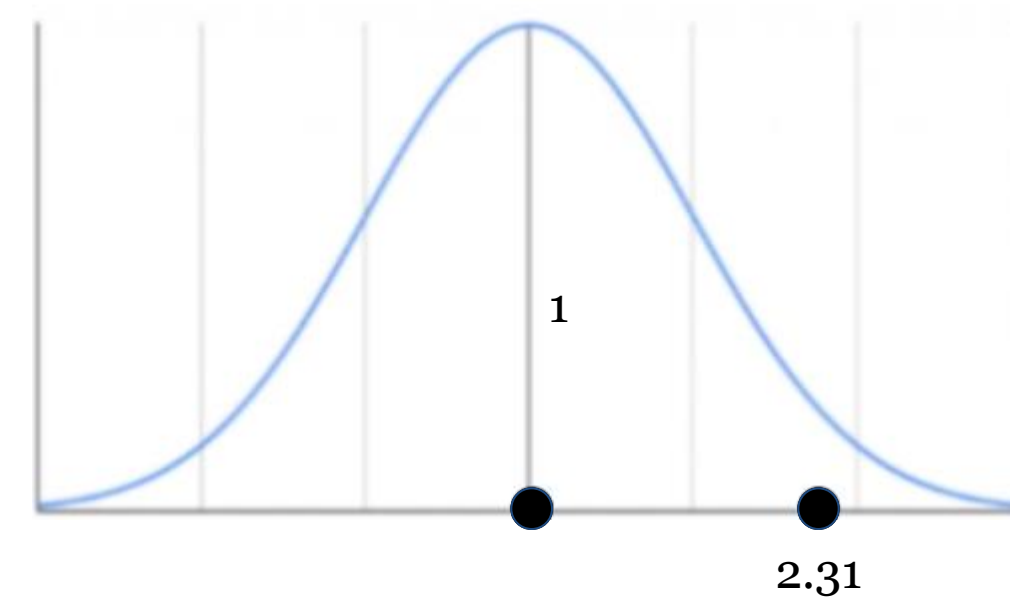
rataan sampel



Standardisasi



rataan sampel

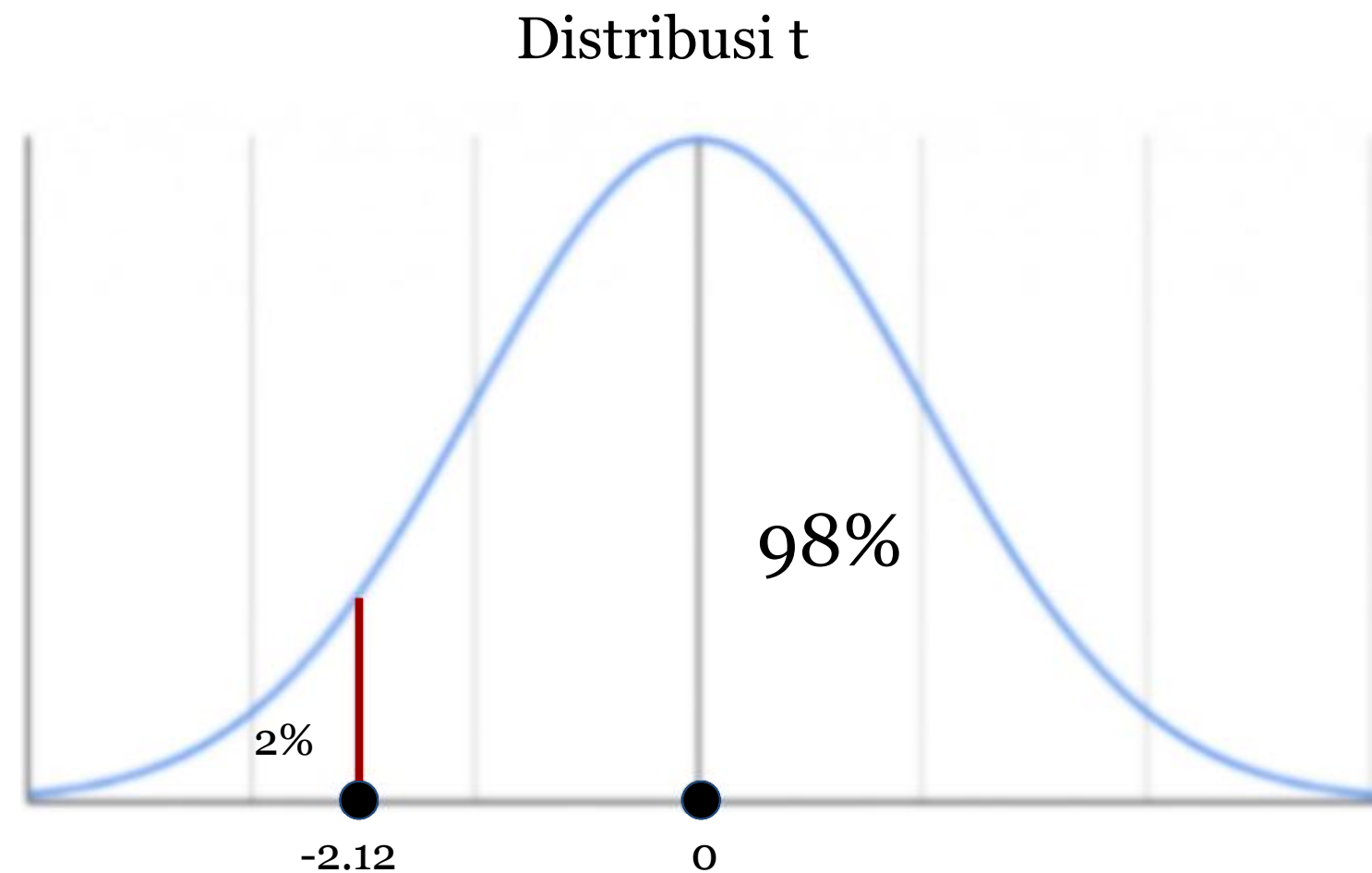


$$(x - \mu) / \sigma$$

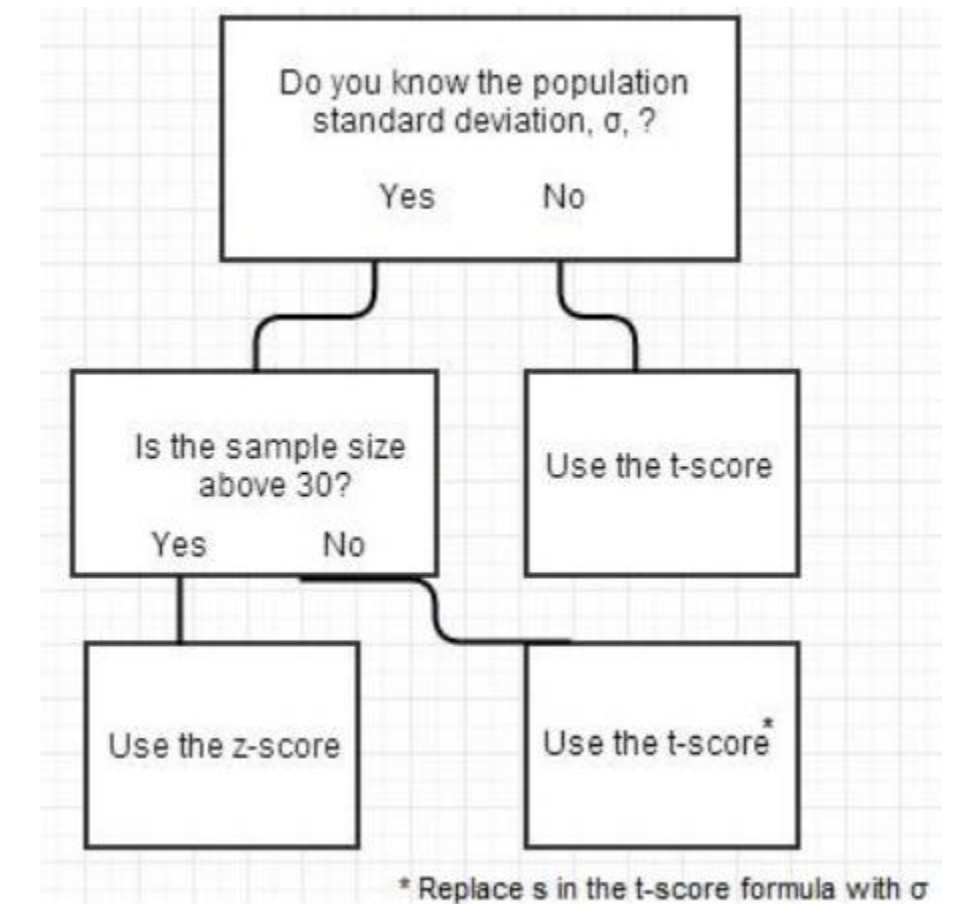
Case 2

Sebuah mobil memiliki garansi 5 tahun. Seorang insinyur percaya bahwa mesin akan rusak kurang dari 5 tahun. Dia mencoba 40 mobil dan ditemukan rata-rata 4.8 tahun dengan standar deviasi 0,5. dengan 98% confidence interval, apakah hipotesis diterima?

- $H_0: \mu \geq 5$
- $H_1: \mu < 5$
- menggunakan distribusi-t karena sampel > 30 dan



variansi populasi tidak diketahui

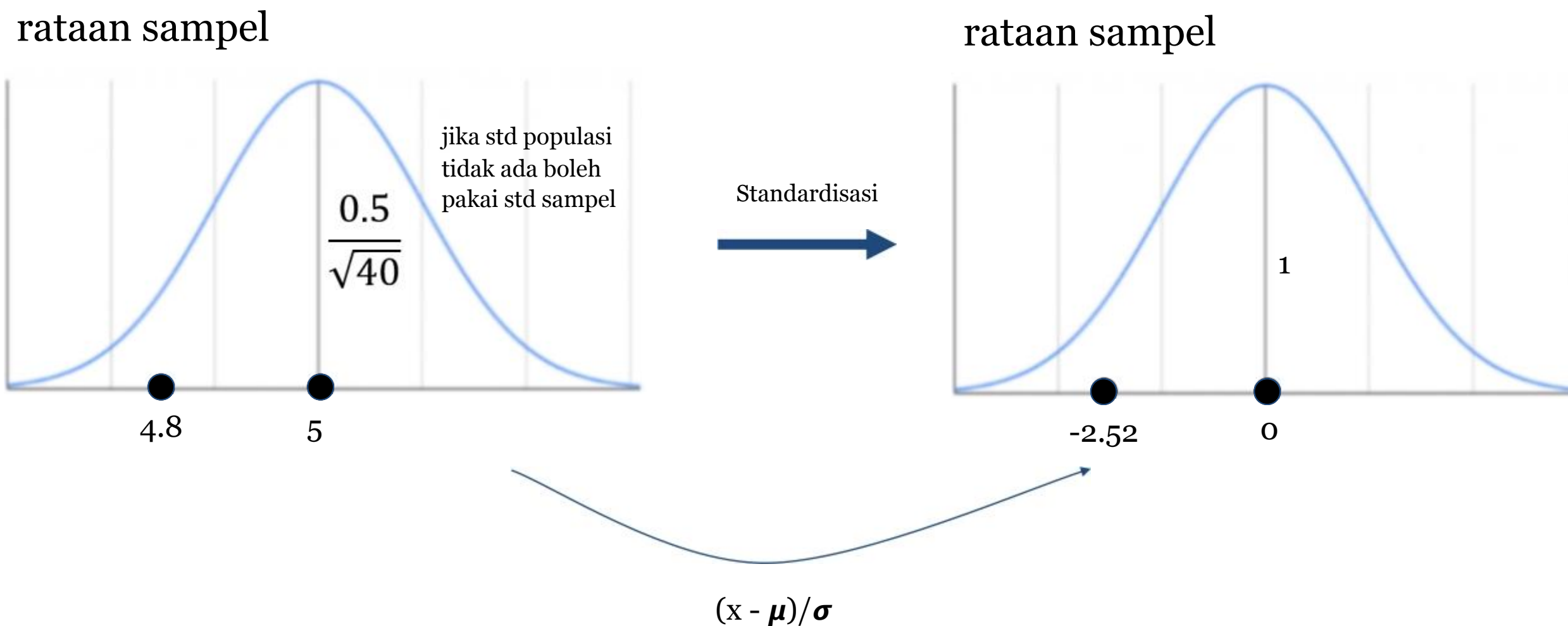


Case 2

Sebuah mobil memiliki garansi 5 tahun. Seorang insinyur percaya bahwa mesin akan rusak kurang dari 5 tahun. Dia mencoba 40 mobil dan ditemukan rata-rata 4.8 tahun dengan standar deviasi 0,5. dengan 98% confidence interval, apakah hipotesis diterima?

$$H_0: \mu_0 \geq 5$$

$$H_1: \mu_0 < 5$$



Summary



Central Tendency

Ukuran pemusatan tergantung dari kondisi data



Variansi

persebaran data (populasi/ sampel)

Kovariansi & korelasi

hubungan linearitas dari dua data

korelasi tidak menyebabkan sebab akibat **Central**



Limit Theorem

rataan sampel selalu berdistribusi normal jika sampel cukup

jika tidak lebih condong ke t distribution

Summary



Probability Mass Function

peluang diskrit

Probability Density Function

peluang kontinu



Ekspektasi

seberapa besar harapan dari suatu eksperimen acak



P-value

Peluang suatu kejadian lebih jarang atau sama dengan kejadian tersebut

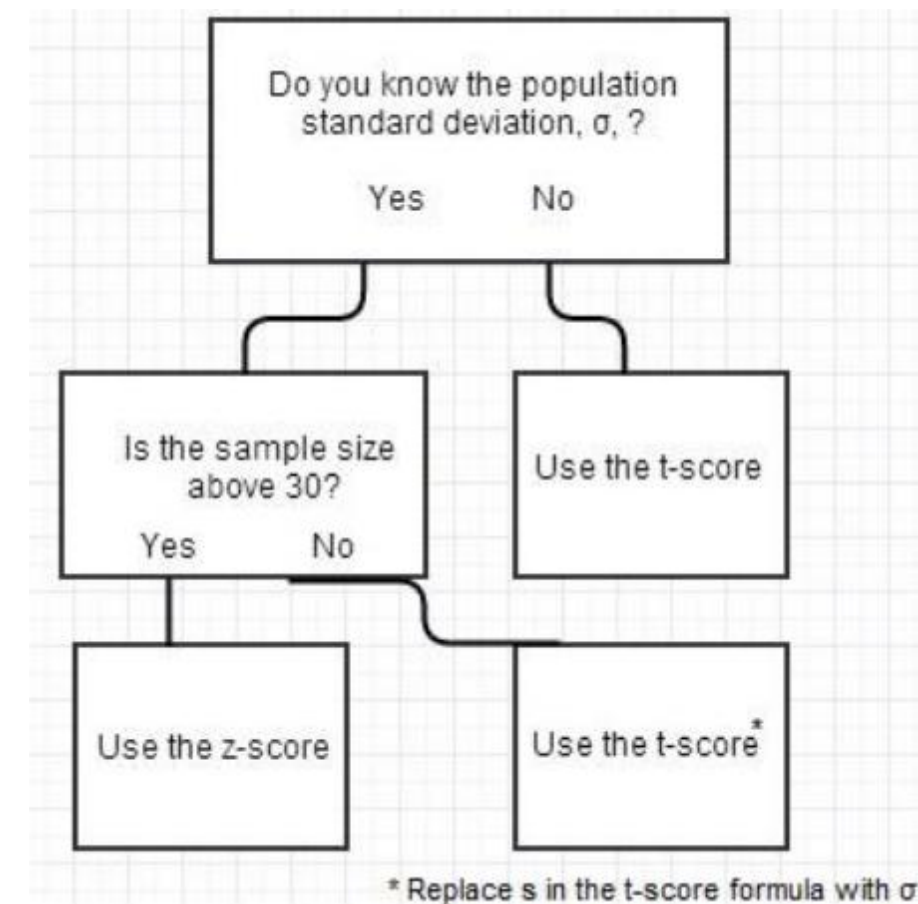


Uji hipotesis metode

1. cari sigma toleransi
2. cari jarak titik hasil sampel
3. jika diluar maka hipotesis ditolak



Uji hipotesis note



Thank you