Penyebaran Data

Tim Ajar Statistik Komputasi

2023/2024

Outlines

- Definisi Penyebaran Data
- Pengukuran Penyebaran Data
- Shifting dan Scaling

Apa itu penyebaran data?

Apa yang dimaksud dengan menyebar? Digunakan untuk apa?

Penyebaran Data

Definisi

- Nilai yang menggambarkan jarak data dari pusat data
- Digunakan untuk mengukur, seberapa baik nilai pemusatan data
- Untuk menentukan nilai pusat data yang paling revelan dengan kondisi data

Pengukuran Penyebaran Data #1

Pengukuran penyebaran data dapat diukur dengan rentang (range) dan interquartile range (IQR). Apa itu?

Rentang (Range)

Nilai Rentang

- Nilai selisih antara amatan terbesar dengan amatan terkecil
- Digunakan untuk penekanan pada nilai ekstrim
- Nilainya dipengaruhi oleh nilai ekstrim (sangat kecil atau sangat besar)

Persamaan

$$rentang = x_{max} - x_{min}$$

Rentang (Range) – Contoh

Nilai Matematika Kelas 12 SMA XYZ

23 56 45 65 59 55 62 54 85 25

Hasil

$$rentang = 85 - 23 = 62$$

Python

```python
import numpy as np

data = np.array([23, 56, 45, 65, 59, 55, 62, 54, 85,25])
data_max = max(data)
data_min = min(data)
range = data_max - data_min
print(range)
```

Interquartile Range (IQR)

Kuartil (Quartile)

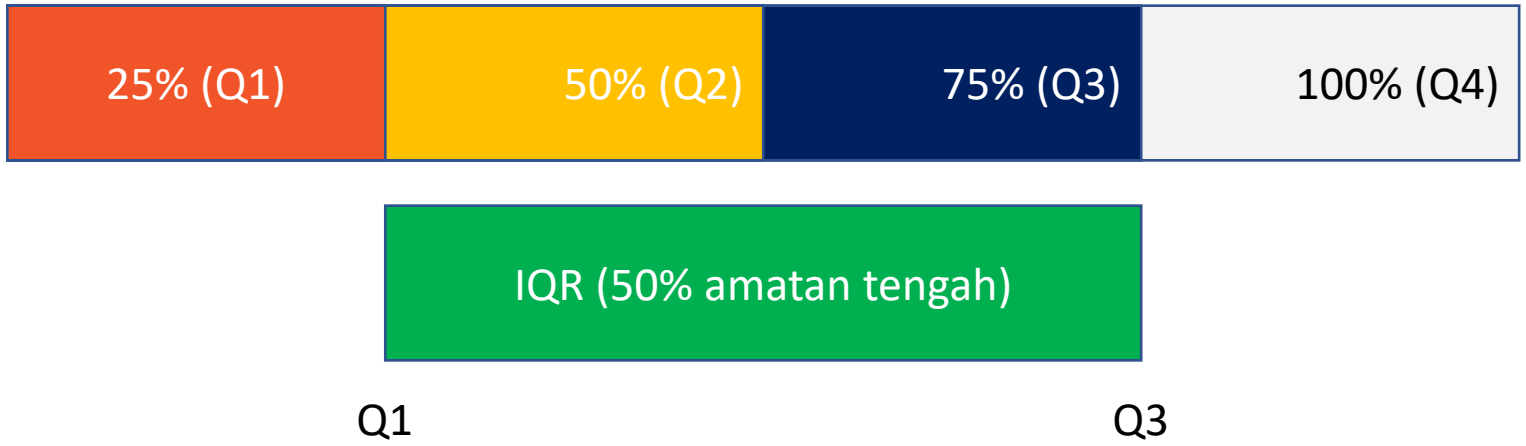
Membagi data kedalam 4 kelompok, 25% pertama data (Q1), 50% data (Q2), 75% data (Q3), dan 100% data (Q4)

Nilai IQR

Nilai yang digunakan untuk mendeskripsikan 50% amatan tengah

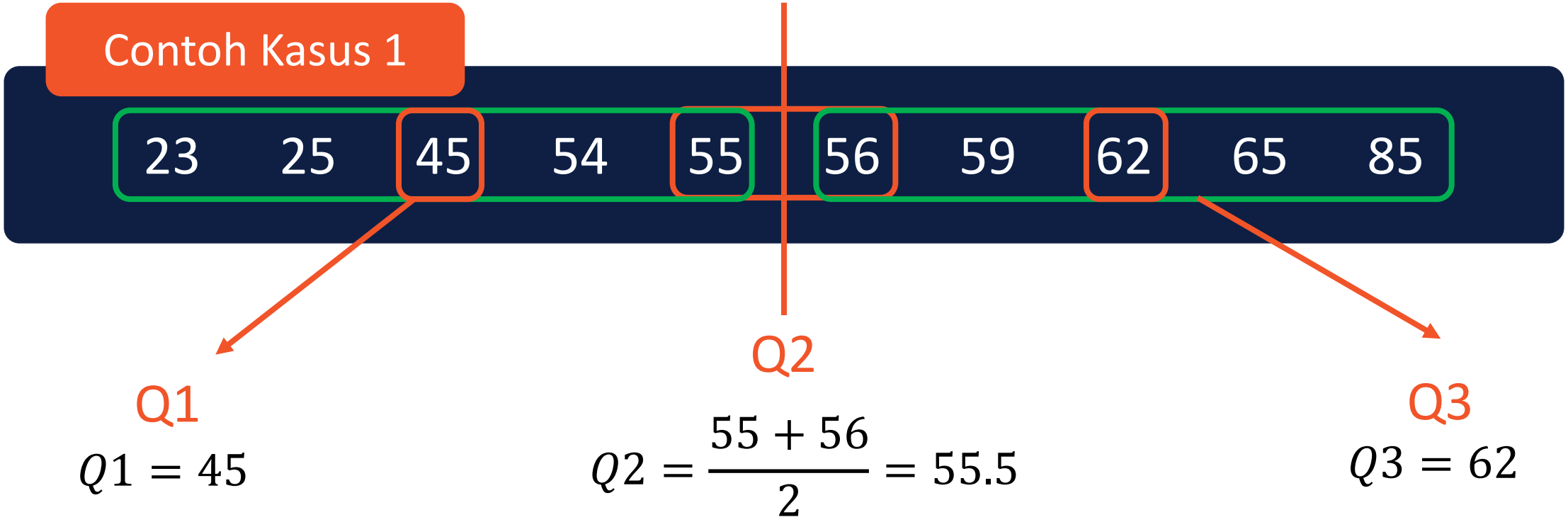
Definisi Matematis

$$IQR = Q3 - Q1$$


25% (Q1)
50% (Q2)
75% (Q3)
100% (Q4)
IQR (50% amatan tengah)
Q1
Q3

Bagaimana mendapatkan nilai Q3 dan Q1? (1)

Contoh Kasus 1



$Q1 = 45$

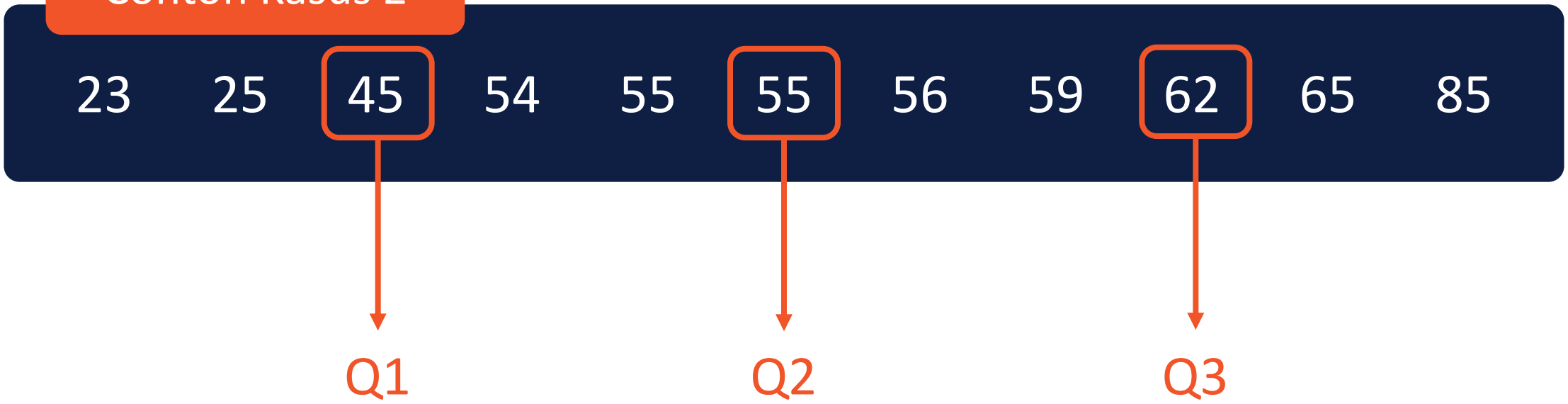
$$Q2 = \frac{55 + 56}{2} = 55.5$$

$Q3 = 62$

$$IQR = 62 - 45 = 17$$

Bagaimana mendapatkan nilai Q3 dan Q1? (3)

Contoh Kasus 2



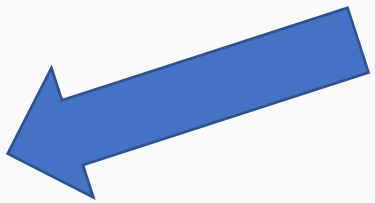
Bagaimana mendapatkan nilai Q3 dan Q1? (3)

Python

```python
import numpy as np
import pandas as pd

data_baru = np.array([23, 56, 45, 65, 59, 55, 62, 54, 85, 25, 55])
print(pd.DataFrame(data_baru).describe())

'''
Output
count  11.000000
mean   53.090909
std    17.466852
min    23.000000
25%    49.500000
50%    55.000000
75%    60.500000
max    85.000000
'''
```

Mengapa hasil pada Python Pandas berbeda?

```python
import numpy as np
import pandas as pd

data_baru = np.array([23, 56, 45, 65, 59, 55, 62, 54, 85, 25, 55])
print(pd.DataFrame(data_baru).describe())

'''
Output
count  11.000000
mean   53.090909
std    17.466852
min    23.000000
25%    49.500000
50%    55.000000
75%    60.500000
max    85.000000
'''
```



Perhatikan nilai Q1, Q2, dan Q3

Pada Pandas, pendekatan untuk mencari nilai percentil dan quantile (termasuk quartile) yaitu dengan menggunakan nilai **LINEAR INTERPOLATION**

Nilai *linear interpolation* digunakan jika nilai yang dicari berada pada 2 nilai (jumlah data genap)

Nilai linear interpolation didapatkan dari,

$$x_{linear} = i + (j - i) * fraction$$

Untuk mendapatkan nilai rata-rata diantara kedua nilai, pada Pandas harus menggunakan nilai 'midpoint' dimana nilainya adalah $(i + j)/2$

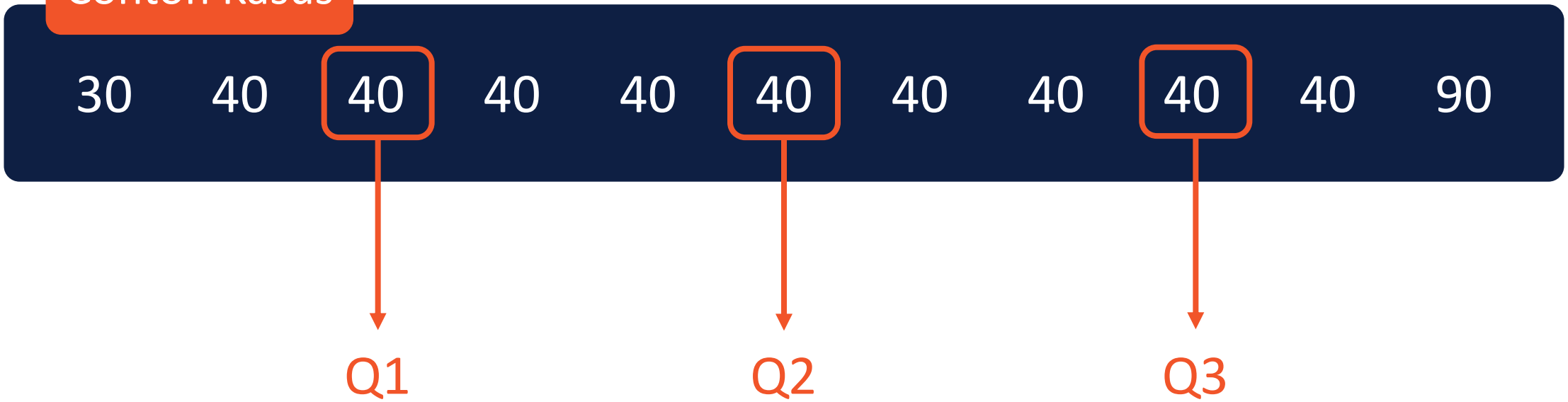
Metode perhitungan percentile, quantile, dan quartile pada library-library Python

Contoh pada Numpy, Pandas, dan Scipy

- Linear → $i + (j - i) * fraction$ (**digunakan secara default**)
- Lower → Nilai i
- Higher → Nilai j
- Nearest → Nilai i atau j tergantung mana yang terdekat
- Midpoint → $(i + j)/2$

Anomali Pada Nilai IQR

Contoh Kasus



$IQR = 40 - 40 = 0$???

Nilai IQR 0 menandakan amatan mempunyai nilai yang identik

Varians (*Variance*)

Definisi

Nilai yang mendeskripsikan seberapa jauh data menyebar dari niai mean

Populasi

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sampel

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Simpangan Baku (*Standard Deviation*) (1)

Definisi

Nilai yang mendeskripsikan seberapa besar variasi data terhadap nilai mean

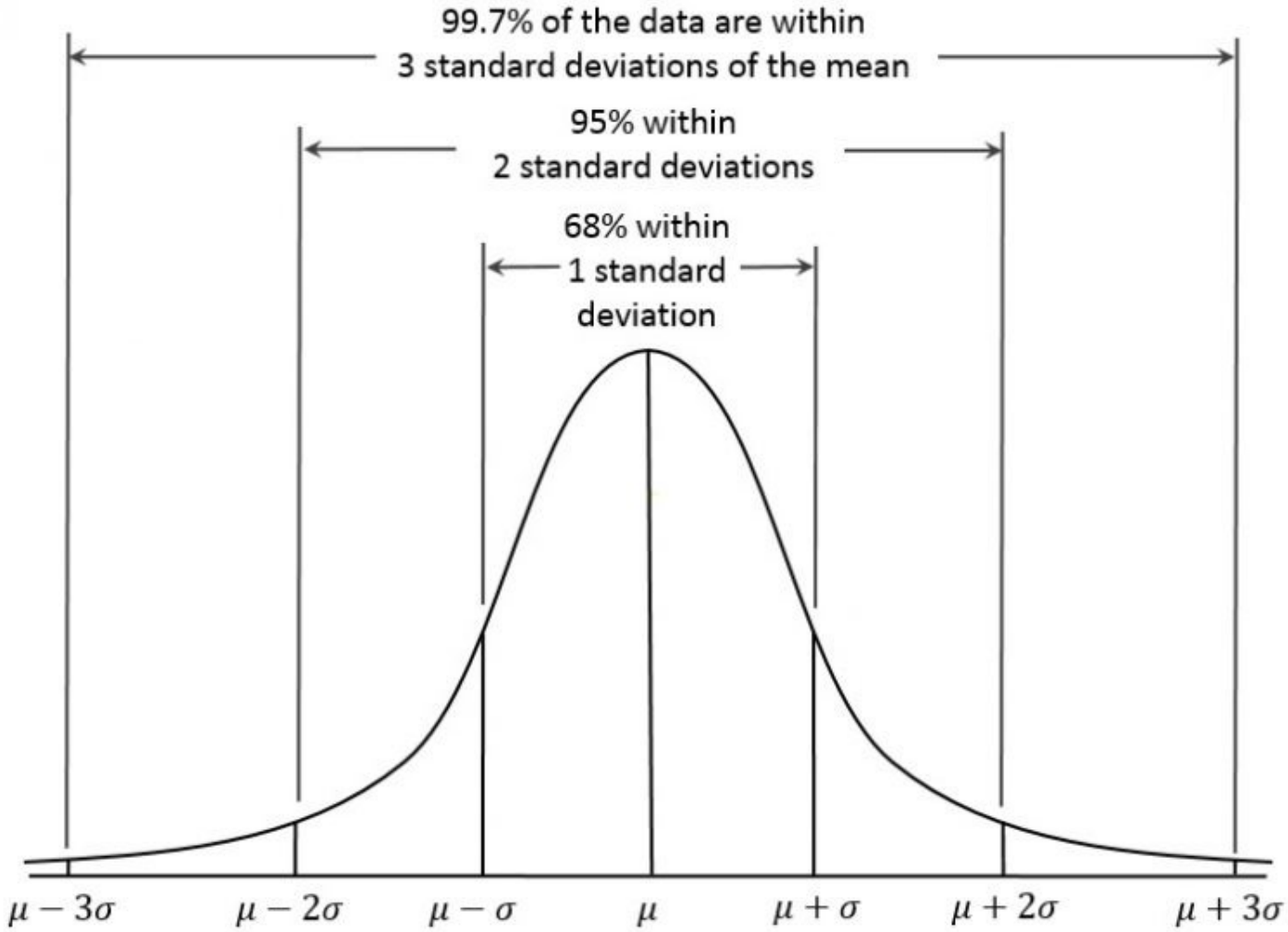
Populasi

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sampel

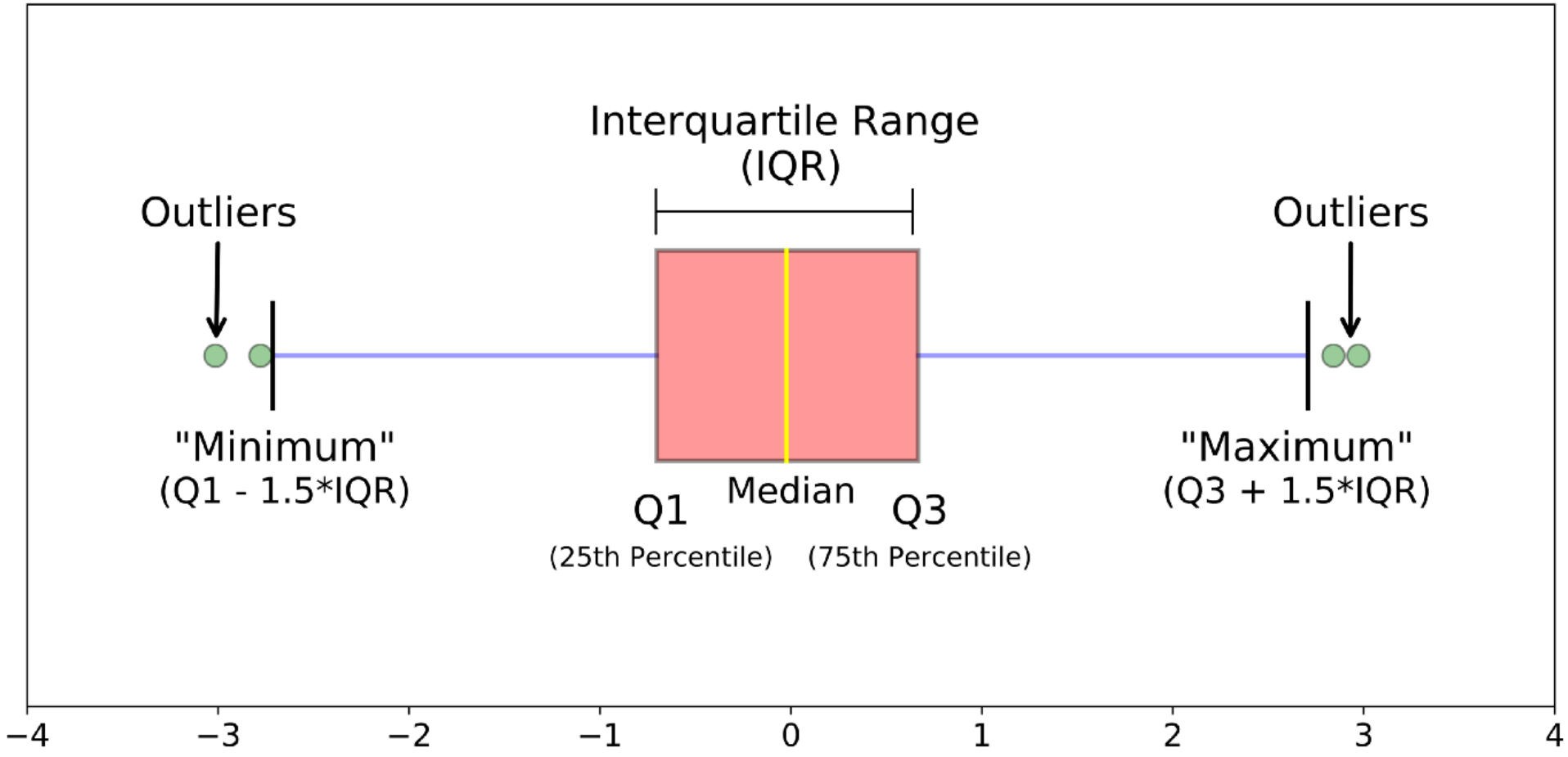
$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Simpangan Baku (*Standard Deviation*) (2)



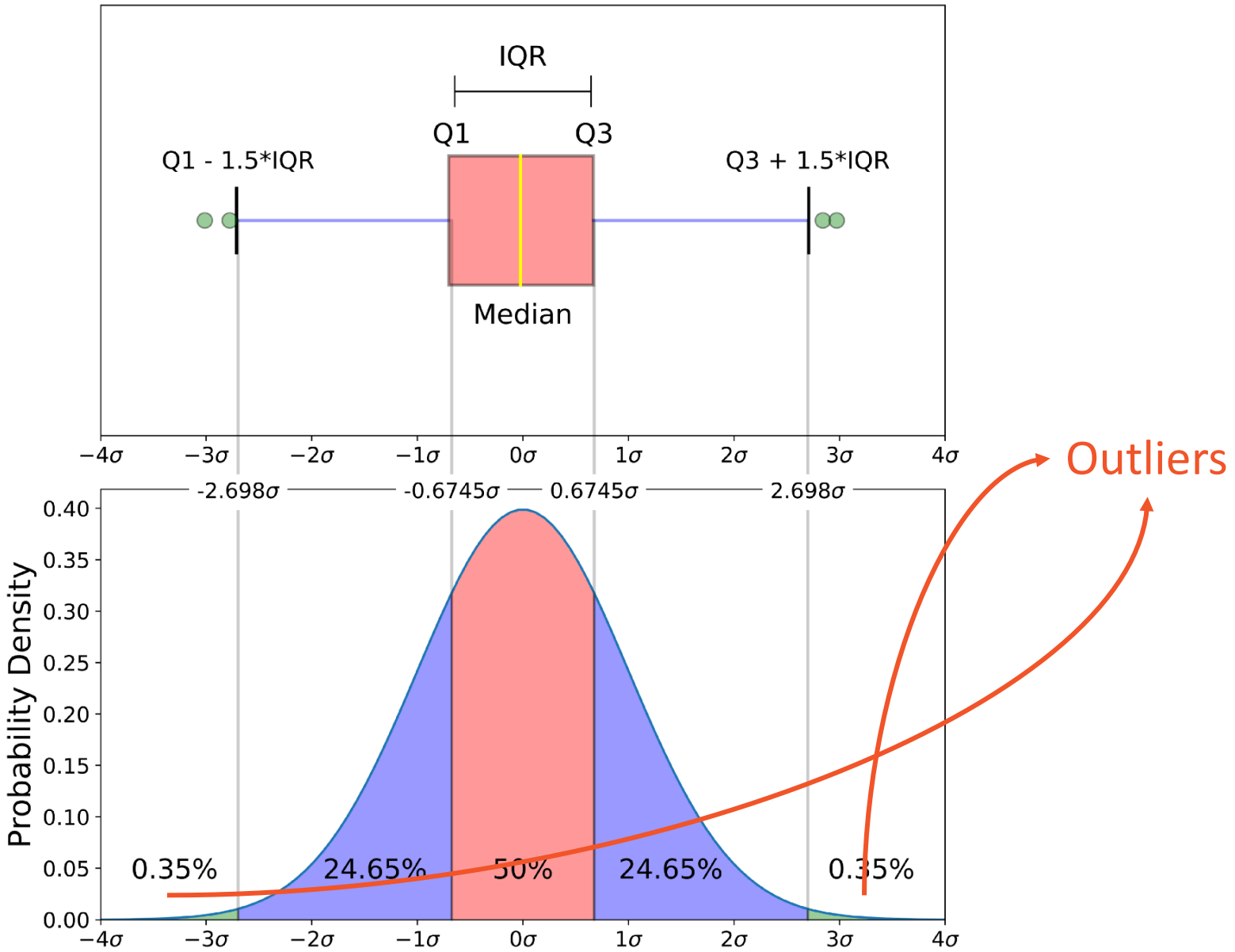
https://s4be.cochrane.org/blog/2018/09/26/a-beginners-guide-to-standard-deviation-and-standard-error/#:~:text=Standard%20deviation%20tells%20you%20how,standard%20deviations%20of%20the%20mean.

Pengenalan Box Plot



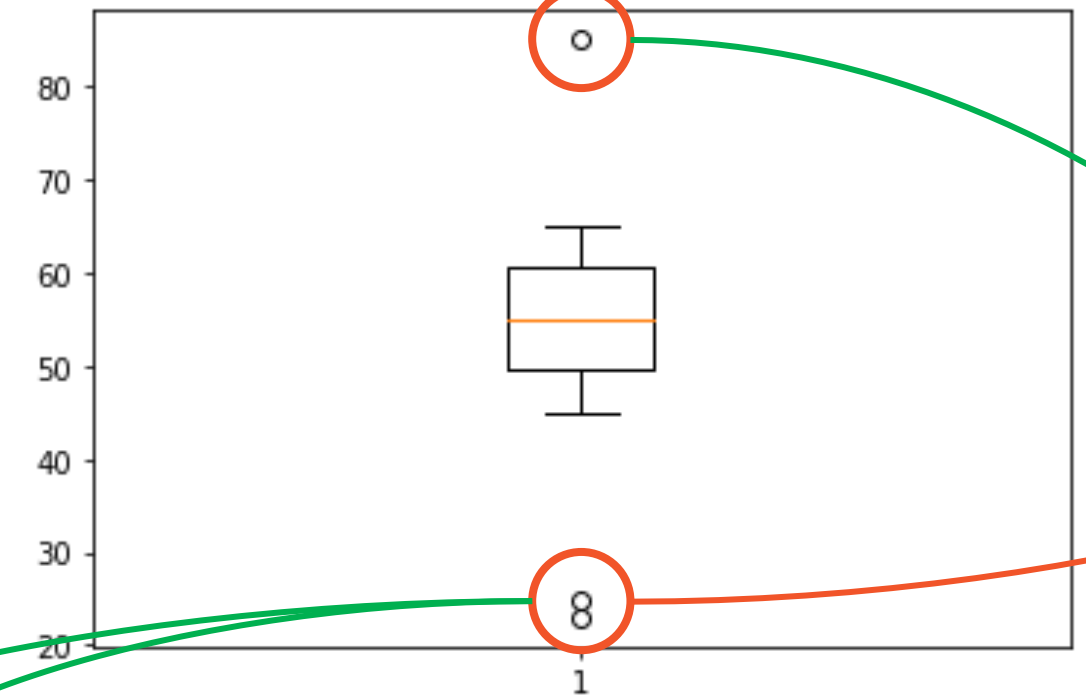
Sumber: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51#:~:text=A%20boxplot%20is%20a%20standardized,and%20what%20their%20values%20are.

Box Plot Pada Distribusi Normal



Sumber: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51#:~:text=A%20boxplot%20is%20a%20standardized,and%20what%20their%20values%20are.

Contoh Kasus Box Plot



Outliers

23 25 45 54 55 55 56 59 62 65 85

Shifting dan *Scaling*

Bagaimana perubahan data mempengaruhi pemusatan dan penyebaran data

Shifting (Pergeseran)

Apa itu *shifting*?

Menggeser nilai pada kelompok data dengan cara **menambahkan** (+) atau **menggurangkan** (-) semua nilai pada kelompok data dengan nilai tertentu

Contoh 1

4, 5, 8, 12 → +2 → 6, 7, 10, 14

Contoh 2

4, 5, 8, 12 → -2 → 2, 3, 6, 10

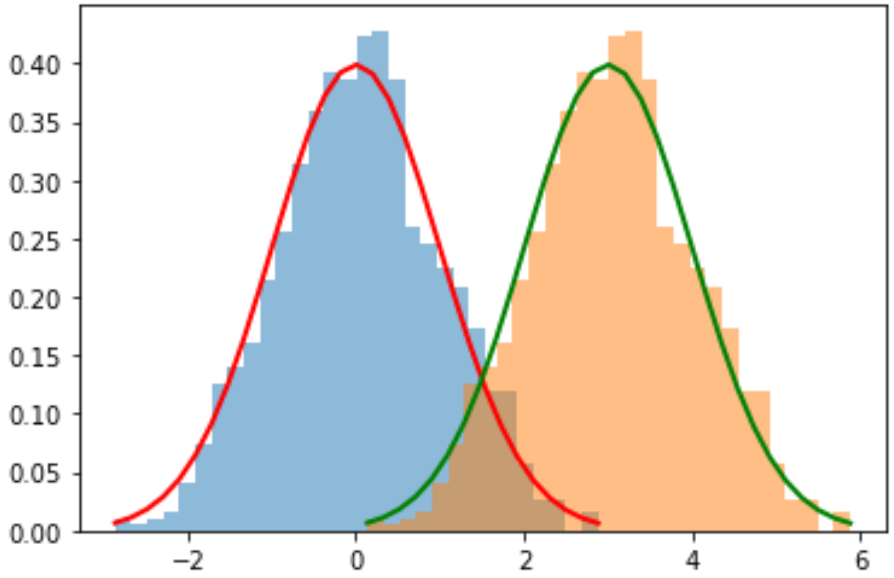
Dampak *Shifting* (1)

Data Asli

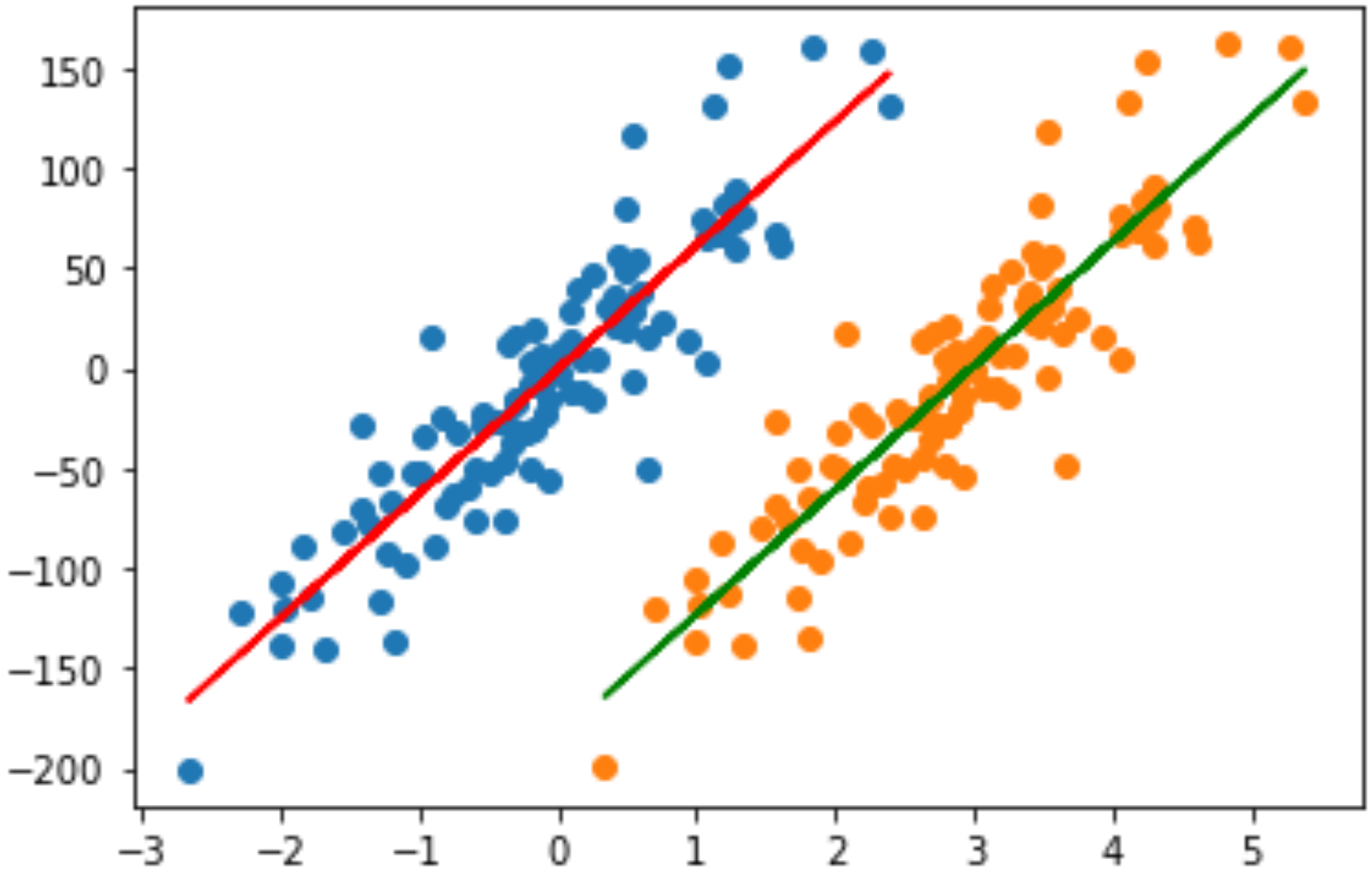
4, 5, 8, 8, 12
Mean = 7.4
Median = 8
Modus = 8
Rentang = 8
IQR = 3
STD = 2.8

Data Shift

+2 → 6, 7, 10, 10, 14
Mean = 9.4
Median = 10
Modus = 10
Rentang = 8
IQR = 3
STD = 2.8



Dampak *Shifting* (2) – Prior Shift di ML



Scaling (Penyekalaan)

Apa itu *scaling*?

Menyekalakan nilai pada kelompok data dengan cara **mengkalikan** atau **membagi** semua nilai pada kelompok data dengan nilai tertentu

Contoh 1

4, 5, 8, 12 → ×2 → 8, 10, 16, 24

Contoh 2

4, 5, 8, 12 → ÷2 → 2, 2.5, 4, 6

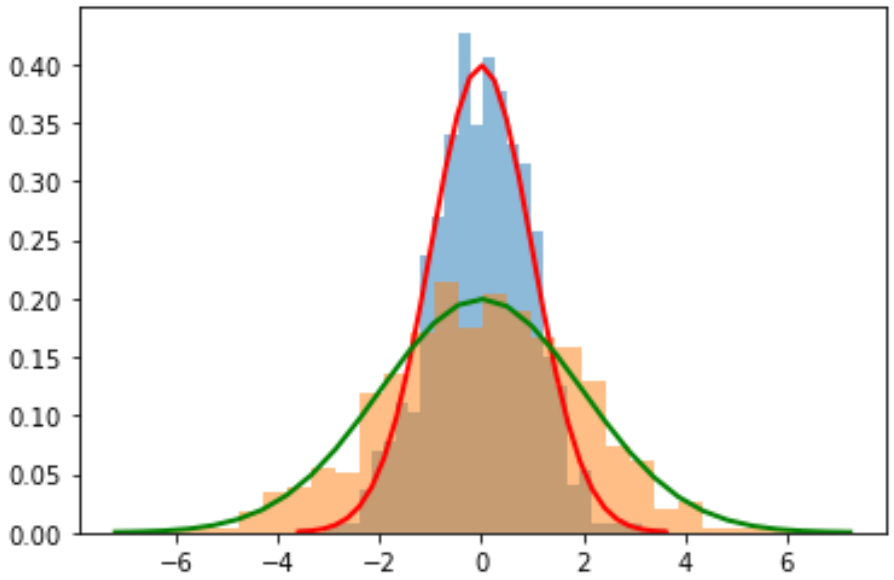
Dampak *Scaling*

Data Asli

4, 5, 8, 8, 12
Mean = 7.4
Median = 8
Modus = 8
Rentang = 8
IQR = 3
STD = 2.8

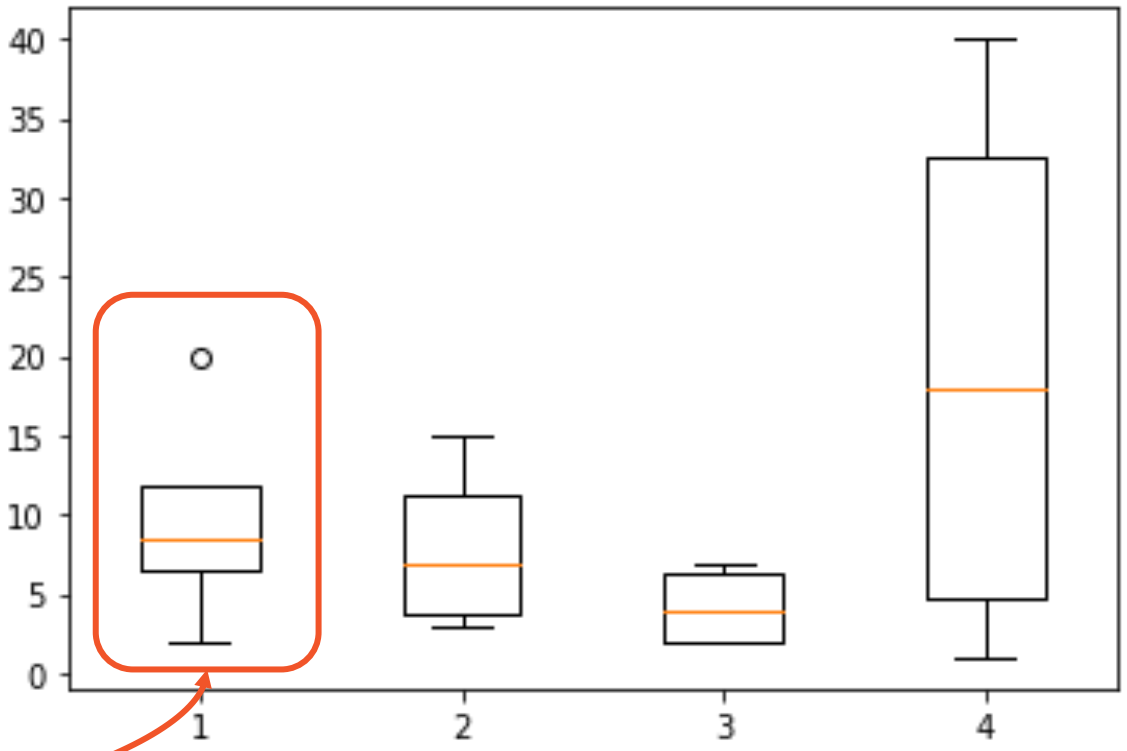
Data Scale

×2 → 8, 10, 16, 16, 24
Mean = 14.8
Median = 16
Modus = 16
Rentang = 16
IQR = 6
STD = 5.6



Manfaat *Scaling* – Normalisasi Data (1)

```python
import numpy as np
import matplotlib.pyplot as plt

data_raw = np.array([
    [2, 3, 7, 30],
    [9, 4, 6, 1],
    [8, 15, 2, 40],
    [20, 10, 2, 6]
])
```

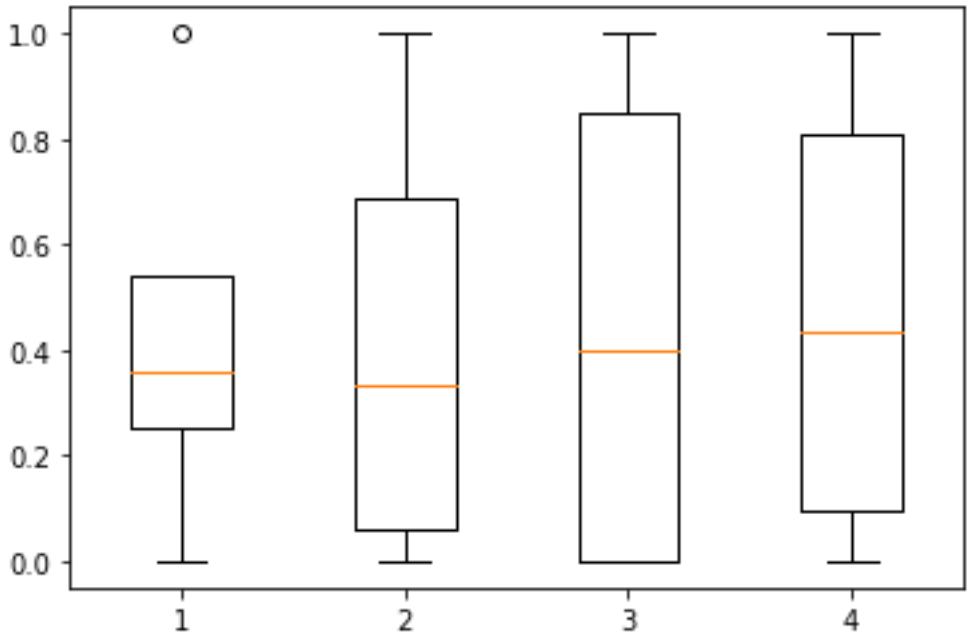


Dibaca secara kolom oleh matplotlib

Manfaat *Scaling* – Normalisasi Data (2)

```python
# Normalisasi Dada - MinMax Scaler
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler

data_raw = np.array([
    [2, 3, 7, 30],
    [9, 4, 6, 1],
    [8, 15, 2, 40],
    [20, 10, 2, 6]
])

scaler = MinMaxScaler()
data_scale = scaler.fit_transform(data_raw)
```

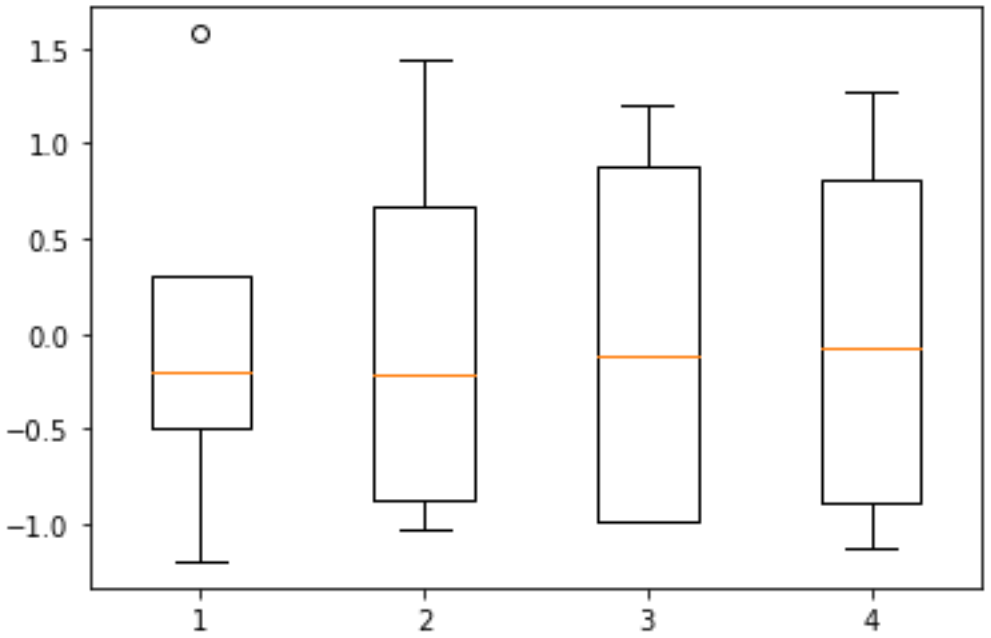
$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$



Manfaat *Scaling* – Standarisasi Data

```python
# Standarisasi Data
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler, StandardScaler

data_raw = np.array([
    [2, 3, 7, 30],
    [9, 4, 6, 1],
    [8, 15, 2, 40],
    [20, 10, 2, 6]
])

stand = StandardScaler()
data_stand = stand.fit_transform(data_raw)

plt.boxplot(data_stand)
```

$$x_{stand} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$




thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
dank je
merci
obrigado
dziękuję
bedankt
hvala
mauruuru
sukriya
kop khun krap
go raibh maith agat
terima kasih
grazie
arigatō
dakujem
tapadh leat
mochchakkeram
спасибо
рахмат
Баярлалаа
faafetai lava
kiitos
dankie
bayarlalaa
gracie
enkosi
sobodi
dekuji
mesi
sagolun
chnorakaloutioun
gracies
sulpay
taiku
takk
djiere dieuf
tau
mamnun
мерси
dhanyavadagalu
shukriya
diolch
rahmat
tanemirt
najis tuke
kam sah hamnida
감사합니다
xiexie
ευχαριστώ
хвала
asante
manana
obrigada
murakoze
tenki
trugarez
merci