

# **Eksplorasi Data**


Pertemuan 6

# Learning Objective

Mahasiswa mengetahui sumber data dan mendapatkannya



Mahasiswa mampu menelaah data dengan statistika



Mahasiswa mampu merepresentasikan data

# Course Materials

Sumber  
Data

Struktur  
Dataset

Penelaahan  
Data

# Sumber Data

Membahas sumber dataset

# Sumber Data

---

Sumber  
Internal

Spreadsheet (Excel, CSV, JSON, dll)

---

Database: diperoleh melalui query SQL

---

Dokumen Multimedia (audio, video)

---

Sumber  
Eksternal

Repositori data terbuka

---

Public domain web pages

---

## (1.2) 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>)
- ☐ **Google Dataset** (<https://datasetsearch.research.google.com>)
- ☐ 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/>)

# (1.3) UCI Machine Learning Repository

← → ↺

archive.ics.uci.edu/ml/index.php

Google

🔍


☆

⚙️

□

M

UCI



Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About

Citation Policy

Donate a Data Set

Contact

Search

Repository

Web

Google


View ALL Data Sets

Check out the [beta version](#) of the new UCI Machine Learning Repository we are currently testing! [Contact us](#) if you have any issues, questions, or concerns. [Click here to try out the new site](#)


Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 622 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:



In Collaboration With:



Latest News:

09-24-2018:

Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidoul

04-04-2013:

Welcome to the new Repository admins Kevin Bache and Moshe Lichman!

03-01-2010:

[Note](#) from donor regarding Netflix data

10-16-2009:

Two new data sets have been added.

09-14-2009:

Several data sets have been added.

03-24-2008:


New data sets have been added!

06-25-2007:


Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Newest Data Sets:


06-05-2021:

 [Average Localization Error \(ALE\) in sensor node localization process in WSNs](#)


05-25-2021:

 [9mers from cullpdb](#)

05-18-2021:


 [TamilSentiMix](#)

05-02-2021:


 [Accelerometer](#)

Most Popular Data Sets (hits since 2007):


5182037:

 [Iris](#)


2703927:

 [Adult](#)

2203374:

 [Dry Bean Dataset](#)

2114385:

 [Wine](#)

# (1.4) Kaggle

Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Search datasets

Filters

All datasets

Computer Science

Education

Classification

Computer Vision

NLP

Data Visualization

Pre-Trained Model

Trending Datasets

See All

**car data**

Athira G · Updated a day ago

Usability 6.5 · 4 kB

1 File (CSV)

13

**Apple's Historical Financials**

The Devastator · Updated 11 day...

Usability 10.0 · 1 kB

1 File (CSV)

20

**PIB-GDP Global by countries since 1980 t...**

Frederick Adolfo Salazar Sanche...

Usability 10.0 · 378 kB

2 Files (CSV)

15

**ipl 2023**

Sudarshan Shinde · Updated 2 d...

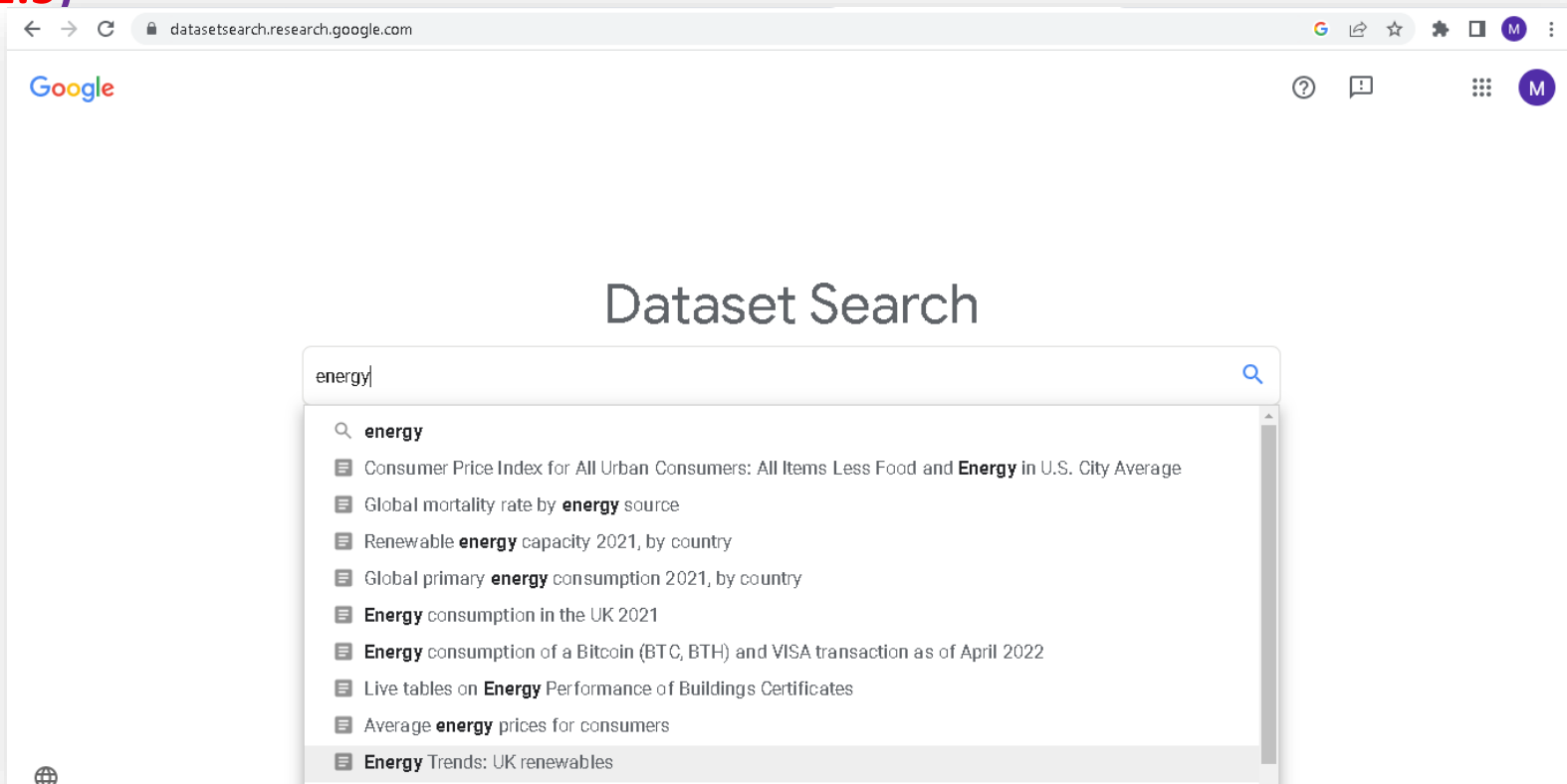
Usability 6.5 · 7 kB

1 File (CSV)

6



# (1.5) Google Dataset



# Struktur Dataset

Membahas struktur dataset dan tipe data

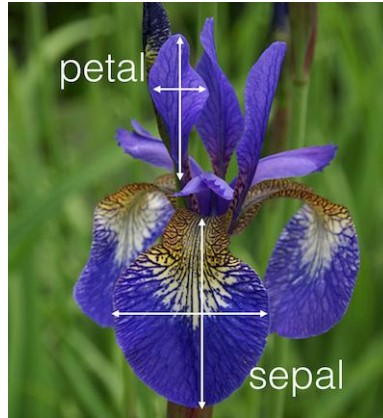
# Struktur Dataset

Attribute/Feature/Dimension

Class/Label/Target

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>

Datum



Record/  
Object/  
Sample/  
Tuple/  
Data

# Istilah Pada Struktur Dataset

- **Dataset** (himpunan data): sekumpulan data
- **Datum** (butir data): satuan terkecil data
- **Record/Object/Sample/Tuple/Data**: kumpulan butir data yang membawa satu kesatuan makna (mendeskripsikan satu objek) tertentu
- **Atribut/Feature/Dimension**: karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi
- **Atribut Class/Label/Target**: atribut yang menjadi tujuan untuk diisi oleh proses data mining

# Tipe Data Berdasarkan Susunannya

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

	Nominal/Kategorikal	Ordinal	Interval	Rasio
Sifat Himpunan awal	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, kuning, dll)	Grade nilai (A, B, C, D, E, dst)	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	$=, \neq$	$=, \neq, <, >$	$=, \neq, <, >, +, -$	$=, \neq, <, >, +, -, \times, \div$

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

# Contoh Model Data Tabular

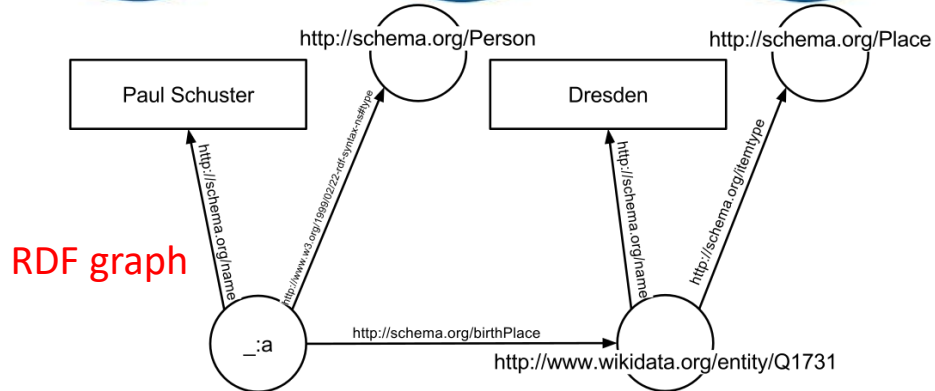
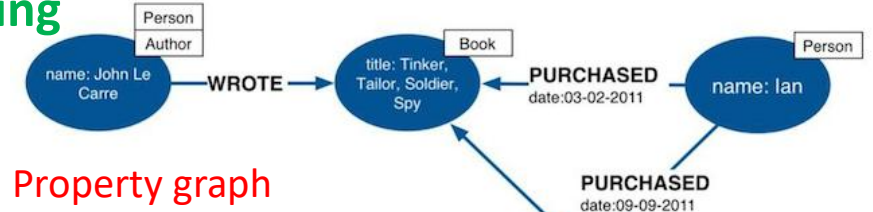
- ❑ Terdiri dari N buah rekord (*record*)
- ❑ Masing-masing rekord mengandung D buah atribut
- ❑ Rekord = baris, *data point*, instans, *example*, transaksi, tupel, entitas, objek, vector fitur.
- ❑ Atribut = kolom, *field*, dimensi, fitur.
- ❑ Atribut yang sama untuk setiap rekord biasanya diasumsikan 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



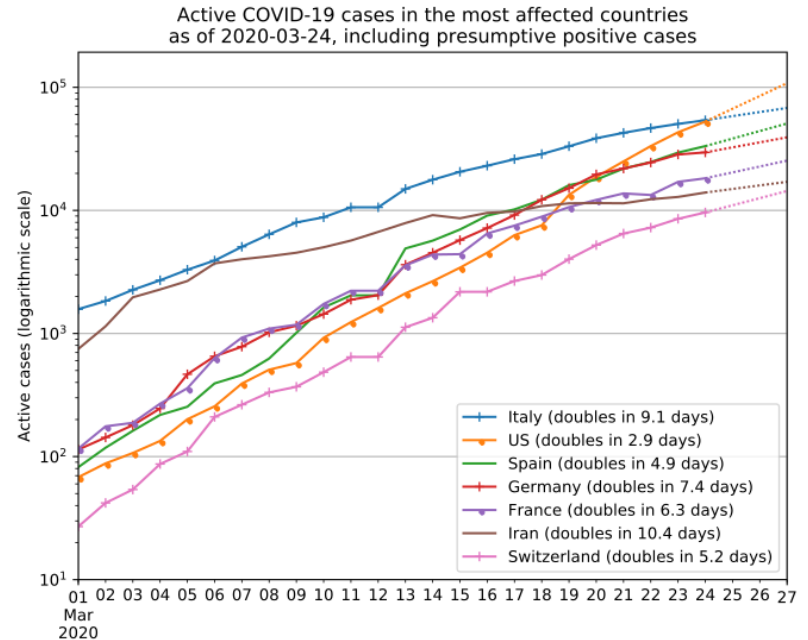
## Contoh Model Data Graf/Jejaring

- ❑ 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/Time Series

- ❑ Tersusun dari record-record yang terhubung secara sekuensial.
- ❑ Contoh: data dari sensor suhu selama suatu rentang waktu.
- ❑ Struktur tersirat dari urutan kemunculan record
- ❑ 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*.



# Telaah Data

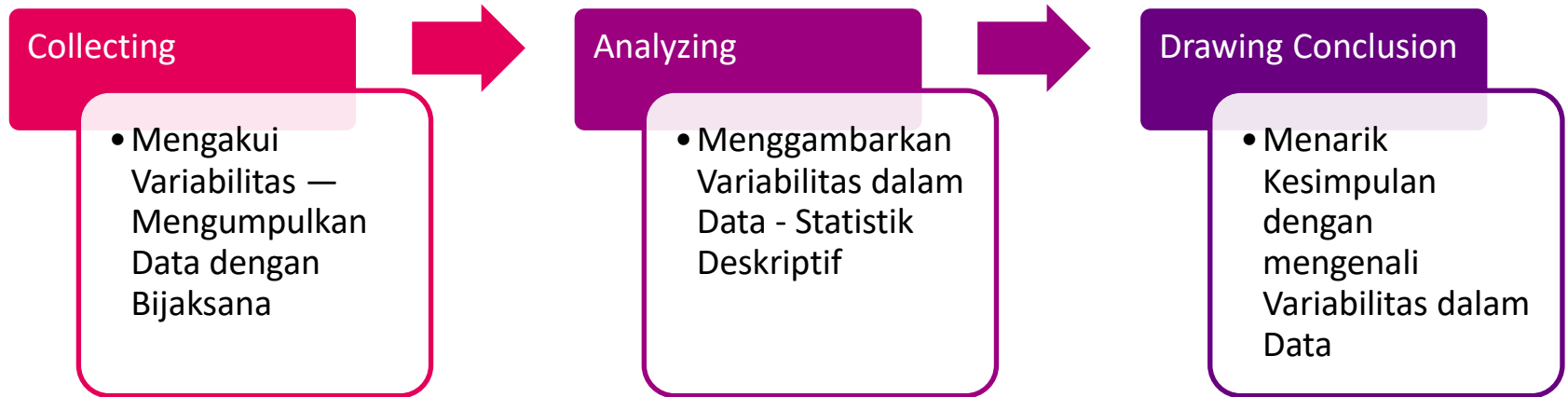
Membahas cara-cara melakukan telaah data

# Statistik

- ❑ Statistik bukan tentang angka.
- ❑ Statistik berbicara **tentang data**, yaitu angka-angka dalam **konteks**. Konteks yang membuat suatu masalah bermakna dan sesuatu yang layak dipertimbangkan.
- ❑ Statistik mempelajari bagaimana **membuat penilaian cerdas dan keputusan** berdasarkan informasi terhadap adanya **ketidakpastian** dan **variasi**.

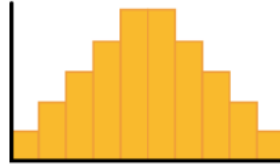
**Statistika** adalah disiplin ilmu yang menyediakan metode untuk membantu memahami data

# Proses Statistik

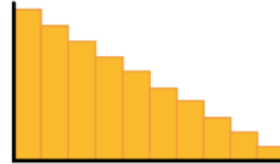


# Distribusi Data

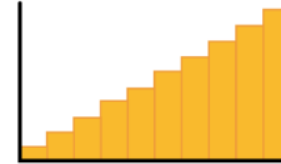
Symmetric (normal)  
vs  
Skewed & Uniform  
Distribution



**Normal distribution**  
(unimodal, symmetric,  
the "bell curve")



**Right-skewed  
distribution**  
(Positively-skewed)



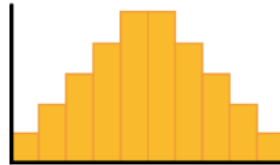
**Left-skewed  
distribution**  
(Negatively-skewed)



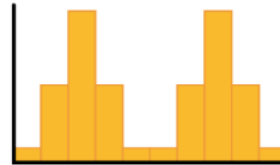
**Uniform distribution**  
(equal spread,  
no peaks)

---

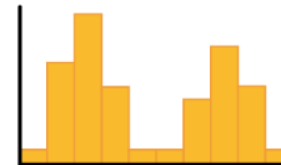
Unimodal  
vs  
Bimodal  
Distribution



**Normal distribution**  
(unimodal, symmetric,  
the "bell curve")



**Symmetric bimodal  
distribution**  
(two modes)



**Non-symmetric  
bimodal distribution**  
(two modes)

# Tendensi Sentral

- ❑ **Tendensi Sentral** merupakan **nilai tunggal** yang menunjukkan **titik tengah** dari suatu dataset untuk mengetahui dimana posisi banyak nilai data berkumpul di dalam distribusi.
- ❑ **Nilai tunggal** tersebut berupa **skor rata-rata (*average*)** dari keseluruhan data.

**Tendensi sentral** adalah **nilai** yang menjadi **pusat** suatu distribusi data

## (4.1) Kegunaan Tendensi Sentral

- ❑ **Menyederhanakan perbandingan** dua atau lebih kelompok data.

### Contoh:

- Membandingkan IPK rata-rata rombel 1, 2, dan 3
- Membandingkan tinggi rata-rata pemain basket tim A dan B

- ❑ memungkinkan kita untuk melakukan proses statistik berikutnya **seperti:**

- Melihat hubungan (korelasi),
- Melihat perbedaan (t-test) antar kelompok,
- dan lain sebagainya.



## (4.2) Pengukuran Tendensi Sentral

Mean  
(rata-rata)

Median  
(nilai tengah)

Mode  
(modus)

## (4.2.1) Mean (Rata-Rata) – Data Tunggal

**Mean** adalah **ukuran pusat data** berupa **rata-rata** yang diperoleh dengan menghitung jumlah semua nilai data dibagi banyaknya data.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

**Keterangan**      $\bar{X}$  = nilai mean  
                       $X_i$  = data ke- $i$  dari variable acak  
                       $n$  = total sampel data

### Contoh

Nilai mata kuliah Data Mining dari 5 mahasiswa.

Mahasiswa	Nilai
1	80
2	83
3	88
4	90
5	95

$$\begin{aligned}\bar{X} &= \frac{\sum X_i}{n} = \frac{80 + 83 + 88 + 90 + 95}{5} \\ \bar{X} &= \frac{436}{5} = 87.2\end{aligned}$$

## (4.2.2) Mean (Rata-Rata) - Data Berkelompok

Jika nilai  $x_1, x_2, \dots, x_n$  masing-masing memiliki **frekuensi**  $f_i$ , maka **nilai mean**

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n}$$

### Keterangan

$\bar{X}$  = nilai mean

$f_i$  = frekuensi data ke- $i$  dari variable acak

$X_i$  = data ke- $i$  dari variable acak

$f$  = total frekuensi data

### Contoh

Data tinggi badan siswa kelas 5 SD .

Tinggi Badan cm	Titik Tengah ( $x_i$ ) cm	Frekuensi ( $f_i$ )	$f_i x_i$
156 - 160	158	6	948
161 - 165	163	10	1630
166 - 170	168	8	1334
171 - 172	173	4	692

$$\bar{X} = \frac{\sum f_i X_i}{f} = \frac{948 + 1630 + 1334 + 692}{6 + 10 + 8 + 4}$$

$$\bar{X} = \frac{4616}{28} = 164.78$$

## (4.2.3) Median (Nilai Tengah) – Data Tunggal [Ganjil]

**Median** adalah nilai tengah dari data yang ada setelah data tersebut diurutkan.

$$M_e = X_{\frac{n+1}{2}}$$

**Keterangan:**

$M_e$  = nilai tengah atau median

$X_{\frac{n+1}{2}}$  = indeks data  $X$  ke  $\frac{n+1}{2}$

$n$  = total sampel data

**Contoh**

Tentukan nilai median dari data berikut:  
70, 65, 50, 40, 35, 45, 70, 80, 90

**Jawaban:**

**Data terurut** = 35, 40, 45, 50, 65, 70, 70, 80, 90

$n = 9 \rightarrow \text{ganjil}$

**Mencari indeks nilai tengah:**

$$M_e = X_{\frac{n+1}{2}} = X_{\frac{9+1}{2}} = X_{\frac{10}{2}} = X_5$$

**Maka:** Nilai **tengah** pada indeks ke-5 ( $X_5$ ) = 65

## (4.2.4) Median (Nilai Tengah) – Data Tunggal [Genap]

$$M_e = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2}$$

**Keterangan:**

$M_e$  = nilai tengah atau median

$X_{\frac{n}{2}}$  = indeks data  $X$  ke  $\frac{n}{2}$

$X_{\frac{n+2}{2}}$  = indeks data  $X$  ke  $\frac{n+2}{2}$

$n$  = total sampel data

**Contoh**

Tentukan nilai median dari data berikut:

3, 2, 5, 2, 4, 6, 6, 7, 9, 6

**Jawaban:**

**Data terurut** = 2, 2, 3, 4, 5, 6, 6, 6, 7, 9

$n = 10 \rightarrow$  *genap*

**Mencari indeks nilai tengah:**

$$M_e = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} = \frac{X_{\frac{10}{2}} + X_{\frac{10+2}{2}}}{2} = \frac{X_5 + X_6}{2}$$

indeks ke-5 ( $X_5$ ) = 5

indeks ke-6 ( $X_6$ ) = 6

**Maka:** Nilai *tengah*  $M_e = \frac{X_5 + X_6}{2} = \frac{5 + 6}{2} = 5.5$

## (4.2.5) Median (Nilai Tengah) – Data Berkelompok

$$M_e = B + \left( \frac{\frac{n}{2} - f_{kum_B}}{f_i} \right) I$$

### Keterangan:

$M_e$  = nilai tengah atau median

$B$  = batas bawah median =  $x_{min} - 0.5$

$f_{kum_B}$  = frekuensi kumulatif data di bawah kelas median

$f_i$  = frekuensi data pada kelas median

$I$  = panjang interval kelas median

$n$  = total sampel data

### Contoh

Hasil pengukuran berat badan sebanyak 26 orang mahasiswa di sebuah universitas ditunjukkan pada tabel di bawah.

Berat Badan (kg)	Frekuensi ( $f_i$ )
46 – 50	3
51 – 55	2
56 – 60	4
61 – 65	5
66 – 70	6
71 – 75	4
76 – 80	1
81 - 85	1

**Hitung median berat badan mahasiwa!**

# (lanjutan) Median (Nilai Tengah) – Data Berkelompok

**Langkah 1:** menentukan  **$n$  data** berdasarkan frekuensi kumulatif

Berat Badan (kg)	Frekuensi ( $f_i$ )	Frekuensi Kumulatif ( $f_{kum}$ )
46 – 50	3	3
51 – 55	2	5
56 – 60	4	9
61 – 65	5	14
66 – 70	6	20
71 – 75	4	24
76 – 80	1	25
81 - 85	1	26

**Langkah 2:** menentukan **batas bawah median ( $B$ )**

$$n = 26 \rightarrow \text{genap}$$

**Mencari indeks nilai tengah:**

$$M_e = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} = \frac{X_{\frac{26}{2}} + X_{\frac{26+2}{2}}}{2} = \frac{X_{13} + X_{14}}{2}$$

- Data ke-**13** dan **14** terletak pada interval 4, yaitu **(61 - 65)**
- $x_{min} = 61$  dan  $x_{max} = 65$
- Kelas interval ke-4 adalah kelas median.

**Maka:** Nilai batas bawah median

$$B = x_{min} - 0.5 = 61 - 0.5$$

$$B = 60.5$$

## (lanjutan) Median (Nilai Tengah) – Data Berkelompok

**Langkah 3:** menentukan frekuensi kumulatif data di bawah kelas median ( $f_{kum_B}$ )

Berdasarkan tabel frekuensi kumulatif, maka

- **Kelas median** terletak pada **interval ke-4**.
- frekuensi kumulatif **sebelum interval ke-4** adalah **9**

**Maka:**  $f_{kum_B} = 9$

**Langkah 4:** menentukan frekuensi kelas median ( $f_i$ )

Berdasarkan tabel frekuensi kumulatif, maka

- **Kelas median** terletak pada **interval ke-4**.
- frekuensi kelas median adalah **5**

**Maka:**  $f_i = 5$

**Langkah 5:** mengidentifikasi panjang kelas interval ( $I$ )

Berdasarkan tabel frekuensi kumulatif, diketahui bahwa panjang interval tiap kelas adalah 5,

- Contoh: pada interval ke-1 nilainya **46 – 50**, maka panjang kelas interval 5

**Maka:**  $I = 5$

**Langkah 6:** menghitung nilai median berkelompok

$$M_e = B + \left( \frac{\frac{n}{2} - f_{kum_B}}{f_i} \right) I = 60.5 + \left( \frac{\frac{26}{2} - 9}{5} \right) 5$$

$$M_e = 60.5 + 4$$

$$M_e = 64.5 \text{ kg}$$



## (4.2.6) Mode (Modus) – Data Tunggal

❑ **Modus** adalah nilai yang paling sering muncul dalam suatu data.

❑ Jenis-jenis modus

- Sejumlah data bisa tidak mempunyai modus.
- Mempunyai 1 modus (unimodal)
- Mempunyai 2 modus (bimodal)
- Mempunyai banyak modus (multimodal)

### Contoh

Tentukan nilai modus dari data tunggal berikut:

- a) 1, 4, 7, 8, 9, 9, 11
- b) 1, 4, 7, 8, 9, 11, 13
- c) 1, 2, 4, 4, 7, 9, 11, 11, 13
- d) 1, 1, 3, 3, 7, 7, 12, 12, 14, 15

### Jawaban:

- a) Modus = 9 (muncul 2 kali)
- b) Modus = tidak ada (semua data muncul 1 kali)
- c) Modus = 4 dan 11 (masing-masing muncul 2 kali)
- d) Modus = 1, 3, 7, dan 12 (masing-masing muncul 2 kali)

## (4.2.7) Mode (Modus) – Data Berkelompok

**Modus** data berkelompok merupakan jenis modus yang ditentukan dari nilai tengah interval yang memiliki nilai terbanyak

$$M_0 = L + \left( \frac{d_1}{d_1 + d_2} \right) I$$

**Keterangan:**

$M_0$  = nilai modus

$L$  = batas bawah kelas modus =  $x_{min} - 0.5$

$d_1$  = Selisih frekuensi kelas modus dengan frekuensi kelas sebelumnya

$d_2$  = Selisih frekuensi kelas modus dengan frekuensi kelas sesudahnya

$I$  = panjang interval kelas

**Contoh**

Tentukan nilai modus dari data pada tabel di bawah

Nilai	Frekuensi
41-45	10
46-50	14
51-55	35
56-60	21
61-65	12
66-70	8
<b>Jumlah</b>	<b>100</b>

## (lanjutan) Mode (Modus) – Data Berkelompok

**Langkah 1:** menentukan kelas modus, yaitu kelas dengan frekuensi terbesar.

Nilai	Frekuensi
41-45	10
46-50	14
51-55	35
56-60	21
61-65	12
66-70	8
Jumlah	100

Frekuensi terbesar

**Maka:**  
kelas modus berada pada interval **51-53** atau **kelas 3**

**Langkah 2:** menentukan tepi bawah kelas ( $L$ )

Berdasarkan kelas modus (kelas 3), maka nilai terkecil  $x_{min} = 51$ .

**Maka:** nilai tepi bawah kelas

$$L = x_{min} - 0.5 = 51 - 0.5 \quad L = 50.5$$

**Langkah 3:** menentukan nilai ( $d_1$ )

Selisih **frekuensi kelas modus** dengan **frekuensi kelas sebelumnya**

$$\text{Maka: } d_1 = 35 - 14 \quad d_1 = 21$$

# (lanjutan) Mode (Modus) – Data Berkelompok

**Langkah 4:** menentukan nilai ( $d_2$ ).

Selisih **frekuensi kelas modus** dengan **frekuensi kelas sesudahnya**.

**Maka:**  $d_2 = 35 - 21$        $d_1 = 14$

**Langkah 5:** mengidentifikasi intrval kelas modus( $I$ ).

**Kelas modus** berada pada **kelas 3** dengan interval **51-55**

**Maka:**  $I = 5$

**Langkah 6:** meghitung modus data berkelompok

$$M_0 = L + \left( \frac{d_1}{d_1 + d_2} \right) I$$

$$M_0 = 50.5 + \left( \frac{21}{21 + 14} \right) 5I$$

$$M_0 = 50.5 + (0.6)5$$

$$M_0 = 50.5 + 3$$

**Maka:**  $M_0 = 53.5$

# Variabilitas

- **Variabilitas** adalah derajat penyebaran nilai-nilai variabel dari tendensi sentralnya dalam suatu distribusi yang menunjukkan seberapa banyak nilai-nilai variabel itu berbeda dari tendensi sentralnya, atau seberapa jauh nilai-nilai variabel itu menyimpang dari tendensi sentralnya (terutama Mean atau rerata).
- **Pengukuran variabilitas** akan memberikan gambaran variasi, jangkauan, serta heterogenitas-homogenitas dari pengukuran suatu kelompok (data).

## (5.1) Kegunaan Pengukuran Variabilitas

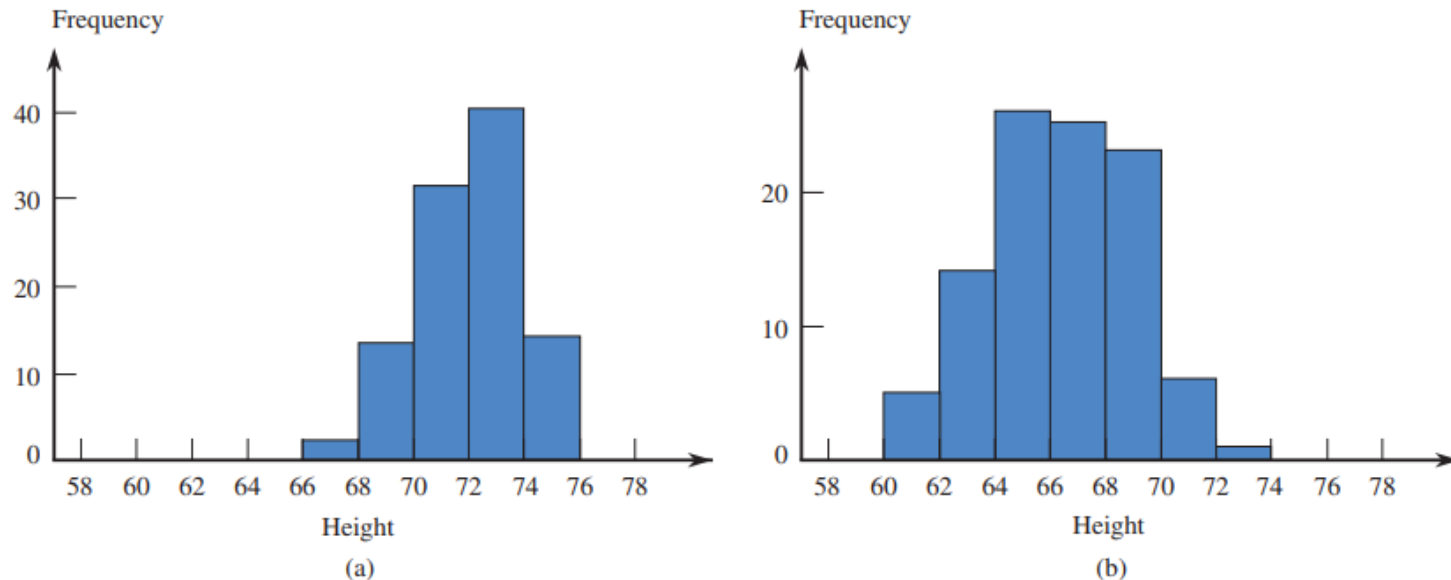
Tendensi sentral hanya memberikan informasi tentang suatu nilai yang menjadi pusat dari nilai-nilai lainnya, tetapi tidak memberikan informasi “seberapa jauh atau seberapa besar nilai-nilai dalam kelompok itu bervariasi”.

A :	25	25	25	25	25	25	25	25	
B :	21	23	23	24	25	26	26	27	30
C :	6	15	15	21	25	27	30	41	45

- Ketiga kelompok data memiliki Mean sama, tetapi karakteristik datanya berbeda.
- Kelompok data A sangat homogen
- kelompok data B lebih homogen dibanding data C.

Dua distribusi yang sama ukuran tendensi sentralnya belum dapat dipergunakan secara meyakinkan bahwa kedua distribusi tersebut sama.

## (5.2) Kasus 1: if the shoe fit

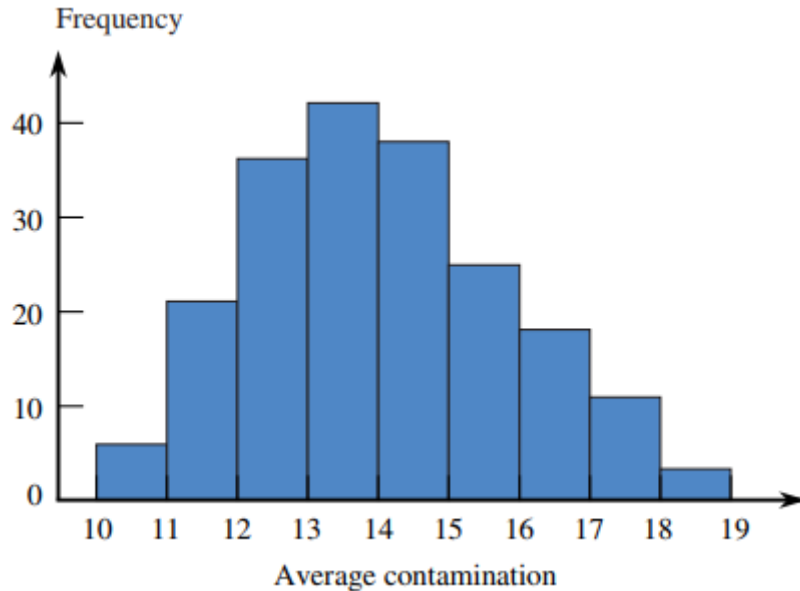


**Figure 1.1** Histograms of heights (in inches) of female athletes: (a) basketball players; (b) gymnasts.

- 1) A tall woman (5 ft 11 in.) tells you she is looking for her sister who is practicing with her team at the gym. Would you direct her to where the basketball team is practicing or to where the gymnastics team is practicing? What reasoning would you use to decide?
- 2) If you found a pair of size 6 shoes left in the locker room, would you first try to return them by checking with members of the basketball team or the gymnastics team?



## (5.2) Kasus 2: Monitoring Water Quality



**Figure 1.2** Frequency of contaminant concentration (in parts per million) in well water.

A chemical spill has occurred at a manufacturing plant 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

One month after the spill, five water specimens are collected from the well, and the average contamination is 15.5 ppm. Considering the variation before the spill, would you take this as convincing evidence that the well water was affected by the spill? What if the calculated average was 17.4 ppm? 22.0 ppm? How is your reasoning related to the graph in Figure 1.2?

## (5.3) Pengukuran Variabilitas Data

Range  
(jangkauan)

Simpangan Baku  
(Standar  
Deviasi)

Pencilan  
(Outlier)

## (5.3.1) Range

**Range** atau **Jangkauan Total (JT)** atau **Rentangan (R)** adalah jarak dari data dengan nilai terendah sampai nilai tertinggi.

$$R = \text{Max} - \text{Min}$$

**Keterangan**

$R$  = Range

$\text{Max}$  = batas nilai tertinggi

$\text{Min}$  = batasa nilai terendah

**Contoh** Tentukan jarak nilai tertinggi dan terendah.

A : 25 25 25 25 25 25 25 25 25

B : 21 23 23 24 25 26 26 27 30

C : 6 15 15 21 25 27 30 41 45

**Jawaban:**

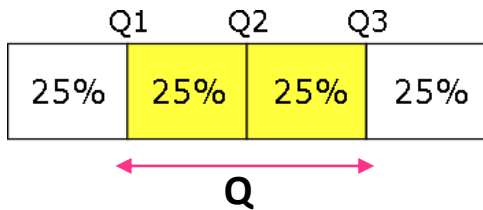
R data A =  $25 - 25 = 0$

R data B =  $30 - 21 = 9$

R data C =  $45 - 6 = 39$

## (5.3.2) Range Semi Kuartil

**Range Semi interkuartil (Q)** adalah distribusi data yang ditunjukkan dipotongnya di kedua ujungnya masing-masing 25%, yang terdapat di antara 3 titik Q1, Q2, dan Q3.



$$Q = \frac{Q_3 - Q_1}{2}$$

### Keterangan

$Q$  = Range semi kuartil

$Q_1$  = Kuartil 1 ( $P_{25}$ )

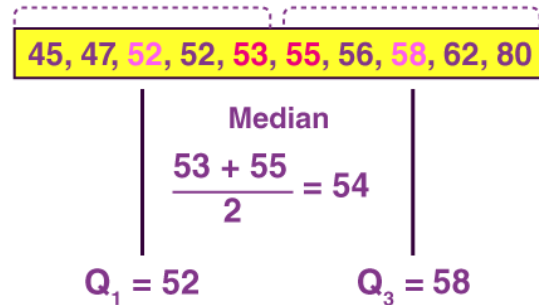
$Q_2$  = Kuartil 2 atau median ( $P_{50}$ )

$Q_3$  = Kuartil 3 ( $P_{75}$ )

**Range antar kuartil (interquartile range)** dapat diketahui dengan menggunakan rumus:

$$Q_3 - Q_1$$

**Contoh** Tentukan Range Semi Kuartil.



**Range Semi Kuartil**

$$Q = \frac{58 - 52}{2} = 3$$

**Interquartile Range**

$$Q_3 - Q_1 = 58 - 52 = 6$$

## (5.3.3) Simpangan Baku (Standard Deviation)

- ❑ **Simpangan baku (*standard deviation*)** adalah salah satu ukuran sebaran data.
- ❑ Dipakai untuk data bertipe interval dan rasio.
- ❑ Untuk kumpulan bilangan  $S = \{x_1, \dots, x_N\}$  dengan **rerata aritmetik**  $\mu_S$ , **simpangan baku**  $\sigma_S$  dari  $S$  adalah:

$$\sigma_S = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_S)^2} = \sqrt{\frac{(x_1 - \mu_S)^2 + \dots + (x_N - \mu_S)^2}{N-1}}$$

- ❑ Kuadrat dari  $\sigma_S$ , yakni  $\sigma_S^2$  disebut sebagai **varian**
- ❑ **Nilai simpangan baku**
  - **Besar** = data secara umum tersebar jauh dari nilai rerata aritmetik
  - **Kecil** = data secara umum terkumpul dekat dengan nilai rerata aritmetik
- ❑ **Simpangan baku** dapat pula dipandang sebagai **derajat ketidakpastian pengukuran data**
  - Contoh: pada pengukuran berulang dengan suatu instrument yang sama, jika simpangan baku data hasil pengukuran bernilai besar, berarti presisi pengukuran rendah.

## (lanjutan) Simpangan Baku

Hitunglah simpangan baku dari data sampel berikut: 5,5,3,4,7,8,9,1,9

Rata-rata

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{5+5+3+4+7+8+9+1+9}{9} = \frac{51}{9} = 5.67$$

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

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(5-5,67)^2 + (5-5,67)^2 + \dots + (5-5,67)^2}{9-1}} =$$

$$\sqrt{\frac{62,0001}{8}} = 2,78$$

## (5.3.4) Simpangan Baku Data Populasi

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

### Keterangan

$\sigma$  (*sigma*) = Simpangan baku populasi

$x_i$  = data ke-i

$\mu$  = Nilai rata-rata populasi

$n$  = jumlah data populasi

### Contoh

Perusahaan produksi kayu jati mengekspor kayu ke Jepang. Datanya sebagai berikut: 234, 321, 231, 332, dan 242 ton. Tentukan nilai simpangan baku!

Berat (Ton)	$\mu$	$x - \mu$	$(x - \mu)^2$
234	272	-38	1.444
321	272	49	2.401
231	272	-41	1.681
332	272	60	3.600
242	272	-30	900
1359	-	-	9.126

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum (x - \mu)^2}{N}} \\ \sigma &= \sqrt{\frac{9.126}{5}} \\ &= \sqrt{1.825,2} \\ &= 42,72 \text{ ton}\end{aligned}$$



## (5.3.5) Menentukan pencilan (secara kasar)

- *3-sigma rule*: Jika data kira-kira terdistribusi normal:
  - $x_i$  adalah pencilan jika  $x_i < \mu_S - 2\sigma_S$  atau  $x_i > \mu_S + 2\sigma_S$   
→ peluang bahwa data berjarak ke rerata lebih jauh dari 2 kali simpangan baku adalah 4.55%.
  - $x_i$  adalah pencilan jika  $x_i < \mu_S - 3\sigma_S$  atau  $x_i > \mu_S + 3\sigma_S$   
→ peluang bahwa data berjarak ke rerata lebih jauh dari 3 kali simpangan baku adalah 0.27%.
  - Kekurangan: (i) asumsi distribusi normal (belum tentu!), (ii) rerata dan simpangan baku dipengaruhi nilai pencilan itu sendiri, dan (iii) tidak dapat mendeteksi pencilan jika jumlah data sedikit (*small sample size*).
- *Tukey's fences*: memakai **rentang antarkuartil** (*interquartile range*)  $IQR = Q_3 - Q_1$ .
  - $x_i$  adalah pencilan jika  $x_i < Q_1 - 1.5(IQR)$  atau  $x_i > Q_3 + 1.5(IQR)$ .
  - $x_i$  adalah pencilan ekstrim jika  $x_i < Q_1 - 3(IQR)$  atau  $x_i > Q_3 + 3(IQR)$ .
- Metode-metode lain (mungkin lebih baik): Visualisasi, Grubb's test, Dixon's Q test, Algoritma Expectation Maximization, Jarak k-Nearest Neighbor, *local outlier factor* berbasis *density* (variasi *density-based clustering*), dll.