

# Statistical Machine Learning's Data Storytelling

spacing = 3 tab

## 1. Part 1 Essential Mathematical Foundations (Deep Dive)

### Chapter 1. Descriptive Statistics

#### 1-1. Introduction To Descriptive Statistics



#### Apa Itu Statistik Deskriptif

01  
**Definisi**

- Cabang statistik yang fokus pada penyajian, pengorganisasian, dan peringkasan data

fisikamodern00-2625220

02  
**Objektif**

- Digunakan untuk memahami karakteristik dasar dari suatu kumpulan data.

#### Apa Itu Statistik Deskriptif



**Tidak digunakan untuk**  
Membuat kesimpulan atau generalisasi terhadap populasi yang lebih luas

# Peran Statistik Deskriptif



Langkah awal dalam eksplorasi data



Membantu mengidentifikasi pola, tren, dan anomali

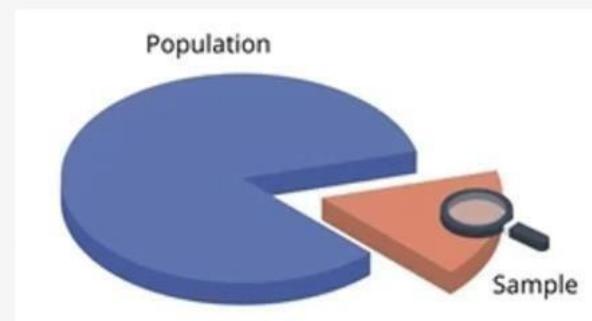
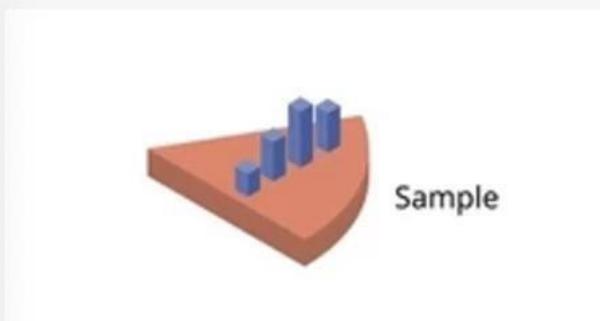


Menyediakan ringkasan numerik dan visual dari data



Dasar untuk analisis statistic lanjutan dan machine learning

## Statistik Deskriptif vs Inferensial



### Deskriptif:

- Fokus pada data yang tersedia
- Tidak membuat prediksi atau generalisasi

### Inferensial:

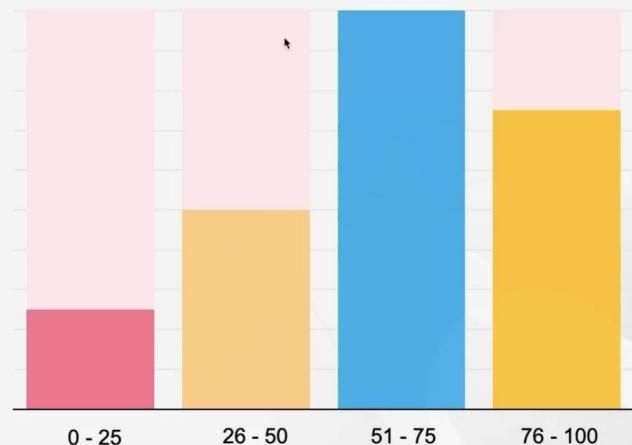
- Menggunakan sampel untuk menyimpulkan tentang populasi.
- Melibatkan estimasi dan pengujian hipotesis

# Contoh Aplikasi Statistik Deskriptif



No	Nilai
1	75
2	73
3	50
4	90
5	75
6	45
7	60
8	80
9	76
10	15

Mean	63,9
Median	74
Standard Deviasi	22,0426355
Minimum	15
Maximum	90
First Quartile	52,5



## 1-2. Measures Of Central Tendency

01

**Ukuran Pemusatan**

02

**Mean, Median, Modus**  
hsikamodem00-2625220

03

**Interpretasi Ukuran Pemusatan**

04

**Contoh Aplikasi**

05

**Penutup**

## Ukuran Pemusatan

- Ukuran pemusatan menggambarkan nilai tengah atau tipikal dari suatu kumpulan data.
- Digunakan untuk meringkas data numerik menjadi satu nilai representatif.
- Tiga ukuran utama: Mean (rata-rata), Median, dan Mode (modus).

## Mean (Rata-rata)

$$x \rightarrow 5 \ 7 \ 6 \ 2 \ 1 \ 4 \quad \text{data} \quad n=6$$

$\underbrace{5+7+6+2+1+4}_{\text{jumlah data}} = 25$

$$\begin{aligned}\bar{x} &= \frac{5+7+6+2+1+4}{6} \\ &= \frac{25}{6} \rightarrow \frac{65}{7} = 9,14... \\ &= 4,16666 \approx 4,17\end{aligned}$$

**Definisi:** Jumlah seluruh nilai dibagi jumlah data.

**Formula:**

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

**Catatan:**

Kelebihan: Menggunakan semua data, mudah dihitung.

Kekurangan: Sangat sensitif terhadap outlier.

## Mean (Rata-rata)

$$x \rightarrow 5 \ 7 \ 6 \ 2 \ 1 \ 4 \quad \text{data} \quad n=6$$

$\underbrace{5+7+6+2+1+4}_{\text{jumlah data}} = 25$

$$\begin{aligned}\bar{x} &= \frac{5+7+6+2+1+4}{6} \\ &= \frac{25}{6} \rightarrow \frac{65}{7} = 9,14... \\ &= 4,16666 \approx 4,17\end{aligned}$$

**Definisi:** Jumlah seluruh nilai dibagi jumlah data.

**Formula:**

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

**Catatan:**

Kelebihan: Menggunakan semua data, mudah dihitung.

Kekurangan: Sangat sensitif terhadap outlier.

# Modus (Mode)

$$X \rightarrow 1 \underline{2} \underline{2} \underline{2} 3 4$$

↓  
modus = 2

$$X \rightarrow 1 \underline{2} \underline{2} \underline{2} 3 \underline{4} \underline{4} \underline{4} 5$$

1 2 2 3 4 4 5 6 6 6

3 3 3      Bimodal.  
multimodal

**Definisi:** Nilai yang sering muncul.

**Formula:**

- Frekuensi / kemunculan data tertinggi
- Dapat memiliki lebih dari satu modus (bimodal/multimodal)

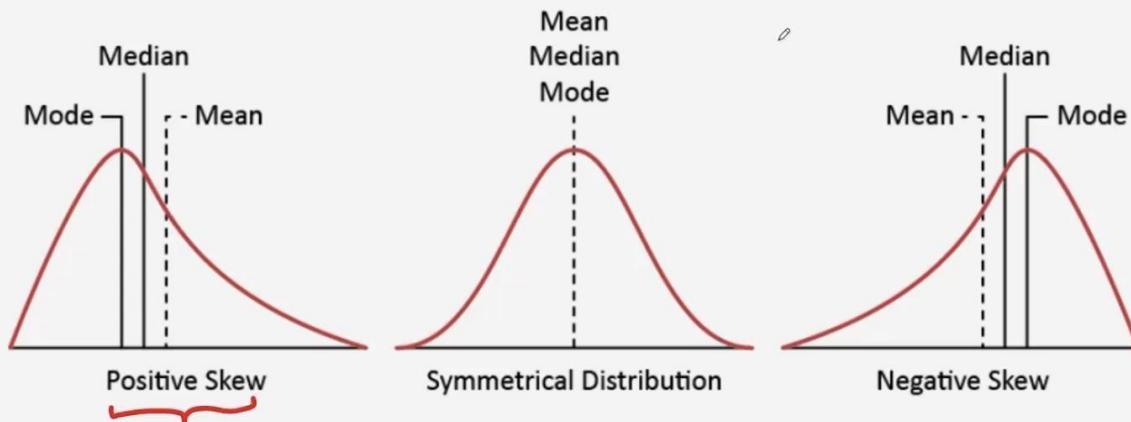
**Catatan:**

Kelebihan: Lebih cocok untuk data kategorikal.

Kekurangan: Tidak selalu ada atau tidak unik.

fisikamodern00-262!

## Perbandingan Mode, Median & Mean



# Contoh Aplikasi Mean, Median & Modus



Data gaji karyawan: gunakan median untuk menghindari bias dari gaji ekstrem.



Nilai ujian siswa: mean untuk melihat performa umum.

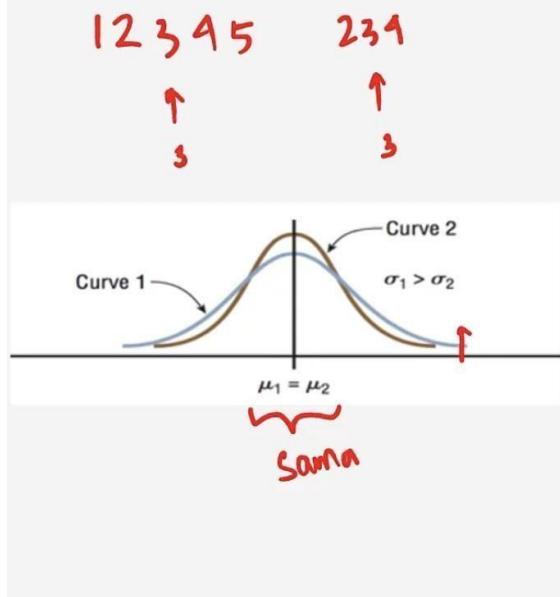


Preferensi produk: mode untuk mengetahui pilihan terbanyak.

## Penutup

- **Mean:** rata-rata aritmatika, sensitif terhadap outlier.
- **Median:** nilai tengah, robust terhadap outlier.
- **Mode:** nilai paling sering muncul, cocok untuk data kategorikal.
- Pilih ukuran pemusatan sesuai konteks data.

# Apa itu ukuran persebaran data



- **Definisi:** Ukuran penyebaran menggambarkan seberapa tersebar data dalam suatu distribusi.
- **Tujuan:**
  - Digunakan untuk memahami variabilitas dan konsistensi data.
  - Deteksi outlier
  - Normalisasi data
  - Evaluasi model prediktif
- **Contoh:** dua dataset dengan rata-rata sama bisa memiliki penyebaran yang berbeda.

## 1-3. Measures Of Dispersion

01	02	03	04	05
<b>Apa itu ukuran persebaran data?</b>	<b>Range (Jangkauan)</b>	<b>Variance (variansi)</b>	<b>Standard Deviation (Simpangan Baku)</b>	<b>Interquartile Range (IQR)</b>

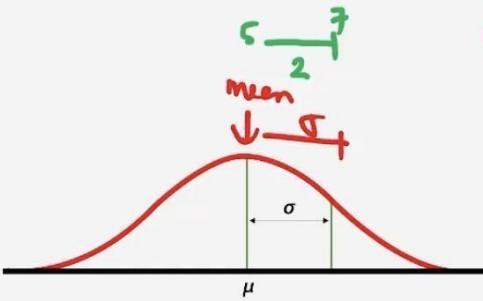
## Range (Jangkauan)

No	Nilai
1	75
2	73
3	50
4	90
5	75
6	45
7	60
8	80
9	76
10	15

- **Definisi:** selisih antara nilai maksimum dan minimum dalam dataset.
- **Formula:**
$$\text{range} = \text{max} - \text{min}$$
- **Kelebihan:** Mudah dihitung
- **Kekurangan:** Sangat dipengaruhi oleh outlier

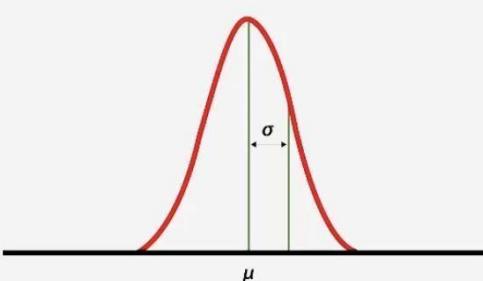
$$\begin{aligned} \text{range} &= \text{max} - \text{min} \\ &= 90 - 15 \\ &= 75 \end{aligned}$$

# Standard Deviation (Simpangan Baku)



- **Definisi:** akar kuadrat dari variance
- **Formula:**

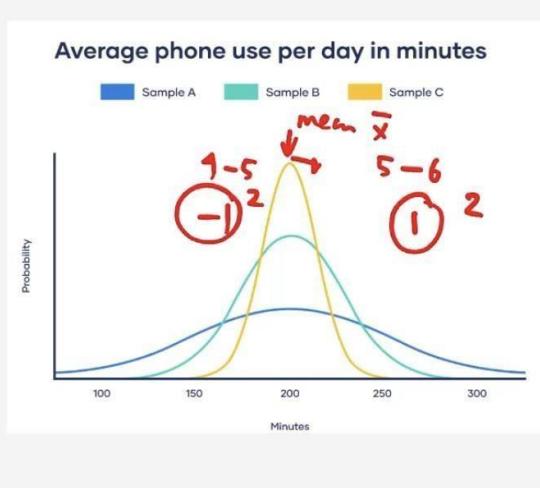
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$



- **Kelebihan:** Satuan sama dengan data asli.
- **Kekurangan:** Masih dipengaruhi oleh outlier



## Variance (Variansi / Ragam)



- **Definisi:** mengukur rata-rata kuadrat deviasi dari mean
- **Formula:**

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

- **Kelebihan:** Menggunakan semua data
- **Kekurangan:** Satuan kuadrat, sulit diinterpretasi langsung

## Interquartile Range (IQR)

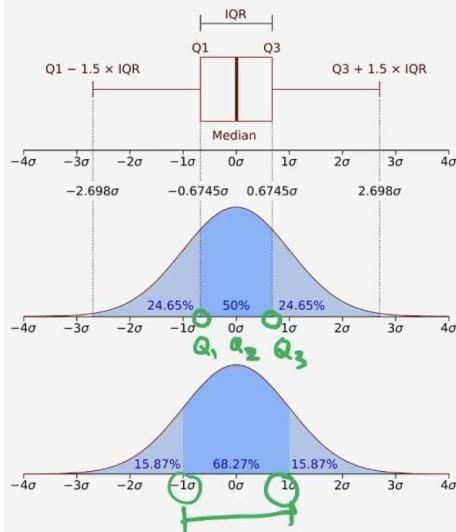


- **Definisi:** selisih antara kuartil ketiga ( $Q_3$ ) dan kuartil pertama ( $Q_1$ ).
- **Formula:**
$$IQR = Q_3 - Q_1$$
- **Kelebihan:** Robust terhadap outlier.
- **Kekurangan:** Tidak menggunakan semua data

# Perbandingan

Range	Mudah Sangat sensitive terhadap outlier
Variance	Menggunakan semua data Satuan kuadrat
Standard Deviation	Menggunakan semua data Interpretasi lebih mudah
Interquartile Range	Robust (Tegar) Terhadap Outlier

## Interquartile Range (IQR)



- **Definisi:** selisih antara kuartil ketiga ( $Q_3$ ) dan kuartil pertama ( $Q_1$ ).

- **Formula:**

$$IQR = Q_3 - Q_1$$

- **Kelebihan:** Robust terhadap outlier.

- **Kekurangan:** Tidak menggunakan semua data



# Bentuk Distribusi Data

- Bentuk distribusi data menggambarkan bagaimana nilai-nilai dalam dataset tersebar.
- Distribusi dapat menunjukkan pola
  - simetris,
  - miring (skewed),
  - seragam,
  - atau memiliki lebih dari satu puncak (bimodal).

fisikamodern00-2625220

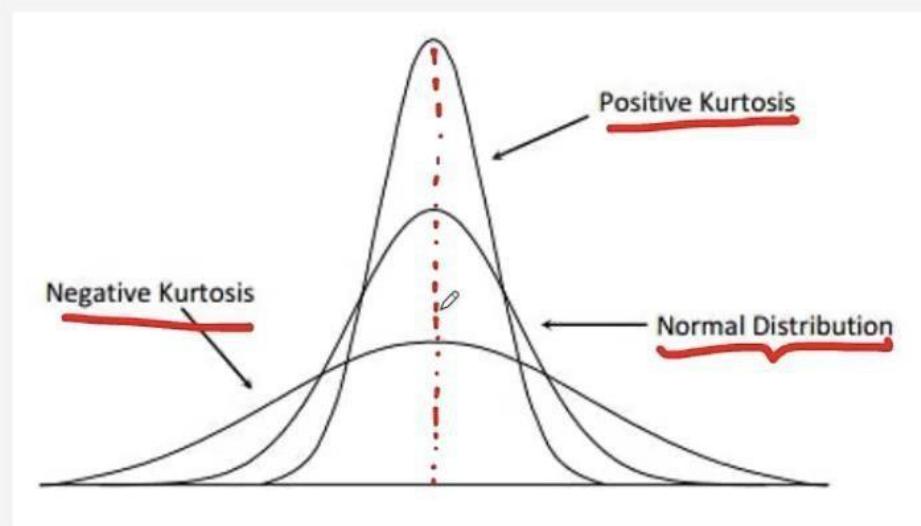
## Aplikasi dalam Data Science dan Statistics

<b>01</b> <b>Menilai Variabilitas Fitur</b>	Ukuran pemasaran seperti mean hanya memberi tahu kita nilai rata-rata, tetapi tidak memberi tahu apakah data homogen atau heterogen. Measures of dispersion seperti standard deviation dan IQR menunjukkan seberapa jauh data menyebar dari pusatnya.
<b>02</b> <b>Mendeteksi Outlier</b>	Ukuran seperti range dan IQR membantu mengidentifikasi nilai ekstrem yang bisa mempengaruhi analisis.
<b>03</b> <b>Menentukan Stabilitas dan Risiko</b>	Dalam konteks bisnis atau keuangan, standard deviation sering digunakan untuk mengukur risiko atau volatilitas.
<b>04</b> <b>Membantu Pemilihan Metode Statistik</b>	Distribusi data yang skewed atau memiliki penyebaran tinggi mungkin tidak cocok untuk metode yang mengasumsikan distribusi normal.
<b>05</b> <b>Meningkatkan Interpretasi Visualisasi</b>	Visualisasi seperti <b>boxplot</b> dan <b>histogram</b> menjadi lebih bermakna jika kita memahami ukuran penyebaran
<b>06</b> <b>Dasar untuk Normalisasi dan Standarisasi</b>	Dalam machine learning, data sering perlu dinormalisasi agar model bekerja optimal. Measures of dispersion digunakan untuk <b>Z-score normalization</b> : menggunakan mean dan standard deviation <b>Min-max scaling</b> : menggunakan range

## 1-4. Understanding Data Distribution Shapes

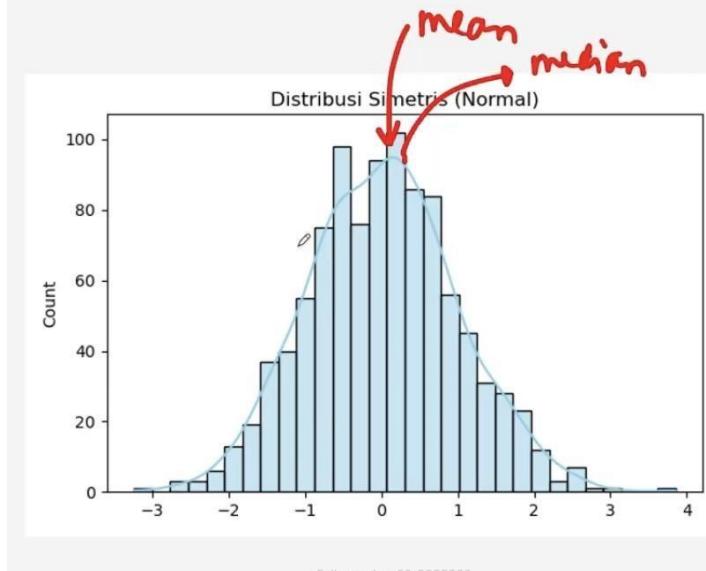
<b>01</b> <b>Definisi</b>	<b>02</b> <b>Jenis-jenis distribusi data</b>	<b>03</b> <b>Skewness &amp; Kurtosis</b>	<b>04</b> <b>Perbandingan Penutup</b>
------------------------------	---	---	--

# Kurtosis



Kurtosis mengukur derajat konsentrasi data di sekitar mean dan frekuensi outlier. Secara teknis, kurtosis adalah momen keempat dari distribusi yang dinormalisasi.

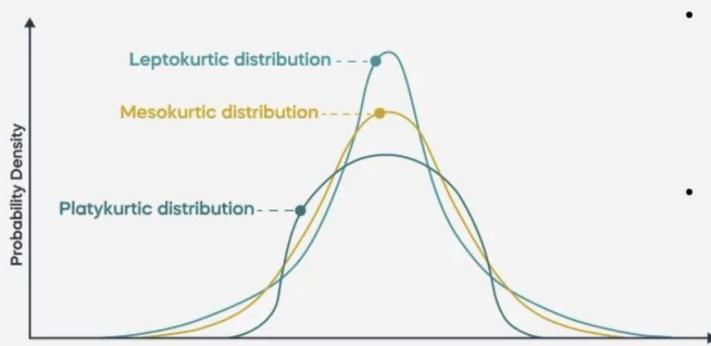
## Distribusi Simetris



- Distribusi yang tersebar dengan titik median dan mean saling berdekatan.
- Contoh dari distribusi ini adalah **distribusi normal**

# Kurtosis

fisik

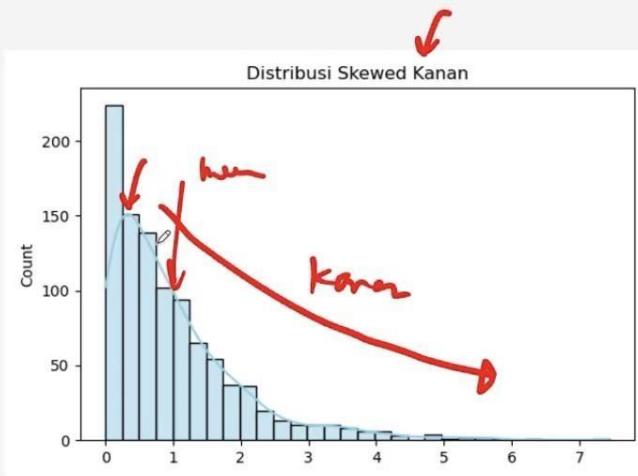


Menghitung kurtosis dengan python

```
from scipy.stats import kurtosis  
kurtosis(data, fisher=False)
```

- Mesokurtic (Kurtosis  $\approx 3$ )
  - Distribusi normal standar.
  - Puncak distribusi **sedang**, tidak terlalu tajam atau datar.
  - Outlier muncul dengan frekuensi normal.
- Leptokurtic (Kurtosis  $> 3$ )
  - Puncak distribusi **tajam dan tinggi**.
  - Data lebih terkonsentrasi di sekitar mean.
  - Outlier lebih sering muncul.
  - Contoh: distribusi keuangan dengan lonjakan ekstrem.
- Platykurtic (Kurtosis  $< 3$ )
  - Puncak distribusi **datar dan lebar**.
  - Data lebih tersebar.
  - Outlier lebih jarang.
  - Contoh: distribusi nilai ujian yang merata.

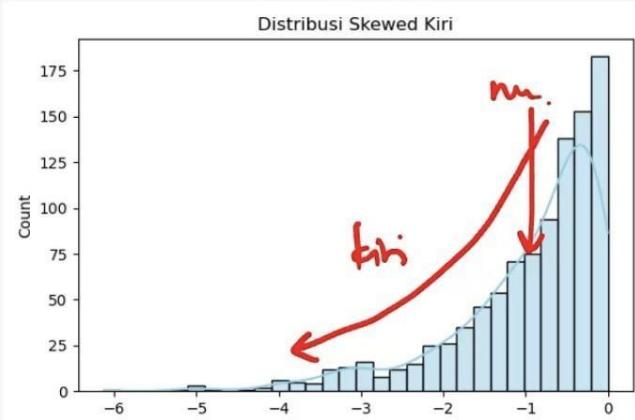
## Distribusi Miring (Skewed)



Skewness  $> 0$ : miring ke kanan

$$skewness = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)\sigma^3}$$

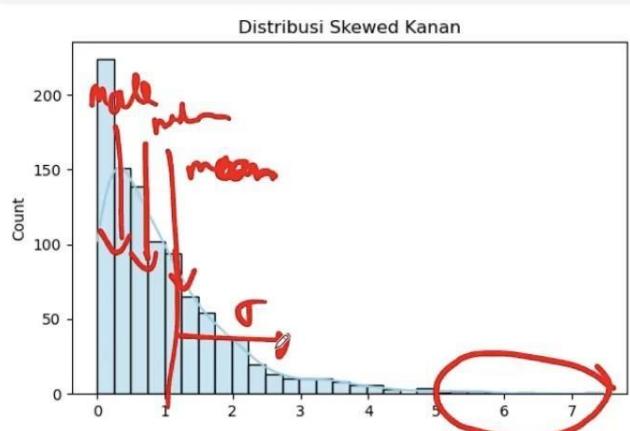
$\bar{x}$  = nilai mean,  
 $x_i$  = Nilai data ke- $i$ ,  
 $\sigma$  = Standard Deviation,  
 $n$  = jumlah data



Skewness  $< 0$ : miring ke kiri

fisikamodern00-2625220

# Distribusi Miring (Skewed)

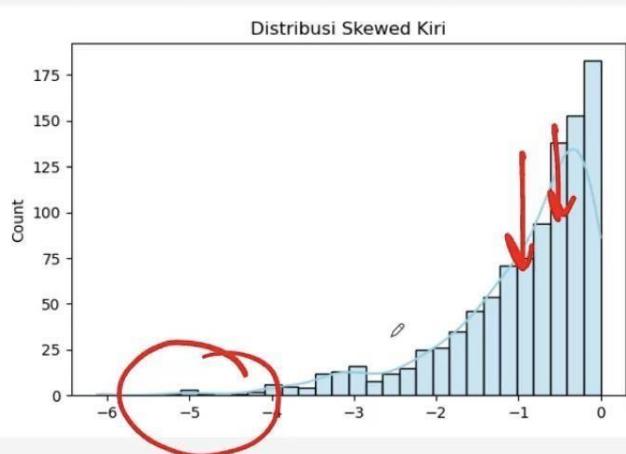


Skewness > 0: miring ke kanan

## Karakteristik:

- Ekor Panjang di kanan
- Mean > Median > Mode
- Contoh:
  - Pendapatan Individu: selegintir orang berpenghasilan sangat tinggi
- Warning:
  - Nilai Mean dapat menyesatkan karena dipengaruhi oleh outliers
  - Median lebih representatif untuk pusat data
  - Standard Deviasi yang sangat besar karena terdapat nilai ekstrem di kanan

# Distribusi Miring (Skewed)



Skewness < 0: miring ke kiri

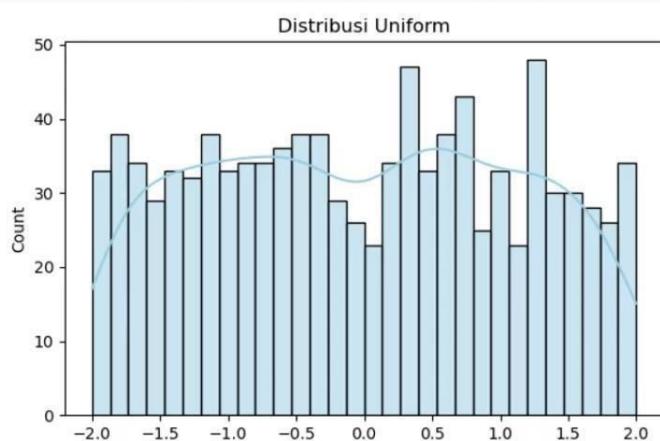
## Karakteristik:

- Ekor Panjang di kiri
- Mean < Median < Mode
- Contoh:
  - Skor ujian dengan nilai tinggi yang banyak dan sedikit yang bernilai rendah
- Warning:
  - Nilai Mean dapat menyesatkan karena dipengaruhi oleh outliers ekstrem minimum
  - Median lebih representatif untuk pusat data
  - Standard Deviasi yang sangat besar karena terdapat nilai ekstrem di kiri

# Penutup

- Bentuk distribusi mempengaruhi analisis data
- Skewness dan kurtosis membantu memahami karakteristik data
- Visualisasi sangatlah penting untuk eksplorasi awal sebaran data dan deteksi outliers.

## Distribusi Uniform



Skewness < 0: miring ke kiri

### Karakteristik:

- Semua nilai memiliki probabilitas yang mirip
- Mean  $\approx$  Median
- Contoh:
  - Hasil lemparan dadu fair  $\rightarrow$  setiap angka memiliki probabilitas 1/6
  - Digunakan untuk menggenerasi nilai random pada nilai awal bobot machine learning.
- Warning:
  - Semakin besar rentang data, semakin besar variansnya.
  - $Var(x) = \frac{(b-a)^2}{12}$ , dengan  $a$  dan  $b$  adalah rentang distribusi

## 1-5. Summary And Practice Problems

### Ringkasan Konsep Utama

- Mean: Rata-rata dari sekumpulan data
- Median: Nilai tengah dari data yang diurutkan
- Mode: Nilai yang paling sering muncul
- Range: Selisih antara nilai maksimum dan minimum
- Variance: Ukuran penyebaran data dari rata-rata
- Standard Deviation: Akar dari variance, menunjukkan sebaran data



# Latihan 1 – Ukuran Pemusatan

Diberikan data: [4, 8, 6, 5, 3, 4, 1, 7]

Hitung mean, median, modus, dan range

$$\text{mean} \rightarrow \bar{x} = \frac{4+8+6+5+3+9+9+7}{8} = \frac{41}{8} = 5,125$$

modus = 4

$$\text{range} = 8 - 3 = 5$$

median  $\rightarrow$  3 4 4 4 5 6 7 8

$$\frac{4+5}{2} = 4,5$$


## Latihan 2 – Ukuran Penyebaran

Diberikan data: [10, 12, 23, 23, 16, 23, 21, 16]

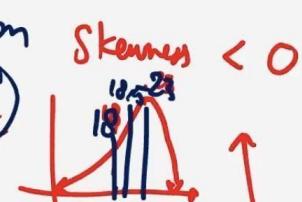
Hitung variance, standard deviation dan skewness

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{(10-18)^2 + (12-18)^2 + \dots + (21-18)^2 + (16-18)^2}{8} = 24\sqrt{8}$$
$$\bar{x} = \frac{149}{8} = 18$$

modus = 23

mean

Skewness < 0



## Tips

- Gunakan mean untuk data simetris tanpa outlier
- Gunakan median jika data memiliki outlier atau skewed
- Mode berguna untuk data kategorikal
- Standard deviation menunjukkan sebaran data dari rata-rata
- Visualisasi seperti histogram membantu memahami distribusi
- Perhatikan bentuk distribusi sebelum memilih ukuran statistik

## Chapter 2. Probability Theory

### 2-1. What Is Probability

01

**Definisi  
Probability**

02

**Sample Space  
Dan  
Event**

03

**Classical  
vs  
Empirical**

## Apa Itu Probabilitas?

**01**

**Definisi**

- Probabilitas adalah ukuran kemungkinan suatu kejadian terjadi

**02**

**Nilai**

- Bernilai 0 saat tidak terjadi.
- Bernilai 1 saat terjadi.

**03**

**Penggunaan**

- Digunakan dalam pengambilan keputusan berbasis data

# Contoh



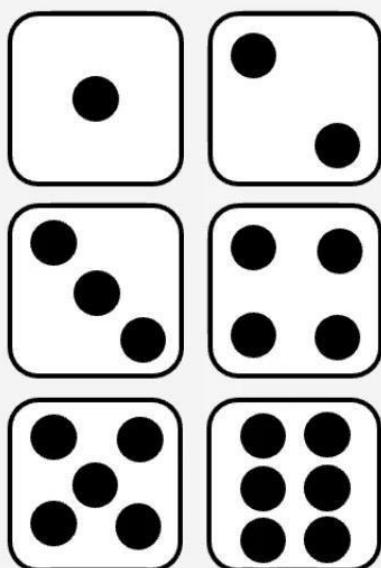
Berapa probabilitas / kemungkinan gambar muncul saat sekali tos?

$$p(\text{gambar}) = \frac{1}{2}$$

$$\underline{p(\text{ongka}) = \frac{1}{2}} +$$

(1)

# Contoh



Berapa probabilitas / kemungkinan dadu berangka 2 muncul saat dilempar?

$$[1, 2, 3, 4, 5, 6]$$

$\downarrow \downarrow \downarrow \downarrow \downarrow \downarrow$

0 1 0 0 0 0

$$P(2) = \frac{1}{6}$$

$$P(1) = \frac{1}{6}$$

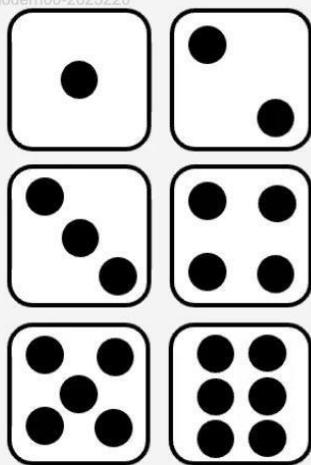
⋮

$$P(6) = \frac{1}{6}$$

# Event / Kejadian

- Event atau kejadian adalah subset dari sample space

fisikamodern00-2625220



Sample Space nya adalah:

[ 1, 2, 3, 4, 5, 6 ]

Event → Angka Ganjil:

[ 1, 3, 5 ]

Event → Angka Genap:

[ 2, 4, 6 ]

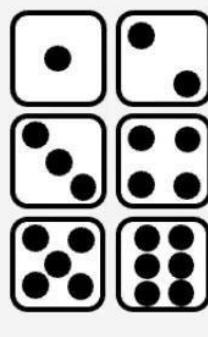
Event → Bilangan Prima:

[ 2, 3, 5 ]

# Sample Space / Ruang Sampel

fisikamodern00-2625220

- Sample Space adalah himpunan semua kemungkinan hasil



Sample Space nya adalah:

[ 1, 2, 3, 4, 5, 6 ]

6

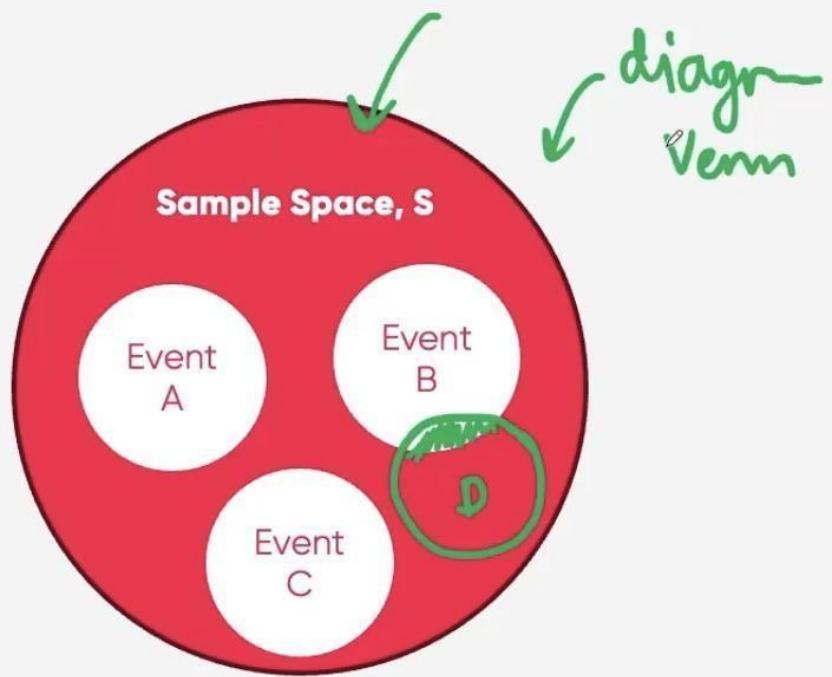


Sample Space nya adalah:

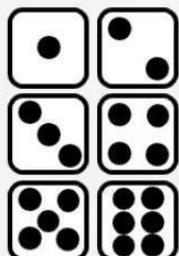
[ angka, gambar ]

2

# Sample Space dan Event



## Classical VS Empirical



Jumlah sampel adalah 6

Kemungkinan angka 3 muncul adalah?

### Classical / Theoretical:

$$P(3) = \frac{\text{kemungkinan } 3}{\text{total kemungkinan}} \\ = \frac{1}{6}$$

### Empirical / Experiment:

$$\begin{array}{cccccc} 1 & 3 & 1 & 4 & 5 & 6 \\ P(1) = \frac{2}{6} = \frac{1}{3} \end{array}$$

## Ringkasan

- **Sample Space:** semua kemungkinan outcome dari eksperimen
- **Eksperimen:** proses yang menghasilkan hasil acak
- **Outcome:** hasil dari satu percobaan

## Kenapa Berbeda?

Probabilitas Teoritical akan sama dengan Probabilitas Empirical saat

Jumlah eksperimen  
sangatlah  
**BESAR!**



Jika nilai probabilitas teoritical tidak diketahui, jumlah observasi yang sangat banyak akan meningkatkan akurasi nilai probabilitas

### Kejadian saling lepas (mutually exclusive)

- Kejadian yang tidak mungkin terjadi bersama-sama:
  - Pelemparan satu dadu menghasilkan angka 2 dan 5
  - Pelemparan coin satu kali menghasilkan gambar dan angka

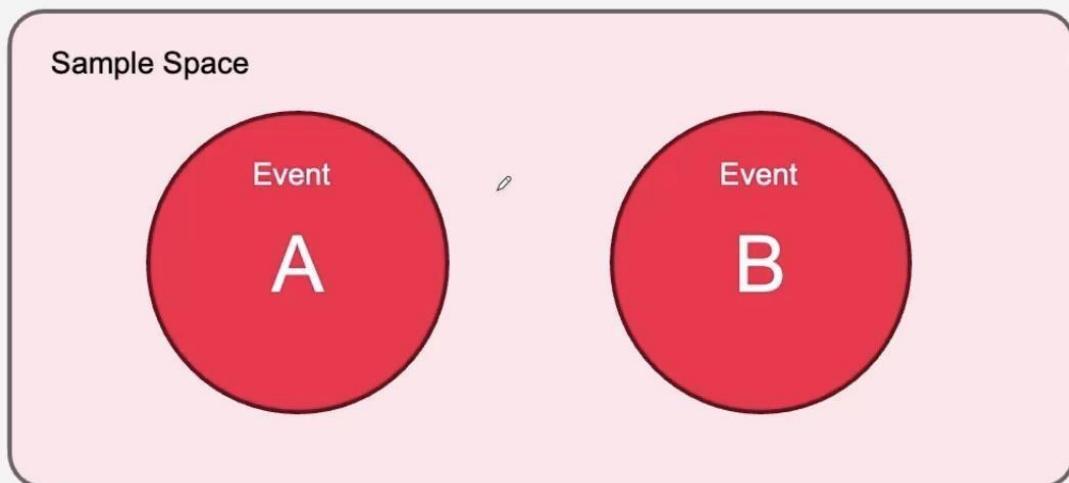


Diagram Venn

fisikamedern00-2625220

### Kejadian tidak saling lepas (Non mutually exclusive)

- Kejadian yang mungkin terjadi bersama-sama:
  - Siswa yang suka fisika dan matematika
  - Siswa yang memakai kacamata dan keriting

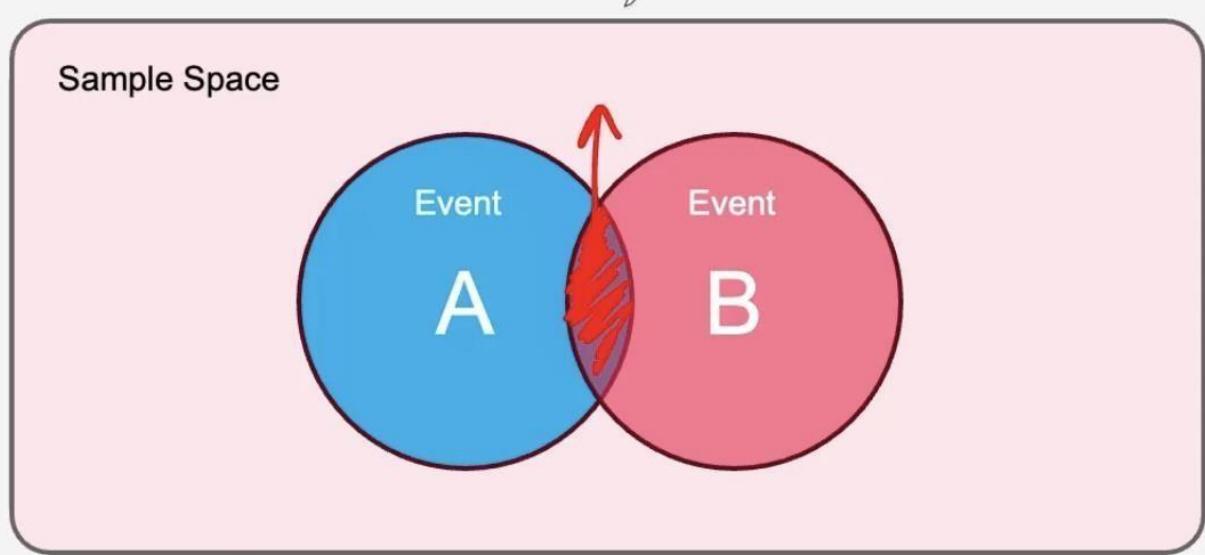
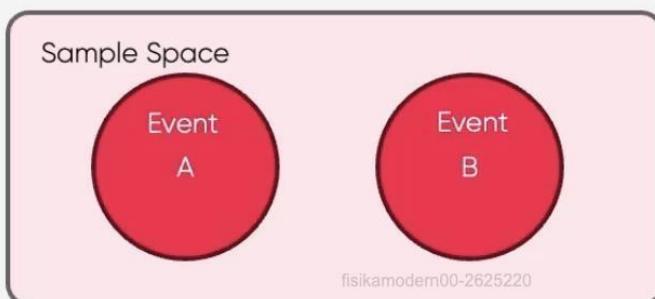


Diagram Venn

# Aturan Penjumlahan (Mutually Exclusive)

- Berapa probabilitas kejadian A atau B. contohnya dalam 1 lempar dadu, muncul angka 1 atau 6 *atau 3*



$$\begin{aligned}P(A \text{ atau } B) &= P(A) + P(B) \\P(1 \text{ atau } 6) &= P(1) + P(6) \\&= \frac{1}{6} + \frac{1}{6} \\&= \frac{2}{6} = \frac{1}{3}\end{aligned}$$

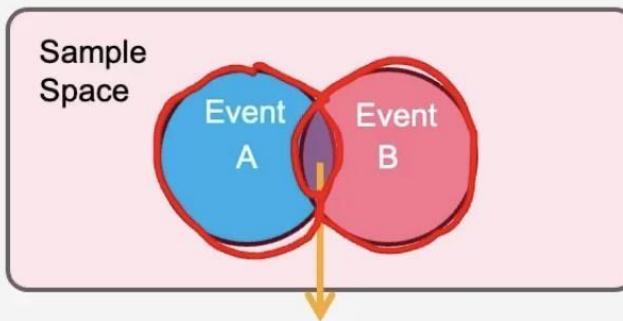
- Formulasi:

$$P(A \text{ atau } B) = P(A \text{ or } B) = P(A \cup B) = P(\bar{A}) + P(B)$$



# Aturan Penjumlahan (Mutually Exclusive)

- Berapa probabilitas kejadian A atau B. contohnya siswa berkacamata atau keriting



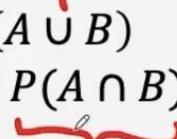
$$P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B)$$



*union .*

- Formulasi:

$$\begin{aligned}P(A \text{ atau } B) &= P(A \text{ or } B) = P(A \cup B) \\&= P(A) + P(B) - P(A \cap B)\end{aligned}$$



# Dependent vs Independent Events

## Dependent

Kejadian yang hasilnya mempengaruhi kejadian selanjutnya

Contoh:

- Pengambilan kartu pada dek kartu tanpa dikembalikan, mempengaruhi pengambilan selanjutnya

## Independent

Kejadian yang hasilnya tidak mempengaruhi kejadian selanjutnya

Contoh:

- Pelemparan coin pertama tidak mempengaruhi pelemparan kedua, dan seterusnya

## Aturan Perkalian (Mutually Exclusive)



- Akan selalu bernilai nol, karena kejadian A dan B terjadi bersamaan tidak ada

Sample Space

Event  
A

Event  
B

$$P(A \cap B) = 0$$
$$P(A) \times P(B) = 0$$

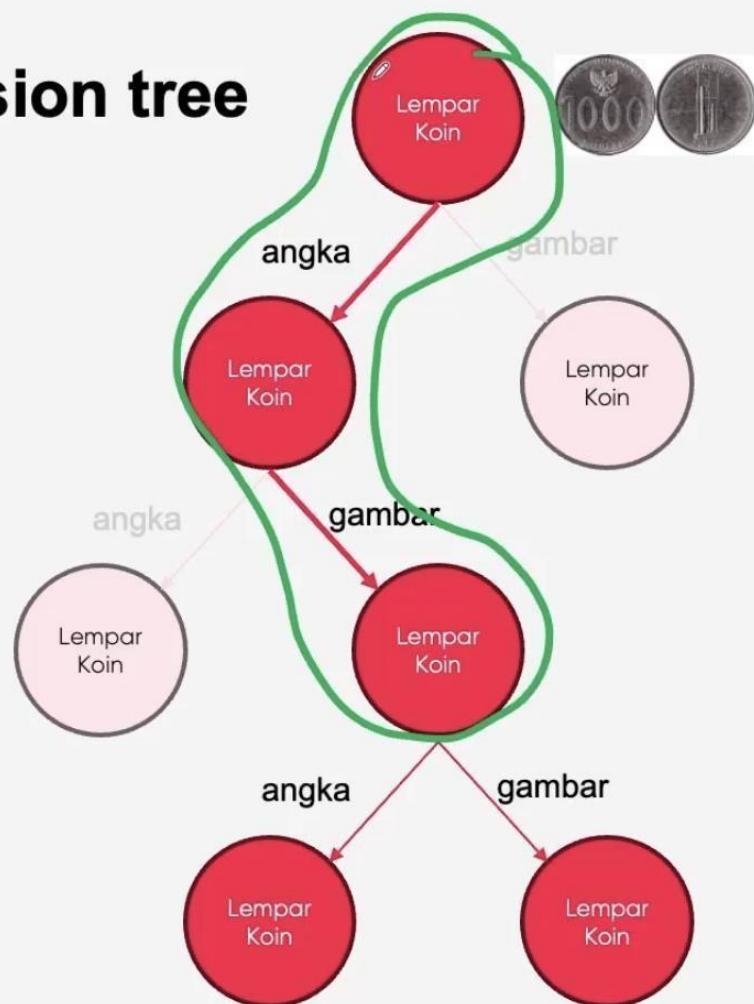
# Aturan Perkalian (Independent)

- Aturan perkalian digunakan untuk menemukan probabilitas bahwa dua atau lebih kejadian saling bebas terjadi secara berurutan atau bersamaan
- Contoh: Melempar koin dua kali
  - Kejadian A: Mendapatkan "Kepala" pada lemparan pertama.  $P(A) = 0.5$
  - Kejadian B: Mendapatkan "Ekor" pada lemparan kedua.  $P(B) = 0.5$
  - Probabilitas mendapatkan kepala lalu ekor:
    - $P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = 0.5 \times 0.5 = 0.25$



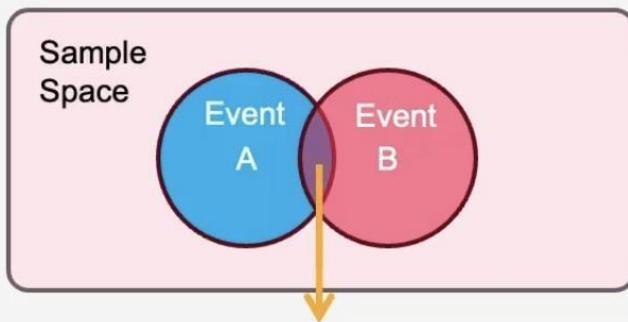
fisikamodem00-2625220

## Ilustrasi decision tree Independent



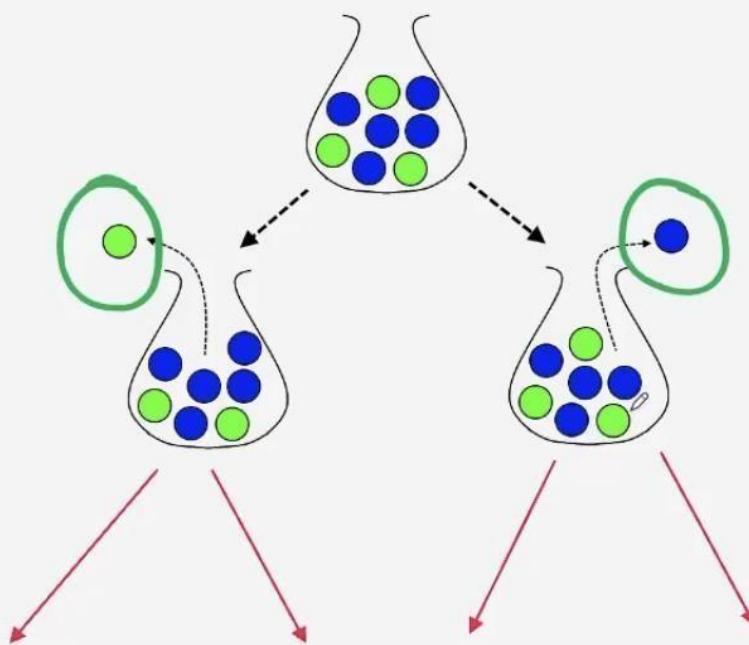
# Aturan Perkalian (Non Mutually Exclusive)

- Akan bernilai tidak nol, karena kejadian A dan B dapat terjadi dalam waktu bersamaan, misal mengambil kartu no 5 dan berwarna merah.



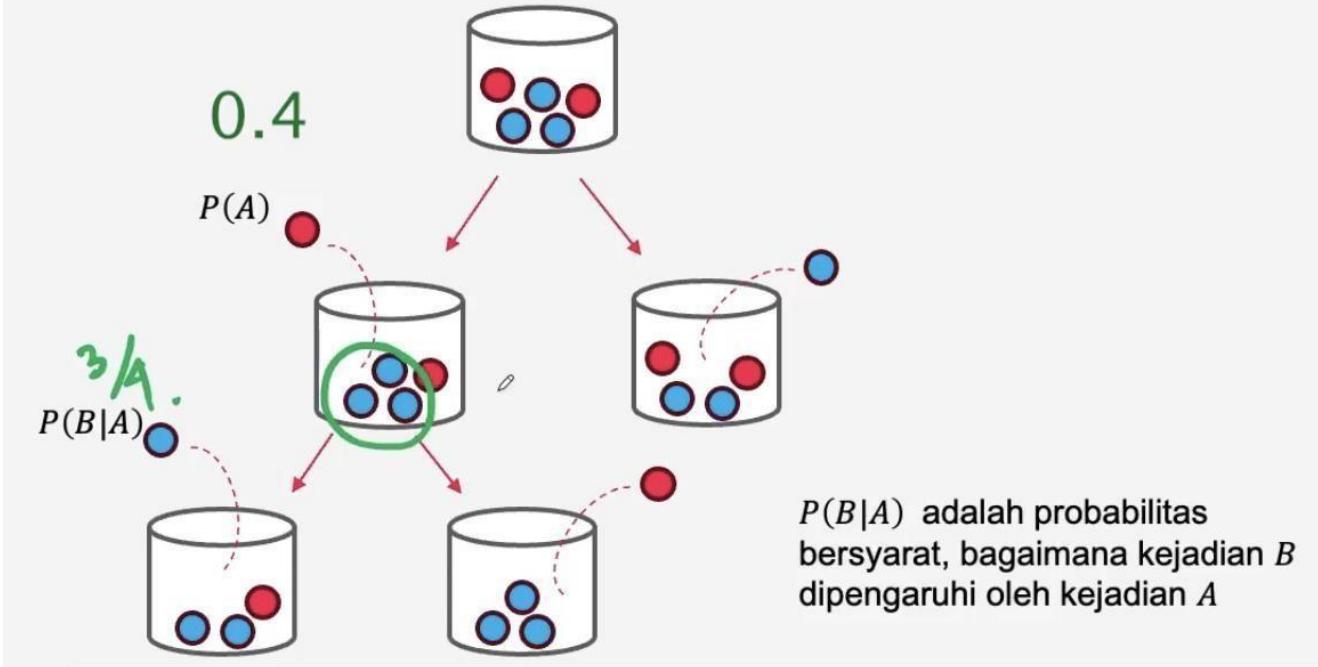
$$P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) \neq 0$$

## Ilustrasi decision tree - Dependent



# Probabilitas Bersyarat

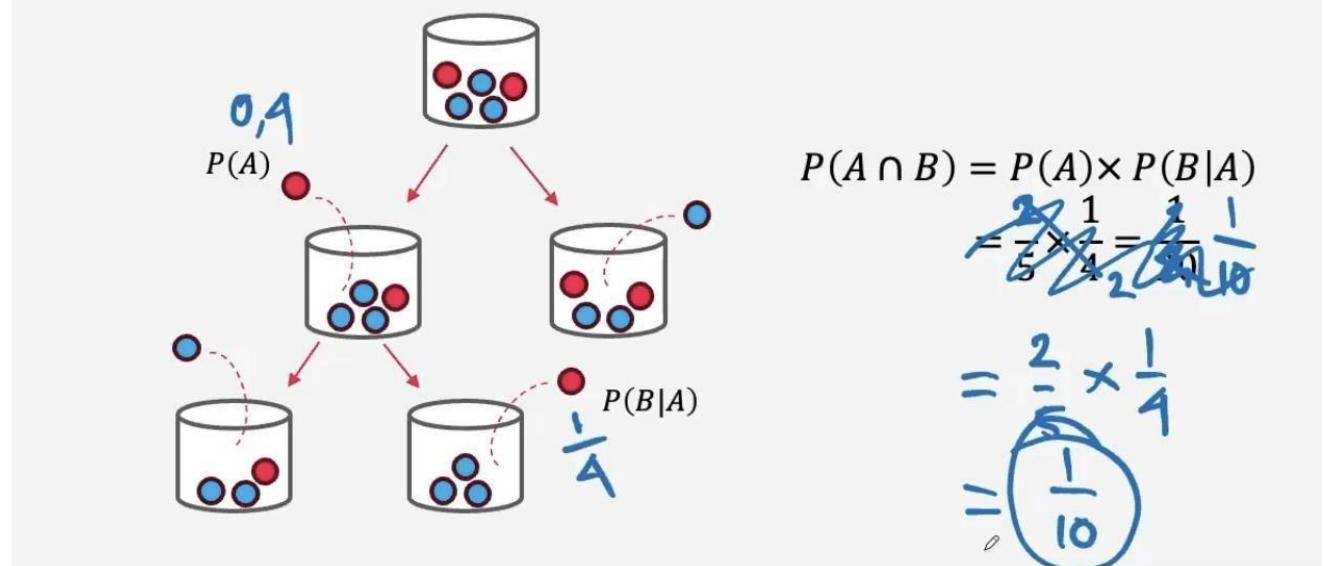
- Sebuah kantong berisi 2 kelereng merah dan 3 kelereng biru
- Kejadian A : Mengambil kelereng merah pada pengambilan pertama
  - $P(A) = \frac{2}{5} = 0.4$
- Kejadian B : Mengambil kelereng biru pada pengambilan kedua
  - $P(B|A) = \frac{\text{jumlah biru}}{\text{kelereng sisa}} = \frac{3}{4}$



$P(B|A)$  artinya peluang B dengan syarat A sudah diketahui (atau terjadi). Urutannya: A diketahui dulu, baru B dihitung.

## Aturan Perkalian - dependent

- Sebuah kantong berisi 2 kelereng merah dan 3 kelereng biru
- Kejadian A : Mengambil kelereng merah pada pengambilan pertama
  - $P(A) = \frac{2}{5} = 0.4$
- Kejadian B : Mengambil kelereng merah pada pengambilan kedua
  - $P(B|A) = \frac{\text{jumlah merah}}{\text{kelereng sisa}} = \frac{1}{4}$



01

## **Memahami konsep conditional probability**

02

## **Menghitung probabilitas bersyarat menggunakan rumus**

03

## **Mengidentifikasi kejadian independen**

04

## **Menggunakan tabel kontingensi dan diagram pohon untuk visualisasi**

# **Apa itu Conditional Probability?**

- Probabilitas suatu kejadian terjadi dengan asumsi bahwa kejadian lain telah terjadi.
- Notasi diberikan oleh:

$P(A|B)$ = Probabilitas A terjadi jika B telah terjadi

- Contoh: Probabilitas seseorang sakit flu jika ia demam

# Conditional Probability Formula

- Ingat bahwa probabilitas kejadian  $A$  terjadi dan kejadian  $B$  terjadi adalah :

$$= P(A) \times P(B|A)$$

$$P(A \cap B) = P(B) \times P(A|B)$$



- Kita bisa menggabarkan bahwa  $P(A|B)$  sebagai seberapa mungkin  $A$  terjadi di antara kejadian  $B$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Contoh: Probabilitas seseorang sakit flu jika ia demam

$$P(\text{flu|demam}) = \frac{P(\text{flu dan demam})}{P(\text{demam})}$$

## Contoh

- Dari 100 pasien, 40 demam, 30 sakit flu, dan 25 keduanya

$$\bullet P(\text{Demam}) = \frac{40}{100} = 0.4$$

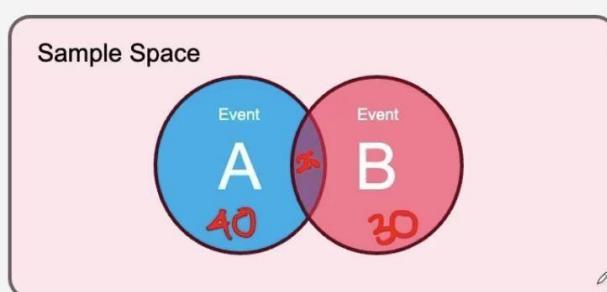
$$P(\text{flu}) = \frac{30}{100} = 0.3$$

$$\bullet P(\text{Sakit} \cap \text{Demam}) = \frac{25}{100} = 0.25$$

$$P(\text{sakit | demam}) = \frac{P(\text{sakit} \cap \text{demam})}{P(\text{demam})}$$

$$\bullet P(\text{Sakit} | \text{Demam}) = \frac{0.25}{0.4} = 0.625$$

$$= \frac{0.25}{0.4} = 0.625$$



$$P(\text{demam | sakit}) = \frac{P(\text{sakit} \cap \text{demam})}{P(\text{sakit})} = \frac{0.25}{0.3} = 0.833$$

# Contingency Table

- Tabel kontingensi membantu menghitung probabilitas bersyarat
- Dari 100 pasien, 40 demam, 30 sakit flu, dan 25 keduanya

$P(\text{Sakit} | \text{Demam}) = \frac{25}{60} = 0,1\ldots$

$P(\text{Sakit} | \text{Demam})$

$\frac{P(\text{Sakit} \cap \text{Demam})}{P(\text{Demam})}$

$= \frac{25}{40}$

$= 0,625$

	Sakit	Tidak Sakit	Total
Demam	25	15	40
Tidak Demam	5	55	60
Total	30	70	100

$P(\text{Demam} | \text{Sakit}) = \frac{25}{30} = 0,1\ldots$

fisikamodern00-2625220

## Mendeteksi Kejadian Independent

fisikamodern00-2625220

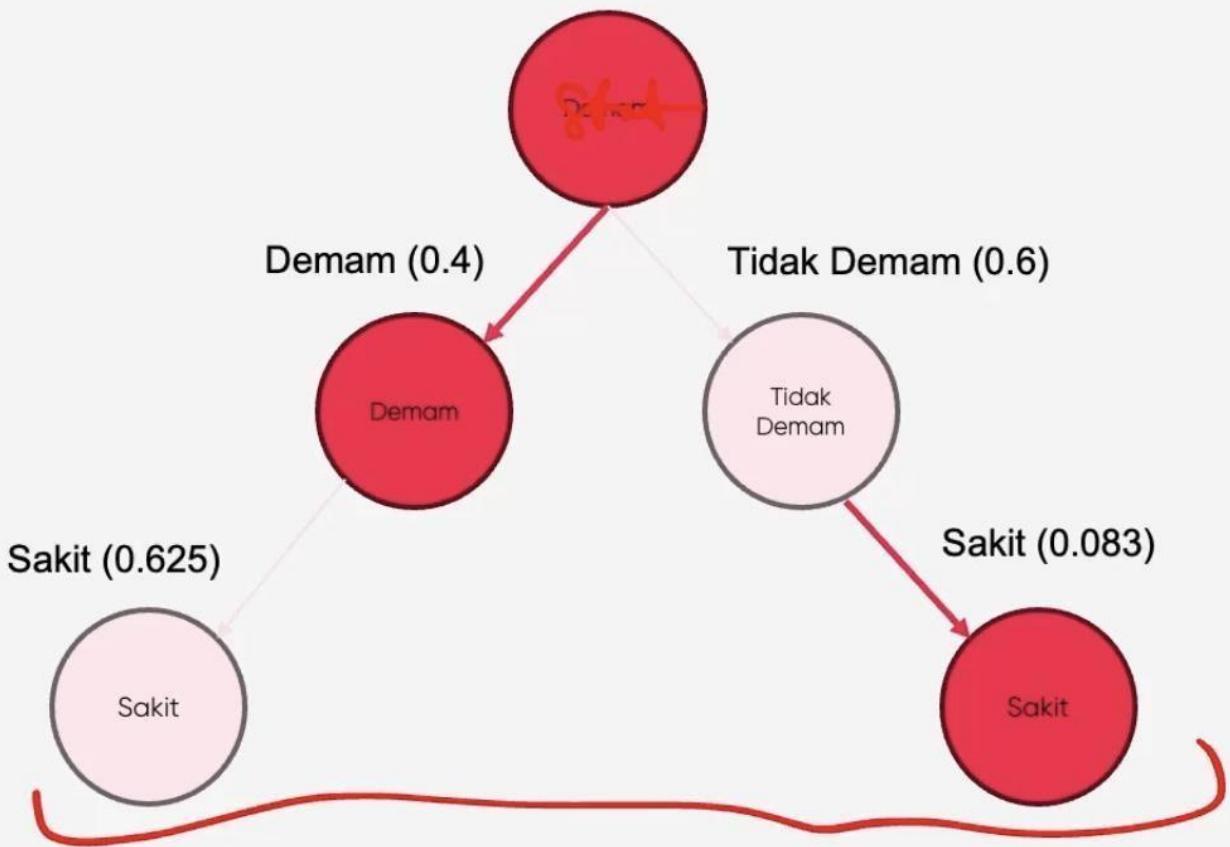
- Dua kejadian A dan B dinyatakan independent jika

$$P(A \cap B) = P(A) \times P(B)$$

- Atau secara ekuivalen

$$P(A|B) = P(A) \text{ dan } P(B|A) = P(B)$$

# Ilustrasi decision tree



## Ringkasan

- Conditional probability memperhitungkan informasi tambahan.
- Kejadian independen tidak saling mempengaruhi.
- Tabel kontingensi dan diagram pohon membantu visualisasi.
- Konsep ini penting dalam model prediktif dan inferensi statistik.

**01 Memahami konsep dasar teorema Bayes**

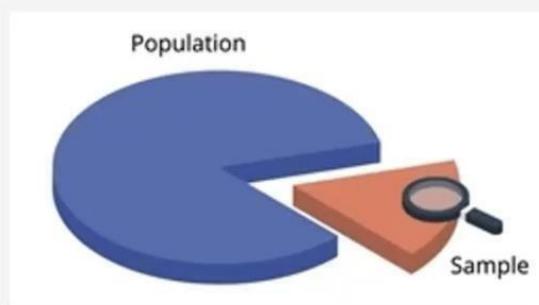
**02 Menginterpretasikan rumus Bayes dalam konteks nyata**

**03 Menerapkan Teorema Bayes pada kasus nyata**

**04 Menghubungkan Teorema Bayes dengan algoritma machine learning**

## Pengantar

- Bayes Theorem adalah alat penting dalam inferensi statistik

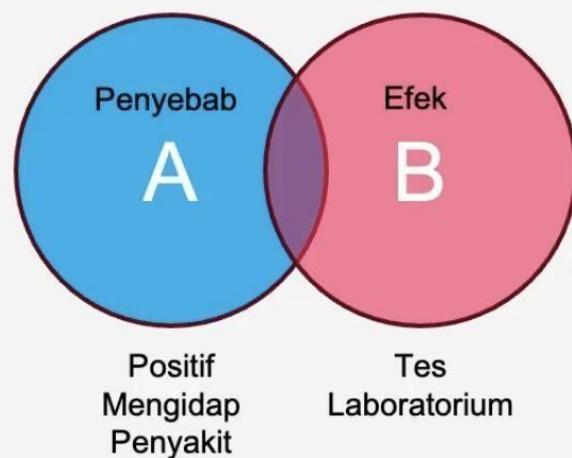


### Inferensial:

- Menggunakan sampel untuk menyimpulkan tentang populasi
- Melibatkan estimasi dan pengujian hipotesis
- Probabilitas diperbaharui setiap ada informasi baru

# Pengantar

- Secara sederhana. Teorema Bayes memberikan cara matematis untuk membalikkan probabilitas bersyarat.
- Memungkinkan untuk menemukan probabilitas suatu "penyebab" ( $A$ ) jika kita mengetahui "efek" ( $B$ ) telah terjadi.



$P(B|A)$  Probabilitas hasil tes positif jika memang sakit

$P(A|B)$  Probabilitas orang sakit jika tes adalah positif

## Teorema Bayes



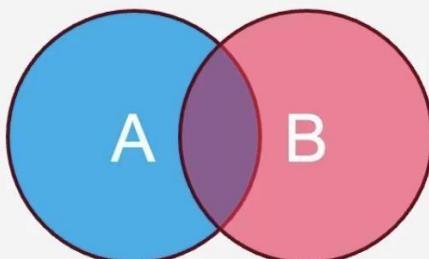
$P(B|A)$  Probabilitas hasil tes positif jika memang sakit

$$P(A \cap B) = P(A) \times P(B|A)$$

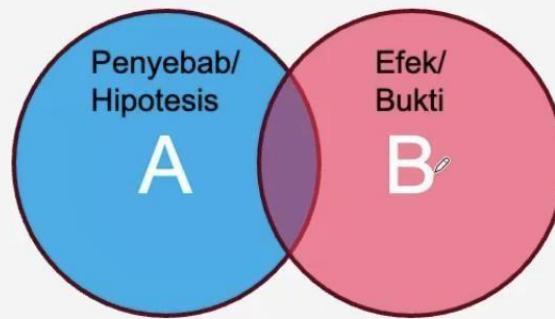
Probabilitas sakit dan tes positif =  
Perobabilitas tes positif  $\times$  Sakit jika hasil tes positif

$$P(A \cap B) = P(B) \times P(A|B)$$

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$



# Teorema Bayes



$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

Inor ↓      bukti. kew. wkt  
bukti. kew. wkt

bukting .

- $P(B|A)$  : **Likelihood** : Probabilitas bukti/terjadi B jika diberikan terjadi A
- $P(A)$  : **Prior** : Probabilitas terjadi A / probabilitas awal
- $P(B)$  : **Bukti/Evidence** : Probabilitas dari bukti B
- $P(A|B)$  : **Posterior** : Probabilitas Hipotesis A atau Penyebab A diberikan bukti/terjadi B

example: orang di test PCR misal hasilnya positif, berapa probabilitas dia untuk terjadinya sakit

misal ada kendaraan dan waktu tempuh

B = sebab = kendaraan

A = akibat = waktu tempuh

$P(A | B)$  = peluang tiba cepat jika naik kendaraan tertentu

$P(B | A)$  = peluang naik kendaraan tertentu jika tiba cepat

Essential Mathematical Foundation – Probability Theory – Bayes' Theorem Explained

**Contoh: Diagnosis Medis**

$P(p) = 1\% = 0,01$   
 $P(p_c) = 99\% = 0,99$   
 $P(t) = P(T|P) \cdot P(p) + P(T|P_c) \cdot P(p_c)$   
 $= 0,99 \cdot 0,01 + 0,05 \cdot 0,99$   
 $= 0,06 \cdot 0,99 = 0,0594$

$P(T|P) = 99\% \rightarrow P(T_c|P) = 1\%$   
 $P(T_c|P_c) = 95\% \rightarrow P(T|P_c) = 5\%$

- 1% populasi mengidap penyakit P
- Tes memiliki sensitivitas 99% dan spesifikitas 95%

$P(p|t) = \frac{P(t|p) \cdot P(p)}{P(t)}$   
 $= \frac{P(t|p) \cdot P(p)}{P(t|p) \cdot P(p) + P(t|p_c) \cdot P(p_c)}$   
 $= \frac{0,99 \cdot 0,01}{0,99 \cdot 0,01 + 0,05 \cdot 0,99}$   
 $= 0,16667 \rightarrow 16,67\%$

## Contoh Teorema Bayes

Diketahui Kotak :

1. Kotak A berisi 10 cokelat manis dan 5 cokelat pahit
2. Kotak B berisi 4 cokelat manis dan 16 cokelat pahit

Diketahui Kebiasaan Kamu :

1. 3 dari 5 kali. kamu memilih Kotak A.  $P(\text{manis} \mid A) = \frac{10}{15} = \frac{2}{3}$
2. 2 dari 5 kali. kamu memilih Kotak B.  $P(\text{manis} \mid B) = \frac{4}{20} = \frac{1}{5}$

Pertanyaan :

1. berapa peluang bahwa cokelat manis itu berasal dari Kotak A

---

Jawaban

1. Peluang Dasar (dari isi kotak) :

$$* P(\text{manis} \mid A) = \frac{10}{15} = \frac{2}{3}$$

$$* P(\text{manis} \mid B) = \frac{4}{20} = \frac{1}{5}$$

2. Hitung total probabilitas dapat cokelat manis  $P(\text{manis})$  :

$$* P(\text{manis}) = P(\text{manis} \mid A) \cdot P(A) + P(\text{manis} \mid B) \cdot P(B)$$

$$* P(\text{manis}) = \frac{2}{3} \cdot \frac{3}{5} + \frac{1}{5} \cdot \frac{2}{5} = \frac{12}{25}$$

3. Sekarang hitung  $P(A \mid \text{manis})$  pakai teorema bayes :

$$* P(A \mid \text{manis}) = \frac{P(\text{manis} \mid A) \cdot P(A)}{P(\text{manis})}$$

$$* P(A \mid \text{manis}) = \frac{\frac{2}{3} \cdot \frac{3}{5}}{\frac{12}{25}} = \frac{50}{60} = \frac{5}{6} = 0.833 = 83.3\%$$

4. Kesimpulan :

setelah kamu tahu cokelat manis, peluang besar bahwa kamu ambil dari kotak A adalah  $\frac{5}{6}$  atau sekitar 83.3%

## Langkah-langkah



Tentukan Prior  $P(A)$



Hitung likelihood  $P(B|A)$



Hitung Evidence  $P(B)$



Gunakan rumus bayes untuk mendapatkan posterior  $P(A|B)$

# Naive Bayes Classifier (Machine Learning)

- Algoritma naïve Bayes memanfaatkan teorema Bayes untuk tugas klasifikasi. Dengan Asumsi penyederhanaan yang NAÏVE!

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} \quad \leftarrow \text{teorema bayes}$$

- Jika bukti yang ingin dipakai banyak, atau bukti ini biasanya disebut fitur.

$$P(A|b_1, b_2, \dots, b_n) = \frac{P(A) \times P(b_1, b_2, \dots, b_n|A)}{P(b_1, b_2, \dots, b_n)} \quad \begin{matrix} \text{class} \\ \text{fitur} \end{matrix}$$

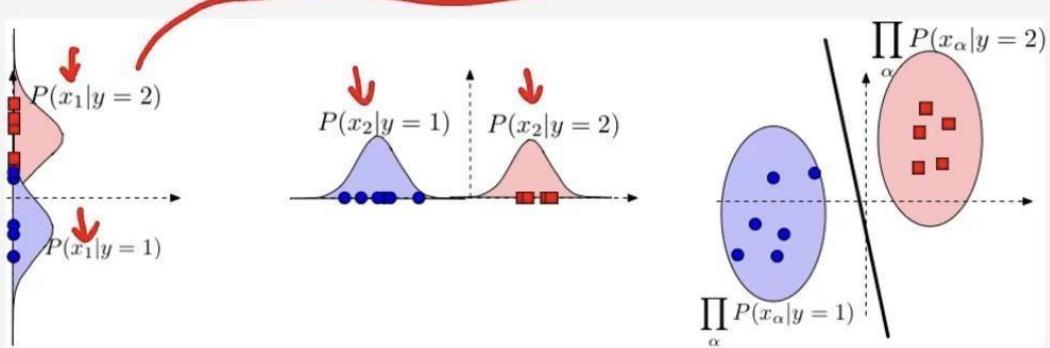
- Naïve Bayes mengasumsikan fitur-fitur ini independent, maka dapat disederhanakan:

$$\begin{aligned} P(b_1, b_2, \dots, b_n|A) &= P(b_1|A) \times P(b_2|A) \times P(b_3|A) \times \dots \times P(b_n|A) \\ &= \prod_{i=1}^n P(b_i|A) \end{aligned}$$

# Naive Bayes Classifier (Machine Learning)

- Naïve Bayes mengasumsikan fitur-fitur ini independent, maka dapat disederhanakan:

$$P(b_1, b_2, \dots, b_n|A) = \prod_{i=1}^n P(b_i|A)$$
$$P(A_m|b_1, b_2, \dots, b_n) = \frac{P(A_m) \times \prod_{i=1}^n P(b_i|A_m)}{P(b_1, b_2, \dots, b_n)}$$



# Contoh

Email ID	Teks Email	Label
1	Beli obat murah	Spam
2	Diskon besar sekarang	Spam
3	Halo apa kabar	Bukan Spam
4	Pertemuan tim besok	Bukan Spam
5	Obat terlaris diskon	Spam

Prediksi email baru: "Murah Sekarang"

$A$  = Label / Kelas

- Jumlah total email: 5
- Jumlah email "Spam": 3
- Jumlah email "Bukan Spam": 2

$$P(\text{spam}) = P(A_1) = \frac{\text{jumlah spam}}{\text{total email}} = \frac{3}{5} = 0.6$$

$$P(\text{Bukan spam}) = P(A_2) = \frac{\text{jumlah bukan spam}}{\text{total email}} = \frac{2}{5} = 0.4$$

# Contoh

Tabel Fitur/likelihood kata pada kelas spam

$$\begin{aligned} P(\text{kata|spam}) &= P(\text{kata}|A_1) = \frac{\text{Jumlah kemunculan kata} + 1}{\text{Total kata di kelas spam} + \text{total kosakata}} \\ &= \frac{\text{jumlah kemunculan kata}}{9 + 13} \end{aligned}$$

Kata	Jumlah di spam	$P(\text{kata spam})$
beli	1	$\frac{(1+1)}{22} = 0.0909$
obat	2	0.1364
murah	1	0.0909
diskon	2	0.1364
besar	1	0.0909
Sekarang	1	0.0909
Terlaris	1	0.0909
(kata lain)	0	0.0455

# Contoh

Tabel Fitur/likelihood kata pada kelas bukan spam

$$P(\text{kata|bukan spam}) = P(\text{kata}|A_2) = \frac{\text{Jumlah kemunculan kata} + 1}{\text{Total kata di kelas bukan spam} + \text{total kosakata}}$$

$$= \frac{\text{jumlah kemunculan kata}}{6 + 13}$$

Kata	Jumlah di bukan spam	$P(\text{kata bukan spam})$
halo	1	$\frac{(1+1)}{19} = 0.1053$
apa	1	0.1053
kabar	1	0.1053
pertemuan	1	0.1053
tim	1	0.1053
besok	1	0.1053
(kata lain)	0	0.0526

## Contoh

Prediksi email baru: "Murah Sekarang"

$$P(\text{spam}) = P(A_1) = \frac{\text{jumlah spam}}{\text{total email}} = \frac{3}{5} = 0.6$$

$$P(\text{Bukan spam}) = P(A_2) = \frac{\text{jumlah bukan spam}}{\text{total email}} = \frac{2}{5} = 0.4$$

$$P(\text{Spam|email baru}) > P(\text{Bukan Spam|email baru})$$

$$0.004958 \times \frac{1}{P(\text{semua kata})} > 0.001107 \times \frac{1}{P(\text{semua kata})}$$

$$0.004958 > 0.001107$$

**SPAM!**

$$P(\text{Spam|email baru}) = \frac{P(\text{Spam}) \cdot P(\text{murah|spam}) \cdot P(\text{sekarang|spam})}{P(\text{semua kata})}$$

$$= 0.6 \times \frac{2}{22} \times \frac{2}{22} \times \frac{1}{P(\text{semua kata})}$$

$$= 0.004958 \times \frac{1}{P(\text{semua kata})}$$

$$P(\text{Bukan Spam|email baru}) = \frac{P(\text{Bukan Spam}) \cdot P(\text{murah|bukan spam}) \cdot P(\text{sekarang|bukan spam})}{P(\text{semua kata})}$$

$$= 0.4 \times \frac{1}{19} \times \frac{1}{19} \times \frac{1}{P(\text{semua kata})}$$

$$= 0.001107 \times \frac{1}{P(\text{semua kata})}$$



# Naïve Bayes



Naïve Bayes Classifier menggunakan teorema Bayes



Fitur bersifat independen



Cepat dan efektif untuk klasifikasi teks



Contoh: `sklearn.naive_base.GaussianNB.MultinomialNB`

## Ringkasan

- Teorema Bayes menggabungkan informasi baru dengan pengetahuan awal
- Digunakan dalam banyak aplikasi nyata dan algoritma Machine Learning
- Penting untuk memahami asumsi dan interpretasi hasil

## Chapter 3. Random Variables Probability Distribution

### 3-1. What are random variables

01

## Memahami konsep Random Variable

02

## Membedakan antara fenomena diskrit dan kontinu

03

## Memahami pentingnya Random Variable untuk pemodelan

### Apa Itu Variables/Variabel?

Simbol yang mewakili kuantitas yang dapat berubah

Contoh: Suhu, Jumlah\_penjualan, Usia

$$\text{Suhu} = 72^\circ$$



### Apa Itu Random Variables?

Dalam Probabilitas, nilai dari variable ditentukan oleh **proses acak** atau **eksperimen**

Contoh:

- Jika kita melempar koin, hasilnya bisa "Heads" (H) atau "Tails" (T). Ini bukan angka.
- Bagaimana jika kita ingin menganalisis hasilnya secara matematis? Kita perlu mengubahnya menjadi angka.

**CONTOH!**

**Percobaan:** Melempar sebuah koin satu kali

**Ruang Sampel ( $S$ ):** {Angka, Gambar}



↓      ↓  
1      0

**Solusi:**

definisi:

$X=1 \rightarrow$  hasil  $\rightarrow$  angka

$X=0 \rightarrow$  hasil  $\rightarrow$  gambar

$$P(X=1) = ? \frac{1}{2} \quad P(X=0) = ? \frac{1}{2}$$

$$\begin{aligned} P(\text{angka}) &\rightarrow P(X=1) \\ P(\text{gambar}) &\rightarrow P(X=0). \end{aligned}$$

**Tantangan:** Analisis secara kuantitatif

**CONTOH!**

**Percobaan:** Melempar sebuah koin **dua** kali

**Ruang Sampel ( $S$ ):**

{AA, AG, GA, GG}

0,25 0,25 0,25 0,25



$$P(X=1) = P(AG) + P(GA)$$

$$= 0,25 + 0,25$$

$$= \cancel{0,5}$$

**Tantangan:** Analisis secara kuantitatif

**Solusi:**

definisi ~~R~~  $\rightarrow X =$  jumlah angka yg muncul

$X=0 \rightarrow$  Angka 0

$X=1 \rightarrow$  Angka 1 kali

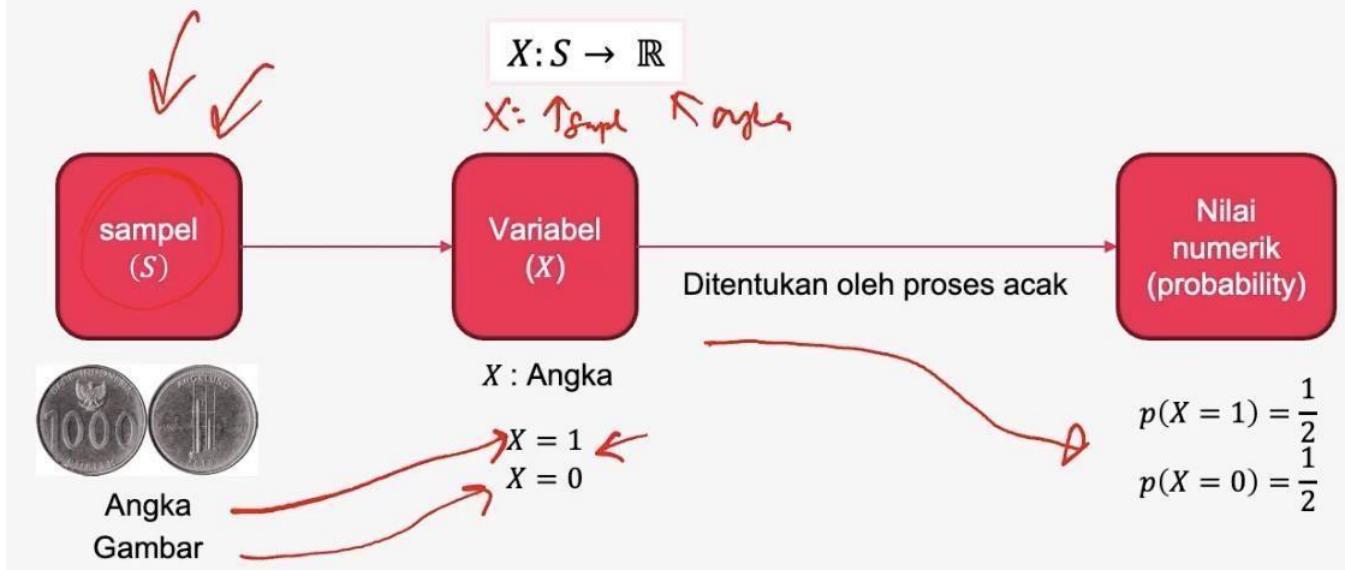
$$X = \{0, 1, 2\}$$

$X=2 \rightarrow$  Angka 2 kali muncul.



# Definisi Formal?

Dilambangkan dengan huruf kapital misalkan  $X, Y$  atau  $Z$ , yang memetakan sample ( $S$ ) ke bilangan real ( $\mathbb{R}$ )



## Diskrit VS Kontinu

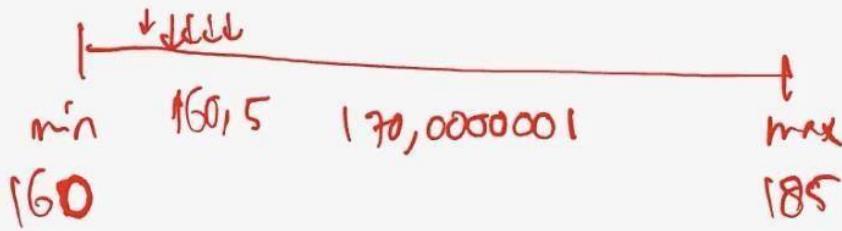
fisikamodern00-2625220

### Discrete Random Variable (Variabel Acak Diskrit)

- Mengambil nilai-nilai yang dapat dihitung atau terpisah.  $1, 2, 3, 4, \dots$
- Jumlah terbatas atau tak terhingga, tapi masih bisa dihitung (misalnya 1, 2, 3, ...)
- Contoh: Jumlah anak di sebuah keluarga, jumlah mobil yang lewat di persimpangan

### Continuous Random Variable (Variabel Acak Kontinu)

- Mengambil nilai dalam sebuah interval yang tak terhitung jumlahnya.
- Nilai-nilai diukur, bukan dihitung.
- Contoh: Tinggi badan seseorang, suhu ruangan, waktu yang dibutuhkan untuk menyelesaikan sebuah tugas



# Let's Differentiate!

Diskrit/Kontinu

Waktu yang dibutuhkan sebuah baterai untuk habis

Kontinu

Jumlah kecelakaan mobil dalam sehari di Bandung

Diskrit

Tingkat gula darah seseorang

Kontinu

Jumlah buku yang terjual di sebuah toko dalam sehari

Diskrit

## RECAP

**Apa itu random variables:** Sebuah fungsi yang mengubah hasil non-numerik dari sebuah percobaan acak menjadi nilai numerik

**Tujuannya:** Memungkinkan kita untuk menggunakan alat-alat matematika dan statistic untuk menganalisa dan memodelkan fenomena acak

**Dua tipe utama:**

- Diskrit: Nilai yang dapat dihitung (count)
- Kontinu: Nilai yang diukur dalam sebuah interval

**Next up:** Kita akan membahas lebih dalam tentang distribusi probabilitas untuk variable acak diskrit dan kontinu

### 3-2. Discrete Probability Distribution - 1

01

## Memahami dan menginterpretasikan Probability Mass Function (PMF) untuk menghitung probabilitas

02

## Mengidentifikasi dan menerapkan model distribusi binomial dan distribusi poisson dalam masalah nyata

03

## Menghitung Expected Value (nilai harapan) dan Variance (varians) untuk distribusi binomial dan poisson)

### Pengantar Probability Mass Function (PMF)

$$F(x) = \dots$$

#### Apa itu PMF?

- PMF adalah sebuah fungsi yang memberikan probabilitas bahwa sebuah variabel acak diskrit akan sama dengan nilai tertentu.
- Singkatnya, ini adalah peta yang menunjukkan seberapa sering setiap nilai yang mungkin muncul.

#### Definisi

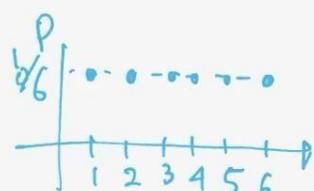
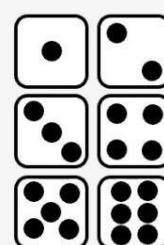
- Untuk variabel acak diskrit  $X$ , PMF dilambangkan sebagai  $P(X = x)$
- Probabilitas untuk setiap nilai harus non-negative:  $P(X = x) \geq 0$
- Total semua probabilitas:  $\sum_x P(X = x) = 1$

lisikamodern00-2625220

### Contoh never dies!

#### Percobaan

Melempar sebuah dadu standar enam sisi



#### Hitung

- Variabel acak  $X$ : hasil lemparan dadu
- Sampel  $S$ :  $\{1, 2, 3, 4, 5, 6\}$
- PMF untuk setiap hasil (asumsi dadu adil):

$$P(X=1) = ?$$

total probabilitas.

$$P(X=1) = \frac{1}{6}$$

$$\sum P(X=x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$P(X=2) = \frac{1}{6}$$

$$+ \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$P(X=6) = \frac{1}{6}$$

$$= \frac{1}{6}$$

# Distribusi Binomial

Digunakan ketika kita memiliki percobaan dengan jumlah pengulangan yang tetap ( $n$ ), di mana setiap percobaan hanya memiliki dua hasil yang mungkin (sukses atau gagal)

Syarat Percobaan Binomial:

1. Ada  $n$  percobaan yang independen.
2. Setiap percobaan memiliki dua hasil: Sukses atau Gagal.
3. Probabilitas sukses ( $p$ ) tetap konstan di setiap percobaan.

guru atau ayah



$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Di mana  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Annotations:  
- "jumlah berhasil" points to  $k$   
- "jumlah percobaan" points to  $n$   
- "probabilitas" points to  $p$

## Distribusi Binomial

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Annotations:  
-  $n=1$ ,  $k=1$ ,  $p=0,5$   
- 1 percobaan, 1 berhasil  
-  $P(X = 1) = \frac{1!}{1!(1-1)!} (0,5)^1 (1-0,5)^{1-1} = 0,5$   
-  $n-k$   
-  $0,5 \cdot 0,5 (0,5)^0 = 0,5$   
-  $k=0$   
- 1 percobaan, 1 gagal



## Distribusi Binomial

$$n=2, k=1, p=0.5$$

2 percobaan, 1 berhasil lalu 1 gagal

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$P(X = 1) = \frac{2!}{1!(2-1)!} (0.5)^1 (1-0.5)^{2-1} = 0.5$$

$$n=2, k=2, p=0.5$$

2 percobaan, 2 berhasil

$$P(X = 2) = \frac{2!}{2!(2-2)!} (0.5)^2 (1-0.5)^{2-2} = 0.25$$

$$n=2, k=0, p=0.5$$

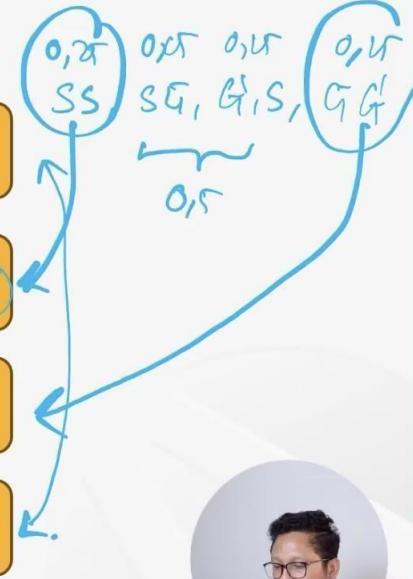
2 percobaan, 2 gagal

$$P(X = 0) = \frac{2!}{0!(2-0)!} (0.5)^0 (1-0.5)^{2-0} = 0.25$$

$$n=2, k=1, p=0.5$$

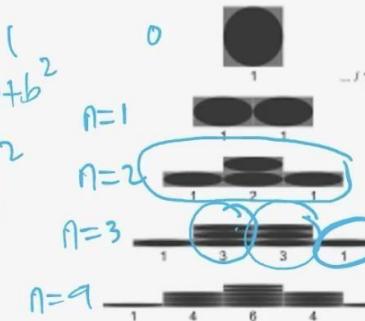
2 percobaan, 1 gagal lalu 1 berhasil

$$P(X = 1) = \frac{2!}{1!(2-1)!} (0.5)^1 (1-0.5)^{2-1} = 0.5$$



## Distribusi Binomial

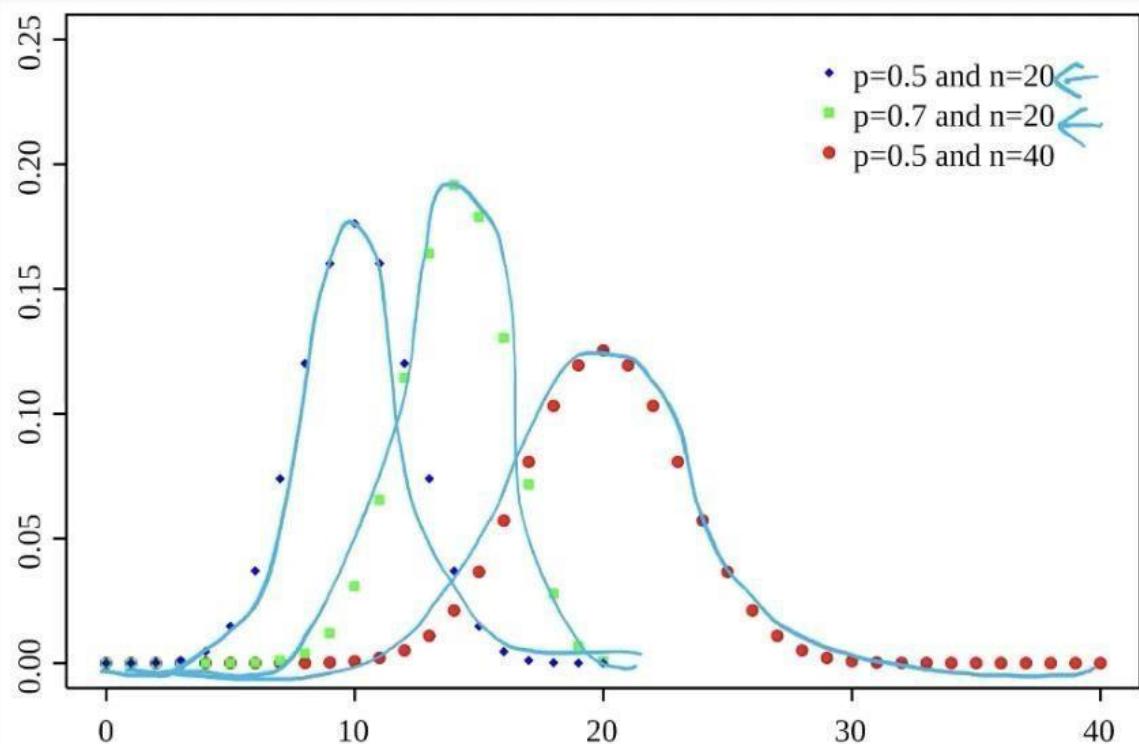
$$\begin{aligned} & 1 \quad a^2 + 2ab + b^2 \quad n=0 \\ & (a+b)^2 \quad n=1 \\ & (a+b)^3 \quad n=2 \\ & (a+b)^9 \quad n=3 \\ & \vdots \quad \vdots \end{aligned}$$



$$\begin{aligned} & S \quad G \\ & SS \quad \{ SG, GS \}, GG \\ & SSS \quad \{ GGS, SSG \} \\ & \vdots \quad \vdots \\ & SSSS \quad \{ GGG, SGS, GSS, SGD \} \\ & \vdots \quad \vdots \\ & SSSSS \quad \{ GGGG, SGGS, SSGS, SGS, GSSG, SGD, GGD \} \\ & \vdots \quad \vdots \\ & SSSSSS \quad \{ \dots \} \end{aligned}$$

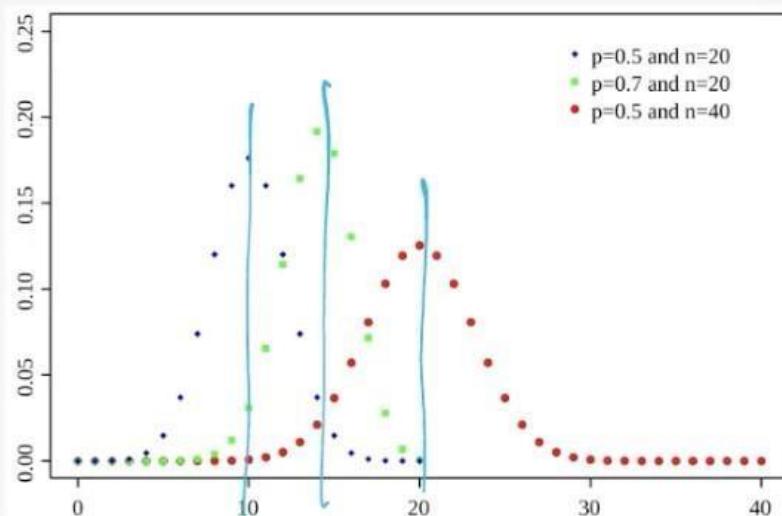


# Distribusi Binomial



## Expected Value

Expected Value ( $E[X]$ ): Rata-rata atau nilai yang diharapkan dari variable acak jika percobaan diulang berkali-kali



$$E[X] = n \cdot p$$

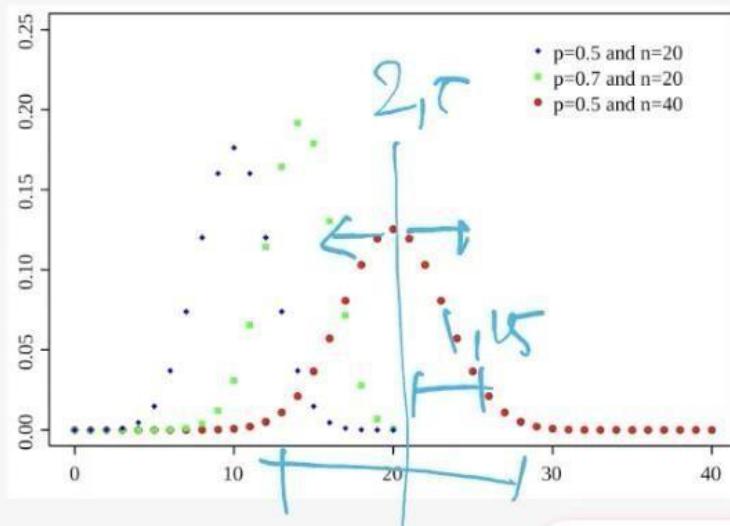
$n$ : Jumlah percobaan  
 $p$ : Probabilitas

Contoh: Melempar koin 5 kali

$$E[X] = 0.5 \cdot 5 = 2.5$$

# Variance

Variance ( $Var[X]$ ): Mengukur seberapa jauh nilai-nilai yang mungkin tersebar dari nilai harapan.



$$Var[X] = n \cdot p \cdot (1 - p)$$

n: Jumlah percobaan

p: Probabilitas

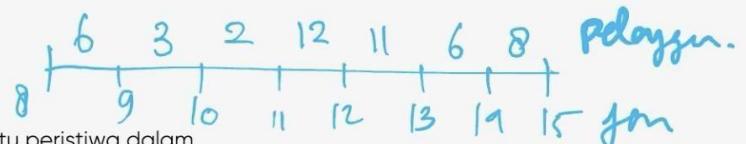
Contoh: Melempar koin 5 kali

$$\begin{aligned} Var(X) &= 2\pi \cdot (1 - 0.5) \\ &= 1.25 \end{aligned}$$

## 3-3. Discrete Probability Distribution - 2

### Distribusi Poisson

Digunakan ketika kita menghitung jumlah kejadian suatu peristiwa dalam interval waktu atau ruang yang tetap, dengan asumsi kejadian-kejadian terjadi secara independent pada Tingkat rata-rata yang konstan ( $\lambda$ )



Parameter ( $\lambda$ ): Rata-rata jumlah kejadian dalam interval yang diberikan

Contoh: Rata-rata 3 pelanggan tiba di sebuah toko per jam

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Di mana

$$e \approx 2.71828$$

## Distribusi Poisson

Jika probabilitas sukses untuk setiap percobaan adalah  $p = \frac{\lambda}{n}$ , di mana  $\lambda$  adalah rata-rata (mean) dan  $n$  adalah jumlah percobaan

$$P(X = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Karena percobaan pada domain kontinu maka  $n$  berjumlah tak hingga

$$\begin{aligned} \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} &= \lim_{n \rightarrow \infty} \left[ \frac{n(n-1) \dots (n-k+1)}{k!} \right] \left[ \frac{\lambda^k}{n^k} \right] \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \left[ \frac{\lambda^k}{k!} \right] \left[ \frac{n(n-1) \dots (n-k+1)}{n^k} \right] \left(1 - \frac{\lambda}{n}\right)^{n-k} \end{aligned}$$

## Distribusi Poisson

$$\begin{aligned} \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} &= \lim_{n \rightarrow \infty} \left[ \frac{n(n-1) \dots (n-k+1)}{k!} \right] \left[ \frac{\lambda^k}{n^k} \right] \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \left[ \frac{\lambda^k}{k!} \right] \left[ \frac{n(n-1) \dots (n-k+1)}{n^k} \right] \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \left[ \frac{\lambda^k}{k!} \right] \left[ \frac{n \cdot n-1 \cdot \dots \cdot n-k+1}{n^n} \right] \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \left[ \frac{\lambda^k}{k!} \right] \left[ \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \right] \\ P(X = k) &= \left[ \frac{\lambda^k}{k!} \right] e^{-\lambda} \end{aligned}$$

# Distribusi Poisson

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$e \approx 2.71828$

**Skenario:** Rata-rata 4 pelanggan tiba di sebuah toko per jam

**Tantangan:** Berapa probabilitas tepat 2 pelanggan tiba dalam jam berikutnya

## Parameter:

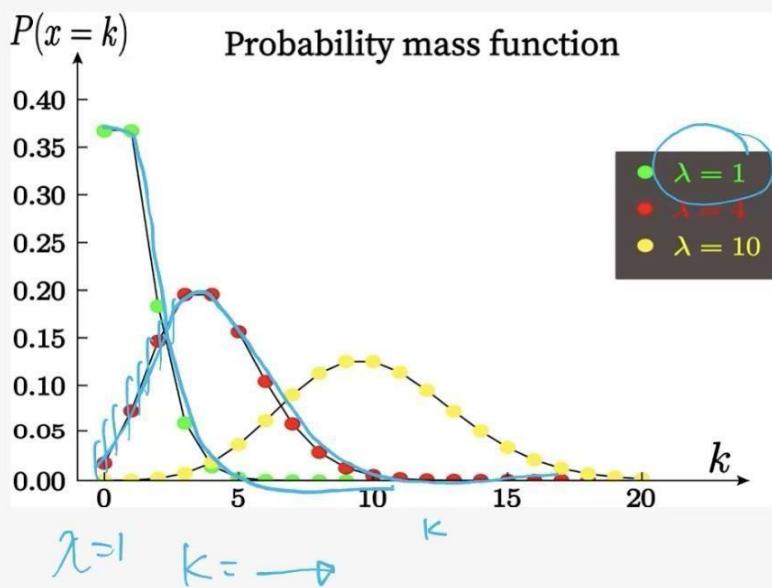
- $\lambda = 4$  (rata-rata pelanggan per jam)
- $k = 2$  (jumlah kejadian yang diinginkan)

### Perhitungan:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$P(X = 2) = \frac{e^{-4} \cdot 4^2}{2!} = \frac{e^{-4} \cdot 16}{2 \cdot 1} = 0,229 \quad 22,9\%$$

# Distribusi Poisson



Distribusi Poisson menunjukkan bahwa semakin tinggi nilai lambda (rata-rata) maka distribusi akan bergeser ke kanan.

Mean  $E[X] = \text{Variance } Var[X] = \lambda$

## Recap

- **Probability Mass Function** : Fungsi untuk menghitung probabilitas nilai spesifik dari variable acak
- **Binomial**: Digunakan untuk percobaan dengan  $n$  tetap, dua hasil (sukses/gagal), dan probabilitas sukses yang konstan.
- **Poisson**: Digunakan untuk menghitung jumlah kejadian dalam interval dengan rata-rata konstan

### Next Up:

Setelah memahami distribusi diskrit, selanjutnya kita akan menyelami dunia Continuous Probability Distributions, termasuk Cumulative Distribution Function (CDF).

### 3-4. Continuous Probability Distribution

01

 **Memahami bagaimana Probability Density Function (PDF) menggambarkan variabel acak kontinu**

02

**Mengaplikasikan distribusi normal dan distribusi uniform dalam konteks praktis**

03

**Menginterpretasikan area di bawah kurva sebagai probabilitas dan memahami konsep probabilitas total**

## Pengantar Probability Density Function (PDF)

### Apa itu PDF?

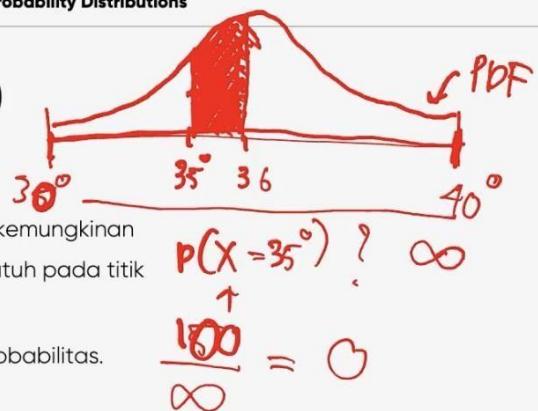
- PMF adalah sebuah fungsi yang menggambarkan kemungkinan relatif bahwa sebuah variabel acak kontinu akan jatuh pada titik tertentu.
- Berbeda dengan PMF, nilai PDF ( $f(x)$ ) bukanlah probabilitas. Probabilitas dihitung untuk sebuah **rentang** nilai

### Penting!

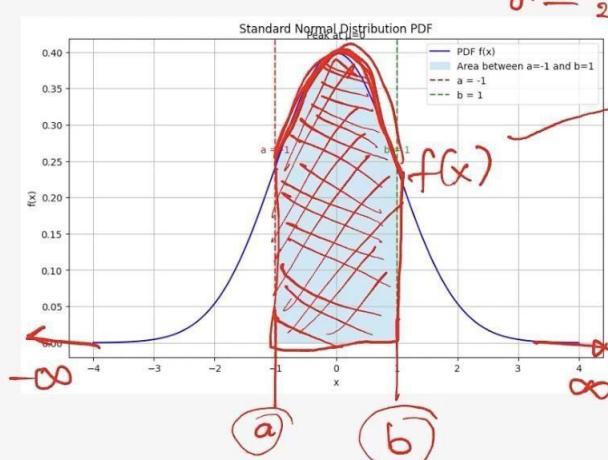
- Untuk variabel acak kontinu, probabilitas bahwa  $X$  sama dengan nilai tertentu adalah nol

$$P(X = x) = 0$$

- Ini karena ada jumlah tak terhingga dari nilai yang mungkin dalam sebuah interval



## Area di Bawah Kurva



$$\text{Luas} = \int_0^2 y \cdot dx = \int_0^2 1 \cdot dx = x \Big|_0^2 = 2 - 0 = 2$$

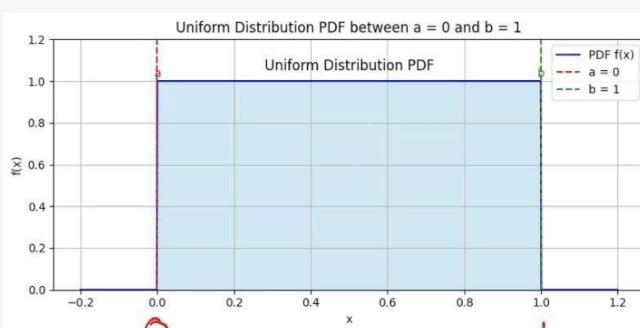
- Probabilitas  $X$  berada dalam interval  $[a, b]$  dihitung sebagai area di bawah kurva PDF dari  $a$  hingga  $b$

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

- Probabilitas Total adalah luas area seluruh kurva PDF  $f(x)$  dari minus tak hingga ke plus tak hingga

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

## Distribusi Uniform (Seragam)



- **Kapan digunakan?** Ketika setiap nilai dalam sebuah interval memiliki probabilitas yang sama untuk terjadi.

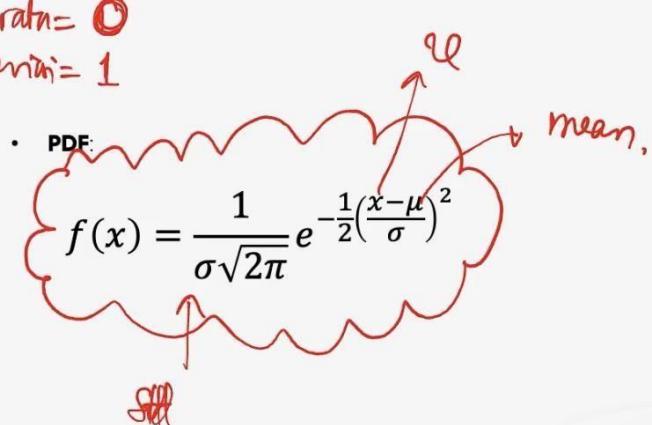
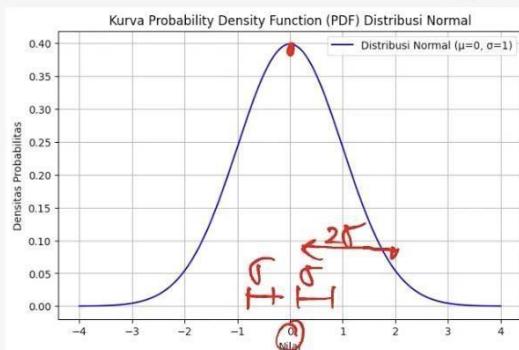
### Ciri-ciri:

- o PDF-nya berbentuk persegi panjang atau "datar"  $a \leq x \leq b$
- o  $f(x) = \frac{1}{b-a}$  untuk  $a \leq x \leq b$   $\rightarrow f(x) = \frac{1}{b-a}$
- o  $f(x) = 0$  di luar interval

- **Contoh:** Waktu tunggu untuk bus yang datang setiap 10 menit. Waktu tunggu antara 0 dan 10 menit memiliki probabilitas yang sama

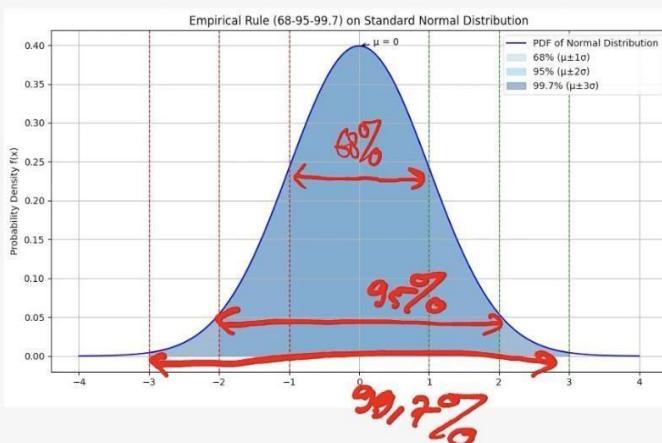


## Distribusi Normal (Gaussian)



- Kapan digunakan?** Ini adalah distribusi paling umum dalam statistik dan data science. Banyak fenomena alami mengikuti pola ini.
- Ciri-ciri:**
  - Bentuk kurva lonceng (bell curve)
  - Simetris sekitar nilai rata-rata atau mean  $\mu$
  - Ditentukan oleh dua parameter: mean  $\mu$  dan standard deviasi  $\sigma$

## Aturan Empiris (68 – 95 – 99.7)



Dalam distribusi Normal, probabilitas suatu nilai jatuh dalam rentang tertentu dari rata-rata dapat diestimasi dengan mudah:

- Sekitar 68% data berada dalam 1 standar deviasi dari rata-rata ( $\mu \pm 1\sigma$ ).
- Sekitar 95% data berada dalam 2 standar deviasi dari rata-rata ( $\mu \pm 2\sigma$ ).
- Sekitar 99.7% data berada dalam 3 standar deviasi dari rata-rata ( $\mu \pm 3\sigma$ ).

Aturan ini adalah cara cepat untuk memahami sebaran data.

## Ringkasan

- PDF** mendeskripsikan probabilitas untuk variabel acak kontinu, di mana probabilitas dihitung sebagai area di bawah kurva.
- Distribusi Uniform** memiliki probabilitas yang sama di seluruh interval, dengan bentuk persegi panjang.
- Distribusi Normal** adalah distribusi berbentuk lonceng yang sangat umum, ditentukan oleh **mean ( $\mu$ )** dan **standard deviation ( $\sigma$ )**.
- Area di bawah kurva PDF** adalah representasi visual dari probabilitas. Total area selalu 1.

NEXT UP → Cumulative Distribution Function (CDF).

### 3-5. Cummulative Distribution Function (CDF)

01

## Menggunakan Cumulative Distribution Function (CDF) untuk menghitung probabilitas

02

## Memvisualisasikan dan menginterpretasikan probabilitas kumulatif melalui grafik

### Pengantar Cumulative Distribution Function (CDF)

#### Apa itu CDF?

- CDF adalah sebuah fungsi yang memberikan probabilitas bahwa sebuah variabel acak ( $X$ ) akan mengambil nilai yang kurang dari atau sama dengan nilai tertentu ( $x$ ).
- CDF dilambangkan sebagai ( $F(x)$ )

#### Formulasi

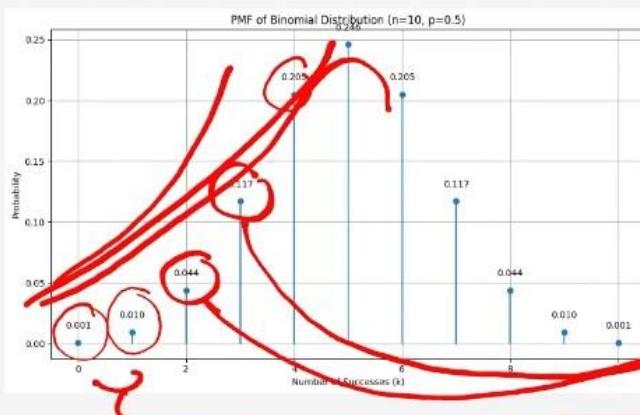
$$F(x) = P(X \leq x)$$



$$F(x) = P(X \leq x)$$

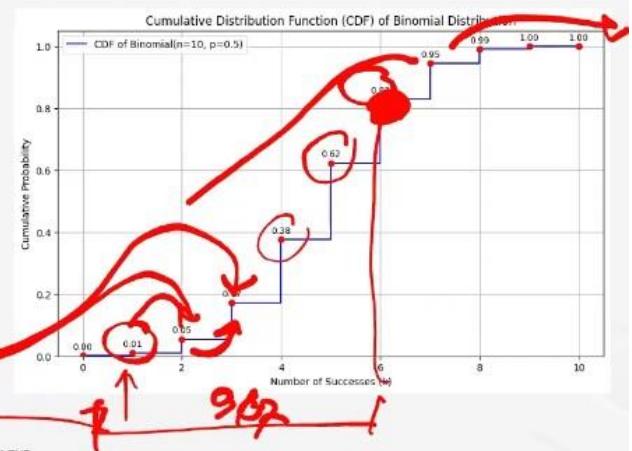
- CDF memberitahu kita total probabilitas yang "terakumulasi" hingga titik tertentu.
- Nilai CDF selalu berada di antara 0 dan 1
- Fungsi CDF selalu non-turun (Monotonically non-decreasing)

### Variabel Diskrit PMF dan CDF



↓ PMF

↓ CDF



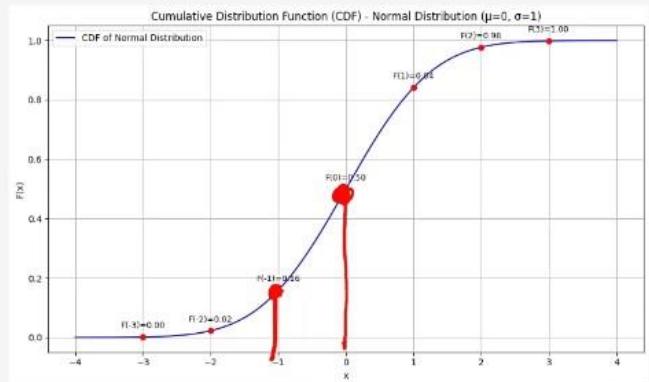
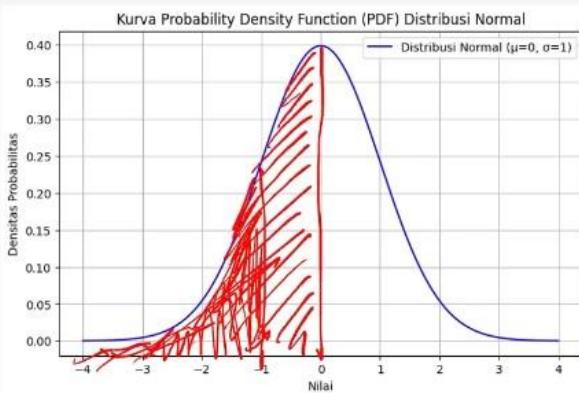
- Hubungan antara PMF dan CDF Adalah bahwa CDF merupakan jumlah dari semua nilai PMF hingga titik  $x$

$$F(x) = \sum_{k \leq x} P(X=k) = P(X=0) + P(X=1) + P(X=2) + \dots + P(X=x)$$

$$F(x) = \sum_{k \leq x} P(X=k)$$

↑ PMF

## Variabel Kontinu PDF dan CDF



- Hubungan antara PDF dan CDF Adalah bahwa CDF merupakan integral fungsi PDF hingga titik  $x$  dari  $-\infty$

$$F(x) = \int_{-\infty}^x f(k)dk$$

fisikamodem00-2025220

## Kenapa CDF?



- Dengan CDF kita dengan mudah menghitung probabilitas di antara dua nilai

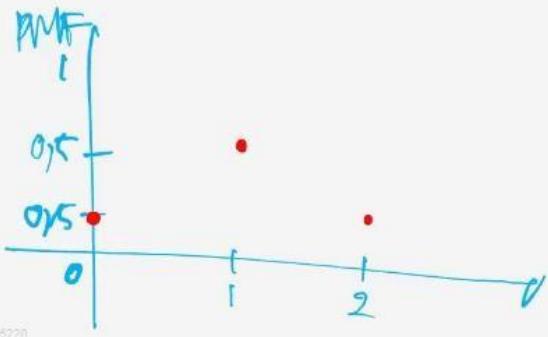
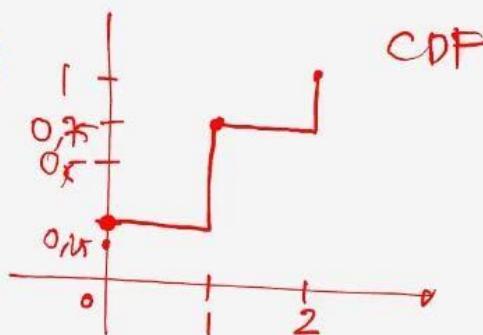
$$\begin{aligned} P(a \leq X \leq b) &= F(b) - F(a) \\ &= F(b) - F(a). \end{aligned}$$

- Berlaku untuk distribusi diskrit maupun kontinu

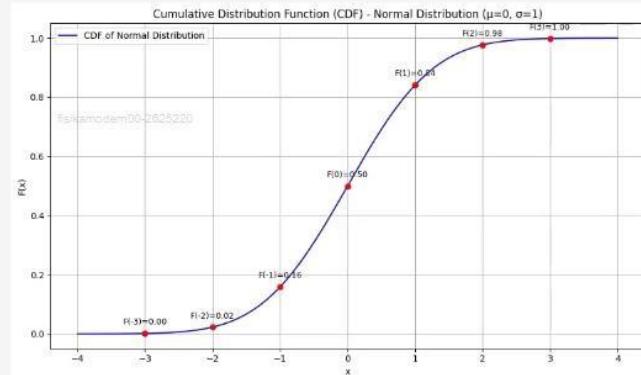
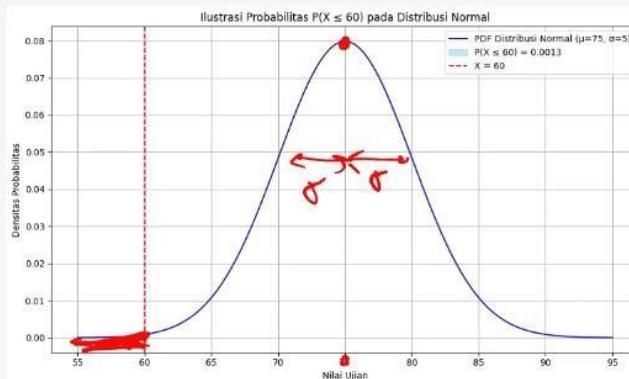
## Contoh diskrit:

- Skenario:** Melempar dua koin,  $A$  = jumlah Angka
- Nilai yang mungkin:**  $a = \{0,1,2\}$
- PMF:**  $P(A = 0) = 0.25, P(A = 1) = 0.5, P(A = 2) = 0.25$
- CDF ( $F(a)$ ):**

$\text{AA } \text{AG } \text{GA } \text{GG}$   
 $P(A=0) = 0,25$   
 $P(A=1) = 0,5$   
 $P(A=2) = 0,25$



## Contoh kontinu:



- Seorang mahasiswa ingin mengetahui probabilitas bahwa nilai ujian statistiknya (yang mengikuti distribusi normal standar dengan rata-rata  $\mu = 75$  dan standar deviasi  $\sigma = 5$ ) ~~lebih~~ dari 60.

$$P(X \leq 60) = \int_{-\infty}^{60} f(x) dx$$

- Seorang mahasiswa ingin mengetahui probabilitas bahwa nilai ujian statistiknya (yang mengikuti distribusi normal standar dengan rata-rata  $\mu = 75$  dan standar deviasi  $\sigma = 5$ ) kurang dari 60.
- Z-score: Mengubah distribusi normal umum menjadi distribusi normal standar (dengan  $\mu = 0, \sigma = 1$ )

$$Z = \frac{X - \mu}{\sigma} = \frac{60 - 75}{5} = -3$$

↓      ↓      ↗  
-15      5

- Sehingga

$$P(X \leq 60) = P(Z \leq -3) = 0.0013$$

$$= 0,13\%$$

Z	CDF P(Z ≤ z)	Z	CDF P(Z ≤ z)
-3.4	0.0003	0.1	0.5398
-3.3	0.0005	0.2	0.5793
-3.2	0.0007	0.3	0.6179
-3.1	0.0010	0.4	0.6554
-3.0	0.0013	0.5	0.6915
-2.9	0.0019	0.6	0.7257
-2.8	0.0026	0.7	0.7580
-2.7	0.0035	0.8	0.7881
-2.6	0.0047	0.9	0.8159
-2.5	0.0062	1.0	0.8413
-2.4	0.0082	1.1	0.8643
-2.3	0.0107	1.2	0.8849
-2.2	0.0139	1.3	0.9032
-2.1	0.0179	1.4	0.9192
-2.0	0.0228	1.5	0.9332
-1.9	0.0287	1.6	0.9452
-1.8	0.0359	1.7	0.9554
-1.7	0.0446	1.8	0.9641
-1.6	0.0548	1.9	0.9713
-1.5	0.0668	2.0	0.9772
-1.4	0.0808	2.1	0.9821
-1.3	0.0968	2.2	0.9861
-1.2	0.1151	2.3	0.9893
-1.1	0.1357	2.4	0.9918
-1.0	0.1587	2.5	0.9938
-0.9	0.1841	2.6	0.9953
-0.8	0.2119	2.7	0.9965
-0.7	0.2420	2.8	0.9974
-0.6	0.2743	2.9	0.9981
-0.5	0.3085	3.0	0.9987
-0.4	0.3446	3.1	0.9990
-0.3	0.3821	3.2	0.9993
-0.2	0.4207	3.3	0.9995
-0.1	0.4602	3.4	0.9997
0.0	0.5000		

## Contoh CDF untuk PDF normal dengan Z-score:

- Seorang mahasiswa ingin mengetahui probabilitas bahwa nilai ujian statistiknya (yang mengikuti distribusi normal standar dengan rata-rata  $\mu = 75$  dan standar deviasi  $\sigma = 5$ ) kurang dari 80 dan lebih besar dari 70
- Z-score: Mengubah distribusi normal umum menjadi distribusi normal standar (dengan  $\mu = 0, \sigma = 1$ )

$$Z_1 = \frac{X - \mu}{\sigma} = \frac{80 - 75}{5} = 1$$

$$Z_2 = \frac{X - \mu}{\sigma} = \frac{70 - 75}{5} = -1$$

- Sehingga

$$P(70 \leq X \leq 80) = P(X \leq 80) - P(X \leq 70)$$

68%

$$= P(z \leq 1) - P(z \leq -1)$$

$$= 0,8913 - 0,1587 = 0,6822$$

Z	CDF P(Z ≤ z)	Z	CDF P(Z ≤ z)
-3.4	0.0003	0.1	0.5398
-3.3	0.0005	0.2	0.5793
-3.2	0.0007	0.3	0.6179
-3.1	0.0010	0.4	0.6554
-3.0	0.0013	0.5	0.6915
-2.9	0.0019	0.6	0.7257
-2.8	0.0026	0.7	0.7580
-2.7	0.0035	0.8	0.7881
-2.6	0.0047	0.9	0.8159
-2.5	0.0062	1.0	0.8413
-2.4	0.0082	1.1	0.8643
-2.3	0.0107	1.2	0.8849
-2.2	0.0139	1.3	0.9032
-2.1	0.0179	1.4	0.9192
-2.0	0.0228	1.5	0.9332
-1.9	0.0287	1.6	0.9452
-1.8	0.0359	1.7	0.9554
-1.7	0.0446	1.8	0.9641
-1.6	0.0548	1.9	0.9713
-1.5	0.0668	2.0	0.9772
-1.4	0.0808	2.1	0.9821
-1.3	0.0968	2.2	0.9861
-1.2	0.1151	2.3	0.9893
-1.1	0.1357	2.4	0.9918
-1.0	0.1587	2.5	0.9938
-0.9	0.1841	2.6	0.9953
-0.8	0.2119	2.7	0.9965
-0.7	0.2420	2.8	0.9974
-0.6	0.2743	2.9	0.9981
-0.5	0.3085	3.0	0.9987
-0.4	0.3446	3.1	0.9990
-0.3	0.3821	3.2	0.9993
-0.2	0.4207	3.3	0.9995
-0.1	0.4602	3.4	0.9997
0.0	0.5000		

# Ringkasan

- **CDF:** Fungsi esensial yang mengukur probabilitas kumulatif  $P(X \leq x)$ , berperan sebagai "total berjalan" probabilitas dari sebuah variabel acak.
- **Korelasi PMF/PDF:** CDF adalah bentuk terintegrasi dari PMF (untuk diskrit) atau PDF (untuk kontinu), menghubungkan keduanya dalam satu kerangka kerja.
- **Perhitungan Interval:** Memudahkan perhitungan probabilitas untuk suatu rentang,  $P(a < X \leq b) = F(b) - F(a)$ , dengan mengurangi probabilitas kumulatif.
- **Grafik CDF:** Visualisasi fundamental yang menunjukkan perbedaan antara distribusi Diskrit (grafik tangga) dan Kontinu (kurva halus berbentuk S).

Felikmodern00-2625220

## 3-6. Comparing And Choosing Distribution

01

**Mengidentifikasi dan memilih distribusi probabilitas yang sesuai untuk skenario yang berbeda berdasarkan sifat data.**

02

**Memahami perbedaan utama dalam bentuk, nilai rata-rata (mean), dan varians di antara distribusi diskrit dan kontinu yang umum.**

## Perbedaan Kunci: Diskrit vs. Kontinu

Kriteria	Distribusi Diskrit	Distribusi Kontinu
Variabel Acak	Mengambil nilai yang dapat dihitung (count), terpisah. Contoh: jumlah mobil di tempat parkir, jumlah pelanggan.	Mengambil nilai dalam rentang yang tak terhitung jumlahnya (measured). Contoh: tinggi badan, suhu, waktu tempuh.
Probabilitas	Diberikan oleh PMF ( $P(X = x)$ ). Probabilitas untuk satu titik bisa lebih besar dari nol. Total probabilitas dihitung dengan menjumlahkan PMF.	Diberikan oleh PDF ( $f(x)$ ). Probabilitas untuk satu titik adalah nol: $P(X = x) = 0$ . Total probabilitas dihitung dengan mengintegrasikan PDF.
Grafik PDF	Grafik berbentuk tangga.	Kurva halus dan berkelanjutan.
Contoh	Binomial, Poisson	Uniform, Normal

## Kapan Menggunakan Distribusi Diskrit

- Distribusi Binomial:
  - Gunakan ketika Anda menghitung jumlah sukses dalam jumlah percobaan tetap ( $n$ ) yang independen. Setiap percobaan memiliki dua hasil yang mungkin (sukses/gagal), dan probabilitas sukses ( $p$ ) harus konstan. Empat asumsi ini harus dipenuhi untuk menerapkan model Binomial.
- Contoh:
  - Perusahaan e-commerce menguji tampilan website baru. Dari 1000 pengunjung ( $n = 1000$ ), berapa probabilitas tepat 50 orang ( $k = 50$ ) melakukan pembelian jika probabilitas pembelian rata-rata adalah 4% ( $p = 0.04$ )?
- Properti:
  - Rata-rata ( $\mu$ ) dari distribusi ini adalah  $n \cdot p$ . Varians ( $\sigma^2$ ), yang mengukur sebaran data, adalah  $n \cdot p \cdot (1 - p)$ .

## Kapan Menggunakan Distribusi Diskrit

- Distribusi Poisson:
  - Gunakan ketika Anda menghitung jumlah kejadian dalam interval waktu atau ruang tetap, dengan tingkat rata-rata kejadian yang konstan ( $\lambda$ ). Asumsi utamanya adalah kejadian-kejadian tersebut bersifat independen.
- Contoh:
  - Rata-rata 7 mobil melewati tol per menit ( $\lambda = 7$ ). Berapa probabilitas bahwa akan ada tepat 5 mobil yang lewat dalam menit berikutnya? Contoh lain: Menghitung jumlah cacat per meter persegi pada lembaran baja.
- Properti:
  - Karakteristik unik dari distribusi Poisson adalah nilai rata-rata ( $\mu$ ) dan varians ( $\sigma^2$ ) yang sama, yaitu  $\lambda$ .

- Distribusi Uniform (Seragam):
  - Gunakan ketika setiap nilai dalam sebuah rentang tertentu memiliki probabilitas yang sama untuk terjadi. PDF-nya berupa garis datar di antara batas bawah ( $a$ ) dan batas atas ( $b$ ).
- Contoh:
  - Sebuah sistem pembangkit bilangan acak menghasilkan angka antara 0 dan 10. Probabilitas untuk mendapatkan angka apa pun di antara 0 dan 10 adalah sama. Contoh lain adalah waktu tunggu untuk bus, di mana waktu tunggu bisa di mana saja dalam interval kedatangan.
- Properti:
  - Rata-rata ( $\mu$ ) berada tepat di tengah interval, yaitu  $\frac{a+b}{2}$ . Variansnya dihitung sebagai  $\frac{(b-a)^2}{12}$ .

## Kapan Menggunakan Distribusi Kontinu

- Distribusi Normal:
  - Ini adalah distribusi paling umum dalam statistik. Gunakan ketika data memiliki bentuk lonceng yang simetris di sekitar nilai rata-rata ( $\mu$ ). Banyak fenomena alami, sosial, dan teknis mengikuti pola ini. Penting untuk dicatat bahwa peran vital distribusi Normal akan dijelaskan secara rinci dalam bab **Central Limit Theorem**.
- Contoh:
  - Tinggi badan populasi dewasa, skor tes standar, atau error pengukuran pada eksperimen ilmiah. Kurva lonceng yang terbentuk menunjukkan bahwa sebagian besar data berkumpul di sekitar rata-rata.
- Properti:
  - Ditentukan oleh dua parameter utama: Mean ( $\mu$ ) yang menunjukkan pusat distribusi, dan Standard Deviation ( $\sigma$ ) yang mengukur seberapa jauh data menyebar dari pusat.

# Perbedaan Kunci: Diskrit vs. Kontinu

Distribusi	Tipe	Parameter Kunci	Mean $E[X]$	Variance $Var[X]$
Binomial	Diskrit	$n$ (jumlah percobaan), $p$ (probabilitas sukses)	$np$	$np(1 - p)$
Poisson	Diskrit	$\lambda$ (rata-rata kejadian per interval)	$\lambda$	$\lambda$
Uniform	Kontinu	$a$ (batas bawah), $b$ (batas atas)	$\frac{a + b}{2}$	$\frac{(b - a)^2}{12}$
Normal	Kontinu	$\mu$ (rata-rata), $\sigma^2$ (varians)	$\mu$	$\sigma^2$

## Rekap

- Pilihan distribusi yang tepat adalah seni dan sains. Ini bergantung pada sifat variabel acak (diskrit vs. kontinu) dan jenis pertanyaan yang Anda ajukan.
- Proses berpikir dalam memilih distribusi:
  1. Tentukan apakah variabel Anda merupakan hasil hitungan (diskrit) atau pengukuran (kontinu).
  2. Jika diskrit, apakah Anda menghitung sukses dari jumlah percobaan tetap (Binomial) atau kejadian dalam interval tertentu (Poisson)?
  3. Jika kontinu, apakah probabilitasnya seragam (Uniform) atau terkonsentrasi di sekitar rata-rata (Normal)?
- Distribusi adalah bahasa untuk menceritakan kisah data Anda. Memilih distribusi yang tepat berarti Anda menggunakan model yang paling akurat untuk "menceritakan" apa yang terjadi.

## Chapter 4. Central Limit Theorem (CLT)

### 4-1. What Is Central Limit Theorem

01

**Memahami konsep dasar Teorema Limit Pusat (CLT) dan mengapa ini begitu penting.**

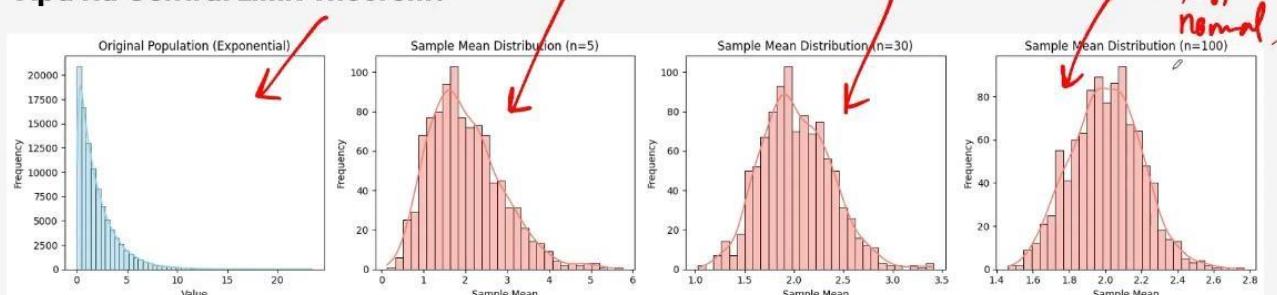
02

**Membedakan antara distribusi populasi dan distribusi sampel.**

03

**Menggunakan CLT untuk membuat kesimpulan yang akurat tentang populasi dari sampel.**

#### Apa itu Central Limit Theorem?



Central Limit Theorem (CLT) adalah pilar fundamental dalam statistika modern. Secara formal, CLT menyatakan bahwa jika Anda mengambil sampel acak yang cukup besar ( $n \geq 30$ ) dari populasi mana pun, distribusi rata-rata sampelnya akan mendekati distribusi Normal (kurva lonceng).

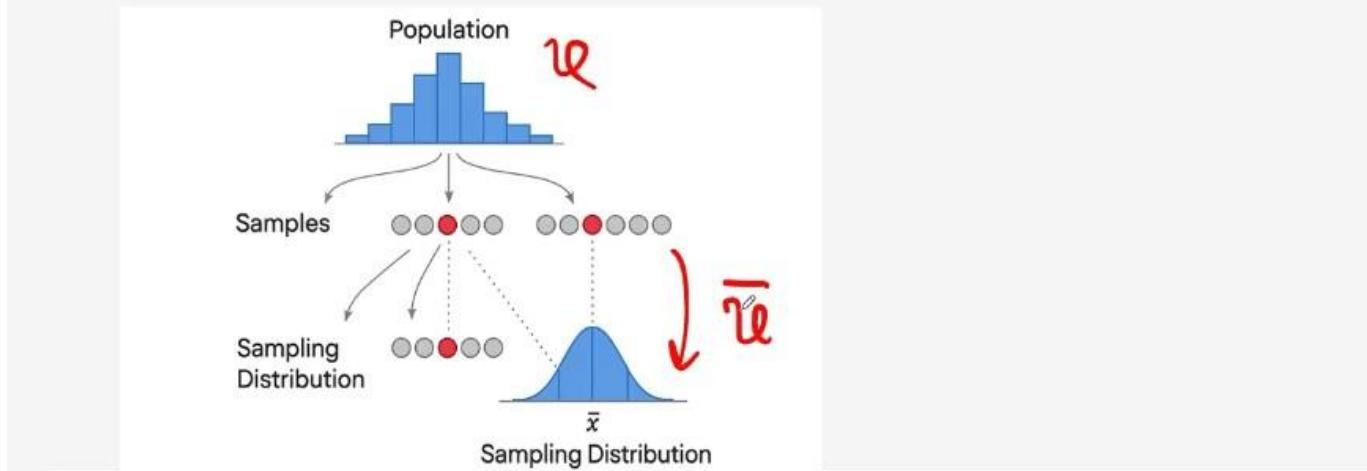
#### Apa itu Central Limit Theorem?

- Mengapa Ini Luar Biasa?** Teorema ini berlaku meskipun populasi aslinya tidak berdistribusi normal—bisa saja sangat miring (skewed), seragam (uniform), atau bentuk apa pun yang tidak terduga. CLT seperti sebuah "konverter ajaib" yang selalu menghasilkan kurva lonceng, selama Anda memberinya sampel yang cukup besar.
- Makna Praktis:** Ini adalah kunci yang memungkinkan kita bekerja dengan rata-rata sampel seolah-olah mereka berasal dari distribusi normal yang dapat diprediksi. Ini sangat penting karena sifat-sifat distribusi normal sangat dipahami dan mudah untuk dianalisis secara matematis.

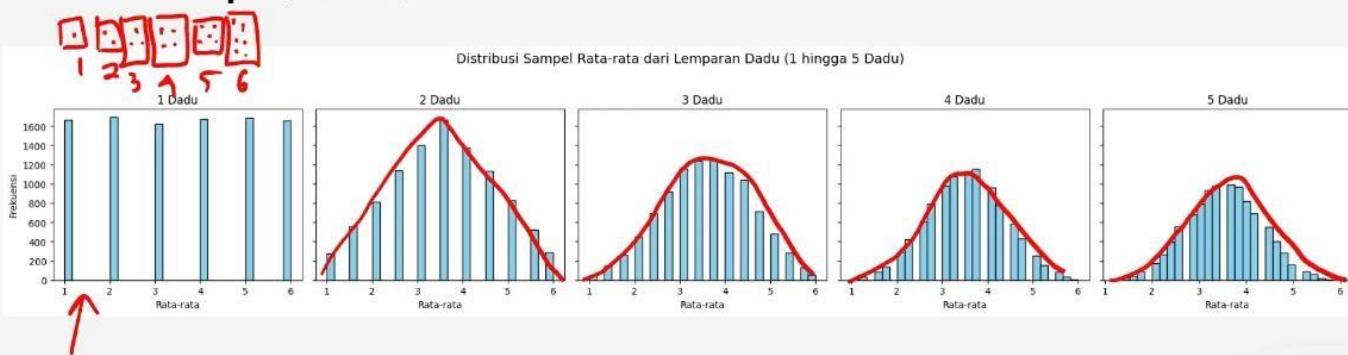
## Distribusi Sampel

$$\begin{array}{lll} K=1 & k=2 & k=L \\ N=5 & N=5 & N=5 \\ M=? & M=? & \mu=? \end{array}$$

- Distribusi sampel adalah konsep penting dalam statistik inferensial yang menggambarkan distribusi probabilitas dari suatu statistik (seperti rata-rata, proporsi, atau varians) yang dihitung dari banyak sampel acak yang diambil dari populasi yang sama.
- Distribusi sampel (sampling distribution) adalah distribusi dari nilai-nilai statistik (misalnya rata-rata sampel  $\bar{x}$ ) yang diperoleh dari berulang kali mengambil sampel acak dari suatu populasi.



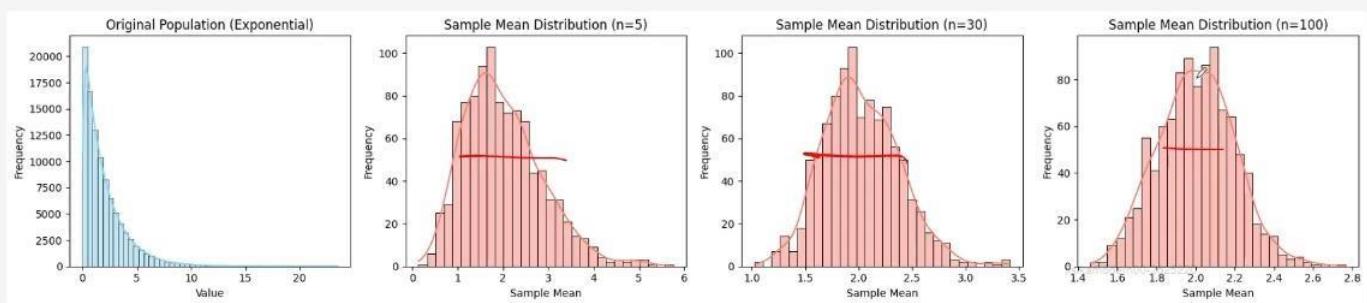
### Distribusi Sampel (Contoh)



- Setiap panel menunjukkan histogram rata-rata hasil lemparan dari ribuan sampel.
- Semakin banyak dadu yang dilempar (semakin besar  $n$ ), distribusi rata-rata hasilnya menjadi:
  - Lebih simetris ✓
  - Lebih mendekati distribusi normal ✓
  - Lebih sempit (variasi lebih kecil) ↗



## Distribusi Sampel (Contoh)



- Panel tengah dan kanan menunjukkan distribusi rata-rata sampel dari populasi tersebut untuk ukuran sampel:
  - n=5: distribusi masih mirip populasi asli, tapi mulai terlihat lebih simetris.
  - n=30: distribusi rata-rata sampel mulai menyerupai distribusi normal.
  - n=100: distribusi rata-rata sampel sangat mendekati distribusi normal.



## Proses membangun distribusi sampel

1. Ambil Sampel: Ambil sebuah sampel acak berukuran  $n$  dari populasi.
2. Hitung Rata-rata: Hitung rata-rata sampel pertama ( $\bar{x}_1$ ). Catat nilai ini.
3. Ulangi Proses: Kembalikan sampel ke populasi. Ulangi langkah 1-2 berkali-kali (misalnya 1000 kali) untuk mendapatkan serangkaian rata-rata sampel:  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{1000}$ .
4. Visualisasikan: Plot semua rata-rata sampel ini dalam sebuah histogram. Histogram inilah yang secara visual mewakili distribusi sampel.

## Properti Kunci Distribusi Sampel

Ketika CLT berlaku, distribusi rata-rata sampel (yang berdistribusi Normal) akan memiliki properti-properti yang sangat spesifik dan dapat diprediksi:

- Mean (Rata-rata): Rata-rata dari semua rata-rata sampel ( $\mu_{\bar{x}}$ ) akan sama persis dengan rata-rata populasi aslinya ( $\mu$ )

$$\mu_{\bar{x}} = \mu$$

- Standard Error: Standar deviasi dari distribusi rata-rata sampel disebut Standard Error ( $\sigma_{\bar{x}}$ ). Nilainya lebih kecil dari standar deviasi populasi aslinya ( $\sigma$ ).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$\sigma_{\bar{x}} = \frac{5}{\sqrt{10}}$$
$$\sigma = 5$$
$$n = 10$$

### Contoh BVAANG!!!!

$$\sigma = 5 \text{ gram}, n = 36, \mu = 100 \text{ gram}$$

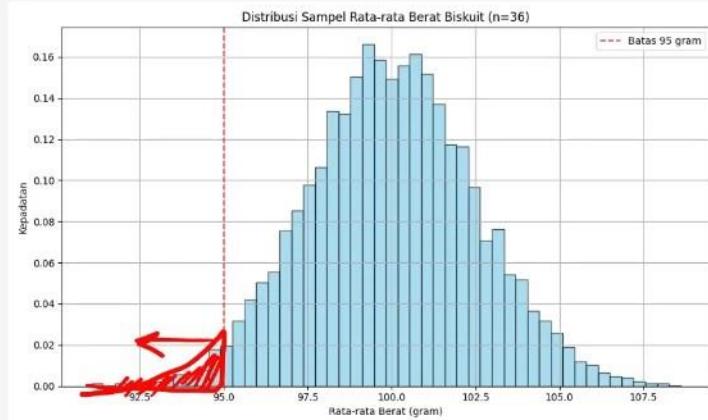
Sebuah pabrik makanan memproduksi biscuit dalam kemasan kecil. Berat biscuit dalam satu kemasan bervariasi dan mengikuti distribusi tidak normal, dengan berat rata-rata Adalah 100 gram, dan standar deviasi populasi Adalah 15 gram. Lalu seorang quality control engineer mengambil sampel acak 36 kemasan setiap jam untuk memantau kualitas produksi.

1. Mengapa engineer dapat menggunakan distribusi normal untuk menganalisis rata-rata berat sampel meskipun distribusi populasi tidak normal?  $n > 30$
2. Hitung standar deviasi dari distribusi rata-rata sampel (standard error)
3. Berapa probabilitas bahwa rata-rata berat dari 36 kemasan biscuit kurang dari 95 gram  $P(\bar{x} < 95)$  ?
4. Jika probabilitas pada soal nomor 3 sangat kecil, apa yang bisa disimpulkan?

Jika probabilitas pada soal nomor 3 sangat kecil, apa yang bisa disimpulkan

- Jika proses berjalan dengan normal, kejadian ini sangat jarang terjadi

## Contoh BVAANG!!!:



Probabilitas bahwa rata-rata berat sampel kurang dari 95 gram adalah sekitar:

$$P(\bar{X} < 95) = P(Z < -2) \approx 0.0228 = 2.28\%$$

- Diambil 10.000 sampel acak, masing-masing berukuran 36 kemasan.
- Histogram menunjukkan distribusi rata-rata berat dari tiap sampel.
- Garis merah menunjukkan batas 95 gram.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5 \text{ gram}$$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{95 - 100}{2.5} = -2$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{95 - 100}{2.5} = -2$$

- Jika rata-rata sample di ambil, misal rata-ratanya bisa dibawah 95, berarti kemungkinan berarti ada kesalahan produksi

## Rasionalisasi

- CLT berfungsi sebagai jembatan logis, memungkinkan kita membuat kesimpulan yang andal tentang populasi dari sampel yang terbatas, karena sulit untuk mengumpulkan data dari seluruh populasi. Ini adalah fondasi teoritis untuk teknik statistika inferensial.
- CLT adalah dasar untuk metode penting seperti Uji Hipotesis (menguji klaim populasi dari data sampel) dan Interval Kepercayaan (memperkirakan rentang parameter populasi).
- Keunggulan utamanya adalah fleksibilitasnya; teorema ini tetap berlaku meskipun distribusi populasi asalnya tidak diketahui atau tidak normal, asalkan ukuran sampelnya cukup besar (umumnya  $n \geq 30$ ).

# Aplikasi 1: Interval Kepercayaan

## Apa itu?

- Interval kepercayaan adalah rentang nilai yang kemungkinan besar berisi parameter populasi yang sebenarnya (misalnya, rata-rata populasi,  $\mu$ ).

## Bagaimana CLT membantunya?

- CLT memberi tahu kita bahwa distribusi rata-rata sampel adalah normal.
- Kita tahu 95% dari semua rata-rata sampel akan berada dalam  $\pm 1.96$  standard error dari rata-rata populasi.  
 $\pm 1.96$  ↑ z-score
- Ini memungkinkan kita untuk membangun interval di sekitar rata-rata sampel kita untuk mengestimasi lokasi  $\mu$  yang tidak diketahui.

## 4-3. Application Of The CLT

01 Menerapkan Central Limit Theorem pada masalah statistik di dunia nyata.

02 Mengidentifikasi kapan asumsi CLT valid dan dapat digunakan dengan tepat.

03 Memahami bagaimana CLT menjadi dasar untuk Interval Kepercayaan (Confidence Intervals) dan Uji Hipotesis (Hypothesis Testing).

## Menggunakan CLT untuk inferensi

**Ide Utama:** Karena CLT menjamin bahwa distribusi rata-rata sampel akan normal, kita dapat menggunakan alat-alat statistik yang dirancang untuk distribusi normal (seperti z-score dan tabel z) untuk membuat kesimpulan tentang populasi.

### Proses:

- Ambil sampel dari populasi. ✓
- Hitung rata-rata sampel ( $\bar{x}$ ). ✓
- Dengan CLT, kita tahu bahwa  $\bar{x}$  berasal dari distribusi normal.
- Kita bisa menggunakan sifat-sifat distribusi normal untuk menghitung probabilitas atau mengestimasi parameter populasi.

z-score

## Aplikasi 1: Interval Kepercayaan

### Contoh:

- Misalkan Anda adalah manajer kualitas di sebuah pabrik yang memproduksi botol air 500 ml. Anda ingin menaksir rata-rata volume air yang sebenarnya diisi ke dalam botol, tetapi Anda tidak dapat memeriksa setiap botol. Anda mencurigai bahwa proses pengisian tidak terdistribusi secara normal.
- Pengambilan data sampel sebanyak 100 botol dalam sehari.
- Rata-rata sampel Adalah 498 ml, dengan simpangan baku 5 ml
- Anda menginginkan Tingkat kepercayaan 95%

$$n = 100 \text{ botol}$$

$$\bar{x} = 498 \text{ ml}$$

$$\sigma = 5 \text{ ml}$$

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{5}{\sqrt{100}} \\ &= 0,5 \text{ ml}\end{aligned}$$

$$\begin{aligned}\text{Interval kepercayaan} &= \bar{x} \pm 1,96 \sigma_{\bar{x}} \\ &= 498 \pm 1,96 \times 0,5 \\ &= 498 \pm 0,98 \\ \text{batas atas} &\rightarrow 498,98 \\ \text{batas bawah} &\rightarrow 497,02 \\ 497,02 \leq \mu \leq 498,98\end{aligned}$$

## Aplikasi 2: Uji Hipotesis

### Apa itu?

- Uji hipotesis adalah metode statistik untuk membuat keputusan tentang populasi berdasarkan data sampel.

### Bagaimana CLT membantunya?

- CLT memungkinkan kita menghitung probabilitas untuk mendapatkan rata-rata sampel tertentu jika hipotesis nol (klaim awal) itu benar.
- Jika probabilitas ini sangat rendah (misalnya, kurang dari 5%), kita dapat menolak hipotesis nol dan menyimpulkan bahwa efek yang kita amati pada sampel kemungkinan besar nyata di populasi.

## Aplikasi 2: Uji Hipotesis

### Contoh:

- Misalkan ada sebuah universitas yang mengklaim bahwa rata-rata berat badan mahasiswanya adalah 65 kg. Anda sebagai peneliti tidak yakin dengan klaim tersebut dan ingin menguji hipotesis ini. Sampel diambil acak sebanyak 100 Mahasiswa, didapatkan rata-rata sampel adalah 67 kg dengan simpangan baku Adalah 10 kg. Dipilih kepercayaan hipotesis nol Adalah 95%.

faktor moden  $\bar{x} \pm 1,96\sigma_{\bar{x}}$       statistik uji  $z$ .       $z_{\text{hitung}} < z_{\text{kritis}}$ .

- Hipotesis 0  $\rightarrow \mu = 65 \text{ kg}$   $\rightarrow 95\%$        $z_{\text{hitung}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{67 - 65}{10/\sqrt{100}} = 2$        $2 < 1,96$

- Hipotesis 1  $\rightarrow \mu \neq 65 \text{ kg}$   $\rightarrow 5\%$       ~~1 < 1%~~

$n = 100$

$\bar{x} = 67 \text{ kg}$

$\sigma_{\bar{x}} = 10 \text{ kg}$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \text{ kg}$

$|z| = 2$

## Contoh Nyata dalam Berbagai Bidang

### Bisnis dan Ekonomi:

- Mengukur rata-rata kepuasan pelanggan dari survei sampel untuk memprediksi kepuasan populasi.
- Menguji apakah kampanye iklan baru meningkatkan rata-rata penjualan secara signifikan.

### Sains dan Medis:

- Mengukur efektivitas obat baru dengan menguji sampel pasien dan menguji hipotesis bahwa obat tersebut memiliki efek yang signifikan.
- Mengestimasi rata-rata tinggi tanaman setelah perlakuan pupuk.

### Penelitian Sosial:

- Menggunakan survei sampel untuk memprediksi hasil pemilu atau mengukur rata-rata pendapatan populasi.

# Ringkasan & Kapan CLT Valid?

- CLT adalah jembatan antara sampel dan populasi, memungkinkan kita menggunakan distribusi normal untuk inferensi.
- CLT Valid jika:
  - Sampel diambil secara acak.
  - Sampel memiliki ukuran yang cukup besar, biasanya  $n \geq 30$ .
- Peringatan: Jika ukuran sampel sangat kecil ( $n < 30$ ), distribusi rata-rata sampel mungkin tidak normal, terutama jika populasi asalnya tidak normal.  
Dalam kasus ini, teknik lain seperti t-distribution mungkin lebih sesuai.

## Next Up

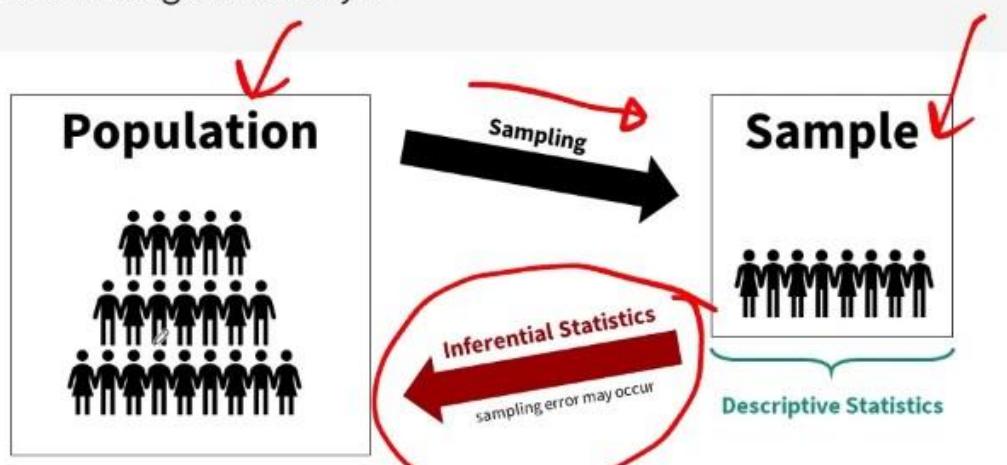
Pemahaman tentang CLT adalah fondasi penting untuk seluruh statistika inferensial. Bab selanjutnya dari kursus ini akan membahas lebih rinci tentang Inferential Statistics dan bagaimana konsep ini digunakan dalam model machine learning.

## Chapter 5. Inferential Statistics

### 5-1. What Is Inferential Statistics

## Pengantar: Apa itu Statistika Inferensial?

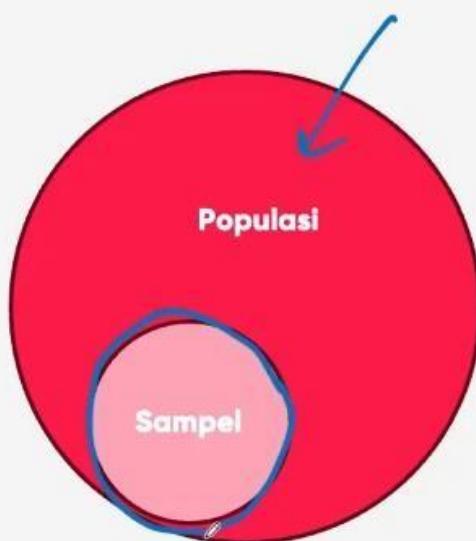
Statistika Inferensial adalah cabang statistika yang memungkinkan kita untuk membuat **kesimpulan**, **prediksi**, atau **generalisasi** tentang suatu **populasi** berdasarkan data yang hanya diambil **dari sampel**. Ini adalah alat untuk "membaca pikiran" populasi yang lebih besar hanya dengan mengamati sebagian kecilnya.



# Statistika Deskriptif vs. Inferensial: Perbedaan Mendasar

Kriteria	Statistik Deskriptif	Statistik Inferensial
Tujuan	Menggambarkan dan meringkas karakteristik data yang sudah ada dalam sampel atau populasi	Membuat kesimpulan, prediksi, atau generalisasi tentang populasi berdasarkan data dari sampel.
Fokus	Data yang diamati (sampel itu sendiri) ↪	Parameter populasi yang tidak diketahui ↪
Pertanyaan Khas	"Berapa rata-rata usia pelanggan kami di survei ini?"	"Berapa rata-rata usia semua pelanggan kami?"
Metode	Mean, median, modus, standar deviasi, frekuensi, histogram, grafik batang ↪	Uji hipotesis, interval kepercayaan, regresi, ANOVA. ↪
Contoh	"Rata-rata tinggi badan siswa di kelas ini Adalah 165 cm."	"Berdasarkan sampel, kami menyimpulkan bahwa rata-rata tinggi badan semua siswa di sekolah ini kemungkinan besar antara 163 cm dan 167 cm"

## Populasi vs. Sampel: Mengapa Inferensi Diperlukan?



- Populasi: Seluruh kelompok individu atau objek yang ingin Anda pelajari. Ini adalah target utama penelitian Anda. (Contoh: semua pemilih di suatu negara, semua bola lampu yang diproduksi oleh suatu pabrik).
- Sampel: Sebagian kecil, sub-kelompok dari populasi yang dipilih untuk analisis. Sampel harus representatif agar kesimpulan inferensial valid.

# Populasi vs. Sampel: Mengapa Inferensi Diperlukan?

Kebutuhan Inferensi:

- Keterbatasan Sumber Daya: Seringkali tidak mungkin secara fisik, finansial, atau waktu untuk mengumpulkan data dari seluruh populasi.
- Efisiensi: Mengumpulkan dan menganalisis data dari sampel jauh lebih efisien.
- Generalisasi: Inferensi memungkinkan kita untuk mengambil temuan dari sampel kecil dan menggeneralisasikannya ke populasi yang lebih besar dengan tingkat kepercayaan tertentu.

## Proses Statistika Inferensial



# Pentingnya Inferensi dalam Data Science

- Pengambilan Keputusan Berbasis Data
- Prediksi & Pemodelan ✓
- Validasi Hipotesis ✓
- Peran Central Limit Theorem (CLT)

## Rangkuman

- **Statistika Inferensial** adalah tentang membuat kesimpulan yang valid dan dapat diandalkan tentang populasi yang besar, hanya dari data sampel yang terbatas.
- Ini adalah perbedaan kunci dari **Statistika Deskriptif**, yang hanya meringkas data yang ada.
- **Inferensi** diperlukan karena keterbatasan praktis dalam mengumpulkan data populasi penuh dan untuk memungkinkan generalisasi temuan.
- **CLT** adalah kunci yang memungkinkan inferensi ini, dengan menjamin normalitas distribusi rata-rata sampel.
- **Next Up:** Di sub-bab berikutnya, kita akan menyelami salah satu aplikasi utama statistika inferensial: Interval Kepercayaan (Confidence Intervals), yang merupakan cara untuk mengestimasi parameter populasi dengan rentang nilai.

01

**Membangun dan menginterpretasi interval kepercayaan (confidence intervals).**

02

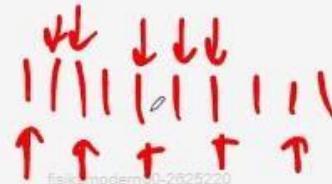
**Memahami peran ukuran sampel dan variabilitas dalam interval kepercayaan.**

03

**Mengidentifikasi perbedaan antara margin of error dan tingkat kepercayaan.**

### Apa itu Interval Kepercayaan?

#### Ide Utama:



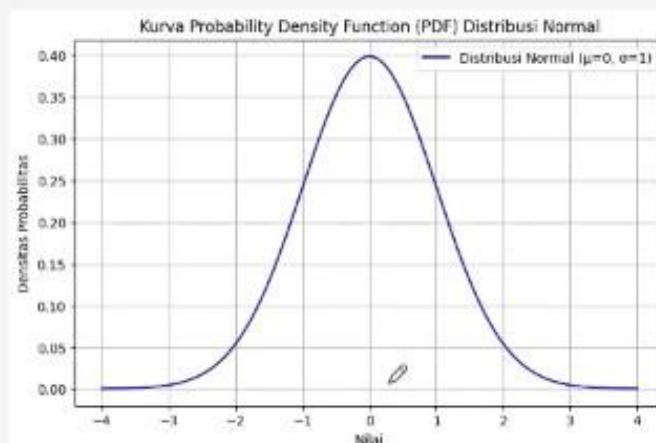
Dalam statistika inferensial, tujuan kita adalah membuat kesimpulan tentang parameter populasi

- seperti rata-rata populasi  $\mu$ ,
- proporsi populasi  $p$ , atau
- standar deviasi populasi  $\sigma$

menggunakan data yang kita kumpulkan dari sampel. Namun, rata-rata sampel ( $\bar{x}$ ) yang kita hitung hanyalah satu perkiraan titik (point estimate). Sangat kecil kemungkinannya rata-rata sampel ini akan sama persis dengan rata-rata populasi yang sebenarnya.

# Apa itu Interval Kepercayaan?

## Mengapa Perkiraan Titik Saja Tidak Cukup?



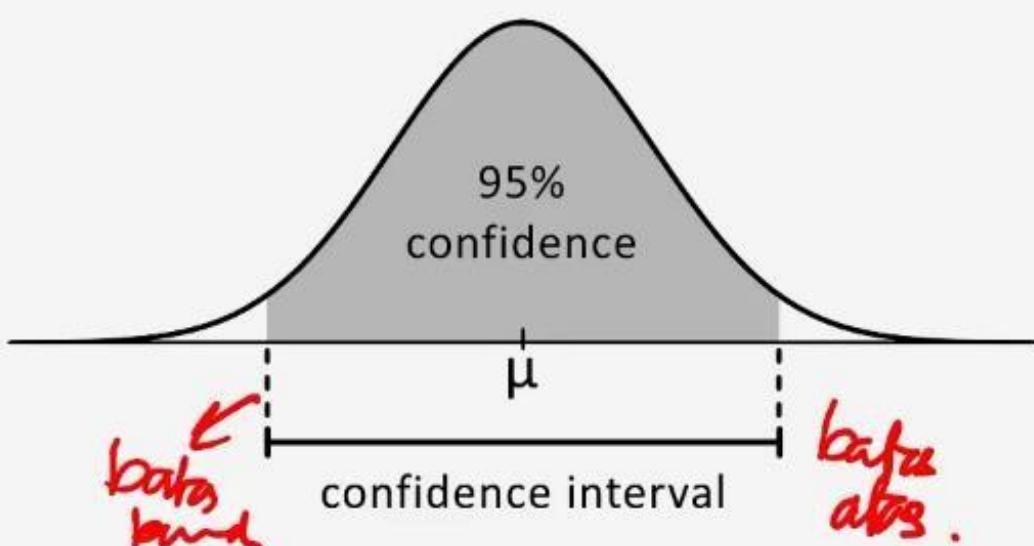
Sampel 1

Sampel 2

Sampel 3

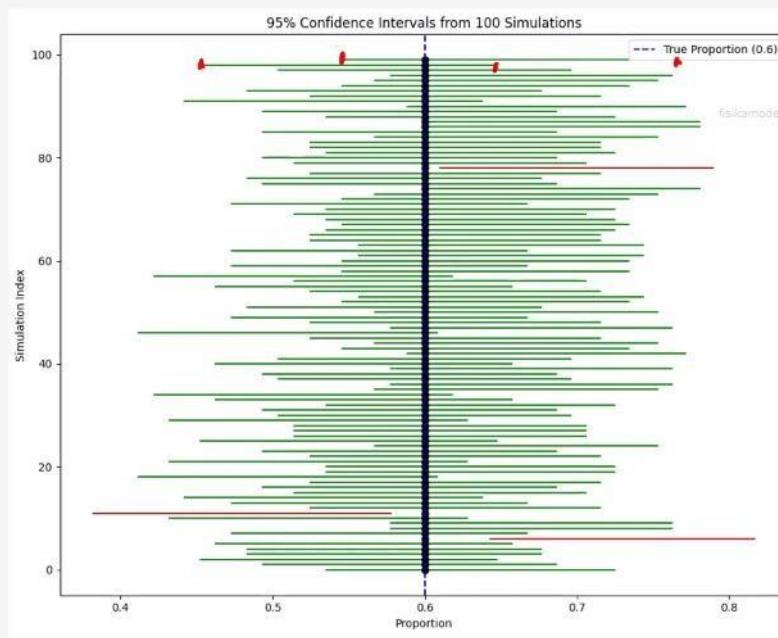
## Apa itu Interval Kepercayaan?

### Interval Kepercayaan (Confidence Interval - CI)



"Kami tidak tahu persis berapa nilai parameter populasi, tapi kami cukup yakin nilainya ada di antara X dan Y."

# Ilustrasi Confidence Interval



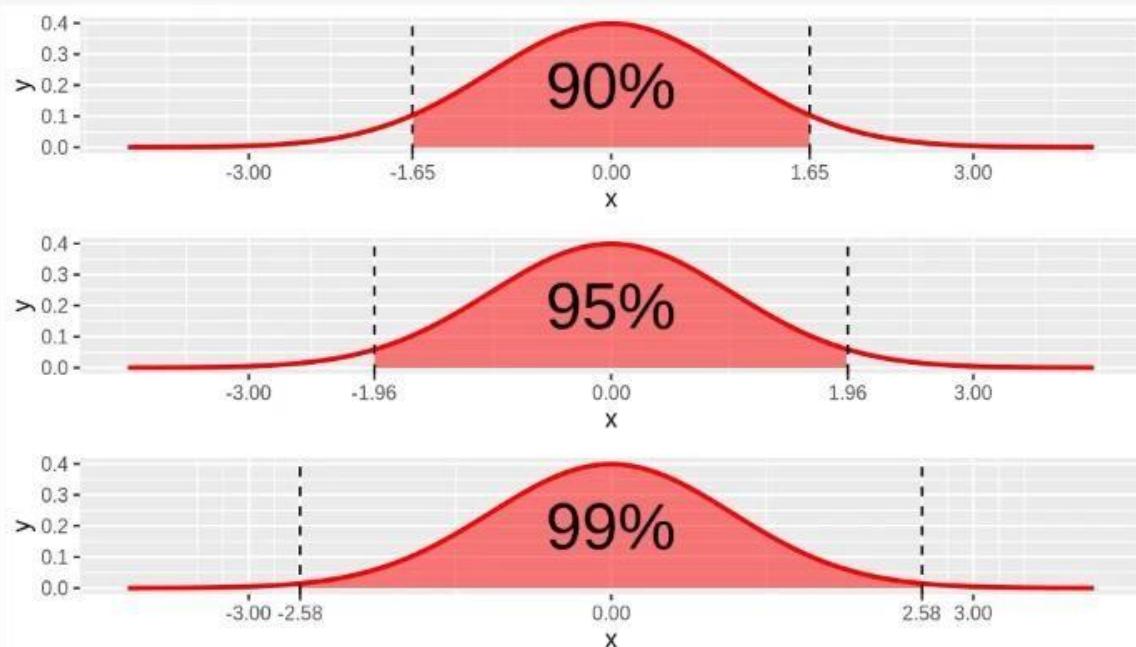
## Penjelasan Grafik:

- Setiap garis horizontal mewakili satu interval kepercayaan dari satu simulasi.
- Garis hijau menunjukkan interval yang mencakup nilai proporsi sebenarnya (0.6).
- Garis merah menunjukkan interval yang tidak mencakup nilai sebenarnya.
- Garis biru putus-putus adalah proporsi sebenarnya dari populasi (0.6).
- Titik hitam menunjukkan posisi nilai sebenarnya pada setiap simulasi.

## Confidence Level

↙ ↘ ↙

Tingkat kepercayaan yang paling sering digunakan adalah 90%, 95%, dan 99%. Pilihan tingkat kepercayaan bergantung pada konsekuensi kesalahan; untuk aplikasi medis atau keuangan yang berisiko tinggi, tingkat kepercayaan yang lebih tinggi (misalnya 99%) mungkin diperlukan.



## Komponen Utama: Margin of Error (Batas Kesalahan)

Margin of Error (MoE) adalah jumlah yang ditambahkan dan dikurangkan dari perkiraan titik untuk membentuk interval kepercayaan. Ini secara langsung mengukur presisi perkiraan kita. Semakin kecil MoE, semakin sempit intervalnya, dan semakin presisi estimasi kita.

$$\text{Margin of Error (MoE)} = \text{Nilai Kritis} \times \text{Standard Error}$$

Confidence Level (CL)	$\alpha$ (significance Level)	$Z_{\alpha/2}$ (Nilai Kritis)
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Jika standar deviasi populasi diketahui

$$\text{Standard Error} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$t-tot$

Jika standar deviasi populasi tidak diketahui, gunakan standar deviasi sampel  $s$

$$\text{Standard Error} = \sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$t-tot$

## Komponen Utama: Confidence Interval (Interval Kepercayaan)

$$\text{Interval Kepercayaan} = \bar{x} \pm \text{Margin of Error}$$

$$CI = \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Confidence Level (CL)	$\alpha$ (significance Level)	$Z_{\alpha/2}$ (Nilai Kritis)
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Formula ini digunakan ketika standard deviasi populasi atau deviasi sampel diketahui atau ketika ukuran sampel sangat besar  $> 30$ , sehingga CLT dipenuhi.

dipenuhi  
normal

## Contoh

Sebuah perusahaan e-commerce ingin mengestimasi rata-rata waktu yang dihabiskan pelanggan di situs mereka per hari. Mereka mengambil sampel acak 200 pelanggan dan menemukan rata-rata waktu yang dihabiskan adalah 35 menit, dengan standar deviasi populasi 10 menit. Mereka menghitung interval kepercayaan 95% untuk rata-rata waktu yang dihabiskan populasi dan mendapatkan hasil

$$\begin{aligned} n &= 200 \\ \bar{x} &= 35 \text{ menit} \\ \sigma &= 10 \text{ menit} \\ CL &= 95\% \end{aligned}$$
$$\begin{aligned} CI &= \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \\ &= 35 \pm 1.96 \times \frac{10}{\sqrt{200}} \\ &= 35 \pm 1.96 \times \frac{10}{14.14} \\ &= 35 \pm 1.96 \times 0.7 \\ &= 35 \pm 1.372 \end{aligned}$$

33.628      **35**      36.372

Kami 95% yakin bahwa interval (33.62 menit, 36.37 menit) berisi rata-rata waktu yang dihabiskan oleh seluruh populasi pelanggan yang sebenarnya.

## Ringkasan

- **Interval Kepercayaan** memberikan rentang perkiraan yang andal untuk parameter populasi yang tidak diketahui, bukan hanya satu titik.
- **Tingkat Kepercayaan** adalah probabilitas jangka panjang bahwa metode kita akan berhasil "menangkap" parameter populasi.
- **Margin of Error** menentukan lebar interval, dipengaruhi oleh tingkat kepercayaan, Standard Error, dan pada akhirnya, ukuran sampel dan variabilitas data.
- Untuk mendapatkan interval kepercayaan yang lebih sempit, kita membutuhkan ukuran sampel yang lebih besar dan variabilitas yang lebih rendah



Next Up: Setelah memahami Interval Kepercayaan, kita akan melanjutkan ke konsep kunci lainnya dalam statistika inferensial: Uji Hipotesis (Hypothesis Testing), yang memungkinkan kita membuat keputusan formal tentang klaim populasi.

## 5-3. Hypothesis Testing Basics

01 Memformulasikan hipotesis untuk pengujian statistik.

02 Memahami konsep kesalahan Tipe I dan kesalahan Tipe II.

03 Menentukan dan menginterpretasi tingkat signifikansi (alpha).

### Apa itu Uji Hipotesis?

**Ide Utama:** Uji hipotesis adalah metode formal untuk membuat keputusan tentang populasi berdasarkan data dari sampel. Ini adalah proses sistematis untuk menentukan apakah ada cukup bukti untuk mendukung suatu klaim atau teori.

**Proses:** Uji hipotesis membandingkan dua pernyataan yang saling bertentangan tentang suatu parameter populasi: **Hipotesis Nol** dan **Hipotesis Alternatif**.

**Contoh Sederhana:** Apakah rata-rata waktu yang dihabiskan siswa di media sosial lebih dari 2 jam per hari? Uji hipotesis akan membantu kita menjawabnya.



### Null and Alternative Hypotheses

**Setiap uji hipotesis dimulai dengan formulasi dua hipotesis:**

**Hipotesis Nol ( $H_0$ ):** Pernyataan awal yang diasumsikan benar.

- Contoh: Rata-rata waktu yang dihabiskan siswa di media sosial sama dengan 2 jam per hari
- $H_0: \mu = 2$  jam

Tujuan kita adalah mengumpulkan bukti statistik (dari sampel) untuk memutuskan apakah akan menolak  $H_0$  atau tidak.

**Hipotesis Alternatif ( $H_a$  atau  $H_1$ ):** Pernyataan yang bertentangan dengan Hipotesis Nol.

- Contoh: Rata-rata waktu yang dihabiskan siswa di media sosial lebih dari 2 jam per hari
- $H_0: \mu > 2$  jam



# Kesalahan Tipe I dan Tipe II

Dalam uji hipotesis, selalu ada risiko membuat keputusan yang salah. Ada dua jenis kesalahan:

Keputusan	$H_0$ Sebenarnya Benar	$H_0$ Sebenarnya Salah
Menolak $H_0$	Kesalahan tipe I (False Positive)	Keputusan Tepat
Tidak Menolak $H_0$	Keputusan Tepat	Kesalahan tipe II (False Negative)

Kesalahan Tipe I (False Positive): Menolak hipotesis nol padahal sebenarnya benar.

- Contoh: Menyimpulkan bahwa obat baru lebih efektif, padahal sebenarnya tidak.

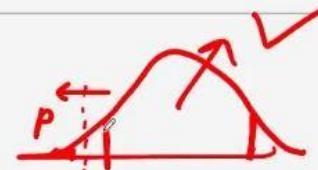
↑  
Tentu saja

Kesalahan Tipe II (False Negative): Tidak menolak hipotesis nol padahal sebenarnya salah.

- Contoh: Menyimpulkan bahwa obat baru tidak efektif, padahal sebenarnya efektif.

↑  
Tidak nyata .

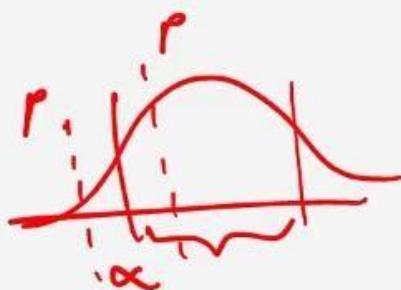
## Tingkat Signifikansi ( $\alpha$ )



- Definisi: Tingkat signifikansi, dilambangkan dengan  $\alpha$  (alpha), adalah probabilitas maksimum untuk membuat Kesalahan Tipe I. Ini adalah ambang batas yang kita tetapkan sebelum pengujian untuk menentukan seberapa banyak risiko yang kita bersedia ambil.  
 $\downarrow 95\%$     $\downarrow 99\%$     $\downarrow 90\%$
- Nilai Umum: Nilai  $\alpha$  yang paling umum adalah 0.05 (5%), 0.01 (1%), atau 0.1 (10%).
- Hubungan dengan  $H_0$ : Jika nilai  $p$  (probabilitas hasil sampel) yang kita hitung lebih kecil dari  $\alpha$ , kita memiliki bukti yang cukup untuk menolak  $H_0$ . Sebaliknya, jika nilai  $p$  lebih besar dari  $\alpha$ , kita tidak memiliki bukti yang cukup untuk menolak  $H_0$ .
- Trade-off: Menurunkan  $\alpha$  (misalnya, dari 0.05 ke 0.01) mengurangi kemungkinan Kesalahan Tipe I, tetapi meningkatkan kemungkinan Kesalahan Tipe II.

# Proses Uji Hipotesis

- Formulasikan Hipotesis: Tulis  $H_0$  dan  $H_a$ .
- Pilih Tingkat Signifikansi: Tentukan  $\alpha$  (misalnya, 0.05).
- Hitung Statistik Uji: Kumpulkan data sampel dan hitung statistik uji (misalnya, z-score atau t-score) dan nilai  $p$  ( $p$ -value).
- Ambil Keputusan: Bandingkan nilai  $p$  dengan  $\alpha$ .
  - o Jika  $p \leq \alpha$ : Tolak  $H_0$ . Ada bukti yang signifikan.
  - o Jika  $p > \alpha$ : Tidak menolak  $H_0$ . Tidak ada bukti yang signifikan.



## Contoh

fsikamodem00-2825220

Skenario: Sebuah perusahaan mengklaim bahwa rata-rata waktu yang dihabiskan pengguna di aplikasi mereka adalah 30 menit per hari.

Tantangan: Sebagai analis, Anda menduga rata-rata waktu sebenarnya lebih dari 30 menit. Anda mengambil sampel acak 100 pengguna dan menemukan rata-rata waktu yang dihabiskan adalah 32 menit. (Asumsi: standar deviasi populasi,  $\sigma = 9.76$ )

$$H_0: \mu = 30 \text{ menit}, \quad n = 100$$

$$H_a: \mu > 30 \text{ menit.} \quad \bar{U} = 32$$

$$\sigma = 9.76$$

## Jawab

Langkah 1: Formulasikan Hipotesis:

-  $H_0$  (Hipotesis Nol) : Rata-rata waktu yang dihabiskan Adalah 30 menit ( $\mu = 30$ ) ✓

-  $H_a$  (Hipotesis Alternatif) : Rata-rata waktu yang dihabiskan lebih dari 30 menit ( $\mu > 30$ ) ✓

Langkah 2: Pengujian dan Tingkat Signifikansi *Tingkat 5%*

- Kita memilih Tingkat signifikansi  $\alpha = 0.05$  ↙

Langkah 3: Detail Perhitungan Statistik Uji

- Standar Error (SE): Mengukur seberapa banyak rata-rata sampel kita bervariasi

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{9.76}{\sqrt{100}} = 0.976$$

*std dev*  
*populasi*

Fast campus

## Jawab

Langkah 3: Detail Perhitungan Statistik Uji

- Standar Error (SE): Mengukur seberapa banyak rata-rata sampel kita bervariasi

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{9.76}{\sqrt{100}} = 0.976$$

✓ *z̄*

- Z-score: Mengukur seberapa jauh rata-rata sampel kita (32) dari rata-rata yang diklaim (30) dalam satuan Standard Error

$$Z = \frac{\bar{x} - \mu}{SE} = \frac{32 - 30}{0.976} = \frac{2}{0.976} \approx 2.05$$


- Gunakan  $\alpha = 0.05$  untuk uji satu sisi.
- Nilai kritis  $Z_{(0.05)} = 1.645$ , Karena  $Z = 2.05 > 1.645$ , maka cukup bukti untuk menolak  $H_0$ .
- Nilai  $p$  dengan  $\alpha: 0.02 \leq 0.05$ , Karena nilai  $p$  lebih kecil dari  $\alpha$ , maka cukup bukti untuk menolak  $H_0$ .

**Kesimpulan:** Ada bukti yang signifikan (pada tingkat signifikansi 5%) untuk menyimpulkan bahwa rata-rata waktu yang dihabiskan pengguna di aplikasi adalah lebih dari 30 menit.

Z	CDF P(Z ≤ z)
...	...
-2.1	0.0179
2.0	0.0228 (nilai p)
-1.9	0.0287
-1.8	0.0359
-1.7	0.0446
-1.6	0.0548
...	...

## Rekap

- Uji Hipotesis adalah metode untuk membuat keputusan tentang populasi dari sampel.
- Hipotesis Nol ( $H_0$ ) adalah pernyataan awal yang kita uji, sementara Hipotesis Alternatif ( $H_a$ ) adalah klaim yang ingin kita buktikan.
- Kesalahan Tipe I (False Positive) adalah menolak  $H_0$  yang benar, dan probabilitasnya dikendalikan oleh Tingkat Signifikansi ( $\alpha$ ).
- Kesalahan Tipe II (False Negative) adalah tidak menolak  $H_0$  yang salah.

- Next Up: Di sub-bab berikutnya, kita akan menerapkan konsep ini secara praktis dengan mempelajari Z-test dan T-test, yang merupakan metode uji hipotesis yang paling umum.

### 5-4. Z-test And T-test In Practice

01

#### Memahami T-Test

02

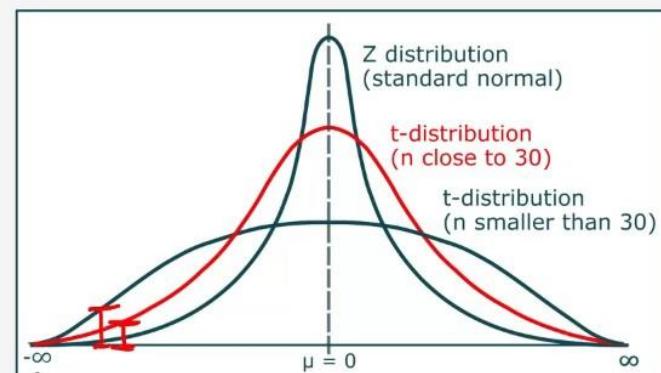
#### Mengidentifikasi kapan harus menggunakan Z-test versus T-test.

03

#### Menginterpretasi hasil uji statistik dan nilai p (p-value).

### T-test: Penjelasan Detail

- Ide Utama: T-test adalah uji hipotesis yang digunakan ketika standar deviasi populasi ( $\sigma$ ) tidak diketahui. Dalam situasi ini, kita harus menggunakan standar deviasi sampel ( $s$ ) sebagai estimasi.
- Distribusi T-Student: Karena menggunakan estimasi dari sampel, T-test menggunakan Distribusi T-Student daripada distribusi Normal. Distribusi t-Student memiliki "ekor" yang lebih tebal, yang mencerminkan ketidakpastian tambahan yang muncul dari estimasi  $\sigma$ .



Tabel T

$\alpha = 0,10$

$\alpha = 0,05$

$\alpha = 0,01$

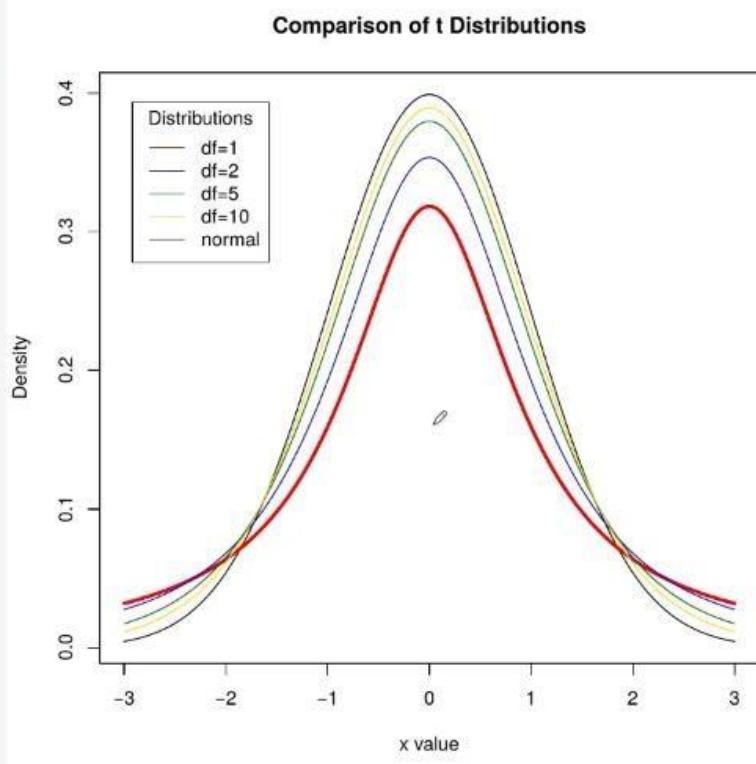
$C_1 = 10\%$

$C_1 = 5\%$

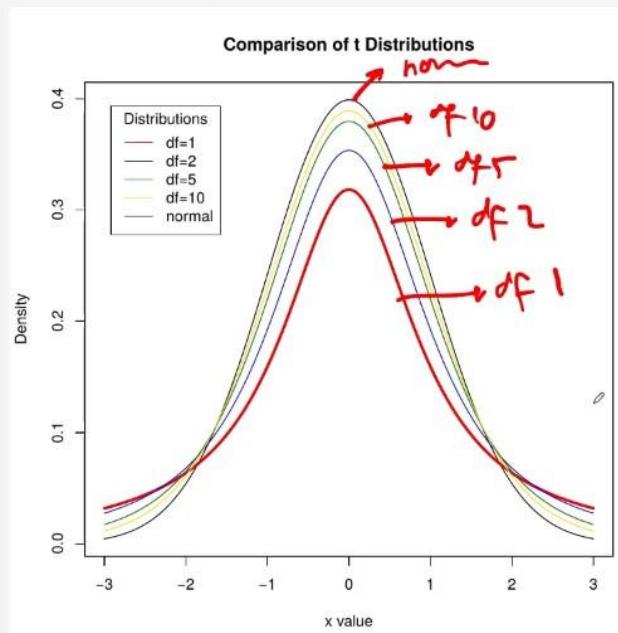
$C_1 = 1\%$

$df$	$t_{0.10}$	$t_{0.05}$	$t_{0.01}$
1	3.078	6.314	31.821
2	1.886	2.920	6.965
5	1.476	2.015	3.365
10	1.372	1.812	2.764
15	1.341	1.753	2.602
20	1.325	1.725	2.528
25	1.316	1.708	2.485
30	1.310	1.697	2.457
40	1.303	1.684	2.423
60	1.296	1.671	2.390
$\infty$	1.282	1.645	2.326

$n=1000$        $n=100$



## T-test: Penjelasan Detail



degree of freedom  
↓ dof

- Derajat Kebebasan ( $df$ ): Bentuk distribusi t-Student dipengaruhi oleh derajat kebebasan ( $df$ ), yang dihitung sebagai  $df = n - 1$ . Semakin besar  $df$  (artinya semakin besar ukuran sampel), semakin mirip distribusi t-Student dengan distribusi Normal.

# T-test Satu Sampel

Seorang dosen ingin mengetahui apakah rata-rata nilai ujian mahasiswa lebih tinggi dari nilai standar kelulusan yaitu 70. Dengan Data Sampel:

- $n = 25$  mahasiswa
- Rata-rata nilai sampel:  $\bar{x} = 73$
- Simpangan baku sampel:  $s = 8$

## ✓ Langkah-langkah Uji Hipotesis:

### 1. Rumusan Hipotesis:

- $H_0$  (Hipotesis nol):  $\mu = 70$  (rata-rata nilai sama dengan standar)
- $H_1$  (Hipotesis alternatif):  $\mu > 70$  (rata-rata nilai lebih tinggi dari standar)

### 2. Statistik Uji: Gunakan rumus **uji-t satu sampel**

$$SE = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad t = \frac{73 - 70}{8/\sqrt{25}} = 1.875$$

## Jenis-jenis T-Test

1. T-test Satu Sampel: Membandingkan rata-rata sampel dengan rata-rata populasi yang diklaim.

1. T-test Dua Sampel Independen: Membandingkan rata-rata dari dua kelompok yang tidak berhubungan.

1. T-test Berpasangan: Membandingkan rata-rata dari kelompok yang sama di dua waktu yang berbeda (misalnya, sebelum dan sesudah perlakuan).

# Memilih Uji yang Tepat

Kriteria	Z-test	T-test
Kapan digunakan	Standar deviasi populasi ( $\sigma$ ) diketahui atau $n$ sangat besar ( $n \geq 30$ )	Standar deviasi populasi ( $\sigma$ ) tidak diketahui
Distribusi	Distribusi Normal (CLT Valid)	Distribusi t-student (dengan derajat kebebasan, $df = n - 1$ )
Penerapan	Skenario yang lebih jarang di dunia nyata di mana $\sigma$ diketahui	Sangat umum di dunia nyata, di mana kita hampir selalu menggunakan standar deviasi sampel

## T-test Satu Sampel

Seorang dosen ingin mengetahui apakah rata-rata nilai ujian mahasiswa lebih tinggi dari nilai standar kelulusan yaitu 70. Dengan Data Sampel:

- $n = 25$  mahasiswa
- Rata-rata nilai sampel:  $\bar{x} = 73$
- Simpangan baku sampel:  $s = 8$

✓ Langkah-langkah Uji Hipotesis:

3. Keputusan:

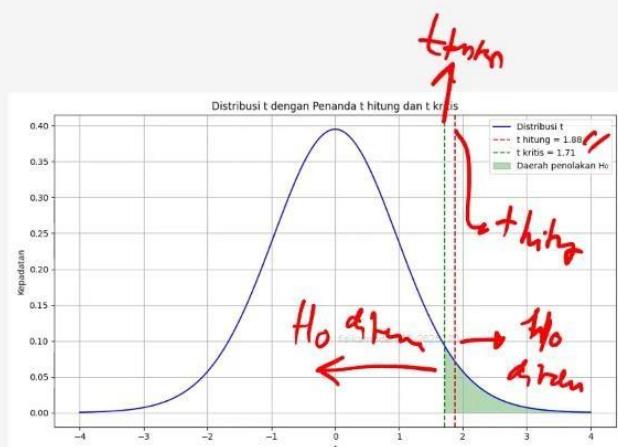
- Derajat kebebasan  $df = n - 1 = 24$
- Nilai kritis untuk  $\alpha = 0.05$  (Uji satu sisi  $t_{0.05}$ ) dan  $df = 24$  adalah 1.711

4. Kesimpulan:

- Karena  $t = 1.875 > 1.711$ , maka tolak  $H_0$

$$t = \frac{73 - 70}{8/\sqrt{25}} = 1.875$$

Ada cukup bukti untuk menyatakan bahwa rata-rata nilai mahasiswa lebih tinggi dari standar kelulusan 70 pada tingkat signifikansi 5%.



# Menginterpretasi Hasil Uji dan Nilai p

Apa itu Nilai p (p-value)?

- Nilai p adalah probabilitas untuk mendapatkan hasil sampel (atau hasil yang lebih ekstrem) yang kita amati, jika Hipotesis Nol ( $H_0$ ) benar.
- Nilai **p** yang kecil menunjukkan bahwa hasil sampel kita sangat tidak mungkin terjadi secara kebetulan jika  $H_0$  benar, sehingga memberikan bukti yang kuat untuk menolak  $H_0$ .

$$\begin{array}{l} z > z_{\text{kritis}} \\ \uparrow \\ t > t_{\text{kritis}} \end{array}$$

Aturan Keputusan:

- Jika **p**-value  $\leq \alpha$  (tingkat signifikansi): Hasilnya signifikan secara statistik. Kita menolak Hipotesis Nol ( $H_0$ ). Ini berarti ada cukup bukti untuk mendukung Hipotesis Alternatif ( $H_a$ ).
- Jika **p**-value  $> \alpha$ : Hasilnya tidak signifikan secara statistik. Kita tidak menolak Hipotesis Nol ( $H_0$ ). Ini berarti tidak ada cukup bukti dari sampel untuk menolak  $H_0$ .

## Next Up

Apa itu Nilai p (p-value)?

- Nilai p adalah probabilitas untuk mendapatkan hasil sampel (atau hasil yang lebih ekstrem) yang kita amati, jika Hipotesis Nol ( $H_0$ ) benar.
- Nilai **p** yang kecil menunjukkan bahwa hasil sampel kita sangat tidak mungkin terjadi secara kebetulan jika  $H_0$  benar, sehingga memberikan bukti yang kuat untuk menolak  $H_0$ .

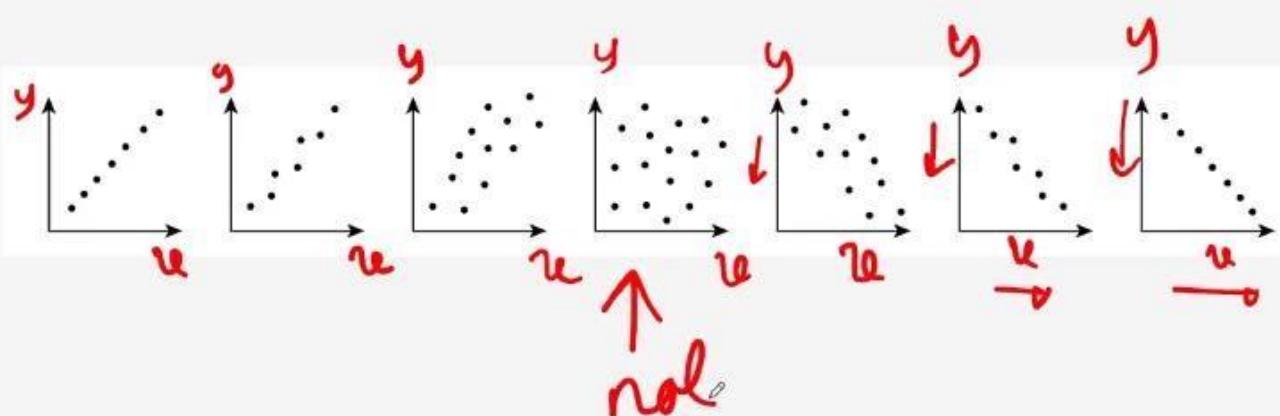
Aturan Keputusan:

- Jika **p**-value  $\leq \alpha$  (tingkat signifikansi): Hasilnya signifikan secara statistik. Kita menolak Hipotesis Nol ( $H_0$ ). Ini berarti ada cukup bukti untuk mendukung Hipotesis Alternatif ( $H_a$ ).
- Jika **p**-value  $> \alpha$ : Hasilnya tidak signifikan secara statistik. Kita tidak menolak Hipotesis Nol ( $H_0$ ). Ini berarti tidak ada cukup bukti dari sampel untuk menolak  $H_0$ .

## Chapter 6. Covariance And Correlation

# Apa Arti Variabel-Variabel Saling Berhubungan?

- Ide Utama: Variabel-variabel dikatakan saling berhubungan ketika ada pola atau tren yang konsisten antara keduanya. Hubungan ini tidak selalu sempurna, tetapi ada kecenderungan yang jelas.
- Contoh: Apakah pengeluaran iklan memengaruhi penjualan? Kita mengamati hubungan dengan melihat apakah penjualan cenderung naik saat pengeluaran iklan meningkat.



## 6-1. Understand Relationship Between Variables

01

**Mengenali pola dalam hubungan data.**



02

**Memahami apa artinya variabel-variabel saling berhubungan.**



03

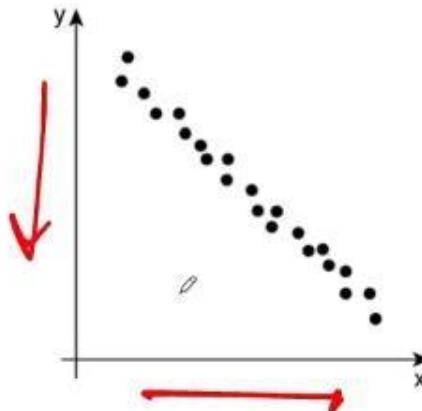
**Menginterpretasi isyarat visual dari korelasi menggunakan scatter plot.**



# Jenis-jenis Hubungan: Positif, Negatif, dan Nol

Hubungan Negatif: Ketika satu variabel meningkat, variabel lain cenderung menurun.

High Degree of Negative Correlation



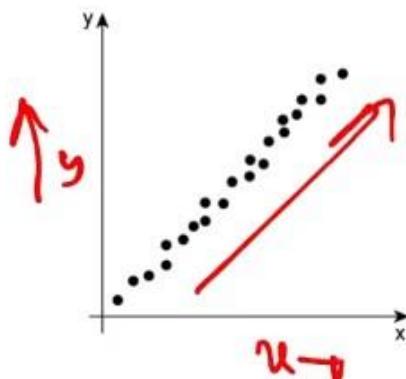
Tidak Ada Hubungan: Perubahan pada satu variabel tidak menunjukkan pola yang konsisten pada variabel lainnya.

## Jenis-jenis Hubungan: Positif, Negatif, dan Nol

Secara visual, kita bisa mengklasifikasikan hubungan antar variabel menjadi tiga jenis utama:

- Hubungan Positif: Ketika satu variabel meningkat, variabel lain juga cenderung meningkat. Garis trennya akan miring ke atas.
- Contoh: Semakin banyak jam belajar, semakin tinggi nilai ujian.

High Degree of Positive Correlation



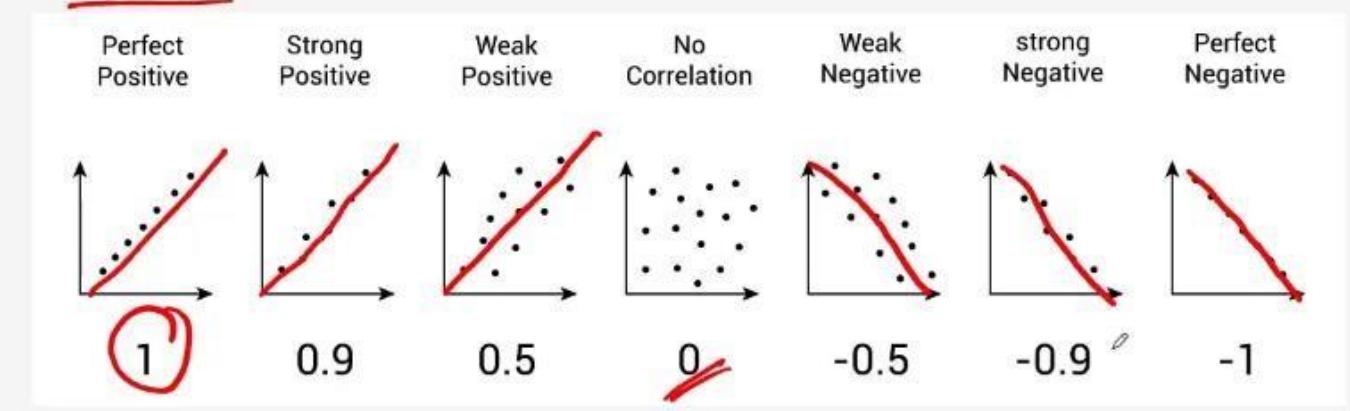
# Memvisualisasikan Hubungan: Scatter Plot

faikamodem00-2625220

Scatter plot adalah alat visual terbaik untuk mendeteksi hubungan antara dua variabel kuantitatif.

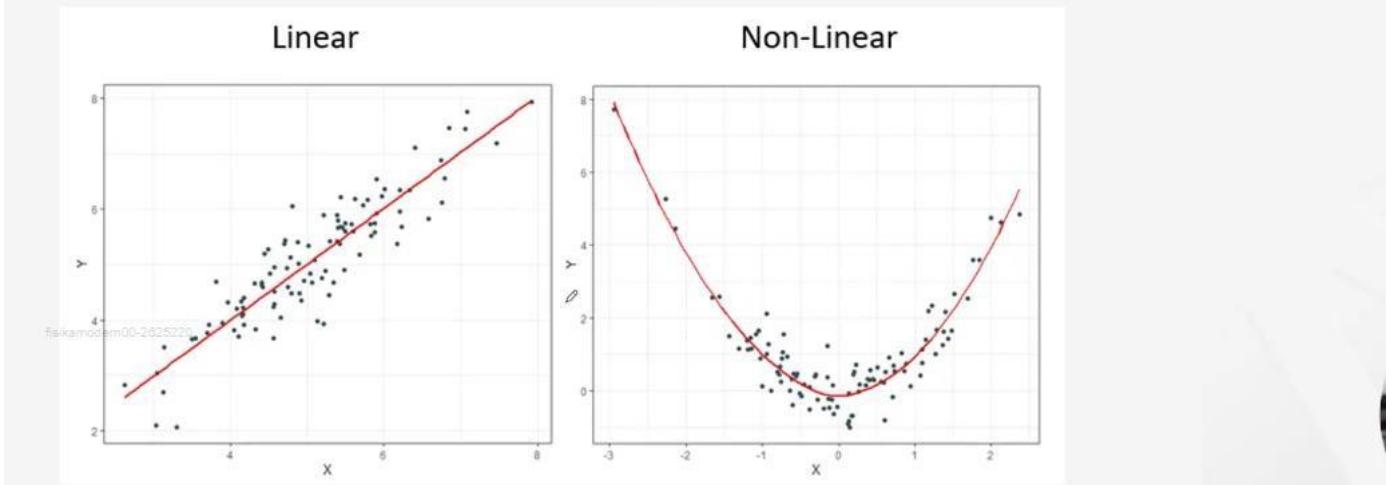
## Cara Membaca Scatter Plot:

- Pola: Perhatikan apakah titik-titik data membentuk pola garis lurus (linear) atau melengkung (non-linear).
- Arah: Arah pola (naik atau turun) menunjukkan jenis hubungan (positif atau negatif).
- Kekuatan: Seberapa rapat titik-titik data di sekitar pola menunjukkan kekuatan hubungan.



## Mengidentifikasi Hubungan Non-Linear

- Tidak semua hubungan adalah garis lurus. Terkadang, hubungan bisa lebih kompleks.
- Contoh: Hubungan antara suhu dan konsumsi es krim mungkin meningkat hingga batas tertentu, lalu menurun. Hubungan ini tidak dapat ditangkap dengan korelasi linear sederhana.
- Penting: Visualisasi sangat krusial untuk mendeteksi jenis hubungan ini sebelum melakukan analisis statistik formal.



## **Ringkasan & Next Up**

- Memahami hubungan antar variabel adalah langkah pertama dalam analisis data yang lebih mendalam.
- Visualisasi, terutama scatter plot, adalah alat yang sangat efektif untuk mengeksplorasi hubungan ini.
- Meskipun visualisasi memberikan petunjuk, kita memerlukan metrik kuantitatif untuk mengukur kekuatan dan arah hubungan ini secara akurat.
- Next Up: Di sub-bab berikutnya, kita akan mempelajari metrik kuantitatif pertama untuk mengukur hubungan: Covariance (Kovarians).

### **6-2. What Is Covariance**

**01**

**Menghitung kovarians antara dua variabel.**

**02**

**Memahami apa yang diinformasikan oleh kovarians (dan apa yang tidak).**

**03**

**Mengidentifikasi keterbatasan kovarians sebagai metrik.**

## Definisi dan Rumus Kovarians

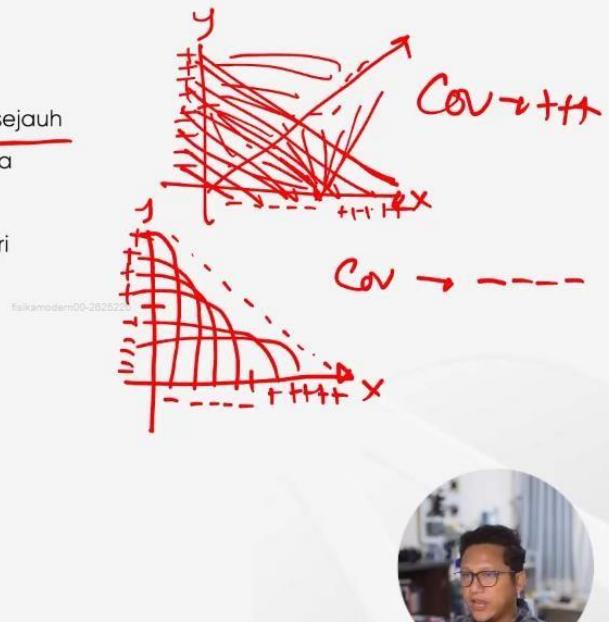
- Ide Utama: Kovarians adalah metrik statistik yang mengukur sejauh mana dua variabel berubah bersama-sama. Ini membantu kita memahami arah hubungan antara dua variabel.
- Bagaimana cara kerjanya? Kovarians menghitung produk dari perbedaan setiap titik data dari rata-ratanya sendiri.
- Rumus Kovarians Sampel:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})(Y_i - \bar{y})$$

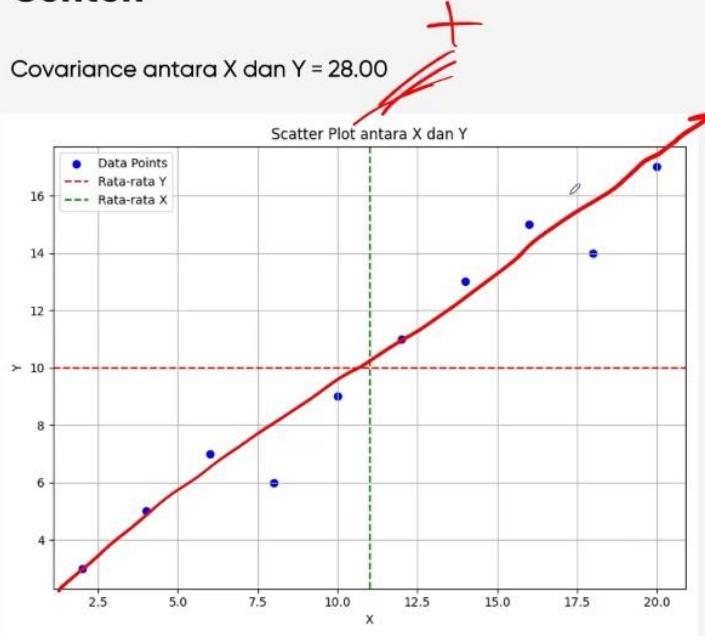
$X_i, Y_i$  = Titik data individual

$\bar{x}, \bar{y}$  = Rata-rata sampel dari  $X, Y$

$n$  = Ukuran Sampel



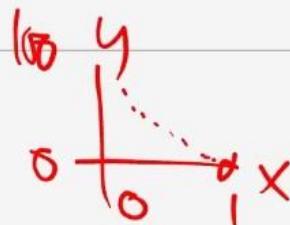
## Contoh



### Penjelasan Visualisasi:

- Titik-titik biru menunjukkan pasangan data ( $X, Y$ ).
- Garis hijau menunjukkan rata-rata X.
- Garis merah menunjukkan rata-rata Y.
- Pola titik menunjukkan bahwa saat X meningkat, Y juga cenderung meningkat  $\rightarrow$  covariance positif.





## Kekurangan Covariance

1. Tidak Terstandarisasi:
  - Covariance bergantung pada unit dan skala variabel.
  - Tidak bisa digunakan untuk membandingkan kekuatan hubungan antar pasangan variabel yang berbeda.
2. Tidak Menunjukkan Kekuatan Hubungan:
  - Covariance hanya menunjukkan arah hubungan, bukan seberapa kuat hubungan tersebut.
3. Sensitif terhadap Outlier:
  - Nilai covariance bisa sangat dipengaruhi oleh outlier karena melibatkan selisih dari rata-rata.
4. Tidak Menunjukkan Kausalitas:
  - Covariance hanya menunjukkan hubungan linier, bukan sebab-akibat.

## Ringkasan

- Kovarians mengukur arah hubungan antara dua variabel.
- Tanda kovarians (+ atau -) menunjukkan apakah hubungan itu positif atau negatif.
- Besaran kovarians tidak mudah diinterpretasi karena tidak distandardisasi.
- Keterbatasan utama: Kovarians hanya efektif untuk hubungan linear dan sulit digunakan untuk perbandingan.
- Next Up: Keterbatasan kovarians mengarah pada kebutuhan akan metrik yang lebih baik **Korelasi**. Di sub-bab berikutnya, kita akan membahas Koefisien Korelasi Pearson dan Spearman.

01

## Menghitung dan menginterpretasi Korelasi Pearson dan Korelasi Rank Spearman.

02

## Memilih metode korelasi yang tepat berdasarkan jenis dan distribusi data.

### Mengapa Kita Membutuhkan Korelasi?

- Keterbatasan Kovarians: Seperti yang kita pelajari, kovarians hanya memberikan arah hubungan (+ atau -) tetapi besarnya sulit diinterpretasi karena tidak dandardisasi.
- Solusi: Korelasi: Korelasi adalah versi kovarians yang dandardisasi. Nilainya selalu berada dalam rentang  $[-1, 1]$ . Ini membuatnya jauh lebih mudah untuk diinterpretasi dan dibandingkan antar pasangan variabel.
- Korelasi adalah metrik yang paling umum digunakan untuk mengukur kekuatan dan arah hubungan linear.

#### Koefisien Korelasi Pearson ( $r$ )

 $r$ -value

Ide Utama: Koefisien Korelasi Pearson ( $r$ ) mengukur kekuatan dan arah hubungan linear antara dua variabel.

Rumus (secara konseptual):

$$r = \frac{Cov(X, Y)}{s_x s_y}$$

Cov( $X, Y$ ): Kovarians antara  $X$  dan  $Y$   
 Standar deviasi  $X$  dan  $Y$

Interpretasi Nilai:

$r = 1$ : Hubungan positif linear sempurna.

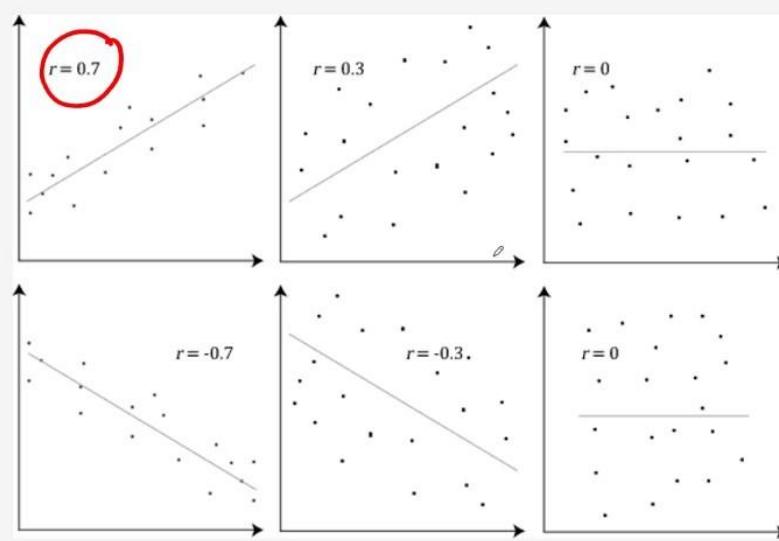
$r = -1$ : Hubungan negatif linear sempurna.

$r = 0$ : Tidak ada hubungan linear.

Semakin dekat nilai  $r$  ke 1 atau -1, semakin kuat hubungan linearinya.

$\overline{0,99}$   
 $-0,92$   
 $0,98$

## Koefisien Korelasi Pearson ( $r$ )



Interpretasi Nilai:

- $r = 1$ : Hubungan positif linear sempurna.
  - $r = -1$ : Hubungan negatif linear sempurna.
  - $r = 0$ : Tidak ada hubungan linear.
- Semakin dekat nilai  $r$  ke 1 atau -1, semakin kuat hubungan linearnya.



## Kapan Menggunakan Korelasi Pearson?

Asumsi Kunci: Korelasi Pearson hanya valid jika memenuhi beberapa asumsi:

1. Hubungan Linear: Hubungan antara variabel harus linear (membentuk garis lurus). ✓
2. Variabel Kontinu: Kedua variabel harus berskala interval atau rasio (data kuantitatif). ✓
3. Distribusi Normal: Data harus berdistribusi normal (atau mendekati normal). ✓

Contoh: Mengukur hubungan antara tinggi badan dan berat badan. Ini adalah hubungan linear, dan kedua variabelnya kontinu.

180,5 cm      96,1 kg  
175,2 cm      85,2 kg

## Korelasi Rank Spearman ( $\rho$ )

- Ide Utama: Korelasi Rank Spearman ( $\rho$ ) mengukur kekuatan dan arah hubungan monotonik antara dua variabel. Ini bekerja dengan mengubah data mentah menjadi peringkat (ranks).

- Apa itu Hubungan Monotonik?

- Ketika satu variabel meningkat, variabel lain cenderung meningkat (atau menurun) secara konsisten, tetapi tidak harus dalam pola garis lurus.
- Interpretasi Nilai: Sama seperti Pearson, nilainya berada dalam rentang  $[-1, 1]$ .

- $\rho = 1$ : Hubungan monotonik sempurna.
- $\rho = -1$ : Hubungan monotonik sempurna.
- $\rho = 0$ : Tidak ada hubungan monotonik.

$$\begin{array}{r} 4 \rightarrow 1 \rightarrow 1 \\ \hline 2 \rightarrow 2 \\ \hline 3 \rightarrow 3 \\ \hline 1 \rightarrow 1 \\ \hline 5 \rightarrow 5 \\ \hline \end{array}$$

$$\begin{array}{r} 1 \rightarrow 1 \\ \hline 2 \rightarrow 2 \\ \hline 5 \rightarrow 5 \\ \hline 3 \rightarrow 3 \\ \hline 1 \rightarrow 1 \\ \hline \end{array}$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Di mana:

- $d_i$  = selisih peringkat antara  $X_i$  dan  $Y_i$
- $n$  = jumlah observasi



## Kapan Menggunakan Korelasi Spearman?

- Fleksibilitas: Korelasi Spearman jauh lebih fleksibel daripada Pearson. Gunakan Spearman ketika:
- Asumsi Linearitas Gagal: Hubungan antara variabel non-linear tetapi masih monotonik.
- Variabel Ordinal: Salah satu atau kedua variabel berskala ordinal (peringkat, misalnya: sangat setuju, setuju, netral).
- Adanya Outlier: Korelasi Spearman kurang sensitif terhadap outlier.
- Contoh: Mengukur hubungan antara peringkat nilai tes (peringkat 1, 2, 3, ...) dan peringkat jam belajar.

## Contoh Perhitungan Pearson



Data :

$$X = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

$$Y = \{15, 25, 35, 45, 55, 65, 75, 85, 95, 105\}$$

Rata-rata:

$$\bar{X} = 55$$

$$\bar{Y} = 60$$

Hitung Covariance:

$$\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{8250}{9} = 916.67$$

Simpangan baku:

$$\sigma_X = \sigma_Y = 30.28$$

Pearson  $r$ :

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} = \frac{916.67}{30.28 \times 30.28} = 1$$



## Contoh Perhitungan Spearman

Data :

$$X = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

$$Y = \{15, 25, 35, 45, 55, 65, 75, 85, 95, 105\}$$

Ranks (berikan ranking sesuai urutan):

$$\text{Rank} = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

Hitung Selisih Ranking:

$$d_i = \text{Rank}(X_i) - \text{Rank}(Y_i)$$

Hitung Spearman  $\rho$ :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - 0 = 1$$

↑  
000000  
↓  
10 (99)

X	Y	X_rank	Y_rank	d	$d^2$
10	15	1.0	1.0	0.0	0.0
20	25	2.0	2.0	0.0	0.0
...	...	...	...	...	...
100	105	10.0	10.0	0.0	0.0

$\rho_{X-Y}$

$2d^2$



## Ringkasan & Perbandingan

Kriteria	Korelasi Pearson ( $r$ )	Korelasi Spearman ( $\rho$ )
Mengukur	Hubungan Linear	Hubungan Monotonik
Jenis Data	Kontinu (Rasio/Interval)	Kontinu atau Ordinal
Sensitivitas	Sangat Sensitif	Kurang Sensitif
Outlier		
Kondisi	Memerlukan asumsi normalitas dan linearitas	Lebih Fleksibel, tidak memerlukan asumsi normalitas
Contoh penggunaan	Hubungan tinggi-berat	Hubungan peringkat nilai-peringkat jam belajar

### 6-4. Correlation VS Causation

01

**Memahami batasan analisis korelasi.**



02

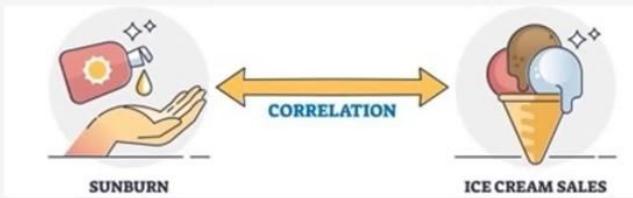
**Mengidentifikasi mengapa korelasi tidak sama dengan kausalitas.**

03

**Mengenali variabel perancu (confounding variables) dan korelasi palsu (spurious correlations).**

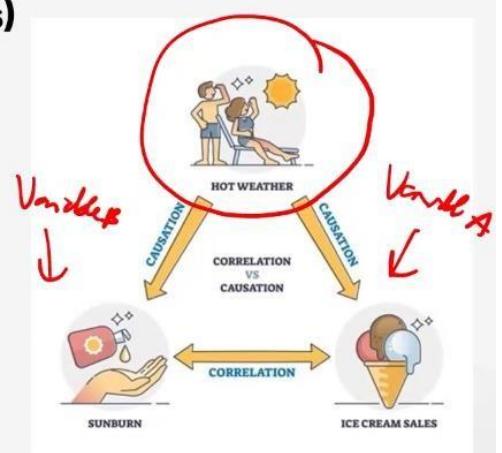
## Mengapa Korelasi Bukan Kausalitas?

- Ide Utama: Korelasi hanya mengukur kekuatan dan arah hubungan antara dua variabel. Korelasi TIDAK membuktikan bahwa satu variabel menyebabkan perubahan pada variabel lainnya.
- Contoh Klasik:
  - Ada korelasi positif yang kuat antara penjualan es krim dan kasus sunburn di musim panas.
  - Apakah penjualan es krim menyebabkan orang sunburn? Tentu tidak.
  - Ini adalah contoh di mana dua variabel berkorelasi, tetapi tidak ada hubungan sebab-akibat langsung di antara mereka.



## Masalah Variabel Perancu (Confounding Variables)

- Variabel Perancu: Ini adalah variabel tersembunyi atau tidak terukur yang memengaruhi kedua variabel yang kita amati, sehingga menciptakan korelasi palsu.
- Kembali ke Contoh Es Krim:
  - Variabel A: Penjualan Es Krim
  - Variabel B: Kasus Sunburn
  - Variabel Perancu (C): Suhu Udara di Musim Panas
  - Hubungan yang Sebenarnya: Suhu yang lebih tinggi menyebabkan penjualan es krim meningkat DAN suhu yang lebih tinggi menyebabkan lebih banyak orang Sunburn.
  - Korelasi antara penjualan es krim dan kasus sunburn adalah hasil dari pengaruh variabel perancu ini, bukan kausalitas langsung.



# Korelasi Palsu (Spurious Correlations)

- Korelasi Palsu: Ini adalah korelasi yang terjadi hanya karena kebetulan. Hubungan statistik yang kuat ada, tetapi tidak ada hubungan logis atau kausal yang masuk akal di antara variabel-variabel tersebut.
- Contoh:
  - Ada korelasi kuat antara jumlah film yang dibintangi Nicolas Cage dan jumlah kasus tenggelam di kolam renang.
  - Secara statistik, korelasi ini mungkin signifikan, tetapi secara logis, tidak ada alasan sama sekali untuk percaya bahwa satu menyebabkan yang lain. Ini adalah murni kebetulan.
- Penting: Selalu gunakan akal sehat dan pengetahuan domain Anda saat menginterpretasi korelasi.

## Cara Membedakan Korelasi dan Kausalitas

Korelasi	Kausalitas
Dapat diukur dengan Koefisien Pearson atau Spearman. $r$	Tidak dapat dibuktikan hanya dengan korelasi.
Menunjukkan seberapa dekat dua variabel bergerak bersama. 	Memerlukan bukti dari eksperimen terkontrol (misalnya, A/B testing) di mana satu variabel dimanipulasi secara sengaja sementara faktor lain dikontrol.
Tidak memerlukan manipulasi variabel.	Ada urutan waktu: penyebab harus mendahului akibat. 

- Kesimpulan: Korelasi adalah langkah pertama yang hebat untuk mengidentifikasi potensi hubungan, tetapi bukan bukti kausalitas.

# **Ringkasan & End of Chapter**

- Korelasi hanya mengukur hubungan, bukan penyebab.
- Selalu waspada terhadap variabel perancu yang dapat menciptakan korelasi palsu.
- Kausalitas hanya dapat dibuktikan melalui desain eksperimental yang ketat, seperti eksperimen terkontrol.
- Penting untuk Data Science: Jangan pernah secara otomatis menyimpulkan bahwa ada hubungan sebab-akibat hanya karena Anda menemukan korelasi yang kuat.
- End of Chapter 6: Dengan ini, kita menyelesaikan Chapter 6 tentang Kovarians dan Korelasi. Anda sekarang memiliki alat untuk mengukur hubungan dan pemahaman kritis tentang batasan alat-alat ini.

## **Chapter 7. Basic Of Regression Analysis**

### **7-1. What Is Regression Analysis**

**01**

**Memahami tujuan utama Analisis Regresi dalam statistik dan data science.**

**02**

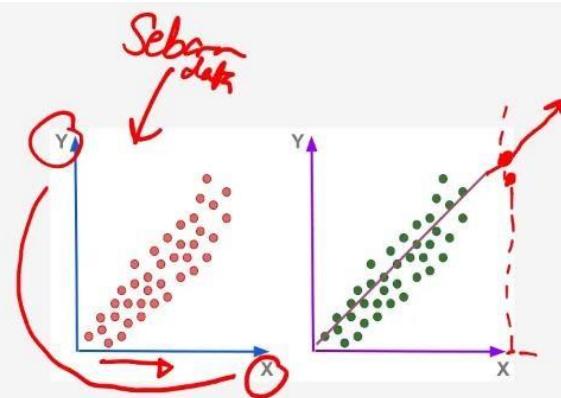
**Mengidentifikasi dan membedakan antara variabel dependen dan variabel independen.**

**03**

**Mengenali aplikasi regresi di dunia nyata.**

# Apa itu Analisis Regresi?

- Tujuan Utama: Analisis regresi adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel dependen (target) dan satu atau lebih variabel independen (prediktor).
- Fungsinya: Analisis regresi melampaui korelasi dengan membangun sebuah persamaan matematika. Tujuan utamanya adalah:
  1. Prediksi: Memperkirakan nilai variabel dependen berdasarkan nilai variabel independen.
  2. Penjelasan: Mengukur kekuatan dan arah hubungan tersebut secara kuantitatif untuk memahami dampak dari setiap variabel independen.



## Variabel Dependen vs. Independen

Memahami peran variabel-variabel ini adalah fondasi dari regresi.

- Variabel Dependen (Dependent Variable - Y):
  - Variabel yang ingin kita prediksi atau jelaskan.
  - Contoh: Harga rumah, penjualan produk, nilai ujian siswa.
- Variabel Independen (Independent Variable - X):
  - Variabel yang kita gunakan untuk memprediksi atau menjelaskan variabel dependen.
  - Contoh: Luas rumah, pengeluaran iklan, jam belajar.
- Analogi Sederhana: Variabel independen adalah "penyebab" atau "faktor" (meskipun tidak selalu kausal), dan variabel dependen adalah "akibat" atau "hasil" yang ingin kita pelajari.



## Aplikasi Nyata Analisis Regresi

Analisis regresi adalah salah satu alat yang paling sering digunakan dalam data science karena kemampuannya untuk menjawab pertanyaan-pertanyaan penting di berbagai bidang.

- Bisnis: Memprediksi penjualan produk di masa depan berdasarkan pengeluaran iklan dan harga.
- Keuangan: Memprediksi risiko gagal bayar pinjaman pelanggan berdasarkan pendapatan dan riwayat kredit mereka.
- Ilmu Pengetahuan: Mengukur hubungan antara dosis obat dan respons pasien untuk menentukan efektivitas.
- Kesehatan: Memprediksi berat badan bayi baru lahir berdasarkan berat badan ibu dan variabel kesehatan lainnya.
- Sosial: Mengestimasi pendapatan tahunan seseorang berdasarkan tingkat pendidikan dan pengalaman kerja.



# Ringkasan & Next Up

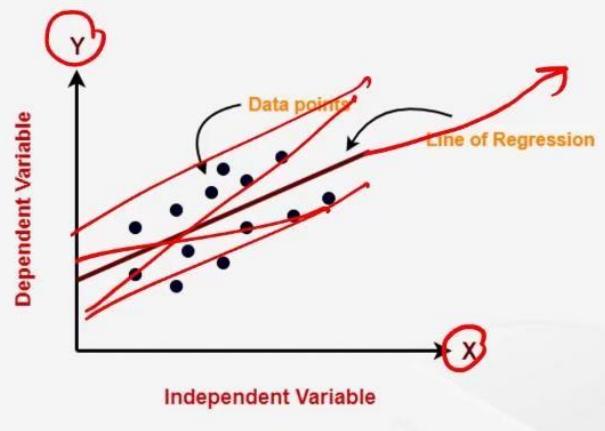
- Analisis Regresi adalah alat pemodelan yang kuat untuk memprediksi dan menjelaskan hubungan antar variabel.
- Ini memungkinkan kita untuk mengukur hubungan secara kuantitatif, melampaui sekadar korelasi.
- Next Up: Di sub-bab berikutnya, kita akan menyelami jenis regresi yang paling dasar: Simple Linear Regression, termasuk konsep dan persamaannya.

## 7-2. Simple Linear Regression - Concept & Equation

- 01 **Menginterpretasi komponen-komponen dari model linear.**
- 02 **Memahami bagaimana garis regresi terbentuk.**
- 03 **Mengidentifikasi peran kemiringan (slope) dan perpotongan (intercept) dalam model.**

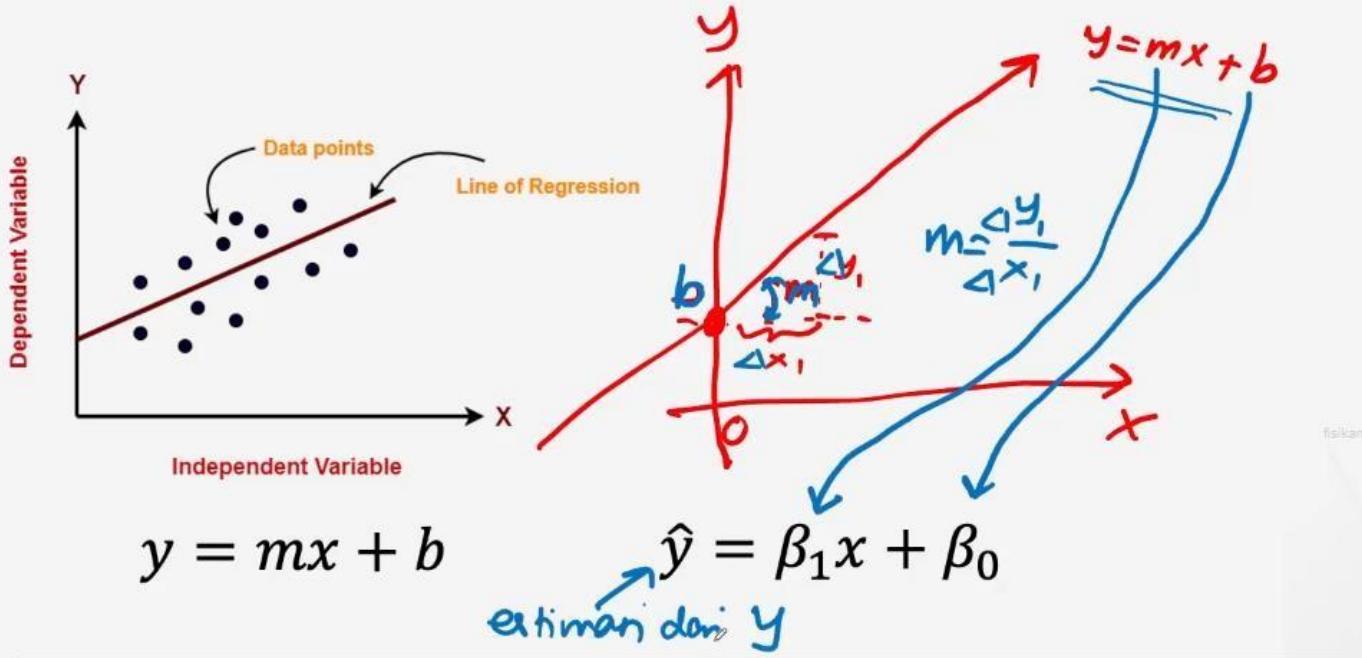
## Apa itu Simple Linear Regression?

- Ide Utama: Regresi Linear Sederhana adalah jenis analisis regresi paling dasar. Ini memodelkan hubungan antara dua variabel kuantitatif: satu variabel dependen ( $Y$ ) dan satu variabel independen ( $X$ ).
- Tujuannya: Untuk menemukan garis lurus yang paling pas (best-fit line) yang menggambarkan hubungan antara variabel  $X$  dan  $Y$ . Garis ini kemudian dapat digunakan untuk membuat prediksi.
- Hubungan Linear: Asumsi utama di sini adalah bahwa hubungan antara  $X$  dan  $Y$  dapat diwakili secara memadai oleh garis lurus.



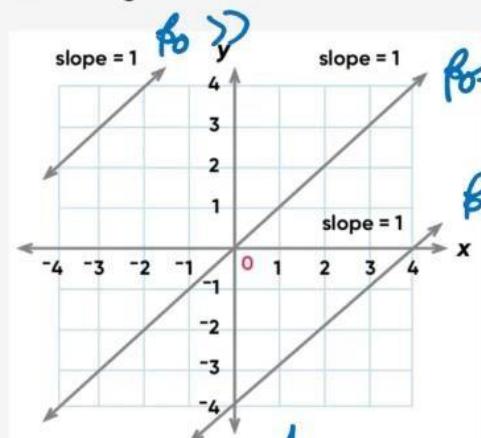
## Persamaan Garis Linear

Analisis regresi linear sederhana didasarkan pada persamaan garis lurus yang sudah familiar.



## Menginterpretasi Intercept ( $\beta_0$ )

Analisis regresi linear sederhana didasarkan pada persamaan garis lurus yang sudah familiar.

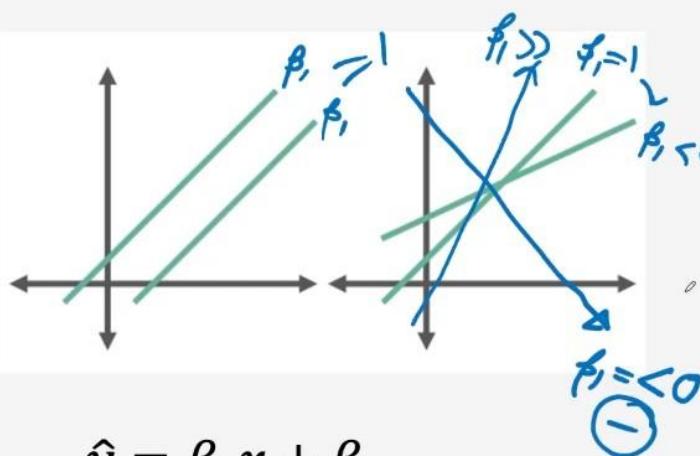


Definisi: Intercept ( $\beta_0$ ) adalah titik di mana garis regresi memotong sumbu Y. Ini adalah nilai yang diprediksi untuk Y ketika nilai X adalah nol.

$$\hat{y} = \beta_1 x + \beta_0$$

## Menginterpretasi Slope ( $\beta_1$ )

Analisis regresi linear sederhana didasarkan pada persamaan garis lurus yang sudah familiar.



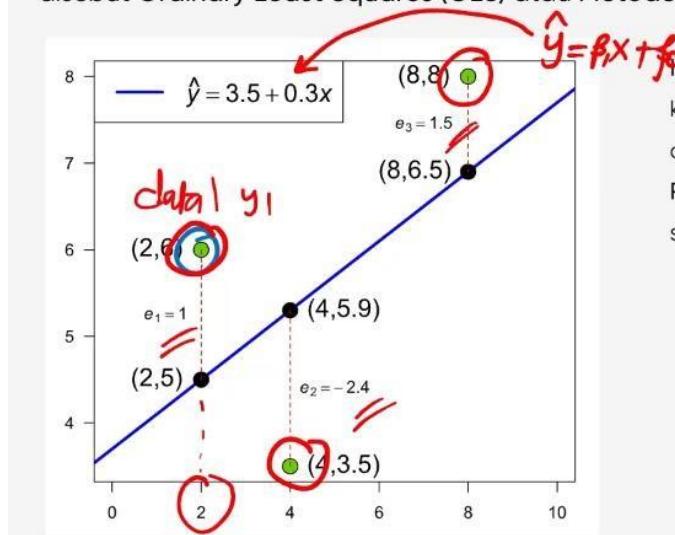
Definisi: Slope ( $\beta_1$ ) mengukur kemiringan garis regresi. Ini adalah nilai yang paling penting dalam analisis regresi.

- Tanda (Sign): Tanda dari  $\beta_1$  menunjukkan arah hubungan.
- Positif ( $\beta_1 > 0$ ): Hubungan positif.
  - Negatif ( $\beta_1 < 0$ ): Hubungan negatif.

$$\hat{y} = \beta_1 x + \beta_0$$

## Garis "Best Fit" (Metode Kuadrat Terkecil)

Bagaimana kita menemukan garis yang paling pas? Analisis regresi menggunakan metode yang disebut Ordinary Least Squares (OLS) atau Metode Kuadrat Terkecil.



$$\hat{y} = \beta_0 + \beta_1 x$$

Penjelasan: Metode ini mencari garis lurus yang meminimalkan jumlah kuadrat dari jarak vertikal (disebut **residual**) antara setiap titik data dan garis regresi.

Residual: Residual adalah perbedaan antara nilai Y yang sebenarnya dan nilai Y yang diprediksi oleh garis regresi.

$$e = y_i - \hat{y}_i$$



## Ringkasan & Next Up

- Regresi Linear Sederhana adalah tentang menemukan garis lurus yang paling pas untuk memodelkan hubungan antara dua variabel.
  - Persamaan regresi adalah  $\hat{y} = \beta_0 + \beta_1 x$ .
  - Intercept ( $\beta_0$ ) adalah nilai Y ketika X=0.
  - Slope ( $\beta_1$ ) adalah perubahan Y per satu unit perubahan X.
  - Garis terbaik ditemukan menggunakan Metode Kuadrat Terkecil, yang meminimalkan residual.
- Intercept*      *Slope*
- Next Up: Setelah memahami konsep dan persamaan, kita akan belajar bagaimana melakukan fitting pada model linear dan menghitung koefisien  $\beta_0$  dan  $\beta_1$  secara praktis.

## 7-3. Fitting A Linear Model (Least Square Method)

01

**Memahami bagaimana parameter model (koefisien) diestimasi.**

02

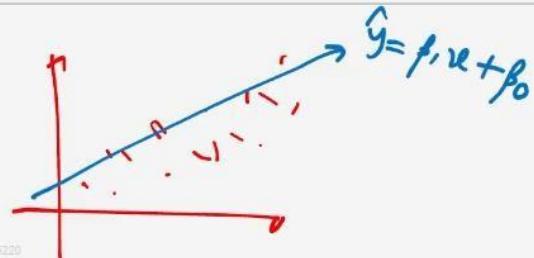
**Menginterpretasi residual dan konsep kesalahan.**

03

**Memahami intuisi di balik Metode Kuadrat Terkecil.**

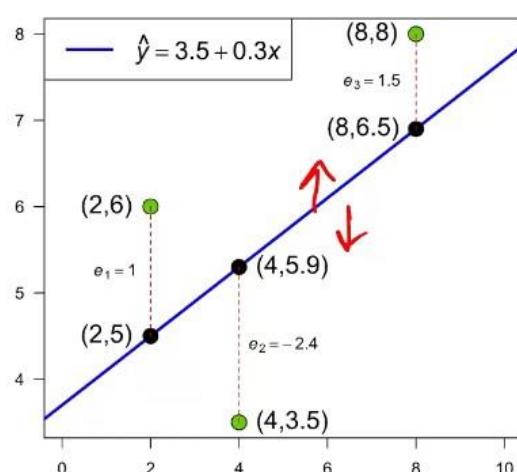
### Mengapa "Fitting" itu Penting?

- Ide Utama: Setelah kita tahu persamaannya adalah  $\hat{y} = \beta_0 + \beta_1 x$ , pertanyaan berikutnya adalah: bagaimana kita menemukan nilai terbaik untuk  $\beta_0$  dan  $\beta_1$  dari data kita?
- Tujuan: Proses "fitting" adalah proses menemukan nilai koefisien ini yang menghasilkan garis "best fit".
- Metode yang paling umum dan intuitif untuk melakukan ini adalah Metode Kuadrat Terkecil (Least Squares Method) atau yang juga dikenal sebagai Ordinary Least Squares (OLS).



### Konsep Residuals dan Error Terms

Residual (Kesalahan): Residual adalah perbedaan antara nilai aktual dari variabel dependen ( $y$ ) dan nilai yang diprediksi oleh model kita ( $\hat{y}$ ).



Residual: Residual adalah perbedaan antara nilai Y yang sebenarnya dan nilai Y yang diprediksi oleh garis regresi.

$$e = y_i - \hat{y}_i \rightarrow e_1 + e_2 + e_3 + e_4 + e_5 = 0.1$$

Tujuan OLS: OLS menemukan garis yang meminimalkan Jumlah Kuadrat Kesalahan (Sum of Squared Errors - SSE).

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= 1^2 + (-2.4)^2 + (2.5)^2 \\ &= 1 + 5.76 + 6.25 \end{aligned}$$

# Ringkasan Perhitungan

Nilai koefisien regresi ( $\beta_0$  dan  $\beta_1$ ) yang meminimalkan SSE dihitung menggunakan kalkulus.

Slope ( $\beta_1$ ):

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Intercept ( $\beta_0$ ):

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Annotations:

- Handwritten notes: "data", "rata-rata", "Covarians (X,Y)", "Cov(X,Y) / Var(X)", "rata-rata<sup>2</sup>X", "rata-rata<sup>2</sup>" are written around the equations.
- Small text at the bottom left: "Salakmoderm00-2625220"

## Ringkasan & Next Up

- Fitting model regresi adalah proses menemukan koefisien yang paling cocok.
- Metode Kuadrat Terkecil (OLS) mencapai ini dengan meminimalkan jumlah kuadrat residual.
- Residuals adalah kesalahan atau jarak vertikal antara nilai data aktual dan nilai yang diprediksi.
- OLS memberikan koefisien yang optimal, memungkinkan kita untuk membuat prediksi terbaik.

01

**Membangun dan menginterpretasi model regresi linear berganda.**

02

**Mengenali kapan regresi berganda sesuai untuk digunakan.**

03

**Memahami konsep dasar multikolinearitas.**

### Dari Sederhana ke Berganda

- Regresi Linear Sederhana: Memodelkan hubungan antara satu variabel dependen ( $Y$ ) dan satu variabel independen ( $X$ ).
- Regresi Linear Berganda (Multiple Linear Regression):  
Memodelkan hubungan antara satu variabel dependen ( $Y$ ) dan dua atau lebih variabel independen ( $X_1, X_2, \dots, X_p$ ).
- Mengapa Menggunakan Regresi Berganda?
  - Dunia nyata jarang sesederhana satu variabel.
  - Regresi berganda memungkinkan kita untuk memperhitungkan pengaruh gabungan dari banyak faktor terhadap variabel target. ( $Y$ )

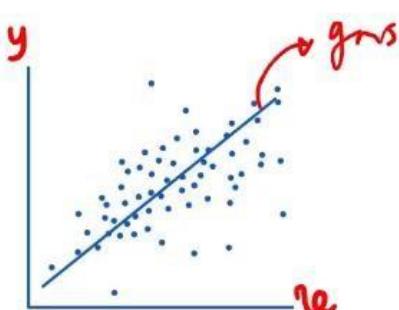
## Persamaan Regresi Berganda

- Persamaan regresi berganda adalah perpanjangan dari persamaan sederhana.
- Persamaan:

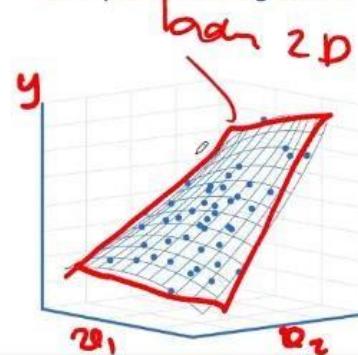
$$\hat{y} = \beta_0 + \beta_1 u_1 + \beta_2 u_2 \dots$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = [\beta_0 \ \beta_1 \ \beta_2 \dots \ \beta_p] \begin{bmatrix} 1 \\ u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}$$

Simple Linear Regression



Multiple Linear Regression



## Menginterpretasi Koefisien

- Penting: Interpretasi koefisien di regresi berganda sedikit berbeda dari regresi sederhana.
- Contoh: Dalam model yang memprediksi harga rumah ( $Y$ ) berdasarkan luas rumah ( $X_1$ ) dan jumlah kamar tidur ( $X_2$ ), jika  $\beta_1 = 50$ :
  - Interpretasinya bukan "untuk setiap 1 unit tambahan luas rumah, harga naik \$50."
  - Interpretasi yang benar adalah: "untuk setiap 1 unit tambahan luas rumah, harga diprediksi meningkat \$50, dengan jumlah kamar tidur tetap konstan."
- Frasa "memegang variabel lain tetap konstan" adalah kunci. Ini memungkinkan kita untuk mengisolasi pengaruh setiap prediktor.

$$y = \beta_1 u_1$$
$$y = \beta_0 + \beta_1 u_1 + \beta_2 u_2$$

# Menghitung Koefisien Regresi Berganda

- Ide Utama: Menemukan koefisien ( $\beta$ ) untuk regresi berganda adalah masalah optimasi yang diselesaikan dengan meminimalkan jumlah kuadrat residual (seperti dalam regresi sederhana).  $SSE = \sum (y - \hat{y})^2$
- Solusi Matriks: Secara matematis, solusi ini paling efisien ditemukan menggunakan aljabar linear dan notasi matriks. Persamaan untuk menghitung vektor koefisien  $\hat{\beta}$  adalah:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Penjelasan Komponen:

- $\hat{\beta}$ : Vektor koefisien yang berisi  $\beta_0, \beta_1, \dots, \beta_p$ .
- $Y$ : Vektor dari nilai-nilai variabel dependen ( $y$ ).
- $X$ : Matriks desain yang berisi nilai-nilai dari semua variabel independen, dengan kolom pertama berisi angka 1 untuk *intercept*.
- $X^T$ : Transpose dari matriks  $X$ .
- $(\ )^{-1}$ : Inverse dari matriks.

## Asumsi Utama Regresi Linear Berganda

Seperti halnya regresi sederhana, regresi berganda juga memiliki asumsi yang harus dipenuhi untuk hasil yang valid.

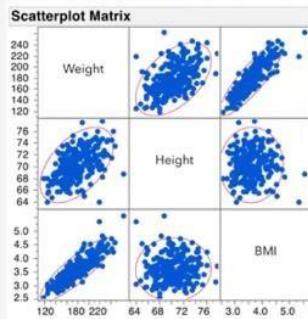
- Linearitas:** Hubungan antara variabel independen dan dependen harus linear.  $u_1 - y$   $u_2 - y$
- Independensi Residual:** Residual tidak berhubungan satu sama lain.
- Homoskedastisitas:** Variabilitas kesalahan prediksi harus seragam di seluruh rentang nilai.
- Normalitas Residual:** Residual harus berdistribusi normal.
- Multikolinearitas Rendah:** Tidak ada variabel independen yang sangat berkorelasi kuat dengan variabel independen lainnya.  $u_1$   $u_2$

## Multikolinearitas: Masalah dan Solusi

$$Y = \beta_0 + \beta_1 V_1 + \beta_2 V_2$$

Multivariate Correlations	$V_2$	$V_1$	$y$
Weight	1.0000	0.5129	0.8668
Height	0.5129	1.0000	0.0220
BMI	0.8668	0.0220	1.0000

- Apa itu Multikolinearitas? Ini adalah masalah di mana dua atau lebih variabel independen dalam model sangat berkorelasi satu sama lain.
- Mengapa Itu Masalah?
  - Multikolinearitas membuat koefisien regresi menjadi tidak stabil dan sulit diinterpretasi karena model tidak dapat mengisolasi efek unik dari setiap variabel independen yang berkorelasi tinggi. Hal ini juga menggembungkan kesalahan standar, membuat uji hipotesis tidak dapat diandalkan.
- Cara Mengidentifikasi: Periksa matriks korelasi antar variabel independen.
- Solusi: Hapus salah satu variabel yang berkorelasi tinggi atau gabungkan menjadi satu variabel.



## Ringkasan & Next Up

- Regrasi Linear Berganda memperluas regresi sederhana dengan menggunakan beberapa prediktor.  $V_1, V_2, \dots$
- Persamaannya memungkinkan kita untuk memprediksi dan memahami pengaruh gabungan dari banyak variabel.
- Interpretasi koefisien harus selalu memperhitungkan efek "memegang variabel lain tetap konstan".
- Multikolinearitas adalah masalah umum di mana prediktor berkorelasi, membuat model sulit diinterpretasi.

$V_1 V_2$

01

**Mengidentifikasi asumsi-asumsi kunci dari regresi linear.**

02

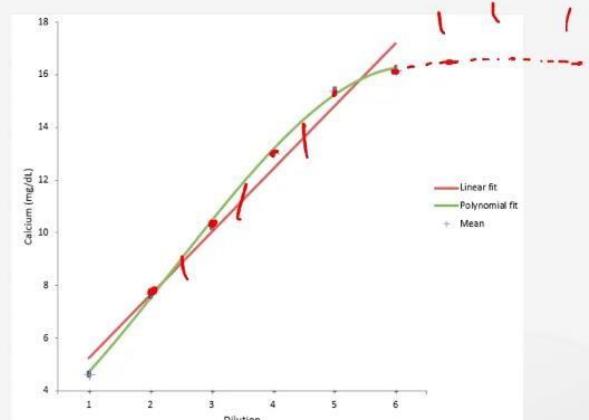
**Menggunakan plot diagnostik untuk menilai validitas model.**

### Mengapa Asumsi Itu Penting?

- Ide Utama: Asumsi regresi linear adalah kondisi yang harus dipenuhi oleh data kita agar model regresi dapat dipercaya dan inferensi statistik (seperti uji hipotesis dan interval kepercayaan) menjadi valid. Jika asumsi ini dilanggar, kesimpulan dari model kita bisa salah atau menyesatkan.
- Analogi Sederhana: Menggunakan model regresi tanpa memeriksa asumsinya sama seperti membangun rumah di atas fondasi yang tidak stabil.
- Ada empat asumsi utama yang perlu kita periksa.

#### Asumsi 1: Linearitas

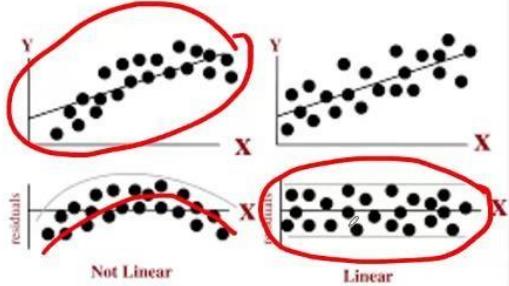
- Konsep: Hubungan antara variabel dependen (Y) dan variabel independen (X) harus linear (garis lurus). Jika hubungan tersebut melengkung, model linear tidak akan dapat menangkap pola yang sebenarnya.
- Bagaimana Cara Memeriksanya?
  - o Buat scatter plot dari variabel dependen terhadap setiap variabel independen.
  - o Periksa apakah titik-titik data tampak mengikuti pola garis lurus.
- Visualisasi: Jika Anda melihat pola yang jelas melengkung (misalnya, bentuk "U" atau "J terbalik"), asumsi linearitas mungkin dilanggar.



## Asumsi 2: Independensi Residual

- Konsep: Residual (kesalahan) model harus independen satu sama lain. Artinya, residual untuk satu titik data tidak boleh memengaruhi atau dipengaruhi oleh residual untuk titik data lainnya.
- Bagaimana Cara Memeriksanya?
  - Plot residual terhadap urutan data atau variabel waktu.
  - Periksa apakah ada pola atau tren yang jelas. Jika residual tinggi diikuti oleh residual tinggi, ini adalah pelanggaran.
- Visualisasi: Plot residual vs. waktu/urutan harus terlihat seperti "kabut" acak tanpa pola yang jelas. Adanya pola (misalnya, gelombang) menunjukkan pelanggaran asumsi.

Residual Analysis for Linearity

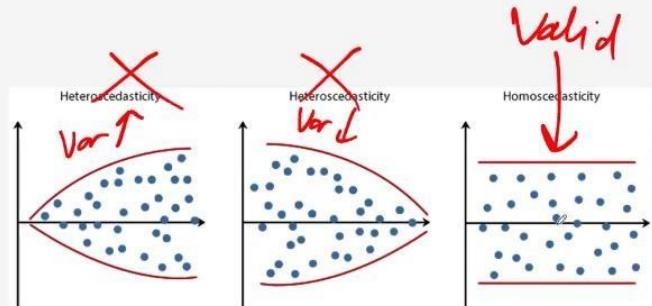


risikamodem00-2020520



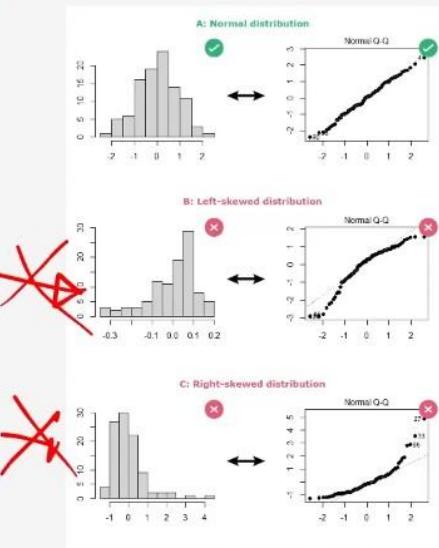
## Asumsi 3: Homoskedastisitas

- Konsep: Varians dari residual harus konstan di seluruh rentang nilai variabel independen. Dengan kata lain, kesalahan prediksi harus konsisten, tidak membesar atau mengecil seiring bertambahnya nilai X.
- Bagaimana Cara Memeriksanya?
  - Buat plot residual vs. nilai prediksi (fitted values).
  - Periksa apakah titik-titik data tersebar secara acak di sekitar sumbu horizontal (homoskedastisitas).
- Visualisasi:
  - Homoskedastisitas (Benar): Terlihat seperti "kabut" acak.
  - Heteroskedastisitas (Salah): Terlihat seperti corong, di mana residual membesar seiring bertambahnya nilai prediksi.



## Asumsi 4: Normalitas Residual

- Konsep: Residual harus berdistribusi normal. Ini adalah asumsi penting untuk uji hipotesis dan interval kepercayaan.
- Bagaimana Cara Memeriksanya?
  - Buat **histogram** dari residual. Periksa apakah bentuknya menyerupai kurva lonceng.
  - Buat Q-Q Plot (Quantile-Quantile Plot) dari residual. Plot ini membandingkan residual kita dengan distribusi normal teoritis. Jika titik-titik data membentuk garis lurus, asumsi terpenuhi.
- Visualisasi:
  - Histogram:** Cari bentuk kurva lonceng yang simetris.
  - Q-Q Plot:** Cari garis lurus. Penyimpangan dari garis lurus (misalnya, kurva di ujungnya) menunjukkan pelanggaran normalitas.



## Ringkasan Asumsi dan Plot Diagnostik

Asumsi	Apa yang dikatakan	Cara memeriksanya (Visual)
Linearitas	Hubungan X dan Y Adalah garis lurus	Scatter Plot: Lihat pola linier
Independensi	Residual tidak berkorelasi	Plot Residual vs Waktu: cari kabut acak
Homoskedastisitas	Varians residual konstan.	Plot Residual vs Fitted: Cari kabut acak, bukan bentuk corong
Normalitas	Residual Berdistribusi Normal	Histogram atau Q-Q Plot: Cari bentuk lonceng/garis lurus

Data  
tidak ada pola  
o Residual

- 01 **Membaca dan menginterpretasi hasil regresi dari perangkat lunak statistik.**
- 02 **Membuat kesimpulan yang tepat dari output model.**

### Membaca Laporan Model

- Ide Utama: Setelah Anda menjalankan regresi (misalnya, di Python, R, atau Excel), perangkat lunak akan menghasilkan ringkasan yang berisi semua informasi kunci tentang model Anda. p-value,  $R^2$ ,  $\beta_1$ ,  $\beta_0$
- Memahami laporan ini adalah keterampilan praktis yang penting. Ini memungkinkan Anda untuk:
  - Mengukur seberapa baik model Anda cocok dengan data.
  - Menentukan variabel mana yang signifikan secara statistik.
  - Menarik kesimpulan yang andal tentang hubungan antar variabel.

# Komponen Output Utama

Ada beberapa bagian penting dari output regresi yang perlu Anda perhatikan:

1. Koefisien ( $\beta_0, \beta_1, \dots$ ): Nilai yang diestimasi untuk intercept dan slope.
2. Statistik Uji Koefisien (t-statistic, p-value): Digunakan untuk menentukan apakah setiap koefisien signifikan secara statistik.
3. R-squared ( $R^2$ ): Metrik untuk mengukur seberapa baik model cocok dengan data.
4. Standar Error dan Statistik Model Lainnya: Memberikan informasi tentang presisi estimasi.

Essential Mathematical Foundation – Basic of Regression Analysis – **Interpreting Regression Output**

## Koefisien, Standard Error, dan p-value

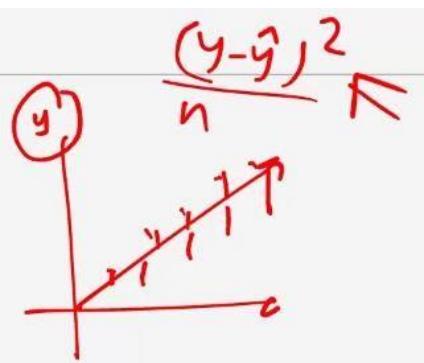
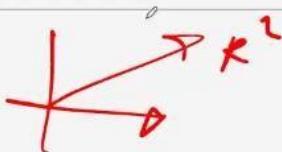
Fitur	Penjelasan	Contoh Interpretasi
Koefisien	Nilai yang diestimasi untuk <u>intercept</u> dan <u>slope</u> . Ini adalah angka yang kita masukkan ke dalam persamaan regresi	$\beta_1 = 2.5$ berarti setiap satu unit kenaikan $X_1$ meningkatkan $Y$ sebesar 2.5
Standard Error	Mengukur presisi koefisien. Semakin kecil nilai SE, semakin presisi estimasi koefisien	SE yang kecil menunjukkan bahwa koefisien tersebut adalah perkiraan yang andal.
Nilai p <i>↓ Y<sub>i</sub> hipotesis</i>	Probabilitas untuk mengamati koefisien sebesar itu (Atau lebih ekstrim) jika nilai koefisien yang sebenarnya adalah nol (yaitu, tidak ada hubungan)	Nilai p kecil ( $\leq \alpha$ ) menunjukkan bahwa koefisien tersebut signifikan secara statistik dan kemungkinan besar bukan nol

Aturan Praktis: Jika nilai p koefisien  $\leq 0.05$  (untuk  $\alpha=0.05$ ), kita menyimpulkan bahwa variabel independen tersebut memiliki hubungan yang signifikan dengan variabel dependen.

$$\begin{aligned} & \text{Koefisien} \\ & y = 2,5x \\ & y = 2,5 \\ & y = 5 + 2,5x \end{aligned}$$



## R-Squared ( $R^2$ )



- Apa itu?  $R^2$  (R-kuadrat) adalah metrik yang mengukur proporsi varians dalam variabel dependen yang dapat dijelaskan oleh model regresi kita.
- Nilai: Nilainya berkisar antara 0 dan 1.
  - $R^2 = 0$ : Model tidak menjelaskan varians sama sekali.
  - $R^2 = 1$ : Model menjelaskan semua varians dengan sempurna.
- Interpretasi: Jika  $R^2 = 0.75$ , ini berarti model regresi Anda menjelaskan 75% variasi dalam variabel dependen. Semakin tinggi  $R^2$ , semakin baik model Anda.
- Peringatan:  $R^2$  bisa menyesatkan karena nilainya selalu meningkat saat Anda menambahkan lebih banyak variabel independen ke model, bahkan jika variabel tersebut tidak signifikan.

## Signifikansi Statistik vs. Signifikansi Praktis

$y=0.05 \uparrow$

- Signifikansi Statistik: Dinyatakan oleh nilai  $p$ . Ini hanya memberi tahu kita apakah hubungan yang diamati kemungkinan besar bukan karena kebetulan.
  - Contoh: Nilai  $p < 0.05$ . ↗ signifikan  $\alpha$
- Signifikansi Praktis: Dinyatakan oleh ukuran koefisien dan konteks masalah. Ini memberi tahu kita apakah besaran efek yang diamati cukup penting di dunia nyata.
  - Contoh: Sebuah model regresi menunjukkan bahwa setiap unit tambahan pengeluaran iklan meningkatkan penjualan sebesar \$0.001, dengan nilai  $p < 0.05$ . Hubungan ini signifikan secara statistik, tetapi tidak signifikan secara praktis karena efeknya terlalu kecil untuk berdampak pada bisnis.

# Ringkasan & Next Up

- Membaca output regresi adalah tentang menafsirkan koefisien untuk efeknya, nilai  $p$  untuk signifikansinya, dan  $R^2$  untuk kecocokan model.
- Selalu bedakan antara signifikansi statistik (apakah efeknya nyata) dan signifikansi praktis (apakah efeknya penting)
- Next Up: Setelah Anda dapat menginterpretasi model regresi, langkah selanjutnya adalah mengevaluasinya. Di lesson berikutnya, kita akan membahas Analisis Residual dan Diagnostik Model.

## 7-9. Residual Analysis And Model Diagnostics

01

**Menggunakan residual untuk mengevaluasi kinerja model.**



02

**Mengidentifikasi potensi masalah dalam kecocokan model.**



03

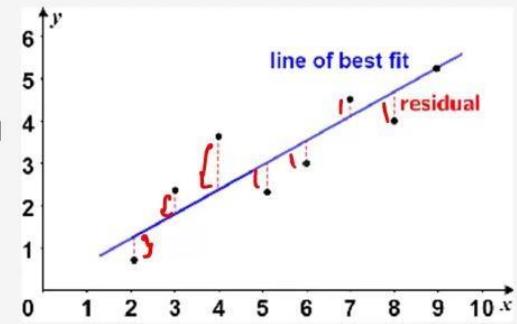
**Mendeteksi outlier dan leverage point menggunakan plot diagnostik.**



## Apa itu Analisis Residual?

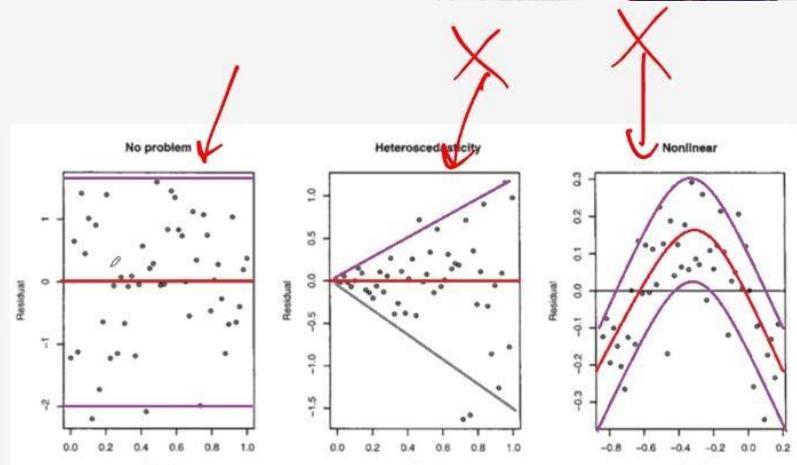
$$e = y - \hat{y}$$

- Ide Utama: Residual adalah perbedaan antara nilai aktual dari variabel dependen ( $y$ ) dan nilai yang diprediksi oleh model ( $\hat{y}$ ). Analisis residual adalah proses memeriksa residual ini untuk melihat apakah asumsi model regresi terpenuhi.
- Mengapa Itu Penting? Jika model kita berfungsi dengan baik, residual haruslah acak dan tidak menunjukkan pola. Jika ada pola dalam residual, itu adalah tanda bahwa model kita memiliki kelemahan atau bahwa salah satu asumsi regresi telah dilanggar.
- Analisis residual adalah cara kita mengonfirmasi bahwa kesimpulan dari model regresi kita valid.



## Plot Residual Dasar

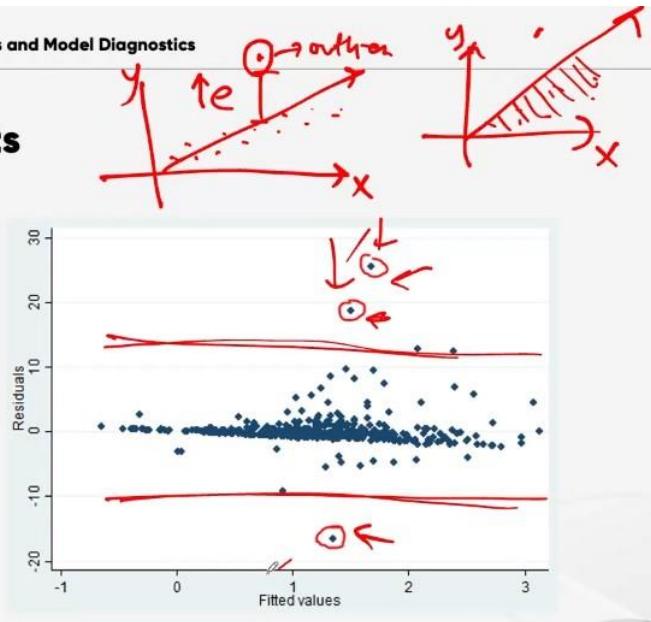
- Plot: Residual vs. Nilai Prediksi (Fitted Values).
- Interpretasi: Jika asumsi regresi terpenuhi, plot ini seharusnya terlihat seperti "kabut" acak tanpa pola yang jelas. Residual harus tersebar secara merata di sekitar garis horizontal di nol.
- Potensi Masalah (Pelanggaran Asumsi):
  - Pola Corong: Residual melebar saat nilai prediksi meningkat. Ini menunjukkan pelanggaran homoskedastisitas.
  - Pola Melengkung: Residual membentuk pola melengkung, menunjukkan pelanggaran linearitas.



Essential Mathematical Foundation – Basic of Regression Analysis – **Residual Analysis and Model Diagnostics**

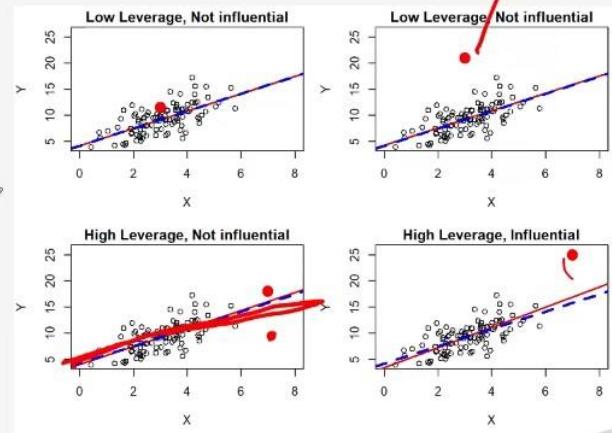
## Mendeteksi Outlier dan Leverage Points

- Outlier: Outlier adalah titik data dengan nilai residual yang sangat besar. Titik ini berada jauh secara vertikal dari garis regresi. Outlier dapat memiliki pengaruh yang besar pada koefisien regresi.
- Leverage Point: Leverage point adalah titik data yang memiliki nilai variabel independen ( $x$ ) yang sangat ekstrem. Titik ini berada jauh secara horizontal dari pusat data. Titik-titik ini dapat "menarik" garis regresi ke arahnya.
- Influential Point: Titik data yang memiliki outlier dan leverage tinggi disebut *influential point*. Titik ini memiliki dampak signifikan pada koefisien regresi.



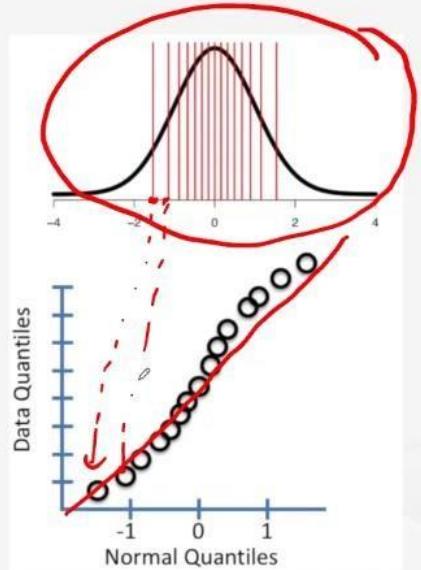
## Mendeteksi Outlier dan Leverage Points

- **Outlier:** Outlier adalah titik data dengan nilai residual yang sangat besar. Titik ini berada jauh secara vertikal dari garis regresi. Outlier dapat memiliki pengaruh yang besar pada koefisien regresi.
- **Leverage Point:** Leverage point adalah titik data yang memiliki nilai variabel independen ( $x$ ) yang sangat ekstrem. Titik ini berada jauh secara horizontal dari pusat data. Titik-titik ini dapat "menarik" garis regresi ke arahnya.
- **Influential Point:** Titik data yang memiliki outlier dan leverage tinggi disebut *influential point*. Titik ini memiliki dampak signifikan pada koefisien regresi.



## Menggunakan Q-Q Plot untuk Normalitas

- Tujuan: Untuk memeriksa asumsi bahwa residual berdistribusi normal.
- **Q-Q Plot:** Plot ini membandingkan kuantil dari residual kita dengan kuantil dari distribusi normal teoritis.
- Interpretasi:
  - **Model Baik:** Jika titik-titik data membentuk garis lurus diagonal, residual berdistribusi normal.
  - **Model Buruk:** Penyimpangan yang signifikan dari garis lurus (misalnya, melengkung di ujungnya) menunjukkan bahwa asumsi normalitas dilanggar



# **Ringkasan Diagnostik Model**

- Kecocokan Model (Model Fit):
  - R-squared: Mengukur seberapa baik model menjelaskan varians. → *output*
  - Analisis Residual: Mengonfirmasi bahwa asumsi model terpenuhi dan model dapat diandalkan. → *error*
- Ringkasan Plot Diagnostik:
  - Residual vs. Fitted: Periksa linearitas dan homoskedastisitas.
  - Q-Q Plot Residual: Periksa normalitas.
  - Plot Residual vs. Leverage: Deteksi outlier dan titik berpengaruh (influential points).

## **Ringkasan**

- Analisis Residual adalah langkah akhir yang penting dalam validasi model regresi.
- Ini memungkinkan kita untuk secara visual memeriksa asumsi yang menjadi dasar model.
- Dengan mengidentifikasi pola dalam residual, kita dapat menemukan kelemahan dalam model kita dan membuat perbaikan yang diperlukan.

**End of Part 01: Selamat! Anda telah menyelesaikan**

Essential Mathematical Foundation

**Next Up – Part 02: Python for Data Science & Machine Learning (Visualization)**

## 2. Part 2 Python For Data Science & Machine Learning (Visualization)

### Chapter 1. Introduction To Data Visualization In Python

#### 1-1. Why Data Visualization Matters

01

**Memahami mengapa visualisasi adalah pilar fundamental dalam analisis data dan bukan sekadar hiasan.**

**Masalahnya: Kita Tenggelam dalam Data**

A	B	C	D	E	F	G	H	I	J	K	L	M
2758	4663	7984	1137	6656	2250	4007	7247	4729	3130	8407	2486	5741 4
4602	4763	5846	6681	2036	4910	1591	6806	5717	4496	5881	7296	6380 2
7155	5094	3963	4736	1510	8701	5364	3837	6126	3966	1943	3778	3462 1
5215	3953	4674	4938	6216	1755	3702	4248	5819	4674	6132	5509	6346 1
3603	6075	6533	2677	3339	7318	6769	4767	5944	3685	6104	3836	6844 4
6492	6162	5918	6048	4258	5850	7971	3771	4730	3197	5149	7314	6765 3
5967	5389	5889	6877	6200	6356	3673	4732	9559	4828	9434	5133	3632 1
5894	4756	6366	6523	1185	5527	9198	5113	3022	6306	3619	8626	1803 4
6483	4917	7516	5726	4843	3746	3837	5886	9271	3959	5494	6006	4523 1
2969	4935	6574	3954	4629	6232	6278	3664	3379	3196	5633	5157	6216 2
5529	3875	6428	5903	6869	1949	5710	3796	5742	4348	5966	4214	5936 2
5209	4671	5235	6713	1824	1322	4779	5153	8113	4638	5387	4306	3064 4
4389	6513	3558	5394	4916	5579	2523	3246	1763	5028	6919	6173	3108 1
5343	6469	4794	5454	6005	3782	4919	4951	5352	3191	5403	4876	6385 3
5386	4948	5781	6241	5168	4331	6060	1314	6803	3582	6312	7112	5709 1
6162	1193	6473	3926	2216	5639	7229	7302	6394	4926	6475	7145	8957 4
5392	4791	6753	5391	5085	5984	3731	6725	8963	4885	6344	5836	6155 3
5365	1348	3966	4425	3117	5987	1982	4437	3987	6210	4363	5927	3290 2
4483	4184	8676	1077	6317	6381	4306	4121	2793	1956	5135	3259	4621 4
7162	6981	5473	6365	5251	6335	5165	3874	4004	6116	6373	4962	6980 1
4317	5646	2912	1563	3339	3277	3859	5033	5261	3407	6396	3601	3144 8
2461	6141	4083	4876	4233	5034	3491	6218	5585	4638	8236	8379	2343 1
5261	3991	2386	4726	1966	3144	8606	1008	4483	7999	9147	1796	1008 3

fasikamodern00-2025220

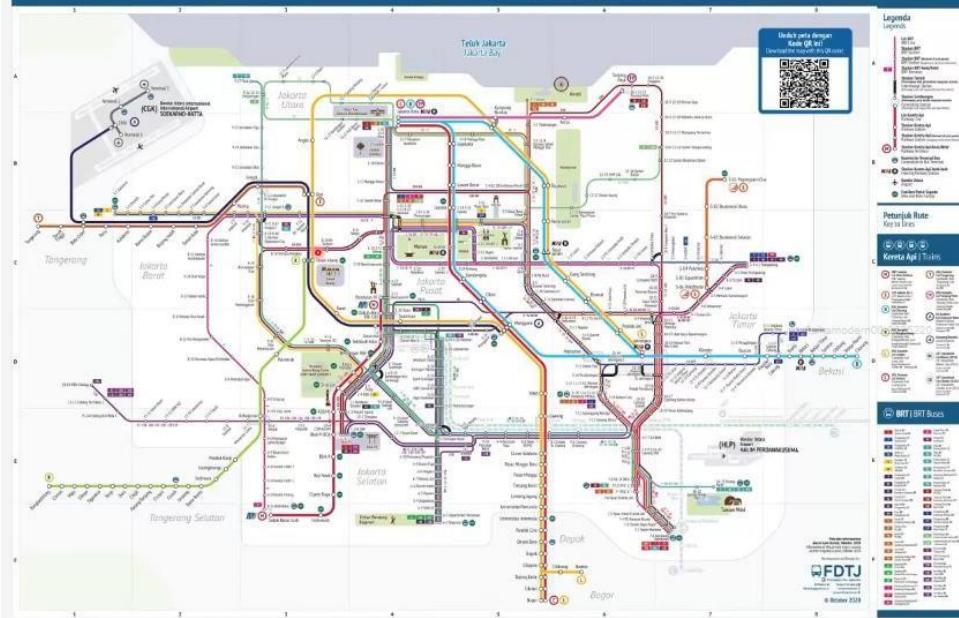


**Solusinya: Otak Kita Adalah Prosesor Visual**



Visualisasi data adalah proses menerjemahkan informasi kompleks (angka, teks) ke dalam konteks visual (grafik, peta) yang dapat kita pahami secara instan.

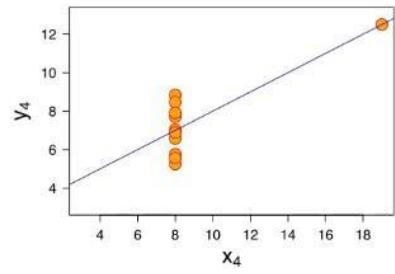
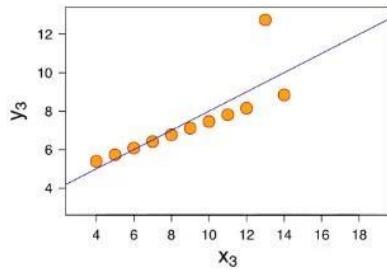
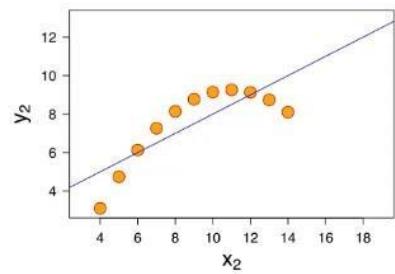
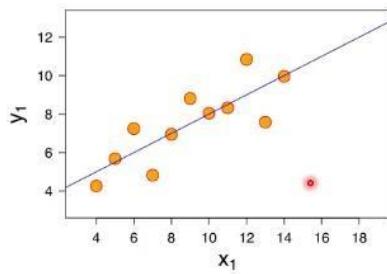
- Mengapa ini sangat efektif?
  - Kecepatan: Kita dapat mengidentifikasi tren dan *outlier* dalam hitungan detik, bukan jam.
  - Pemahaman: Grafik mengungkapkan hubungan, korelasi, dan distribusi yang tersembunyi.
  - Ingatan: Orang cenderung lebih mengingat wawasan yang disajikan secara visual.



Fast campus



## Bukti Nyata: Kuartet Anscombe



Secara statistik, mereka hampir identik:

- Rata-rata X: 9.0
- Rata-rata Y: 7.5
- Korelasi: 0.816



# Peran Visualisasi dalam Data Science

## Alat Wajib di Setiap Tahapan

Visualisasi bukan hanya untuk laporan akhir. Ini adalah alat kerja harian seorang *data scientist*.

### 1. Eksplorasi Data (EDA – Exploratory Data Analysis)

- Tujuan: "Berkenalan" dengan data Anda.
- Contoh: Membuat histogram untuk melihat distribusi usia pelanggan, atau *scatter plot* untuk melihat hubungan antara harga dan rating produk. Di sini kita **menemukan cerita**.

### 2. Pemodelan & Evaluasi

- Tujuan: Memahami seberapa baik performa model *machine learning* kita.
- Contoh: Memplot nilai prediksi vs. nilai aktual untuk melihat di mana model kita membuat kesalahan.

### 3. Komunikasi Hasil (Storytelling)

- Tujuan: Menyajikan temuan Anda kepada orang lain (manajer, klien).
- Contoh: Membuat dasbor yang jelas untuk menunjukkan pertumbuhan penjualan bulanan kepada tim bisnis. Di sini kita **menceritakan cerita** yang kita temukan.

## Dari Grafik ke Cerita

Sebuah visualisasi yang efektif lebih dari sekadar gambar; ia adalah sebuah argumen atau sebuah narasi.

Grafik Biasa (Hanya Menampilkan)



Ini adalah penjualan kita.

Visualisasi Efektif (Menceritakan)



Penjualan Meroket 40% di Bulan Agustus Berkat Kampanye "Merdeka Sale"!

# Ringkasan

- Visualisasi mengubah data yang membingungkan menjadi wawasan yang jelas.
- Ia adalah alat analisis yang krusial, bukan sekadar dekorasi.
- Tujuan akhirnya adalah storytelling: mengkomunikasikan temuan secara efektif untuk mendorong tindakan dan keputusan.

**Selanjutnya:** Kita akan menyiapkan lingkungan Python kita untuk mulai membuat visualisasi-visualisasi ini dengan *library* seperti **Matplotlib** dan **Seaborn**.

## Chapter 6. Ensemble Method For Regression

### 6-1. Matching The Visual To The Question

#### **Selalu Mulai dengan Pertanyaan**

Sebelum Anda menulis import seaborn as sns, berhenti sejenak dan tanyakan:

**"Apa pertanyaan spesifik yang ingin saya jawab dengan data ini?"** ✓

Jawaban Anda adalah kompas yang akan menuntun Anda ke jenis visualisasi yang paling efektif.

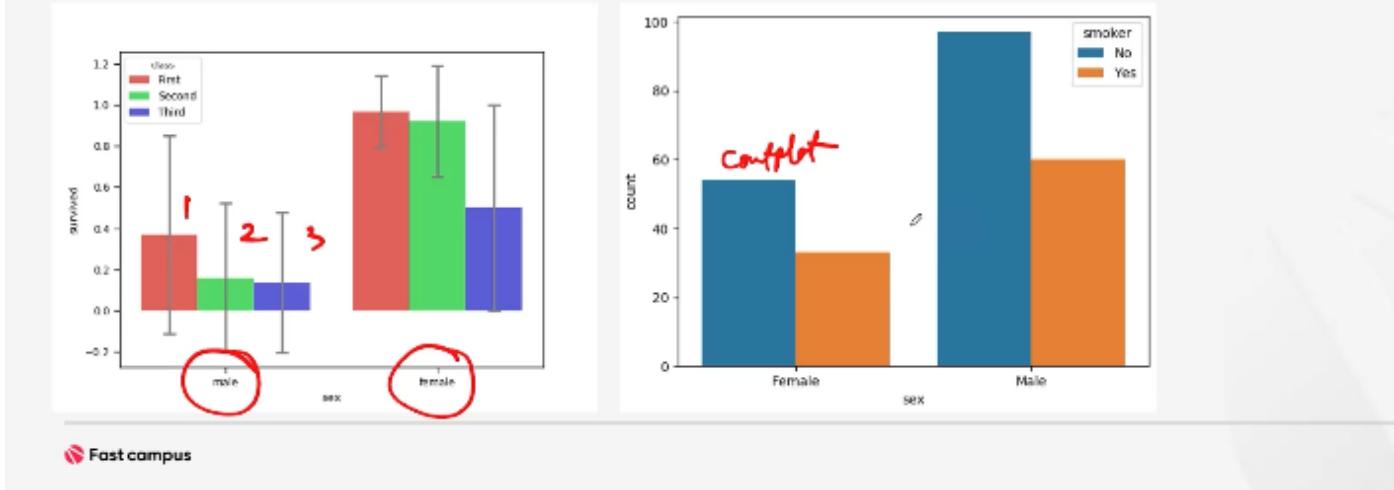
**Bukan tentang grafik yang paling keren.  
Ini tentang grafik yang paling jelas.**

## Tujuan 1: Perbandingan ✓

Pertanyaan: "Berapa banyak?" atau "Manakah yang lebih besar?"

Ketika Anda ingin membandingkan nilai antar beberapa kategori.

### Alat Utama Anda: Grafik Batang ✓

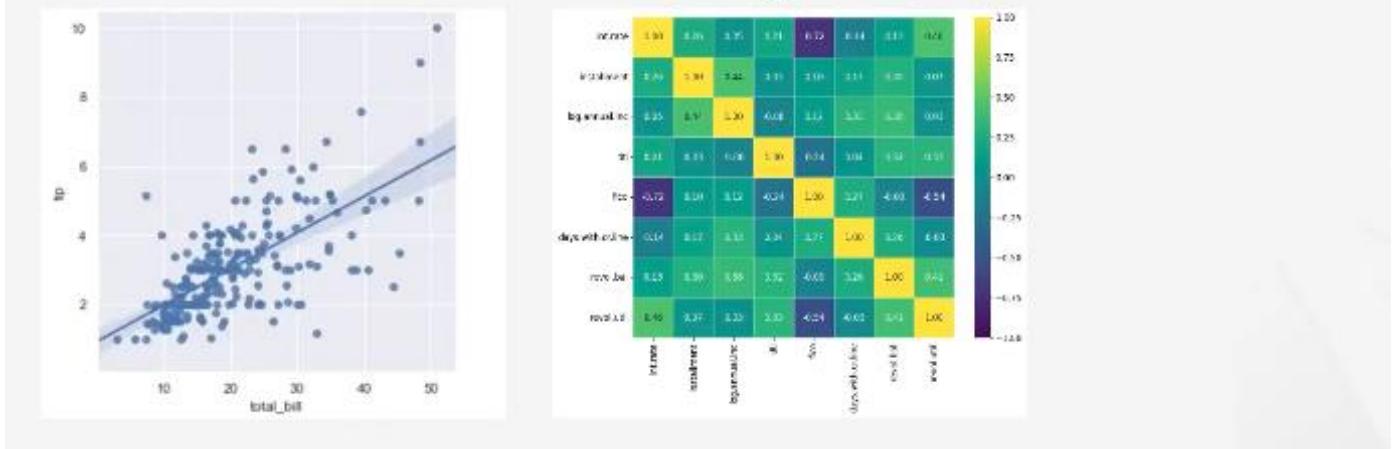


## Tujuan 2: Hubungan

Pertanyaan: "Bagaimana variabel A memengaruhi variabel B?"

Ketika Anda ingin menyelidiki korelasi atau tren antara dua atau lebih variabel numerik.

### Alat Utama Anda: Scatter & Heatmap

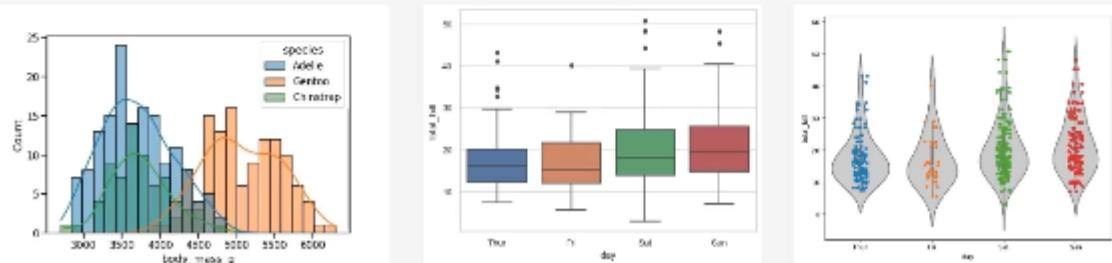


## Tujuan 3: DISTRIBUSI

Pertanyaan: "Seperti apa sebaran data saya?"

Ketika Anda ingin memahami variabilitas, rentang, dan *outlier* dalam data numerik Anda.

### Alat Utama Anda: Box & Violin



## Tujuan 4: KOMPOSISI

Pertanyaan: "Bagaimana bagian-bagian menyusun sebuah keseluruhan?"

Ketika Anda ingin menunjukkan proporsi atau persentase dari setiap kategori.

### Alat Utama Anda: Pie & Stacked Bar



fsikamodem00-2625220

# Kesimpulan

## Pilih Alat yang Tepat untuk Pekerjaan yang Tepat ✓

- Tidak ada satu grafik "terbaik" untuk semua situasi.
- Memulai dengan pertanyaan yang jelas akan menghemat waktu Anda dan membuat analisis Anda lebih tajam.
- Visualisasi yang tepat akan membuat wawasan Anda bersinar.

### 6-2. Avoiding Misleading Visuals

## Visualisasi Adalah Alat Persuasi yang Kuat

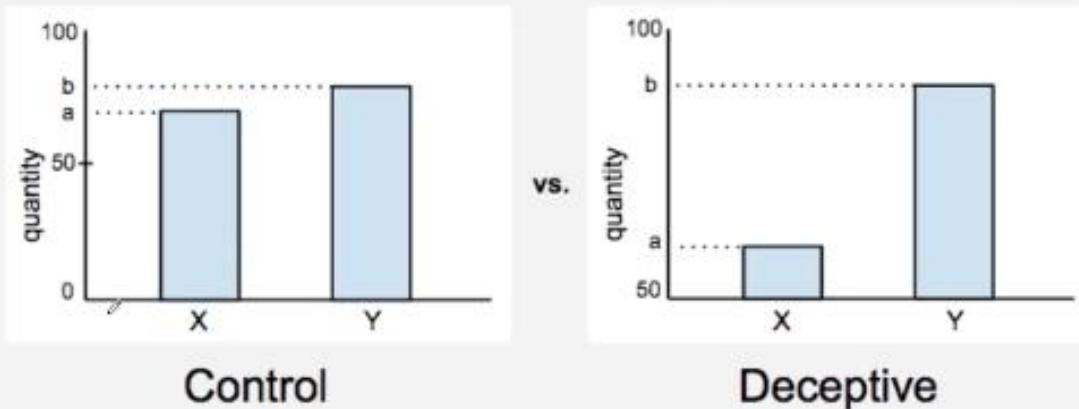
Sebuah gambar bisa lebih meyakinkan daripada seribu angka. Karena itu, visualisasi dapat dengan mudah disalahgunakan—**sengaja atau tidak**—untuk menipu atau menyesatkan audiens.

Tujuan kita sebagai praktisi data:  
**Mewakili kebenaran seakurat mungkin.**  
**Membangun kepercayaan dengan audiens kita.**  
**Mengiluminasi, bukan memanipulasi.**

## Dosa #1: Memotong Sumbu Y (Truncated Y-Axis)

Ini adalah trik yang paling umum dan paling menipu. Dengan tidak memulai sumbu Y dari angka nol pada grafik batang, perbedaan kecil dapat terlihat sangat dramatis.

### Contoh: Perbandingan penjualan



**Aturan Emas:** Untuk grafik batang, sumbu Y harus selalu dimulai dari nol.

## Dosa #2: Masalah Skala: Sumbu yang Menipu

Selain memotong sumbu, ada cara lain untuk memanipulasi persepsi melalui skala.

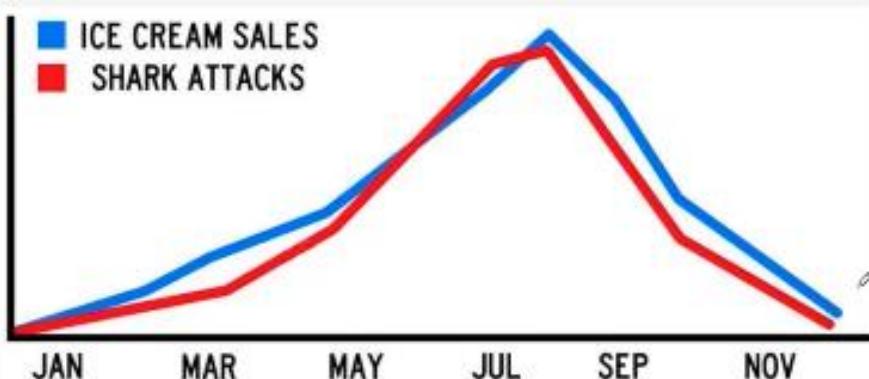
- **Skala Logaritmik yang Tidak Jelas:** Berguna untuk data dengan rentang sangat besar, tetapi jika tidak diberi label dengan jelas, dapat mengecilkan perbedaan besar.
- **Sumbu Y Ganda (Dual Y-Axis):** Menempatkan dua variabel dengan skala yang sangat berbeda pada satu grafik dapat menciptakan ilusi hubungan yang sebenarnya tidak ada. Sangat mudah untuk memanipulasi skala agar dua garis terlihat bergerak bersamaan.

**Intinya:** Selalu pastikan skala sumbu Anda jelas, diberi label, dan sesuai dengan konteks data.

## Dosa #3: Korelasi ≠ Kausalitas

*"Hanya karena dua hal terjadi bersamaan, bukan berarti yang satu menyebabkan yang lain."*

Contoh Klasik: Sebuah grafik menunjukkan bahwa penjualan es krim dan serangan hiu meningkat tajam di bulan yang sama.



Tugas Anda adalah menyajikan korelasi, tetapi berhati-hati dalam menarik kesimpulan tentang sebab-akibat.

## Dosa #4: Cherry-picked Data

Ini adalah praktik memilih-milih data yang hanya mendukung narasi Anda, sambil mengabaikan data lain yang mungkin bertentangan.

*Contoh: Menampilkan tren harga saham.*



## Dosa #5: Jebakan Visual: Efek 3D & "Chart Junk"

Elemen visual yang tidak perlu dapat mengganggu dan bahkan mendistorsi data.

- **Efek 3D:** Membuat elemen yang lebih dekat (di latar depan) terlihat lebih besar daripada yang lebih jauh, meskipun nilainya sama. Ini sangat menipu pada *pie chart*.
- **"Chart Junk":** Istilah dari Edward Tufte untuk semua elemen dekoratif yang tidak menambahkan informasi (latar belakang gambar, gradien warna yang berlebihan, ikon yang tidak relevan).



## The Right Things To Do

Sebelum Anda mempublikasikan grafik, tanyakan pada diri sendiri:

- Apakah sumbu Y saya (untuk bar chart) dimulai dari nol?
- Apakah skala sumbu saya jelas dan tidak dimanipulasi?
- Apakah judul dan label saya bersifat deskriptif, bukan persuasif?
- Apakah saya sudah menampilkan konteks data yang cukup (tidak *cherry-picking*)?
- Apakah saya membedakan antara korelasi dan kausalitas dalam narasi saya?
- Apakah saya sudah menghilangkan semua "sampah visual" yang tidak perlu?

# Kesimpulan

Tujuan Anda: Illuminasi, Bukan Manipulasi

- Visualisasi data yang etis adalah tentang kejujuran intelektual.
- Audiens Anda memercayai Anda untuk menyajikan data secara akurat.
- Dengan mengikuti praktik terbaik, Anda memastikan bahwa visualisasi Anda adalah alat untuk pemahaman yang lebih dalam, bukan untuk kebingungan atau penipuan.

## 6-3. Principles Of Effective Visual Design

### Jembatan dari Etika ke Desain

Setelah "Apa" dan "Mengapa", Sekarang "Bagaimana"

fakkomdem00-2925220

Di subchapter sebelumnya, kita belajar cara membuat visualisasi yang jujur. Sekarang, kita akan belajar cara membuatnya efektif.

- **Grafik yang Jujur:** Menampilkan data secara akurat.
- **Grafik yang Efektif:** Memastikan audiens memahami data tersebut dengan cepat dan mudah.

Tujuannya adalah mengurangi beban kognitif audiens, sehingga mereka bisa fokus pada wawasan, bukan pada mencoba memahami grafik Anda.

# Prinsip #1: Decluttering (Merampingkan)

Hilangkan Semua yang Tidak Perlu

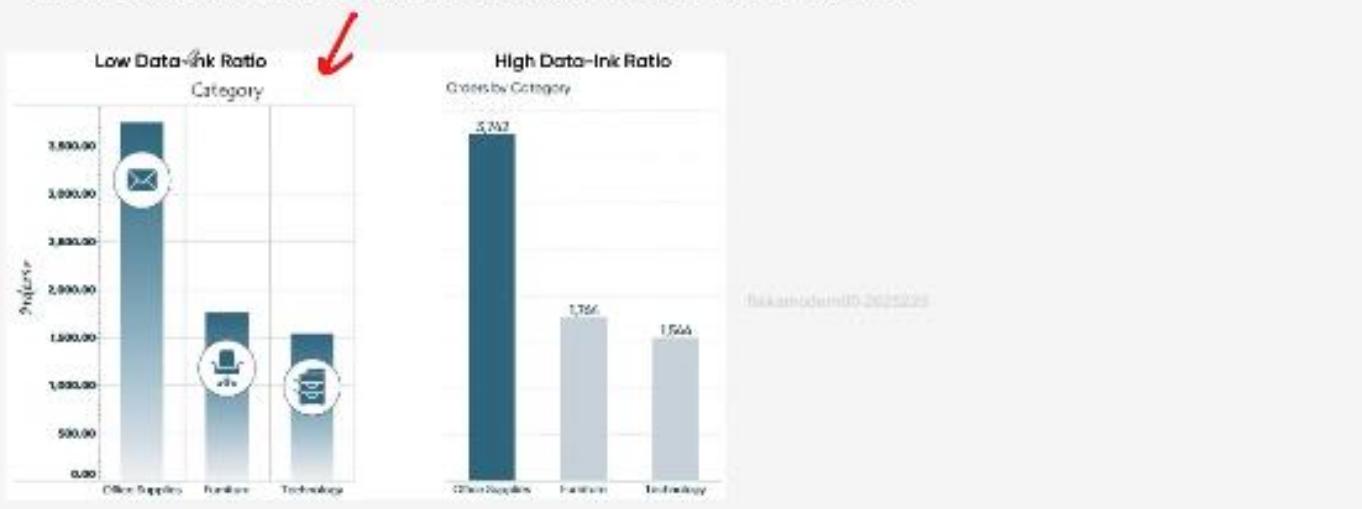


Setiap elemen dalam visualisasi Anda harus memiliki tujuan. Jika tidak, ia hanya menjadi "sampah visual" (*chart junk*) yang mengganggu.

Konsep Kunci: Data-Ink Ratio (Edward Tufte) Maksimalkan rasio "tinta" yang digunakan untuk menampilkan data, dan minimalkan tinta untuk elemen non-data (dekorasi, grid berlebihan, dll.).

## Prinsip #1: Decluttering (Merampingkan)

Konsep Kunci: Data-Ink Ratio (Edward Tufte) Maksimalkan rasio "tinta" yang digunakan untuk menampilkan data, dan minimalkan tinta untuk elemen non-data (dekorasi, grid berlebihan, dll.).



# Prinsip #1: Decluttering (Merampingkan)

Konsep Kunci: Data-Ink Ratio (Edward Tufte) Maksimalkan rasio "tinta" yang digunakan untuk menampilkan data, dan minimalkan tinta untuk elemen non-data (dekorasi, grid berlebihan, dll.).



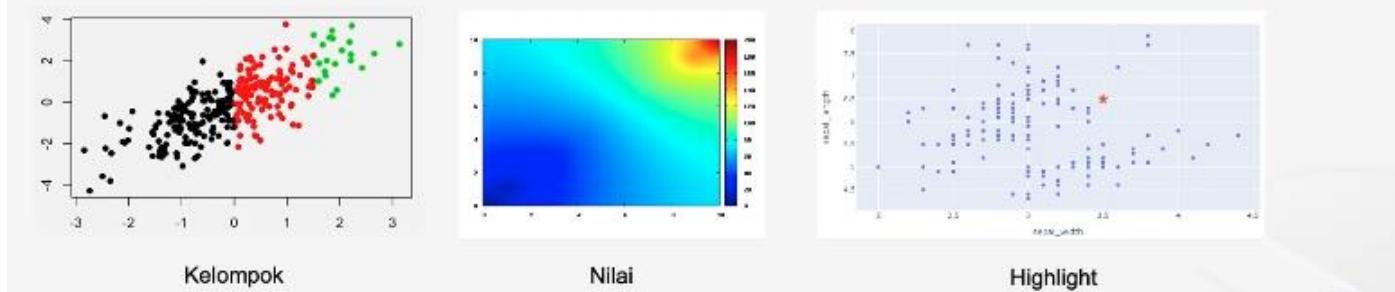
## Prinsip #1: Decluttering (Merampingkan)

- Hapus Latar Belakang:** Latar belakang berwarna atau bergambar jarang sekali membantu. Gunakan latar putih atau abu-abu muda.
- Hapus Border Grafik:** Garis batas di sekeliling area plot biasanya tidak diperlukan.
- Ringankan atau Hapus Gridlines:** Jika Anda butuh grid, buat warnanya abu-abu muda dan tipis agar tidak mendominasi. Seringkali, grid horizontal saja sudah cukup.
- Hapus Redundansi:** Jika Anda sudah memiliki label data di atas batang, apakah Anda masih memerlukan sumbu Y? Mungkin tidak.
- Gunakan Label Langsung:** Alih-alih menggunakan legenda, seringkali lebih jelas untuk memberi label langsung pada garis atau batang.

## Prinsip #2: Penggunaan Warna yang Strategis

Warna Bukanlah Dekorasi, Melainkan Alat Komunikasi

Jangan gunakan warna hanya karena "terlihat bagus". Setiap warna harus memiliki makna dan tujuan.



## Prinsip #2: Penggunaan Warna yang Strategis

Gunakan Palet yang Tepat:

- Sekuensial: Gradien satu warna, bagus untuk data numerik dari rendah ke tinggi.
- Diverging: Dua warna yang bertemu di tengah netral, bagus untuk menunjukkan penyimpangan dari titik tengah (misalnya, positif vs. negatif).
- Kategorikal: Warna-warna berbeda yang mudah dibedakan, untuk data kualitatif.

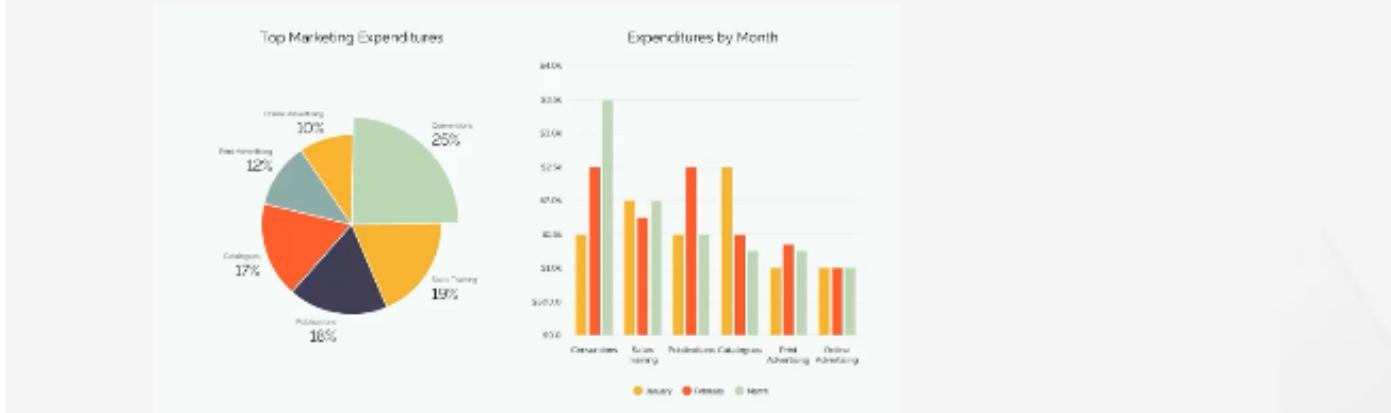
Perhatikan Buta Warna: Sekitar 8% pria memiliki buta warna (terutama merah-hijau). Gunakan palet yang ramah buta warna. Seaborn memiliki banyak palet bawaan seperti ini ('colorblind').

Jaga Konsistensi: Jika Anda menggunakan biru untuk "Apple" di satu grafik, gunakan biru untuk "Apple" di semua grafik Anda.

## Prinsip #3: Memandu Perhatian dengan Hierarki Visual

Kontrol ke Mana Mata Audiens Melihat Terlebih Dahulu

Hierarki visual adalah seni mengatur elemen desain untuk menunjukkan urutan kepentingannya. Anda ingin audiens melihat judul terlebih dahulu, lalu wawasan utama, baru kemudian detail pendukung.



## Prinsip #3: Memandu Perhatian dengan Hierarki Visual

Cara Menciptakan Hierarki:

- **Ukuran:** Elemen yang lebih besar akan lebih menarik perhatian.  
Buat judul Anda lebih besar dari label sumbu.
- **Warna & Kontras:** Warna yang cerah dan kontras tinggi akan menonjol. Gunakan ini untuk data yang paling penting.
- **Posisi:** Mata kita secara alami membaca dari kiri atas ke kanan bawah. Letakkan informasi terpenting di jalur ini.

# Kesimpulan

Desain yang Baik Membuat Data Tidak Terlihat Menakutkan

- Declutter: Buang semua yang tidak perlu.
- Gunakan Warna dengan Tujuan: Untuk menyoroti, mengelompokkan, atau menyatakan nilai.
- Ciptakan Hierarki Visual: Pandu mata audiens Anda ke wawasan yang paling penting.

Dengan menerapkan prinsip-prinsip ini, Anda tidak hanya menyajikan data, tetapi Anda membuatnya dapat diakses, dapat dipahami, dan mudah diingat.

## 6-4. Telling A Story With Data

# Telling a Story with Data

Sebagai seorang analis, Anda bekerja dalam dua mode:

- Analisis Eksplorasi (Untuk Diri Sendiri):
  - Tujuan: Memahami data, menemukan pola, mencari wawasan.
  - Proses: Anda mencoba puluhan jenis grafik, memfilter data, dan mengikuti setiap petunjuk. Prosesnya berantakan dan tidak linear.
- Analisis Penjelasan (Untuk Audiens):
  - Tujuan: Mengkomunikasikan wawasan spesifik kepada orang lain.
  - Proses: Anda sudah tahu ceritanya. Tugas Anda adalah menyusun visualisasi secara strategis untuk menceritakan kisah itu dengan jelas dan persuasif.

# Struktur Cerita Klasik



Setiap Cerita yang Baik Memiliki Awal, Tengah, dan Akhir

Dari film blockbuster hingga dongeng, narasi yang efektif mengikuti struktur tiga babak:

- Babak I: The Setup (Pengenalan)
  - Memperkenalkan karakter, latar, dan konflik awal. Menciptakan konteks.
- Babak II: The Confrontation (Konfrontasi)
  - Konflik memuncak. Karakter menghadapi rintangan. Terjadi *rising action* dan *turning point*.
- Babak III: The Resolution (Penyelesaian)
  - Konflik diselesaikan. Ada kesimpulan dan perubahan yang terjadi.

## Struktur Tiga Babak untuk Data

Babak	Dalam Cerita Film	Dalam Cerita Data
Babak I: The setup	✓ Memperkenalkan dunia dan masalahnya	Menyajikan Konteks. Apa masalah bisnisnya? Data apa yang kita lihat? Apa status quo saat ini?
Babak II: The Confrontation	Pahlawan melawan penjahat. <small>https://www.pexels.com/photo/a-hero-fight-with-a-villain-in-a-city-422220/</small>	Menyajikan Analisis & Temuan. Di sinilah "aha!" momen terjadi. Tunjukkan anomali, hubungan, atau tren yang Anda temukan.
Babak III: The Resolution	Pahlawan menang dan dunia berubah.	Menyajikan Kesimpulan & Rekomendasi. Jelaskan arti dari temuan Anda ("So what?"). Apa yang harus dilakukan selanjutnya?

# Langkah 1: Kenali Audiens & Tentukan "Big Idea" Anda

## 1. Siapa Audiens Saya?

- Apakah mereka eksekutif (butuh gambaran besar)? Atau analis (butuh detail teknis)? Sesuaikan kedalaman cerita Anda.

## 2. Apa "Big Idea" Saya?

- Rangkum pesan utama Anda dalam satu kalimat lengkap. Ini adalah inti dari cerita Anda.
- Contoh Buruk: "Grafik penjualan bulanan." (Ini hanya deskripsi)
- Contoh Baik: "Meskipun penjualan secara keseluruhan naik, profitabilitas kita menurun karena diskon yang berlebihan pada produk kategori A." (Ini memiliki sudut pandang dan kesimpulan)

"Big Idea" Anda adalah tujuan akhir dari cerita Anda.

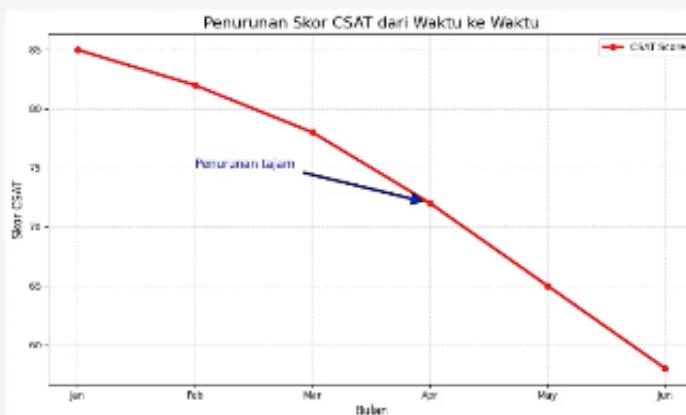
## Studi Kasus: Misteri Penurunan Kepuasan Pelanggan

Mari kita terapkan struktur tiga babak.

### Babak I: The Setup (Konteks)

Masalah: "Tim manajemen khawatir karena skor kepuasan pelanggan (CSAT) kita terus menurun selama tiga kuartal terakhir."

Visual Anda: Sebuah lineplot sederhana yang secara dramatis menunjukkan tren penurunan ini.

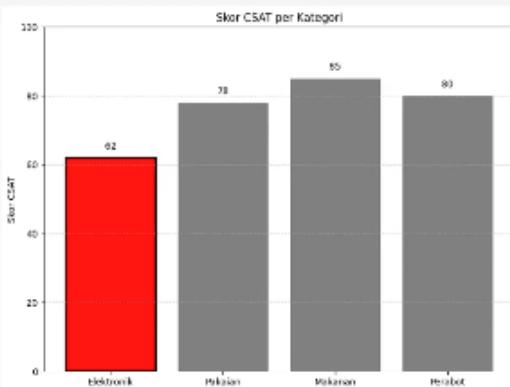


## Studi Kasus: Babak II (Konfrontasi & Temuan)

Sekarang kita tunjukkan analisis kita. Kita memandu audiens melalui penemuan kita, selangkah demi selangkah.

Visual 1: barplot CSAT per Kategori Produk ✓

"Kami memecah skor CSAT berdasarkan kategori produk dan menemukan bahwa penurunan terbesar terjadi pada kategori 'Elektronik'."

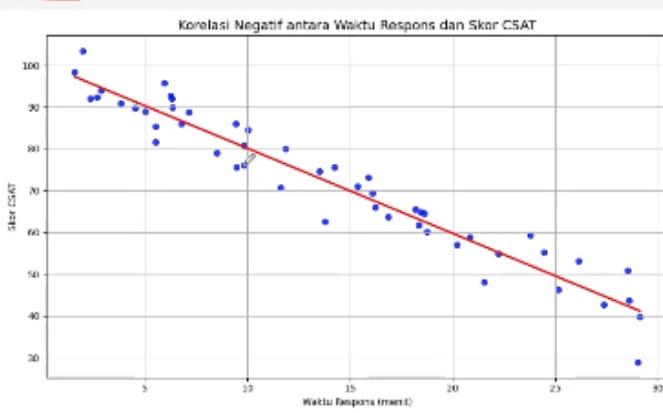


## Studi Kasus: Babak II (Konfrontasi & Temuan)

Sekarang kita tunjukkan analisis kita. Kita memandu audiens melalui penemuan kita, selangkah demi selangkah.

Visual 2: scatterplot Waktu Respons vs. CSAT (Hanya untuk Elektronik)

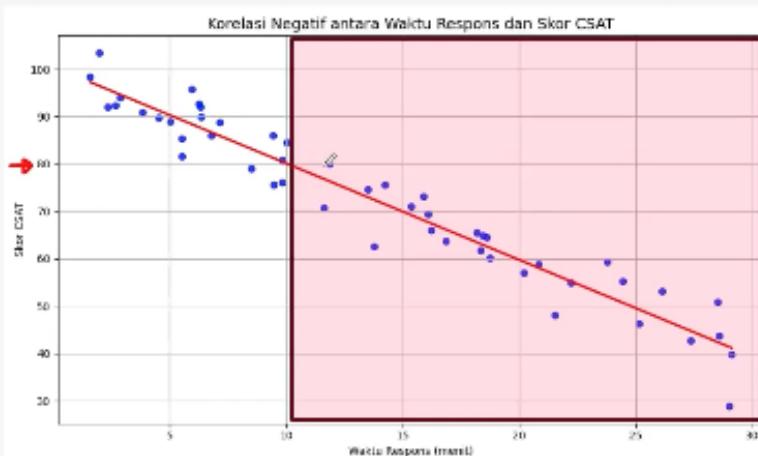
"Lebih dalam lagi, kami menemukan korelasi negatif yang kuat: semakin lama waktu respons layanan pelanggan, semakin rendah skor CSAT, terutama untuk produk elektronik."



## Studi Kasus: Babak III (Resolusi & Rekomendasi)

Saatnya menutup cerita dengan kesimpulan yang jelas dan seruan untuk bertindak (call to action).

Visual Anda: Bisa jadi visual yang sama dari Babak II, tetapi sekarang dengan anotasi yang memperkuat pesan utama.



## Studi Kasus: Babak III (Resolusi & Rekomendasi)

Kesimpulan (The "So What?"):

"Penurunan skor CSAT kita secara langsung disebabkan oleh lambatnya waktu respons tim layanan pelanggan untuk tiket terkait produk elektronik."

Rekomendasi (The "Now What?"):

"Kami merekomendasikan untuk memprioritaskan pelatihan dan penambahan staf untuk tim layanan pelanggan produk elektronik untuk menargetkan waktu respons di bawah 10 menit."

# Kesimpulan

Anda Bukan Hanya Analis, Anda Adalah Pencerita

- Jangan hanya menunjukkan grafik, jelaskan artinya.
- Gunakan struktur tiga babak untuk memandu audiens Anda secara logis dari masalah ke solusi.
- Mulai dengan pertanyaan, bukan dengan data.
- Akhiri dengan rekomendasi yang dapat ditindaklanjuti.

Visualisasi yang dirangkai dalam sebuah cerita yang kuat adalah hal yang mengubah data menjadi keputusan dan tindakan.

## 3. Part 3 Core Machine Learning Concept & Algorithms (Regression)

### Chapter 1. Introduction To Supervised Learning & Regression

#### 1-1. What Is Supervised Learning

#### Tujuan Pembelajaran Chapter Ini



Menjelaskan konsep inti dari **Supervised Learning** menggunakan analogi yang mudah dipahami.



Mengidentifikasi perbedaan fundamental antara **Input (Fitur)** dan **Output (Label)** dalam sebuah dataset.



Membedakan dua tugas utama dalam **Supervised Learning**: **Regresi** (memprediksi angka) dan **Klasifikasi** (memprediksi kategori).



Memberikan contoh-contoh nyata aplikasi **Supervised Learning** yang Anda temui setiap hari.

# Analogi Sederhana: Belajar dengan Kartu Flash

## PERTANYAAN

"Ibu Kota Prancis?"

"Gunung Tertinggi?"

...

"Ibu Kota Italia?"

SUPERVISED  
LEARNING

## JAWABAN

Paris

Everest

...

????

## SUPERVISED LEARNING

*Supervised Learning* (Pembelajaran Terarah) adalah cabang dari *machine learning* di mana kita "mengajari" sebuah model komputer dengan memberinya **data yang sudah memiliki jawaban yang benar (label)**. Data ini berfungsi sebagai "kunci jawaban" yang digunakan model untuk belajar.

## KOMPONEN

- **Input (Fitur / Features):** Ini adalah "pertanyaan" atau data mentah yang kita berikan ke model. Fitur adalah variabel-variabel individual yang mendeskripsikan setiap data.
- **Output (Label/Target):** Ini adalah "jawaban yang benar" yang ingin kita prediksi.

$f(g, \theta)$  ✓

### Data Berlabel

Kita memberikan ribuan contoh "pertanyaan dan jawaban"

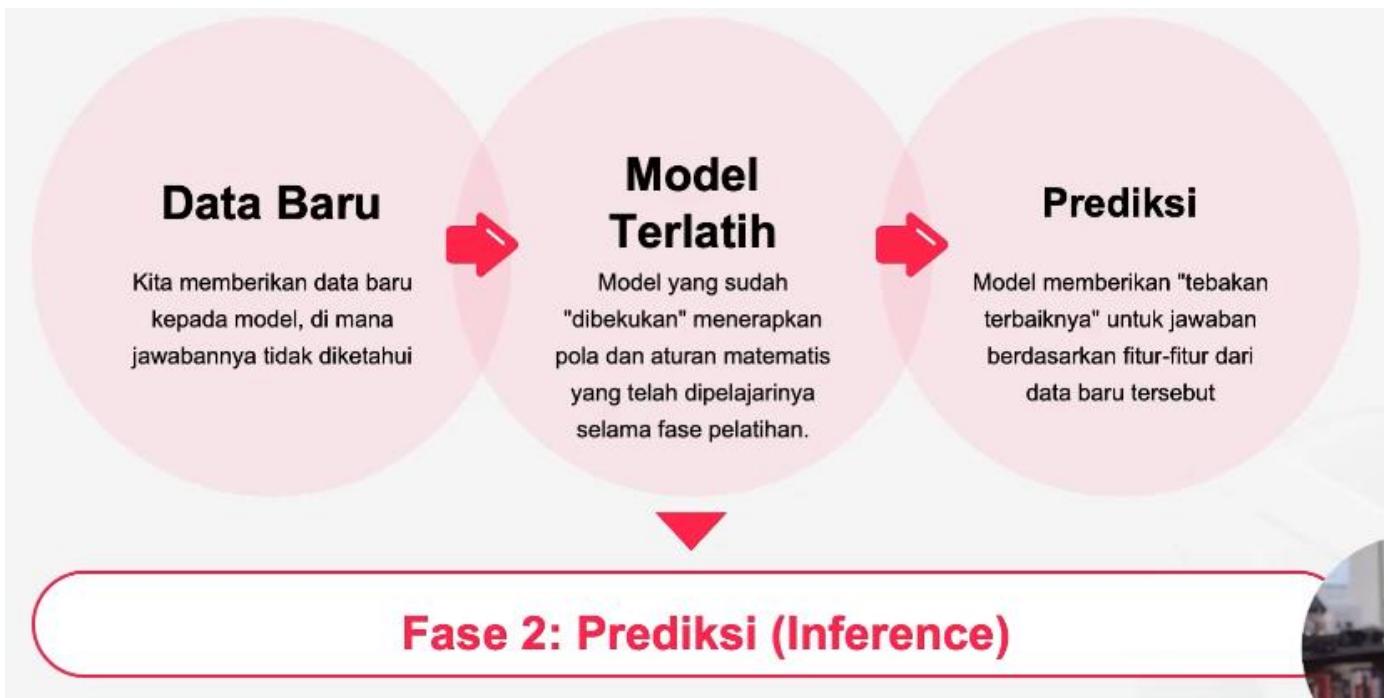
### Algoritma

Algoritma *machine learning* (seperti Regresi Linear) menganalisis data ini. Prosesnya bersifat iteratif.

### Model Terlatih

Setelah proses iteratif ini selesai, hasilnya adalah sebuah "model" yang telah "belajar" pola dari data

Fase 1: Pelatihan (Training)



## Dua Tugas Utama dalam Supervised Learning

*Supervised Learning* memiliki dua jenis tugas utama, tergantung pada jenis "jawaban" yang ingin kita prediksi.

Jenis Tugas	Pertanyaan yang Dijawab	Contoh
Regression (Regresi)	"Berapa banyak?" atau "Berapa nilainya?"	Memprediksi harga rumah, suhu besok, jumlah penjualan, durasi penerbangan, atau estimasi permintaan listrik.
Classification (Klasifikasi)	"Yang mana?" atau "Apa jenisnya?"	Memprediksi apakah email itu <b>spam/bukan spam</b> , apakah penumpang <b>Titanic selamat/tidak selamat</b> , atau mengidentifikasi objek dalam gambar ( <b>kucing/anjing</b> ).

## Contoh di Dunia Nyata

### Contoh 1

#### Filter Spam Email (Klasifikasi)

- **Input (Fitur):** Teks email (kata-kata seperti "gratis", "promo", "pemenang"), reputasi pengirim, subjek, apakah ada lampiran yang aneh.
- **Output (Label):** Kategori ("Spam" atau "Bukan Spam").

### Contoh 2

#### Prediksi Harga Rumah (Regresi)

- **Input (Fitur):** Luas rumah, jumlah kamar tidur, jumlah kamar mandi, lokasi (kode pos), usia bangunan, jarak ke sekolah terdekat.
- **Output (Label):** Harga rumah (misalnya, Rp 1.5 Miliar).

### Contoh 3

#### Rekomendasi Film (Klasifikasi/Regresi)

- **Input (Fitur):** Film yang pernah Anda tonton, rating yang Anda berikan, genre film, aktor yang bermain.
- **Output (Label):** Bisa berupa prediksi rating yang akan Anda berikan untuk film baru (Regresi), atau apakah Anda akan suka/tidak suka (Klasifikasi).

### Contoh 4

#### Diagnosis Medis (Klasifikasi)

- **Input (Fitur):** Data dari gambar rontgen, hasil tes darah, detak jantung pasien, riwayat kesehatan keluarga.
- **Output (Label):** Kategori diagnosis ("Sakit" atau "Sehat", atau jenis penyakit spesifik).

# Ringkasan

## Poin Kunci untuk Diingat



- *Supervised Learning* belajar dari **data berlabel** (contoh dengan jawaban yang sudah diverifikasi).
- Tujuannya adalah untuk **menggeneralisasi** pola agar bisa **memprediksi jawaban** pada data baru yang tak terlihat.
- Dua tugas utamanya adalah **Regresi** (memprediksi angka kontinu) dan **Klasifikasi** (memprediksi kategori diskrit).
- Kualitas model sangat bergantung pada kualitas dan kuantitas data pelatihannya. Prinsip "**Garbage In, Garbage Out**" sangat berlaku di sini: data yang buruk akan menghasilkan model yang buruk.

sisikamadan00-2025220

## Langkah Selanjutnya

**Selanjutnya:** Kini setelah Anda memahami konsep dasarnya, kita akan menyelami lebih dalam perbedaan antara Regresi dan Klasifikasi. Setelah itu, kita akan memperkenalkan salah satu konsep paling fundamental dan wajib dalam setiap alur kerja *machine learning*: **Train-Test Split**, sebuah teknik untuk menguji seberapa baik model kita sebenarnya.

### 2-1. Regression VS Classification

#### Tujuan Pembelajaran

sisikamadan00-2025220

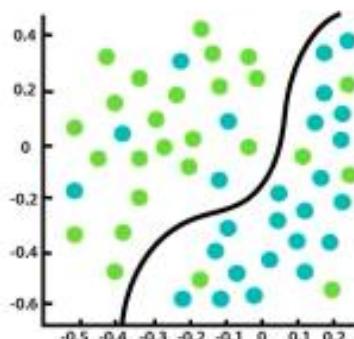
- ✓ **Membedakan** secara jelas antara masalah **Regresi** dan **Klasifikasi** berdasarkan sifat variabel targetnya.
- ✓ **Mengidentifikasi** dengan percaya diri jenis masalah *machine learning* dari deskripsi sebuah skenario bisnis atau ilmiah.
- ✓ **Menjelaskan** mengapa perbedaan ini sangat penting dan bagaimana ia secara fundamental memengaruhi pilihan algoritma, cara model "berpikir", dan metrik evaluasi yang digunakan.
- ✓ **Memberikan** contoh-contoh praktis yang beragam untuk setiap jenis masalah, melampaui contoh-contoh yang paling umum.

# Recap

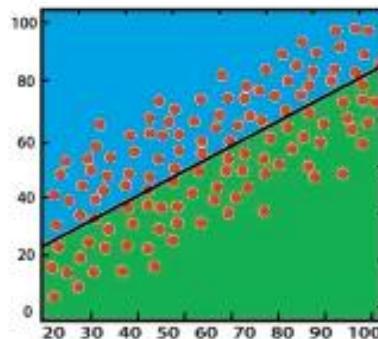
Kita tahu bahwa *Supervised Learning* adalah tentang belajar dari data berlabel

## Pertanyaan Kunci

Apakah kita mencoba memprediksi sebuah kuantitas yang dapat diukur, atau sebuah label dari kelompok tertentu?



Classification



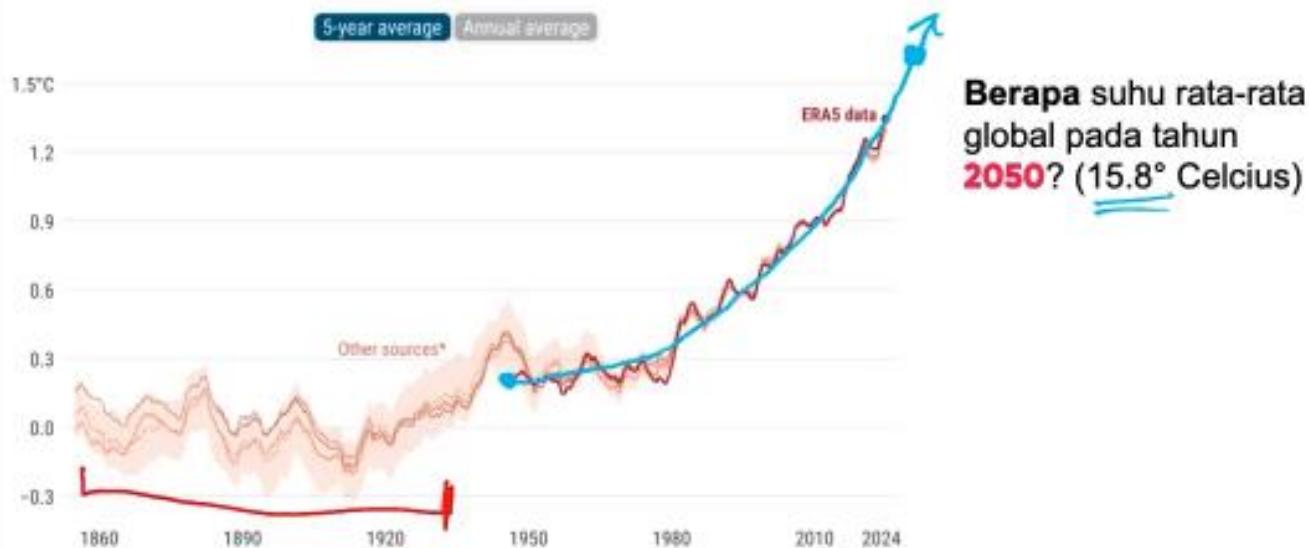
Regression

## REGRESSION - Menjawab "Berapa Banyak?"

Regresi digunakan ketika target prediksi kita adalah nilai **numerik kontinu**.

Global surface temperature: increase above pre-industrial ✓

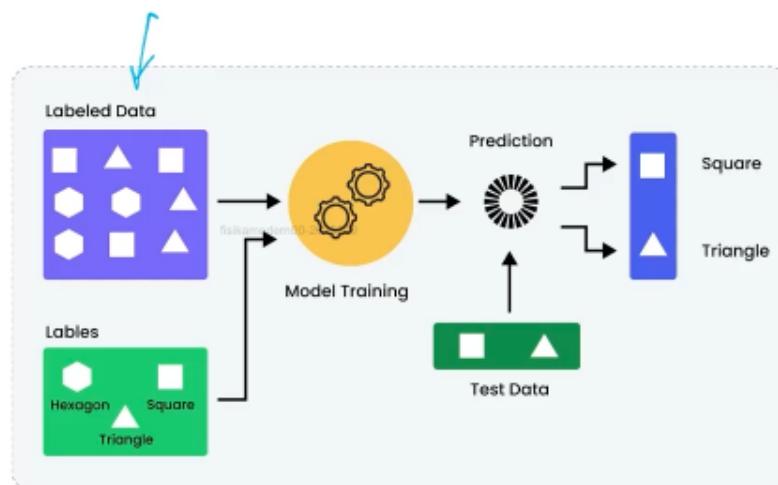
Reference period: pre-industrial (1850–1900) • Credit: C3S/ECMWF



\*Other sources comprise JRA-30, GISTEMPv4, NOAA GlobalTempv6, Berkeley Earth and the HadCRUT5 ensemble mean. Shading shows the range of the HadCRUT5 ensemble.

# CLASSIFICATION - Menjawab "Yang Mana?"

Klasifikasi digunakan ketika target prediksi kita adalah **kategori diskrit** atau **kelas**.



"Diskrit" berarti jawabannya harus dipilih dari daftar pilihan yang terbatas, berbeda, dan telah ditentukan sebelumnya. Tidak ada nilai "di antara" dua kategori.

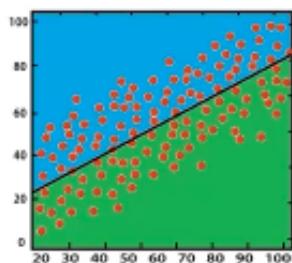
## Perbandingan Langsung: Sifat Output

Inilah perbedaan paling fundamental antara keduanya, yang memengaruhi segalanya.

Fitur	Regression	Classification
Jenis Output	Numerik Kontinu (Kuantitatif)	Kategorial/Diskrit (Kualitatif)
Sifat Jawaban	Bisa berupa angka apa saja dalam sebuah rentang. Jumlah kemungkinan tak terbatas	Harus dipilih dari daftar kelas yang terbatas dan telah ditentukan
Contoh Jawaban	173.5 cm, Rp. 5.432, -5.2	"Ya"/"Tidak "Tipe A"/"Tipe B"/"Tipe C"
Visualisasi Khas	Scatter plot dengan garis tren. Tujuannya Adalah untuk melihat seberapa dekat titik-titik data dengan garis prediksi	Titik data yang dikelompokan, dipisahkan oleh batas. Tujuannya Adalah untuk melihat seberapa baik model memisahkan kelompok

# Bagaimana Model "Berpikir" Secara Berbeda

Karena tujuan dan sifat outputnya berbeda, cara model "berpikir" atau fungsi objektif yang dioptimalkannya juga berbeda secara fundamental.

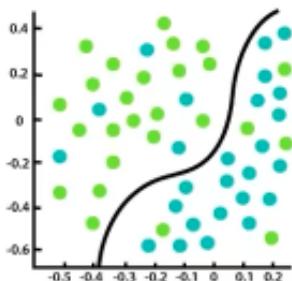


Regression

## Tujuan:

"Bagaimana cara saya menggambar garis atau kurva yang paling pas (best-fit line) yang meminimalkan total jarak (kesalahan) antara garis prediksi saya dan semua titik data aktual?"

Model ini dihukum berdasarkan seberapa *jauh* prediksinya dari nilai sebenarnya.



Classification

## Tujuan:

"Bagaimana cara saya menggambar batas keputusan (decision boundary) yang paling efektif untuk memisahkan titik-titik data ke dalam kelompok yang benar?"

Model ini dihukum berdasarkan apakah ia menempatkan titik data di sisi batas yang *benar* atau *salah*.

## Latihan Cepat: Regresi atau Klasifikasi?

Mari kita uji pemahaman Anda. Tentukan jenis masalah untuk setiap skenario berikut:

Problem	Regresi	Klasifikasi
<b>Memprediksi jumlah pengunjung sebuah situs web besok.</b>	✓	
<b>Menentukan apakah ulasan produk bersifat positif, negatif, atau netral.</b>		✓
<b>Mendiagnosis apakah seorang pasien menderita penyakit tertentu (Ya/Tidak).</b>		✓
<b>Memprediksi rating bintang (1, 2, 3, 4, atau 5) yang akan diberikan pengguna pada sebuah film.</b>		✓

# Ringkasan

	Regression	Classification
Tujuan	Memprediksi nilai kontinu	Memprediksi label kategori
Pertanyaan	Berapa banyak?	Yang Mana
Contoh	Harga, Suhu	Selamat, Spam

## Mengapa ini Sangat Penting

Memilih antara Regresi dan Klasifikasi adalah **langkah pertama dan paling krusial** dalam setiap proyek *supervised learning*. Pilihan ini akan menentukan seluruh jalur proyek Anda:

- **Algoritma** yang bisa Anda gunakan: Anda akan menggunakan **Linear Regression** atau **Decision Tree Regressor** untuk masalah regresi, tetapi **Logistic Regression** atau **Decision Tree Classifier** untuk masalah klasifikasi. Keduanya bekerja dengan cara yang sangat berbeda.
- **Metrik** yang Anda gunakan untuk mengevaluasi keberhasilan model: Untuk regresi, Anda akan mengukur "seberapa jauh kesalahan prediksi" dengan metrik seperti **Mean Squared Error (MSE)**. Untuk klasifikasi, Anda akan mengukur "berapa banyak tebakan yang benar" dengan metrik seperti **Akurasi, Presisi, dan Recall**.
- Salah membingkai masalah—misalnya, menggunakan model regresi untuk memprediksi kategori—akan membawa Anda ke jalan yang salah sejak awal dan menghasilkan model yang tidak masuk akal.

### 3-1. The Machine Learning Workflow And Train Test Split

## Tujuan Pembelajaran



Menggambarkan alur kerja standar dalam sebuah proyek *machine learning* dari awal hingga akhir, memahami bahwa pemodelan hanyalah salah satu bagian dari proses yang lebih besar.



Menjelaskan secara mendalam mengapa menguji model pada data yang sama dengan data pelatihannya adalah kesalahan metodologis yang fatal, dengan menggunakan analogi yang kuat untuk mengilustrasikan konsep **menghafal (memorization)** versus **memahami (generalization)**.

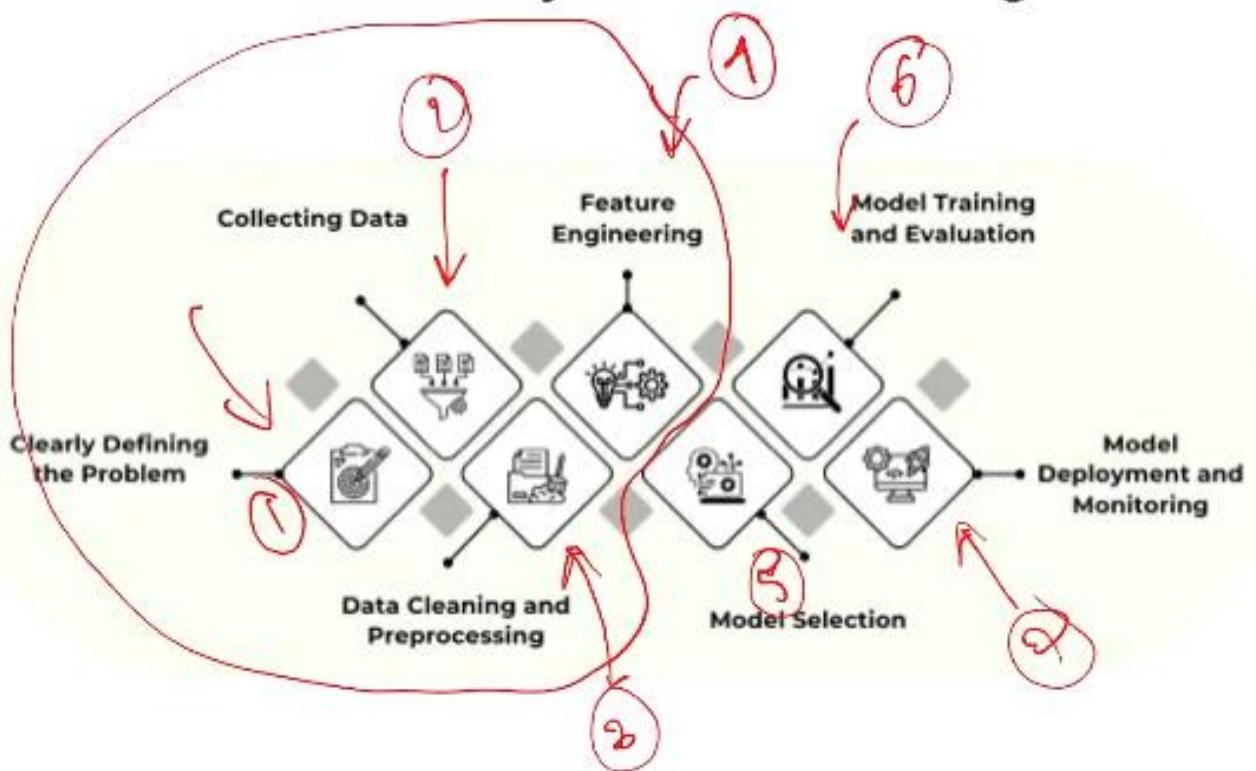


Memahami konsep, tujuan, dan mekanisme dari **Train-Test Split** sebagai pilar utama untuk evaluasi model yang objektif dan tidak bias.

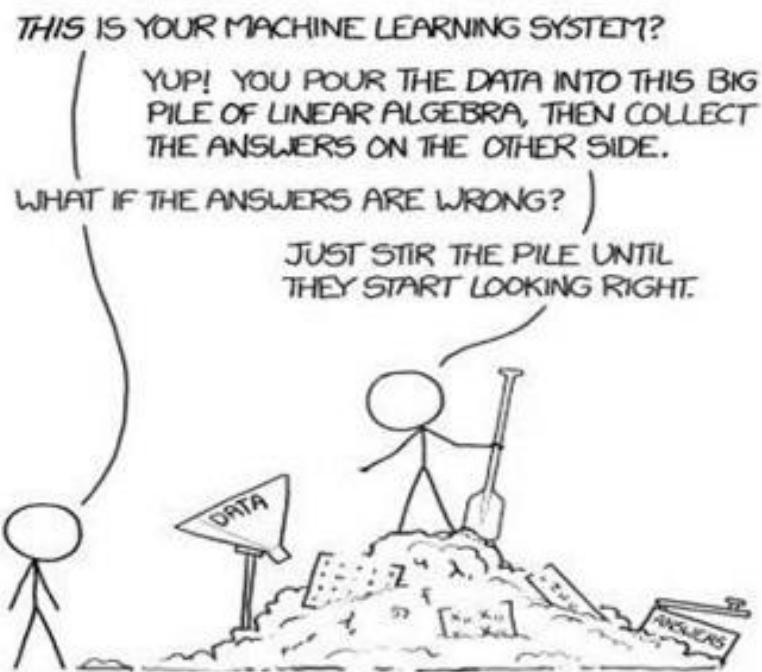


Mengimplementasikan `train_test_split` menggunakan library Scikit-learn, serta menjelaskan arti dan pentingnya parameter kunci seperti `test_size` dan `random_state`.

## Gambaran Besar: Alur Kerja Machine Learning



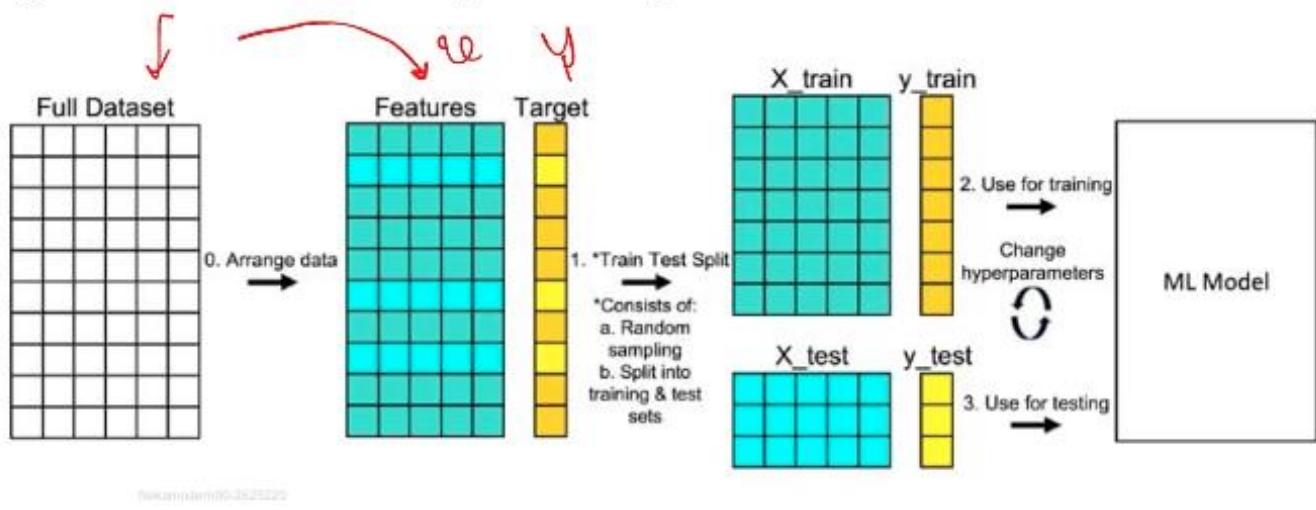
## Masalah Kritis: Bagaimana Kita Tahu Model Kita Bagus?



## Masalah Kritis: Bagaimana Kita Tahu Model Kita Bagus?



## Bagaimana Train-Test Split Bekerja



## Implementasi dengan Scikit-learn

### Ringkasan & Langkah Selanjutnya

#### Poin Kunci untuk Ditunggu

- Jangan pernah menguji model pada data yang digunakannya untuk belajar.  
Ini akan memberikan rasa percaya diri yang palsu dan model yang tidak berguna di dunia nyata.
- Train-Test Split adalah praktik standar industri untuk mendapatkan evaluasi yang jujur tentang kemampuan **generalisasi** model pada data baru.
- Lakukan pemisahan ini di **awal** alur kerja Anda, sebelum melakukan persiapan data lainnya, untuk mencegah *data leakage*.
- Gunakan `random_state` untuk memastikan hasil Anda dapat direproduksi.

**Selanjutnya:** Kita sekarang memiliki semua konsep dasar yang kita butuhkan. Di chapter berikutnya, kita akan mulai menerapkan alur kerja ini untuk **menyiapkan data** kita dengan benar (menangani teks, nilai yang hilang, dll.) dan kemudian **membangun model regresi pertama kita!**

*X\_train  
Y\_train*