

spacing = 3 tab

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

### Chapter 1. Descriptive Statistics

#### 1-1. Introduction To Descriptive Statics

01	02	03	04	05
<b>Apa Itu Statistik Deskriptif</b>	<b>Peran Statistik Deskriptif</b>	<b>Statistik Deskriptif vs Inferential</b>	<b>Contoh Aplikasi Statistik Deskriptif</b>	<b>Visualisasi Distribusi Data</b>

#### 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

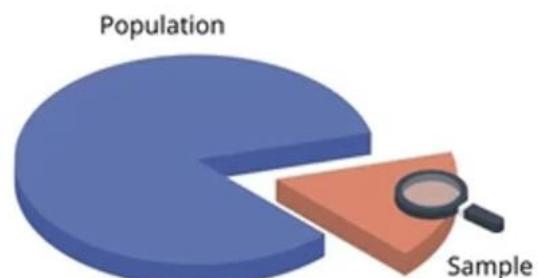


Menyediakan ringkasan numerik dan visual dari data



Dasar untuk analisis statistic lanjutan dan machine learning

## Statistik Deskriptif vs Inferensial



### Deskriptif:

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

### Inferensial:

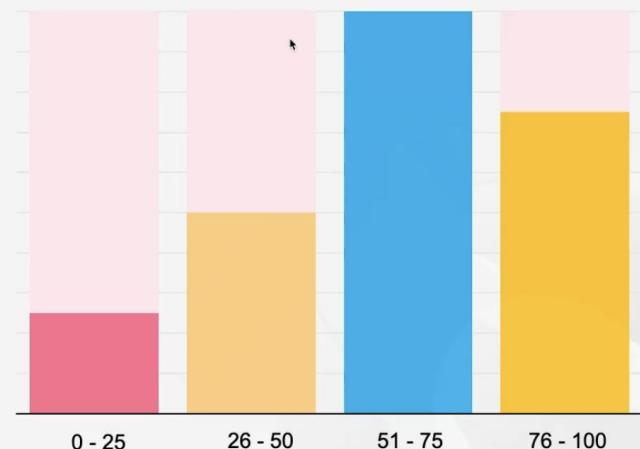
- Menggunakan sampel untuk menyimpulkan tentang populasi.
- Melibatkan estimasi dan pengujian hipotesis

# 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**  
fisikamodem00-2625220

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)

$$x \rightarrow 5 \ 7 \ 6 \ 2 \ 1 \ 4 \quad 40$$

$\uparrow$   
data  
 $n=6$

$$\begin{aligned}\bar{x} &= \frac{5+7+6+2+1+4}{6} + 40 \\ &= \frac{25}{6} \rightarrow \frac{65}{7} = 9,14... \\ &= 4,16666 \approx 4,17\end{aligned}$$

**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)

$$x \rightarrow 5 \ 7 \ 6 \ 2 \ 1 \ 4 \quad 40$$

$\uparrow$   
data  
 $n=6$

$$\begin{aligned}\bar{x} &= \frac{5+7+6+2+1+4}{6} + 40 \\ &= \frac{25}{6} \rightarrow \frac{65}{7} = 9,14... \\ &= 4,16666 \approx 4,17\end{aligned}$$

**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)

$X \rightarrow 1 \underline{2} 2 \underline{2} 3 \ 4$   
↓  
modus=2

$X \rightarrow 1 \underline{2} 2 \underline{2} 3 \underline{4} \underline{4} 4 \ 5$   
 $\begin{matrix} 1 & 2 & 2 & 3 & 4 & 4 & 5 & 6 & 6 & 6 \\ \swarrow & \swarrow \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{matrix}$   
Bimodal.  
multimodal

**Definisi:** Nilai yang sering muncul.

**Formula:**

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

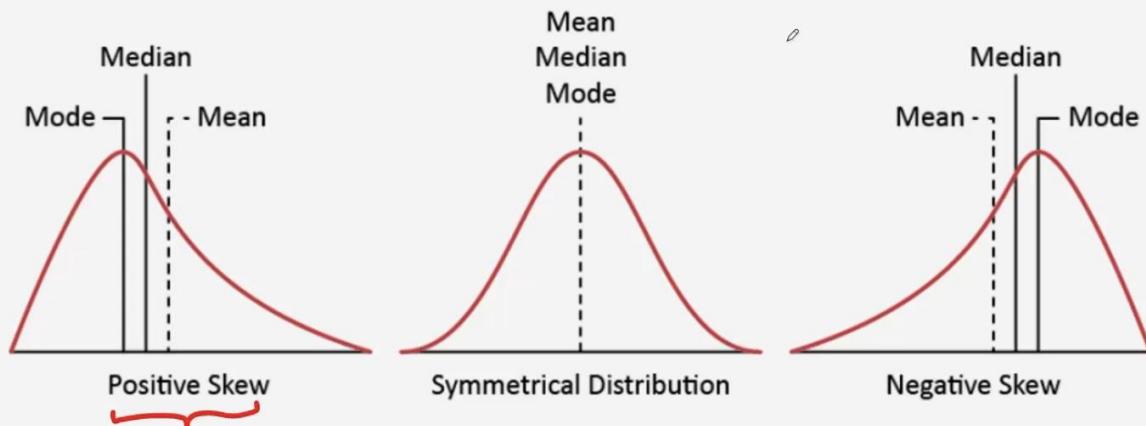
**Catatan:**

Kelebihan: Lebih cocok untuk data kategorikal.

Kekurangan: Tidak selalu ada atau tidak unik.

fisikamodem00-2621

## 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.

### 1-3. Measures Of Dispersion

01

Apa itu ukuran persebaran data?

02

Range (Jangkauan)

03

Variance (variansi)

04

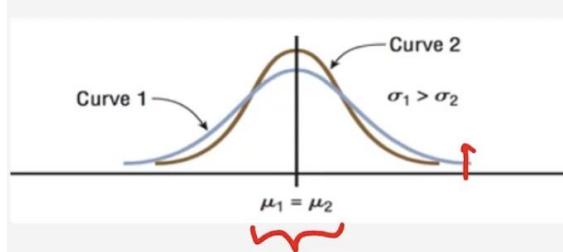
Standard Deviation (Simpangan Baku)

05

Interquartile Range (IQR)

# Apa itu ukuran persebaran data

12 3 4 5      2 3 1  
↑                ↑  
3                3



- **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.

## Range (Jangkauan)

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

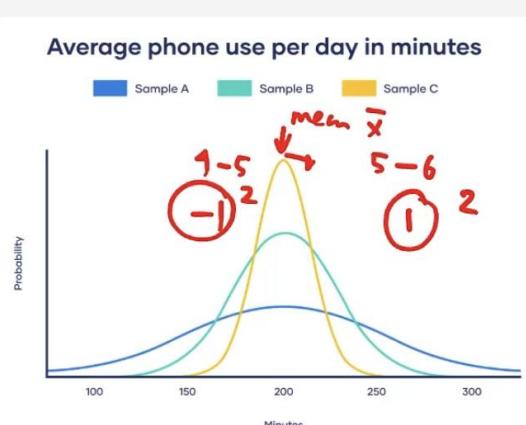
115

$$\text{range} = \max - \min$$
$$= 90 - 15$$
$$= 75$$

max      min

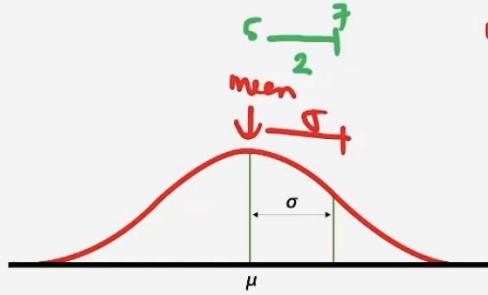
- **Definisi:** selisih antara nilai maksimum dan minimum dalam dataset.
- **Formula:**
$$\text{range} = \max - \min$$
- **Kelebihan:** Mudah dihitung
- **Kekurangan:** Sangat dipengaruhi oleh outlier

## Variance (Variansi / Ragam)



- **Definisi:** mengukur rata-rata kuadrat deviasi dari mean
- **Formula:**
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$
- **Kelebihan:** Menggunakan semua data
- **Kekurangan:** Satuan kuadrat, sulit diinterpretasi langsung

# Standard Deviation (Simpangan Baku)

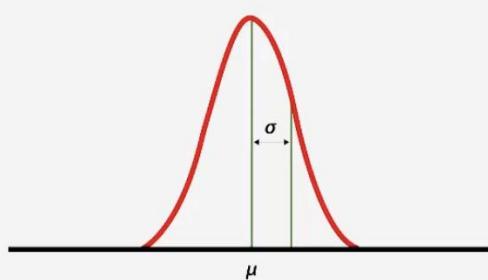


$$\text{mean} = \mu$$
$$\sigma = 2$$

- **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

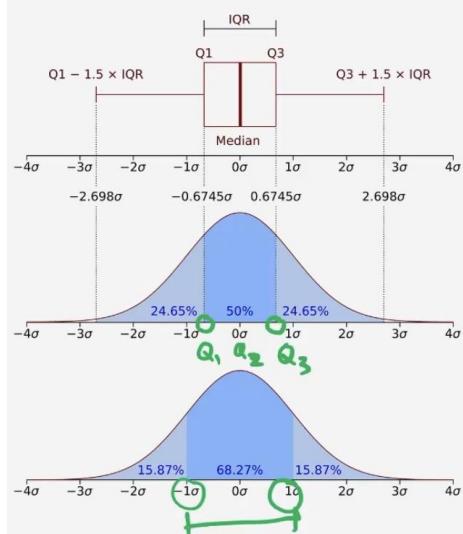


## Interquartile Range (IQR)



- **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

## 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

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

## Aplikasi dalam Data Science dan Statistics

<b>01</b> <b>Menilai Variabilitas Fitur</b>	Ukuran pemasatan 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

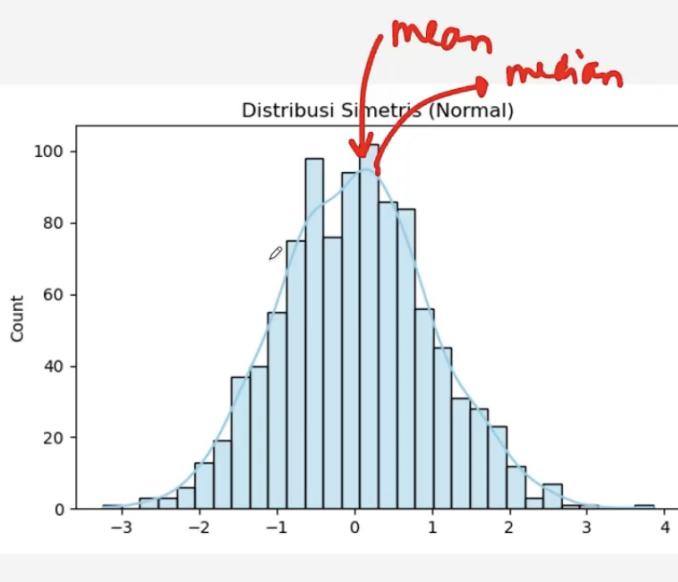
<b>01</b> <b>Definisi</b>	<b>02</b> <b>Jenis-jenis distribusi data</b>	<b>03</b> <b>Skewness &amp; Kurtosis</b>	<b>04</b> <b>Perbandingan Penutup</b>
------------------------------	---	---	--

# 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).

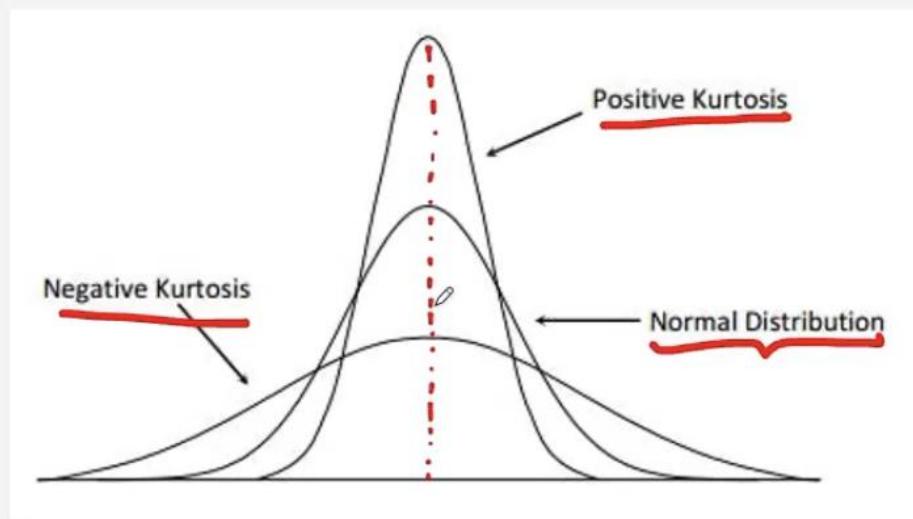
fisikamodem00-2625220

## Distribusi Simetris



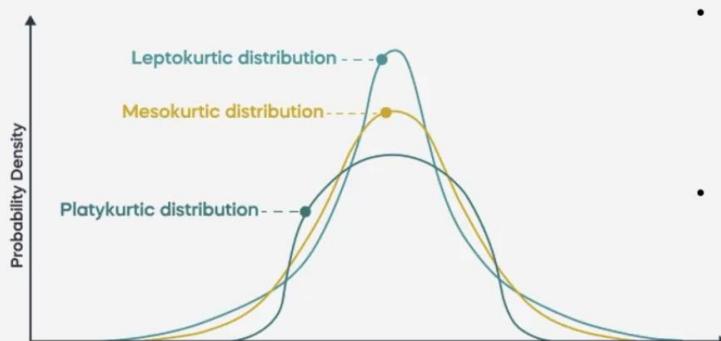
- Distribusi yang tersebar dengan titik median dan mean saling berdekatan.
- Contoh dari distribusi ini adalah **distribusi normal**

# Kurtosis



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

## Kurtosis

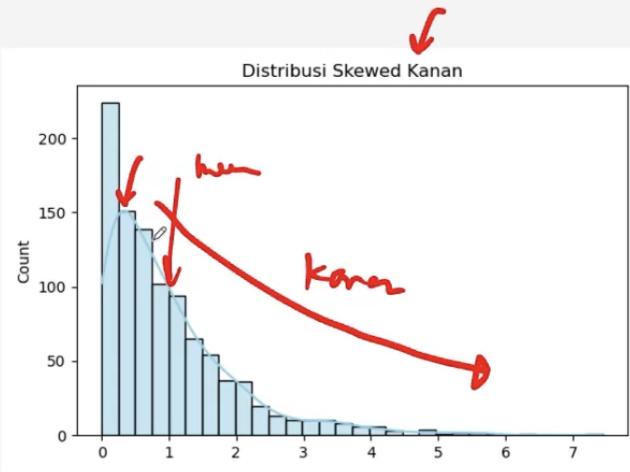


Menghitung kurtosis dengan python

```
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.

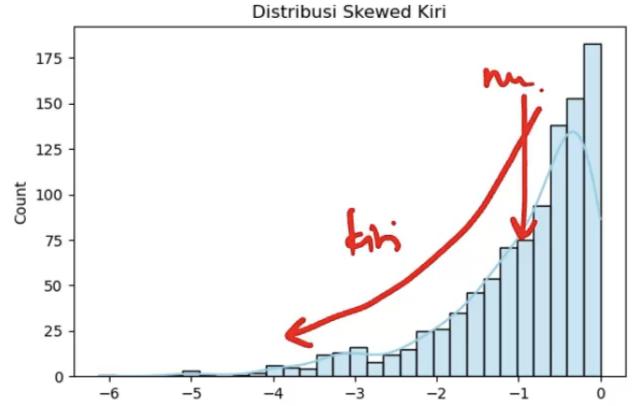
# Distribusi Miring (Skewed)



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

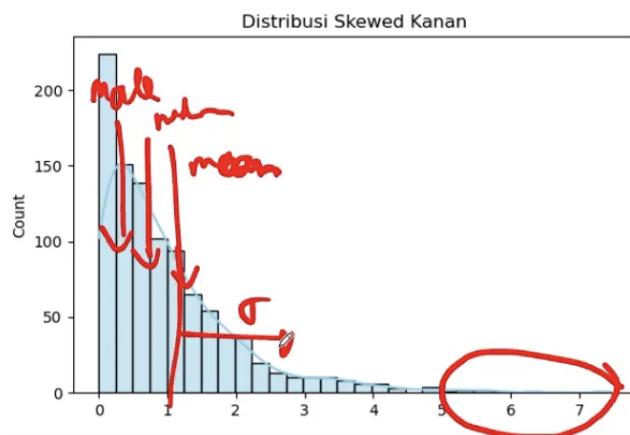


Skewness < 0: miring ke kiri

fisikamodern00-2625220

# Distribusi Miring (Skewed)

fisikamodern00-2625220

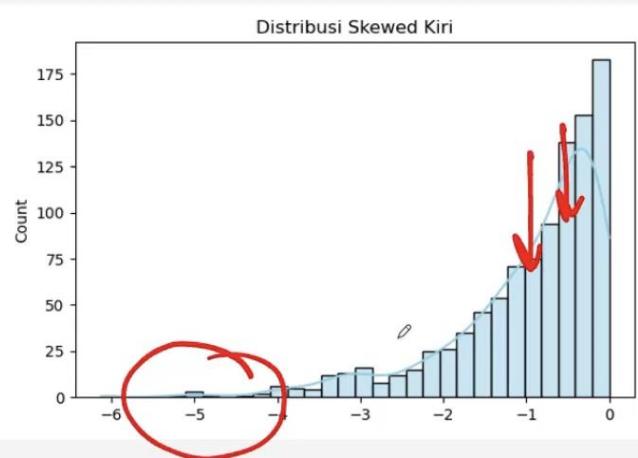


Skewness > 0: miring ke kanan

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)

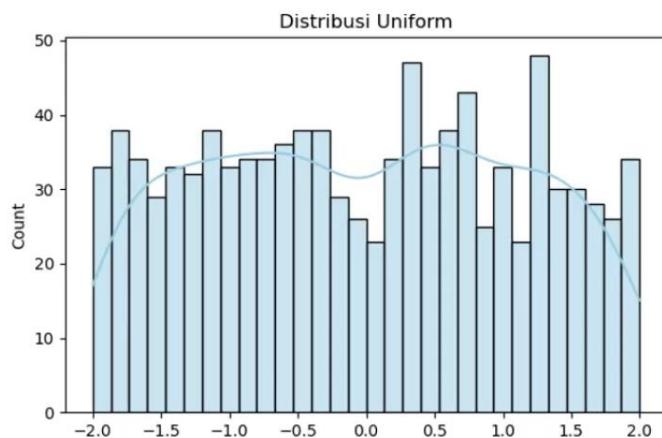


Skewness < 0: miring ke kiri

## 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

# Distribusi Uniform



Skewness < 0: miring ke kiri

## Karakteristik:

- Semua nilai memiliki probabilitas yang mirip
- Mean ≈ Median
- Contoh:
  - Hasil lemparan dadu fair → 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

# Penutup

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

## 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]

fisikamodern00-2625220

Hitung mean, median, modus, dan range

$$\text{mean} \rightarrow \bar{x} = \frac{4+8+6+5+3+4+1+7}{8} = \frac{41}{8} = 5,125$$

modus = 4

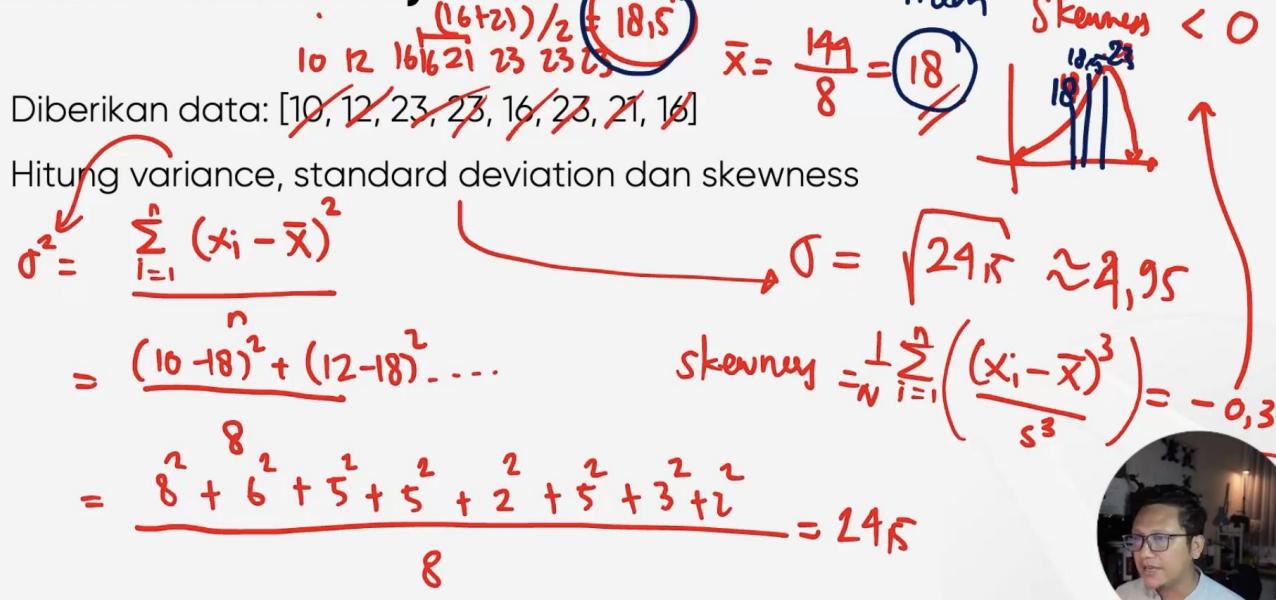
$$\text{range} = 8 - 3 = 5$$

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

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

A hand-drawn bell-shaped curve is shown below the data list.

## Latihan 2 – Ukuran Penyebaran



### 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**

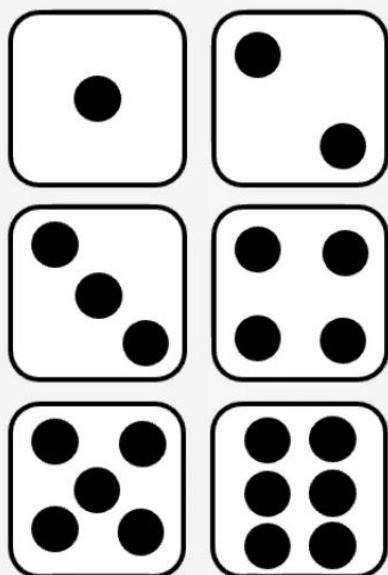
03

**Classical  
vs  
Empirical**

# Apa Itu Probabilitas?

- |           |   |
|-----------|---|
| <b>01</b> | • Probabilitas adalah ukuran kemungkinan suatu kejadian terjadi |
| <b>02</b> | • Bernilai 0 saat tidak terjadi.<br>• Bernilai 1 saat terjadi.  |
| <b>03</b> | • Digunakan dalam pengambilan keputusan berbasis data           |

## Contoh



Berapa probabilitas / kemungkinan dadu berangka 2 muncul saat dilempar?

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

↓ ↓ ↓ ↓ ↓ ↓

0 1 0 0 0 0

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

# Contoh



Berapa probabilitas / kemungkinan gambar muncul saat sekali tos?

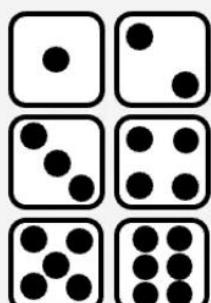
$$p(\text{gambar}) = \frac{1}{2}$$

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

(1)

## Sample Space / Ruang Sampel

- Sample Space adalah himpunan semua kemungkinan hasil



Sample Space nya adalah:

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

$\underbrace{\phantom{000}}_6$



Sample Space nya adalah:

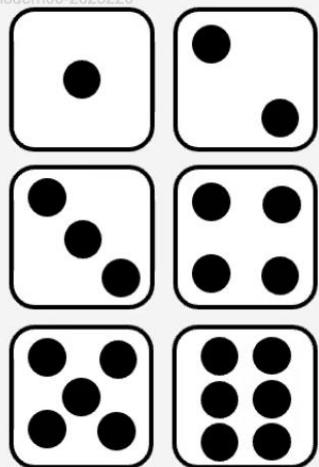
[ angka, gambar ]

$\underbrace{\phantom{00}}_2$

# 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 ]

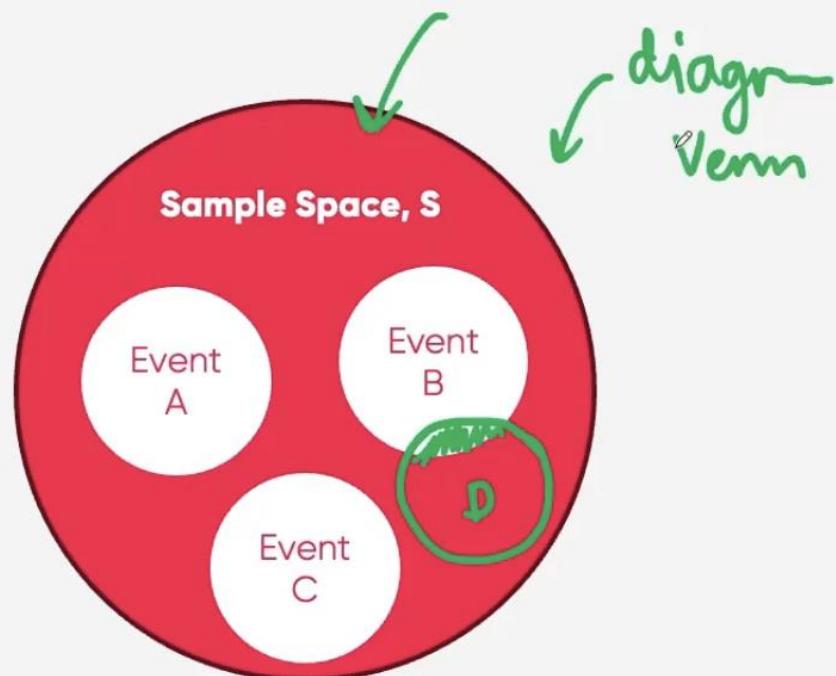
Event → Angka Genap:

[ 2, 4, 6 ]

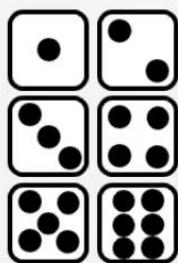
Event → Bilangan Prima:

[ 2, 3, 5 ]

## Sample Space 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:**

$$1 \ 3 \ 1 \ 4 \ 5 \ 6 \\ P(1) = \frac{2}{6} = \frac{1}{3}$$

## Kenapa Berbeda?

Probabilitas Teoritical akan sama dengan Probabilitas Empirical saat

jumlah ekperimen

**Jumlah ekperimen**

**sangatlah**

**BESAR!**



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

## Ringkasan

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

## 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

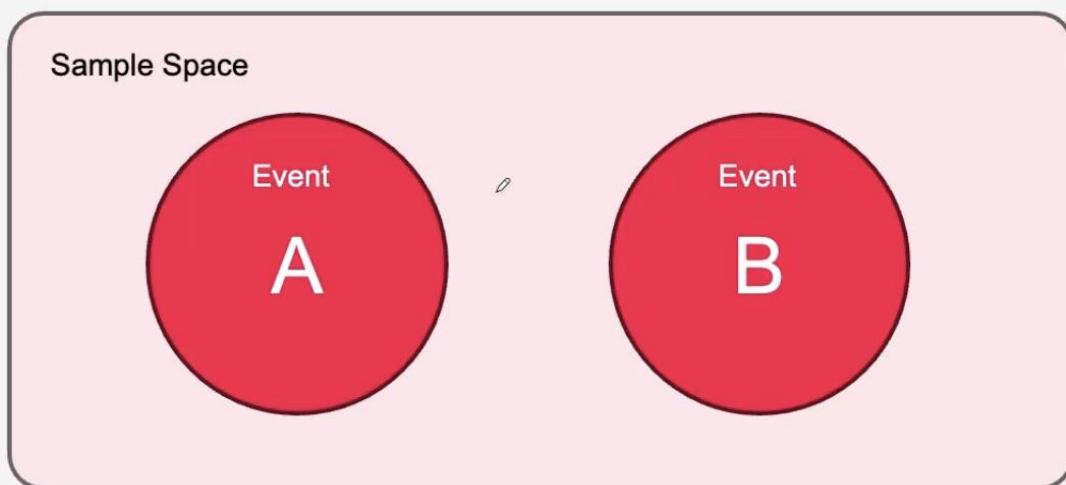


Diagram Venn

fisikamodern00\_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

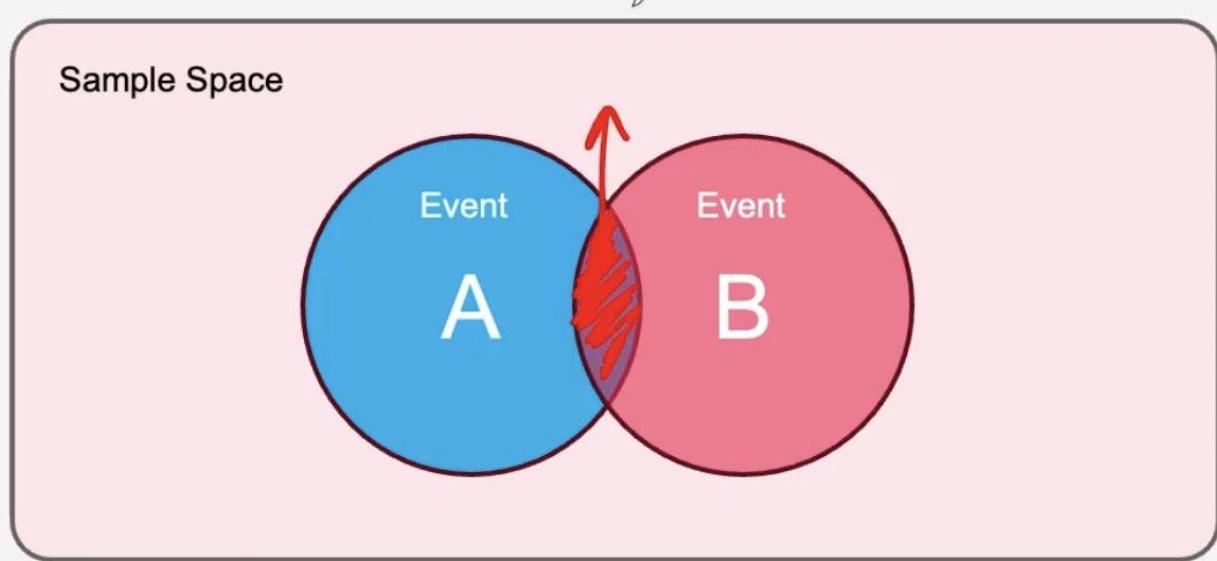
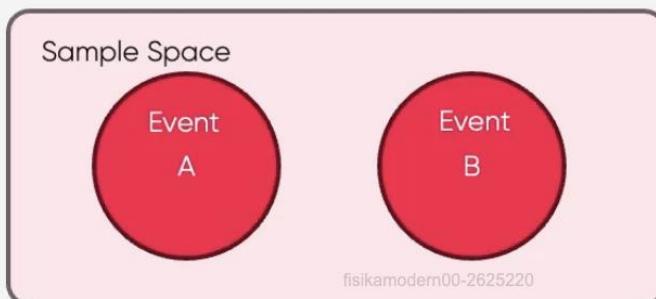


Diagram Venn

# Aturan Penjumlahan (Mutually Exclusive)

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



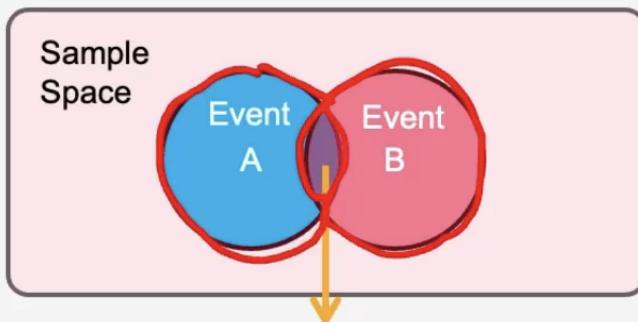
$$\begin{aligned}P(A \text{ atau } B) &= P(A) + P(B) \\P(1 \text{ atau } 6) &= P(1) + P(6) \\&= \frac{1}{6} + \frac{1}{6} \\&= \frac{2}{6} = \frac{1}{3}\end{aligned}$$

- 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



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

- Formulasi:

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

union .

# 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

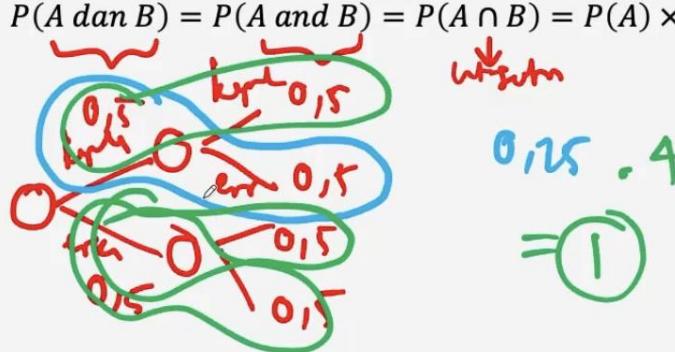
Event  
B

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

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

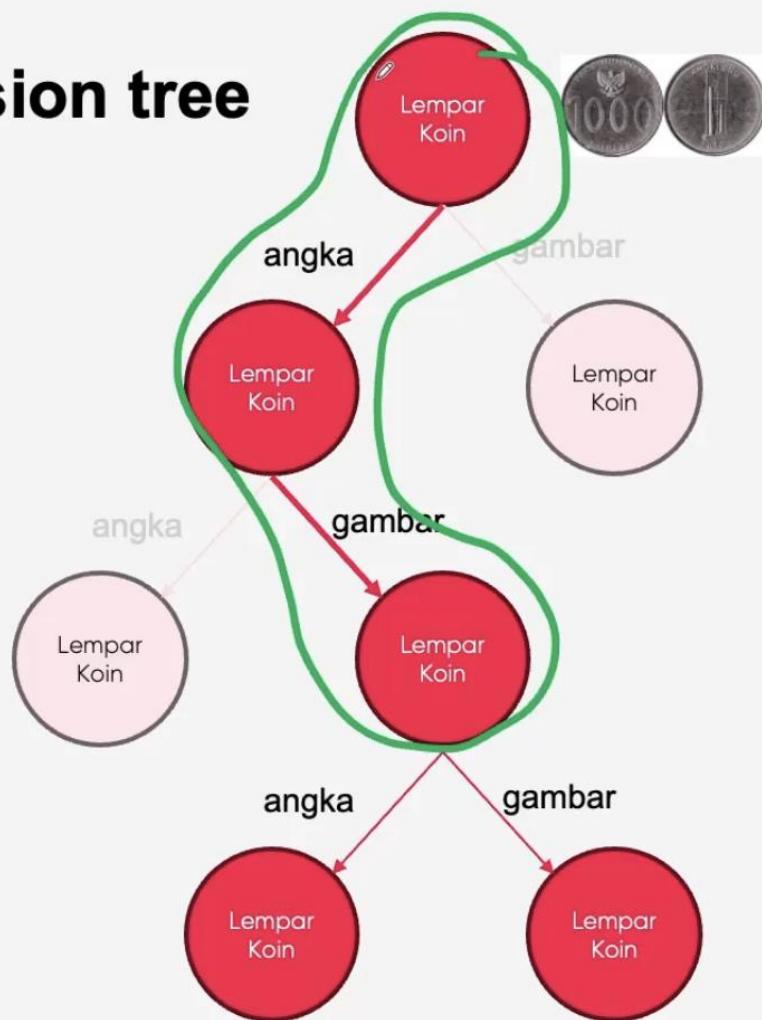
# Aturan Perkalian (Independent)

- 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:
    - $P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = 0.5 \times 0.5 = 0.25$



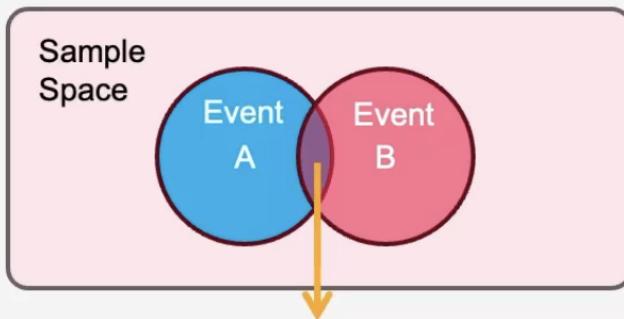
fisikamodern00-2625220

## Ilustrasi decision tree Independent



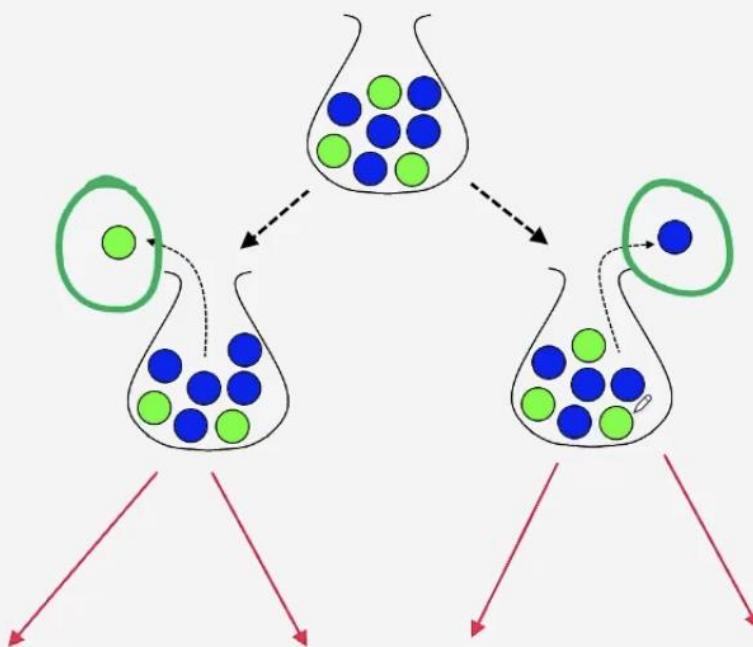
# Aturan Perkalian (Non Mutually Exclusive)

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



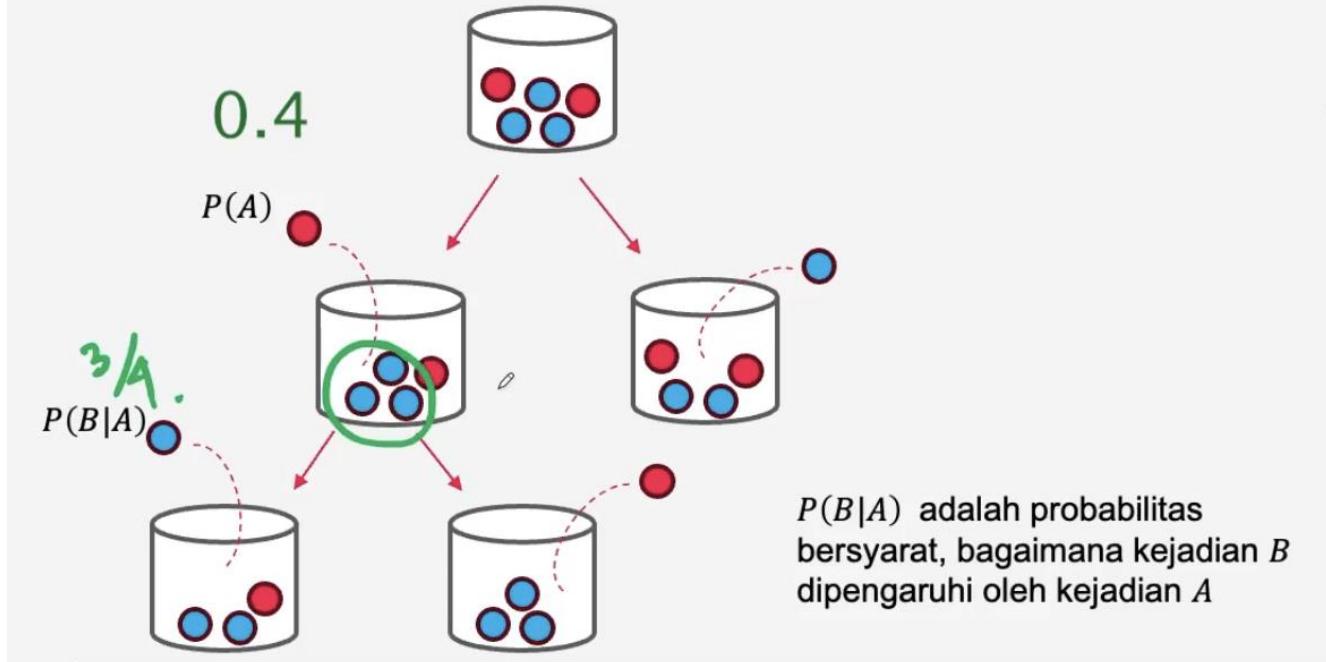
$$P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) \neq 0$$

## Ilustrasi decision tree - Dependent



## Probabilitas Bersyarat

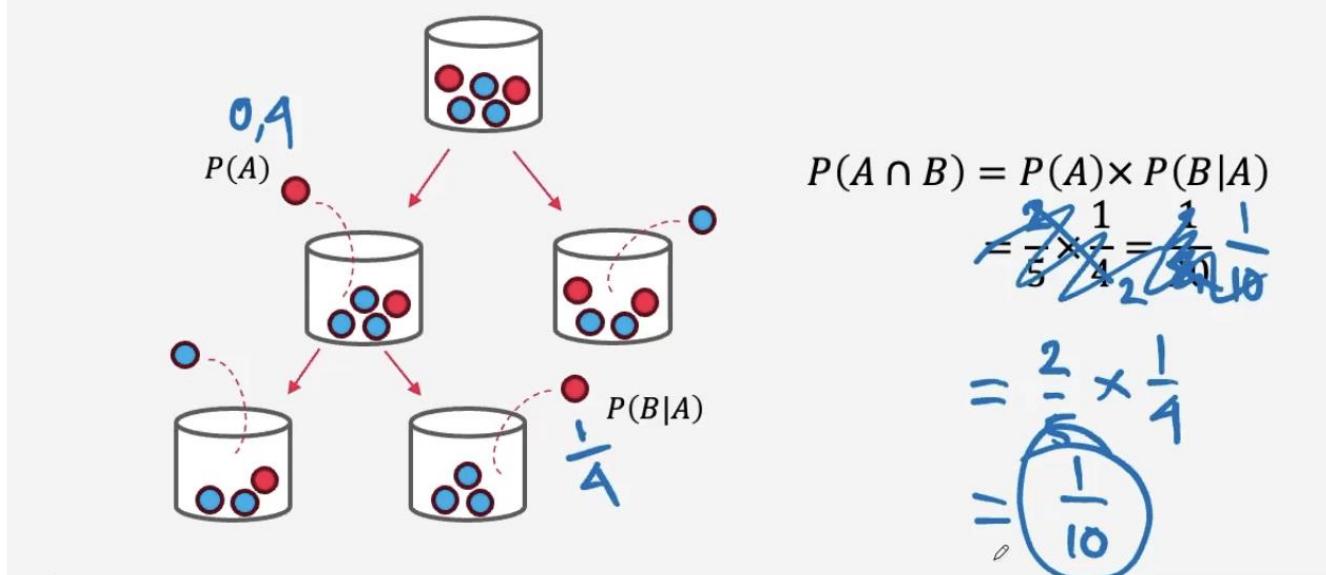
- Sebuah kantong berisi 2 kelereng merah dan 3 kelereng biru
  - Kejadian A : Mengambil kelereng merah pada pengambilan pertama
    - $P(A) = \frac{2}{5} = 0.2$
  - Kejadian B : Mengambil kelereng biru pada pengambilan kedua
    - $P(B|A) = \frac{\text{jumlah biru}}{\text{kelereng sisa}} = \frac{3}{4}$



$P(B|A)$  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
    - $P(B|A) = \frac{\text{jumlah merah}}{\text{kelereng sisa}} = \frac{1}{4}$



## 2-3. Conditional Probability & Independence

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

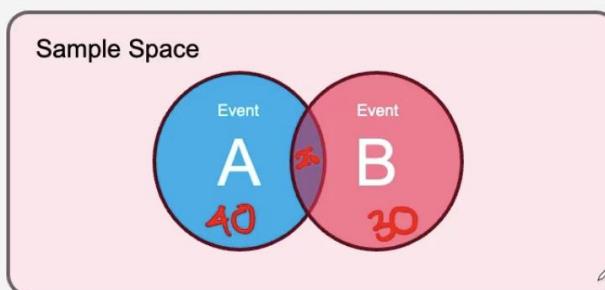
- Dari 100 pasien, 40 demam, 30 sakit flu, dan 25 keduanya

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

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

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

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



$$\begin{aligned} P(\text{sakit} | \text{demam}) &= \frac{P(\text{sakit} \cap \text{demam})}{P(\text{demam})} \\ &= \frac{0.25}{0.4} \\ P(\text{demam} | \text{sakit}) &= \frac{P(\text{sakit} \cap \text{demam})}{P(\text{sakit})} \\ &= \frac{0.25}{0.3} = 0.833 \end{aligned}$$

# Contingency Table

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

$$P(\text{Sakit} | \text{Demam}) = \frac{15}{40} = 0,375$$
$$P(\text{Sakit} | \text{Demam}) = \frac{P(\text{Sakit} \cap \text{Demam})}{P(\text{Demam})}$$
$$= \frac{25}{40}$$
$$= 0,625$$

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

$$P(\text{Demam} | \text{Sakit}) = \frac{25}{30} = 0,833$$

fisikamodern00-2625220

## Mendeteksi Kejadian Independent

fisikamodern00-2625220

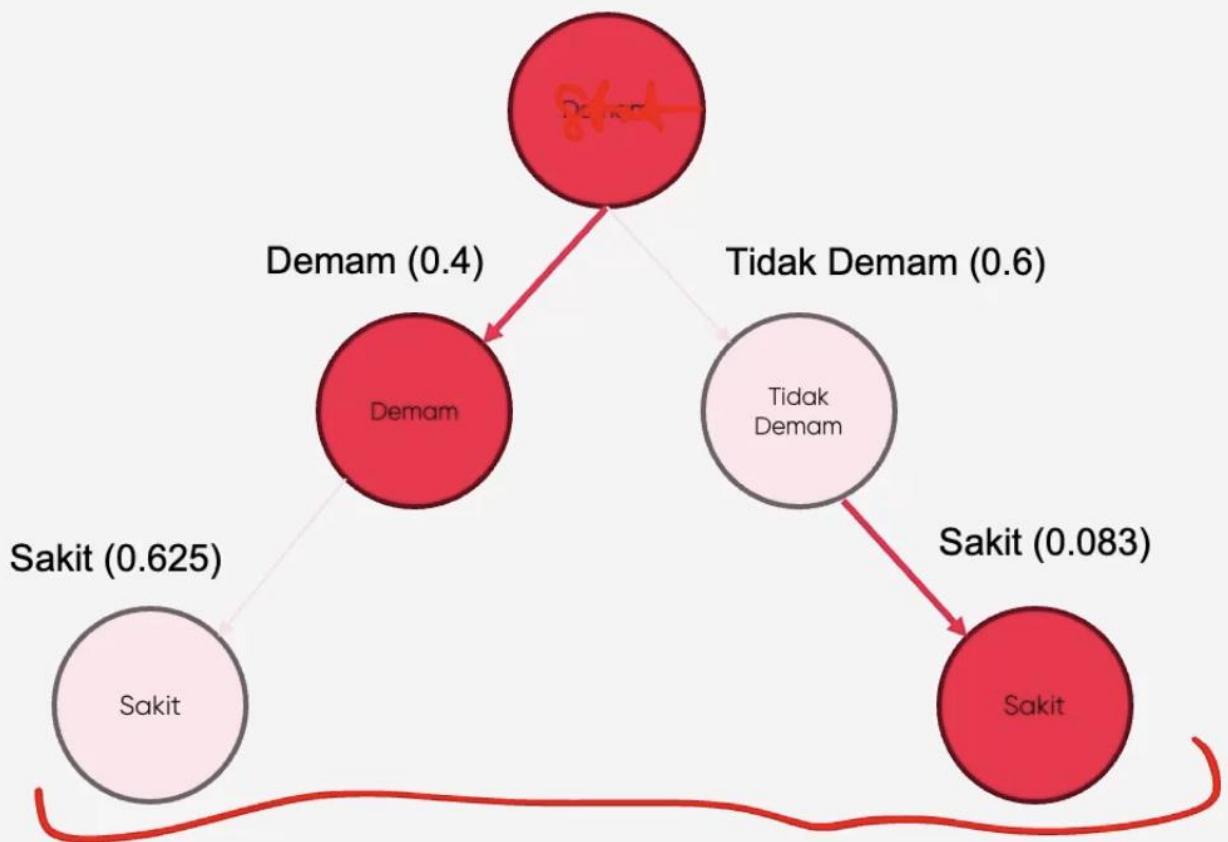
- Dua kejadian A dan B dinyatakan independent jika

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

- Atau secara ekuivalen

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

# Ilustrasi decision tree



$$\text{Probabilitas total sakit} = (0.4 \times 0.625) + (0.6 \times 0.083) = 0.25 + 0.05 = 0.30$$

## 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

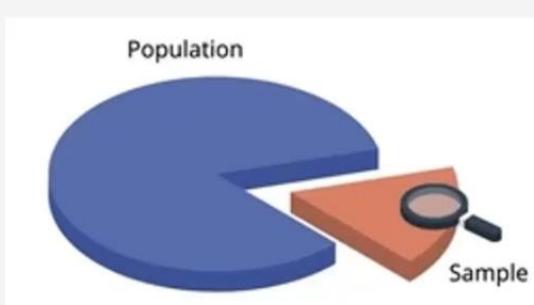
**Menerapkan Teorema Bayes pada kasus nyata**

04

**Menghubungkan Teorema Bayes dengan algoritma machine learning**

## Pengantar

- Bayes Theorem adalah alat penting dalam inferensi statistik

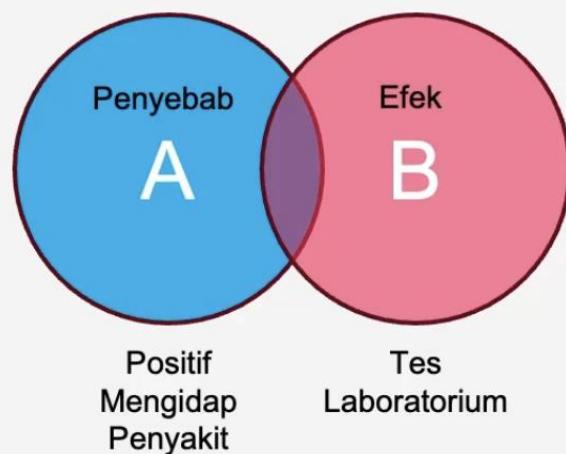


### Inferensial:

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

# Pengantar

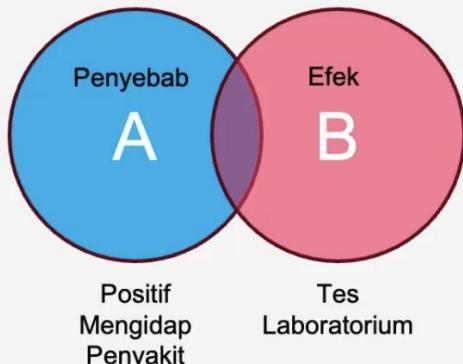
- Secara sederhana. Teorema Bayes memberikan cara matematis untuk membalikkan probabilitas bersyarat.
- Memungkinkan untuk menemukan probabilitas suatu "penyebab" ( $A$ ) jika kita mengetahui "efek" ( $B$ ) telah terjadi.



$P(B|A)$  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

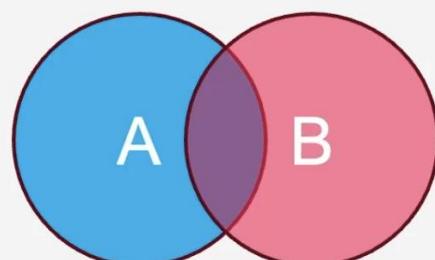
$$P(A \cap B) = P(A) \times P(B|A)$$

Probabilitas sakit dan tes positif =

Perprobabilitas tes positif  $\times$  Sakit jika hasil tes positif

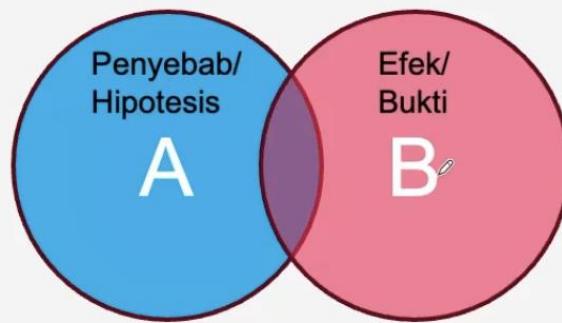
$$P(A \cap B) = P(B) \times P(A|B)$$

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



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

# Teorema Bayes



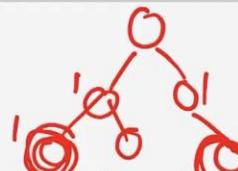
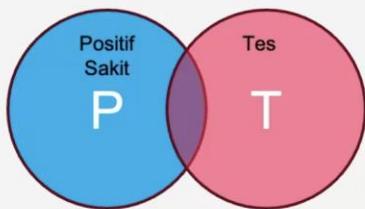
$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

Prior  $\downarrow$   
 bukti.  
 kira  
 kira  
 wkt  
 banting.

- $P(B|A)$  : **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

Essential Mathematical Foundation – Probability Theory – Bayes' Theorem Explained

## Contoh: Diagnosis Medis



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

$$P(p) = 1\% = 0,01$$

$$P(p_c) = 99\% = 0,99$$

$$P(t) = P(T|p) \cdot P(p)$$

$$= 0,99 \cdot 0,01 + 0,05 \cdot 0,99$$

$$= 0,06 \cdot 0,99 = 0,0594$$

$$P(T|p) = 99\% \rightarrow P(T_c|p) = 1\%$$

$$P(T_c|p_c) = 95\% \rightarrow P(t|p_c) = 5\%$$

$$+ P(t|p_c) \cdot P(p_c) \cdot P(p|t)$$

$$= 0,05 \cdot 0,05 \cdot 0,99$$

$$= 0,0594$$

$$\begin{aligned}
 P(p|t) &= \frac{P(t|p) \cdot P(p)}{P(t)} \\
 &= \frac{99\% \cdot 1\%}{0,0594} \\
 &= 0,16667 \rightarrow 16,67\%
 \end{aligned}$$



## 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.

$$P(A|b_1, b_2, \dots, b_n) = \frac{P(A) \times P(b_1, b_2, \dots, b_n|A)}{P(b_1, b_2, \dots, b_n)}$$

← class  
← fitur

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

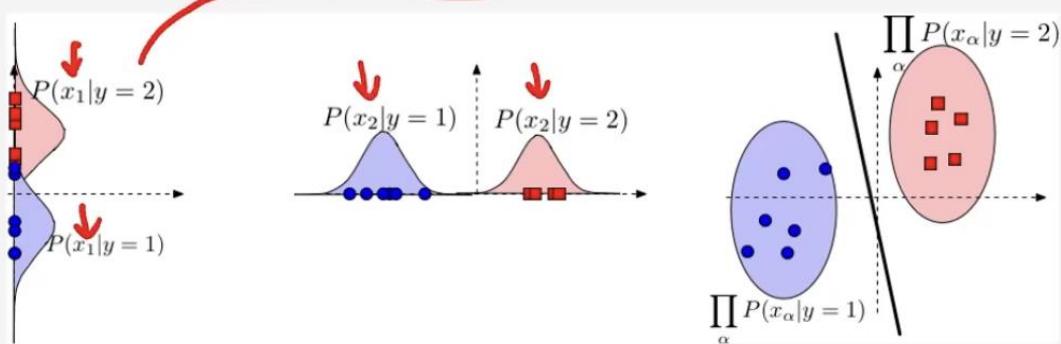
$$\begin{aligned} P(b_1, b_2, \dots, b_n|A) &= P(b_1|A) \times P(b_2|A) \times P(b_3|A) \times \dots \times P(b_n|A) \\ &= \prod_{i=1}^n P(b_i|A) \end{aligned}$$

## 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