Statistical Machine Learning & Data Storytelling

spacing = 3 tab

1. Part 1 Essential Mathematical Foundations (Deep Live)

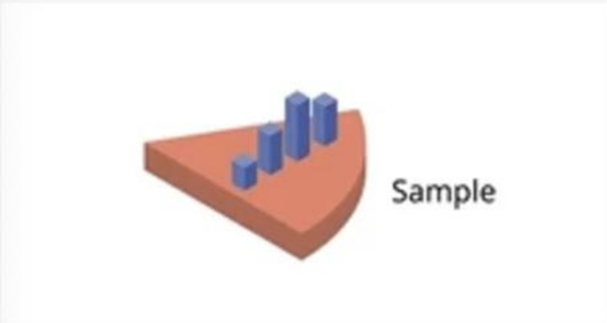
Chapter 1. Descriptive Statistics

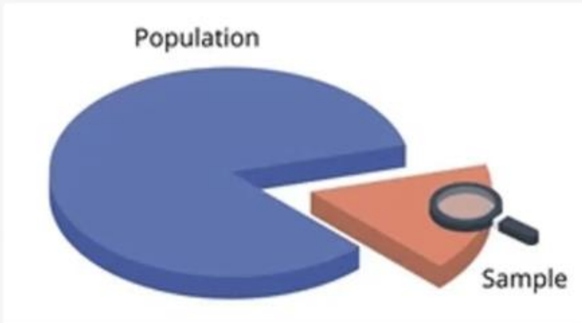
1-1. Introduction To Descriptive Statics

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

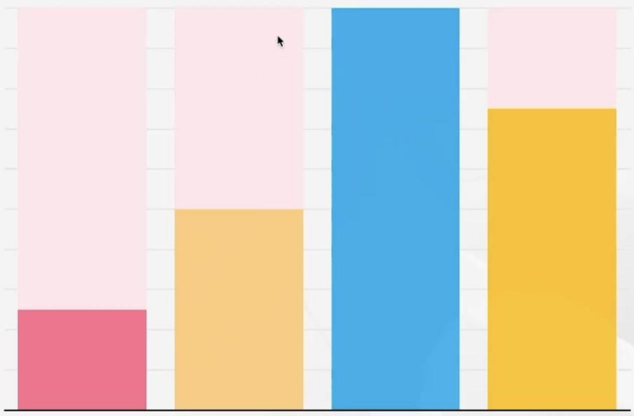
Inferensial:

- 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**

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)

$x \rightarrow \underbrace{5\ \ 7\ \ 6\ \ 2\ \ 1\ \ 4}_{n=6}\ \ 40$

data

$\bar{x} = \dfrac{5+7+6+2+1+4}{\cancel{6}\ 7} + 40$

$= \dfrac{25}{6} \rightarrow \dfrac{65}{7} = 9{,}14\ldots$

$= 4{,}16666 \approx 4{,}17$

Definisi: Jumlah seluruh nilai dibagi jumlah data.

Formula:

nilai

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

jumlah data

Catatan:

Kelebihan: Menggunakan semua data, mudah dihitung.

Kekurangan: Sangat sensitif terhadap outlier.

Mean (Rata-rata)

$x \rightarrow \underbrace{5\ \ 7\ \ 6\ \ 2\ \ 1\ \ 4}_{n=6}\ \ 40$

data

$\bar{x} = \dfrac{5+7+6+2+1+4}{\cancel{6}\ 7} + 40$

$= \dfrac{25}{6} \rightarrow \dfrac{65}{7} = 9{,}14\ldots$

$= 4{,}16666 \approx 4{,}17$

Definisi: Jumlah seluruh nilai dibagi jumlah data.

Formula:

nilai

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

jumlah data

Catatan:

Kelebihan: Menggunakan semua data, mudah dihitung.

Kekurangan: Sangat sensitif terhadap outlier.

Modus (Mode)

X → 1 $\underbrace{2\ 2\ 2}$ 3 4

↓

modus = 2

X → 1 $\underbrace{2\ 2\ 2}$ 3 $\underbrace{4\ 4\ 4}$ 5

↓

Bimodal.

1 2 2 2 3 4 4 4 5 6 6 6

3 3 3

multimodal

Definisi: Nilai yang sering muncul.

Formula:

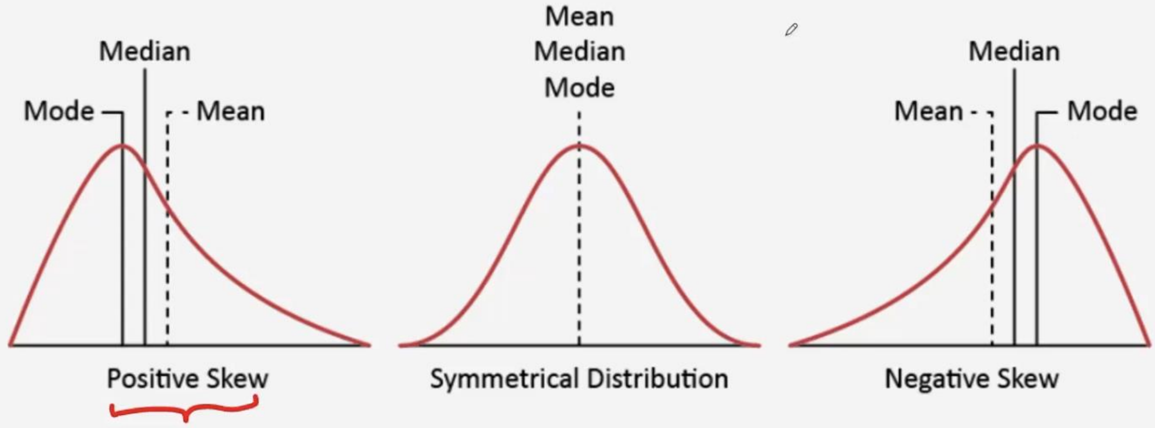
- Frekuensi / kemunculan data tertinggi
- Dapat memiliki lebih dari satu modus (bimodal/multimodal)

Catatan:

Kelebihan: Lebih cocok untuk data kategorikal.

Kekurangan: Tidak selalu ada atau tidak unik.

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)**

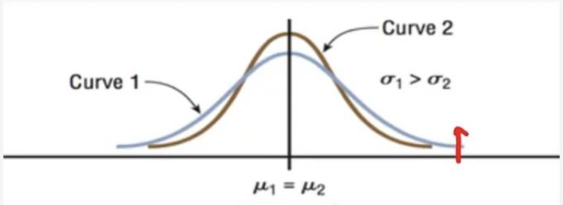
04 **Standard Deviation (Simpangan Baku)**

05 **Interquartile Range (IQR)**

Apa itu ukuran persebaran data

12345 234

3 3



sama

- **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)

115

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

max

$$range = \max - min$$
$$= 90 - 15$$
$$= 75$$

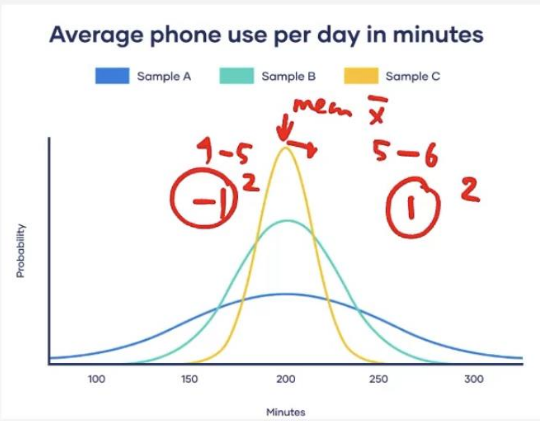
min

- **Definisi**: selisih antara nilai maksimum dan minimum dalam dataset.
- **Formula**:

$$range = max - min$$

- **Kelebihan**: Mudah dihitung
- **Kekurangan**: Sangat dipengaruhi oleh outlier

Variance (Variansi / Ragam)

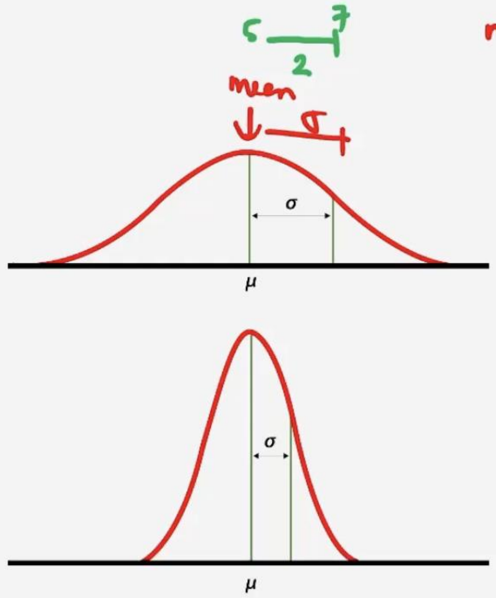


- **Definisi**: mengukur rata-rata kuadrat deviasi dari mean
- **Formula**:

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

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

Standard Devitation (Simpangan Baku)



men = 5

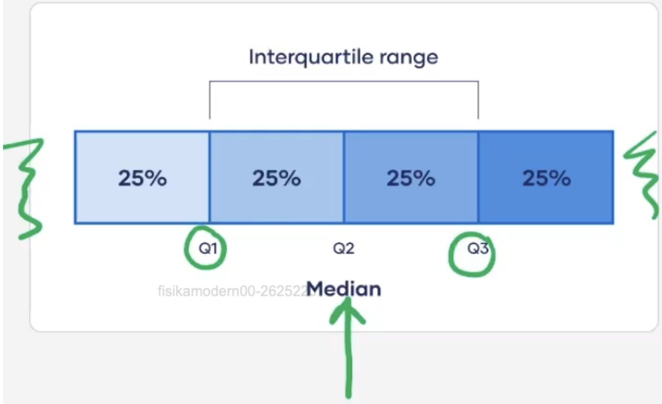
σ = 2

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

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

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

Interquartile Range (IQR)

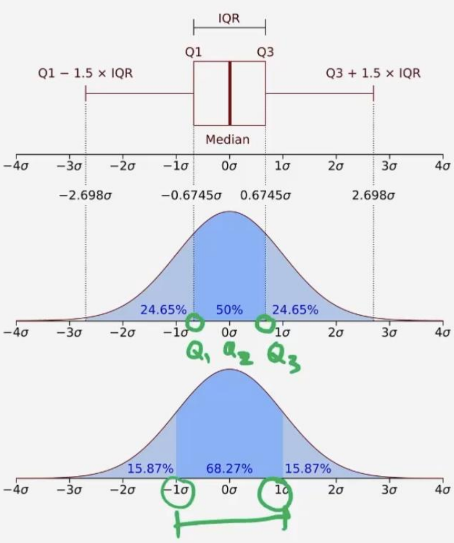


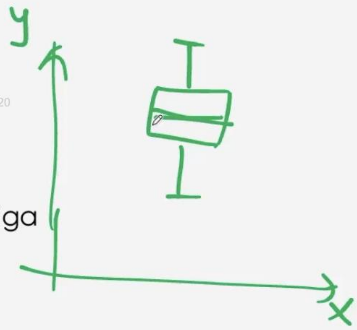
- **Definisi**: selisih antara kuartil ketiga (Q_3) dan kuartil pertama (Q_1).
- **Formula**:

$$IQR = Q_3 - Q_1$$

- **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

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

Aplikasi dalam Data Science dan Statistics

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

1-4. Understanding Data Distribution Shapes

01 **Definisi**

02 **Jenis-jenis distribusi data**

03 **Skewness & Kurtosis**

04 **Perbandingan**

05 **Penutup**

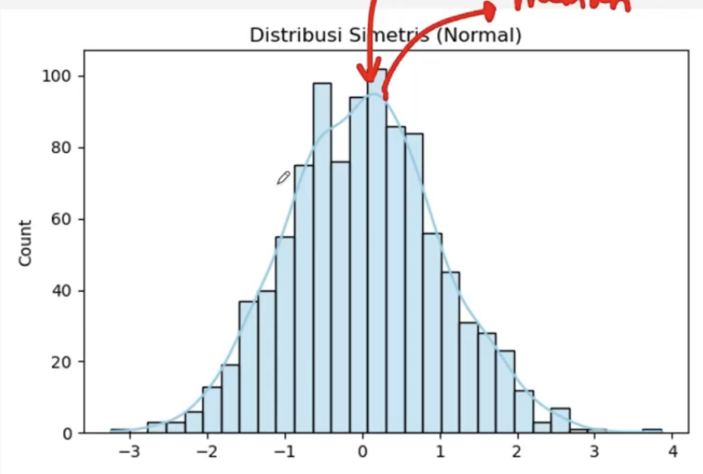
Bentuk Distribusi Data

- 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).

Distribusi Simetris

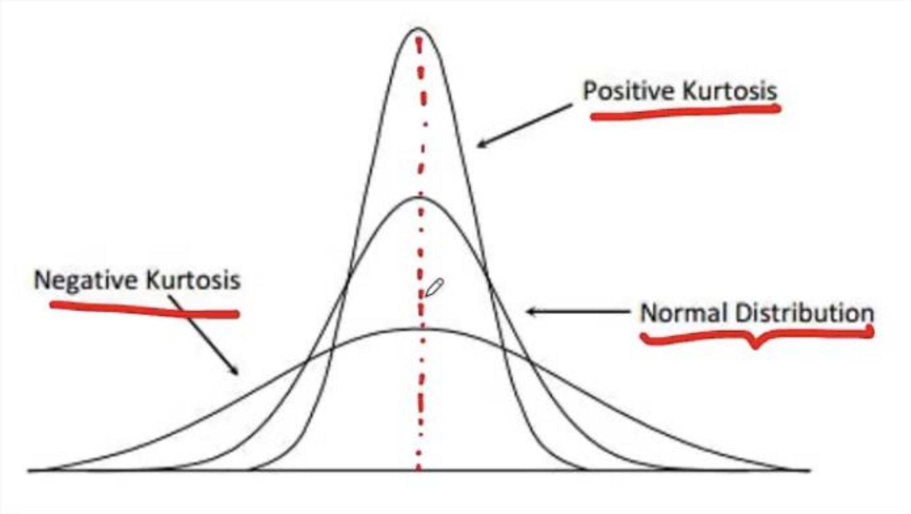
mean

median



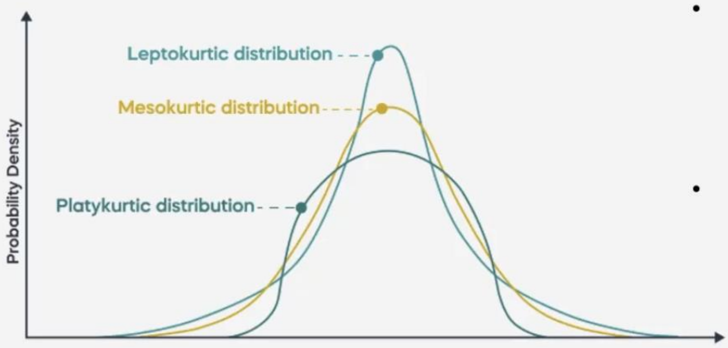
- Distribusi yang tersebar dengan titik **median** dan **mean** saling **berdekatan.**
- Contoh dari distribusi ini adalah **distribusi normal**

Kurtosis



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

Kurtosis



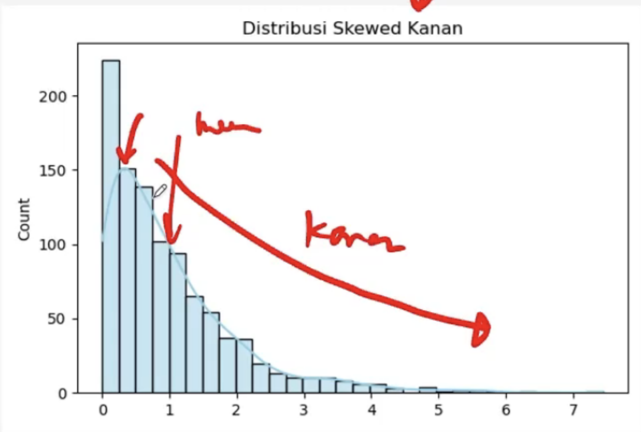
library

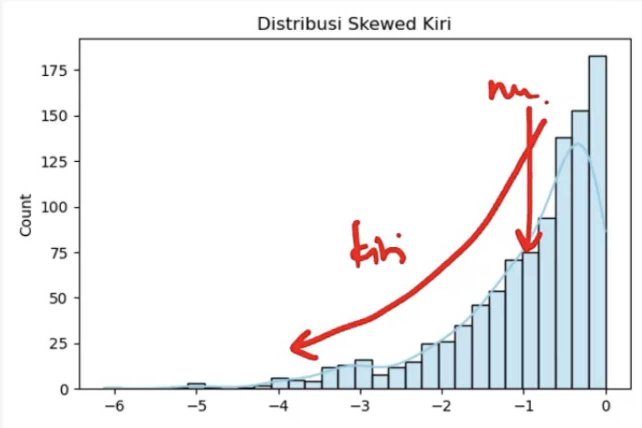
Menghitung kurtosis dengan python

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

- Mesokurtic (Kurtosis ≈ 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)





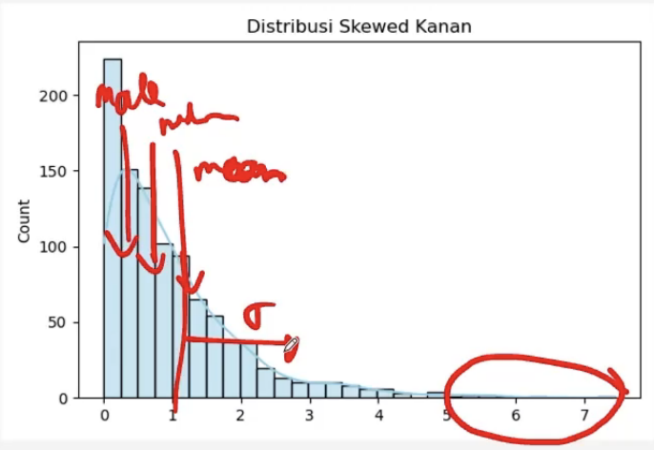
Skewness > 0: miring ke kanan

Skewness < 0: miring ke kiri

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

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

Distribusi Miring (Skewed)

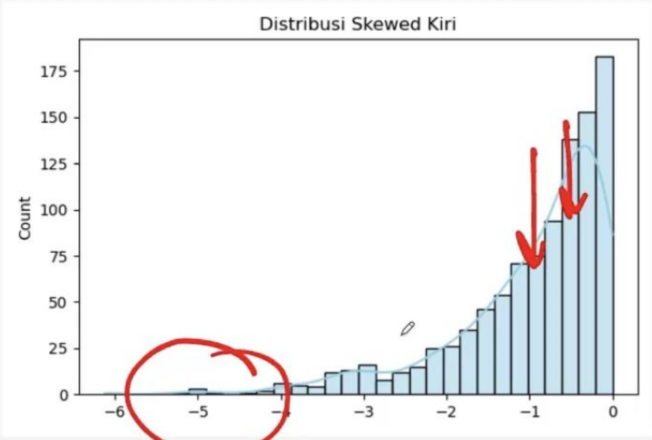


Skewness > 0: miring ke kanan

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)

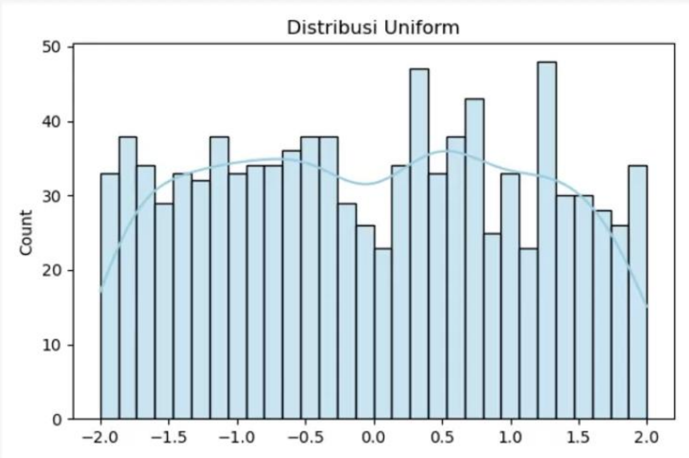


Skewness < 0: miring ke kiri

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

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 variansinya.
  - $Var(x) = \frac{(b-a)^2}{12}$, 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, 4, 7]

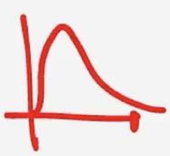
Hitung mean, median, modus, dan range

mean → $\bar{x} = \frac{4+8+6+5+3+4+4+7}{8}$

$= \frac{41}{8} = 5,125$

median → 3 4 4 $\underbrace{4\ 5}$ 6 7 8

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



modus = 4

range = 8 − 3

= 5

Latihan 2 – Ukuran Penyebaran

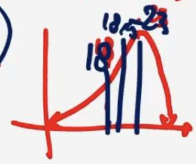
median: $\frac{(16+21)}{2} = 18,5$

modus = 23

mean: $\bar{x} = \frac{144}{8} = 18$

10 12 16 16 21 23 23 23

Skewness < 0



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

Hitung variance, standard deviation dan skewness

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

$$= \frac{(10-18)^2 + (12-18)^2 \ldots}{8}$$

$$= \frac{8^2 + 6^2 + 5^2 + 5^2 + 2^2 + 5^2 + 3^2 + 2^2}{8} = 24,5$$

$$\sigma = \sqrt{24,5} \approx 4,95$$

$$\text{skewness} = \frac{1}{N} \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})^3}{s^3} \right) = -0,3$$

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

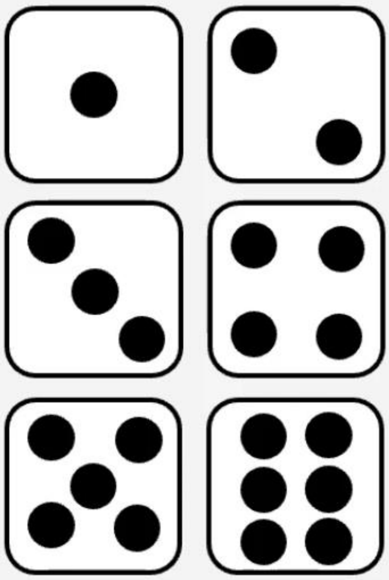
2-1. What Is Probability

01	02	03
Definisi Probability	**Sample Space Dan Event**	**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 dadu berangka 2 muncul saat dilempar?

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

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

0 1 0 0 0 0

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

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

$$\vdots$$

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

1

Contoh

angka ↓ gambar ↓

Berapa probabilitas / kemungkinan gambar muncul saat sekali tos?

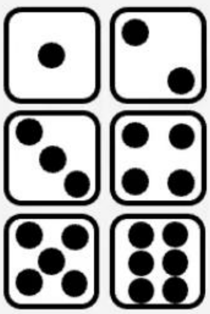
$$P(\text{gambar}) = \frac{1}{2}$$

$$P(\text{angka}) = \frac{1}{2} \quad +$$

$$1$$

Sample Space / Ruang Sampel

- Sample Space adalah himpunan semua kemungkinan hasil



Sample Space nya adalah:

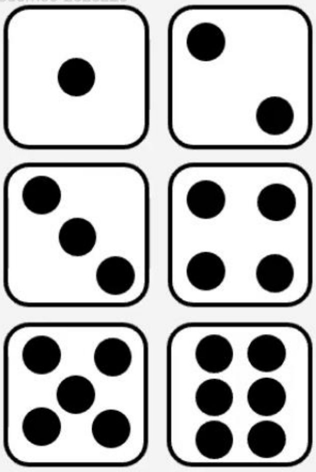
$\underbrace{[\,1, 2, 3, 4, 5, 6\,]}_{6}$

Sample Space nya adalah:

$\underbrace{[\,\text{angka, gambar}\,]}_{2}$

Event / Kejadian

- Event atau kejadian adalah subset dari sample space



Sample Space nya adalah:

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

Event → Angka Ganjil:

[1, 3, 5]

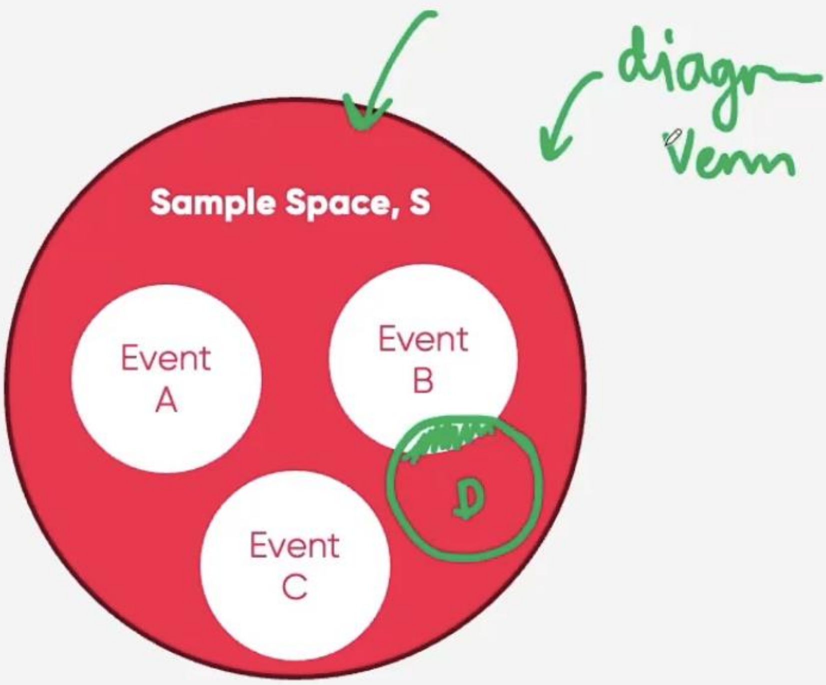
Event → Angka Genap:

[2, 4, 6]

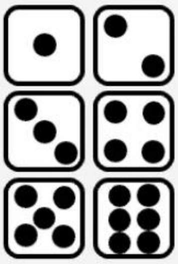
Event → Bilangan Prima:

[2, 3, 5]

Sample Space dan Event



Classical VS Empirical



Jumlah sampel adalah 6

Kemungkinan angka 3 muncul adalah?

Classical / Theoretical:

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

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 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

2-2. Basic Probability Rules

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

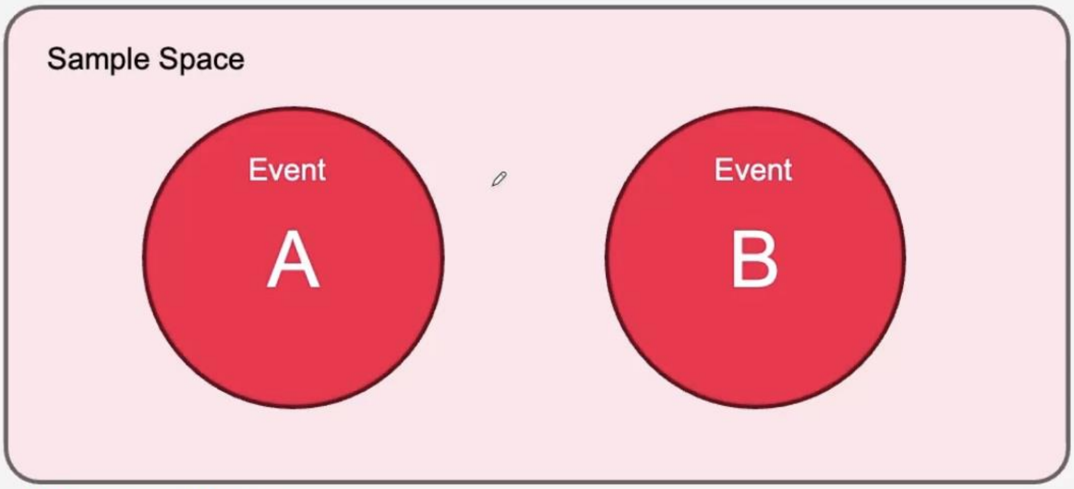


Diagram Venn

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

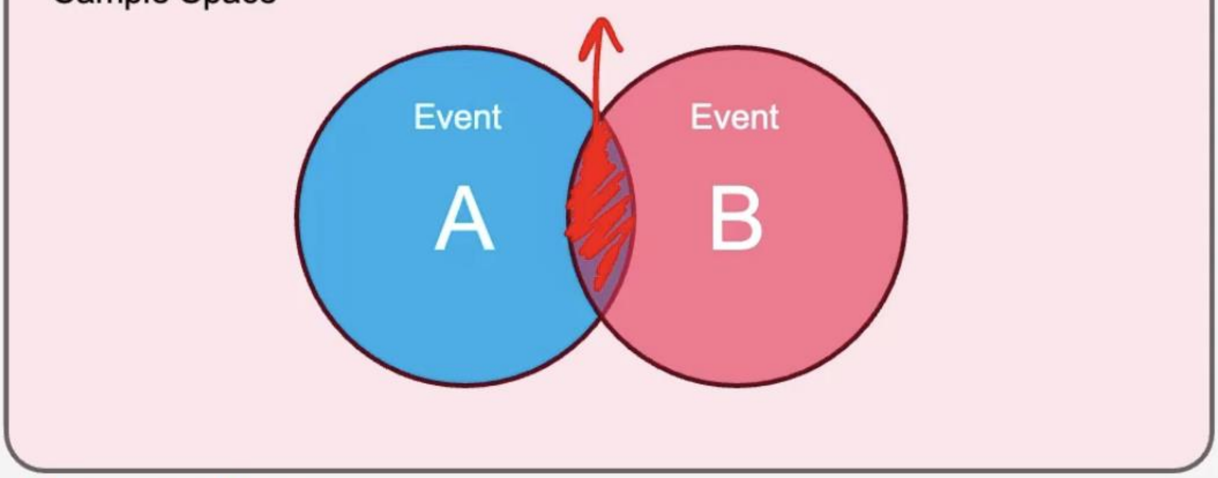
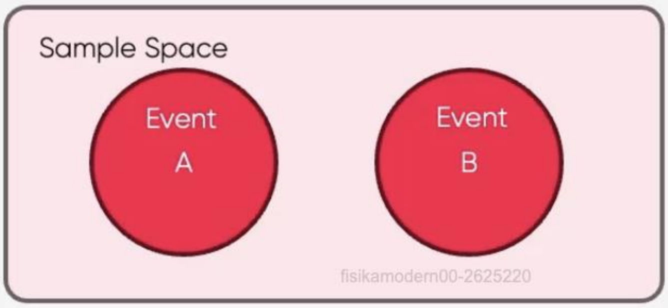


Diagram Venn

Aturan Penjumlahan (Mutually Exclusive)

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



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

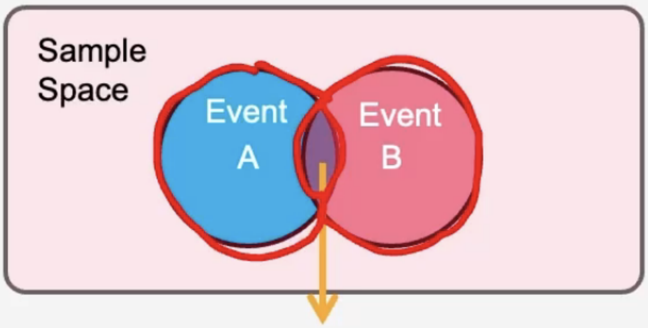
- Formulasi:

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

NON

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:

$$P(A \text{ atau } B) = P(A \text{ or } B) = P(A \cup B)$$
$$= P(A) + P(B) - \underbrace{P(A \cap B)}$$

Dependent vs Independent Events

Dependent

Kejadian yang hasilnya mempengaruhi kejadian selanjutnya

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

Independent

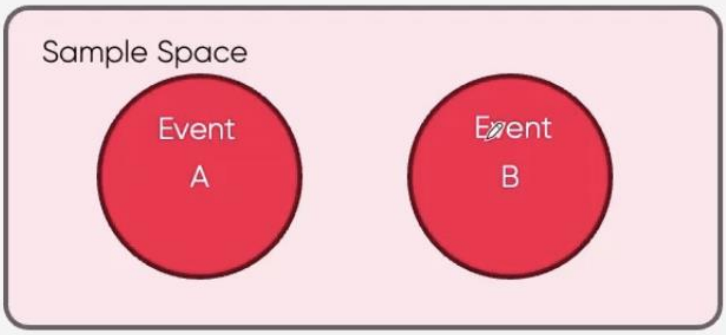
Kejadian yang hasilnya tidak mempengaruhi kejadian selanjutnya

Contoh:
- Pelemparan coin pertama tidak mempengahuri pelemparan kedua, dan seterusnya

Aturan Perkalian (Mutually Exclusive)

dan $\cap$

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

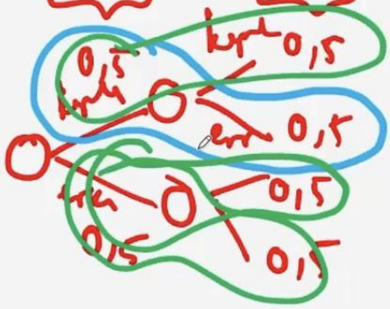


$$P(A \cap B) = 0$$
$$P(A) \times P(B) = 0$$

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

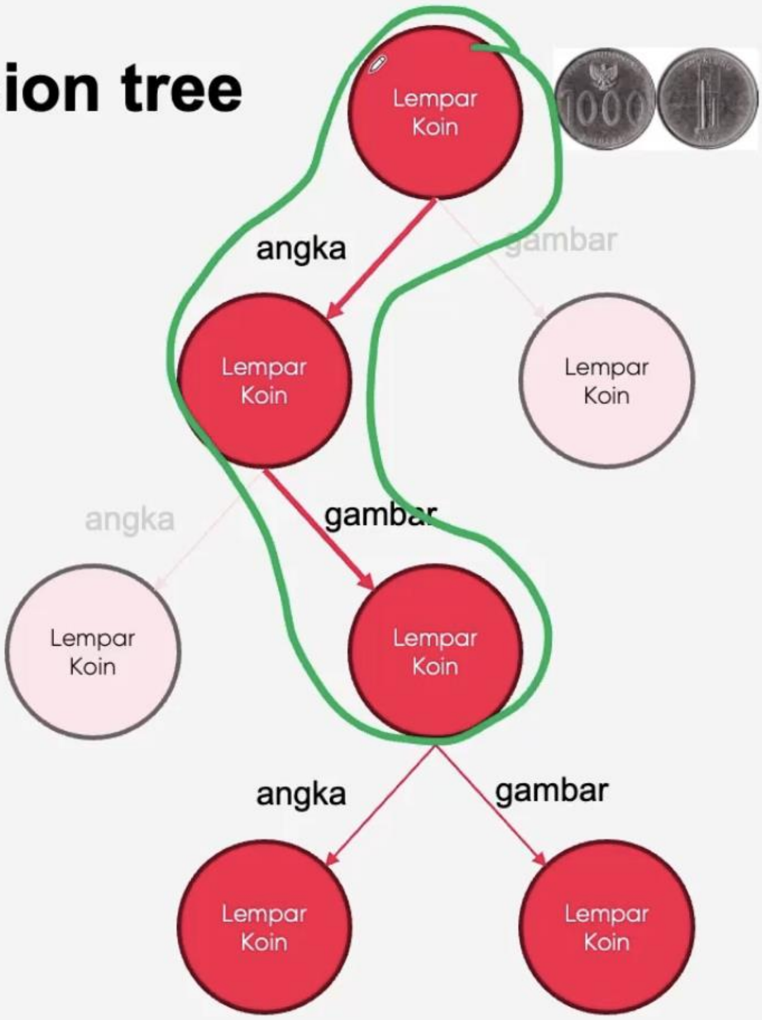
urutan



0,25 . 4

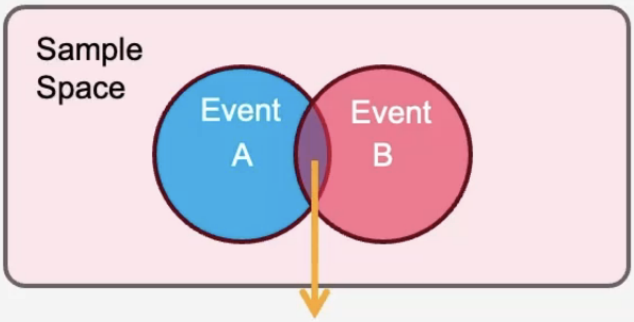
= 1

Ilustrasi decision tree Independent



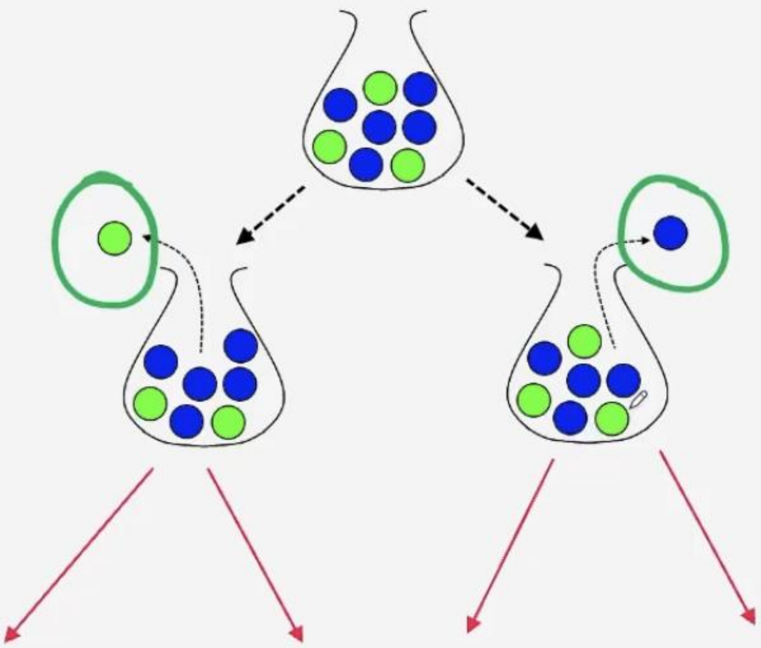
Aturan Perkalian (Non Mutually Exclusive)

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



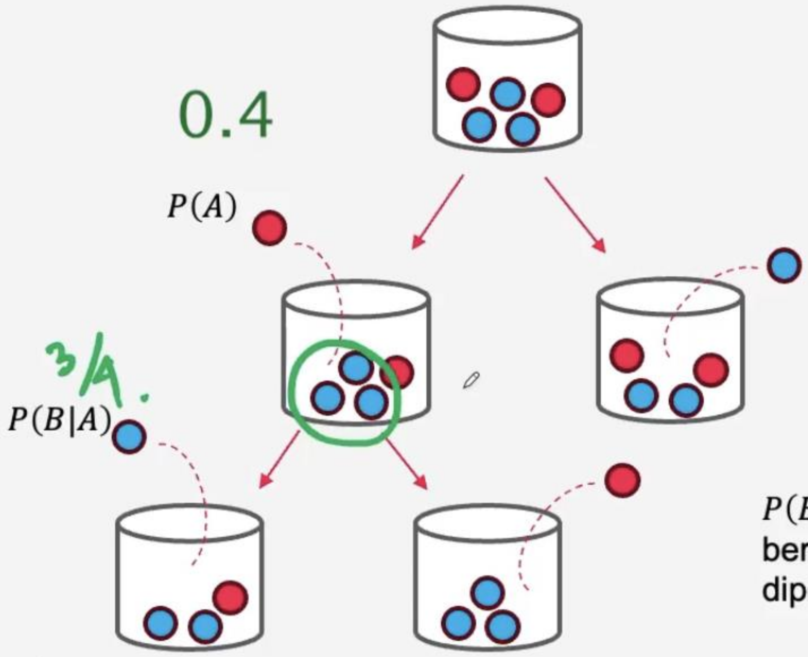
$$P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) \neq 0$$

Ilustrasi decision tree - Dependent



Probabilitas Bersayarat

- 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
  - $P(B|A) = \frac{jumlah\ biru}{kelereng\ sisa} = \frac{3}{4}$

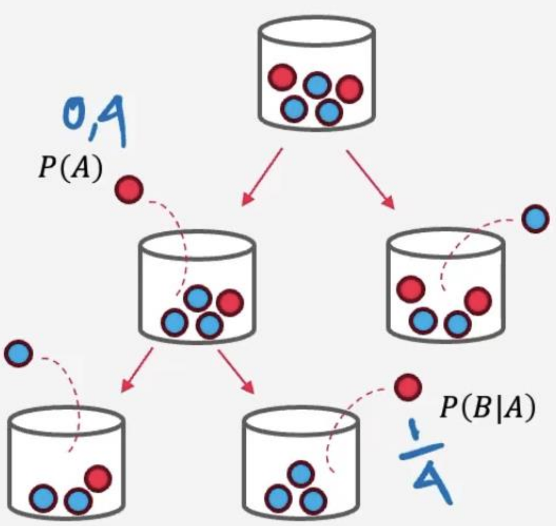


$P(B|A)$ adalah probabilitas bersyarat, bagaimana kejadian B dipengaruhi oleh kejadian A

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



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

$$= \frac{2}{5} \times \frac{1}{4}$$

$$= \frac{1}{10}$$

2-3. Conditional Probabilty & 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 :

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

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

$P(B)$

$P(A|B)$

- 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

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

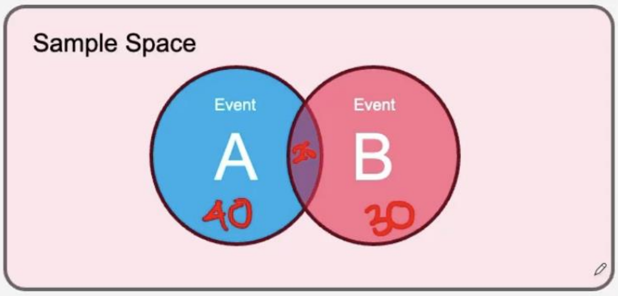
Contoh

- Dari 100 pasien, 40 demam, 30 sakit flu, dan 25 keduanya
- $P(Demam) = \frac{40}{100} = 0.4$
- $P(Sakit\ \cap Demam) = \frac{25}{100} = 0.25$
- $P(Sakit\ |\ Demam) = \frac{0.25}{0.4} = 0.625$

$$P(flu) = \frac{30}{100} = 0{,}3$$

$$P(sakit\ |\ dem) = \frac{P(sakit \cap dem)}{P(dem)} = \frac{0{,}25}{0{,}4}$$

$$P(dem\ |\ saki) = \frac{P(sakit \cap dem)}{P(sakit)} = \frac{0{,}25}{0{,}3} = \cancel{0{,}0\ldots} = 0{,}1\ldots$$



Contigency Table

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

$p(\text{tidak sakit} \mid \text{tidak demam}) = \frac{55}{60} = 0{,}1\ldots$

$P(\text{sakit} \mid \text{demam}) = \frac{P(\text{sakit} \cap \text{demam})}{P(\text{demam})} = \frac{25}{40} = 0{,}625$

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

$p(\text{demam} \mid \text{sakit}) = \frac{25}{30} = 0{,}1\ldots$

Mendeteksi Kejadian Independent

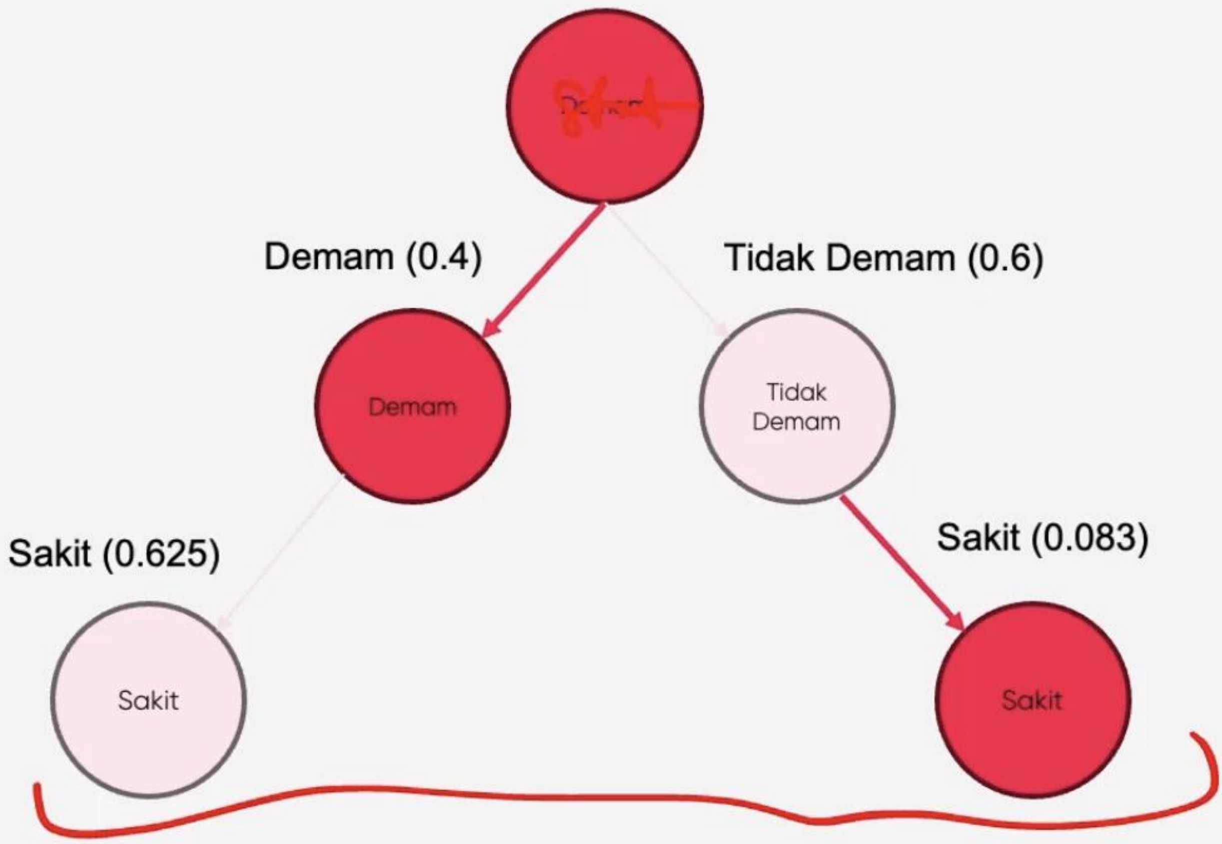
- Dua kejadian A dan B dinyatakan independent jika

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

- Atau secara ekuivalen

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

Ilustrasi decision tree



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.

2-4. Bayes' Theorem Explained - 1

01 **Memahami konsep dasar teorema Bayes**

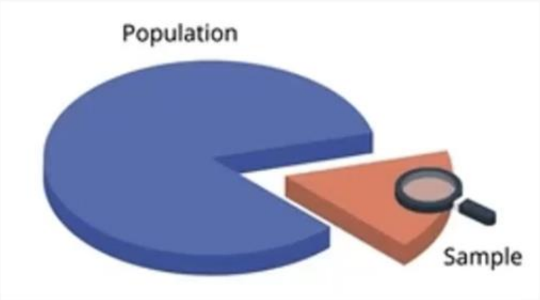
02 **Menginterpretasikan rumus Bayes dalam konteks nyata**

03 **Menerapkan Teorema Bayes pada kasus nyata**

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

Pengantar

- Bayes Theorem adalah alat penting dalam inferensi statistik

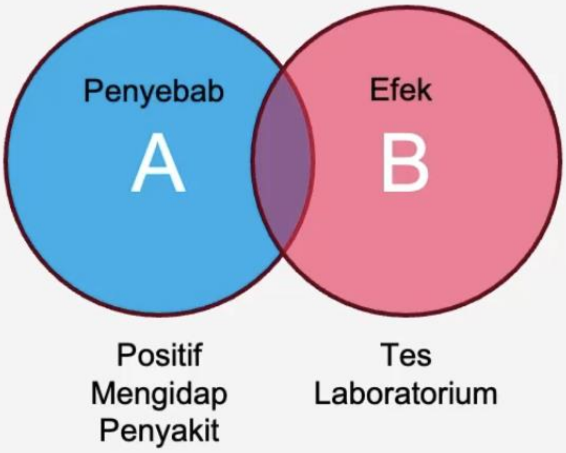


Inferensial:

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

Pengantar

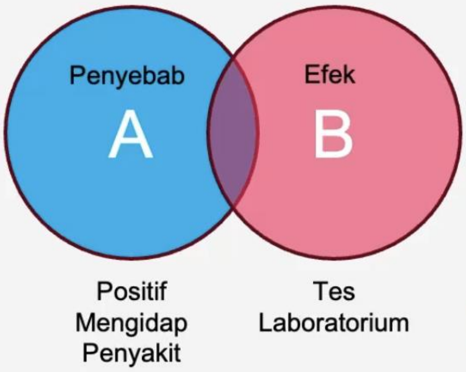
- Secara sederhana. Teorema Bayes memberikan cara matematis untuk membalikkan probabilitas bersyarat.
- Memungkinkan untuk menemukan probabilitas suatu **"penyebab"** (A) jika kita mengetahui **"efek"** (B) telah terjadi.



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

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

Teorema Bayes



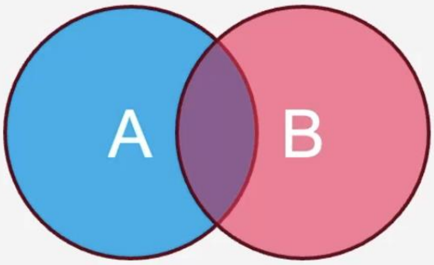
$P(B|A)$ Probabilitas hasil tes positif jika memang sakit

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

Probabilitas sakit dan tes positif =
Perobabilitas tes positif × Sakit jika hasil tes positif

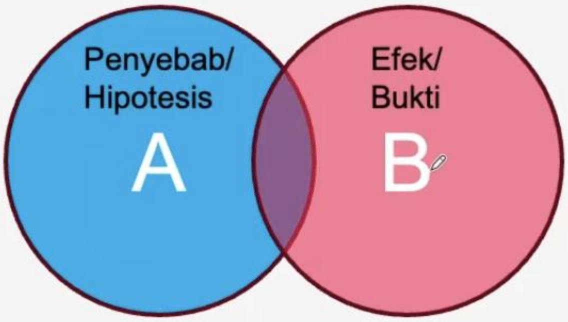
$$P(A \cap B) = P(B) \times P(A|B)$$

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



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

Teorema Bayes

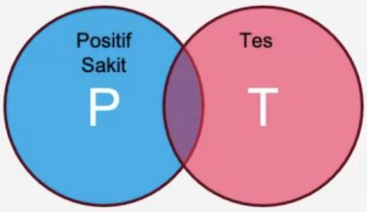


Prior, Bukti.

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

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

Contoh: Diagnosis Medis



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

$P(P) = 1\% = 0{,}01$

$P(P_c) = 99\% = 0{,}99$

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

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

$$P(T) = P(T|P) \cdot P(P) + P(T|P_c) \cdot P(P_c)$$
$$= 0{,}99 \cdot 0{,}01 + 0{,}05 \cdot 0{,}99$$
$$= 0{,}06 \cdot 0{,}99 = 0{,}0594$$

$$P(P|T) = \frac{P(T|P) \cdot P(P)}{P(T)} = \frac{99\% \cdot 1\%}{0{,}0594} = 0{,}16667 \rightarrow 16{,}67\%$$

Langkah-langkah

- Tentukan Prior $P(A)$
- Hitung likelihood $P(B|A)$
- Hitung Evidence $P(B)$
- Gunakan rumus bayes untuk mendapatkan posterior $P(A|B)$

2-4. Bayes' Theorem Explained - 2

Naive Bayes Classifier (Machine Learning)

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

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

teorema bayes

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

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

class

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

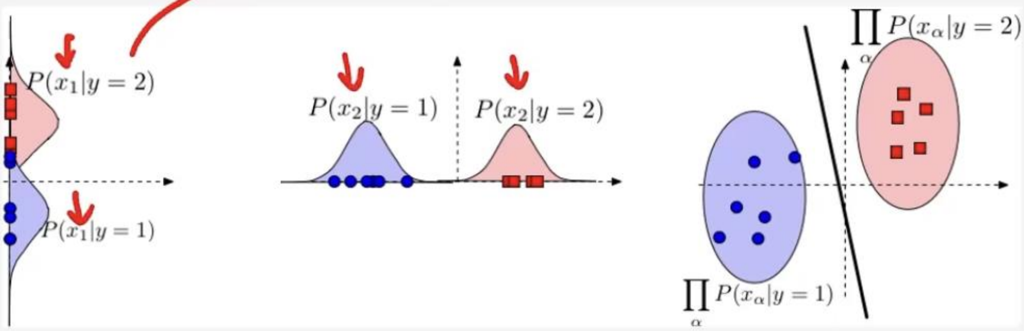
$$\begin{aligned} P(b_1, b_2, \dots, b_n|A) &= P(b_1|A) \times P(b_1|A) \times P(b_1|A) \times \dots \times P(b_n|A) \\ &= \prod_{i=1}^{n} P(b_i|A) \end{aligned}$$

Naive Bayes Classifier (Machine Learning)

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

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

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



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

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

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

Contoh

Tabel Fitur/likelihood kata pada kelas spam

$$P(kata|spam) = P(kata|A_1) = \frac{Jumlah\ kemunculan\ kata + 1}{Total\ kata\ di\ kelas\ spam + total\ kosakata}$$

$$= \frac{jumlah\ kemunculan\ kata}{9 + 13}$$

Kata	Jumlah di spam	$P(kata\|spam)$
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

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

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

Kata	Jumlah di bukan spam	$P(kata\|bukan\ spam)$
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"

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

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

$$P(Spam|email\ baru) = \frac{P(Spam).P(murah|spam).P(sekarang|spam)}{P(semua\ kata)}$$
$$= 0.6 \times \frac{2}{22} \times \frac{2}{22} \times \frac{1}{P(semua\ kata)}$$
$$= 0.004958 \times \frac{1}{P(semua\ kata)}$$

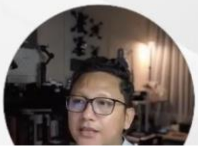
$$P(Bukan\ Spam|email\ baru) = \frac{P(Bukan\ Spam).P(murah|bukan\ spam).P(sekarang|bukan\ spam)}{P(semua\ kata)}$$
$$= 0.4 \times \frac{1}{19} \times \frac{1}{19} \times \frac{1}{P(semua\ kata)}$$
$$= 0.001107 \times \frac{1}{P(semua\ kata)}$$

$$P(Spam|email\ baru) > P(Bukan\ Spam|email\ baru)$$

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

$$0.004958 > 0.001107$$

SPAM!



Naive Bayes

- Naïve Bayes Classifier menggunakan teorema Bayes
- Fitur bersifat independen
- Cepat dan efektif untuk klasifikasi teks
- Contoh **`sklearn.naive_base.GaussianNB.MultinominalNB`**

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