Statistical Machine Learning s Data Storytelling

spacing = 3 tab

1. Part 1 Essential Mathematical Foundations (Deep Live)

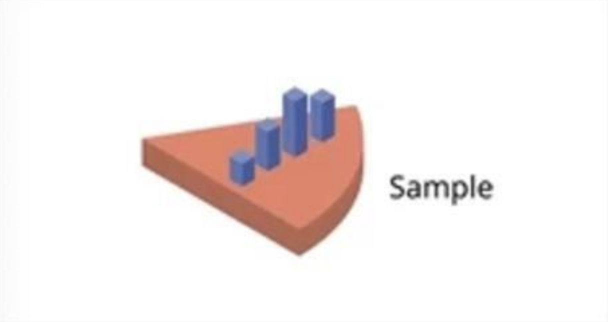
Chapter 1. Descriptive Statistics

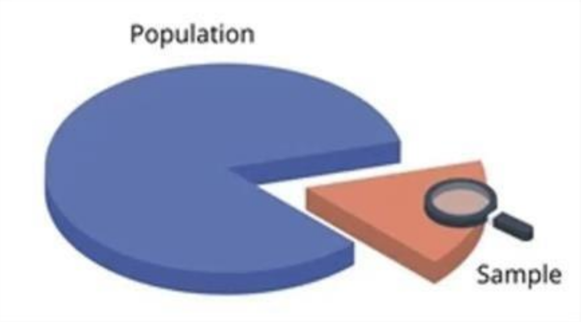
1-1. Introduction To Descriptive Statics

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

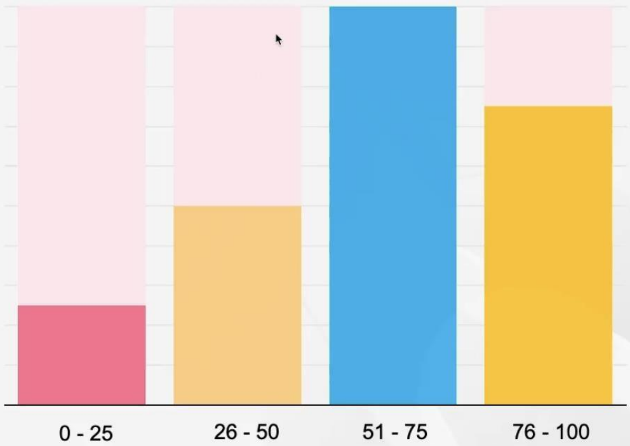
Inferensial:

- 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

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

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

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

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

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

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

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

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)

$X \to 1\ \underbrace{2\ 2\ 2}\ 3\ 4$

modus = 2

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

Bimodal.

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

multimodal

Definisi: Nilai yang sering muncul.

Formula:

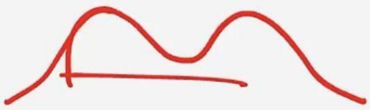
- Frekuensi / kemunculan data tertinggi
- Dapat memiliki lebih dari satu modus (bimodal/multimodal)

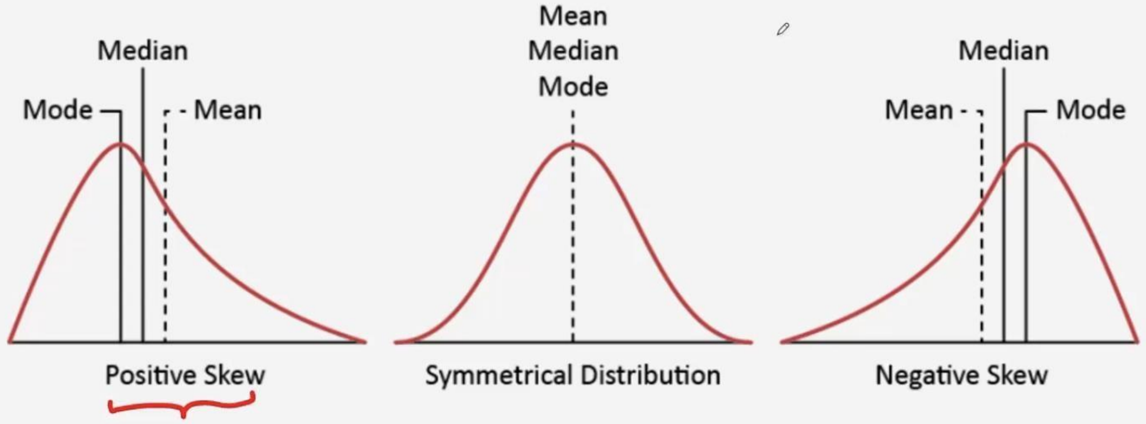
Catatan:

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)**

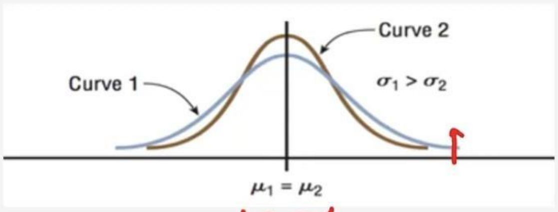
04 **Standard Deviation (Simpangan Baku)**

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

min

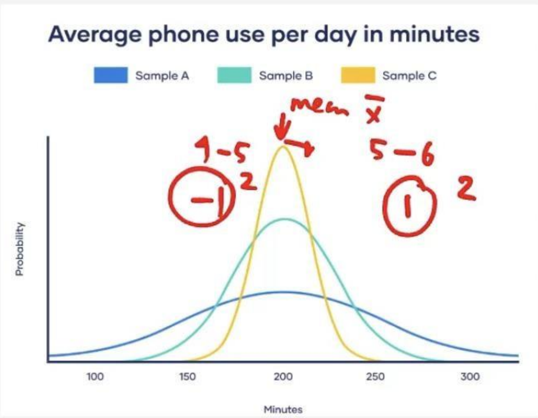
$$\begin{aligned} range &= \max - min \\ &= 90 - 15 \\ &= 75 \end{aligned}$$

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

$$range = max - min$$

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

Variance (Variansi / Ragam)



mean $\bar{x}$

4-5   5-6

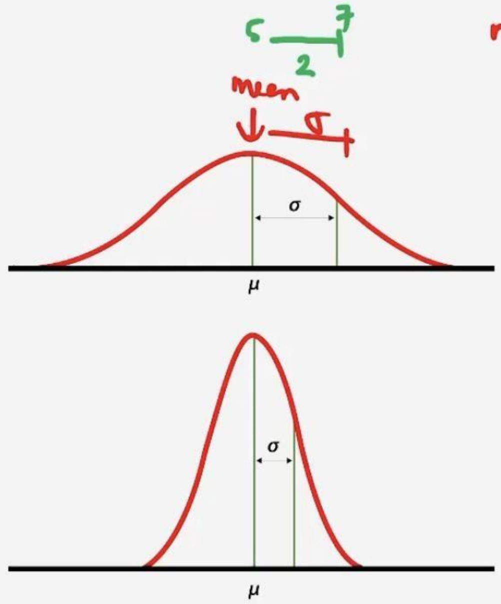
$(-1)^2$ $(1)^2$

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

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

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

Standard Devitation (Simpangan Baku)



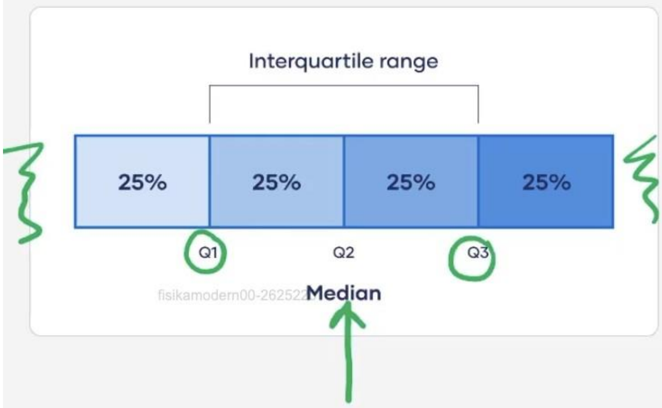
mean = 5
σ = 2

- **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)

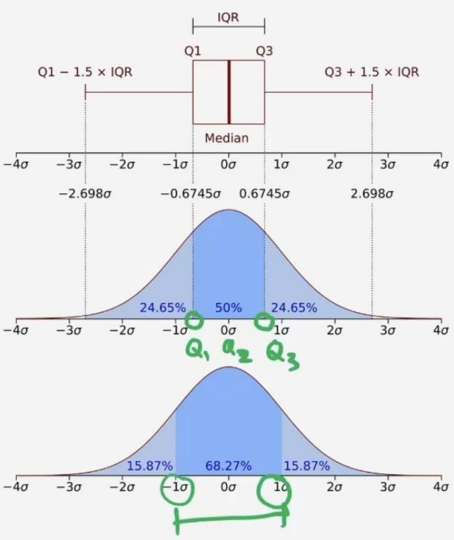


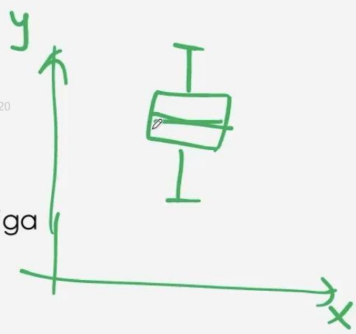
- **Definisi**: selisih antara kuartil ketiga (Q_3) dan kuartil pertama (Q_1).
- **Formula**:

$$IQR = Q_3 - Q_1$$

- **Kelebihan**: Robust terhadap outlier.
- **Kekurangan**: Tidak menggunakan semua data

Interquartile Range (IQR)

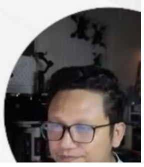




- **Definisi**: selisih antara kuartil ketiga (Q_3) dan kuartil pertama (Q_1).
- **Formula**:

$$IQR = Q_3 - Q_1$$

- **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	02	03	04	05
Definisi	**Jenis-jenis distribusi data**	**Skewness & Kurtosis**	**Perbandingan**	**Penutup**

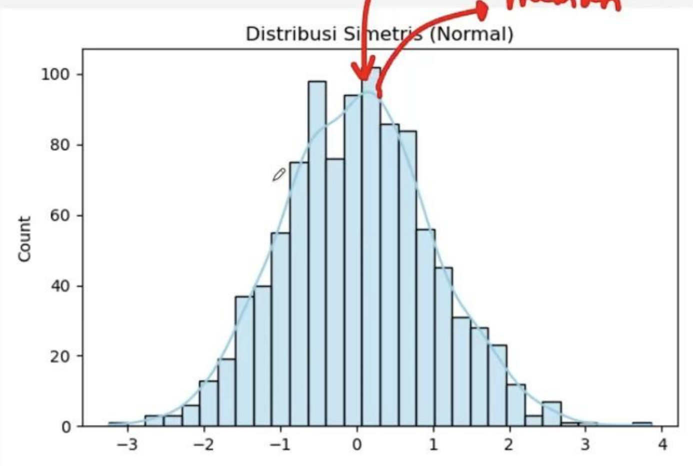
Bentuk Distribusi Data

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

Distribusi Simetris

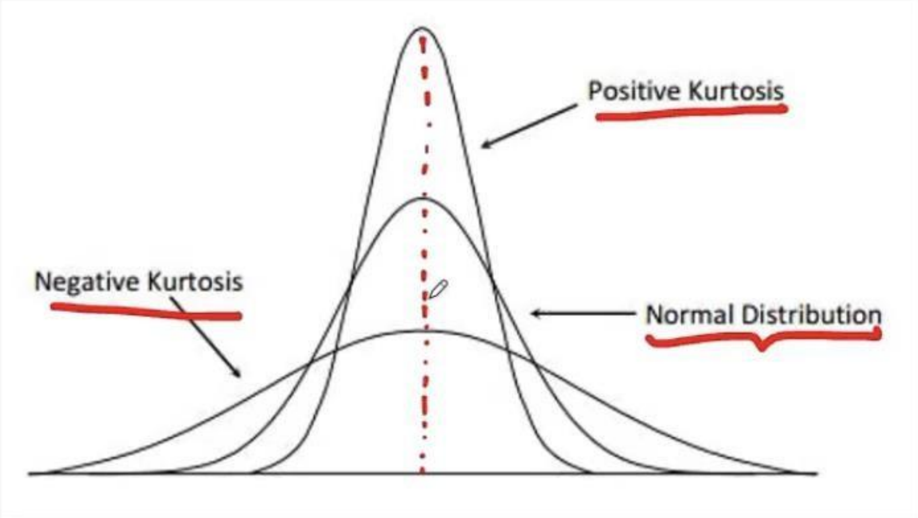
mean

median



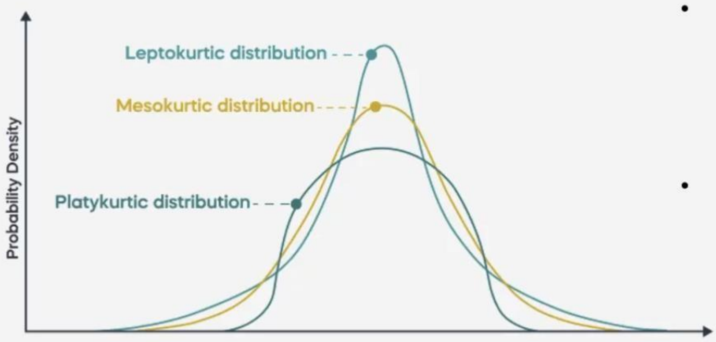
- Distribusi yang tersebar dengan titik **median** dan **mean** saling **berdekatan.**
- Contoh dari distribusi ini adalah **distribusi normal**

Kurtosis



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

Kurtosis



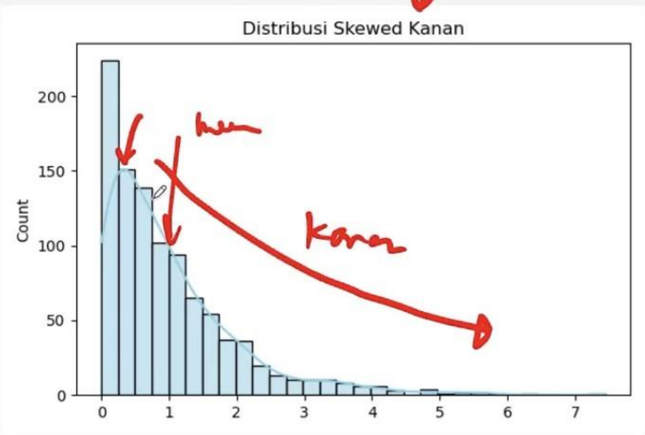
library

Menghitung kurtosis dengan python

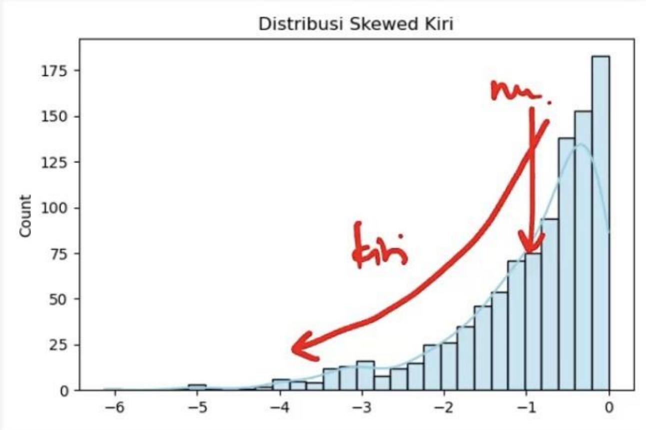
```
from scipy.stats import kurtosis
kurtosis(data, fisher=False)
```

- 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

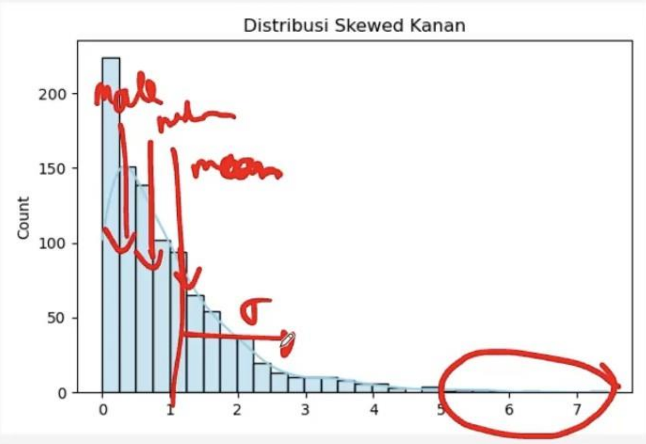


Skewness < 0: miring ke kiri

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

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

Distribusi Miring (Skewed)

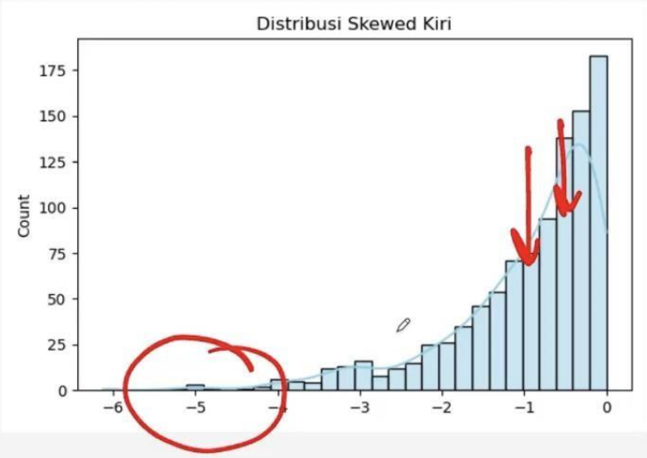


Skewness > 0: miring ke kanan

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)

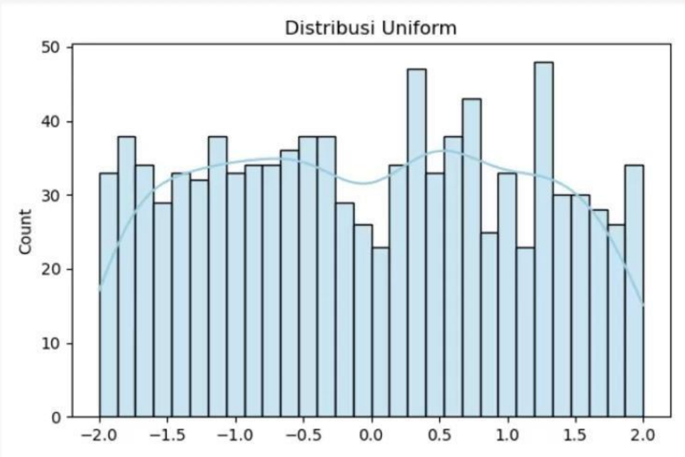


Skewness < 0: miring ke kiri

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]

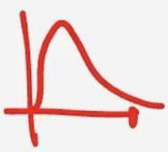
Hitung mean, median, modus, dan range

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

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

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

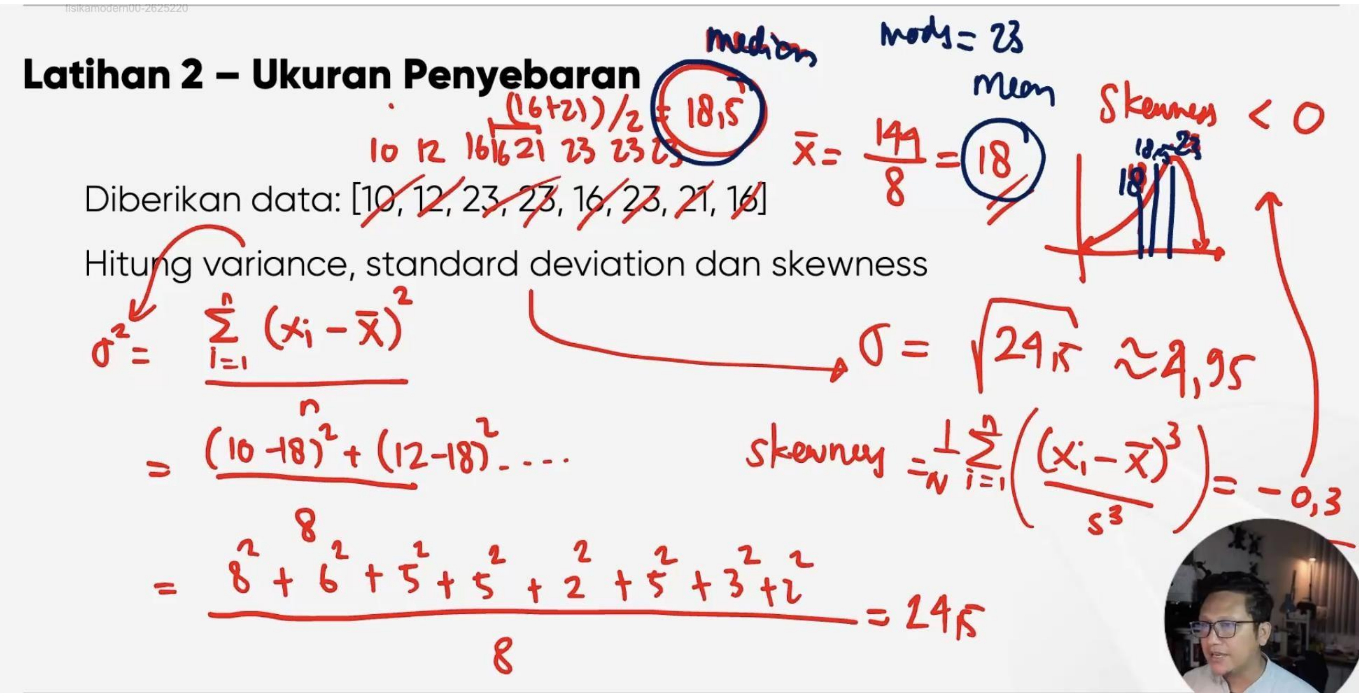
$\dfrac{4+5}{2} = 4{,}5$



modus = 4

range = 8 − 3

= 5



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

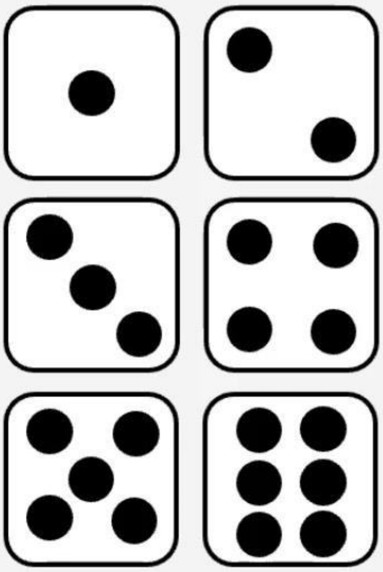
Chapter 2. Probability Theory

2-1. What Is Probability

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?

$$
\begin{array}{cccccc}
[1, & 2, & 3, & 4, & 5, & 6] \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
0 & 1 & 0 & 0 & 0 & 0
\end{array}
$$

$$\left.\begin{array}{l} P(2) = \frac{1}{6} \\ P(1) = \frac{1}{6} \\ \vdots \\ P(6) = \frac{1}{6} \end{array}\right\} 1$$

Contoh

angka ↓ gambar ↓

Berapa probabilitas / kemungkinan gambar muncul saat sekali tos?

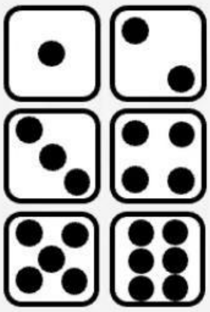
$P(\text{gambar}) = \frac{1}{2}$

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

1

Sample Space / Ruang Sampel

- Sample Space adalah himpunan semua kemungkinan hasil



Sample Space nya adalah:

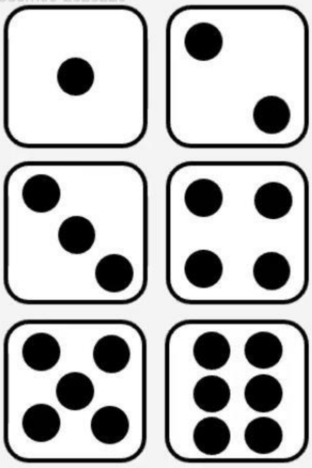
$\underbrace{[\,1, 2, 3, 4, 5, 6\,]}_{6}$

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]

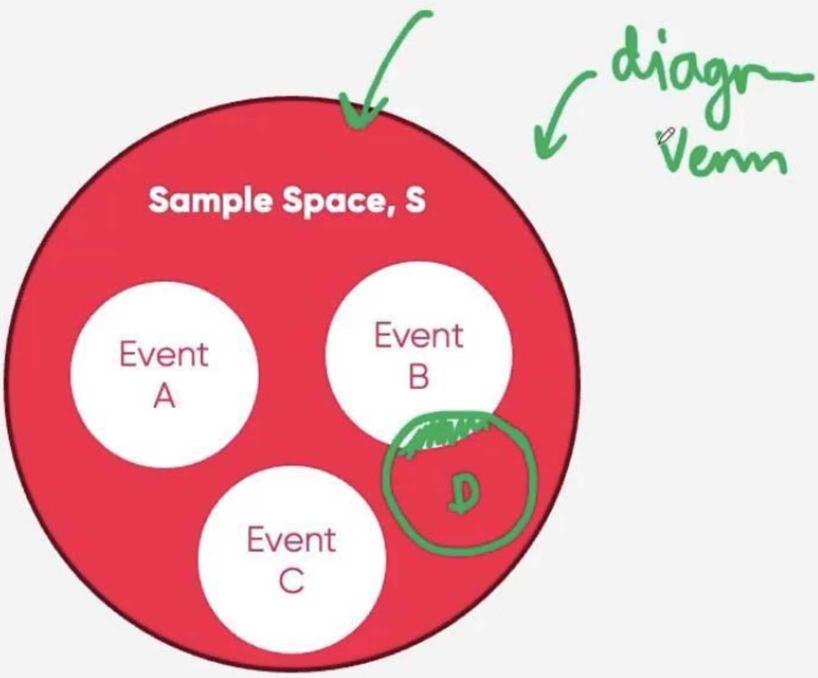
Event → Angka Genap:

[2, 4, 6]

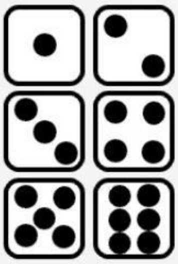
Event → Bilangan Prima:

[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

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

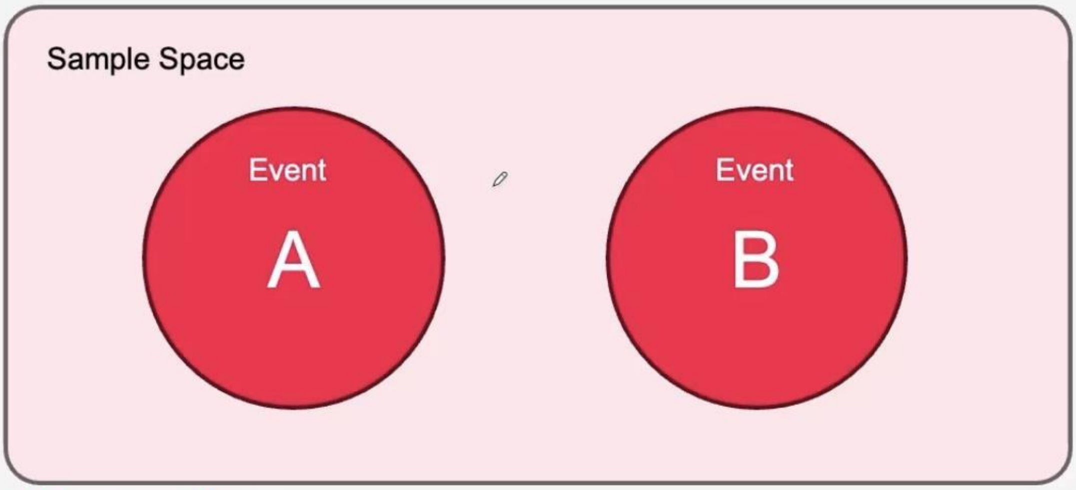


Diagram Venn

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

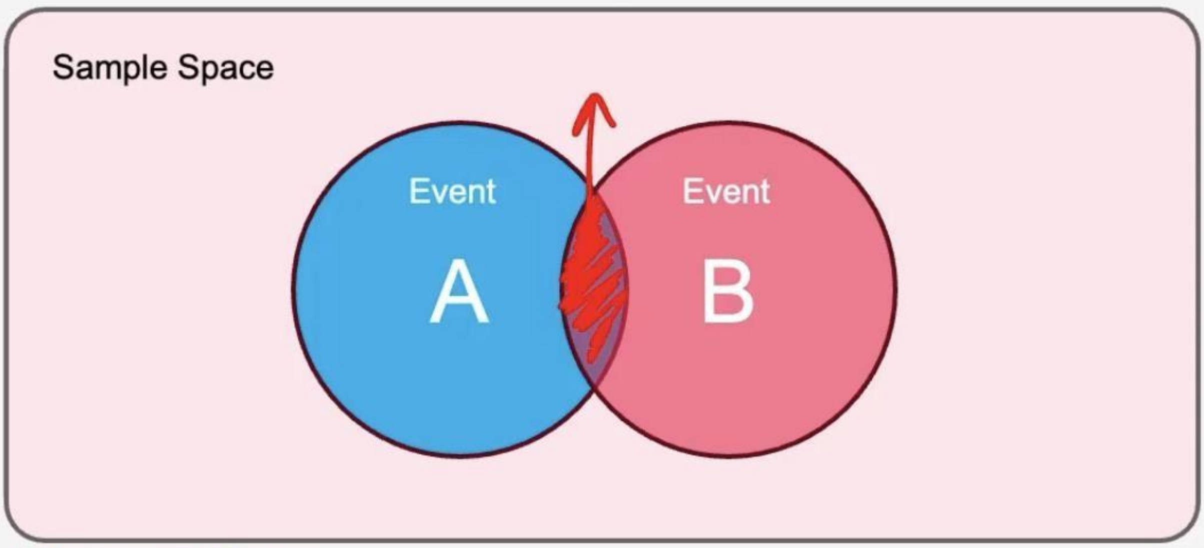
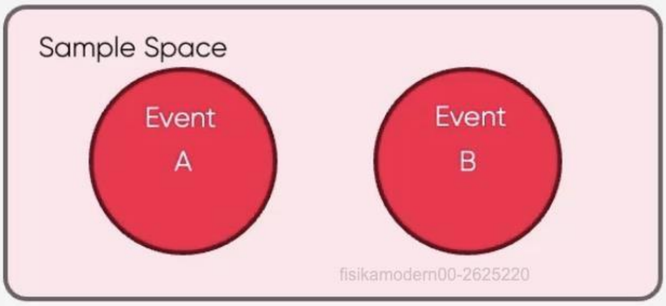


Diagram Venn

Aturan Penjumlahan (Mutually Exclusive)

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



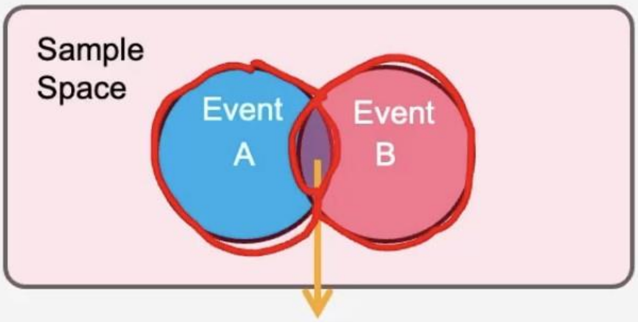
$$P(A \text{ atau } B) = P(A) + P(B)$$
$$P(1 \text{ atau } 6) = P(1) + P(6)$$
$$= \frac{1}{6} + \frac{1}{6}$$
$$= \frac{2}{6} = \frac{1}{3}$$

- Formulasi:

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

Aturan Penjumlahan NON (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:

union.

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

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

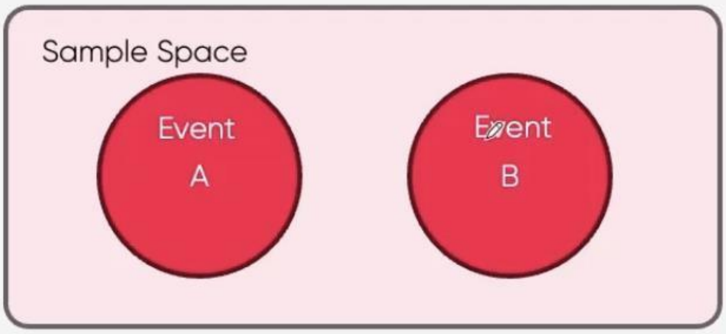
Contoh:

- 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

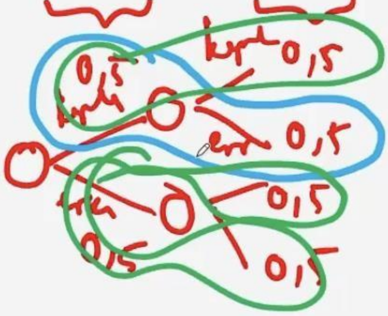


$$P(A \cap B) = 0$$
$$P(A) \times P(B) = 0$$

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

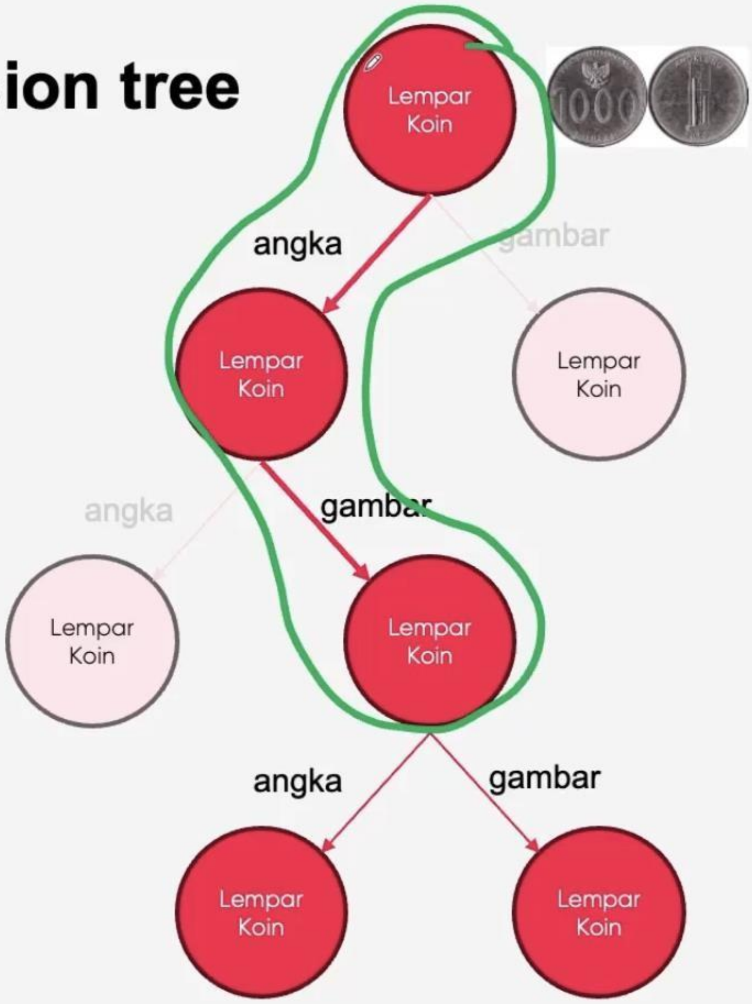
urutan



0,25 . 4

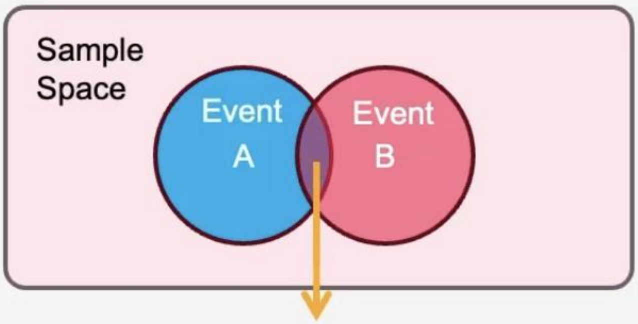
= 1

Ilustrasi decision tree Independent



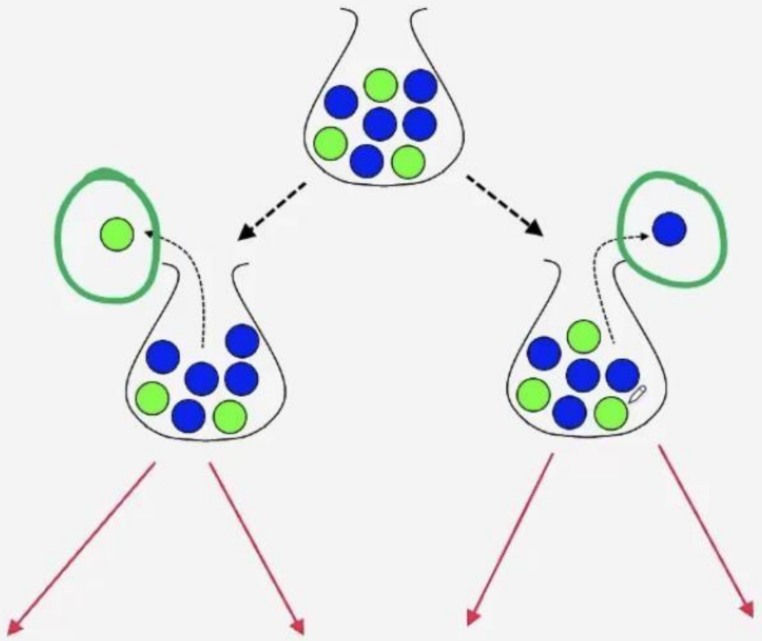
Aturan Perkalian (Non Mutually Exclusive)

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



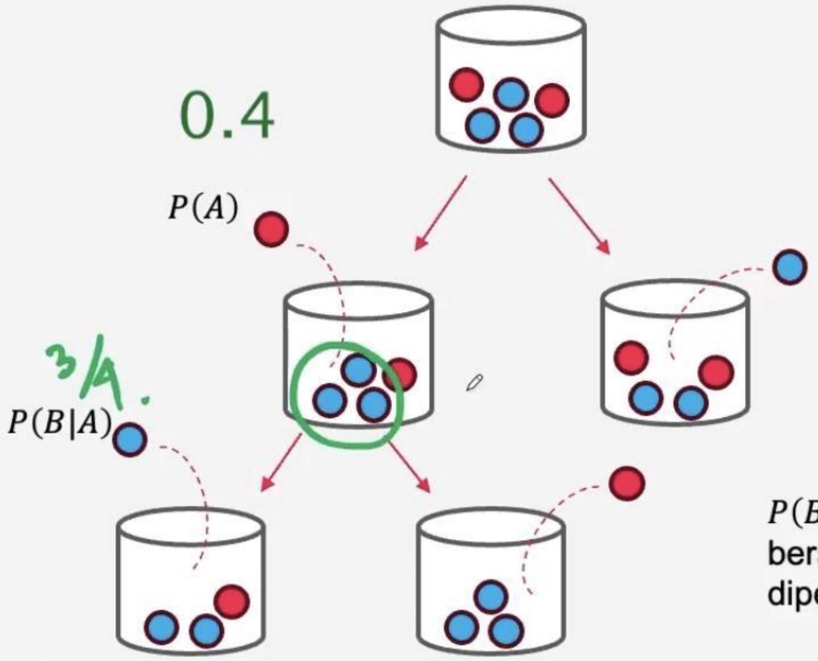
$$P(A \text{ dan } B) = P(A \text{ and } B) = P(A \cap B) \neq 0$$

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

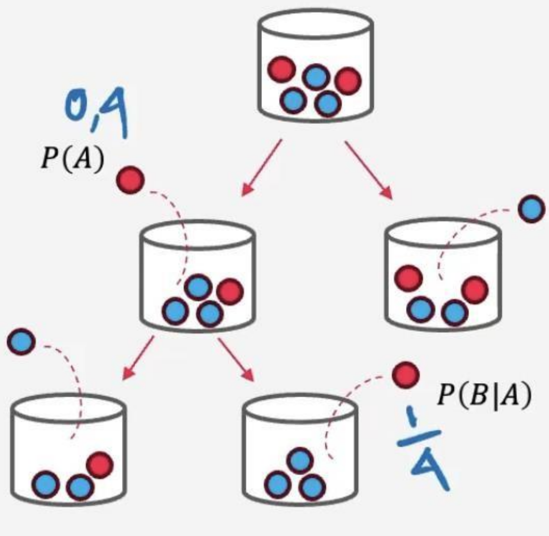


$P(B|A)$ adalah probabilitas bersyarat, bagaimana kejadian B dipengaruhi oleh kejadian A

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



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

$$= \cancel{\frac{2}{5}} \times \cancel{\frac{1}{4}} = \cancel{\frac{2}{20}}\ \frac{1}{10}$$

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

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

2-3. Conditional Probabilty s 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 :

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

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

$P(A|B)$

- 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

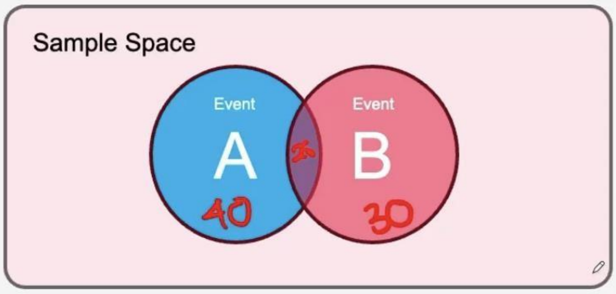
$$P(\text{flu}|\text{demam}) = \frac{P(\text{flu dan demam})}{P(\text{demam})}$$

Contoh

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

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

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



$P(dem \mid saki) = \frac{P(sakit \cap dem)}{P(sakit)} = \frac{0{,}25}{0{,}3} = \cancel{0{,}0\ldots7} = 0{,}1\ldots$

Contigency Table

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

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

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

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

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

Mendeteksi Kejadian Independent

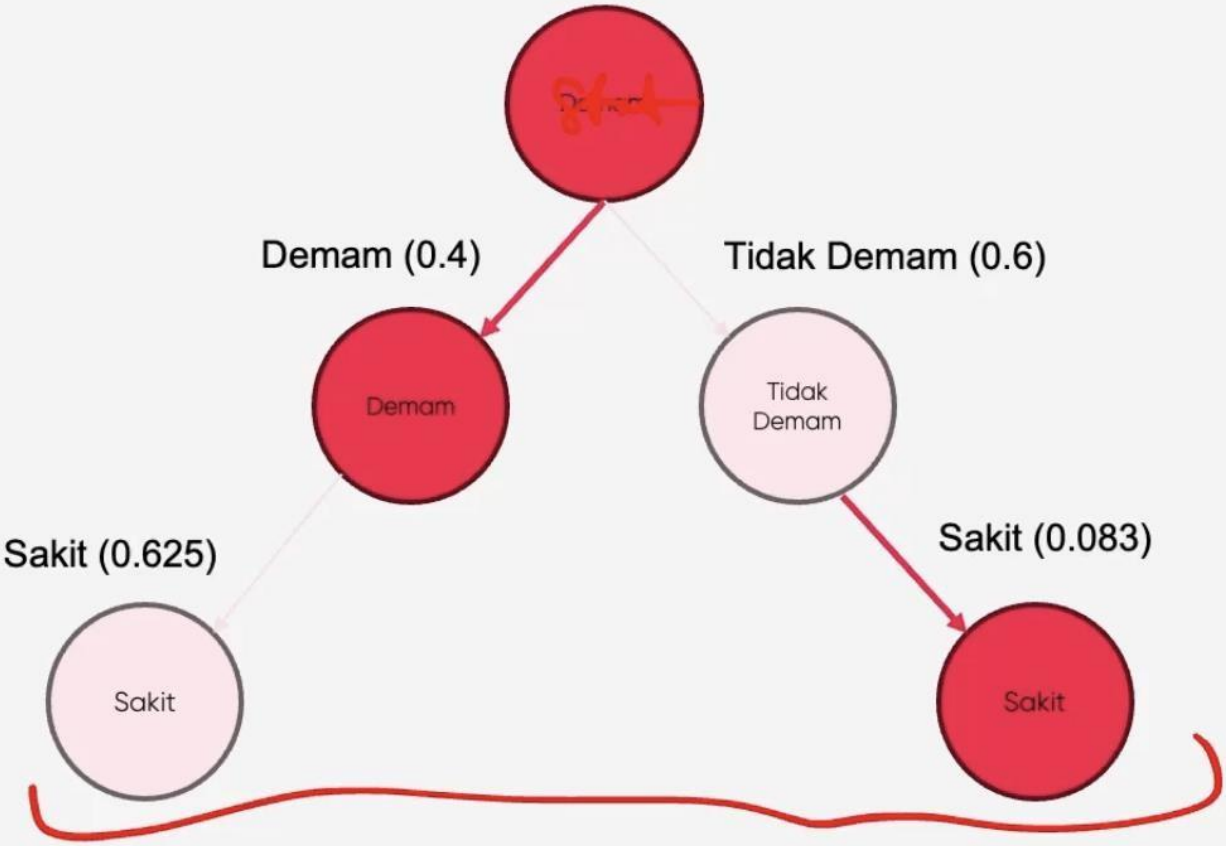
- Dua kejadian A dan B dinyatakan independent jika

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

- Atau secara ekuivalen

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

Ilustrasi decision tree



Probabilitas total sakit = (0.4×0.625) + (0.6×0.083) = 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**

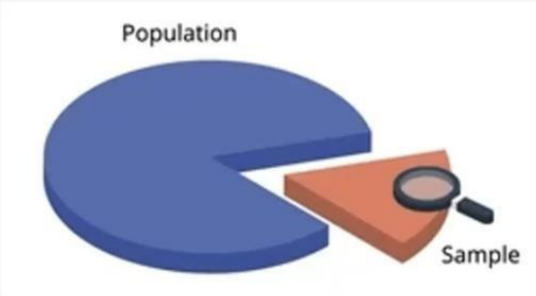
02 **Menginterpretasikan rumus Bayes dalam konteks nyata**

03 **Menerapkan Teorema Bayes pada kasus nyata**

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

Pengantar

- Bayes Theorem adalah alat penting dalam inferensi statistik

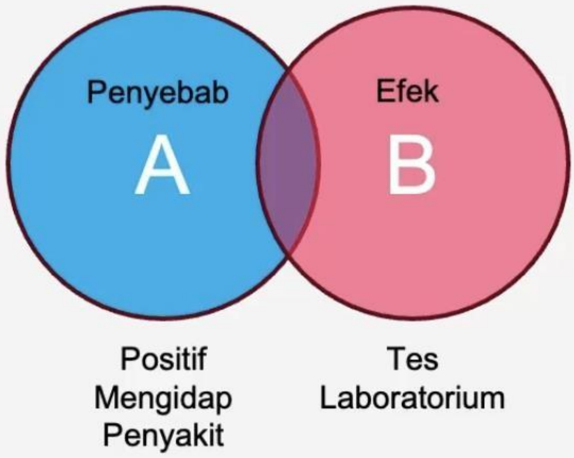


Inferensial:

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

Pengantar

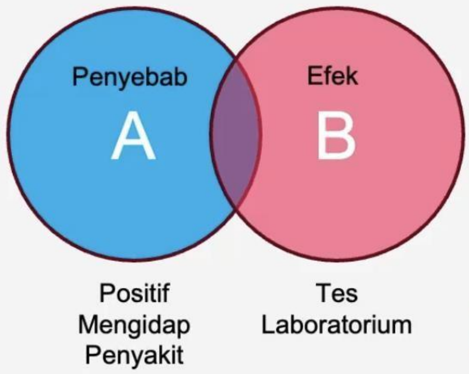
- Secara sederhana. Teorema Bayes memberikan cara matematis untuk membalikkan probabilitas bersyarat.
- Memungkinkan untuk menemukan probabilitas suatu **"penyebab"** (A) jika kita mengetahui **"efek"** (B) telah terjadi.

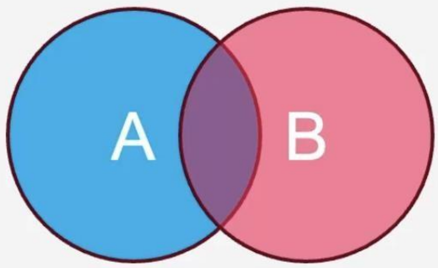


$P(B|A)$ Probabilitas hasil tes positif jika memang sakit

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

Teorema Bayes





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

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

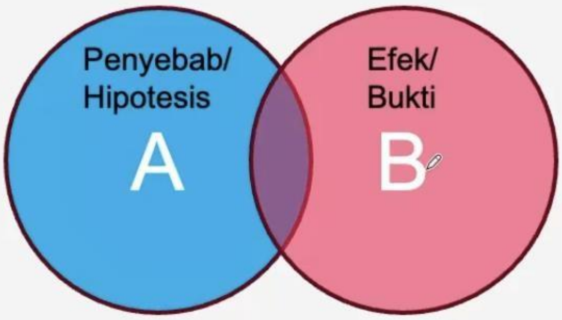
Probabilitas sakit dan tes positif =
Perobabilitas tes positif × Sakit jika hasil tes positif

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

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

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

Teorema Bayes



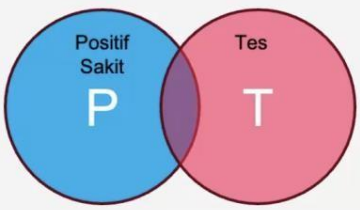
Prior — Bukti.

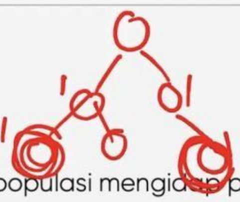
$$\underbrace{P(A|B)} = \frac{P(A)\times P(B|A)}{\underbrace{P(B)}_{\text{bobing.}}}$$

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

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

Contoh: Diagnosis Medis





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

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

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

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

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

$P(T) = P(T|P) \cdot P(P) + P(T|P_c) \cdot P(P_c)$. $P(P|T)$

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

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

$P(P|T)$

$= \dfrac{P(T|P) \cdot P(P)}{P(T)}$

$= \dfrac{99\% \cdot 1\%}{0{,}0594}$

$= 0{,}16667 \rightarrow 16{,}67\%$

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

Pertanyaan :

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

Jawaban

1. Peluang Dasar (dari isi kotak) :

* $P(manis \mid A) = \frac{10}{15} = \frac{2}{3}$
* $P(manis \mid B) = \frac{4}{20} = \frac{1}{5}$

2. Hitung total probabilitas dapat cokelat manis $P(manis)$:

* $P(manis) = P(manis \mid A) \,.\, P(A) + P(manis \mid B) \,.\, P(B)$
* $P(manis) = \frac{2}{3} \,.\, \frac{3}{5} + \frac{1}{5} \,.\, \frac{2}{5} = \frac{12}{25}$

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

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

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 $P(A|B)$

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!

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

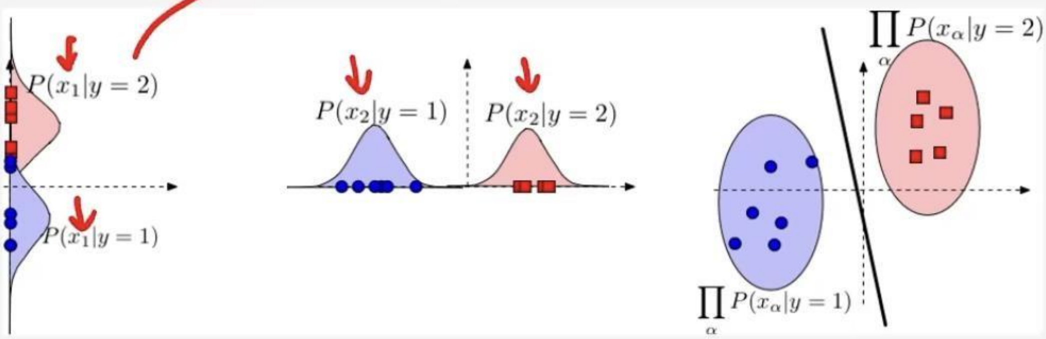
$$P(b_1, b_2, \dots, b_n|A) = P(b_1|A) \times P(b_1|A) \times P(b_1|A) \times \cdots \times P(b_n|A)$$
$$= \prod_{i=1}^{n} P(b_i|A)$$

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

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

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

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"

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

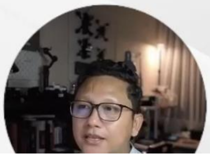
$$P(Bukan\ spam) = P(A_2) = \frac{jumlah\ bukan\ spam}{total\ email} = \frac{2}{5} = 0.4$$

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

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

$$P(Spam|email\ baru) > P(Bukan\ Spam|email\ baru)$$
$$0.004958 \times \frac{1}{P(semua\ kata)} > 0.001107 \times \frac{1}{P(semua\ kata)}$$
$$0.004958 > 0.001107$$

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

Chapter 3. Random Variable s Probability Distribution

3-1. What are random variables

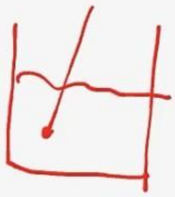
01 **Memahami konsep Random Variable**

02 **Membedakan antara fenomena diskrit dan kontinu**

03 **Memahami pentingnya Random Variable untuk pemodelan**

Apa Itu Variables/Variabel?

Suhu = 72°



Simbol yang mewakili kuantitas yang dapat berubah

Contoh: Suhu, Jumlah_penjualan, Usia

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): $\{Angka, Gambar\}$

$\downarrow \quad \downarrow$

$1 \quad 0$

Tantangan: Analisis secara kuantitatif

Solusi:

definisikan:

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

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

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

$P(\text{angka}) \rightarrow P(X=1)$

$P(\text{gambar}) \rightarrow P(X=0)$

CONTOH!

Percobaan: Melempar sebuah koin **dua** kali

Ruang Sampel (S): $\{AA, AG, GA, GG\}$

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

Tantangan: Analisis secara kuantitatif

Solusi:

definisikan RV $\rightarrow X$ = jumlah angka yg muncul

$X=0 \rightarrow$ Angka 0

$X=1 \rightarrow$ Angka 1 kali

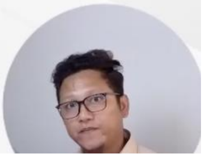
$X=2 \rightarrow$ Angka 2 kali muncul.

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

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

$= 0,25 + 0,25$

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



Definisi Formal?

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

$$X: S \rightarrow \mathbb{R}$$

X: ↑sampl ↑angka

sampel (S) → Variabel (X) → Nilai numerik (probability)

Ditentukan oleh proses acak

Angka

Gambar

X : Angka

$X = 1$

$X = 0$

$p(X = 1) = \frac{1}{2}$

$p(X = 0) = \frac{1}{2}$

Diskrit VS Kontinu

Discrete Random Variable (Variabel Acak Diskrit)

- Mengambil nilai-nilai yang dapat dihitung atau terpisah. 1, 2, 3, 4 .....
- 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

↓dddd

min 160 160,5 170,0000001 max 185

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 Probabily 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.

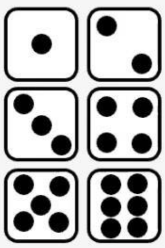
Definisi

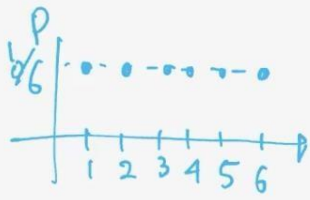
- Untuk variabel acak diskrit X, PMF dilambangkan sebagai

$$\boxed{P(X = x)}$$

- Probabilitas untuk setiap nilai harus non-negative: $P(X = x) \geq 0$
- Total semua probabilitas: $\sum_x P(X = x) = 1$

Contoh never dies!





Percobaan Melempar sebuah dadu standar enam sisi

Hitung

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

$P(X = x) = ?$

$P(X = 1) = 1/6$

$P(X = 2) = 1/6$

$\vdots$

$P(X = 6) = 1/6$

total propabilitas

$\sum P(X = x) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6$

$= 1$

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.

jumlah berhasil

jumlah percobaan

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

probabilitas

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

Distribusi Binomial

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

$n=1$ $k=1$ $p=0{,}5$

p p $n-k$

1 percobaan, 1 berhasil

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

$\frac{\cancel{1!}}{1!\cdot\cancel{1!}}\cdot 0{,}5\,(0{,}5)^0=0{,}5$

1 percobaan, 1 gagal

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

k k k k $k=0$

Distribusi Binomial

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

$n=2,\ k=1,\ p=0{,}5$

2 percobaan, 1 berhasil ~~lalu~~ 1 gagal

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

$n=2,\ k=2,\ p=0{,}5$

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

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

0,25 SS   0,25 SG, 0,25 GS, 0,25 GG

$\underbrace{SG,\ GS}_{0{,}5}$

Distribusi Binomial

1 2 1 0

$a^2+2ab+b^2$

$(a+b)^2$

$(a+b)^3$

$(a+b)^4$

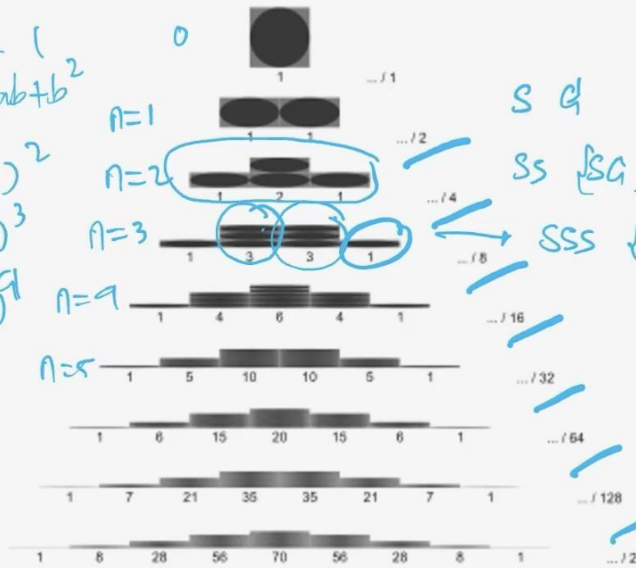
n=1

n=2

n=3

n=4

n=5



S G

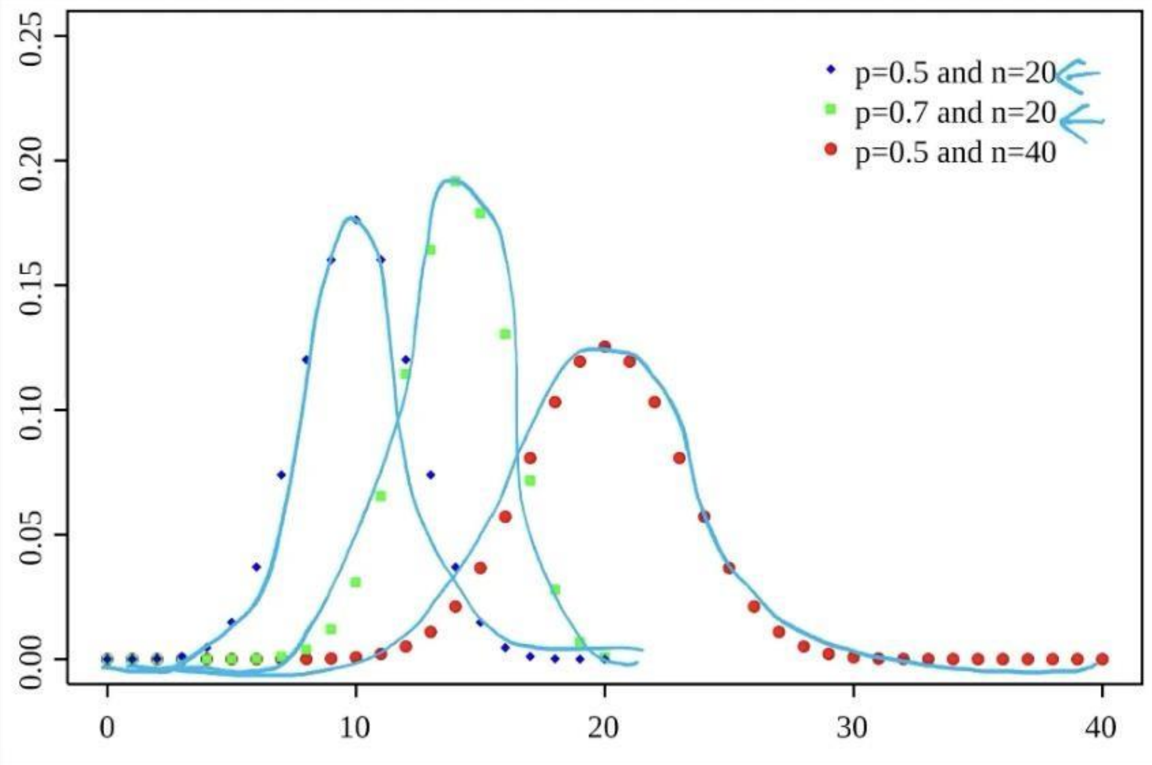
SS {SG, GS}, GG

SSS {GGS, SSG}, GGG

GGS, SGG, GSG } 3

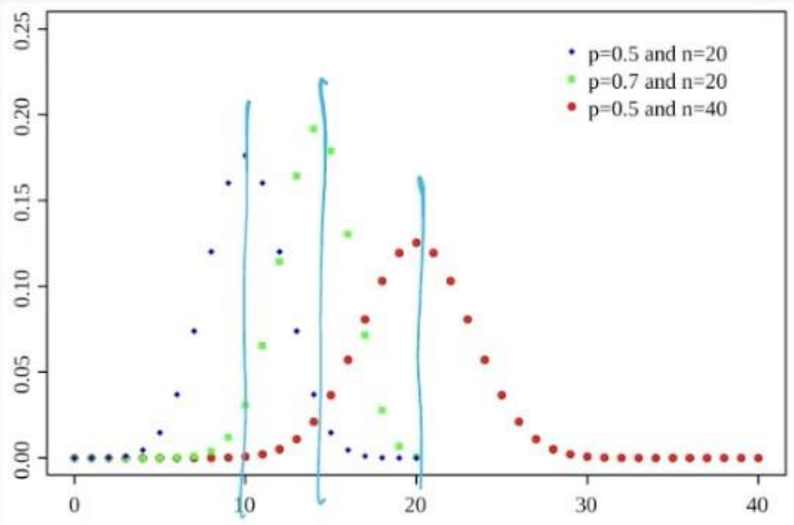
SSG, SGS, GSS } 3

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

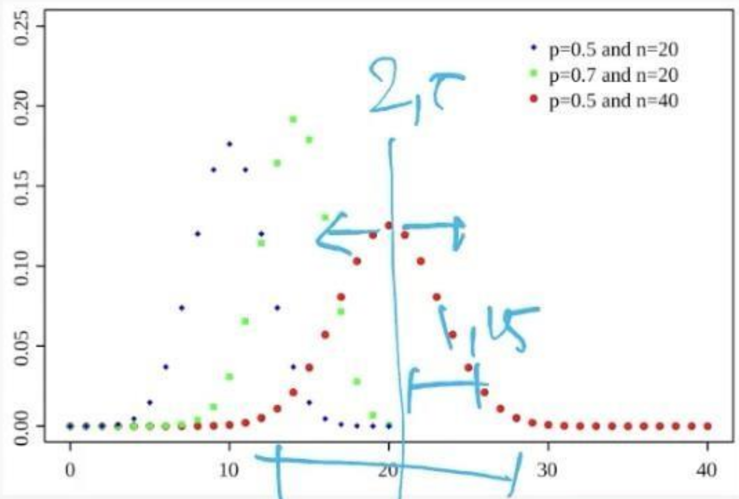
Contoh: Melempar koin 5 kali

$E(x) = 0{,}5 \cdot 5 = 2{,}5$

Variance

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

$E(x)\,(1-p)$,



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

n: Jumlah percobaan
p: Probabilitas

Contoh: Melempar koin 5 kali

$Var(X) = 2{,}5 \cdot (1 - 0{,}5)$
$= 1{,}25$

3-3. Discrete Probabily Distribution - 2

Distribusi Poisson

6 3 2 12 11 6 8 pelanggan.
8 9 10 11 12 13 14 15 jam

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 (λ)

Parameter (λ): Rata-rata jumlah kejadian dalam interval yang diberikan
Contoh: Rata-rata 3 pelanggan tiba di sebuah toko per jam

e

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

rata-rata.

kejadian

Di mana

$e \approx 2.71828$

Distribusi Poisson

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

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

p   1 − p

Karena percobaan pada domain kontinu maka n berjumlah tak hingga

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

$$= \lim_{n\to\infty}\left[\frac{\lambda^k}{k!}\right]\left[\frac{n(n-1)\dots(n-k+1)}{n^k}\right]\left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Distribusi Poisson

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

$$= \lim_{n\to\infty}\left[\frac{\lambda^k}{k!}\right]\left[\frac{n(n-1)\dots(n-k+1)}{n^k}\right]\left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \lim_{n\to\infty}\left[\frac{\lambda^k}{k!}\right]\left[\frac{n}{n}\cdot\frac{n-1}{n}\cdot\dots\cdot\frac{n-k+1}{n}\right]\left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \lim_{n\to\infty}\left[\frac{\lambda^k}{k!}\right]\left[\frac{\left(1 - \frac{\lambda}{n}\right)^{n}}{\left(1 - \frac{\lambda}{n}\right)^{k}}\right]$$

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

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)

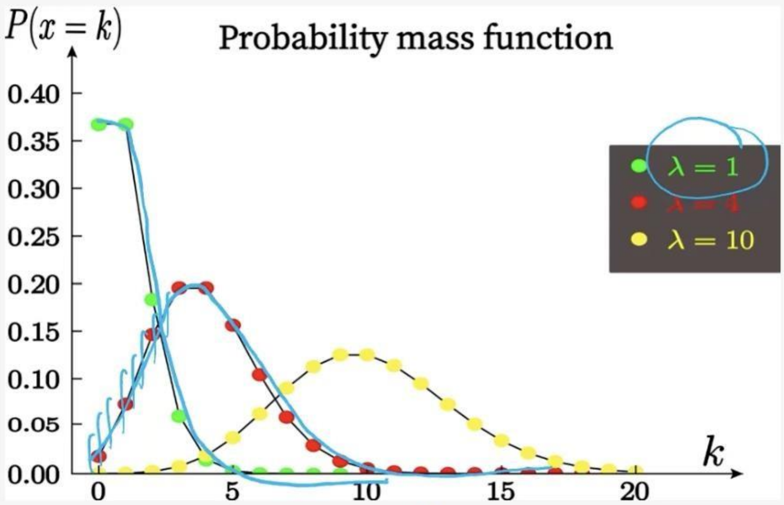
Perhitungan:

$$P(X=k) = \frac{e^{-4}\, 4^k}{k!}$$

$$p(X=2) = \frac{e^{-4} \cdot 4^2}{2!} = \frac{e^{-4}\, \cancel{16}\, 8}{\cancel{2} . 1} = 0{,}224$$

22,4%

Distribusi Poisson



$\lambda = 1$ $k = \longrightarrow$ k

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

Mean $E[X]$ = Variance $Var[X]$ = λ

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).

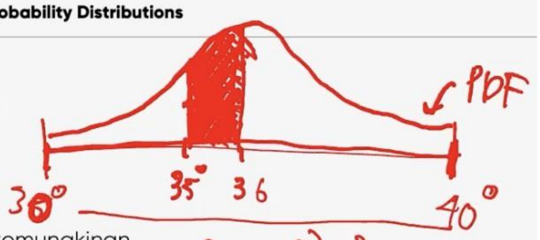
3-4. Continuous Probabiliy Distribution

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?

- PDF 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

$P(X = 35°)$? ∞

$\frac{100}{\infty} = 0$

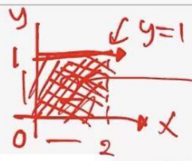
Penting!

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

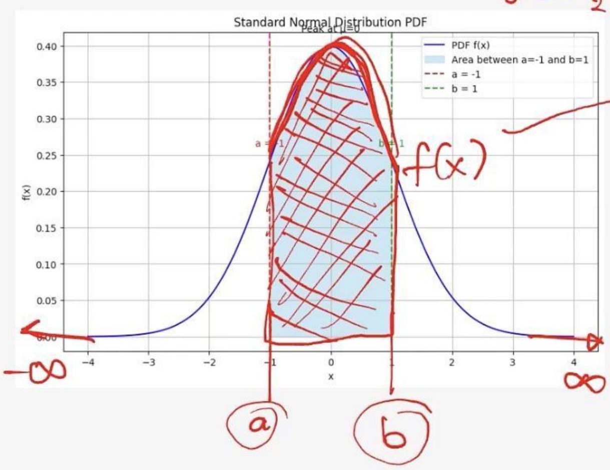
$$\boxed{P(X = x) = 0}$$

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

Area di Bawah Kurva



$\text{Luas} = \int_0^2 y \cdot dx = \int_0^2 1 \cdot dx = x\Big|_2 - x\Big|_0$
$= 2 - 0 = 2$



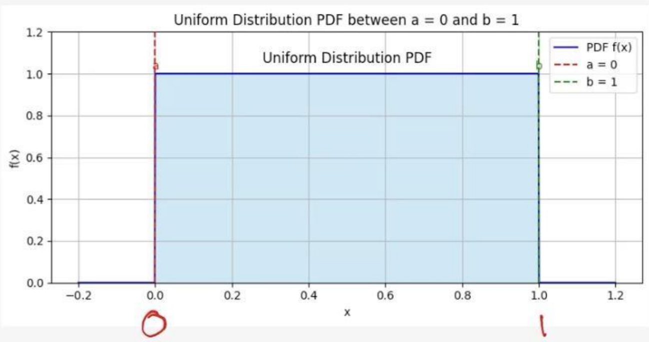
- Probabilitas X berada dalam interval $[a, b]$ dihitung sebagai area di bawah kurva PDF dari a hingga b

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

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

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

Distribusi Uniform (Seragam)

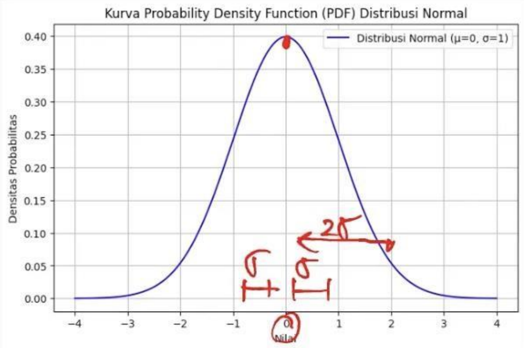


- **Contoh**: Waktu tunggu untuk bus yang datang setiap 10 menit. Waktu tunggu antara 0 dan 10 menit memiliki probabilitas yang sama

- **Kapan digunakan?** Ketika setiap nilai dalam sebuah interval memiliki probabilitas yang sama untuk terjadi.
- **Ciri-ciri**:
  - PDF-nya berbentuk persegi Panjang atau "datar" $a \le x \le b$
  - $f(x) = \frac{1}{b-a}$ untuk $a \le x \le b$ → $f(x) = \frac{1}{b-a}$
  - $f(x) = 0$ di luar interval

Distribusi Normal (Gaussian)

Rata-rata = 0
Standr deviasi = 1



- **PDF**:

e

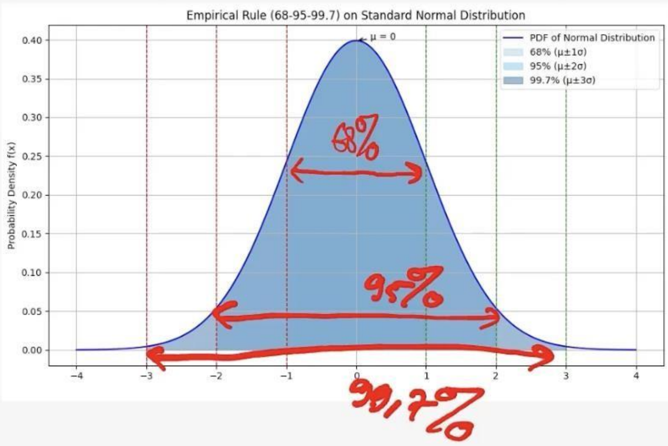
mean,

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

std

- **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 μ
  - Ditentukan oleh dua parameter: mean μ dan standard deviasi σ

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 (μ)** dan **standard deviation (σ)**.
- **Area di bawah kurva PDF** adalah representasi visual dari probabilitas. Total area selalu 1.

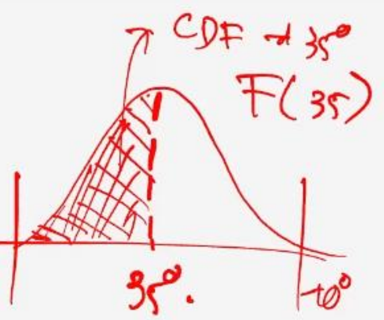
NEXT UP → Cumulative Distribution Function (CDF).

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

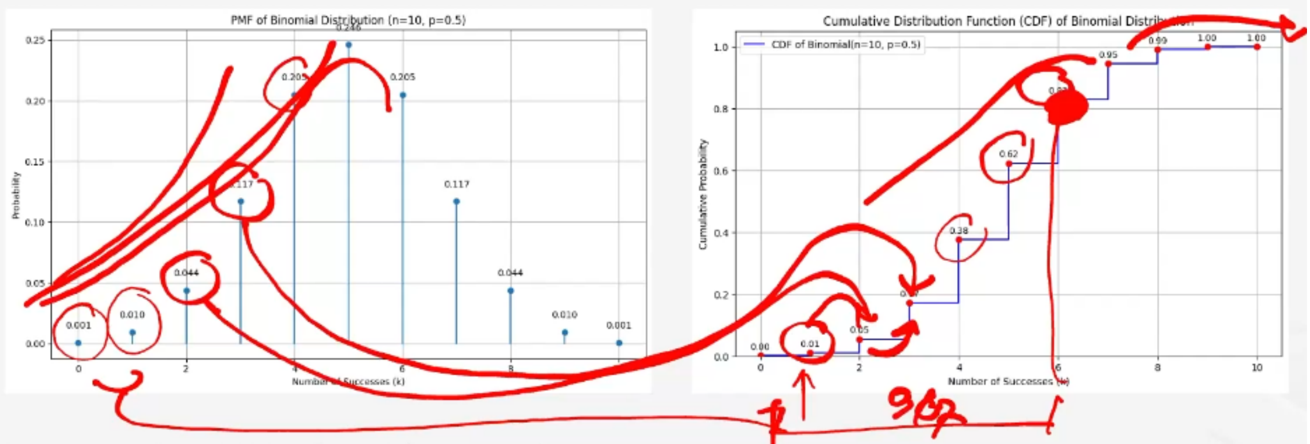
$$F(x) = P(X \leq x)$$

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

- 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



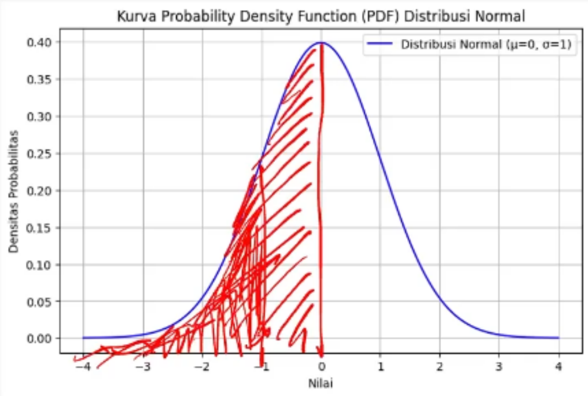
- Hubungan antara PMF dan CDF Adalah bahwa CDF merupakan jumlah dari semua nilai PMF hingga titik x

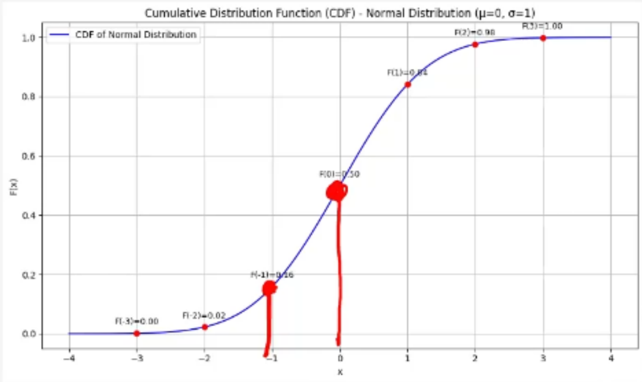
$$F(6) = \sum_{k \leq 6} P(X=k) = P(X=1) + P(X=2)$$
$$P(X=3) \ldots$$
$$P(X=6)$$

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

↑ ↑PMF

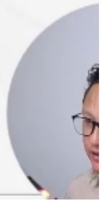
Variabel Kontinu PDF dan CDF



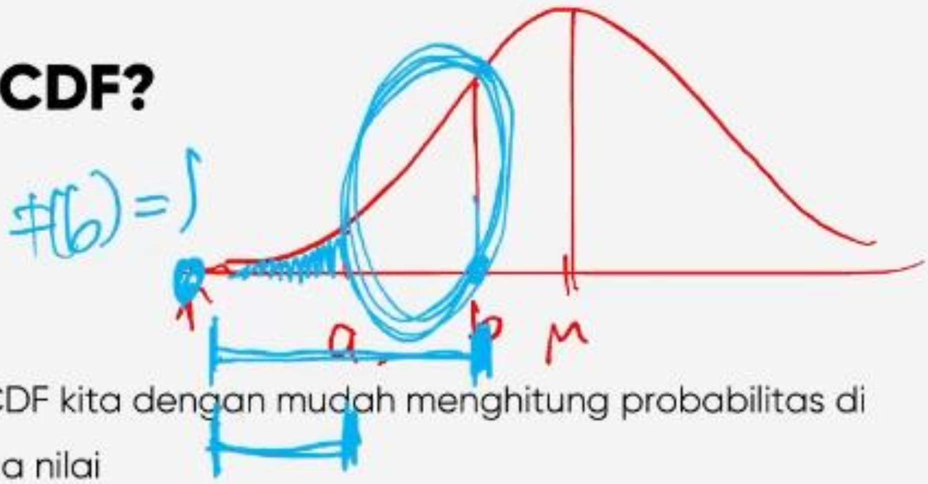


- Hubungan antara PDF dan CDF Adalah bahwa CDF merupakan integral fungsi PDF hingga titik x dari $-\infty$

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



Kenapa CDF?



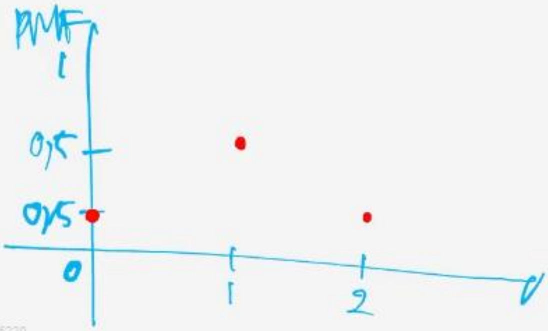
- Dengan CDF kita dengan mudah menghitung probabilitas di antara dua nilai

$$P(a \leq X \leq b) = F(b) - F(a)$$

$= F(b) - F(a).$

- Berlaku untuk distribusi diskrit maupun kontinu

Contoh diskrit:



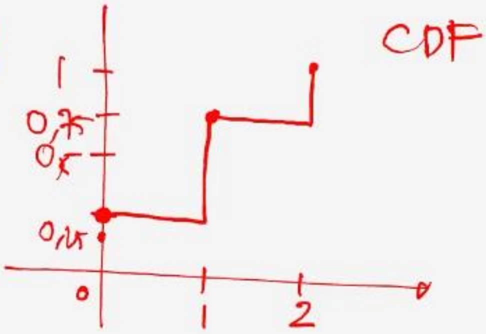
- **Skenario**: Melempar dua koin, A = jumlah Angka
- **Nilai yang mungkin**: $a = \{0,1,2\}$
- **PMF**: $P(A = 0) = 0.25, P(A = 1) = 0.5, P(A = 2) = 0.25.$
- **CDF** $(F(a))$:

AA AG GA GG

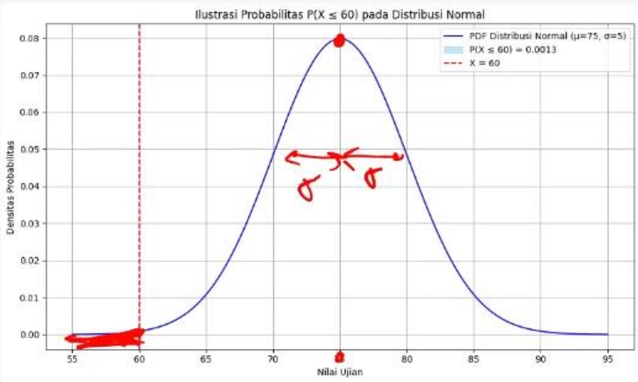
$P(A=0) = 0{,}25$

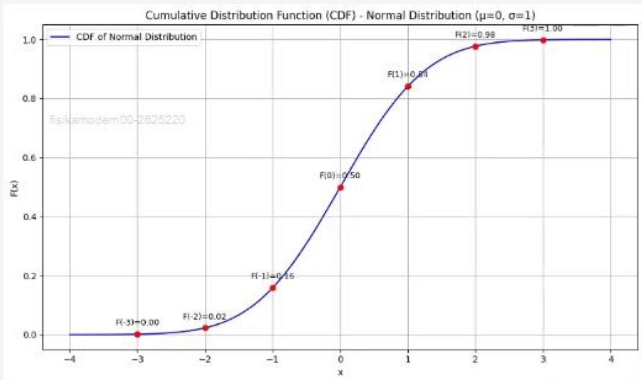
$P(A=1) = 0{,}5$

$P(A=2) = 0{,}25$



Contoh kontinu:





- Seorang mahasiswa ingin mengetahui probabilitas bahwa nilai ujian statistiknya (yang mengikuti distribusi normal standar dengan rata-rata $\mu = 75$ dan standar deviasi $\sigma = 5$) kurang 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 $\boxed{\text{normal } \textbf{standar} \text{ (dengan } \mu = 0, \sigma = 1)}$

$$-15 \quad 5$$

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

- Sehingga

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

$$= 0{,}13\%$$

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

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

$$68\%$$

$$= P(Z < 1) - P(Z \le -1)$$
$$= 0{,}8413 - 0{,}1587 = 0{,}6827$$

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).

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. CDF
PMF Grafik PDF /CDF	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 (μ) dari distribusi ini adalah $n \cdot p$. Varians (σ^2), yang mengukur sebaran data, adalah $n \cdot p \cdot (1 - p)$.

$\mu(1-p)$

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 (λ). 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 (μ) dan varians (σ^2) yang sama, yaitu λ.

- 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 (μ) berada tepat di tengah interval, yaitu $\frac{a+b}{2}$. Variansnya dihitung sebagai $\frac{(b-a)^2}{12}$. $\frac{(b-a)^2}{12}$

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 (μ). 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** (μ) yang menunjukkan pusat distribusi, dan **Standard Deviation** (σ) 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	λ (rata-rata kejadian per interval)	λ	λ
Uniform	Kontinu	a (batas bawah), b (batas atas)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	Kontinu	μ (rata-rata), σ^2 (varians)	μ	σ^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.

4-1. What Is Central Limit Theorem

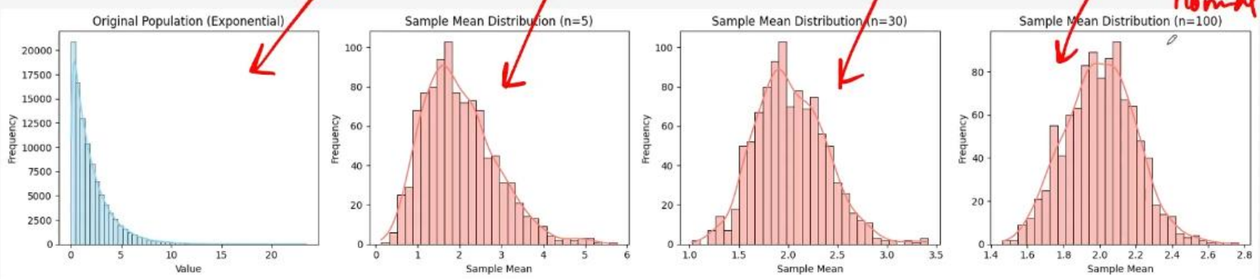
01 **Memahami konsep dasar Teorema Limit Pusat (CLT) dan mengapa ini begitu penting.**

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?

distribusi normal



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?

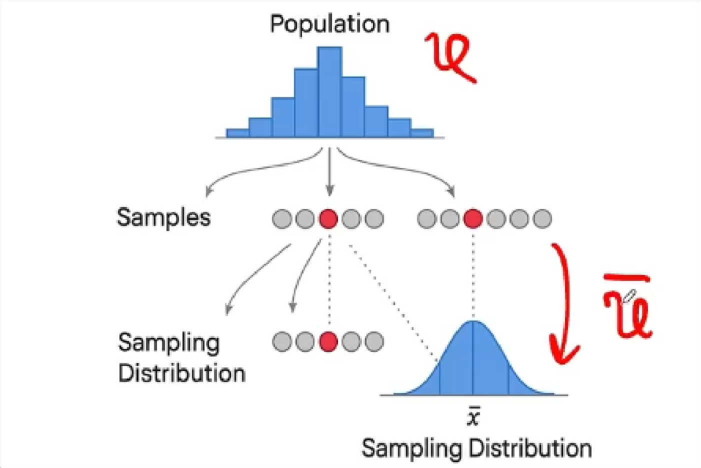
- **Mengapa Ini Luar Biasa?** Teorema ini berlaku meskipun populasi aslinya tidak berdistribusi normal—bisa saja sangat miring (**skewed**), seragam (uniform), atau bentuk apa pun yang tidak terduga. CLT seperti sebuah "*konverter ajaib*" yang selalu menghasilkan kurva lonceng, selama Anda memberinya 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.

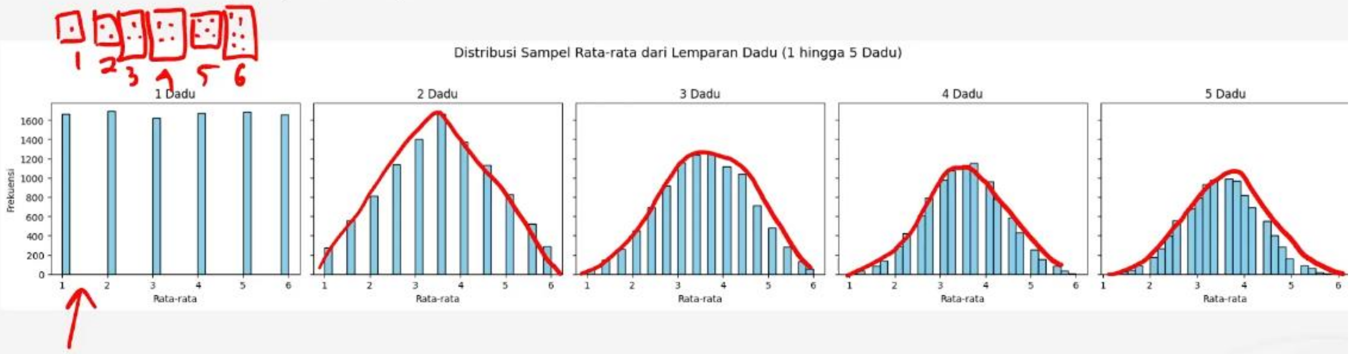
Distribusi Sampel

K=1 N=5 M=? k=2 N=5 M=? K=L N=5 μ=?

- **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.

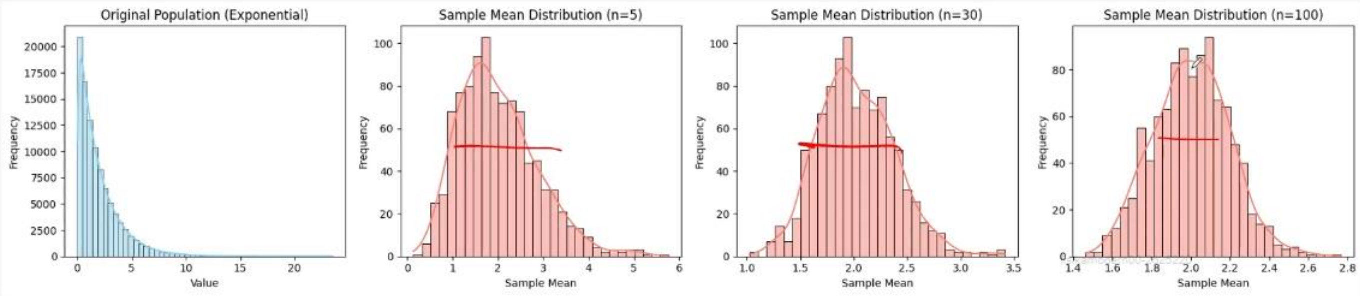


Distribusi Sampel (Contoh)



- 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: $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{1000}$.
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 (μ)

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

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

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

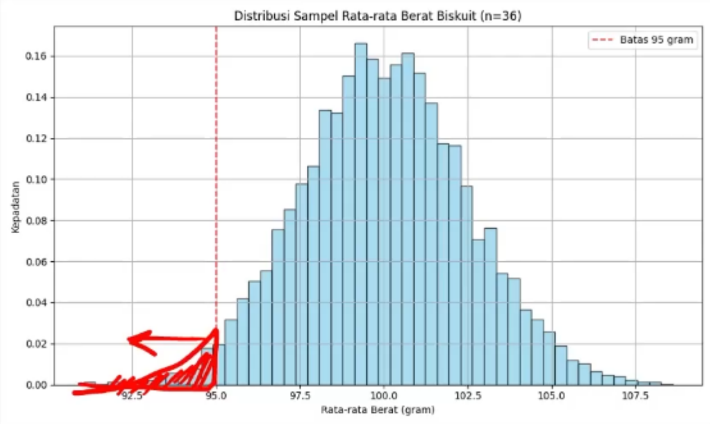
$\sigma = 5$
$n = 10$

Contoh BVAANG!!!: $\sigma = 15$ gram, $n = 36$, $\mu = 100$ 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 kuaitas produksi.

1. Mengapa engineer dapat menggunakan distribusi normal untuk menganalisis rata-rata berat sampel meskipun distribusi populasi tidak normal? $n \geq 30$
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(x...<95)~~ $P(X < 95)$?
4. Jika probabilitas pada soal nomor 3 sangat kecil, apa yang bisa disimpulkan?

Contoh BVAANG!!!:



- 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.

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

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

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{95 - 100}{2{,}5} = -2$$

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

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

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

- Jika proses berjalan dengan normal, kejadian ini sangat jarang terjadi
- Jika rata rata sample di ambil, missal rata rata nya 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≥30).

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:

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

z-score

Aplikasi 1: Interval Kepercayaan

Apa itu?

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

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 ±1.96 standard error dari rata-rata populasi.

$\pm 1{,}96\sigma$ ↑ z-score

- Ini memungkinkan kita untuk membangun interval di sekitar rata-rata sampel kita untuk mengestimasi lokasi μ yang tidak diketahui.

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%

$n = 100$ botol

$\bar{x} = 498$ ml

$\sigma = 5$ ml

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

Interval kepercayaan $= \bar{x} \pm 1{,}96\,\sigma_{\bar{x}}$

$= 498 \pm 1{,}96 \times 0{,}5$

$= 498 \pm 0{,}98$

batas atas → 498,98

batas bawah → 497,02

$497{,}02 \le \mu \le 498{,}98$

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%.

$\bar{x} \pm 1{,}96\,\sigma_{\bar{x}}$

- Hipotesis 0 → $\mu = 65\,kg$ → 95%
- hipotesis 1 → $\mu \neq 65\,kg$ → 5%

$n = 100$

$\bar{x} = 66\,kg$

$\sigma = 10\,kg$

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

Statistik uji z.

$z_{hitung} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{66 - 65}{1\,kg} = -1$

$|z| = 1$

$z_{hitung} < z_{kritis}.$

$z_{hitung} < 1{,}96$

$\cancel{1 < 1{,}96}$

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.