DISTRIBUSI SAMPLING

Prof. Dr. Arna Fariza, S.Kom., M.Kom.

Topik Bahasan

1. Recap: Populasi vs Sampel
2. Pengertian Distribusi Sampling
3. Distribusi Sampling dari Mean
4. Standard Error (SE)
5. Teorema Limit Pusat (Central Limit Theorem)
6. Mengapa CLT Sangat Penting di Bidang IT?
7. Contoh Kasus Terapan (Server & Jaringan)
8. Latihan Soal & Kesimpulan

RECAP - POPULASI VS SAMPEL

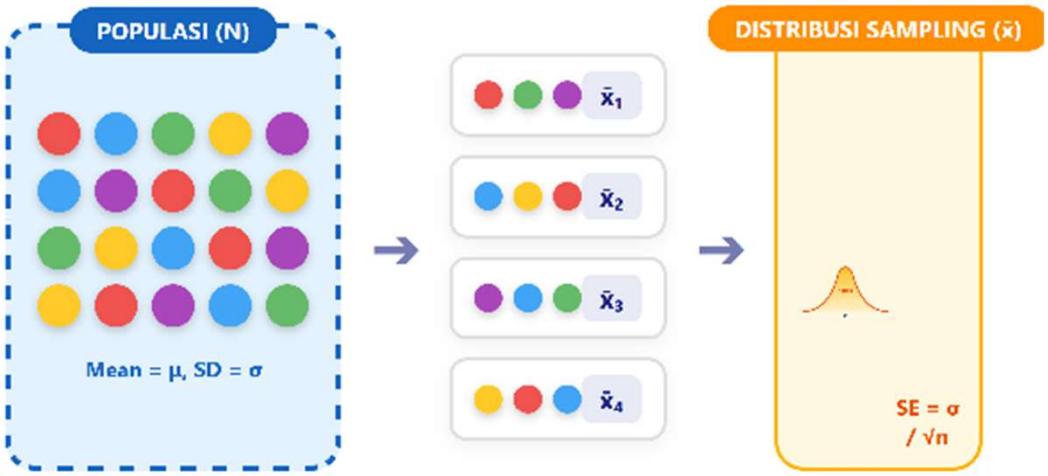
Mengapa Kita Mengambil Sampel?
Meneliti seluruh populasi seringkali tidak mungkin karena keterbatasan waktu, biaya, dan tenaga. Kita mengambil sebagian (sampel) untuk mewakili keseluruhan (populasi).

Aspek	**Populasi**	**Sampel**
Ukuran	N(total)	n(sebagian)
Mean	μ(mu)	$\bar{x}$(x-bar)
Standar Deviasi	σ(sigma)	s
Sifat	Parameter (nilai sebenarnya)	Statistik (estimasi)

Pertanyaan Kunci:
Jika kita mengambil banyak sampel berbeda dari satu populasi yang sama, apakah nilai rata-ratanya ($\bar{x}$) akan selalu sama? **Tidak.** Nilai $\bar{x}$akan bervariasi. Variasi inilah yang dipelajari dalam **Distribusi Sampling**.

PENGERTIAN DISTRIBUSI SAMPLING

Definisi:
Distribusi sampling adalah **distribusi probabilitas** dari suatu statistik (misalnya mean, proporsi, atau standar deviasi) yang dihitung dari **semua kemungkinan sampel** berukuran $n$yang diambil dari suatu populasi.



Kunci: Kita tidak melihat data aslinya, melainkan melihat **bagaimana nilai statistik (misal: rata-rata) bervariasi** antar sampel.

DISTRIBUSI SAMPLING DARI MEAN ($\bar{X}$)

Fokus Utama: Distribusi dari nilai rata-rata sampel ($\bar{x}$).

Sifat-Sifat Penting:

1. **Mean dari Distribusi Sampling:**
   Rata-rata dari semua nilai $\bar{x}$akan sama dengan rata-rata populasi.

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

2. **Standar Deviasi dari Distribusi Sampling:**
   Disebut sebagai **Standard Error** (SE).

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

Perbedaan Penting:

- σ= Variasi nilai individu dalam populasi.
- $\sigma_{\bar{x}}$**(SE)** = Variasi nilai **rata-rata sampel** dari sampel ke sampel.

STANDARD ERROR (SE) - DEFINISI & RUMUS

Definisi:
Standard Error (SE) adalah **standar deviasi dari distribusi sampling**. SE mengukur seberapa jauh rata-rata sampel ($\bar{x}$) diperkirakan akan menyimpang dari rata-rata populasi sebenarnya (μ).

Rumus: $SE = \frac{\sigma}{\sqrt{n}}$

Keterangan:

- σ= Standar deviasi populasi
- n= Ukuran sampel

Jika σ tidak diketahui (dan $n \geq 30$):

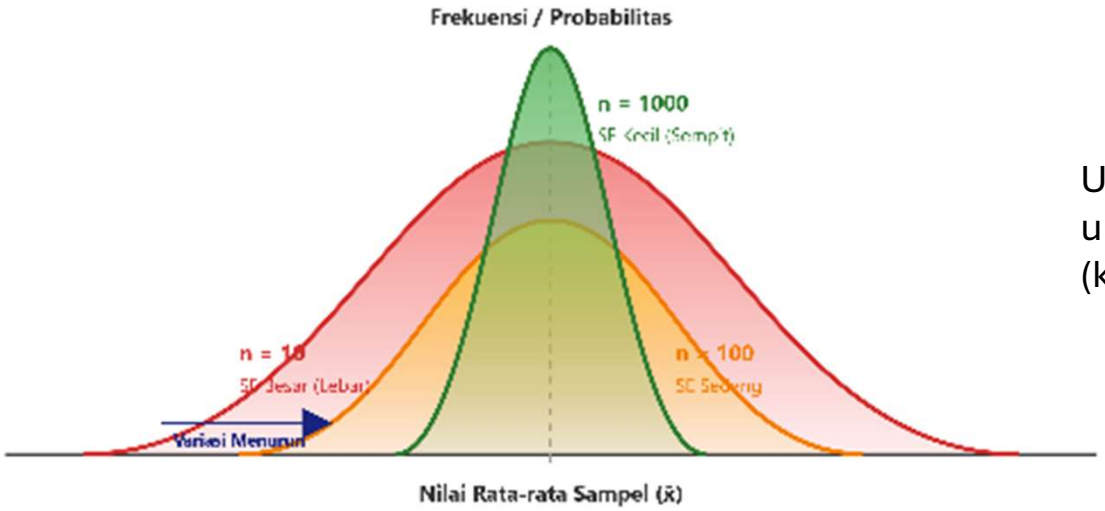
Gunakan standar deviasi sampel (s). $SE = \frac{s}{\sqrt{n}}$

Interpretasi:
Semakin kecil nilai SE, semakin presisi rata-rata sampel kita dalam mengestimasi rata-rata populasi.

FAKTOR YANG MEMPENGARUHI STANDARD ERROR

Bagaimana Ukuran Sampel (n) Mempengaruhi SE?

Ukuran Sampel (n)	Standard Error (SE)	Interpretasi
$n = 10$	$SE = \sigma/\sqrt{10} \approx 0.316\sigma$	Variasi masih cukup besar
$n = 100$	$SE = \sigma/\sqrt{100} = 0.1\sigma$	Variasi lebih kecil
$n = 1000$	$SE = \sigma/\sqrt{1000} \approx 0.031\sigma$	Sangat presisi



Untuk menurunkan SE menjadi setengahnya, ukuran sampel harus dinaikkan **4 kali lipat** (karena akar kuadrat).

TEOREMA LIMIT PUSAT (CLT) - KONSEP UTAMA

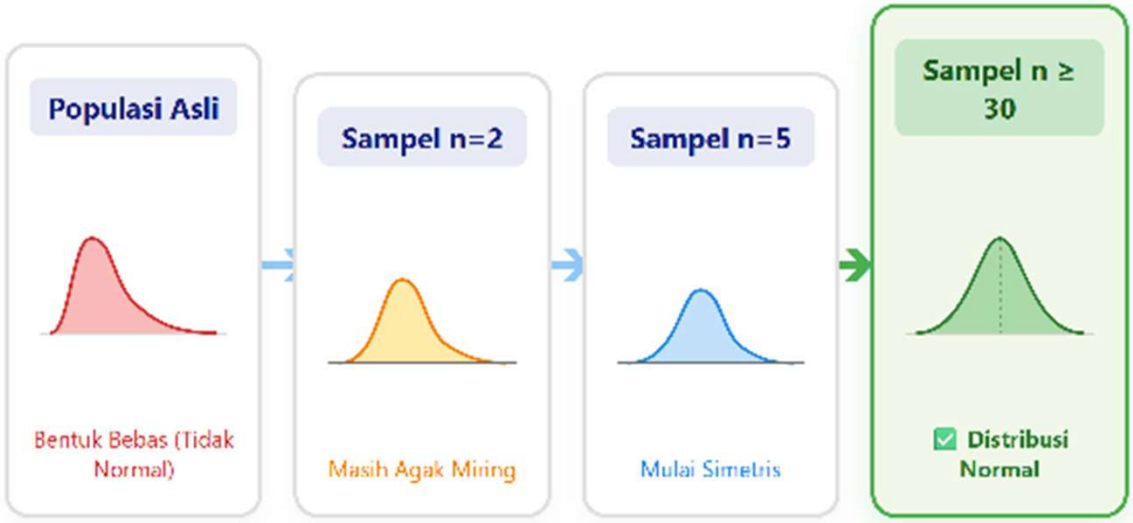
Apa itu Teorema Limit Pusat (Central Limit Theorem)?
Ini adalah salah satu teorema paling penting dalam statistika. CLT menyatakan bahwa:
"Jika ukuran sampel $n$cukup besar (biasanya $n \geq 30$), maka distribusi sampling dari mean ($\bar{x}$) akan mendekati Distribusi Normal, TERLEPAS dari bentuk distribusi populasi aslinya.“

Mengapa Ini Luar Biasa?

1. Populasi asli bisa berbentuk apa saja (miring, seragam, bimodal).
2. Selama kita mengambil sampel yang cukup besar ($n \geq 30$), rata-rata sampelnya akan membentuk **Kurva Lonceng (Normal)**.
3. Ini memungkinkan kita menggunakan **Tabel Z** dan rumus Normal untuk menganalisis sampel.

VISUALISASI TEOREMA LIMIT PUSAT (CLT)

Simulasi Perubahan Bentuk Distribusi:



Kunci Visual:

Apa pun bentuk populasi aslinya (kiri), distribusi rata-rata sampel (kanan) akan membentuk lonceng normal. Semakin besar n, semakin sempurna bentuk loncengnya.

Aturan Praktis $n \geq 30$:

- Jika populasi asli **sudah Normal**, CLT berlaku bahkan untuk $n$kecil.
- Jika populasi asli **tidak Normal**, kita butuh $n \geq 30$agar distribusi sampling mendekati Normal.

MENGAPA CLT SANGAT PENTING DI BIDANG IT?

1. **A/B Testing & Analisis Konversi**
   Saat menguji dua desain UI, data konversi (ya/tidak) tidak berdistribusi normal. Namun, karena kita mengambil sampel pengguna ($n > 30$), kita bisa menggunakan CLT untuk menguji apakah perbedaan konversi signifikan.
2. **Evaluasi Kinerja Server & SLA**
   Waktu respons server seringkali *skewed* (banyak request cepat, sedikit yang lambat). Dengan CLT, kita bisa mengestimasi rata-rata waktu respons sistem secara keseluruhan dari sampel log server.
3. **Machine Learning & Data Science**
   Algoritma Machine Learning sering mengasumsikan data berdistribusi normal. CLT memungkinkan kita menggunakan statistik parametrik untuk mengevaluasi model pada dataset yang tidak normal asalkan jumlah sampelnya besar.
4. **Deteksi Anomali Jaringan**
   Kita bisa membuat *confidence interval* untuk lalu lintas jaringan normal. Jika rata-rata sampel saat ini berada di ekor distribusi (Z-score ekstrem), itu bisa menjadi indikasi serangan siber.

CONTOH KASUS TERAPAN (IT)

Kasus: Kinerja Server Web
Sebuah server web memiliki waktu respons yang berdistribusi tidak normal (miring ke kanan) dengan rata-rata populasi $\mu = 200$ms dan standar deviasi populasi $\sigma = 40$ms. Seorang engineer mengambil sampel acak sebanyak $n = 64$**request**.

Pertanyaan:
Berapa peluang rata-rata waktu respons sampel ($\bar{x}$) kurang dari **190 ms**?

Penyelesaian:
1. **Cek Asumsi CLT:** Karena $n = 64 \geq 30$, maka distribusi $\bar{x}$mendekati Normal (meskipun data asli tidak normal).
2. **Hitung Standard Error (SE):**

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = \frac{40}{8} = \mathbf{5\ ms}$$

CONTOH KASUS TERAPAN (IT)

3. **Hitung Z-score untuk sampel:**

$$Z = \frac{\bar{x} - \mu}{SE} = \frac{190 - 200}{5} = \frac{-10}{5} = \mathbf{-2.00}$$

4. **Cari Peluang di Tabel Z:**

$$P(\bar{x} < 190) = P(Z < -2.00) = \mathbf{0.0228}$$

Kesimpulan: Peluang rata-rata waktu respons sampel kurang dari 190 ms adalah **2.28%**.
(Ini adalah kejadian yang cukup langka, mungkin perlu diinvestigasi).

LATIHAN SOAL

Soal 1:
Sebuah populasi memiliki rata-rata 100 dan standar deviasi 15. Jika diambil sampel berukuran $n = 25$, hitunglah Standard Error (SE)!

Soal 2:
Sebuah server database memiliki waktu query rata-rata 50 ms dengan standar deviasi 10 ms. Diambil sampel 36 query. Berapa peluang rata-rata waktu query sampel berada di antara 48 ms dan 52 ms? (Gunakan pendekatan CLT).

Soal 3:
Jelaskan mengapa Teorema Limit Pusat (CLT) sangat berguna dalam analisis data jaringan komputer yang seringkali tidak berdistribusi normal.

JAWABAN LATIHAN SOAL

Jawaban Soal 1:

- Diketahui: $\sigma = 15, n = 25$
- $SE = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = \mathbf{3}$

Jawaban Soal 2:

- Diketahui: $\mu = 50, \sigma = 10, n = 36$
- $SE = 10/\sqrt{36} = 10/6 \approx 1.667$
- $Z_1 = (48 - 50)/1.667 = -1.20$
- $Z_2 = (52 - 50)/1.667 = 1.20$
- $P(-1.20 < Z < 1.20) = P(Z < 1.20) - P(Z < -1.20)$
- $= 0.8849 - 0.1151 = \mathbf{0.7698}$atau **76.98%**

Jawaban Soal 3:

Karena data jaringan (seperti latensi atau ukuran paket) seringkali tidak simetris (miring/heavy-tailed), kita tidak bisa langsung menggunakan rumus normal. Namun, dengan CLT, selama kita mengambil sampel data yang cukup besar ($n \geq 30$), **distribusi dari rata-rata sampel** akan menjadi normal. Ini memungkinkan engineer untuk menggunakan uji statistik standar (seperti Z-test atau T-test) untuk menganalisis kinerja jaringan.

Ringkasan

1. **Distribusi Sampling** adalah distribusi probabilitas dari suatu statistik (misal: mean) dari semua kemungkinan sampel.
2. **Distribusi Sampling dari Mean** memiliki rata-rata $\mu_{\bar{x}} = \mu$.
3. **Standard Error (SE)** mengukur variasi rata-rata sampel, dengan rumus $SE = \sigma/\sqrt{n}$. Semakin besar n, semakin kecil SE (semakin presisi).
4. **Teorema Limit Pusat (CLT)** menyatakan bahwa untuk $n \geq 30$, distribusi rata-rata sampel akan mendekati **Normal**, terlepas dari bentuk populasi aslinya.
5. **Aplikasi IT:** Sangat penting untuk A/B Testing, evaluasi kinerja server (SLA), dan analisis data yang tidak berdistribusi normal.

TUGAS

Soal 1:
Sebuah populasi data memiliki rata-rata 75 dan standar deviasi 12. Jika diambil sampel acak berukuran $n = 144$, hitunglah Standard Error (SE) dari distribusi sampling mean-nya!

Soal 2:
Waktu eksekusi sebuah algoritma memiliki rata-rata 30 detik dan standar deviasi 5 detik. Jika diambil sampel sebanyak 100 kali eksekusi, berapa peluang rata-rata waktu eksekusi sampel lebih dari 31 detik? (Asumsikan distribusi mendekati normal karena CLT).

Soal 3 (Aplikasi IT):
Sebuah ISP (Internet Service Provider) mencatat bahwa rata-rata penggunaan bandwidth per pengguna adalah 50 Mbps dengan standar deviasi 15 Mbps. Jika diambil sampel 50 pengguna, berapa peluang rata-rata penggunaan bandwidth mereka kurang dari 46 Mbps?