Sampling & Distribusi Sampling

Jembatan inferensial dari data sampel terbatas menuju populasi: Teknik sampling, Kesalahan Standar (SE), dan Teorema Limit Pusat (CLT)

Statistika dan Probabilitas · Pertemuan 07

Peta Pertemuan 07

1 Pondasi Sampling & Bias Representasi

Mengapa sensus mustahil, pemisahan parameter vs statistik, dan bahaya bias seleksi.

2 Teknik Penarikan Sampel (Sampling Methods)

Probabilitas (SRS, Stratified, Cluster, Systematic, Multistage) & Non-Probabilitas (Convenience, Purposive, Snowball, Quota).

3 Distribusi Sampling & Standard Error

Perilaku rata-rata sampel ($\overline{X}$), penaksir tak bias ($\mu_{\overline{X}} = \mu$), dan reduksi galat ($\frac{\sigma}{\sqrt{n}}$).

4 Teorema Limit Pusat (Central Limit Theorem)

Konvergensi kurva lonceng normal dari populasi apa pun saat $n \geq 30$, serta distribusi proporsi.

Blok 1 — Mengapa Perlu Sampling?

Keterbatasan sensus, pemisahan parameter vs statistik, dan bahaya bias

Pemantik Diskusi: Mengetes Rasa Kuah Sup

PERTANYAAN PEMICU

- Ketika seorang juru masak memasak kuah sup di dalam panci raksasa bervolume 50 liter…
- Apakah juru masak harus meminum **seluruh 50 liter kuah sup sampai habis** untuk memastikan rasanya sudah pas?
- Tentu tidak! Cukup mengaduk sup tersebut hingga merata, lalu mencicipi **satu sendok makan saja**.

Analogi Sup dalam Bahasa Statistika

Analogi Kuliner

- **Panci 50 Liter:**
  Seluruh kuah yang ada di dalam panci.
- **Mengaduk Merata:**
  Memastikan bumbu tersebar merata di setiap sudut.
- **Satu Sendok Cicipan:**
  Bagian kecil yang diambil untuk dirasakan.
- **Rasa di Sendok:**
  Kesan rasa dari satu sendok cicipan.

Konsep Statistika Inferensial

- **Populasi (N):**
  Semesta objek yang ingin disimpulkan.
- **Random Sampling:**
  Setiap unit memiliki kesempatan setara untuk terpilih.
- **Sampel (n):**
  Subhimpunan terukur yang diamati langsung.
- **Statistik ($\overline{x}$):**
  Ukuran sampel untuk menduga **Parameter** (μ).

Mengapa Sensus Lengkap Sering Mustahil?

Ada 4 kendala utama mengapa kita **tidak bisa** mengukur seluruh populasi:

1. Biaya & Sumber Daya

Mengukur jutaan pengguna atau perangkat membutuhkan anggaran miliaran dan server raksasa.

2. Kecepatan Keputusan

Sensus butuh waktu berbulan-bulan; tren data sains dan bisnis berubah dalam hitungan jam.

3. Uji Bersifat Merusak (Destructive)

Crash test mobil atau uji umur baterai. Jika disensus, seluruh barang hancur dan tidak ada yang bisa dijual!

4. Populasi Tak Hingga

Aliran paket data jaringan (**streaming**), klik iklan web, atau transaksi yang terus diproduksi tanpa henti.

Pembeda Mutlak: Parameter vs Statistik

Statistika membedakan secara tegas simbol dan makna antara populasi dan sampel:

Ukuran Data	Parameter (Populasi)	Statistik (Sampel)	Karakteristik Kunci
Banyaknya Data	N (Kapital)	n (Kecil)	Sampel adalah bagian dari populasi ($n < N$)
Rata-Rata (**Mean**)	μ (**mu**)	$\overline{x}$ (**x-bar**)	μ bernilai tetap; $\overline{x}$ berubah tiap sampel
Varians	σ^2	s^2	Varians sampel menggunakan pembagi $(n-1)$
Simpangan Baku	σ	s	Satuan persebaran asli data
Proporsi Sukses	p	$\hat{p}$ (**p-hat**)	Rasio sukses terhadap total: $\hat{p} = \frac{X}{n}$

Prinsip Emas Statistika Inferensial

Prinsip Emas yang Wajib Diingat Selamanya

Parameter Populasi (μ, σ, p) bernilai tetap dan pasti, namun hampir tidak pernah kita ketahui nilainya di dunia nyata.

Sebaliknya, **Statistik Sampel ($\bar{x}$, s, $\hat{p}$)** dapat kita hitung secara presisi dari data di tangan, tetapi nilainya **berubah-ubah (berfluktuasi)** tergantung sampel acak mana yang terambil.

Musuh Terbesar Sampling 1: Bias Seleksi

KASUS NYATA: PILPRES AMERIKA SERIKAT 1936

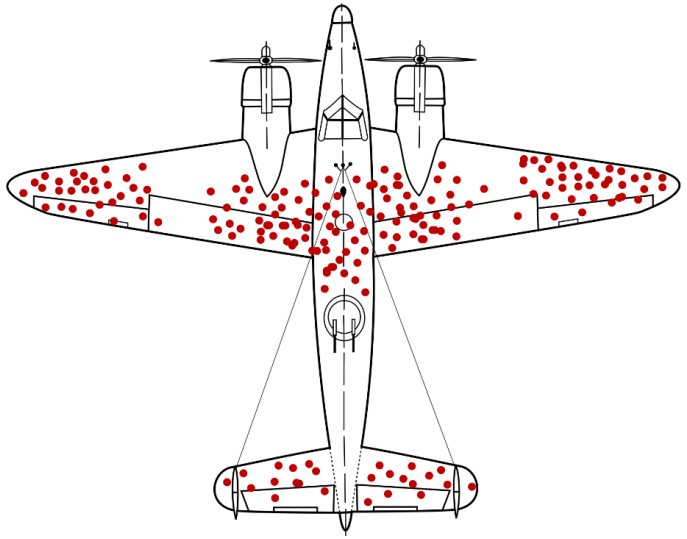
Majalah ternama **Literary Digest** menyurvei **2,4 juta responden** melalui nomor telepon dan data kepemilikan mobil. Hasil survei meramalkan Alf Landon akan menang mutlak 57%.

Fakta Lapangan: Franklin D. Roosevelt menang telak dengan perolehan 62%!

Mengapa Sampel 2,4 Juta Orang Bisa Meleset Total?

Di tengah krisis ekonomi **Great Depression 1936**, hanya golongan orang kaya yang sanggup memiliki telepon dan mobil. Sampel berukuran jutaan orang sekalipun **tidak ada artinya jika proses pengambilannya cacat dan bias sistematis**!

Musuh Terbesar Sampling 2: Kasus Pesawat Tempur PD II



Pola sebaran lubang peluru (titik merah) pada pesawat tempur yang berhasil kembali ke pangkalan.

KASUS NYATA: ANGKATAN UDARA PD II

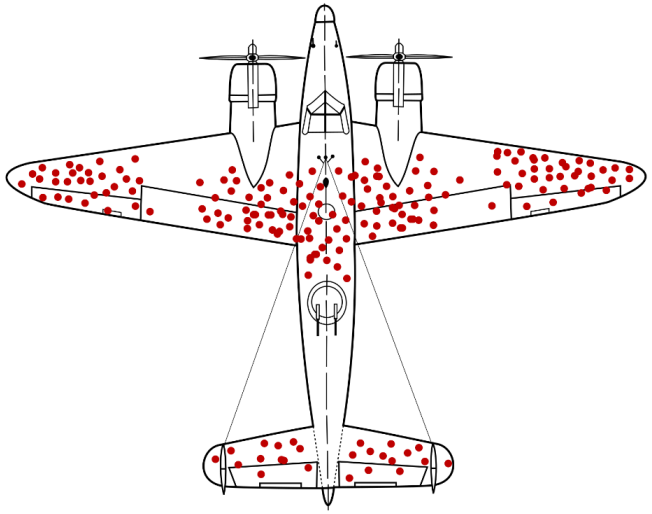
Militer memeriksa pesawat tempur yang selamat kembali ke pangkalan. Sebagian besar lubang peluru terkonsentrasi di **sayap**, **badan**, dan **ekor**.

Rencana Awal Pihak Militer

Memasang lapis baja pelindung ekstra pada **bagian sayap dan ekor** yang paling banyak terkena tembakan peluru musuh.

Pertanyaan: Apakah keputusan militer ini tepat secara logika sampling?

Koreksi Brilian Abraham Wald: Bias Kesintasan



Titik fatal sesungguhnya: Mesin & Kokpit

Koreksi Abraham Wald: "Keputusan Militer Salah Fatal!"

"Pesawat yang tertembak di mesin TIDAK PERNAH kembali pulang untuk dihitung ke dalam sampel!"

Mengapa Disebut Survivorship Bias?

- Lubang peluru di sayap membuktikan pesawat **masih sanggup bertahan dan terbang pulang** meski tertembak.
- Bagian mesin tampak "mulus" **karena pesawat yang tertembak di mesin langsung jatuh di medan perang** (hilang dari sampel).
- **Solusi Tepat:** Lapis baja pelindung harus dipasang pada **mesin dan kokpit**!

Hukum Mutlak Kualitas Data

Sampel berukuran $n = 500$ yang representatif

JAUH LEBIH BERHARGA & AKURAT

daripada sampel $n = 5.000.000$ yang mengandung bias seleksi!

Studi Kasus Blok 1 (Kasus 1): Tragedi Polling Pilpres 1936

KASUS SEJARAH: LITERARY DIGEST VS GEORGE GALLUP

Pada Pilpres AS 1936, majalah ternama **Literary Digest** menyebar 10 juta kartu pos survei dan menerima "2,3" juta balasan ($n = 2.300.000$). Mereka meramalkan Alf Landon menang telak 57% mengalahkan Franklin D. Roosevelt.

Sebaliknya, George Gallup hanya mengambil sampel acak $n = 50.000$ orang. Hasil pemilu nyata: Roosevelt menang telak dengan raihan suara 62%!

Pertanyaan Diskusi

1. Mengapa sampel raksasa "2,3" juta responden bisa salah total, sedangkan sampel 50 ribu orang milik Gallup terbukti akurat?
2. Sebutkan dan jelaskan dua jenis bias metodologis utama yang menjebak **Literary Digest**!

Pembahasan Kasus 1: Mengapa Sampel “2,3” Juta Responden Gagal?

1. Selection Bias (Kerangka Sampel Cacat)

• **Daftar Sampel Cacat:** Majalah mengumpulkan alamat dari buku telepon dan registrasi mobil mewah. Pada masa Depresi Besar (1936), hanya kalangan kaya yang memiliki telepon & mobil (mayoritas pendukung Landon).
• Rakyat miskin yang merupakan pemilih mayoritas sama sekali tidak terdaftar di **sampling frame**.

2. Non-Response Bias (Responden Sukarela)

• Dari 10 juta kartu pos, hanya 23% yang membalas. Orang yang sukarela membalas umumnya mereka yang sangat kecewa terhadap kebijakan ekonomi pemerintah Roosevelt (opini ekstrem).

> **Pelajaran Rekayasa Data: Ukuran sampel raksasa (n jutaan) tidak ada gunanya jika mekanismenya cacat (Garbage In, Garbage Out). Kualitas representasi jauh lebih utama daripada kuantitas!**

Studi Kasus Blok 1 (Kasus 2): Evaluasi Retensi Fitur AI Mobile

SKENARIO STARTUP REKAYASA PERANGKAT LUNAK

Tim Data Science sebuah startup meluncurkan fitur asisten AI baru di aplikasi mobile mereka. Untuk mengukur kepuasan pengguna, tim mengumpulkan survei dan log aktivitas dari pengguna aktif harian (**Daily Active Users** / DAU) di **minggu ke-4** setelah peluncuran.

Hasilnya sangat memuaskan: skor kepuasan mencapai 92% dan pengguna melaporkan latensi AI sangat responsif. Manajemen langsung memutuskan memperluas fitur ini ke jutaan pengguna.

Pertanyaan Diskusi

1. Kesalahan fatal apa yang dilakukan tim dalam mendefinisikan sampel survei?
2. Mengapa skenario ini merupakan contoh nyata dari **Survivorship Bias** (Bias Kesintasan)?

Pembahasan Kasus 2: Jebakan Survivorship Bias pada Pengguna

Manifestasi Survivorship Bias dalam Produk Digital

- Pengguna yang mengalami **crash**, respons AI ngawur, atau latensi lemot telah **menghapus aplikasi (churn)** pada minggu pertama atau kedua.
- Pengguna yang masih bertahan aktif di minggu ke-4 (**survivors**) hanyalah mereka yang kebetulan memiliki koneksi internet kencang atau merasa fitur tersebut cocok.
- Tim hanya menyurvei "pesawat yang selamat kembali ke pangkalan", mengabaikan pengguna yang "gugur di medan tempur".

Solusi Metodologis yang Tepat

- **Cohort Analysis:** Sampel harus ditarik sejak Hari ke-1 peluncuran (**Day-1 cohort**) dan memantau seluruh pengguna, termasuk yang berhenti memakai aplikasi.
- Lakukan wawancara khusus pada kelompok pengguna yang melakukan **uninstall** untuk mengetahui titik kegagalan sistem.

Blok 2 — Teknik Penarikan Sampel

Membedakan metode probabilitas ilmiah versus non-probabilitas

Dua Kategori Besar Penarikan Sampel

1. Probability Sampling (Ilmiah)

- Peluang tiap unit untuk terpilih **diketahui secara pasti** dan bukan nol.
- Menghilangkan bias subjektif peneliti.
- **Syarat mutlak** untuk menerapkan statistika inferensial dan pengujian hipotesis.
- Memungkinkan estimasi **margin of error**.

2. Non-Probability Sampling

- Peluang unit untuk terpilih **tidak diketahui** secara matematis.
- Rentan terhadap bias subjektif (**convenience** atau **self-selection**).
- **Tidak dapat** digunakan untuk generalisasi populasi formal.
- Sangat berguna untuk studi eksplorasi awal, **pilot project**, atau uji purwarupa (**usability testing**).

Metode 1: Simple Random Sampling (SRS)

Konsep Dasar Simple Random Sampling

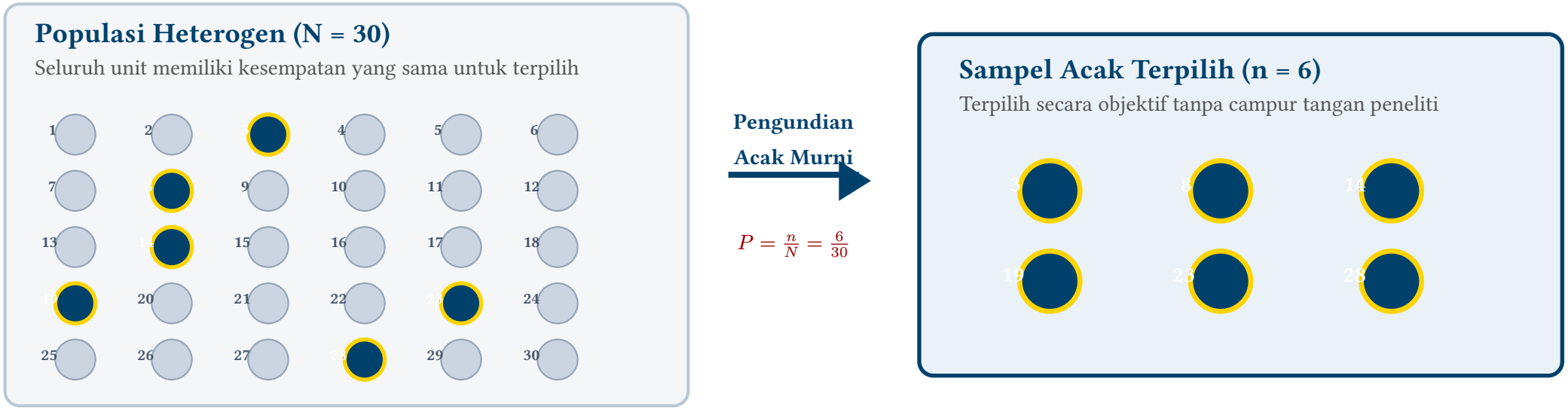
Setiap unit tunggal di dalam populasi memiliki peluang yang **setara persis** untuk terpilih ke dalam sampel.

Peluang Terpilih: Setiap unit memiliki peluang $P = \frac{n}{N}$.

Cara Kerja & Implementasi Komputasi

1. Berikan nomor identitas unik 1 hingga N pada seluruh anggota populasi (**sampling frame**).
2. Bangkitkan n angka acak tanpa pengembalian menggunakan **Random Number Generator** (RNG).
3. Di Python: `numpy.random.choice(populasi, size=n, replace=False)`.
4. Di SQL: `SELECT * FROM users ORDER BY RAND() LIMIT n;`.

Ilustrasi Visual: Simple Random Sampling (SRS)



Karakteristik Utama: Sangat ideal jika kerangka sampel (sampling frame) lengkap dan populasi relatif seragam.

Metode 2: Stratified Random Sampling (Berstrata)

Konsep Dasar Stratified Sampling

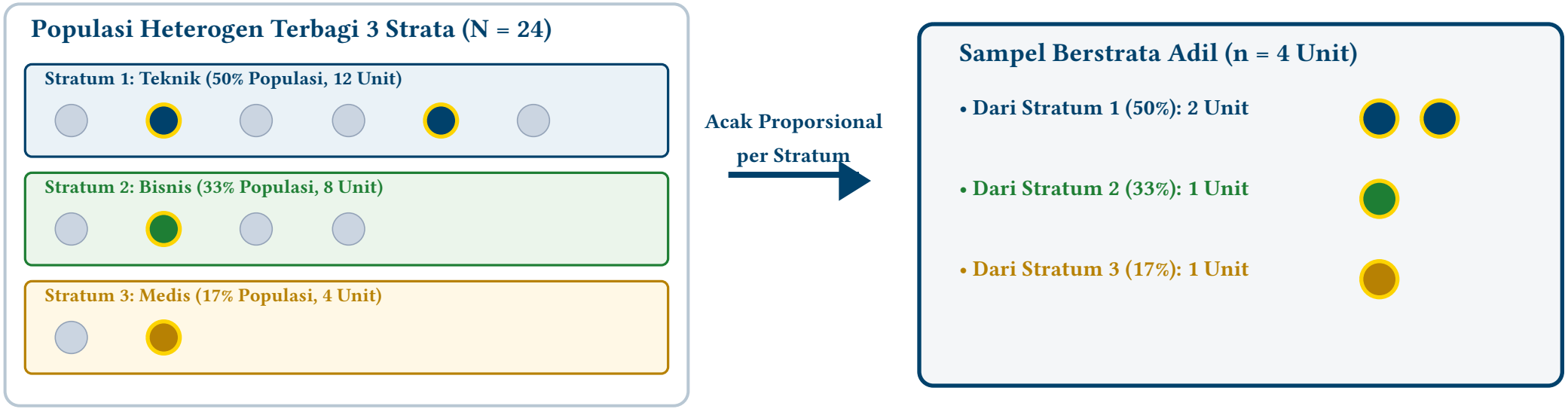
Populasi yang **heterogen** dipecah terlebih dahulu menjadi kelompok-kelompok yang **homogen** di dalamnya (**disebut Strata**).

Lalu dari setiap stratum, sampel acak diambil secara **proporsional** sesuai ukuran kelompoknya di populasi.

Contoh Nyata & Keunggulan

- **Kasus Survei Mahasiswa Satu Universitas:**
Jangan acak murni! Bagi dulu per fakultas (Teknik 40%, Ekonomi 30%, Kedokteran 20%, Hukum 10%). Ambil sampel dengan proporsi yang sama.
- **Keunggulan Utama:** Menjamin kelompok minoritas **pasti terwakili secara adil** dan tidak tenggelam oleh kelompok mayoritas.

Ilustrasi Visual: Stratified Random Sampling (Berstrata)



Keunggulan Mutlak: Sub-populasi minoritas (seperti Medis) dijamin terwakili adil dan tidak tertinggal.

Metode 3: Cluster Sampling (Sampling Klaster / Area)

Konsep Dasar Cluster Sampling

Populasi dibagi menjadi kelompok-kelompok alamiah (**Klaster**, misal: sekolah, kelurahan, gedung kantor).

Alih-alih mengacak individu, peneliti **mengacak klaster secara utuh**. Seluruh anggota dalam klaster terpilih akan diukur!

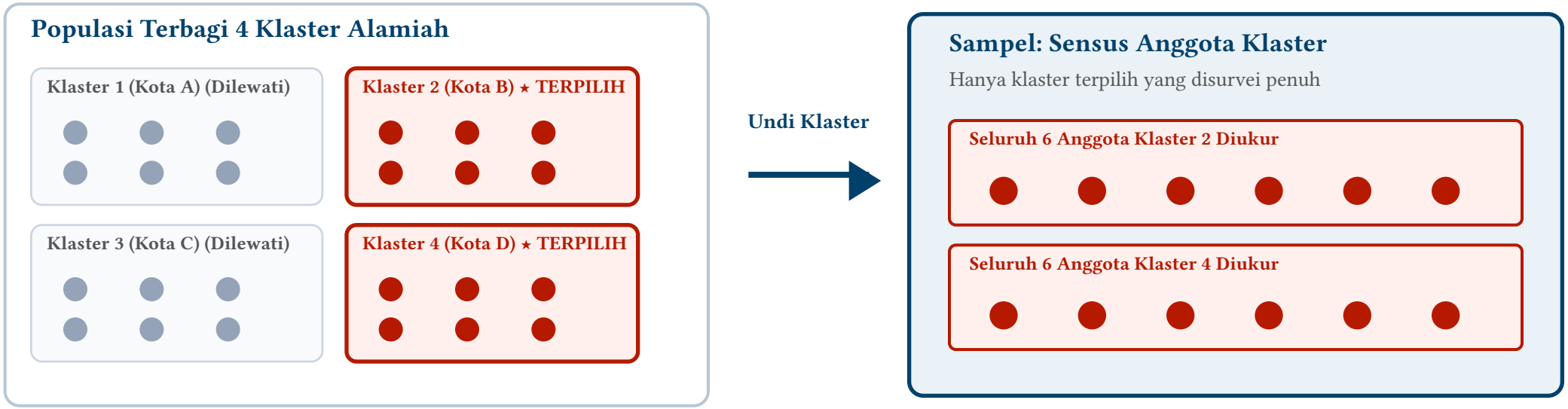
Stratified Sampling

- Strata = **Homogen di dalam**.
- Semua strata diambil sampelnya.
- Tujuan: Presisi representasi kelompok.

Cluster Sampling

- Klaster = **Heterogen di dalam** (miniatur populasi).
- Hanya klaster terpilih yang diamati.
- Tujuan: Efisiensi biaya & logistik geografis.

Ilustrasi Visual: Cluster Sampling (Sampling Klaster)



Efisiensi Lapangan: Peneliti cukup mendatangi Kota B dan Kota D, menghemat 50% biaya perjalanan!

Metode 4: Systematic Sampling (Sistematis)

Konsep Dasar Systematic Sampling

Memilih unit sampel secara periodik dengan interval keteraturan yang tetap:

$$k = \frac{N}{n} \quad \text{(Interval Sampling)}$$

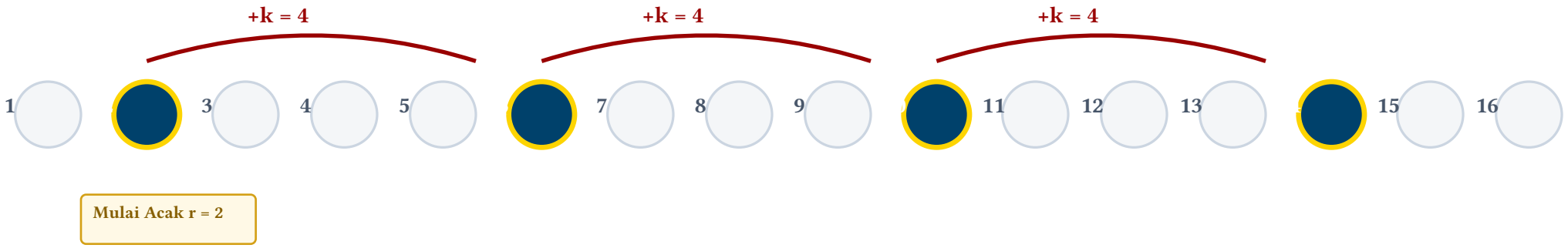
Prosedur Kerja & Jebakan Pola Periodisitas

1. Tentukan nomor awal r secara acak antara 1 s.d. k.
2. Ambil sampel berikutnya pada urutan: $r, r+k, r+2k, r+3k, \ldots$
3. **Peringatan Fatal:** Jangan gunakan metode ini jika data memiliki siklus berulang (misal mengambil data penjualan setiap hari ke-7 akan selalu jatuh di hari Minggu!).

Ilustrasi Visual: Systematic Sampling (Sistematis)

Populasi Tersusun Rapi (N = 16) dengan Titik Awal r = 2 dan Interval k = 4

Menghitung interval sampling: k = N / n = 16 / 4 = 4



Sampel Sistematis Terpilih: Unit 2, Unit 6, Unit 10, Unit 14

- Titik awal r = 2 diundi secara acak antara 1 s.d. 4. Lalu setiap unit kelipatan 4 diambil secara otomatis.
- **Peringatan: Sangat cepat pada lini perakitan pabrik, namun hindari jika data memiliki pola musiman berulang!**

Metode 5: Multistage Sampling (Sampling Bertingkat)

Konsep Dasar Multistage Sampling

Kombinasi hierarkis bertingkat dari beberapa teknik sampling untuk survei berskala nasional atau geografis luas. Penarikan sampel dilakukan bertahap dari unit wilayah terbesar hingga unit terkecil.

Tahapan Kerja Survei Nasional (Contoh: Sensus BPS / Quick Count)

- **Tahap 1 (Provinsi):** Pilih 10 dari 38 provinsi secara acak (**Klaster Wilayah**).
- **Tahap 2 (Kabupaten/Kota):** Dari tiap provinsi terpilih, pilih 5 kabupaten/kota acak.
- **Tahap 3 (TPS / Desa):** Dari tiap kabupaten terpilih, pilih 4 TPS secara acak.
- **Tahap 4 (Rumah Tangga):** Di tiap TPS, ambil 10 KK dengan **Simple Random Sampling**.

> **Keunggulan: Peneliti tidak perlu menyusun daftar 270 juta nama warga negara di awal. Sampling frame hanya perlu dibuat untuk unit di TPS terpilih pada tahap akhir!**

Transisi: Ranah Non-Probability Sampling

Kapan Non-Probability Boleh & Tepat Digunakan?

- **Studi Eksplorasi & Pilot Study:** Menguji kuesioner atau konsep awal sebelum riset skala besar berbiaya miliaran rupiah.
- **Populasi Tersembunyi (**Hard-to-reach**):** Kelompok tanpa **sampling frame** resmi (misal: komunitas **ethical hacker**, korban perundungan).
- **Pengujian Perangkat Lunak (UI/UX Usability):** Menguji aplikasi pada 10 pengguna di lab untuk menemukan kendala alur kerja (**bugs**).

Peringatan Ilmiah Mutlak: Batasan Inferensial

1. **Hukum Peluang Gugur:** Peluang unit untuk terpilih tidak diketahui (P tidak terukur).
2. **Dilarang Melakukan Generalisasi:** Hasil riset **TIDAK BISA** digeneralisasikan ke populasi luas secara matematis.
3. **Margin of Error (MoE) dan p-value Tidak Berlaku:** Tidak ada dasar matematis untuk klaim inferensial formal.

Metode Non-Prob 1: Convenience Sampling (Akses Seadanya)

Konsep Dasar Convenience Sampling

Memilih anggota sampel semata-mata karena mereka **paling mudah diakses**, paling dekat secara fisik, atau paling cepat dijangkau oleh peneliti.

Sering juga disebut **Accidental Sampling** atau **Opportunity Sampling**.

Contoh Riil & Jebakan Fatal Bias

• **Contoh Mahasiswa:** Membagikan tautan kuesioner skripsi hanya ke grup WhatsApp teman seangkatan atau mencegat mahasiswa yang lewat di depan kantin.
• **Contoh Startup:** Menanyai rekan satu ruangan kantor untuk menilai kenyamanan desain landing page baru.
• **Jebakan Fatal:** Sampel sangat homogen dan tidak representatif. Mengambil keputusan strategis dari data ini ibarat menyetir mobil dengan kaca depan tertutup kabut!

Metode Non-Prob 2: Purposive / Judgmental Sampling

Konsep Berdasarkan Pertimbangan Kriteria Khusus

Peneliti memilih sampel secara **sengaja** berdasarkan keahlian, penilaian profesional (**expert judgment**), atau kriteria ketat yang telah ditentukan sebelumnya.

Fokus utamanya bukan pada kuantitas acak, melainkan pada **kedalaman informasi** dari informan kunci (**key informants**).

Contoh Penerapan Nyata

- **Riset Cybersecurity:** Mewawancarai mendalam 12 Chief Information Security Officer (CISO) dari bank multinasional mengenai strategi mitigasi **zero-day exploit**.
- **Audit Sistem Informasi:** Memeriksa secara spesifik 30 transaksi dengan nilai di atas Rp 500 juta yang dieksekusi di luar jam kerja normal.
- **Nilai Manfaat:** Memberikan wawasan teknis yang sangat tajam dan tidak mungkin diperoleh dari orang awam secara acak.

Metode Non-Prob 3: Snowball Sampling (Bola Salju)

Konsep Rantai Rujukan (Chain Referral)

Pengambilan sampel dimulai dari sejumlah kecil informan awal (**seed**). Setelah diwawancarai, informan ini diminta mereferensikan kenalan mereka yang memenuhi kriteria riset.

Proses rujukan berlanjut dari generasi ke generasi sehingga jejaring sampel membesar seperti bola salju yang menggelinding.

Kapan Snowball Menjadi Satu-Satunya Pilihan?

- **Populasi Tersembunyi & Sensitif (**Hard-to-reach**):** Tidak ada daftar nama atau dokumen publik yang mencatat anggota populasi tersebut.
- **Contoh Kasus:**
- Meneliti penderita penyakit langka (**rare genetic disorder**) yang saling terhubung di grup dukungan privat.
- Wawancara komunitas peretas etis (**bug bounty hunters**) yang mengutamakan anonimitas dan kepercayaan personal.

Ilustrasi Visual: Snowball Sampling (Rantai Rujukan)



Keunggulan: Menjangkau populasi langka / tersembunyi (ahli keamanan siber, penderita penyakit langka, komunitas tertutup).

Kelemahan Kritis: Rentan **Homophily Bias** (responden hanya mereferensikan teman yang sepemikiran / satu lingkaran).

Metode Non-Prob 4: Quota Sampling (Sampling Kuota)

Konsep Dasar Quota Sampling

Peneliti menetapkan target kuota responden berdasarkan sub-kelompok demografi (misal: 50 pengguna Android dan 50 pengguna iOS).

Pengumpulan data berhenti segera setelah kuota masing-masing kelompok terpenuhi.

Perbedaan Kritis: Quota vs Stratified Sampling!

- **Stratified Sampling (Probabilitas):** Kelompok dibagi rapi, lalu anggota tiap kelompok **diambil secara acak (RNG)** dari **sampling frame** resmi.
- **Quota Sampling (Non-Probabilitas):** Kelompok dibagi rapi, namun pengisian kuota dilakukan **seadanya / siapa cepat dia dapat** tanpa pengacakan formal.
- **Konteks Pemakaian:** Sangat lazim dipakai oleh agensi riset pasar komersial untuk mengejar tenggat waktu cepat dan biaya terjangkau.

Metode Non-Prob 5: Voluntary Response Sampling

Konsep Responden Sukarela (Self-Selection)

Sampel terbentuk karena responden **memilih dirinya sendiri secara sukarela** untuk mengisi survei, bukan karena dipilih secara aktif oleh peneliti.

Contoh Nyata & Bahaya Polarisasi Ekstrem

- **Contoh Populer:** Polling di Twitter/X atau Instagram Story, kolom ulasan bintang 1 s.d. 5 di Google Play Store atau Google Maps.
- **Jebakan Polarisasi Ekstrem:**
  - Orang yang sukarela meluangkan waktu umumnya hanya mereka yang memiliki **opini atau emosi yang sangat ekstrem** (sangat membenci atau sangat memuja).
  - Mayoritas pengguna moderat yang "puas wajar" tidak akan pernah repot-repot mengisi ulasan.
- **Akibat:** Data ulasan online hampir selalu membentuk kurva terdistorsi tajam di ujung ekstrim!

Matriks Panduan: Memilih Metode Sampling yang Tepat

Dimensi Evaluasi	Probability Sampling	Non-Probability Sampling
Sampling Frame	Wajib ada daftar lengkap populasi	Tidak memerlukan kerangka populasi
Peluang Terpilih	Diketahui pasti ($P > 0$) & terukur	Tidak diketahui, subjektif peneliti
Generalisasi Populasi	**Sah secara matematis (Inferensial)**	**Hanya berlaku untuk sampel itu saja**
Margin of Error	Dapat dihitung persis ($\pm$ MoE)	Mustahil dihitung secara valid
Biaya & Waktu	Tinggi, butuh protokol ketat	Cepat, hemat, dan fleksibel
Contoh Kasus Ideal	Quick Count Pemilu, BPS, Uji Klinis	Uji Usability UI/UX, Riset Ahli, Pilot Study

Studi Kasus Blok 2 (Kasus 1): Audit Fraud Transaksi E-Commerce

SKENARIO PENGAWASAN RISIKO FINANSIAL

Platform e-commerce memproses 500.000 transaksi per hari. Tim **Fraud Detection** ingin menarik sampel audit harian sebesar $n = 2.500$ transaksi untuk diinvestigasi mendalam oleh analis. Data transaksi terbagi atas:

- Segmen Mikro (< Rp 50.000): 70% transaksi (350.000 transaksi).
- Segmen Retail (Rp 50.000 s.d. Rp 2.000.000): 28% transaksi (140.000 transaksi).
- Segmen Korporat (> Rp 50.000.000): 2% transaksi (10.000 transaksi, risiko kerugian terbesar).

Pertanyaan Diskusi

1. Jika menggunakan **Simple Random Sampling** (SRS), apa risiko fatal bagi pengawasan transaksi korporat?
2. Metode sampling apa yang wajib diterapkan dan bagaimana strategi alokasi sampelnya?

Pembahasan Kasus 1: Mengapa Wajib Stratified Oversampling?

1. Kelemahan Fatal Simple Random Sampling (SRS)

- Dengan SRS, transaksi korporat (proporsi 2%) hanya akan memperoleh sampel sekitar $2\% \times 2.500 = 50$ transaksi per hari.
- Jumlah $n = 50$ terlampau kecil untuk mendeteksi variasi modus pembobolan fraud yang kompleks, padahal kerugian finansial per insiden di segmen ini bernilai miliaran rupiah!

2. Solusi: Stratified Sampling Disproporsional (Oversampling)

- Bentuk 3 strata berdasarkan nilai nominal transaksi.
- Lakukan **oversampling** pada stratum korporat: alokasikan misalnya 500 transaksi korporat (5% dari total korporat), 1.000 transaksi retail, dan 1.000 transaksi mikro.
- **Manfaat Rekayasa:** Seluruh segmen risiko tinggi terlindungi optimal tanpa mengabaikan transaksi lainnya.

Studi Kasus Blok 2 (Kasus 2): Inspeksi Wilayah vs Komunitas Tertutup

DUA KEBUTUHAN RISET DENGAN TANTANGAN BERBEDA

Kasus A (Audit Server Nasional): Badan Siber Pemerintah perlu mengaudit keamanan fisik server di 1.200 kantor cabang yang tersebar di 300 kota di Indonesia dengan anggaran dan waktu perjalanan auditor yang sangat terbatas.

Kasus B (Riset Sindikat Kejahatan Siber): Peneliti akademis ingin meneliti modus operandi jaringan pembuat **ransomware** dan broker data bocor di forum **dark web** di mana identitas mereka dirahasiakan dan tidak memiliki direktori publik.

Pertanyaan Diskusi

Tentukan metode penarikan sampel yang paling tepat untuk masing-masing kasus (Kasus A dan Kasus B) beserta alasan metodologisnya!

Pembahasan Kasus 2: Cluster / Multistage vs Snowball Sampling

Kasus A: Cluster Sampling atau Multistage Sampling

- **Alasan:** Jika memakai SRS, auditor harus terbang ke ratusan kota berbeda hanya untuk memeriksa 1–2 server (biaya logistik membengkak).
- **Solusi:** Kelompokkan berdasarkan kota (klaster). Undi acak misalnya 20 kota, lalu audit seluruh cabang di 20 kota terpilih. Menghemat 80% biaya dan waktu logistik!

Kasus B: Snowball Sampling (Non-Probability)

- **Alasan:** Populasi target adalah **hidden population** yang menolak kontak publik dan tidak memiliki **sampling frame** resmi.
- **Solusi:** Peneliti membangun kepercayaan dengan 1 informan kunci (**seed**), lalu meminta rekomendasi akses (**chain referral**) ke narasumber terpercaya berikutnya secara berantai.

Blok 3 — Distribusi Sampling & Standard Error

Memahami perilaku statistik rata-rata sampel ketika pengambilan diulang

Eksperimen Mental: Memahami Distribusi Sampling

Misalkan kita punya populasi 100.000 mahasiswa dengan IPK rata-rata $\mu = 3{,}20$:

Simulasi Penarikan Berulang 10.000 Kali

1. Ambil sampel acak $n = 50$, hitung rata-ratanya: dapat $\overline{x}_1 = 3{,}24$.
2. Ambil sampel acak ke-2 ($n = 50$): dapat $\overline{x}_2 = 3{,}18$.
3. Ambil sampel acak ke-3 ($n = 50$): dapat $\overline{x}_3 = 3{,}21$.
4. Ulangi proses ini sebanyak 10.000 kali hingga kita memiliki 10.000 angka rata-rata!

Distribusi Sampling dari $\overline{X}$ adalah distribusi probabilitas dari kumpulan 10.000 angka rata-rata ($\overline{x}$) tersebut!

Sifat 1: Rata-Rata Sampel Bersifat Tak Bias (Unbiased)

Teorema Penaksir Tak Bias

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = \mu$$

Rata-rata dari seluruh kemungkinan nilai rata-rata sampel **selalu sama persis dengan rata-rata populasi asalnya**.

Meskipun satu sampel tertentu bisa meleset di atas atau di bawah, secara jangka panjang nilai rata-rata sampel tidak memiliki bias sistematis ke arah mana pun.

Sifat 2: Kesalahan Standar (Standard Error / SE)

Rumus Kesalahan Standar Rata-Rata

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Simpangan baku dari distribusi rata-rata sampel dinamakan **Standard Error** (SE).

Ukuran ini mengukur seberapa jauh rata-rata sampel ($\overline{x}$) diperkirakan meleset dari rata-rata populasi sebenarnya (μ) akibat faktor kebetulan penarikan sampel.

Mengapa Nilai Rata-Rata Jauh Lebih Ramping?

Perhatikan keberadaan faktor pembagi $\sqrt{n}$ pada penyebut formula:

• Pada data individu ($n = 1$), sangat mungkin kita menjumpai nilai ekstrem langka.

• Namun pada kelompok $n = 100$ orang acak, peluang bahwa **seluruh 100 orang tersebut ekstrem bersamaan** adalah mendekati nol.

• Nilai tinggi dan nilai rendah **saling menetralkan** satu sama lain saat dirata-ratakan!

Penyusutan Galat (SE)

Sampel (n)	**Galat Baku (SE)**
$n = 1$	σ (Data Asli)
$n = 4$	$\frac{\sigma}{2} = 0{,}50\,\sigma$
$n = 16$	$\frac{\sigma}{4} = 0{,}25\,\sigma$
$n = 100$	$\frac{\sigma}{10} = 0{,}10\,\sigma$

Faktor Koreksi Populasi Terhingga (FPC)

Kapan Rumus Standar Harus Dikoreksi?

Jika sampel diambil dari **populasi terbatas tanpa pengembalian**, dan ukuran sampel mencakup porsi besar populasi:

$$\frac{n}{N} > 0{,}05 \quad \text{(Fraksi Sampling lebih dari 5\%)}$$

Rumus FPC & Makna Fisis

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Makna Fisis: Semakin banyak porsi populasi yang kita amati, semakin sedikit ketidakpastian yang tersisa. Jika $n = N$ (sensus lengkap), nilai $\sqrt{\frac{N-N}{N-1}} = 0$, sehingga SE $= 0$.

Studi Kasus Blok 3 (Kasus 1): Pengetatan SLA Latensi API

SKENARIO SITE RELIABILITY ENGINEERING (SRE)

Seorang SRE mengukur waktu respons (latensi) gateway API sistem perbankan. Variabilitas waktu respons diketahui memiliki simpangan baku populasi $\sigma = 60$ ms.

Pada tahap uji awal, tim mengambil sampel acak $n_1 = 25$ request, yang menghasilkan galat baku SE_1. Untuk memenuhi standar SLA baru dari regulator, manajemen menuntut agar nilai galat baku dipotong menjadi **seperempatnya** ($\frac{1}{4}$) dari galat semula.

Pertanyaan Diskusi

1. Berapakah nilai **Standard Error** (SE_1) pada uji awal?
2. Berapa ukuran sampel baru (n_2) yang wajib diambil agar $\mathrm{SE}_2 = \frac{1}{4}\,\mathrm{SE}_1$?
3. Jelaskan hukum kelipatan kuadrat (**Square Root Law**) yang mendasari hasil ini!

Pembahasan Kasus 1: Aturan Kuadrat Pengurangan Galat Baku

1. Perhitungan SE Sampel Awal ($n_1 = 25$)

$$SE_1 = \frac{\sigma}{\sqrt{n_1}} = \frac{60}{\sqrt{25}} = \frac{60}{5} = 12\,\text{ms}$$

2. Menentukan Ukuran Sampel Baru (n_2)

Target kita adalah $SE_2 = \frac{12}{4} = 3$ ms:

$$\frac{60}{\sqrt{n_2}} = 3 \rightarrow \sqrt{n_2} = \frac{60}{3} = 20 \rightarrow n_2 = 20^2 = 400\,\text{request}$$

Prinsip Hukum Kuadrat (Diminishing Returns): Untuk memotong galat menjadi $\frac{1}{k}$ kali lipat, sampel harus diperbesar sebesar k^2 kali lipat! Memotong SE menjadi $\frac{1}{4}$ butuh data $4^2 = 16$ kali lipat ($25 \times 16 = 400$).

Studi Kasus Blok 3 (Kasus 2): Audit Repositori Modul Terbatas

SKENARIO SOFTWARE QUALITY ASSURANCE (QA)

Sebuah tim QA melakukan audit keamanan terhadap repositori perangkat lunak **on-premise** yang memiliki total $N = 300$ modul microservice. Simpangan baku kompleksitas siklomatis modul adalah $\sigma = 10$ poin.

Tim mengambil sampel acak tanpa pengembalian sebanyak $n = 60$ modul untuk diperiksa secara mendalam.

Pertanyaan Diskusi

1. Hitung fraksi sampling $\frac{n}{N}$. Apakah Faktor Koreksi Populasi Terhingga (**FPC**) wajib digunakan? Jelaskan alasannya!
2. Hitung nilai **Standard Error** (SE) dengan menyertakan faktor FPC!

Pembahasan Kasus 2: Pengaruh Faktor FPC pada Populasi Terbatas

1. Uji Fraksi Sampling

$$\frac{n}{N} = \frac{60}{300} = 0{,}20\,(20\%) > 0{,}05\,(5\%)$$

Karena sampel mencakup lebih dari 5% populasi, rumus galat standar wajib dikalikan FPC.

2. Perhitungan SE dengan FPC

$$\text{FPC} = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{300-60}{300-1}} = \sqrt{\frac{240}{299}} = \sqrt{0{,}8027} \approx 0{,}8959$$

$$\sigma_{\overline{X}} = \frac{10}{\sqrt{60}} \times 0{,}8959 = 1{,}291 \times 0{,}8959 \approx 1{,}157 \,\text{poin}$$

Catatan: Tanpa FPC, galat baku bernilai 1,291. Berkat FPC, ketidakpastian menyusut sekitar 10,4% karena seperlima populasi sudah diperiksa secara nyata.

Blok 4 — Teorema Limit Pusat (CLT)

Hukum fundamental inferensi: Mengapa dunia konvergen menuju kurva lonceng normal

Bunyi Teorema Limit Pusat (Central Limit Theorem)

TEOREMA PALING AGUNG DALAM ILMU STATISTIKA

Bila sampel-sampel acak berukuran n diambil dari populasi berdistribusi apa pun dengan rata-rata μ dan varians σ^2 terhingga:

Maka ketika ukuran sampel cukup besar ($n \geq 30$), distribusi sampling dari rata-rata sampel $\overline{X}$ akan mendekati Distribusi Normal:

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Keajaiban Teorema Limit Pusat

Mengapa CLT Merupakan Terobosan Monumental?

Di dunia nyata, kita **hampir tidak pernah tahu bentuk distribusi asli populasi**. Populasi bisa jadi sangat miring (gaji karyawan), seragam (waktu tunggu bus), atau bimodal.

CLT membebaskan kita dari keharusan menebak distribusi populasi!

Asalkan ukuran sampel kita cukup ($n \geq 30$), rata-rata sampel ($\overline{X}$) dipastikan patuh pada hukum kurva lonceng simetris Normal Gauss.

Standarisasi Z untuk Rata-Rata Sampel ($\overline{X}$)

Rumus Transformasi Skor Z Distribusi Sampling

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Perhatikan Penyebutnya:

Penyebut rumus bukan lagi simpangan baku data asli σ, melainkan **Standard Error** ($\frac{\sigma}{\sqrt{n}}$).

Perbandingan Z Individu vs Z Rata-Rata Kelompok

Z untuk Nilai Individu (Pertemuan 6)

$$Z = \frac{X - \mu}{\sigma}$$

- Digunakan saat kita mengambil **1 orang / 1 item saja**.
- Mengukur deviasi nilai tunggal X terhadap populasi.
- Variabilitas besar (kurva lebar).

Z untuk Rata-Rata Sampel (Pertemuan 7)

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Digunakan saat kita mengambil **kelompok berukuran n**.
- Mengukur deviasi rata-rata $\overline{X}$ terhadap populasi.
- Variabilitas kecil (kurva ramping).

Distribusi Sampling Proporsi ($\hat{P}$)

Karakteristik Data Kategorik Biner ($\hat{p} = \frac{X}{n}$)

Jika data berbentuk proporsi (misal: tingkat klik iklan, proporsi pemilih):

1. **Pusat Rata-Rata:** $\mu_{\hat{p}} = E(\hat{p}) = p$ (Penaksir Tak Bias)
2. **Standard Error Proporsi:**

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Syarat Normalitas CLT Proporsi

Distribusi $\hat{p}$ dapat dianggap normal jika jumlah kejadian sukses dan gagal keduanya cukup besar:

$$np \geq 10 \quad \text{dan} \quad n(1-p) \geq 10$$

Standarisasi Z untuk Proporsi ($\hat{p}$)

Rumus Transformasi Skor Z Proporsi

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Formula ini adalah pilar utama yang dipakai oleh lembaga survei (**Quick Count** pemilu), praktisi **Digital Marketing** (A/B testing CTR), dan industri manufaktur (pengendalian cacat mutu).

Studi Kasus Blok 4 (Kasus 1): Mitigasi Lonjakan Bandwidth Kantor

SKENARIO INFRASTRUKTUR JARINGAN KOMPUTER

Konsumsi bandwidth karyawan kantor berdistribusi **sangat miring ke kanan (right-skewed)** dengan rata-rata $\mu = 15$ Mbps dan simpangan baku $\sigma = 12$ Mbps (mayoritas memakai bandwidth kecil, namun ada segelintir **power users** yang mengunduh file video/dataset raksasa).

Router utama gedung melayani secara simultan $n = 36$ karyawan independen. Kapasitas maksimum router dirancang mampu menahan beban rata-rata hingga 19 Mbps per karyawan.

Pertanyaan Diskusi

1. Apakah sah menggunakan kurva Normal untuk menghitung peluang ini meski distribusi konsumsi bandwidth asli sangat miring? Jelaskan alasannya!
2. Berapakah peluang router mengalami **overload** (rata-rata pemakaian 36 orang melampaui 19 Mbps)? ($\Phi(2{,}00) = 0{,}9772$)

Pembahasan Kasus 1: Menghitung Risiko Overload dengan CLT

1. Validitas Teorema Limit Pusat (CLT)

• **Sah & Valid:** Berdasarkan CLT, karena ukuran sampel $n = 36 \geq 30$, distribusi dari rata-rata pemakaian ($\overline{X}$) dipastikan berdistribusi **Normal simetris**, tidak peduli seberapa miring distribusi populasi individualnya!

2. Perhitungan Skor Z dan Peluang Overload

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2\,\text{Mbps}$$

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{19 - 15}{2} = +2 \rightarrow Z = 2{,}00$$

$$P\left(\overline{X} > 19\right) = 1 - \Phi(2{,}00) = 1 - 0{,}9772 = 0{,}0228\,(2{,}28\%)$$

Wawasan Rekayasa: Peluang router jebol akibat pemakaian serempak rata-rata melampaui 19 Mbps sangat kecil, hanya "2,28"%. Kapasitas router dinyatakan aman!

Studi Kasus Blok 4 (Kasus 2): Uji Signifikansi A/B Testing CTR

SKENARIO PRODUCT GROWTH & REKAYASA WEB

Tim produk e-commerce menguji desain tombol **checkout** baru (Varian B). Rasio konversi historis adalah $p = 0{,}10$ (10%). Desain baru diuji coba pada $n = 400$ pengunjung acak dan menghasilkan $X = 52$ transaksi sukses ($\hat{p} = \frac{52}{400} = 0{,}13$ atau 13%).

Pertanyaan Diskusi

1. Periksa apakah syarat normalitas CLT proporsi ($np \geq 10$ dan $n(1-p) \geq 10$) terpenuhi!
2. Jika tombol baru sebenarnya tidak memberi efek apapun (p asli tetap "0,10"), berapa peluang munculnya sampel dengan konversi $\hat{p} \geq 13\%$ murni akibat faktor kebetulan acak? ($\Phi(2{,}00) = 0{,}9772$)

Pembahasan Kasus 2: Uji Validitas Normalitas & Z Proporsi

1. Uji Syarat Kelayakan CLT Proporsi

- $np = 400 \times 0{,}10 = 40 \geq 10$ (Terpenuhi)
- $n(1-p) = 400 \times 0{,}90 = 360 \geq 10$ (Terpenuhi)

Kedua syarat terpenuhi, sehingga distribusi sampling $\hat{P}$ valid didekati dengan Normal Gauss.

2. Perhitungan Standard Error dan Skor Z

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}10 \times 0{,}90}{400}} = \frac{0{,}30}{20} = 0{,}015\,(1{,}5\%)$$

$$Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}} = \frac{0{,}13 - 0{,}10}{0{,}015} = \frac{0{,}03}{0{,}015} = +2{,}00$$

$$P(\hat{p} \geq 0{,}13) = 1 - \Phi(2{,}00) = 1 - 0{,}9772 = 0{,}0228\,(2{,}28\%)$$

Pelajaran Penting: Karena peluang kebetulan acak hanya "2,28"% ($< 5\%$), tim memiliki bukti empiris kuat bahwa tombol baru benar-benar meningkatkan konversi (fondasi Pengujian Hipotesis P10!).

Penutup & Jembatan ke Estimasi Parameter

Rekapitulasi paruh pertama semester & persiapan Ujian Tengah Semester (UTS)

Rekapitulasi Teori Sampling & Jembatan Paruh Kedua

Fondasi Paruh Pertama (P01–P07)

- P01: Pengantar data & skala pengukuran
- P02–P03: Pemusatan & penyebaran
- P04: Teorema Bayes & probabilitas
- P05: Peubah diskret (Binomial, Poisson)
- P06: Peubah kontinu (Normal, Z-Score)
- P07: Teorema Limit Pusat & Sampling

Aplikasi Paruh Kedua (P08–P16)

- **P08: Ujian Tengah Semester (UTS)**
- P09: Estimasi Parameter & Confidence Interval
- P10–P11: Uji Hipotesis (Z, t-test, A/B Testing)
- P12: Uji Chi-Kuadrat & ANOVA
- P13–P15: Regresi & Machine Learning

Selamat mempersiapkan diri menghadapi Evaluasi UTS Minggu Depan!