Pengantar Statistika dan Jenis Data

Problem-based: dari kasus nyata ke konsep — apa yang diukur, siapa yang diwakili, dari mana sumber datanya

Statistika dan Probabilitas · Pertemuan 01

Peta pertemuan — tiap topik dimulai dari masalah

1	Mengapa angka saja tidak cukup untuk memutuskan	Kasus berita survei
2	Merangkum data vs menggeneralisasi ke populasi	Kasus nilai UTS
3	Populasi, sampel, unit, dan teknik sampling	Kasus klaim “nasional”
4	Skala pengukuran menentukan operasi yang sah	Kasus ranking salah
5	Sumber data primer vs sekunder	Kasus dataset siap pakai

> Alur tiap blok: **kasus → pertanyaan → konsep**.

Blok 1 — Mengapa belajar statistika

Kasus: klaim persentase di berita

news.uhb.ac.id / survei-kuliah-daring

Senin, 7 Sep 2026

KAMPUS NEWS

Beranda · Akademik · Survei · Opini

AKADEMIK

Survei: 78% mahasiswa setuju kuliah daring penuh

Oleh Redaksi Kampus News · 2 jam lalu · 1.2k pembaca

Berdasarkan survei daring awal pekan ini, mayoritas mahasiswa menyatakan setuju jika perkuliahan digelar daring penuh semester depan.

Rektorat diminta mempertimbangkan hasil ini sebagai bahan kebijakan.

*Sumber grafis: olahan redaksi dari hasil survei internal yang diisi **43 mahasiswa dari satu kelas**.*

INFOGRAFIS SURVEI

78%

setuju kuliah daring penuh

Komposisi jawaban (n = 43)



Setuju 34 Tidak 9

Pemantik diskusi

PERTANYAAN PEMICU

- Apakah angka 78% tersebut mewakili opini **seluruh** mahasiswa?
- Informasi krusial apa yang tidak disampaikan oleh **headline** sehingga bisa memicu kesimpulan yang salah?
- Sebagai pimpinan kampus, hal apa yang perlu Anda pastikan terlebih dahulu sebelum mengambil kebijakan dari berita tersebut?

Konsep: Statistika sebagai alat validasi klaim data

Akar masalah pada kasus berita tadi adalah **hilangnya konteks pengambilan sampel**. Dalam statistika, pengambilan data dapat dilakukan terhadap sebagian populasi saja (sampel). Namun apakah responden (43 mahasiswa dari satu kelas) cukup mewakili **seluruh** mahasiswa?

Judul berita menyembunyikan fakta ketidakpastian, sehingga pembaca berasumsi itu adalah kebenaran mutlak.

Belajar statistika akan membentuk tiga pola pikir utama:

1

Kritis menanyakan **siapa** subjek pengukurannya dan **berapa jumlah** pastinya.

2

Mampu memisahkan antara fakta dari kelompok sampel dengan kesimpulan untuk seluruh populasi.

3

Transparan dalam menyatakan sejauh mana batas berlakunya suatu kesimpulan.

Contoh 1 — Kasus

PRAKIRAAN CUACA

Probabilitas hujan 70% hari ini

Bagaimana Anda membaca klaim ini?

Contoh 1 — Anggapan yang umum (tapi keliru)

Anggapan A

“70% berarti **pasti** hujan — jadi saya wajib bawa payung / batalkan rencana.”

Anggapan B

“Kemarin juga 70% tapi cerah. Berarti angka ini tidak bisa dipercaya.”

Anggapan C

“Kalau tidak hujan hari ini, prediksi 70% itu salah.”

Contoh 1 — Makna sesungguhnya

> Ini pernyataan **peluang**, bukan ramalan mutlak untuk satu hari tunggal.

Dengan kondisi atmosfer yang **serupa** seperti hari ini, hujan umumnya terjadi pada sekitar **7 dari 10** peristiwa semacam itu.

Hari ini bisa hujan atau tidak — keduanya tetap konsisten dengan peluang 70%. Satu hasil tunggal tidak "membuktikan" atau "membantah" persentase tersebut.

Yang Anda putuskan (bawa payung atau tidak) adalah mitigasi risiko di bawah ketidakpastian — bukan reaksi terhadap kepastian palsu.

Contoh 1 — Pola pikir yang dibentuk

1. Persentase (%) adalah estimasi peluang, bukan jaminan.
2. Satu kali prediksi meleset tidak mematahkan konsep peluang.
3. Aksi (membawa payung) = langkah mitigasi risiko, bukan bukti kepastian.

> Alur: data prediksi ⟶ tindakan ⟶ fatal jika persentase diyakini sebagai kepastian.

Contoh 2 — Kasus

DISKON TOKO

Potongan 50% dari harga normal Rp400.000

Bagaimana Anda membaca klaim hemat ini?

Contoh 2 — Anggapan yang umum (tapi keliru)

Anggapan A

“Harga normal Rp400.000 pasti harga pasar — jadi saya benar-benar hemat 50%.”

Anggapan B

“Angka diskon di etalase sudah cukup; tidak perlu bandingkan ke toko lain.”

Anggapan C

“Semakin besar persen potongan, semakin besar keuntungan saya.”

Contoh 2 — Makna sesungguhnya

> “Harga normal” adalah **baseline** yang bisa dipilih penjual — bukan otomatis harga pasar.

Tanpa kejelasan, patokan “normal” bisa berupa harga semu yang jarang dipakai (**mark-up**), atau harga eksklusif satu toko.

Klaim hemat baru bermakna jika ada pembanding riil: harga kompetitor, riwayat harga, atau rata-rata pasar, bukan hanya selisih dari angka di label.

Persen besar dengan baseline palsu bisa lebih mahal daripada harga kompetitor tanpa “diskon”.

Contoh 2 — Pola pikir yang dibentuk

1. Selalu kritis: “normal” ini standar **siapa** dan **kapan**?
2. Angka tunggal tanpa data pembanding sangat rawan manipulasi.
3. Klaim “lebih murah” belum tentu mencerminkan harga asli pasar.

> Alur: janji promosi ⟶ dorongan membeli ⟶ bahaya keliru meyakini harga dasar (baseline).

Contoh 3 — Studi Kasus

LAPORAN MEDIS & ANGKA KESEHATAN

“Hasil tes melampaui batas normal” atau “tingkat efektivitas vaksin 95%”

Pertanyaan kritis apa yang wajib kita ajukan sebelum menyimpulkan angka tersebut?

Contoh 3 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Lewat dari angka normal berarti pasti sakit parah — langsung panik tanpa mengecek konteks acuan medisnya.”

Miskonsepsi B

“Efektivitas 95% dianggap jaminan kekebalan total bagi semua orang terhadap segala macam risiko penularan.”

Miskonsepsi C

“Laporan efek samping yang sedikit dianggap 100% aman; atau sebaliknya, muncul satu kasus langsung divonis berbahaya bagi publik.”

Contoh 3 — Fakta dan interpretasi yang tepat

> Data klinis baru bisa dipahami jika kita mengetahui persis apa variabelnya, populasi rujukannya, dan besaran sampelnya.

Kata “normal” hanyalah batas referensi dari sampel acuan, bukan vonis diagnosis mutlak. Begitu pula klaim “efektivitas 95%” harus jelas: **apa** yang berhasil dicegah (gejala ringan atau rawat inap RS?) dan **siapa** kelompok yang diuji.

Kejadian efek samping umumnya berskala sangat kecil dalam uji klinis. Jika kita menelan angka persentase tanpa tahu jumlah sampel (**n**), standar kriteria, dan pembaginya, kita bisa terjebak pada kepanikan buta atau rasa aman yang semu.

Prinsip dasar: periksa apa metriknya, siapa targetnya, seberapa besar datanya, dan apa kriteria pengujiannya.

Contoh 3 — Membangun kebiasaan berpikir kritis

1. Perjelas dulu **definisi operasional variabel** sebelum mengambil kesimpulan medis.
2. Angka proporsi atau persentase wajib memiliki populasi acuan dan penyebut (**denominator**) yang jelas.
3. Label “normal” murni merupakan rentang batas dari sampel referensi, bukan patokan kondisi ideal mutlak.

> Alur observasi: data klinis → pengambilan tindakan → kewaspadaan terhadap bahaya generalisasi.

Contoh 4 — Studi Kasus

EKSPERIMEN KONTEN DIGITAL (A/B TESTING)

Baru diuji 20 tayangan, thumbnail A unggul 3 klik — haruskah langsung dipakai?

Apakah data minim ini valid untuk mengganti thumbnail video YouTube?

Contoh 4 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Thumbnail A sudah menang di data awal — langsung pasang di semua video.”

Miskonsepsi B

“Keunggulan 3 klik membuktikan thumbnail A lebih disukai seluruh penonton channel.”

Miskonsepsi C

“Kalau CTR naik sedikit saja, berarti eksperimen thumbnail itu sukses besar.”

Contoh 4 — Fakta dan interpretasi yang tepat

> Tren awal pada ukuran sampel (**n**) yang sangat kecil sering kali murni faktor kebetulan (**sampling error**) — bukan sinyal asli dari desain thumbnail.

Memutuskan thumbnail default menuntut jumlah tayangan/impresi yang memadai, metrik yang spesifik (klik? CTR? watch time?), dan kesadaran bahwa penonton uji awal belum tentu mewakili seluruh audiens channel.

“Menang di uji terbatas” tidak menjamin akan “menang di seluruh upload.” Di materi selanjutnya, kita akan menggunakan uji hipotesis untuk menyaring sinyal nyata dari kebisingan data (**noise**).

Jebakan **false positive**: terlalu cepat merayakan kemenangan dari data yang belum stabil.

Contoh 4 — Membangun kebiasaan berpikir kritis

1. Hindari mengambil kesimpulan akhir dari jumlah sampel (**n**) yang terlampau kecil.
2. Bedakan menang di uji thumbnail terbatas vs menang di seluruh populasi penonton.
3. Ke depan: terapkan uji hipotesis dan ukur **effect size** untuk memvalidasi keputusan.

> Alur observasi: uji thumbnail skala kecil ⟶ ganti desain default ⟶ ancaman **false positive**.

Contoh 5 — Studi Kasus

OPERASIONAL & TELEMETRI SISTEM

Rata-rata latency terpantau 180 ms, tingkat error 5xx = 0,4%

Apakah angka di layar pantau ini cukup valid untuk menyatakan sistem “aman” atau justru “kritis”?

Contoh 5 — Miskonsepsi yang sering terjadi

Miskonsepsi A

"Rata-rata 180 ms itu normal — ada **spike** 2 detik sesekali tak perlu dipusingkan."

Miskonsepsi B

"Terdapat satu lonjakan di malam hari ⟶ pasti seluruh layanan sedang **down**."

Miskonsepsi C

"Persentase **error** 0,4% itu sangat kecil, jadi volume **request**-nya tidak perlu dicek lagi."

Contoh 5 — Fakta dan interpretasi yang tepat

> Menyimpulkan performa hanya dari nilai tengah (**mean**) bisa menyesatkan — nilai ekstrem (ekor distribusi) dan jumlah pembaginya sama pentingnya.

Angka rata-rata sering menutupi **request** yang sesungguhnya berjalan sangat lambat. Begitu pula proporsi **error** tanpa total **request**; 0,4% dari 100 **request** memiliki tingkat urgensi yang jauh berbeda dengan 0,4% dari sejuta **request**.

Sebelum bertindak, tentukan parameter sebaran (p95/p99), rentang waktu pengamatan, serta batas toleransi yang disepakati. Ini penting agar fluktuasi yang wajar tidak dianggap sebagai insiden kritis.

Tim operasional yang tanggap tahu cara memfilter dinamika harian (**noise**) dari gangguan yang riil.

Contoh 5 — Membangun kebiasaan berpikir kritis

1. Satu metrik ringkasan saja tidak memadai (pusat data tidak bisa mewakili nilai ekstrem).
2. Pahami cara memisahkan fluktuasi sistem yang wajar dari sinyal anomali.
3. Menilai tingkat kesalahan (**error rate**) selalu membutuhkan penyebut (total beban **request**).

> Alur observasi: pantauan metrik ⟶ respons teknis ⟶ bahaya alarm palsu.

Pola yang sama di kelima contoh



Statistika = bahasa untuk mengelola risiko itu — mulai dari kasus portal berita di awal blok ini.

Blok 2 — Deskriptif vs inferensial

Kasus: mean UTS satu kelas dipakai untuk kebijakan angkatan

Blok 2 — Studi Kasus

STATISTIKA DESKRIPTIF VS INFERENSIAL

Rata-rata UTS Kelas A = 72 (n = 40).
Muncul usulan: hapus program remedial untuk seluruh angkatan.

Padahal, Kelas B dan C (sekitar 180 mahasiswa) belum ikut ujian. Dari situasi ini, mana yang disebut pernyataan deskriptif dan mana yang inferensial?

Blok 2 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Nilai rata-ratanya sudah didapat, berarti kesimpulan untuk performa satu angkatan sudah jelas.”

Menyamakan rangkuman fakta (deskriptif) sebagai kesimpulan menyeluruh (inferensi).

Miskonsepsi B

“Statistika deskriptif dan inferensial tidak ada bedanya, toh keduanya sama-sama menghitung nilai rata-rata.”

Miskonsepsi C

“Selama angkanya tampak valid dan tinggi, kita tidak perlu repot memisahkan antara fakta sampel dan populasi.”

Blok 2 — Fakta: dua jenis klaim di kasus yang sama

Klaim deskriptif — akurat

"Rata-rata nilai UTS mahasiswa di Kelas A (n = 40) adalah 72."

Pernyataan ini murni memotret data yang ada di tangan. Sama sekali tidak menyinggung kondisi Kelas B, C, atau seluruh angkatan.

Klaim inferensial — belum terbukti

"Performa angkatan ini sudah bagus, jadi kelas remedial ditiadakan."

Ini adalah lompatan generalisasi ke seluruh populasi (±180 orang). Langkah ini membutuhkan metode **sampling** dan analisis peluang yang belum dilakukan pada kasus ini.

> Akar masalahnya bukanlah pada nilai 72, melainkan pada **keputusan besar** yang secara prematur dipaksakan dari angka tersebut.

Blok 2 — Membangun kebiasaan berpikir kritis

1. Saat disodorkan sebuah angka, tanyakan selalu: apakah ini sekadar **deskriptif** (merangkum sampel), atau sedang mencoba **inferensial** (menebak populasi)?
2. Menyebutkan rata-rata satu kelas adalah fakta deskriptif, tetapi menjadikannya dasar kebijakan satu angkatan adalah tindakan inferensial.
3. Hindari mencampuradukkan kedua jenis klaim ini dalam satu kalimat laporan tanpa memberikan batasan yang jelas.

> Alur keliru: memotret data deskriptif ⟶ dipaksakan menjadi kesimpulan inferensial ⟶ menghasilkan kebijakan salah sasaran.

Konsep Dasar: Merangkum Fakta vs Menarik Kesimpulan

Penjelasan teoretis dari studi kasus di slide sebelumnya:

Statistika deskriptif

Fokus menyajikan dan meringkas data yang sudah terkumpul.

Sesuai kasus: rata-rata nilai 72 khusus untuk 40 mahasiswa Kelas A.

Klaim yang dibuat mutlak hanya berlaku untuk data yang sedang diolah.

Statistika inferensial

Berusaha menggeneralisasi sampel untuk memprediksi populasi yang lebih luas.

Sesuai kasus: simpulan prematur bahwa “seluruh angkatan aman dari remedial.”

Langkah ini wajib didukung teknik **sampling** dan analisis probabilitas.

Kesalahan usulan tadi adalah mencampuradukkan kedua ranah ini. Alur belajar kita semester ini: pastikan dulu **karakter data** yang diukur, masuk ke konsep peluang, barulah kita menyentuh inferensi formal.

Konsep: parameter (populasi) vs statistik (sampel)

Parameter — biasanya tidak diketahui

Sifat populasi yang ingin diketahui.

- μ mean populasi
- p proporsi populasi
- σ^2 varians populasi

Di kasus: μ = mean **angkatan**.

Statistik — dihitung dari data

Nilai dari sampel yang Anda pegang.

- $\overline{x}$ mean sampel
- $\hat{p}$ proporsi sampel
- s^2 varians sampel

Di kasus: $\overline{x} = 72$ kelas A.

Blok 3 — Memahami Populasi, Sampel, dan Unit

Studi Kasus: Bahaya klaim “pengguna se-Indonesia” dari sampel yang terlalu sempit

Blok 3 — Studi Kasus

MENEBAK POPULASI, SAMPEL, DAN UNIT

Klaim: “90% pengguna se-Indonesia puas!”

Padahal hanya dari 200 survei in-app di satu kota

Konteks tambahan: Survei ini muncul tepat setelah promo **cashback**. Unit observasi yang dihitung hanyalah akun yang mengisi survei hingga tuntas.

Blok 3 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Karena di judul tertulis ‘pengguna Indonesia’, berarti data tersebut otomatis mewakili sentimen seluruh warga se-nusantara.”

Miskonsepsi B

“Tidak perlu repot membedakan antara sampel dan populasi, yang penting angka 90%-nya terlihat meyakinkan untuk diiklankan.”

Miskonsepsi C

“Unit observasi itu pastilah manusia — jadi satu responden yang mengisi otomatis mewakili satu individu asli.”

Blok 3 — Fakta: tiga realitas berbeda di balik satu kasus

Populasi

Klaim target: Seluruh pengguna aplikasi di Indonesia.

Kenyataan di lapangan: Hanya mewakili pengguna aktif di satu kota yang kebetulan berpartisipasi dalam survei.

Sampel

Hanya 200 akun yang mengisi kuesioner sampai selesai.

Angka ini sama sekali bukan hasil pengambilan sampel acak yang mewakili demografi seluruh Indonesia.

Unit Observasi

Akun pengguna aplikasi (**user account**).

Ini tidak sama dengan total jumlah **download**, dan satu akun belum tentu mewakili satu orang yang berbeda.

> Kehadiran promo **cashback** menciptakan bias besar: pengguna yang baru saja diuntungkan tentu lebih cenderung memberi nilai tinggi, sehingga merusak objektivitas angka 90% tersebut.

Blok 3 — Membangun kebiasaan berpikir kritis

1. Biasakan untuk membedakan antara populasi **target** (yang diklaim secara bombastis) dengan populasi **sumber/frame** (dari mana data sungguh-sungguh diambil).
2. Berhentilah berlindung di balik **jargon** populasi besar; sebutkan profil dan batasan sampel Anda secara transparan.
3. Perjelas sejak awal apa unit observasinya: apakah yang dihitung itu orang perorangan, akun terdaftar, durasi interaksi, atau sekadar jumlah unduhan?

> Alur keliru: memakai slogan “se-Indonesia” ⟶ padahal jangkauan sempit dan hanya berbasis akun tertentu ⟶ memicu klaim berlebihan (**overclaim**) yang manipulatif.

Konsep Dasar: Tiga Istilah yang Pantang Tertukar

Mendudukkan kasus **fintech** tadi ke dalam bahasa formal statistika:

Populasi

Target luas yang ingin disimpulkan.

Ingat, target ideal sering berbeda dengan sumber data riil (**frame**). Keduanya wajib ditulis transparan.

Sampel

Sebagian kecil kelompok yang benar-benar diteliti.

Pada kasus tadi: hanya sebatas 200 responden dari dalam aplikasi.

Unit observasi

Satu entitas data tunggal yang diukur.

Bisa berupa orang, akun, atau transaksi. Wajib didefinisikan sejak awal.

Selanjutnya: mari kita bongkar **bagaimana** cara sampel tersebut dipilih (teknik **sampling**).

Teknik Sampling — Studi Kasus

MELANJUTKAN KASUS FINTECH — METODE PENARIKAN DATA

Terpilih 200 responden in-app di satu kota, tepat setelah bagi-bagi promo cashback.

Sebenarnya bagaimana cara sampel ini **dipilih**? Dan mengapa metodenya bisa meruntuhkan keabsahan klaim mereka?

Teknik Sampling — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Angka 200 responden itu sudah cukup besar, jadi wajar saja kalau dianggap sah mewakili suara seluruh Indonesia.”

Miskonsepsi B

“Asalkan orang mengisi survei tanpa paksaan, datanya sudah bisa disebut sampel acak (**random**).”

Miskonsepsi C

“Yang penting tabel datanya sudah terkumpul; urusan metode pencariannya tidak perlu dicantumkan di laporan.”

Teknik Sampling — Fakta dan interpretasi yang tepat

> Validitas generalisasi ditentukan oleh **bagaimana** cara data tersebut dipancing masuk, bukan sekadar seberapa banyak jumlah baris datanya.

Acak Sederhana (Simple Random)

Seluruh anggota populasi memiliki kesempatan yang persis sama untuk ditarik menjadi sampel (mirip sistem undian dari daftar lengkap).

Berstrata (Stratified)

Populasi dipecah dahulu berdasarkan kelompok (misal per kota/prodi), lalu sampel diambil dari tiap kelompok agar porsinya berimbang.

Pada kasus **fintech** tadi: surveinya bersifat sukarela, lokasinya mengelompok di satu kota, dan dilakukan pasca-promo. Ini namanya **convenience** atau **self-selection sampling**, sama sekali bukan sampel acak nasional! Tentu saja ada **bias seleksi**: orang yang baru saja diuntungkan dengan promosi akan lebih antusias mengisi dan memberi nilai bagus.

Pembahasan matematisnya (Teorema Limit Pusat & distribusi **sampling**) akan dibahas di Pertemuan 09.

Teknik Sampling — Membangun kebiasaan berpikir kritis

1. Selidiki dulu: dari kolam data mana (**frame**) sampel ini berasal, dan menggunakan metode apa data ini ditarik?
2. Sebesar apa pun jumlah sampel Anda (**n**), tidak akan bisa menambal cacat dari metode pengambilan data yang sudah bias sejak awal.
3. Jujurlah memaparkan teknik **sampling** di laporan Anda, dan turunkan ego saat membuat klaim generalisasi jika datanya memang terbatas.

> Alur analisis: periksa metode penarikan sampel → tentukan batasan siapa yang valid diwakili → hindari jebakan klaim bombastis (**overclaim**).

Blok 4 — Skala pengukuran

Kasus: ranking yang memakai rata-rata pada data yang salah jenis

Blok 4 — Studi Kasus

SKALA DATA MENENTUKAN CARA HITUNG YANG SAH

Bikin leaderboard “objektif” dari rata-rata kode jurusan dan skor Likert?

Cuplikan hasil survei (ilustrasi)

Resp.	Prodi	Kode	Kepuasan
R01	SI	1	4
R02	TI	2	5
R03	IF	3	3
R04	SI	1	4
R05	TI	2	2
R06	IF	3	5

Kode: SI=1, TI=2, IF=3 · Kepuasan: Likert 1 (sangat tidak puas) – 5 (sangat puas)

Hitungan asisten di rapat

Mean kode prodi

2,00

“Jadi prodi dominan ≈ TI?”

Mean kepuasan

3,83

Dipakai untuk ranking unit

Operasi mana yang melanggar aturan skala?

Blok 4 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Mentang-mentang wujudnya angka, semuanya bebas dirata-rata. Kode prodi kan juga pakai angka 1, 2, 3!”

Miskonsepsi B

“Skala Likert (1–5) itu jarak antar angkanya pasti sama rata seperti penggaris, jadi menghitung nilai **mean**-nya sudah pasti valid tanpa syarat.”

Miskonsepsi C

“Selama hasil akhirnya keluar berupa tabel urutan **ranking** yang rapi, pasti metode pengukurannya sudah benar.”

Blok 4 — Fakta dan interpretasi yang tepat

Kode Prodi — Nominal

Angka di sini murni sekadar label penamaan, tidak punya urutan. SI, TI, dan IF bukanlah sebuah jarak ukur.

Merata-rata kode jurusan sama sekali **tidak ada artinya**.

Operasi yang aman: menghitung jumlah kemunculan (frekuensi), mencari yang terbanyak (modus), atau persentase proporsinya.

Skala Likert 1–5 — Ordinal

Angkanya memang punya urutan (5 lebih baik dari 4), tapi jarak antarskornya tidak dijamin sama besar.

Mencari nilai tengah (median) atau persentil adalah langkah paling aman.

Di dunia kerja, nilai **mean** pada Likert memang sering dipakai agar praktis — tapi ini butuh catatan tebal, bukan lantas dianggap sebagai "jarak pasti".

> Akar masalahnya bukanlah pada wujud tabel urutannya, melainkan pada **operasi matematika** yang memaksakan diri menabrak aturan jenis datanya.

Blok 4 — Membangun kebiasaan berpikir kritis

1. Sebelum gegabah menghitung **mean** atau **median**, pastikan dulu: apa wujud skala datanya: nominal, ordinal, interval, atau rasio?
2. Kritis melihat tabel **spreadsheet**: tidak semua yang tertulis sebagai “angka” bebas ditambah atau dibagi begitu saja.
3. Pilih dulu rumus ringkasan yang memang diizinkan oleh skalanya, baru setelah itu Anda sah menyusun **leaderboard** atau mengambil keputusan.

> Alur berpikir yang benar: kenali variabelnya → pastikan skalanya → pilih operasi yang legal → hasilkan klaim “objektif” yang tak menyesatkan.

Konsep Utama: Empat Skala Pengukuran Data

Mengenal Kerangka Stevens — landasan teori di balik kasus **leaderboard** tadi:

Skala	Karakteristik	Operasi yang Sah	Contoh Kasus
Nominal	Sekadar label, tanpa kasta/urutan	Hitung frekuensi, modus	Kode jurusan
Ordinal	Punya urutan, tapi jarak antar level tidak pasti	Mencari median, persentil	Skor Likert 1–5
Interval	Jarak pasti, tapi angka nol-nya tidak mutlak	Menghitung selisih, **mean***	Suhu ruangan (°C)
Rasio	Jarak pasti dan punya angka nol mutlak	Bisa dibagi (rasio) & semua operasi	Durasi, jumlah klik

*Catatan penting: **Mean** (rata-rata) sangat legal di Interval/Rasio. Di lapangan, data Ordinal sering dipaksa dirata-rata, tetapi ini butuh asumsi ekstra. Namun ingat, jangan pernah merata-rata data Nominal!

Rincian Nominal — frekuensi & modus, bukan mean

KASUS: PENDAFTARAN MATA KULIAH PILIHAN (N = 12)

Data mentah (prodi + kode yang "dipetakan")

Mhs	Prodi	Kode	Mhs	Prodi	Kode
01	SI	1	07	TI	2
02	SI	1	08	IF	3
03	TI	2	09	SI	1
04	IF	3	10	SI	1
05	SI	1	11	TI	2
06	TI	2	12	SI	1

Hitung yang sah vs yang salah

Frekuensi: SI = 6 · TI = 4 · IF = 2

Proporsi SI: $\frac{6}{12} = 50\%$

Modus = SI

Kalau "mean kode" dipaksa:

$\frac{1+1+2+3+1+2+2+3+1+1+2+1}{12} = 1\{,\}67$

Angka 1,67 **tidak** berarti "prodi di antara SI dan TI". Label diganti angka — jaraknya fiktif.

Rincian Ordinal — urutan ada; jarak belum tentu sama

KASUS: SKOR KEPUASAN LAYANAN IT (LIKERT 1–5), N = 9

Skor responden

R01–R03	R04–R06	R07–R09
5 · 4 · 4	3 · 5 · 2	4 · 3 · 5

Urutkan: 2, 3, 3, 4, 4, 4, 5, 5, 5

Median (nilai ke-5) = **4**

Persentil ke-25 ≈ 3 · Persentil ke-75 ≈ 5

Mean sering dipakai — dengan catatan

$\overline{x} = \frac{5+4+4+3+5+2+4+3+5}{9} = 3\{,\}89$

Angka 3,89 tampak "rapi" untuk ranking unit.

Masalahnya: jarak 2⟶3 (tidak puas ⟶ netral) belum tentu sama dengan 4⟶5 (puas ⟶ sangat puas).

Aman default: median / persentil. Mean = pilihan praktis + asumsi jarak.

Rincian Interval — selisih & mean sah; rasio tidak

KASUS: SUHU PENDINGIN RUANG SERVER (°C) TIAP JAM, N = 5

Pembacaan sensor

Jam	°C	Δ dari jam 1
08:00	22	—
09:00	24	+2
10:00	26	+4
11:00	28	+6
12:00	30	+8

Mean: $\frac{22+24+26+28+30}{5} = 26°C$

Selisih 12:00 − 08:00 = **8°C** — bermakna.

Yang tidak boleh diklaim

"30°C dua kali lebih panas dari 15°C"?

Secara aritmetika $\frac{30}{15} = 2$, tapi **bukan** perbandingan suhu termodinamika.

0°C hanya titik beku air (konvensi), bukan "nol energi panas".

Sah: selisih, mean, rentang. Tidak sah: klaim "× kali lebih panas" pada °C.

Rincian Rasio — selisih, mean, dan perbandingan

KASUS: WAKTU MUAT HALAMAN BERANDA (DETIK), N = 5

Hasil uji load time

Uji	Detik	vs uji 1
1	1,2	1,00×
2	1,5	1,25×
3	2,0	1,67×
4	2,4	2,00×
5	3,0	2,50×

Mean: $\frac{1{,}2+1{,}5+2{,}0+2{,}4+3{,}0}{5} = 2{,}02$ dtk

Mengapa rasio di sini sah

0 detik = **tidak ada waktu berlalu** (nol absolut).

Uji 4 (2,4 dtk) memang **dua kali** lebih lambat dari uji 1 (1,2 dtk).

Sah juga: “halaman B 40% lebih cepat” bila baseline jelas.

Rasio = interval + nol mutlak ⟶ perbandingan “× kali” punya arti.

Konsep Pendamping: Kategorik vs Numerik (Bukan Pengganti Skala)

Kelompok Kategorik

Mencakup data berskala Nominal dan Ordinal.

Contoh: Nama program studi, status kelulusan, skor kepuasan, **rating** bintang.

Kelompok Numerik

Mencakup data berbentuk bilangan bulat terhitung (Diskret) atau rentang hasil pengukuran (Kontinu).

Contoh: Jumlah celah keamanan / **bug** (diskret); **latency** server, berat badan, durasi muat halaman (kontinu).

> Langkah terbaik: tetapkan dulu **skala pengukurannya** untuk tahu operasi hitung apa yang legal. Setelah itu, baru Anda jabarkan apakah itu diskret atau kontinu. Keduanya saling melengkapi.

Uji Pemahaman — Mari Mengklasifikasikan Data

A	Status kelulusan mahasiswa: Lulus / Tidak Lulus	Nominal
B	Survei kepuasan pelanggan dengan skala 1–5	Ordinal
C	Suhu pendingin ruang server (dalam °C)	Interval
D	Durasi **loading** halaman website (dalam detik)	Rasio
E	Jumlah barang yang dimasukkan ke **shopping cart**	Rasio · Diskret

Bahan diskusi: mengapa suhu Celsius (°C) masuk kategori Interval, sedangkan Kelvin (K) dianggap skala Rasio?

Diskusi: mengapa °C interval, Kelvin rasio?

Celsius (°C) — Interval

Nolnya **konvensi**: 0°C = titik beku air, bukan "tidak ada suhu".

Selisih tetap bermakna: naik 10°C dari 20 ke 30 sama besarnya dengan dari 30 ke 40.

Rasio **tidak** bermakna: 20°C **bukan** "dua kali lebih panas" dari 10°C.

Itulah ciri skala interval: jarak sah, nol arbitrer.

Kelvin (K) — Rasio

Nolnya **absolut fisis**: 0 K = energi kinetik molekul nol (titik nol absolut).

Selisih tetap bermakna — sama seperti di °C (beda hanya pergeseran nol).

Rasio **bermakna**: 200 K memang dua kali suhu termodinamika 100 K.

Itulah ciri skala rasio: jarak + nol absolut ⟶ perbandingan sah.

> Uji cepat: "Apakah 0 pada skala ini berarti ketiadaan besaran yang diukur?" Jika ya ⟶ kandidat rasio. Jika 0 hanya titik acuan yang disepakati ⟶ biasanya interval.

Blok 5 — Sumber data

Kasus: dataset publik untuk klaim lokal

Blok 5 — Studi Kasus

ASAL USUL DATA MENENTUKAN KEABSAHAN KLAIM

Pakai dataset “mahasiswa” dari Kaggle → berani klaim “berdampak signifikan bagi kampus UHB”?

Yang diunduh

Mengunduh file CSV berisi nilai mahasiswa dari situs Kaggle.

Asal negara, sistem kurikulum, maupun tahun datanya tidak diketahui pasti.

Sama sekali tidak ada survei yang dilakukan di lingkungan kampus UHB.

Yang ditulis di skripsi

“Faktor X memberikan dampak yang signifikan pada prestasi akademik mahasiswa UHB.”

Klaimnya menyasar populasi mahasiswa UHB, padahal datanya sama sekali bukan dari sana.

> Fokus di Blok 5 ini: apakah data tersebut tergolong primer atau sekunder? Lalu, kesimpulan seperti apa yang sebenarnya masuk akal untuk ditarik?

Blok 5 — Miskonsepsi yang sering terjadi

Miskonsepsi A

“Karena nama dataset-nya ‘mahasiswa’, berarti sah-sah saja dipakai untuk menebak performa mahasiswa UHB.”

Miskonsepsi B

“Data sekunder itu ibarat ‘barang bekas’ dan tidak ilmiah. Riset yang valid harus selalu menyebar kuesioner sendiri.”

Miskonsepsi C

“Jumlah baris datanya puluhan ribu dan **p-value**-nya sangat kecil, jadi pasti aman kalau digeneralisasi untuk kampus mana saja.”

Blok 5 — Fakta: primer, sekunder, dan klaim proporsional

Data Primer

Data yang digali secara spesifik **untuk menjawab riset Anda sendiri**.

Contoh: menyebar survei langsung di UHB, melakukan A/B **testing**, atau memasang sensor mandiri.

Kelebihan: Anda punya kendali penuh atas sampel dan definisi variabel. Kekurangan: butuh waktu dan tenaga ekstra.

Data Sekunder

Data hasil pengumpulan pihak lain yang Anda **manfaatkan kembali**.

Contoh: dataset Kaggle, rilis BPS, catatan **log** sistem kampus, hingga file CSV pada kasus di atas.

Kelebihan: sangat praktis dan cepat. Kekurangan: wajib memeriksa ulang tujuan awal, batasan populasi, dan definisi datanya.

> Pada kasus skripsi tadi: memakai data Kaggle sangat dibolehkan. Namun, yang keliru fatal adalah mengklaim bahwa populasi sumber data tersebut mewakili "mahasiswa UHB" tanpa adanya validasi dari data lokal.

Blok 5 — Membangun kebiasaan berpikir kritis

Empat hal yang wajib ditanyakan sebelum mengolah data sekunder:

1. Apa motivasi atau tujuan utama saat data ini pertama kali direkam?
2. Siapa saja subjek yang dilibatkan, dan kelompok mana yang dieksklusi dari kerangka (**frame**) sampelnya?
3. Bagaimana cara mengukur tiap variabelnya (perhatikan satuan ukur, kode angka, dan data yang kosong/**missing**)?
4. Apakah kita memiliki izin resmi dari segi hukum, etika, dan lisensi untuk memakainya ulang?

> Alur ideal: pahami sumbernya ⟶ audit batasan dan definisinya ⟶ sesuaikan klaim dengan kapasitas data ⟶ barulah Anda bisa menyusun kesimpulan yang "objektif".

Rangkuman Konsep: Data Primer vs Sekunder

Data Primer

Digali secara spesifik untuk menjawab riset Anda.

Contoh: menyebar survei, merancang A/B **testing**, observasi lapangan, atau memasang sensor mandiri.

Kelebihan: kendali penuh atas kualitas data. Kekurangan: memakan lebih banyak waktu dan biaya.

Data Sekunder

Warisan pengumpulan pihak lain yang Anda manfaatkan kembali.

Contoh: repositori Kaggle, portal BPS, riwayat **log** sistem kampus, atau publikasi resmi pemerintah.

Kelebihan: praktis dan instan. Kekurangan: wajib mengaudit definisi operasional dan celah biasnya.

Pelajaran dari Kasus Kaggle vs Kampus UHB

Ingatlah, kesimpulan yang valid hanya berlaku pada populasi **asli** tempat data itu diambil — jangan memaksakannya seolah itu cerminan kondisi di kampus Anda sendiri.

Kesimpulan: Intisari Pembelajaran Pertemuan 01

Konsep yang Telah Dikuasai

- Setiap klaim angka wajib memiliki konteks **sampling** yang jelas
- Fakta sampel (deskriptif) ≠ tebakan populasi (inferensial)
- Mengenali batasan antara populasi, sampel, unit observasi, & teknik penarikannya
- Memahami bahwa skala data mengunci operasi matematika apa yang legal
- Tahu cara mengaudit data sekunder sebelum diolah

Jebakan yang Kini Bisa Dihindari

- Percaya buta pada **headline** tanpa mengecek ukuran sampel (**n**) & metodenya
- Memaksakan data satu kelas seolah mewakili performa seluruh angkatan
- Mengklaim “skala nasional” dari survei sukarela yang mengelompok di satu kota
- Sembarangan merata-rata kode jurusan atau skor survei Likert mentah
- Memakai dataset asing demi memvalidasi kondisi kampus lokal

> Agenda Pertemuan 02: Ukuran Pemusatan Data — belajar merumuskan “titik kumpul data” dengan tepat, setelah kita mengantongi aturan main pengukurannya.