UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 4 Nomor 3 Tahun 2026 Halaman 335-345.

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Bayesian and Frequentist Inference Under Small Sample and Non-Normal Conditions: A Full-Factorial Monte Carlo Study

Aflah Zakinov Irta*, Rizal Kurniawan
Departemen Psikologi, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: aflah.zakinov@fpk.unp.ac.id

Submitted : 09 Mei 2026
Revised : 13 Agustus 2026
Accepted : 28 Agustus 2026

ABSTRACT

Small samples and non-normal data are the working reality of behavioural research, yet the conditions under which Bayesian estimation outperforms frequentist procedures remain poorly quantified. This study reports a full-factorial Monte Carlo experiment spanning 144 conditions and 144,000 replications, crossing four sample sizes, four population effect sizes, three data distributions, and three prior specifications. Three procedures are compared: the Welch unequal-variance t test, the Wilcoxon rank-sum test, and conjugate Bayesian estimation whose informative prior is derived from a formal random-effects synthesis of three meta-analytic sources. An informative prior cut mean absolute error by roughly three-fifths at the smallest sample size and lowered root mean square error in every condition examined, with gains shrinking steadily as sample size grew. The direction of the error trade-off reverses with sample size: at small samples the informative prior lowered type one error while raising power, whereas at the largest sample size it inflated type one error above the nominal level. Recalibrating the frequentist threshold to match empirical error rates left the Bayesian power advantage intact. Prior specification mattered far more than the choice of framework, yet its benefit collapsed once the prior was badly located, showing that the gains are conditional rather than general.

Keywords: *Bayesian, Frequentist, Prior Sensitivity, Credible Interval, Statistical Power.*

I. PENDAHULUAN

Penelitian ilmu sosial dan perilaku menghadapi kendala struktural: ukuran sampel jarang mencukupi untuk mencapai presisi yang memadai karena populasi klinis tertentu langka, prosedur eksperimental mahal, dan kriteria inklusi ketat. Marszalek et al. (2011) mengonfirmasinya secara empiris, dengan rerata ukuran sampel dalam empat jurnal psikologi terkemuka selama tiga dekade hanya N = 40. Lakens (2022) menilai kegagalan menjustifikasi ukuran sampel secara a priori sebagai sumber ketidakandalan yang paling dapat dihindari, sementara Vankelecom et al. (2025) mencatat prevalensi analisis daya hanya naik dari 9,5% menjadi 30% antara 2015 dan 2021. Nakagawa et al. (2024) mengidentifikasi siklus bias yang bersifat swasembada: kalkulasi daya dibangun di atas estimasi efek yang telah terkontaminasi selektivitas publikasi.

Kenyataan itu melahirkan paradoks metodologis, sebab perangkat inferensial yang mendominasi analisis data perilaku justru paling rentan pada sampel kecil. Cohen (1994) menegaskan bahwa nilai-p tidak menjawab pertanyaan yang ingin dijawab peneliti: yang diinginkan $P(H_0|D)$, yang tersedia hanyalah $P(D|H_0)$, dan Gigerenzer (2004) menunjukkan kebingungan antara keduanya tertanam secara institusional dalam cara statistik diajarkan. Wasserstein dan Lazar (2016) kemudian menerbitkan pernyataan resmi Asosiasi Statistik Amerika yang memperingatkan bahaya penggunaan mekanis ambang $p < 0,05$, dan Wasserstein et al. (2019) mengadvokasi cara yang lebih bernuansa dalam mengomunikasikan ketidakpastian.

Konsekuensinya mencapai titik kritis ketika Open Science Collaboration (2015) melaporkan hanya 36% dari 100 studi psikologi berhasil direplikasi. Button et al. (2013) mengidentifikasi akar strukturalnya, yakni daya statistik median dalam neurosains hanya 21%, yang dikombinasikan dengan tekanan publikasi menciptakan literatur yang melebih-lebihkan besaran efek (Ioannidis, 2005; Simmons et al., 2011; Schäfer dan Schwarz, 2019). Persoalan itu bertumpang tindih dengan persoalan distribusional: seluruh dari 440 dataset psikometri berskala besar yang ditelaah Micceri (1989)

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

menyimpang dari normalitas, dan Blanca et al. (2013) mengonfirmasi 74% variabel psikologi melanggar asumsi normalitas secara bermakna.

Inferensi Bayesian menawarkan alternatif yang menarik secara konseptual. Berbeda dari nilai-p yang bergantung pada data hipotetis yang tidak pernah teramati, distribusi posterior menguantifikasi ketidakpastian tentang parameter dengan memadukan pengetahuan awal dan informasi data (van de Schoot et al., 2021), dan Kruschke (2013) memperlihatkan bahwa estimasi Bayesian menghasilkan inferensi yang lebih kaya daripada uji-t konvensional. Klaim superioritas itu tetap perlu disikapi kritis. McNeish (2016) menunjukkan prior yang terlampau difus justru dapat memperburuk estimasi pada sampel kecil, dan Smid et al. (2020) menegaskan keunggulan Bayesian bersifat kondisional: konsisten muncul ketika prior dispesifikasikan cermat, namun menghilang ketika prior difus digunakan sembarangan. Depaoli et al. (2020) menekankan analisis sensitivitas prior berpotensi mengubah kesimpulan, sejalan dengan daftar periksa WAMBS dari Depaoli dan van de Schoot (2017), sementara regularisasi moderat seperti diadvokasikan Gelman (2006) menawarkan jalan tengah.

Studi ini mengisi kesenjangan tersebut melalui simulasi Monte Carlo faktorial penuh dengan 144.000 replikasi. Tiga keputusan desain membedakannya dari studi sejenis: kondisi efek nol disertakan secara eksplisit agar Galat Tipe I sejati dapat diestimasi, bukan diproksikan; prior informatif diturunkan melalui sintesis meta-analitik random-effects formal, bukan perataan informal; dan uji Wilcoxon–Mann–Whitney disertakan sebagai pembanding frekuentis nonparametrik agar perbandingan pada kondisi nonnormal bersifat setara. Karena mean prior dipertahankan tetap sementara δ populasi bervariasi, desain ini sekaligus memetakan konsekuensi misspesifikasi prior pada rentang −0,45 hingga +0,35. Berangkat dari kesenjangan tersebut, penelitian ini mengajukan tiga persoalan yang saling terkait. Pertama, perbandingan kinerja: sejauh mana estimasi Bayesian, uji-t Welch, dan uji Wilcoxon berbeda dalam bias bertanda, MAE, RMSE, dan tingkat cakupan ketika ukuran sampel, ukuran efek, dan bentuk distribusi divariasikan sistematis. Kedua, peran prior sebagai pemoderasi, yakni seberapa jauh spesifikasi prior menentukan kinerja Bayesian dan apa konsekuensi misspesifikasinya terhadap akurasi estimasi serta kalibrasi interval. Ketiga, pertukaran antarkriteria, yaitu pada kondisi apa peningkatan daya dibarengi inflasi Galat Tipe I sehingga keduanya perlu ditimbang bersamaan. Ketiga persoalan itu menjadi kerangka penyajian hasil dan pembahasan.

II. METODE PENELITIAN

A. Desain Simulasi dan Pembangkitan Data

Seluruh data dibangkitkan lewat komputasi. Desain faktorial penuh menyilangkan empat faktor secara ortogonal: ukuran sampel ($n = 20, 50, 100, 200$; dibagi seimbang menjadi $n_1 = n_2 = n/2$), ukuran efek populasi ($\delta = 0{,}0; 0{,}2; 0{,}5; 0{,}8$), jenis distribusi (normal, menceng, ekor berat), dan spesifikasi prior (tak-informatif, lemah-informatif, informatif), sehingga terbentuk 144 kondisi unik yang masing-masing direplikasi 1.000 kali dengan benih acak tetap. Karena level prior tidak memengaruhi pembangkitan data, ketiganya diterapkan pada himpunan simulasi yang sama untuk tiap kombinasi (n, δ, distribusi); angka 144.000 menghitung kondisi-replikasi, sedangkan himpunan data unik berjumlah 48.000. Kondisi $\delta = 0{,}0$ sengaja disertakan karena Galat Tipe I hanya dapat diukur ketika hipotesis nihil benar dan tidak dapat diproksikan dari kondisi efek kecil. Replikasi 1.000 menghasilkan SE Monte Carlo 0,0069 di sekitar $p = 0{,}95$, di bawah ambang 0,010 yang direkomendasikan Sigal dan Chalmers (2016).

Alur prosedurnya disajikan pada Gambar 1 dan berjalan dalam dua tingkat pengulangan bersarang. Pengulangan dalam menangani 1.000 replikasi untuk satu kondisi: dua sampel dibangkitkan menurut distribusi dan ukuran efek yang berlaku, lalu dianalisis serentak oleh ketiga prosedur sehingga estimasi, interval, dan status penolakan terekam pada replikasi yang sama; penerapan pada data identik inilah yang memungkinkan kontras antarprosedur dihitung dengan memperhitungkan korelasi antarmetode. Pengulangan luar menelusuri 144 kondisi desain, dan matriks 144 baris yang terbentuk menjadi masukan bagi analisis sensitivitas prior serta pemetaan misspesifikasi. Ketiga distribusi di-*rescaling* agar varians populasi sama dengan satu: normal memakai $N(0,1)$ dan $N(\delta,1)$; menceng memakai Gamma(bentuk = 4; skala = 0,5) yang dipusatkan di nol dengan kemencengan populasi 1,0; ekor berat memakai $t_5 \times \sqrt{(3/5)}$ karena $\mathrm{Var}(t_5) = 5/3$. Verifikasi atas 2.000.000 pengamatan mengonfirmasi ketiga momen tersebut, sekaligus menunjukkan momen orde tinggi yang teramati pada sampel kecil bias ke bawah, misalnya kemencengan Gamma yang hanya 0,51 dari nilai populasi 1,00 pada $h = 10$.

B. Prosedur Estimasi dan Aturan Keputusan

Tiga prosedur dibandingkan. Uji-t Welch mewakili prosedur frekuentis parametrik yang paling lazim digunakan, sedangkan uji Wilcoxon–Mann–Whitney disertakan sebagai pembanding nonparametrik; tanpanya, perbandingan pada kondisi nonnormal menjadi asimetris karena mempertentangkan prosedur yang tangguh dengan prosedur yang

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

mengalami misspesifikasi (Kelter, 2021). Uji Wilcoxon–Mann–Whitney secara formal menguji $Pr(X_1 < X_2) + Pr(X_1 = X_2)/2 = 0{,}5$, bukan selisih rerata atau median (Divine *et al.*, 2018); karena desain ini menerapkan pergeseran lokasi murni dengan bentuk distribusi identik antarkelompok, hipotesis nihil ketiga prosedur berimpit sehingga perbandingan tingkat penolakan tetap sah, meskipun ekuivalensi itu tidak berlaku umum di luar model geser-lokasi.

Estimand pada ketiga prosedur adalah $\delta = (\mu_2 - \mu_1)/\sigma$, yang diduga melalui Hedges' g pada Persamaan (1), dengan s_p simpangan baku gabungan dan J faktor koreksi bias untuk $df = n_1 + n_2 - 2$ (Hedges, 1981):

$$g = \frac{\bar{x}_2 - \bar{x}_1}{s_p} \cdot J, \qquad J = 1 - \frac{3}{4df - 1} \tag{1}$$

Varians *likelihood* memakai ekspresi Hedges dan Olkin (1985) untuk d pada Persamaan (2), tanpa faktor J^2. Pada $n = 20$ kelalaian ini melebihkan varians *likelihood* sekitar 9,0% ($1/J^2$ dari Persamaan (1)), bukan 21%. Presisi *likelihood* yang sedikit lebih rendah menaikkan bobot prior dalam presisi posterior, sehingga reduksi MAE dan penurunan Galat Tipe I yang dilaporkan pada Bagian III sedikit lebih optimistis daripada semestinya, sedangkan cakupan interval kredibel sedikit terlalu lebar; pergeseran bobot prior akibat kelalaian ini tidak lebih dari dua poin persentase pada $n = 20$ dan mengecil pada n yang lebih besar.

$$\sigma^2_{lik} = \frac{1}{n_1} + \frac{1}{n_2} + \frac{g^2}{2(n_1 + n_2)} \tag{2}$$

Prosedur Bayesian memadukan prior $N(\mu_0, \sigma_0^2)$ dengan *likelihood* melalui pembaruan konjugat Normal–Normal: presisi posterior adalah jumlah presisi prior dan presisi *likelihood*, varians posterior kebalikannya, dan mean posterior rata-rata terbobot keduanya, sebagaimana Persamaan (3):

$$\sigma^2_{post} = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2_{lik}}\right)^{-1}, \qquad \mu_{post} = \sigma^2_{post}\left(\frac{\mu_0}{\sigma_0^2} + \frac{g}{\sigma^2_{lik}}\right) \tag{3}$$

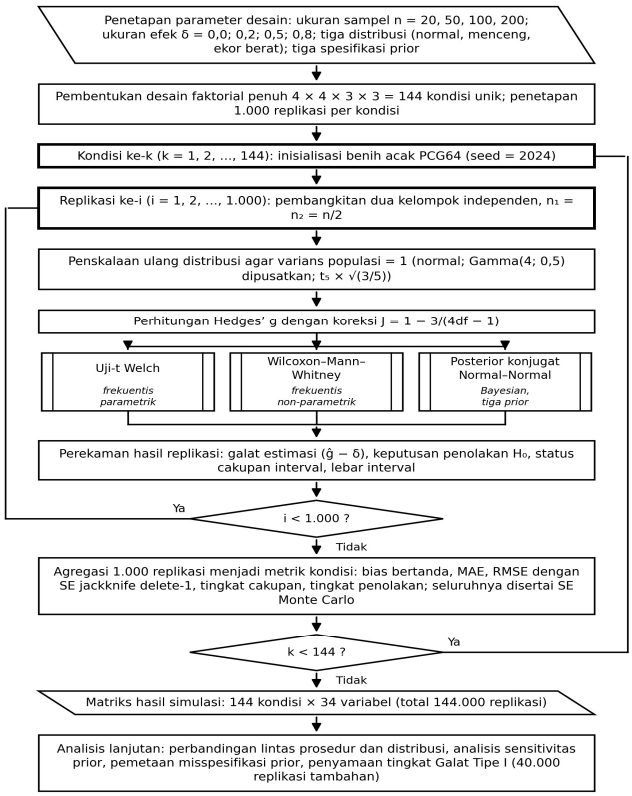
Estimasi titik memakai μ_post sebagai penduga δ, dinilai melalui bias bertanda, MAE, dan RMSE; estimasi interval memakai interval kredibel ekor-sama 95% pada Persamaan (4). Keputusan biner memakai aturan eksklusi nol, yakni H_0 ditolak apabila interval kredibel tidak memuat $\delta = 0$ — setara secara struktural dengan penolakan berbasis interval kepercayaan frekuentis sehingga tingkat penolakan kedua kerangka dapat dibandingkan langsung. Faktor Bayes tidak dievaluasi karena menjawab pertanyaan berbeda, yakni rasio bukti relatif antarmodel (Tendeiro dan Kiers, 2019).

$$IKr_{95\%} = \mu_{post} \pm 1{,}96 \cdot \sigma_{post} \tag{4}$$

Pembaruan konjugat diterapkan seragam lintas ketiga distribusi. Keputusan ini disengaja: pertanyaannya bukan seberapa baik kinerja model Bayesian yang dispesifikasikan sempurna, melainkan apa yang terjadi ketika praktisi menerapkan prosedur Bayesian standar pada data yang melanggar normalitas, sehingga kondisi nonnormal berfungsi sebagai uji ketangguhan terhadap misspesifikasi *likelihood*.

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 1. Diagram alir prosedur simulasi Monte Carlo.

C. Spesifikasi Prior

Prior informatif diturunkan melalui sintesis *random-effects* formal atas tiga sumber meta-analitik, dengan korelasi dikonversi ke skala d melalui $d = 2r/\sqrt{(1-r^2)}$. Ketiganya melaporkan: Richard *et al.* (2003) d = 0,4296 (SE = 0,0147; k = 474, psikologi sosial), Gignac dan Szodorai (2016) d = 0,3871 (SE = 0,0106; k = 708, psikologi diferensial), dan Szucs dan Ioannidis (2017) d = 0,2400 (SE = 0,0037; k = 26.841, psikologi kognitif dan neurosains). Heterogenitasnya ekstrem: $Q(2) = 303{,}8$ ($p < 0{,}001$), $I^2 = 99{,}3\%$ (Higgins dan Thompson, 2002), yang mengonfirmasi ketiganya mengukur populasi efek berbeda sehingga sintesis *random-effects* merupakan pendekatan yang tepat.

Dengan hanya tiga sumber, penduga DerSimonian dan Laird (1986) diketahui menghasilkan interval terlalu sempit dan tingkat galat melampaui nominal (IntHout *et al.*, 2014), sehingga sintesis ini memakai penduga *restricted maximum likelihood* (REML) untuk τ^2 beserta interval Hartung–Knapp–Sidik–Jonkman. Lokasi δ kokoh terhadap pilihan penduga, sebab DerSimonian–Laird, REML, dan Paule–Mandel seluruhnya menghasilkan $\delta \approx 0{,}352$; dengan REML ($\tau^2 = 0{,}0099$) sintesis memberikan $\delta = 0{,}352$ dengan interval 95% [0,104; 0,599]. Lokasi prior ditetapkan pada 0,35, sedangkan SD prior 0,30 dipilih sebanding dengan setengah-lebar interval tersebut (0,248) sehingga prior tidak lebih yakin daripada bukti yang mendasarinya. Mengingat k = 3, sintesis ini diperlakukan sebagai prosedur elisitasi prior yang terdokumentasi, bukan meta-analisis formal yang berdiri sendiri; rekaman pada Szucs dan Ioannidis (2017) juga tidak sepenuhnya independen sehingga SE-nya berpotensi terlalu kecil, meskipun dampaknya terbatas karena pembobotan didominasi τ^2 sehingga bobot ketiga sumber hampir seragam (33,0%, 33,3%, dan 33,7%). Prior informatif $N(0{,}35;\ 0{,}30^2)$ menempatkan 62,4% massa probabilitas pada rentang efek kecil hingga sedang ($0{,}2 < \delta < 0{,}8$) dan hanya 6,7% pada wilayah efek besar ($\delta > 0{,}8$), sehingga tidak bersifat over-optimistik. Dua prior pembanding ditetapkan sebagai acuan dasar dan sebagai regularisasi, masing-masing $N(0;\ 10^2)$ dan $N(0;\ 1^2)$, keduanya simetris di sekitar nol. Perlu diakui bahwa ketiga sumber ini sendiri berasal dari literatur terpublikasi dan berpotensi mewarisi inflasi akibat selektivitas publikasi yang disinggung pada Pendahuluan (Nakagawa *et al.*, 2024); sebagian keunggulan yang dilaporkan pada $\delta = 0{,}5$ karenanya turut mencerminkan kedekatan antara lokasi prior dan efek yang diuji, bukan semata keunggulan struktural prosedur.

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

D. Teknik Analisis Data

Lima metrik dihitung per kondisi: bias bertanda, *Mean Absolute Error*, RMSE beserta SE, tingkat cakupan interval 95%, dan tingkat penolakan H_0 yang diinterpretasikan sebagai Galat Tipe I sejati ketika $\delta = 0{,}0$ dan sebagai daya ketika $\delta > 0$. Empat metrik pertama menuntut penduga δ sehingga hanya dihitung untuk uji-t Welch dan Bayesian; Wilcoxon, yang tidak menduga δ, dievaluasi semata melalui tingkat penolakan. SE *jackknife* dipilih karena deterministik dan tidak menambah variabilitas seperti *bootstrap*. Karena seluruh prosedur diterapkan pada *dataset* yang sama, SE Monte Carlo untuk kontras antarprosedur dihitung sebagai $\mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A, B)$ (Morris *et al.*, 2019); perbedaan dinyatakan bermakna apabila selisihnya melampaui 2×SE-MC kontras tersebut. Cakupan frekuentis diperlakukan sebagai kriteria evaluasi eksternal demi komparabilitas lintas kerangka, bukan klaim bahwa interval kredibel harus mencapai cakupan nominal (Datta dan Mukerjee, 2004; Rousseau, 2016). Interval kredibel dibentuk dari kuantil normal 1,96 sedangkan interval frekuentis dari kuantil-t Welch; pada $n = 20$ keduanya berbeda 6,7% (2,101 versus 1,960), sehingga sebagian selisih lebar interval bersumber dari pilihan kuantil, bukan kontribusi prior. Lebar interval, simpangan baku pusat posterior, dan bobot prior turut direkam untuk mendiagnosis deviasi cakupan. Tiga analisis tambahan dilakukan. Pertama, ambang penolakan uji-t Welch dikalibrasi ulang hingga Galat Tipe I empirisnya menyamai prosedur Bayesian pada kondisi yang sama, lalu daya keduanya dibandingkan melalui 40.000 replikasi tambahan per kondisi pada distribusi normal (SE-MC ≈ 0,0011–0,0012). Kedua, konsekuensi misspesifikasi prior dianalisis melalui selisih $\mu\_prior - \delta$ pada tiap kondisi. Ketiga, karena desain utama hanya mencakup $\delta \geq 0$ sehingga tidak dapat mendeteksi asimetri aturan eksklusi nol terhadap efek berlawanan arah dengan prior, pemeriksaan tambahan pada $\delta = -0{,}2$ dan $-0{,}5$ dijalankan secara independen pada distribusi normal (4.000 replikasi per kondisi, benih terpisah dari simulasi utama; SE-MC ≈ 0,0079). Seluruh perhitungan dijalankan dalam Python 3.12 dengan NumPy 1.26 (pembangkitan acak PCG64, benih 2024 untuk simulasi utama) dan SciPy 1.13; implementasi paralel dalam R 4.4.1 disediakan sebagai suplemen

III. HASIL DAN PEMBAHASAN

A. Akurasi Estimasi

Pada distribusi normal, bias frekuentis berada dalam rentang −0,021 hingga +0,014 dan tidak melampaui 2×SE-MC gabungan di kondisi mana pun, sesuai prediksi teori. Pada ekor berat koreksi itu tidak sepenuhnya menghilangkan bias: pada $n = 20$ dengan $\delta = 0{,}8$ bias mencapai +0,040 (SE-MC = 0,015), melampaui 2×SE-MC, karena faktor koreksi J diturunkan di bawah asumsi normalitas (Hedges, 1981). Estimator Bayesian berprior informatif menunjukkan bias yang mengikuti arah *shrinkage* terhadap mean prior: positif ketika δ di bawah 0,35 dan negatif ketika di atasnya.

Selisih akurasinya besar. Pada $n = 20$ dengan $\delta = 0{,}5$, MAE Bayesian 0,141 (SE-MC 0,003) berbanding 0,376 (SE-MC 0,010) pada frekuentis, yakni reduksi 62,3%; pada $\delta = 0{,}2$ reduksinya setara (62,1%). Keunggulan menyempit menjadi 42,9% ($n = 50$), 28,2% ($n = 100$), dan 16,4% ($n = 200$), konsisten dengan peluruhan kontribusi prior seiring akumulasi informasi data. Bias Bayesian menyusut dari −0,113 ($n = 20$) menjadi −0,030 ($n = 200$); bias frekuentis tetap dalam ±0,010. RMSE, yang menggabungkan bias dan varians, lebih rendah pada seluruh 36 kondisi berefek dengan prior lemah-informatif, menyusut menjadi 30 dari 36 kondisi dengan prior tak-informatif — manfaat regularisasi karenanya memerlukan prior yang setidaknya membatasi ruang parameter. Ketangguhan terhadap pelanggaran distribusional juga nyata: reduksi MAE ($n = 20$, $\delta = 0{,}5$) tetap konsisten yakni 62,3% pada normal, 61,2% pada menceng, dan 63,1% pada ekor berat, meskipun *likelihood* Normal–Normal mengalami misspesifikasi pada dua kondisi terakhir.

B. Kalibrasi Interval

Prosedur frekuentis dan Bayesian berprior difus mencapai cakupan mendekati nominal 0,95 (0,941, sedikit di bawah nominal namun dalam 2×SE-MC; 0,953; dan 0,962 pada $n = 20$), sedangkan prior informatif menunjukkan *over-coverage* substansial yang mencapai 0,995 pada $n = 20$ dengan $\delta = 0{,}5$. Menyimpulkan bahwa interval "terlalu konservatif" akan menyesatkan, sebab interval kredibel Bayesian memang tidak dirancang memenuhi jaminan cakupan frekuentis. Diagnostik lebar interval menjelaskan sebabnya. Pada $n = 20$ dengan prior informatif, bobot prior mencapai 70,1% sehingga pusat posterior sangat stabil (SD 0,137) sementara lebar interval tetap 0,985; rasio setengah-lebar interval terhadap simpangan baku pusatnya 3,60, dibandingkan 1,88 pada prior tak-informatif — interval bergerak terlalu sedikit relatif terhadap lebarnya sehingga hampir selalu memuat nilai sejati. *Over-coverage* ini konsekuensi dominasi prior atas *likelihood*, bukan sifat interval yang sekadar konservatif; peneliti yang menjumpainya sebaiknya

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

menghitung bobot prior, dan bila melampaui sekitar setengah, melaporkan plot tumpang-tindih prior–posterior (Kruschke, 2021).

C. Galat Tipe I dan Daya Statistik

Uji-t Welch menjaga Galat Tipe I pada rentang 0,041–0,055 dan Wilcoxon pada 0,033–0,051, keduanya terkalibrasi baik. Perilaku prior informatif jauh lebih mengejutkan: bertentangan dengan ekspektasi bahwa prior yang menggeser massa ke arah efek positif selalu menginflasi Galat Tipe I, arah *trade-off* justru berbalik seiring bertambahnya ukuran sampel. Pada n = 20, prior informatif menghasilkan Galat Tipe I sebesar 0,031, lebih rendah daripada 0,049 pada uji-t Welch (z = −12,7), sekaligus menaikkan daya dari 0,180 menjadi 0,214, sehingga tampak sebagai dominasi Pareto. Keunggulan itu satu arah: aturan eksklusi nol menolak H_0 hanya ketika pusat posterior melampaui ambang di atas nol, sehingga nyaris tidak berdaya mendeteksi efek berlawanan arah dengan prior. Pemeriksaan tambahan independen pada δ = −0,5 (n = 20) menunjukkan daya uji-t Welch tetap simetris (0,188, sebanding 0,197 pada δ = +0,5), sedangkan daya Bayesian runtuh ke 0,002; pada δ = −0,2 dayanya (0,011) bahkan di bawah Galat Tipe I nominalnya sendiri — dominasi Pareto pada n = 20 dengan demikian bersifat kondisional pada arah efek. Pola inflasi Galat Tipe I sejati (δ = 0,0) sendiri tidak bertahan pada n = 50 (0,053 versus 0,049; z = +2,5), berlanjut pada n = 100 (0,057 versus 0,052) dan n = 200 (0,058 versus 0,049).

Mekanismenya dapat diuraikan. Aturan eksklusi nol menolak H_0 apabila pusat posterior melampaui setengah-lebar interval kredibel. Pada n = 20 presisi *likelihood* rendah sehingga interval tetap lebar meskipun pusatnya tertarik ke arah prior, akibatnya nol jarang tereksklusi ketika δ = 0. Seiring bertambahnya n, interval menyempit lebih cepat daripada meluruhnya tarikan prior sehingga pergeseran lokasi yang tersisa mulai cukup untuk mengeksklusi nol secara keliru, dengan titik balik di sekitar n = 50. Kekhawatiran terhadap inflasi Galat Tipe I akibat prior informatif karenanya paling relevan pada sampel sedang hingga besar, bukan pada sampel yang sangat kecil.

Klaim dominasi Pareto itu masih perlu diperiksa karena aturan eksklusi nol tidak dikalibrasi ke α = 0,05. Analisis kalibrasi pada Tabel 1 menutup celah tersebut dan memperkuat temuan sebelumnya: pada n = 20, penyamaan tingkat galat ke 0,031 menurunkan daya uji-t Welch dari 0,180 menjadi 0,132 sedangkan daya Bayesian tetap 0,214, sehingga kelebihan 0,082 bertahan utuh. Kelebihan itu memuncak pada n = 50 (0,130) dan tetap positif pada n = 100 (0,102) serta n = 200 (0,023), sehingga keunggulan daya Bayesian bersumber dari informasi yang dikandung prior, bukan dari tingkat galat yang lebih longgar.

Tabel 1. Daya pada tingkat Galat Tipe I yang disamakan (distribusi normal, δ = 0,5)

n	Uji-t Welch α = 0,05	Daya uji-t Welch	Bayes-IN α empiris	Daya Bayes-IN	Daya uji-t Welch terkalibrasi	Selisih
20	0,049	0,180	0,031	0,214	0,132	+0,082
50	0,049	0,404	0,053	0,548	0,419	+0,130
100	0,052	0,695	0,057	0,815	0,712	+0,102
200	0,049	0,941	0,058	0,972	0,949	+0,023

D. Ketangguhan terhadap Pelanggaran Asumsi

Penyertaan Wilcoxon terbukti bukan formalitas, seperti terlihat pada Tabel 2. Ketiga prosedur menjaga Galat Tipe I dalam rentang nominal pada seluruh kondisi nonnormal (0,028–0,064). Pada distribusi normal Wilcoxon kehilangan sedikit daya dibandingkan uji-t Welch, sesuai prediksi teori, namun pada distribusi nonnormal hubungan itu berbalik: pada n = 100 distribusi menceng daya Wilcoxon 0,777 melampaui 0,721 milik uji-t Welch, dan pada ekor berat selisihnya lebih besar (0,780 versus 0,692). Perbandingan Bayesian terhadap uji-t Welch saja karenanya melebih-lebihkan keunggulannya pada kondisi nonnormal. Terhadap pembanding yang lebih kuat itu, Bayesian tetap unggul pada sampel kecil hingga sedang (n = 20 distribusi menceng: 0,216 berbanding 0,170; n = 50: 0,563 berbanding 0,445), namun keunggulan itu habis pada sampel besar — pada n = 200 daya keduanya praktis identik baik pada distribusi menceng (0,972 berbanding 0,971) maupun ekor berat (0,978 berbanding 0,977). Pada data nonnormal bersampel besar, prosedur nonparametrik sederhana dengan demikian mencapai kinerja setara tanpa memerlukan justifikasi prior.

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

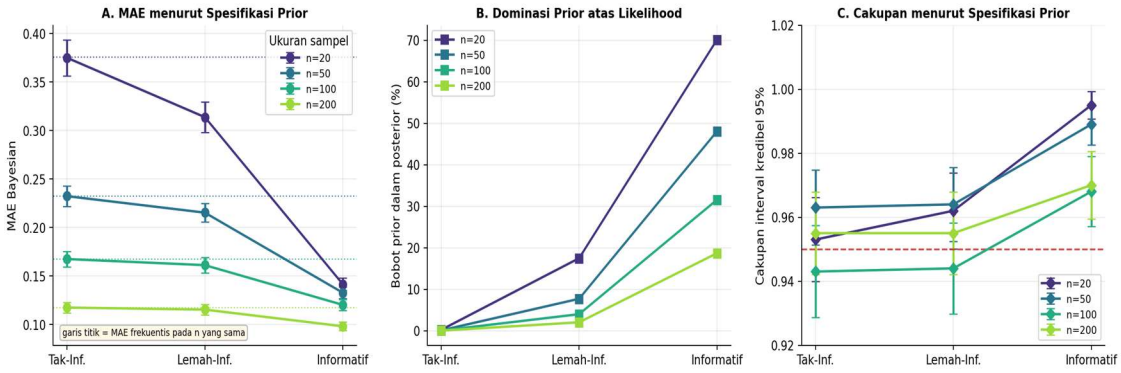
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Tabel 2. Galat Tipe I dan daya tiga prosedur pada distribusi nonnormal

Distribusi	n	Tipe I uji-t Welch	Tipe I Wilcoxon	Tipe I Bayes-IN	Daya uji-t Welch	Daya Wilcoxon	Daya Bayes-IN
Menceng	20	0,055	0,051	0,031	0,171	0,170	0,216
Menceng	50	0,038	0,044	0,051	0,411	0,445	0,563
Menceng	100	0,054	0,054	0,047	0,721	0,777	0,831
Menceng	200	0,045	0,045	0,050	0,947	0,971	0,972
Ekor berat	20	0,045	0,041	0,028	0,192	0,199	0,229
Ekor berat	50	0,042	0,037	0,054	0,410	0,466	0,562
Ekor berat	100	0,046	0,045	0,064	0,692	0,780	0,813
Ekor berat	200	0,060	0,053	0,062	0,944	0,977	0,978

E. Sensitivitas dan Misspesifikasi Prior

Temuan menarik dapat dilihat dari perbandingan langsung antara pilihan kerangka dan pilihan prior. Pada n = 20 dengan δ = 0,5 pada distribusi normal, MAE uji-t Welch 0,3756 sementara MAE Bayesian berprior tak-informatif 0,3748 — selisih 0,0008 yang mengikuti langsung dari Persamaan (3): dengan SD prior 10, bobot prior dalam presisi posterior hanya 0,21%, sehingga posterior nyaris identik dengan *likelihood* apa pun lokasinya. Perbedaan bermakna baru muncul ketika SD prior diperketat, yakni 0,3136 pada SD = 1 dan 0,1415 pada SD = 0,30. Karena bobot prior pada Persamaan (3) hanya ditentukan oleh SD prior, bukan lokasinya, penurunan MAE ini terutama mencerminkan penyempitan skala prior; kedekatan lokasi prior (0,35) dengan δ yang diuji (0,5) turut menyumbang, sehingga perbandingan tiga-prior ini tidak memisahkan efek skala dari efek lokasi secara bersih. Gambar 2 menjelaskan pola bobot ini lebih lanjut. Panel (b) menunjukkan bahwa pada n = 20 prior tak-informatif menyumbang 0,21% presisi posterior, lemah-informatif 17,5%, dan informatif 70,1%, sedangkan pada n = 200 kontribusi prior informatif meluruh menjadi 18,7%. Panel (a) memperlihatkan konsekuensinya, yakni jarak MAE antarspesifikasi prior terlebar pada n = 20 dan menyempit seiring bertambahnya n, sementara panel (c) menunjukkan cakupan justru menjauhi nominal 0,95 tepat ketika bobot prior memuncak. Ketika kontribusi prior mendekati nol, posterior merupakan reparameterisasi *likelihood* sehingga kedua kerangka konvergen; peneliti yang beralih ke Bayesian dengan prior *default* yang difus karenanya tidak memperoleh keuntungan inferensial apa pun.



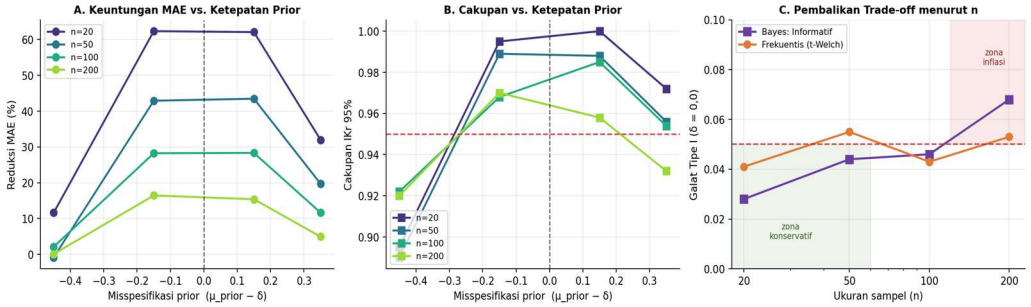
Gambar 2. Sensitivitas prior pada δ = 0,5 (distribusi normal): (a) MAE menurut spesifikasi prior, (b) bobot prior dalam presisi posterior, dan (c) cakupan interval kredibel.

Karena mean prior tetap 0,35 sementara δ populasi bervariasi dari 0,0 hingga 0,8, desain ini otomatis memetakan konsekuensi misspesifikasi prior pada rentang −0,45 hingga +0,35 (Gambar 3). Panel (a) memperlihatkan reduksi MAE memuncak di sekitar misspesifikasi nol dan meluruh ke kedua arah: pada misspesifikasi kecil (δ = 0,2 atau 0,5) reduksi mencapai 62%, sedangkan pada δ = 0,8 (menyimpang −0,45) reduksi tinggal 11,6% dengan cakupan turun ke 0,895 pada panel (b) — *under-coverage* yang bermakna. Pada δ = 0,0 (misspesifikasi +0,35), reduksi MAE tetap 32,0% namun bias bertanda mencapai +0,237 karena *shrinkage* menarik estimasi menjauh dari nol. Panel (c) merangkum pembalikan arah *trade-off* pada subbab C: kurva selisih Galat Tipe I memotong garis nol di sekitar n = 50. Keuntungan

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

itu karenanya hanya berlaku ketika efek sesungguhnya berada dalam rentang yang lazim ditemui dalam literatur perilaku ($\delta \approx 0{,}2$–$0{,}5$); di luarnya peneliti menghadapi risiko *under-coverage* dan bias substansial. Prior lemah-informatif menawarkan kompromi yang lebih aman: pada misspesifikasi terberat ($\delta = 0{,}8$; $n = 20$) prior informatif menghasilkan bias $-0{,}327$ dengan cakupan 0,895, sedangkan prior lemah-informatif menekan bias menjadi $-0{,}149$ dan memulihkan cakupan ke 0,967 dengan reduksi MAE sedikit lebih besar (13,5% berbanding 11,6%).



Gambar 3. Konsekuensi misspesifikasi prior: (a) reduksi MAE dan (b) cakupan menurut selisih $\mu\_prior - \delta$, serta (c) pembalikan arah trade-off Galat Tipe I menurut ukuran sampel.

F. Implikasi Praktis dan Keterbatasan

Panduan berikut bersifat kondisional dan disertai *trade-off* yang harus dilaporkan. Pada $n \leq 50$ dengan efek diantisipasi di kisaran $\delta \approx 0{,}2$–$0{,}5$, prior informatif meta-analitik memberi keuntungan terbesar berupa reduksi MAE 43–62% dengan Galat Tipe I di bawah nominal pada $n = 20$; bobot prior dan hasil di bawah prior alternatif wajib dilaporkan. Pada ukuran sampel sama namun efek melampaui $\delta = 0{,}8$ atau tidak diketahui, prior lemah-informatif lebih dianjurkan karena tetap menurunkan RMSE tanpa risiko misspesifikasi lokasi. Pada $n = 100$ keuntungan MAE moderat (28%); pada $n \geq 200$ keuntungan menyusut ke 16% sementara Galat Tipe I terinflasi sehingga prosedur frekuentis sudah memadai. Pada data sangat menceng, Wilcoxon sebaiknya dijadikan pembanding frekuentis, bukan uji-t Welch. Panduan tersebut dibatasi oleh cakupan simulasi. Kondisi nonnormal hanya mencakup satu tingkat kemencengan populasi (1,0) dan satu derajat ekor berat (t_5), sedangkan desainnya terbatas pada perbandingan dua kelompok independen dengan pergeseran lokasi murni; pelanggaran yang lebih ekstrem menuntut pemodelan distribusional eksplisit melalui MCMC penuh, misalnya dengan brms (Bürkner, 2017). Prior informatif juga diturunkan dari tiga sumber meta-analitik pada ranah psikologi, sehingga rentang efek yang mendasarinya belum tentu berlaku pada disiplin lain, sebagaimana ditunjukkan Delfin (2023) untuk korelasi bivariat psikologis yang menuntut elisitasi prior tersendiri.

IV. KESIMPULAN

Simulasi terhadap 144.000 replikasi menunjukkan bahwa keunggulan inferensi Bayesian atas prosedur frekuentis bersifat kondisional dan sangat ditentukan oleh spesifikasi prior, bukan sekadar pada pilihan kerangka inferensinya. Estimasi Bayesian dengan prior informatif terbukti secara signifikan meningkatkan akurasi titik, memberikan daya uji yang lebih tinggi, serta menekan risiko Galat Tipe I pada ukuran sampel kecil ($n = 20$) asalkan efek sejati selaras dengan prior tersebut. Namun, dominasi ini menyusut pada ukuran sampel sedang hingga besar, di mana risiko inflasi Galat Tipe I justru meningkat dan kinerja metode frekuentis seperti uji Wilcoxon mampu menyamai Bayesian pada data non-normal. Mengingat keunggulan Bayesian menurun drastis jika nilai prior meleset dari nilai sejati, penggunaan prior informatif paling direkomendasikan hanya pada kondisi sampel terbatas. Peneliti juga dituntut untuk memilih metode sesuai kondisi spesifik penelitian dengan melaporkan bobot prior, melakukan analisis sensitivitas, serta mengevaluasi daya uji dan Galat Tipe I secara bersamaan.

Sebagai saran untuk penelitian selanjutnya, cakupan pengujian perlu diperluas pada berbagai tingkat pelanggaran distribusional yang lebih ekstrem, serta memetakan kinerja metode inferensi pada desain penelitian yang lebih kompleks seperti pengukuran berulang dan model multilevel. Selain itu, perumusan prior pada studi mendatang sebaiknya diturunkan langsung dari meta-analisis pada domain yang relevan agar rentang efek yang mendasarinya benar-benar sesuai dengan karakteristik bidang penelitian yang bersangkutan..

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

DAFTAR PUSTAKA

Blanca, M.J., Arnau, J., López-Montiel, D., Bono, R., dan Bendayan, R. (2013), "Skewness and kurtosis in real data samples", *Methodology*, Vol. 9, No. 2, hal. 78–84. https://doi.org/10.1027/1614-2241/a000057

Bürkner, P.-C. (2017), "brms: An R package for Bayesian multilevel models using Stan", *Journal of Statistical Software*, Vol. 80, No. 1, hal. 1–28. https://doi.org/10.18637/jss.v080.i01

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., dan Munafò, M.R. (2013), "Power failure: Why small sample size undermines the reliability of neuroscience", *Nature Reviews Neuroscience*, Vol. 14, No. 5, hal. 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1994), "The earth is round (p < .05)", *American Psychologist*, Vol. 49, No. 12, hal. 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Datta, G.S., dan Mukerjee, R. (2004). *Probability matching priors: Higher order asymptotics*, Springer. https://doi.org/10.1007/978-1-4612-2036-7

Delfin, C. (2023), "Improving the stability of bivariate correlations using informative Bayesian priors: A Monte Carlo simulation study", *Frontiers in Psychology*, Vol. 14, Artikel 1253452. https://doi.org/10.3389/fpsyg.2023.1253452

Depaoli, S., dan van de Schoot, R. (2017), "Improving transparency and replication in Bayesian statistics: The WAMBS-checklist", *Psychological Methods*, Vol. 22, No. 2, hal. 240–261. https://doi.org/10.1037/met0000065

Depaoli, S., Winter, S.D., dan Visser, M. (2020), "The importance of prior sensitivity analysis in Bayesian statistics", *Frontiers in Psychology*, Vol. 11, Artikel 608045. https://doi.org/10.3389/fpsyg.2020.608045

DerSimonian, R., dan Laird, N. (1986), "Meta-analysis in clinical trials", *Controlled Clinical Trials*, Vol. 7, No. 3, hal. 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Divine, G., Norton, H.J., Barón, A.E., dan Juarez-Colunga, E. (2018), "The Wilcoxon–Mann–Whitney procedure fails as a test of medians", *The American Statistician*, Vol. 72, No. 3, hal. 278–286. https://doi.org/10.1080/00031305.2017.1305291

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models", *Bayesian Analysis*, Vol. 1, No. 3, hal. 515–533. https://doi.org/10.1214/06-BA117A

Gigerenzer, G. (2004), "Mindless statistics", *Journal of Socio-Economics*, Vol. 33, No. 5, hal. 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gignac, G.E., dan Szodorai, E.T. (2016), "Effect size guidelines for individual differences researchers", *Personality and Individual Differences*, Vol. 102, hal. 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Hedges, L.V. (1981), "Distribution theory for Glass's estimator of effect size and related estimators", *Journal of Educational Statistics*, Vol. 6, No. 2, hal. 107–128. https://doi.org/10.3102/10769986006002107

Hedges, L.V., dan Olkin, I. (1985). *Statistical methods for meta-analysis*, Academic Press.

Higgins, J.P.T., dan Thompson, S.G. (2002), "Quantifying heterogeneity in a meta-analysis", *Statistics in Medicine*, Vol. 21, No. 11, hal. 1539–1558. https://doi.org/10.1002/sim.1186

IntHout, J., Ioannidis, J.P.A., dan Borm, G.F. (2014), "The Hartung–Knapp–Sidik–Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian–Laird method", *BMC Medical Research Methodology*, Vol. 14, Artikel 25. https://doi.org/10.1186/1471-2288-14-25

Ioannidis, J.P.A. (2005), "Why most published research findings are false", *PLOS Medicine*, Vol. 2, No. 8, Artikel e124. https://doi.org/10.1371/journal.pmed.0020124

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Kelter, R. (2021), "Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests", *WIREs Computational Statistics*, Vol. 13, No. 6, Artikel e1523. https://doi.org/10.1002/wics.1523

Kruschke, J.K. (2013), "Bayesian estimation supersedes the t test", *Journal of Experimental Psychology: General*, Vol. 142, No. 2, hal. 573–603. https://doi.org/10.1037/a0029146

Kruschke, J.K. (2021), "Bayesian analysis reporting guidelines", *Nature Human Behaviour*, Vol. 5, No. 10, hal. 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Lakens, D. (2022), "Sample size justification", *Collabra: Psychology*, Vol. 8, No. 1, Artikel 33267. https://doi.org/10.1525/collabra.33267

Marszalek, J.M., Barber, C., Kohlhart, J., dan Holmes, C.B. (2011), "Sample size in psychological research over the past 30 years", *Perceptual and Motor Skills*, Vol. 112, No. 2, hal. 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

McNeish, D. (2016), "On using Bayesian methods to address small sample problems", *Structural Equation Modeling*, Vol. 23, No. 5, hal. 750–773. https://doi.org/10.1080/10705511.2016.1186549

Micceri, T. (1989), "The unicorn, the normal curve, and other improbable creatures", *Psychological Bulletin*, Vol. 105, No. 1, hal. 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Morris, T.P., White, I.R., dan Crowther, M.J. (2019), "Using simulation studies to evaluate statistical methods", *Statistics in Medicine*, Vol. 38, No. 11, hal. 2074–2102. https://doi.org/10.1002/sim.8086

Nakagawa, S., Lagisz, M., Yang, Y., dan Drobniak, S.M. (2024), "Finding the right power balance: Better study design and collaboration can reduce dependence on statistical power", *PLOS Biology*, Vol. 22, No. 1, Artikel e3002423. https://doi.org/10.1371/journal.pbio.3002423

Open Science Collaboration (2015), "Estimating the reproducibility of psychological science", *Science*, Vol. 349, No. 6251, Artikel aac4716. https://doi.org/10.1126/science.aac4716

Richard, F.D., Bond, C.F., Jr., dan Stokes-Zoota, J.J. (2003), "One hundred years of social psychology quantitatively described", *Review of General Psychology*, Vol. 7, No. 4, hal. 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rousseau, J. (2016), "On the frequentist properties of Bayesian nonparametric methods", *Annual Review of Statistics and Its Application*, Vol. 3, hal. 211–231. https://doi.org/10.1146/annurev-statistics-041715-033523

Schäfer, T., dan Schwarz, M.A. (2019), "The meaningfulness of effect sizes in psychological research", *Frontiers in Psychology*, Vol. 10, Artikel 813. https://doi.org/10.3389/fpsyg.2019.00813

Sigal, M.J., dan Chalmers, R.P. (2016), "Play it again: Teaching statistics with Monte Carlo simulation", *Journal of Statistics Education*, Vol. 24, No. 3, hal. 136–156. https://doi.org/10.1080/10691898.2016.1246953

Simmons, J.P., Nelson, L.D., dan Simonsohn, U. (2011), "False-positive psychology", *Psychological Science*, Vol. 22, No. 11, hal. 1359–1366. https://doi.org/10.1177/0956797611417632

Smid, S.C., McNeish, D., Miočević, M., dan van de Schoot, R. (2020), "Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review", *Structural Equation Modeling*, Vol. 27, No. 1, hal. 131–161. https://doi.org/10.1080/10705511.2019.1577140

Szucs, D., dan Ioannidis, J.P.A. (2017), "Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature", *PLOS Biology*, Vol. 15, No. 3, Artikel e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tendeiro, J.N., dan Kiers, H.A.L. (2019), "A review of issues about null hypothesis Bayesian testing", *Psychological Methods*, Vol. 24, No. 6, hal. 774–795. https://doi.org/10.1037/met0000221

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., dan Yau, C. (2021), "Bayesian statistics and modelling", *Nature Reviews Methods Primers*, Vol. 1, Artikel 1 https://doi.org/10.1038/s43586-020-00001-2

Vankelecom, L., Schacht, O., Laroy, N., Loeys, T., dan Moerkerke, B. (2025), "A systematic review on the evolution of power analysis practices in psychological research", *Psychologica Belgica*, Vol. 65, No. 1, hal. 17–37. https://doi.org/10.5334/pb.1318

Wasserstein, R.L., dan Lazar, N.A. (2016), "The ASA's statement on p-values: Context, process, and purpose", *The American Statistician*, Vol. 70, No. 2, hal. 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R.L., Schirm, A.L., dan Lazar, N.A. (2019), "Moving to a world beyond 'p < 0.05'", *The American Statistician*, Vol. 73, Suppl. 1, hal. 1–19. https://doi.org/10.1080/00031305.2019.1583913

.

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/507