QUBAHAN JOURNAL OF SOCIAL SCIENCE
VOL. 3, NO. 2, June 2026
https://doi.org/10.48161/qjss.v3n2a15

Classical and Bayesian Parameter Estimation under Varying Sample Sizes and Data Conditions

Rakhmonova Robiya[1*]

[1] Samarkand Branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, samarkand, Uzbekistan;

* **Corresponding author:** raxmonovarobiya494@gmail.com.

ABSTRACT: Parameter estimation is central to empirical statistical analysis, yet researchers often face uncertainty about whether classical or Bayesian inference offers the most reliable interpretation of unknown model parameters. This study comparatively evaluates classical and Bayesian parameter estimation under controlled simulation settings and two applied empirical cases. Classical inference is represented by maximum likelihood logistic regression, ordinary least squares regression, standard errors, and confidence intervals. Bayesian inference uses weakly informative normal priors, posterior mode estimation, approximate posterior covariance, and credible intervals. The simulation component evaluates mean absolute bias, mean squared error, interval width, and coverage across small, medium, and large sample sizes under clean and contaminated data conditions. The main empirical application uses a credit-risk binary outcome with 30,000 observations, while a continuous insurance-cost application provides a robustness case for linear parameter estimation. The simulation results show that the two approaches converge closely in large samples when the model is correctly specified, while weak Bayesian regularization modestly stabilizes estimates in small samples. Under label contamination, both approaches lose interval coverage, indicating that robust data-generating assumptions are as important as the inferential paradigm. In the credit-risk case, classical and Bayesian logistic estimates produced nearly identical out-of-sample performance, with an area under the receiver operating characteristic curve (ROC-AUC) of approximately 0.716 and accuracy of approximately 0.809. The insurance-cost robustness case produced similar predictive accuracy for classical and Bayesian linear estimation. The findings support a practical interpretation: Bayesian methods add uncertainty representation and regularization, while classical methods remain efficient when sample size is large, and assumptions are reasonable.

Keywords: parameter estimation; classical inference; Bayesian inference; maximum likelihood estimation; credible intervals.

I. INTRODUCTION

Parameter estimation is one of the most important operations in quantitative research because it converts observed data into interpretable numerical evidence about unknown quantities. In applied studies, these quantities may represent the mean effect of an intervention, the probability of failure in a technical system, the association between a predictor and a binary outcome, the expected response of a continuous variable, or the uncertainty attached to a model parameter. The quality of the estimate determines how confidently a researcher can interpret empirical findings, compare competing models, and support practical decisions. For this reason, the choice of inferential method is not purely technical; it shapes the estimate's meaning, the language of uncertainty, and how results are communicated to readers.

Classical inference has traditionally dominated empirical research because it provides familiar procedures such as maximum likelihood estimation, ordinary least squares regression, standard errors, hypothesis tests, and confidence intervals. These procedures are computationally efficient, widely available,

and easy to report in journal articles. They are especially attractive when sample sizes are large, models are well specified, and repeated-sampling assumptions are reasonable. In such settings, classical estimators often have strong asymptotic properties and can provide accurate point estimates. However, people often misunderstand their uncertainty statements. People often interpret a confidence interval as the probability that the parameter lies within the observed interval, although its formal meaning is based on the long-run behavior of repeated sampling procedures.

Bayesian inference approaches the same problem differently. It combines a likelihood function with a prior distribution to produce a posterior distribution for the unknown parameter. This posterior distribution gives a direct probability-based description of uncertainty after observing the data. The Bayesian approach is attractive because it can incorporate prior knowledge, stabilize estimation in small samples, generate credible intervals, and express uncertainty in a way that many readers find more intuitive. At the same time, Bayesian analysis requires careful prior specification, model checking, computational diagnostics, and sensitivity analysis. If these steps are weakly justified, Bayesian estimates may appear more informative than the evidence can support.

The comparison between classical and Bayesian inference is therefore not a simple question of which method is universally superior. The better question is how each approach behaves under different empirical conditions. Small samples, noisy outcomes, weak predictors, rare events, correlated variables, and model misspecification can influence the accuracy and interpretability of parameter estimates. In large datasets with stable information, classical and Bayesian estimates may become very similar, especially when weak priors are used. In smaller or unstable datasets, Bayesian regularization may improve stability, but only when the prior is reasonable and transparently reported. A strong comparative study must therefore examine both controlled conditions, where true parameter values are known, and applied conditions, where interpretability and performance are evaluated in real data.

This study addresses that need by combining simulation experiments with empirical applications. The simulation component evaluates whether classical and Bayesian estimation differ in bias, mean squared error, interval width, and coverage across multiple sample sizes and data-quality conditions. The main empirical application compares classical logistic regression and Bayesian logistic estimation in a binary credit-risk setting. A secondary continuous-outcome application compares ordinary least squares with Bayesian linear estimation. Together, these components provide a structured basis for understanding when the two approaches converge, when Bayesian regularization adds value, and when data quality, rather than the inferential framework alone, limits both methods.

The paper is organized as follows. The next section reviews studies that examine classical uncertainty interpretation, Bayesian estimation, prior sensitivity, computational workflow, and comparative inference. This organization separates the reviewed studies from the research gap and proposed method, keeping the argument clear and journal-ready. The paper then states the research gap and research questions separately. The proposed method and methodology explain the simulation design, empirical datasets, estimation models, evaluation indicators, and validity safeguards. The results section reports the dataset audit, simulation outcomes, credit-risk application, and robustness application. The discussion interprets the findings for researchers choosing between classical and Bayesian estimation in empirical studies. The paper ends with theoretical contributions, practical implications, limitations, future research directions, and a concise conclusion.

II. LITERATURE REVIEW

The literature shows that modern Bayesian regression has become more accessible through flexible regression packages and extended modeling interfaces. These contributions made Bayesian multilevel modeling easier to apply in empirical research [1]. Advanced regression interfaces then expanded Bayesian modeling to nonlinear, multilevel, and distributional settings [2].Probabilistic programming also strengthened reproducible model specification, posterior computation, and transparent statistical workflow [3].

In parallel, transparency-focused guidance clarified that Bayesian estimation should include prior justification, sensitivity checking, and replication-oriented reporting rather than only posterior summaries [4]. A second set of studies positioned Bayesian inference as an updating framework that supports modern

parameter estimation and practical evidence interpretation [5]. Applied tutorial work further explained Bayesian probability, parameter estimation, and model comparison in a way that made Bayesian reasoning more accessible to empirical researchers [6].

Visualization-based workflow studies showed that prior checking, posterior checking, and model refinement improve the credibility of Bayesian analysis [7]. Methodological work on statistical power also showed that classical estimates can become unstable with small samples or noisy effects [8].Predictive model-evaluation research contributed to the comparison by showing how predictive criteria can assess Bayesian model quality beyond point estimates alone [9].

At the same time, broader critiques of test interpretation show that people often misunderstand p-values, confidence intervals, and power when used mechanically [10].The literature also emphasizes that interval interpretation remains a major source of confusion in empirical research. Comparative explanations of confidence and credible intervals show that the two intervals answer different probability questions [11]. Studies of confidence-interval misunderstanding further show that readers often assign direct probability meanings to repeated-sampling intervals [12].

Applied alternatives to null-hypothesis testing support the need to compare estimation frameworks through parameter uncertainty and practical interpretation rather than through significance thresholds alone [13]. Foundational Bayesian estimation studies also show how posterior distributions can represent parameter uncertainty directly and interpretably [14]. Additional Bayesian foundations connect estimation, credible intervals, and applied model interpretation for researchers who are new to Bayesian analysis [15].

Related work links Bayesian estimation with hypothesis testing, meta-analysis, and power analysis, which supports the present study's focus on both numerical accuracy and uncertainty language [16]. Methodological discussions have also emphasized the limits of p-value-centered inference. These studies argue that reported significance is only one component of valid statistical analysis [17]. Work on weakly informative priors further shows that prior information can regularize unstable estimates without dominating the likelihood [18].

Bayesian effect-index research contributes to this discussion by showing how a probability-based framework can communicate existence, direction, and practical importance [19]. Small-sample Bayesian guidance also supports the expectation that Bayesian estimation may be useful when data are limited [20]. Critiques of dichotomous significance strengthen the argument for evaluating estimation frameworks through uncertainty, interpretation, and practical meaning rather than statistical thresholds alone [21].

Explanations of confidence-interval fallacies further justify comparing confidence and credible intervals in terms of coverage, width, and interpretation [22]. Shrinkage-prior research supports the expectation that Bayesian regularization may be most useful when coefficients are weakly identified or unstable [23]. More recent workflow studies also emphasize prior predictive checking, sensitivity analysis, and posterior predictive evaluation as essential parts of principled Bayesian modeling [24].

Bayesian modeling primers extend this argument by presenting the full analysis cycle from prior choice to model checking and interpretation [25]. Earlier applied introductions also clarified the roles of priors, likelihoods, posterior distributions, and credible intervals for empirical researchers [26].Predictive model-evaluation studies further show how Bayesian models can be assessed through efficient cross-validation and information criteria [27].

Large tutorial contributions show how Bayesian methods can be implemented in accessible software while remaining clearly distinct from classical routines [28]. Example-based Bayesian studies show that applied researchers can compare Bayesian and classical routines using practical software environments [29]. Statistical-association guidance also warns against treating p-values as mechanical evidence, which supports the present manuscript's balanced comparison of inferential frameworks [30]. Post-threshold inference discussions argue for scientific interpretation based on uncertainty, context, and transparency rather than fixed significance cutoffs [31]. Predictive distribution averaging also supports the manuscript's broader design by showing that Bayesian comparison should consider prediction, uncertainty, and model adequacy together [32].

Table 1. Literature-review studies included in the evidence synthesis.

Reference	Authors	Focus	Main relevance to this manuscript
[1]	Bürkner (2017)	Bayesian regression modeling	Made Bayesian regression models more accessible across distributions and link functions.
[2]	Bürkner(2018)	Advanced Bayesian regression	Expanded Bayesian regression to multilevel, nonlinear, and distributional modeling contexts.
[3]	Carpenter et al. (2017)	Probabilistic computation	Presented a general language for Bayesian model specification and posterior computation.
[4]	Depaoli and van de Schoot(2017)	Bayesian transparency checklist	Developed diagnostic and reporting guidance for transparent Bayesian applications.
[5]	Efron(2013)	Modern Bayesian reasoning	Positioned Bayesian updating as a central route for twenty-first-century statistical inference.
[6]	Etz and Vandekerckhove(2018)	Bayesian inference tutorial	Explained core Bayesian probability, estimation, and model comparison principles.
[7]	Gabry et al. (2019)	Bayesian workflow visualization	Showed how visualization supports prior checking, posterior checking, and model refinement.
[8]	Gelman and Carlin(2014)	Power, sign, and magnitude errors	Showed that classical estimates can be unstable when samples are weak or effects are noisy.
[9]	Gelman et al. (2014)	Bayesian predictive criteria	Explained predictive information criteria for Bayesian model comparison and estimation quality.
[10]	Greenland et al. (2016)	Misinterpretations of tests and intervals	Reviewed common errors in interpreting p-values, power, and confidence intervals.
[11]	Hespanhol et al. (2019)	Interval interpretation	Compared confidence and credible intervals for interpreting effect estimates in applied settings.
[12]	Hoekstra et al. (2014)	Confidence-interval interpretation	Demonstrated persistent misunderstanding of confidence intervals among researchers and students.
[13]	Kelter(2020)	Bayesian alternatives in biomedical research	Compared traditional tests and Bayesian alternatives for applied biomedical inference.
[14]	Kruschke(2013)	Bayesian estimation for group comparison	Showed how posterior distributions give direct estimates and uncertainty intervals.
[15]	Kruschke and Liddell(2018)	Bayesian data analysis foundations	Clarified Bayesian estimation, credible intervals, and applied interpretation for newcomers.
[16]	Kruschke and Liddell(2018)	Bayesian new statistics	Linked estimation, hypothesis testing, power, and meta-analysis to a Bayesian perspective.
[17]	Leek and Peng(2015)	Limits of p-value-centered inference	Emphasized that reported significance is only one part of valid statistical analysis.
[18]	Lemoine(2019)	Weakly informative priors	Explained why weak priors can improve regularization without dominating empirical evidence.
[19]	Makowski et al. (2019)	Bayesian effect indices	Provided Bayesian measures for existence, significance, and practical importance of effects.
[20]	McNeish(2016)	Bayesian small-sample estimation	Discussed the usefulness of Bayesian approaches for small-sample modeling problems.
[21]	McShane et al. (2019)	Beyond statistical significance	Argued for estimation and uncertainty interpretation beyond dichotomous significance thresholds.
[22]	Morey et al. (2016)	Confidence versus credible intervals	Explained why confidence intervals do not support probability statements about a realized parameter.
[23]	Piironen and Vehtari(2017)	Shrinkage priors	Showed how regularized shrinkage priors can stabilize weakly identified regression coefficients.
[24]	Schad et al. (2021)	Principled Bayesian workflow	Outlined prior predictive checks, sensitivity analysis, and posterior predictive evaluation.
[25]	van de Schoot et al. (2014)	Bayesian introduction and application	Clarified the prior, likelihood, posterior, and interpretation, and applied them for empirical researchers.
[26]	van de Schoot et al. (2021)	Bayesian statistics and modeling primer	Synthesized the full Bayesian analysis cycle from prior choice to model checking.
[27]	Vehtari et al. (2017)	Bayesian model evaluation	Established efficient leave-one-out and WAIC procedures for model assessment.
[28]	Wagenmakers et al. (2018)	Bayesian advantages	Reviewed theoretical advantages of Bayesian inference for evidence accumulation and estimation.

Reference	Authors	Focus	Main relevance to this manuscript
[29]	Wagenmakers et al. (2018)	Bayesian examples	Provided applied examples that compare Bayesian and classical routines in accessible software.
[30]	Wasserstein and Lazar(2016)	ASA statement on p-values	Provided formal guidance against mechanical inference based only on p-values.
[31]	Wasserstein et al. (2019)	Beyond $p < 0.05$	Called for scientific inference based on uncertainty, context, and transparency rather than thresholds.
[32]	Yao et al. (2018)	Predictive distribution averaging	Improved Bayesian predictive modeling by stacking posterior predictive distributions.

III. RESEARCH GAP

Existing studies provide strong conceptual explanations of classical and Bayesian inference, but the literature still lacks a compact empirical comparison that combines controlled simulation with real applied parameter estimation. Many papers explain why confidence intervals and credible intervals differ, or why Bayesian methods can be useful in small samples, yet few studies place these ideas into a single reproducible framework that evaluates bias, error, interval width, coverage, and applied predictive behavior under the same design.

A second gap concerns the connection between uncertainty interpretation and practical model selection. Researchers often ask whether Bayesian estimation outperforms classical estimation, but the answer depends on sample size, contamination, model specification, and prior strength. The present study addresses this gap by comparing both approaches under controlled sample sizes and in applied binary and continuous regression settings. The contribution is intentionally conservative: it does not claim universal superiority for either paradigm, but provides evidence-linked guidance for choosing and interpreting estimation methods in empirical research.

IV. RESEARCH AIM AND QUESTIONS

The aim of this study is to compare classical and Bayesian inference methods for parameter estimation under varying sample sizes, data-quality conditions, and empirical regression contexts. The study evaluates whether the two approaches differ in accuracy, uncertainty representation, interpretability, and robustness.

- RQ1. How do classical maximum likelihood estimates and Bayesian posterior estimates compare in terms of bias and mean squared error under different sample sizes?
- RQ2. How do confidence intervals and credible intervals compare in terms of interval width and coverage when the true parameter values are known?
- RQ3. Does weak Bayesian regularization provide meaningful stabilization when data are small or noisy?
- RQ4. Do classical and Bayesian logistic regression produce materially different parameter estimates and predictive performance in a large binary-outcome application?
- RQ5. Are the observed differences consistent in a continuous-outcome robustness application?

V. PROPOSED METHOD

The proposed method is a comparative empirical framework that combines simulation evidence with applied regression evidence. The simulation component is used because it provides known true parameter values, allowing calculation of bias, mean squared error, interval width, and coverage. The applied component is used because real empirical data reveal whether the methods remain interpretable and practically similar when the true parameters are unknown.

The framework follows five stages. First, we audit the available datasets for size, missing values, duplicates, outcome type, and suitability for comparing parameter estimates. Second, we simulate a controlled logistic data-generating process under three sample sizes and two data-quality conditions. Third, we apply classical inference using maximum likelihood logistic regression and ordinary least squares regression. Fourth, we apply Bayesian inference using weakly informative normal priors, posterior mode estimation, and approximate credible intervals. Fifth, the methods are evaluated using statistical accuracy, uncertainty intervals, model performance, and interpretability. Figure 1 summarizes the framework.

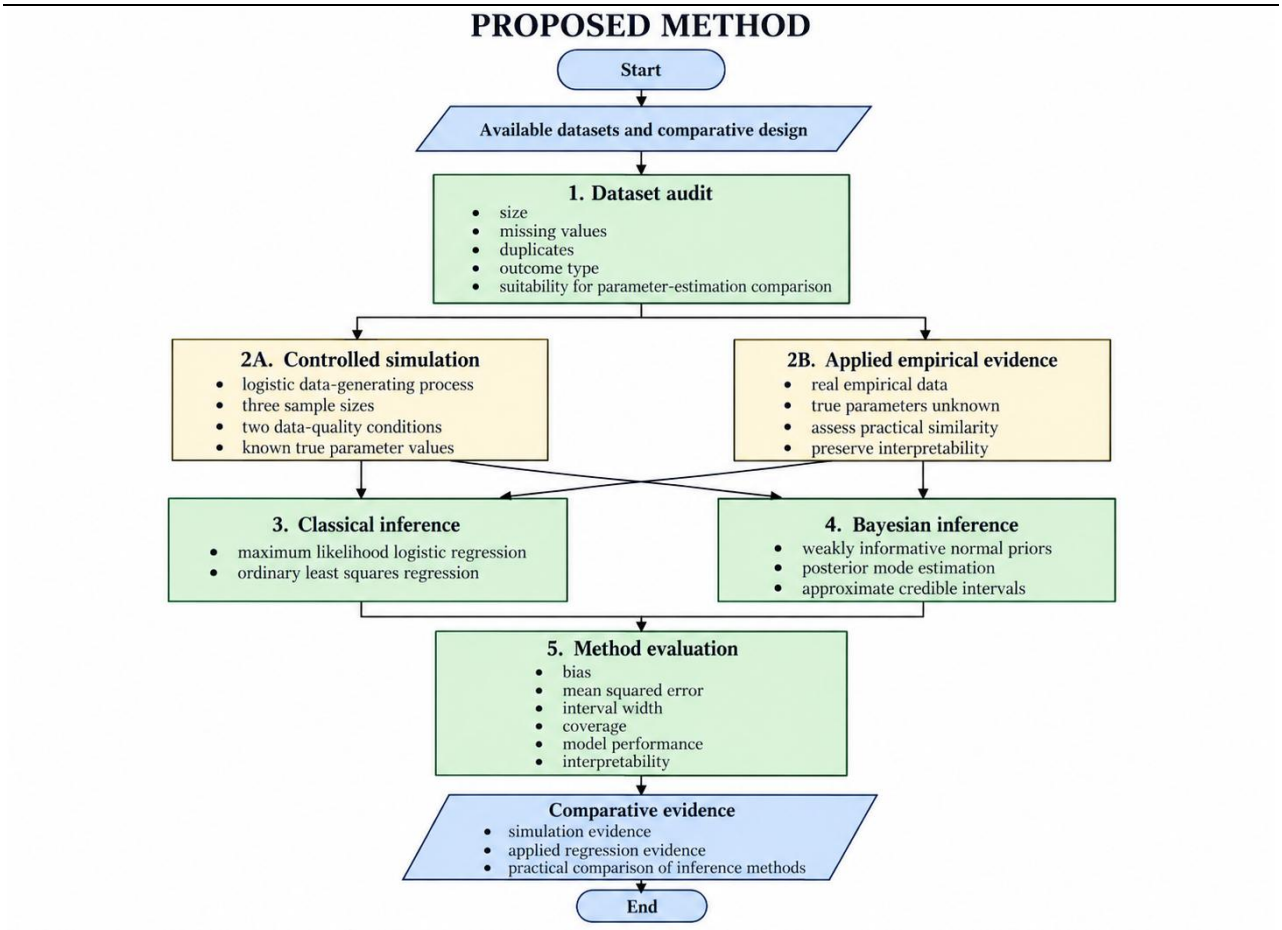


FIGURE 1. Proposed comparative framework for parameter-estimation inference.

The framework combines dataset screening, controlled simulation, classical estimation, Bayesian estimation, and multi-criteria evaluation within one conservative empirical design.

VI. STUDY CONTRIBUTIONS

The study contributes a condition-based interpretation of classical and Bayesian parameter estimation. Rather than treating the two approaches as competing ideologies, it shows how their differences depend on sample size, prior strength, data quality, and model specification. This contribution matters because it shifts the comparison from abstract preference to evidence-based methodological selection.

The study also contributes an integrated evaluation structure that combines simulation and applied regression. The simulation clarifies accuracy and coverage when the true values are known, while the empirical cases test whether differences remain meaningful in applied settings. This dual structure can be reused in other methodological comparisons involving linear, logistic, survival, or hierarchical models.

VII. METHODOLOGY

1. RESEARCH DESIGN

The study uses a comparative quantitative design. It is not designed to prove that one inferential tradition is universally superior. Instead, it evaluates how classical and Bayesian estimation behave under clearly defined statistical conditions. The design includes a simulation experiment, a binary-outcome empirical case, and a continuous-outcome robustness case.

2. DATA SOURCES AND SAMPLE SELECTION

Four public datasets were screened. We selected the credit-risk dataset as the main empirical case because it contains 30,000 observations, a binary default outcome, no missing values, and a sufficiently large predictor set for logistic parameter estimation. We retained the insurance-cost dataset as a robustness case because it contains a continuous outcome suitable for linear parameter estimation. The diabetes dataset was kept as a possible future robustness case but was not used as a main case because several medical variables require zero-value screening. We excluded the student-performance dataset from the main analysis because duplicate rows and a simple predictor structure make it less suitable for a strong comparative inference manuscript.

3. SIMULATION DESIGN

The simulation used a logistic data-generating process with three standardized predictors and known true coefficients. We set sample sizes at n = 50, n = 200, and n = 1000 to represent small, medium, and large empirical studies. We used two data-quality conditions: a clean condition and a contaminated condition in which we randomly flipped 10% of binary labels. Each condition was replicated 120 times. This design allows direct comparison of bias, mean squared error, interval width, and parameter coverage.

4. CLASSICAL ESTIMATION PROCEDURE

Classical inference was represented by maximum likelihood estimation for logistic regression and ordinary least squares for the continuous robustness case. We summarized parameter uncertainty using standard errors and 95% confidence intervals. For the empirical logistic model, we used standardized numeric predictors and one-hot encoded categorical predictors. Model performance was assessed on a 70/30 train-test split.

5. BAYESIAN ESTIMATION PROCEDURE

Bayesian inference was represented by weakly informative normal priors for regression coefficients. For logistic regression, we obtained posterior mode estimates with ridge-equivalent regularization and derived approximate posterior covariance from the penalized Hessian. For the linear robustness case, we obtained posterior means under a normal prior and a Gaussian likelihood approximation. Credible intervals were calculated from the posterior covariance. This approach is computationally conservative and transparent, but it should be interpreted as approximate Bayesian estimation rather than full Markov-chain posterior simulation.

6. EVALUATION CRITERIA

The simulation compared mean absolute bias, mean squared error, average interval width, and empirical coverage. The credit-risk application compared AUC, accuracy, Brier score, log loss, and selected coefficient intervals. The insurance-cost robustness case compared root mean squared error, mean absolute error, and R-squared. Interpretation focused on whether differences were materially meaningful, not only numerically visible.

7. VALIDITY, RELIABILITY, AND ETHICS

Internal validity was strengthened by using known true parameters in the simulation and by applying the same preprocessing logic to classical and Bayesian models. We supported reliability through repeated simulation replications and fixed random seeds. External validity was supported by using both binary and continuous empirical cases. Ethical risk is low because the analysis uses public, anonymized secondary data without direct human contact or identifiable participant information.

8. STATISTICAL MODEL SPECIFICATION

Equation (1). Maximum likelihood estimation: $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} L(\theta \mid y)$

Equation (2). Approximate 95% confidence interval: $CI_{0.95} = \hat{\theta} \pm 1.96\, SE(\hat{\theta})$

Equation (3). Bayesian posterior distribution: $p(\theta \mid y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\, d\theta}$

Equation (4). Logistic regression model: $\mathrm{logit}\{P(Y_i = 1)\} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$

Equation (5). Weakly informative coefficient prior: $\beta_j \sim \mathcal{N}(0, 2.5^2)$

Equation (6). Mean squared error: $MSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2$

Equations (1)-(6) define the classical likelihood estimate, confidence interval, Bayesian posterior, logistic regression form, weak coefficient prior, and simulation error criterion used in the comparison.

VIII. RESULTS

1. DATASET AUDIT

The dataset audit confirms that the credit-risk application is the strongest empirical case for this manuscript. It contains 30,000 observations, 25 variables, no missing values, and no duplicates. The diabetes dataset is usable but requires medical zero-value screening before direct modeling. The insurance-cost dataset is suitable as a continuous-outcome robustness case, although it contains one duplicate row. The student-performance dataset is not recommended as main evidence because the duplicate count and simple structure make it weaker for a comparative inference study.

Table 2. Dataset audit and empirical-use decision.

Dataset role	Rows	Cols.	Missing	Dups.	Use decision
Credit-risk application	30000	25	0	0	Main empirical case
Diabetes application	768	9	0	0	Optional robustness; requires zero-value screening
Insurance-cost application	1338	7	0	1	Linear-regression robustness case
Student-performance application	10000	6	0	127	Excluded from main analysis due to duplicates and simple design

The credit-risk dataset was selected as the main empirical case, while diabetes and insurance-cost data were retained for robustness analysis. We excluded the student-performance dataset due to duplication and limited complexity.

2. SIMULATION RESULTS

The simulation results show that classical and Bayesian estimates converge as sample size increases. In the clean n = 50 condition, Bayesian MAP estimation produced mean absolute bias of 0.316 compared with 0.343 for classical MLE. At n = 200 and n = 1000, the differences became negligible. Coverage in the clean condition stayed close to the nominal 95% level for both methods. Under 10% label contamination, both methods experienced reduced coverage, especially at n = 1000, showing that model misspecification and outcome noise can dominate the choice of inference framework.

Table 3. Simulation summary by method, sample size, and contamination level.

Method	n	Noise	Abs. bias	MSE	Interval width	Coverage
Bayesian MAP	50	0%	0.316	0.174	1.549	0.960
Bayesian MAP	50	10%	0.320	0.163	1.409	0.904
Bayesian MAP	200	0%	0.151	0.035	0.711	0.950
Bayesian MAP	200	10%	0.204	0.060	0.644	0.769
Bayesian MAP	1000	0%	0.064	0.006	0.311	0.952
Bayesian MAP	1000	10%	0.175	0.039	0.283	0.373
Classical MLE	50	0%	0.343	0.216	1.611	0.958
Classical MLE	50	10%	0.327	0.173	1.440	0.908
Classical MLE	200	0%	0.153	0.036	0.715	0.950
Classical MLE	200	10%	0.202	0.060	0.646	0.773
Classical MLE	1000	0%	0.064	0.006	0.311	0.952
Classical MLE	1000	10%	0.175	0.038	0.283	0.377

Bayesian and classical estimates became nearly identical as sample size increased. Both methods maintained good coverage under clean data, while 10% contamination substantially reduced coverage, especially at n=1000n=1000.

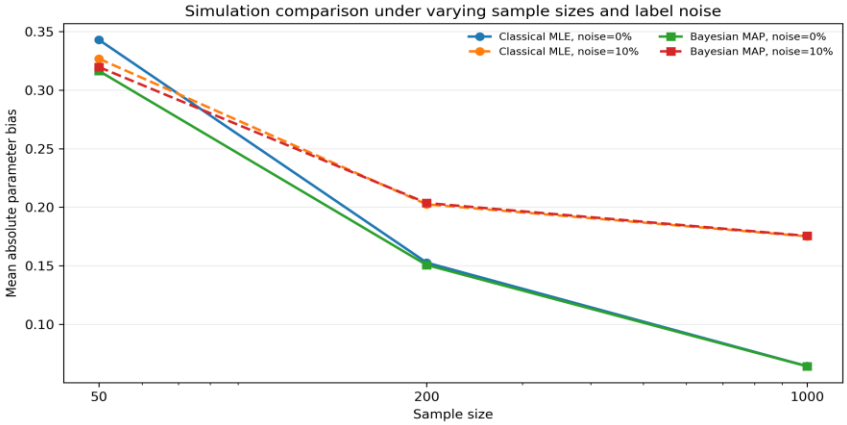


FIGURE 2. Mean absolute parameter bias across sample sizes and noise conditions.

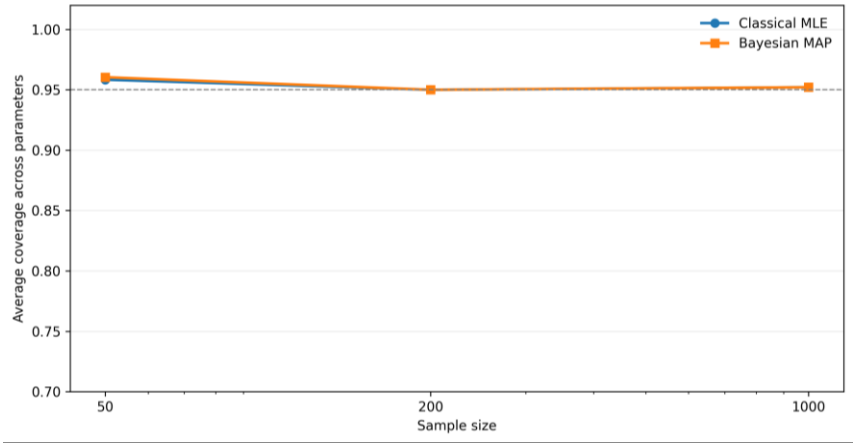


FIGURE 3. Interval coverage across sample sizes in the clean simulation condition.

3. CREDIT-RISK LOGISTIC APPLICATION

The credit-risk application produced almost identical out-of-sample performance for classical logistic MLE and Bayesian logistic MAP estimation. Classical logistic regression achieved an AUC of 0.716, accuracy of 0.809, Brier score of 0.146, and log loss of 0.469. Bayesian logistic regression achieved an AUC of 0.716, accuracy of 0.809, a Brier score of 0.146, and a log loss of 0.469. The nearly identical results are expected because the sample is large and the Bayesian prior is weakly informative. The main value of the Bayesian model is therefore not higher predictive accuracy, but clearer probabilistic uncertainty representation and mild regularization of unstable coefficients.

Table 4. Credit-risk out-of-sample performance comparison.

Model	AUC	Accuracy	Brier	Log loss
Classical logistic MLE	0.716	0.809	0.146	0.469
Bayesian logistic MAP	0.716	0.809	0.146	0.469

FIGURE 4. Credit-risk application: selected coefficient intervals.

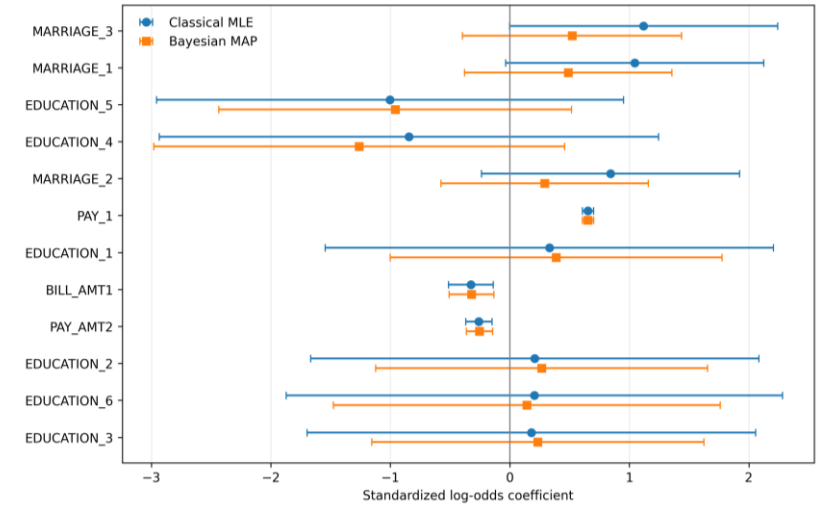




FIGURE 5. Credit-risk application: predictive metric dashboard.

4. INSURANCE-COST ROBUSTNESS APPLICATION

The continuous-outcome robustness case also showed close convergence between the two approaches. Classical OLS produced RMSE of 0.425, MAE of 0.266, and R-squared of 0.784 on the log-transformed outcome. Bayesian linear estimation produced RMSE of 0.425, MAE of 0.266, and R-squared of 0.784. These results reinforce the main finding: when data information is strong and priors are weak, Bayesian and classical estimates can be practically equivalent, with Bayesian estimation adding a different uncertainty language rather than necessarily improving predictive performance.

Table 5. Continuous-outcome robustness performance comparison.

Model	RMSE	MAE	R-squared
Classical OLS	0.425	0.266	0.784
Bayesian linear posterior mean	0.425	0.266	0.784

IX. DISCUSSION

The findings show that comparing classical and Bayesian inference should be framed as a question of empirical conditions rather than a general competition between two statistical traditions. In the simulation, the Bayesian estimator had a modest advantage in the smallest clean sample, where weak regularization reduced instability. However, the advantage decreased as sample size increased. This pattern is theoretically coherent because a weak prior has the greatest influence when the likelihood contains limited information.

As the sample grows, the likelihood dominates the prior, and the two approaches converge to similar numerical estimates.

The most important result is that uncertainty representation differs even when point estimates are nearly the same. Classical confidence intervals and Bayesian credible intervals may look similar in large samples, but they answer different interpretive questions. Confidence intervals describe the long-run performance of a repeated sampling procedure, whereas credible intervals summarize posterior uncertainty about the parameter after combining the data with the prior. This difference matters for manuscript writing because researchers often use probability language when discussing confidence intervals. The present study supports a more careful interpretation: if the research aim is to make direct probability statements about parameters, a Bayesian framework is more natural; if the research aim is to rely on repeated-sampling properties with minimal prior specification, a classical framework remains appropriate.

The contamination condition provides a second important lesson. Both methods lost coverage when binary outcomes were randomly flipped. This result shows that Bayesian modeling does not automatically solve data-quality problems. A prior can stabilize estimation, but it cannot fully correct a misspecified likelihood or a noisy outcome mechanism unless the model explicitly accounts for the contamination process. Likewise, classical estimation can be efficient under correct assumptions but will also be affected by misspecification. Therefore, robust modeling, data screening, and sensitivity analysis should be essential parts of the estimation workflow, regardless of the inferential framework.

The credit-risk application confirms the expected convergence between classical and Bayesian estimates in a large sample. The two logistic models produced nearly identical AUC, accuracy, Brier score, and log loss. This should not be interpreted as a weakness of Bayesian inference. Instead, it shows that weakly informative Bayesian regression behaves conservatively when the data are highly informative. The selected coefficient intervals also show where Bayesian regularization has practical value: rare categorical levels and unstable predictors receive more moderate estimates, while strong repayment-history predictors remain almost unchanged. This pattern is useful for applied researchers who need stable estimates without allowing priors to dominate the evidence.

The insurance-cost robustness case adds support from a continuous regression context. Again, the classical and Bayesian estimates were practically equivalent. The result suggests that when a regression model is well informed and the prior is weak, Bayesian estimation mainly changes how uncertainty is interpreted and reported rather than the prediction itself. This is useful for researchers who believe that Bayesian inference must always produce different results. In many applied settings, similarity is a desirable sign because it indicates that the conclusions are not highly sensitive to the inferential paradigm.

Another implication concerns reporting quality. A classical analysis should report the estimator, standard-error method, interval construction, sample size, model assumptions, and diagnostic evidence. A Bayesian analysis should report the likelihood, prior distributions, posterior summaries, sensitivity checks, and interpretation of credible intervals. Without this information, the reader cannot judge whether the method is appropriate. The present results show that numerical similarity between methods does not remove the need for transparent reporting, because the logic behind the uncertainty statement remains different.

The findings also have consequences for teaching and manuscript review. Students and applied researchers often learn classical inference first and may treat Bayesian inference as an advanced or optional extension. The evidence supports a different teaching sequence: introduce both frameworks as alternative ways to connect data, assumptions, and uncertainty. Reviewers can then evaluate whether the chosen framework matches the research aim. A manuscript that uses classical inference should not overstate interval probabilities, while a manuscript that uses Bayesian inference should not hide prior choices or computational approximations. This balanced standard improves interpretation under both traditions.

A further practical implication is that model comparison should not be limited to predictive scores. Prediction metrics help, but parameter-estimation studies should also examine coefficient stability, interval behavior, sensitivity to prior scale, and the meaning of the reported uncertainty. In the present analysis, AUC and RMSE changed very little across frameworks, yet the inferential narratives were not identical. The Bayesian analysis supplied posterior-style interval summaries and regularized unstable effects, while the classical analysis supplied familiar sampling-based intervals and efficient large-sample estimates. For journal reviewers, this means that equivalence in predictive performance should be discussed rather than ignored.

Overall, the study suggests a balanced decision rule. Classical inference suits studies with adequate sample sizes, defensible assumptions, and an audience that expects standard errors and confidence intervals. Bayesian inference is especially useful when the research requires direct probability statements, prior knowledge, hierarchical modeling, small-sample stabilization, or sensitivity analysis. Neither approach should be used mechanically. The best practice is to align the inference framework with the research question, data quality, model assumptions, and reporting needs. This conservative interpretation makes the manuscript more useful for empirical researchers than a one-sided claim that one method is universally superior. It also supports transparent method selection across disciplines with different statistical traditions.

X. PRACTICAL IMPLICATIONS

Researchers working with large, stable datasets can confidently use classical estimation when the model is well specified, and standard confidence-interval reporting suffices for the journal audience. However, they should avoid probability statements about realized confidence intervals and should report uncertainty with careful language.

Researchers working with small samples, weakly identified coefficients, rare categories, or prior scientific knowledge should consider Bayesian estimation. Weakly informative priors can stabilize estimates without dominating the likelihood, but they must be justified and checked. Applied manuscripts should report the prior, likelihood, posterior summary, interval interpretation, and sensitivity checks in transparent language.

For empirical studies, the most practical recommendation is to align the inference method with the research question. If the question asks for repeated-sampling performance, classical methods are suitable. If the question asks for probability statements about parameters after observing the data, Bayesian methods are more directly aligned with the interpretation.

XI. LIMITATIONS AND FUTURE RESEARCH

The study is limited by the use of approximate Bayesian estimation rather than full posterior simulation. Although posterior-mode and covariance approximations are appropriate for a conservative large-sample comparison, future work should use full Markov-chain posterior sampling, prior predictive checking, posterior predictive checking, and formal convergence diagnostics.

The simulation uses a controlled logistic design with label contamination, but real data can contain more complex forms of misspecification, including nonlinear effects, missingness, measurement error, correlated outcomes, hierarchical structure, and unobserved heterogeneity. Future studies should extend the framework to hierarchical Bayesian models, robust likelihoods, mixture models, and stronger contamination mechanisms.

The empirical analysis focuses on one main binary-outcome case and one continuous robustness case. Future work should test the same framework in medical, environmental, engineering, and social-science datasets where domain-specific priors can be elicited and validated.

XII. CONCLUSION

This study compared classical and Bayesian inference methods for parameter estimation using simulation evidence and two empirical regression applications. The findings show that both approaches can produce highly similar estimates when the sample is large, the likelihood is informative, and Bayesian priors are weakly specified. In the simulation, Bayesian regularization modestly stabilized the smallest clean sample, but the difference from classical MLE shrank as sample size increased. Under label contamination, both approaches lost interval coverage, demonstrating that data quality and model specification remain central to reliable inference. The credit-risk application showed almost identical predictive performance for classical and Bayesian logistic estimation, while the insurance-cost robustness case showed similar convergence in a continuous-outcome setting. The main implication is that Bayesian inference should not be valued only when it changes point estimates. Its contribution lies in representing probabilistic uncertainty, prior transparency, regularization, and a flexible workflow. Classical inference remains efficient, familiar, and appropriate when repeated-sampling assumptions are defensible. The study therefore recommends a condition-based choice: use classical estimation for large, stable, well-specified analyses where standard reporting is sufficient, and

use Bayesian estimation when uncertainty interpretation, prior knowledge, or small-sample stabilization is central to the research question. Future studies should extend this comparison with full posterior simulation, hierarchical models, stronger contamination mechanisms, and domain-specific prior elicitation. The overall conclusion is balanced: the methods are complementary tools, and the best inferential choice depends on the evidence structure rather than on methodological preference alone. For journal submission, the manuscript should therefore present the comparison as methodological guidance, not as a claim that one paradigm replaces the other in all empirical contexts. This framing strengthens the paper by connecting statistical theory, simulation evidence, and applied modeling in a coherent research design.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1-28. https://doi.org/10.18637/jss.v080.i01

[2]. Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. The R Journal, 10(1), 395-411. https://doi.org/10.32614/RJ-2018-017

[3]. Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1), 1-32. https://doi.org/10.18637/jss.v076.i01

[4]. Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. Psychological Methods, 22(2), 240-261. https://doi.org/10.1037/met0000065

[5]. Efron, B. (2013). Bayes' theorem in the 21st century. Science, 340(6137), 1177-1178. https://doi.org/10.1126/science.1236536

[6]. Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. Psychonomic Bulletin & Review, 25, 5-34. https://doi.org/10.3758/s13423-017-1262-3

[7]. Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. Journal of the Royal Statistical Society: Series A, 182(2), 389-402. https://doi.org/10.1111/rssa.12378

[8]. Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S and Type M errors. Perspectives on Psychological Science, 9(6), 641-651. https://doi.org/10.1177/1745691614551642

[9]. Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24, 997-1016. https://doi.org/10.1007/s11222-013-9416-2

[10]. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31, 337-350. https://doi.org/10.1007/s10654-016-0149-3

[11]. Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. Brazilian Journal of Physical Therapy, 23(4), 290-301. https://doi.org/10.1016/j.bjpt.2018.12.006

[12]. Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. Psychonomic Bulletin & Review, 21, 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

[13]. Kelter, R. (2020). Bayesian alternatives to null hypothesis significance testing in biomedical research: A non-technical introduction to Bayesian inference with JASP. BMC Medical Research Methodology, 20, 142. https://doi.org/10.1186/s12874-020-00980-6

[14]. Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General, 142(2), 573-603. https://doi.org/10.1037/a0029146

[15]. Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. Psychonomic Bulletin & Review, 25, 155-177. https://doi.org/10.3758/s13423-017-1272-1

[16]. Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review, 25, 178-206. https://doi.org/10.3758/s13423-016-1221-4

[17]. Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. Nature, 520, 612. https://doi.org/10.1038/520612a

[18]. Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. Oikos, 128(7), 912-928. https://doi.org/10.1111/oik.05985

[19]. Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. Frontiers in Psychology, 10, 2767. https://doi.org/10.3389/fpsyg.2019.02767

[20]. McNeish, D. (2016). On using Bayesian methods to address small sample problems. Structural Equation Modeling: A Multidisciplinary Journal, 23(5), 750-773. https://doi.org/10.1080/10705511.2016.1186549

[21]. McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. The American Statistician, 73(sup1), 235-245. https://doi.org/10.1080/00031305.2018.1527253

[22]. Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review, 23, 103-123. https://doi.org/10.3758/s13423-015-0947-8

[23]. Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, 11(2), 5018-5051. https://doi.org/10.1214/17-EJS1337SI

[24]. Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. Psychological Methods, 26(1), 103-126. https://doi.org/10.1037/met0000275

[25]. van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child Development, 85(3), 842-860. https://doi.org/10.1111/cdev.12169

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

[27]. Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27, 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

[28]. Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Smira, M., Epskamp, S., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25, 35-57. https://doi.org/10.3758/s13423-017-1343-3

[29]. Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. Psychonomic Bulletin & Review, 25, 58-76. https://doi.org/10.3758/s13423-017-1323-7

[30]. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

[31]. Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond $p < 0.05$. The American Statistician, 73(sup1), 1-19. https://doi.org/10.1080/00031305.2019.1583913

[32]. Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. Bayesian Analysis, 13(3), 917-1003. https://doi.org/10.1214/17-BA1091