ROeS 2025 Conference, 34th Conference of the Austro-Swiss Region (ROeS)

STRATOS-TG2 project P 6: Contrasting Bayesian and frequentist model building for descriptive research questions–a paired-design experiment

Mariana Nold, Aliaksandr Hubin, Michael Kammer, Georg Heinze

Date: September 16, 2025

Introduction

STRATOS Initiative and TG2

STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies) is a global collaboration of statistical experts, aiming to provide accessible and evidence-based guidance for the design and analysis of observational studies. `https://stratos-initiative.org`

STRATOS TG2 is one of several topic groups within the initiative. `https://stratostg2.github.io/`

Main Aim TG2: Develop guidance for variable selection and functional form specification in multivariable analyzes.

Thanks to all TG2 members for supporting this project.

Bayesian vs. Frequentist Thinking

Traditional Distinctions:

- **Frequentist:** Probability as long-run frequency; parameters are fixed.
- **Bayesian:** Probability as belief; parameters are random variables.

Established Perspectives:

- Both represent distinct statistical paradigms - frequentism dominated much of the 20th century, while Bayesian approaches have gained momentum since the 1990s, driven by computational advances.
- Disciplinary and national traditions shape terminology and usage.

Diversity Within the Frequentist Paradigm

A Non-Monolithic Paradigm - Selected aspects of internal diversity:

- **Interpretation:** Disputes over p-values (usage and interpretation) and evolving use of confidence intervals instead of p-values (e.g., "New Statistics" [2]).
- **Modeling Practice:** Debates on the selection of variables and functional forms $\rightarrow$ Divergent views on variable selection and model-building criteria [19].
- **Philosophical Variants:** From strict sampling-based frequentism to likelihoodist approaches using LR tests.

The Evolving Meaning of "Bayesian"

- As Fienberg (2006) [5] notes in his seminal paper *When Did Bayesian Inference Become "Bayesian"*, early 20th-century statisticians used Bayes' theorem without referring to Bayesian methods.
- Fienberg's historical analysis shows that the meaning of Bayesian has evolved and continues to change in disciplinary, national, and epistemological contexts.
- Recent computational advances have shifted Bayesian practice towards prediction, with priors increasingly used to stabilize inference [8] $\rightarrow$ Bayesian rationality is evolving in important ways at the moment [13].

Two Developments Motivating our Project

- **Paradigm Diversity:** The frequentist–Bayesian divide is increasingly viewed as a spectrum. Both paradigms encompass diverse traditions shaped by historical and institutional contexts [5, 14, 20].
- **Critique of Modeling Practice:**
  - Many practices reflect the *True Model Myth*, often lacking solid theoretical grounding and clear research questions [1].
  - Cross-disciplinary critiques highlight deficits in scientific rigor and transparency within common modeling practices [10, 9, 21].

Guiding Questions: What distinguishes frequentist and Bayesian reasoning today? How can modeling practice be improved in light of these critiques?

Focus on Descriptive Research Questions

Definition (following [1, 8]): A descriptive research question summarizes statistical patterns that reflect **changes between units** - differences between individuals or observational units - without invoking interventions or counterfactuals. In contrast, a causal research question addresses **changes within units,** to ask how the outcome for the same unit would differ under alternative interventions.

- Descriptive: Differences observed between subjects/units.
- Causal: hypothetical contrasts within subjects/units.
- Describe patterns, not causal mechanisms.
- May inform, but not prove, causality.

Statistical Thinking as a Spectrum: Blurring Boundaries

- As Lin (2024) [14] notes, the frequentist–Bayesian divide is overly simplistic and masks a spectrum of nuanced positions.
- This perspective acknowledges that methodological choices often combine elements from both traditions, depending on context, goals, and epistemological stance.
- Bayesian and frequentist analyses can converge in practice, as shown in Inchausti's textbook [12] *Statistical Modeling With R: A Dual Frequentist and Bayesian Approach for Life Scientists*.

$\Rightarrow$ This raises the question of robustness: When do Bayesian and frequentist approaches yield converging substantive insights, and how is reliable inference defined within each paradigm?

Robustness Across Paradigms: Inspiration from Nuijten (2022)

Nuijten's Retrospective 4-Step Check [16]: A minimal-resource framework to evaluate the robustness of published findings.

1. **Internal Consistency:** Are results coherent?
2. **Reanalysis:** Do they replicate under the original strategy?
3. **Alternative Strategies:** Are conclusions stable across analytical choices?
4. **Replication:** Do the findings hold in a new sample?

$\Rightarrow$ Inspired by Nuijten's logic, we adapt this idea to a proactive setting with a focus on Step 3, the alternative strategy.

Project Research Question

How can analysis plans be designed to integrate frequentist and Bayesian perspectives - embedding robustness checks proactively, to improve modeling practices for descriptive research questions?

- Promotes proactive robustness as a means to encourage methodological openness and improve planning quality.
- Embeds critical reflection from the dialogue of the cross-paradigm.

Methodological Approach

- **Paired design with four statisticians:** Each locates themselves between Bayesian and frequentist thinking (Lin, 2024), with a leaning toward one side.
- **Cross-paradigm robustness check:** Each participant applies a robustness check to a counterpart's analysis plan from the opposite paradigm.
- **Four case studies:** Each study addresses a specific statistical focus and allows the comparison of paradigm-specific modeling strategies. Simplifying assumptions help avoid overlapping challenges.

Study Design: Seven Phases of the Analytical Workflow

Phase 0: Data Cleaning

Phase 1: Draft Statistical Analysis Plan (SAP)

IDA: Initial Data Analysis (IDA) [11] → Refinement based on IDA

Phase 2: Refine SAP based on discussion within paradigm

Phase 3: Robustness check by opposite paradigm

Phase 4: Execute main analysis and robustness analyses

Phase 5: Compare approaches and document insights

Phase 6: Interpret results and reflect on implications

Case study I: School-belonging Case Study

Empirical Application based on PISA 2018 data from Austria

- **Outcome variable:** Binary indicator for school withdrawal ($1 =$ yes), operationalized as a very low sense of belonging to school.
- **Focal predictors:**
  - (1) Bullying victim status (yes/no)
  - (2) School-level truancy (scale 1–4, based on proportion of students with attendance problems per school)

  $\rightarrow$ Combined into **8 distinct profiles**.

Substantive Research Questions:

- How does school belonging vary across bullying/truancy profiles?
- What role do background variables play in understanding differences between these focal groups, especially considering their potentially uneven distribution across profiles?

Background Variables: Finn's Engagement Model

Theoretical Framing, Finn (1989) [6]

Background variables were selected from PISA data based on Finn's engagement model, capturing the risks and protection dimensions within the school alienation cycle.

Core Dimensions

- **Participation**: Presence and involvement
- **Identification**: Sense of connection and belonging
- **Success**: Experience of achievement and competence

Withdrawal Cycle

Low Participation $\leftrightarrow$ Low Identification $\rightarrow$ Withdrawal $\rightarrow$ Negative Outcomes $\rightarrow$ Low Participation $\leftrightarrow$ Low Identification

Variable Structure: Linking Theory and Data

Background Variables (BV)

- Conceptually based on Finn's theory $\rightarrow$ Represent protective or risk factors in the alienation process.
- Scaled such that higher values indicate lower risk ($\uparrow$ = protective).
- $\Rightarrow$ When theoretically aligned, BV act as predictive indicators: **Higher values are expected to reduce withdrawal behavior**

Scaling Approach (Hypothesized Pattern):

- BVs scaled into quartiles (25%, 50%, 75%)
- If consistent with Finn's framework, higher quartiles should indicate stronger protection and thus lower risk of withdrawal
- Empirical consistency to be checked

Data Preparation and Modeling Framework

Methodological Boundaries (for simplicity, to avoid overlapping challenges)

- Dataset treated as random sample and missing data handled via single imputation

Initial Data Analysis (IDA)

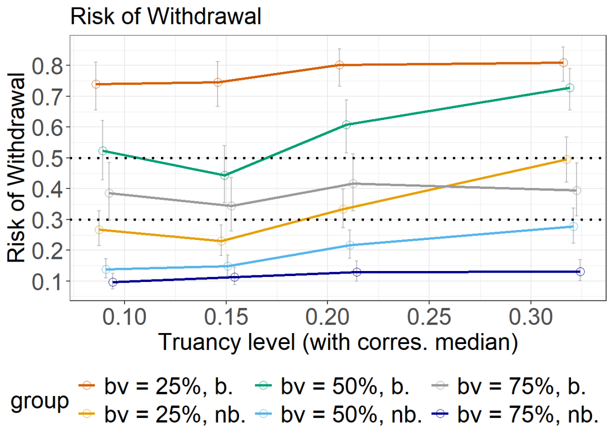
- Two focal variables plus **15 background variables**: $\rightarrow$ 4 binary, 11 continuous (7 individual-v. standardized and 4 school-level-v. are proportions)

First Modeling Step: Parsimonious model to summarize data structure

- Multilevel logistic regression with random intercept (schools as clusters) and interaction between focal variables.

Between-Units Student Profiles: Bullying × School Truancy Level

Bullying emerges as a distinguishing factor, while truancy plays a limited role. The findings are broadly consistent with the Finn theory. Scaled BV (↑ = low risk) → 25%, 50%, 75% quantiles.)



Step 2: Background Model – Are encoded assumptions and information sufficient for prediction?

Induction within Deduction (Gelman & Shalizi, 2011)

"Statistical models are tools for inductive reasoning within a deductive framework."

- Models encode assumptions → Derive model based on assumptions (**deduction**)
- Predictions tested against data → generate predictions (**induction**) and compare to observed data
- Failed predictions expose limits → learn from mismatches (**falsification**)

Core question: Are encoded assumptions and information sufficient to predict the outcome distribution in relevant regions?

Models as Filters - Diagnostics in Practice

Learning based on what the model cannot predict

A model is **useful** when this filtering enables a meaningful insight into the research question.

- Learning arises from mismatches between model predictions and observed data - this is where insight lives.
- Diagnostics are important **purpose-built tools**, crafted to test specific model assumptions relevant to the research question.
- By comparing models, we uncover their blind spots; these limitations inform a deeper understanding.

Background Model and Diagnostic Strategy

- Theory-driven background model (*withdrawal circle*) includes only BVs.
- Focal predictors - `Bull` (binary) and `Truancy level` (categorical) - are **intentionally excluded**.
- **Key question:** Can the background model predict outcome distributions across focal groups - despite being blind to them?
- The distribution of theoretically derived BVs across focal groups is analytically informative, as it supports **understanding** of group-specific dynamics and the plausibility of theoretical explanations based on statistical patterns.
- Model fit assessed via **randomized quantile residuals**.

Basic Idea: Randomized Quantile Residuals

- Residuals are calculated using the cumulative distribution function (CDF) of the fitted model.
- If the model fits well, the observed values behave like random draws, and the residuals appear uniform.
- These are transformed to normality using the probability integral transform (PIT):

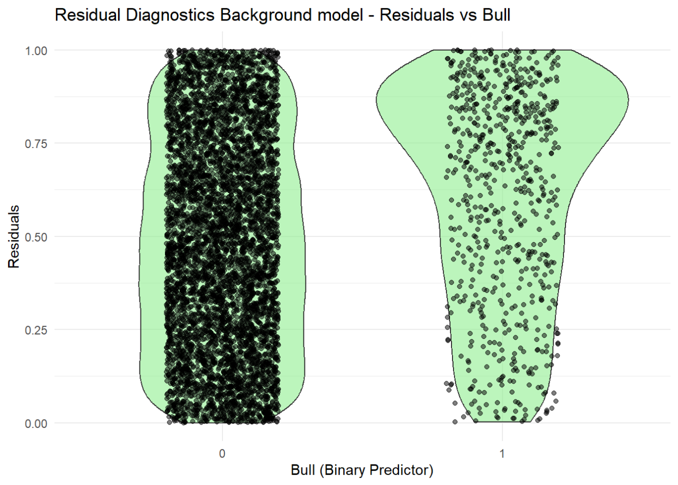
$$r_{Q,i} = \Phi^{-1}\left(F(y_i \mid \text{model})\right)$$

- For discrete outcomes, small random noise ("jittering") ensures smooth residuals.

Sources: Dunn (1996) et al. [3], Feng et al. (2017) [4], Inchausti (2022) [12].

Residual Visualization and Tilt Signature [15]

Tilt signature: Residuals cluster in the upper half of the uniform scale.
→ **Systematic underestimation:** Withdrawal is higher than predicted for victims, and slightly lower for non-victims.



Diagnostic Insight: Limits of the Background Model

Diagnostic Insight

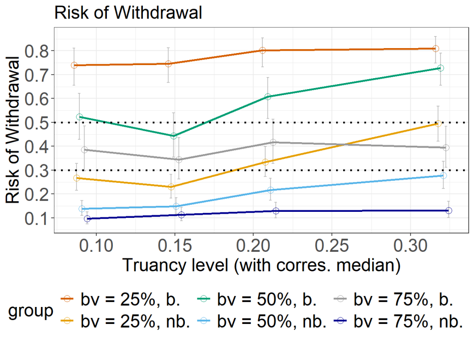
The background model successfully captures predictive patterns within the outcome region of `Truancy level`, but fails to represent structures associated with `Bull` (bullying).

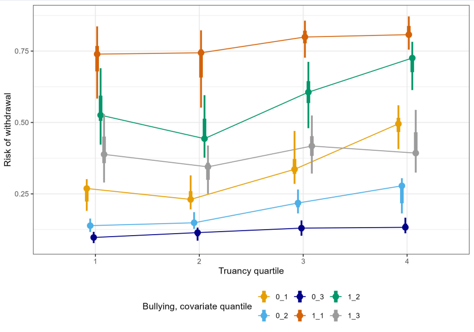
- The residuals for `Bull` show a clear **tilt**, indicating systematic underestimation.
- No such pattern for `Truancy level` – its signal is well captured (not shown).
- This contrast highlights the **limits** of the background model: it cannot account for all regions of the outcome distribution.
- These findings reveal empirically grounded patterns that invite substantive interpretation: they do not prove causality, but may inform causal reasoning and guide further inquiry.

Frequentist Robustness Check I

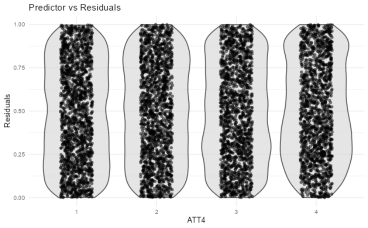
- **Aim**: replicate Bayesian analysis as close as possible
- **Estimand**: Probability of high sense of belonging for representative students at different levels of school truancy and bullying experience
- **Data and variables**: same as Bayesian analysis
- **Methods**: same as Bayesian analysis - Multilevel logistic regression with random intercepts for schools, focal variables including interaction, covariates
  - Background model without focal variables as comparator
  - Bootstrap for uncertainty quantification of primary estimand
- **Diagnostics and performance**: same as Bayesian analysis - Randomized Quantile Residuals
  - Also calibration and c-index, assess normality of random effect
- **Model comparison**: Likelihood ratio test

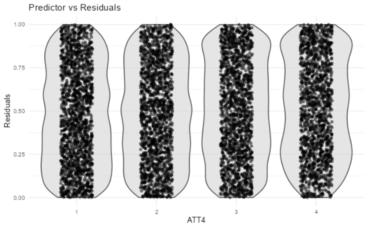
Frequentist Robustness Check II

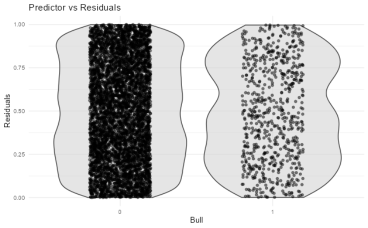


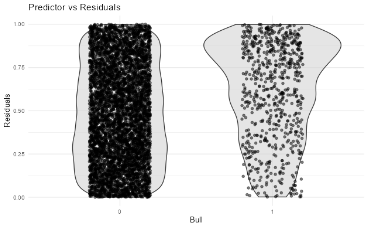


Frequentist Robustness Check III









Conclusion Frequentist vs Bayesian Randomized Quantile Residuals

- **Goal:** Both approaches evaluate model fit by comparing observed data with model-implied distributions.
- **Construction:**
  - **Frequentist:** Uses fixed parameter estimates.
  - **Bayesian:** Averages over posterior samples to account for uncertainty.
- **Interpretation:**
  - **Frequentist:** Residuals reflect fit under point estimates.
  - **Bayesian:** Residuals reflect fit across the posterior.
- **Randomization:** Both apply uniform randomization to handle discrete outcomes.
- **Diagnostics:** Similar plots (e.g., residual vs predictor, QQ plots) are used in both cases.

Consistency Between Bayesian and Frequentist Approaches

- **Residual Diagnostics:** Randomized quantile residuals are nearly identical across Bayesian and frequentist models, indicating strong agreement.
- **Interpretation Frameworks:** Bayesian methods express uncertainty more explicitly, yet key diagnostics remain stable across paradigms.
- **Computational Demands:** Bayesian estimation can be resource-intensive. In this case study, sampling from the posterior caused storage problems, underscoring the need for efficient implementation.
- **Scope of Results:** Shown residual plots represent only a subset. Full diagnostics are available in the extended analysis.
- **Robustness of Conclusions:** Substantive findings are consistent between approaches.

Conclusion

Statistical Pluralism in Practice

Methodological Diversity with Empirical Convergence

Froslie (2019) [7] describes statistics as a **language for reasoning with data**, not just a set of tools. Statistics as theory constitutes an autonomous justification of its own methodology: it establishes a unifying and binding framework [18]. Each statistical school provides its own justification and methodological framework, which sets it apart; yet, their conclusions mainly converge in our case-studies, similar to examples in the textbook of Inchausty [12].

- Each statistical school offers different methodological justifications and specific strengths.
- When aligned through a coherent analysis plan, these approaches often lead to similar conclusions.

Frequentism and the Semantic Predicament

- **Semantic inconsistency** manifests in two interrelated forms:
  - → Terminological ambiguity (polysemy) – similar terms with differing meanings across fields (e.g., psychology vs. biometrics); statistically valid, yet prone to misunderstanding.
  - → Conceptual vagueness (erosion) - terms such as 'control variable' or 'p-value' lose precision and are used without clear reasoning, weakening the study foundations [see e. g. 21, 17].
- Frequentist methods, widely used by nonexperts, foster **conceptual vagueness** → a **semantic predicament** where vague terms persist.
- **STRATOS** counters this by promoting clarity and education-based reasoning in statistical practice.

Proactive Robustness: Revealing Hidden Biases

- **External critique is essential:** It uncovers overinterpretation and blind spots within statistical schools.
- **Early robustness checks:** Joint planning between traditions fosters transparency and trustworthiness.
- **Confronting perspectives sharpens reasoning:** It reveals hidden assumptions and encourages reflection.
- **Comparing modeling assumptions:** Direct contrasts deepen understanding between paradigms.
- **Plurality of methodological views:** Debating what counts as justified methodology helps draw clearer boundaries to less rigorous approaches.

Key Insight: *Robustness becomes a lens for epistemological reflection - not just a technical safeguard.*

Significance and Implications

Why This Matters

- Embedding critical perspectives **before analysis** challenges dominant conventions.
- Opens space for **rethinking modeling norms** and exploring **new research trajectories**.
- Encourages **reflexivity** in statistical practice - moving beyond routine application.

Broader Impact

- Enhances the **robustness** of findings.
- Promotes **methodological pluralism** and **innovation**.
- Supports a more **critical and creative research culture** across paradigms.

Future directions for STRATOS - TG2P6

- In light of the **alleged blurring boundaries** between Bayesian and frequentist approaches, are we truly united by more than what divides us?
- Should we more clearly describe what these boundaries are, namely, the **different justifications** for Bayesian and frequentist methodologies, and what exactly is being blurred?
- Do we need additional **case studies**? If so, what kind of case studies would best illuminate convergence, divergence, or intersection?

That concludes our presentation.
We welcome your questions and comments.

References

[1] John B Carlin and Margarita Moreno-Betancur. **"On the uses and abuses of regression models: a call for reform of statistical practice and teaching".** In: *Statistics in Medicine* 44.13-14 (2025), e10244.

[2] Geof Cumming and Robert Calin-Jageman. **Introduction to the new statistics: Estimation, open science, and beyond.** Routledge, 2024.

[3] Peter K. Dunn and Gordon K. Smyth. **"Randomized Quantile Residuals".** In: *Journal of Computational and Graphical Statistics* 5.3 (1996), pp. 236–244. ISSN: 10618600. DOI: 10.2307/1390802.

[4] Cindy Xin Feng, Alireza Sadeghpour, and Longhai Li. **"Randomized Quantile Residuals: an Omnibus Model Diagnostic Tool with Unified Reference Distribution".** In: 2017. URL: https://api.semanticscholar.org/CorpusID:126143700.

[5] Stephen E Fienberg. **"When did Bayesian inference become" Bayesian"?"** In: *Bayesian Analysis* (2006).

[6] Jeremy D. Finn. **"Withdrawing From School".** In: *Review of Educational Research* 59.2 (1989), pp. 117–142. DOI: 10.3102/00346543059002117. eprint: https://doi.org/10.3102/00346543059002117. URL: https://doi.org/10.3102/00346543059002117.

[7] Kathrine Frey Frøslie and Jo Røislien. **"Sprechen sie statistik?"** In: *Tidsskrift for Den norske legeforening* (2019).

[8] Andrew Gelman, Jennifer Hill, and Aki Vehtari. **Regression and other stories.** Analytical Methods for Social Research. Cambridge, UK: Cambridge University Press, 2021. ISBN: 1107676517.

[9] Andrew Gelman and Eric Loken. **"The Statistical Crisis in Science".** In: *American Scientist* 102.6 (2014), p. 460. ISSN: 0003-0996. DOI: 10.1511/2014.111.460.

[10] Georg Heinze et al. **"Phases of methodological research in biostatistics—building the evidence base for new methods".** In: *Biometrical Journal* 66.1 (2024), p. 2200222.

[11] Georg Heinze et al. **"Regression without regrets–initial data analysis is a prerequisite for multivariable regression".** In: *BMC Medical Research Methodology* 24.1 (2024), p. 178.

[12] Pablo Inchausti. **Statistical Modeling With R: a dual frequentist and Bayesian approach for life scientists.** Oxford University Press, Nov. 2022. ISBN: 9780192859013. DOI: 10.1093/oso/9780192859013.001.0001. URL: https://doi.org/10.1093/oso/9780192859013.001.0001.

[13] Johannes Lenhard. **"A transformation of Bayesian statistics: computation, prediction, and rationality".** In: *Studies in History and Philosophy of Science* 92 (2022), pp. 144–151.

[14] Hanti Lin. **"To be a frequentist or Bayesian? Five positions in a spectrum".** In: *Harvard Data Science Review* 6.3 (2024).

[15] Alan BH Nguyen et al. **"LOO-PIT: A sensitive posterior test".** In: *Journal of Cosmology and Astroparticle Physics* 2025.01 (2025), p. 008.

[16] Michèle B Nuijten. **"Assessing and improving robustness of psychological research findings in four steps".** In: *Avoiding questionable research practices in applied psychology*. Springer, 2022, pp. 379–400.

[17] Don van Ravenzwaaij et al. **"Perspectives on scientific error".** In: *Royal Society Open Science* 10 (July 2023). DOI: 10.1098/rsos.230448.

[18] Bernhard Rüger. **Test-und Schätztheorie: Band I: Grundlagen.** R. Oldenbourg Verlag, 1999.

[19] Willi Sauerbrei et al. **"State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues".** In: *Diagnostic and prognostic research* 4.1 (2020), p. 3.

[20] Jordi Vallverdú. **Bayesians versus frequentists: a philosophical debate on statistical reasoning.** Springer, 2015.

[21] Ben Van Calster et al. **"Methodology over metrics: current scientific standards are a disservice to patients and society".** In: *Journal of Clinical Epidemiology* 138 (2021), pp. 219–226. ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2021.05.018. URL: https://www.sciencedirect.com/science/article/pii/S0895435621001700.