Indices of Effect Existence and Significance in the Bayesian Framework

There is now a general agreement that the Bayesian statistical framework is the right way to go for psychological science. Nevertheless, its flexible nature is its power and weakness, for there is no agreement about what indices should be computed or reported. Moreover, the lack of a consensual index of effect existence, such as the frequentist *p*-value, possibly contributes to the unnecessary murkiness that many non-familiar readers perceive in Bayesian statistics. Thus, this study describes and compares several indices of effect existence, provide intuitive visual representation of the “behavior” of such indices in relationship with traditional metrics such as sample size and frequentist significance. The results contribute to develop the intuitive understanding of the values that researchers report and allow to draw recommendations for Bayesian statistics description, critical for the standardization of scientific reporting.

# Introduction

The Bayesian framework is quickly gaining popularity among psychologists and neuroscientists (Andrews & Baguley, 2013). Reasons to prefer this approach are reliability, better accuracy in noisy data, better estimation for small samples, less proneness to type I error, the possibility of introducing prior knowledge into the analysis and, critically, results intuitiveness and their straightforward interpretation (Dienes & Mclatchie, 2018; Etz & Vandekerckhove, 2016; Kruschke, 2010; Kruschke, Aguinis, & Joo, 2012; Wagenmakers et al., 2018; Wagenmakers, Morey, & Lee, 2016). The frequentist approach has been associated with the focus on null hypothesis testing, and the misuse of *p*-values has been shown to critically contribute to the reproducibility crisis of psychological science (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Szucs & Ioannidis, 2016). There is a general agreement that the generalization of the Bayesian approach is a way of overcoming this issue (Benjamin et al., 2018; Etz & Vandekerckhove, 2016; Maxwell, Lau, & Howard, 2015; Wagenmakers et al., 2017).

While the probabilistic reasoning promoted by the Bayesian framework is pervading most of data science aspects, it is already well established for statistical modelling. This facet, on which psychology massively rely, could roughly be grouped into two soft-edged categories; predictive and structural modelling. Although a statistical model can (often) serve both purposes, predictive modelling is devoted to build and find the best model that accurately predicts a given outcome. It is centered around the concepts such as fitting metrics, predictive accuracy and model comparison. At the extrema of this dimension lie machine and deep learning models, used for their strong predictive power, often at the expense of Human readability (these models has been often referring to as “black-boxes”, emphasizing the difficulty to appraise their internal functioning; Burrell, 2016; Castelvecchi, 2016; Snoek, Larochelle, & Adams, 2012). On the other side, psychologists are often using more simple models (for instance related to the general linear framework) to explore their data. Within this framework, the goal switches from building the best model to understanding the parameters inside the model. In reality, the methodological pipeline often starts with predictive modelling involving model comparison (“what is the best model of the world (*i.e.*, the observed variable)”) and then seemingly transit to structural modelling: “given this model of the world, how the effects (*i.e.*, model parameters) are influencing the outcome”. For this last part, they often rely on an index of effect “existence”.

Indeed, while one of the strengths of the Bayesian framework is its probabilistic parameter estimation, allowing to quantify the inherent uncertainty associated with each estimation, psychologists are also interested in parameter *existence*: A decision criterion that allows them to conclude if an effect is “different from 0” (statistically corresponding to either “not negligible” or “not of the opposite direction”). In other words, to know, before taking interest in the importance, relevance or strength of the effect, whether it is related to the outcome in a given direction. This need has led to the wide adoption of the frequentist *p*-value, used as an index of effect existence, and its acceptance was accompanied with the creation of arbitrary clusters for its classification (.05, .01 and .001). Unfortunately, these heuristics have severely rigidified, becoming a goal and threshold to reach rather than a tool for understanding the data (Cohen, 2016; Kirk, 1996).

Thus, the ability of the Bayesian framework to answer psychological questions without the need of such null-hypothesis testing indices is often promoted as the promise of a “a new world without *p*-value” as opposed to the old and flawed frequentist one. Nonetheless, it seems that “effect existence” indices and criteria are useful for Humans to gain an intuitive understanding of the interactions and structure of their data. It is thus unsurprising that the development of Bayesian user-friendly implementations was accompanied with the promotion of the Bayes Factor (BF), an index reflecting the predictive performance of a model against another (*e.g.*, the null vs. the alternative hypothesis; Jeffreys, 1998; Ly, Verhagen, & Wagenmakers, 2016). It provides many advantages over the *p* value, having a straightforward interpretation (“the data were 3 times (*BF* = 3) more likely to occur under the alternative than the null hypothesis”) and allowing to make statements about the alternative, rather than just the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). Moreover, recent mathematical developments allow its computation for complex models (Gronau, Van Erp, et al., 2017; Gronau et al., 2017). Although the BF lives up to the expectations of a solid, valid, intuitive and better index compared to the p value, its use for model selection is still a matter of debate (Piironen & Vehtari, 2017). Indeed, as the predictions used for its computation are generated from the prior distributions on the model parameters, it is highly dependent on priors specification (Etz, Haaf, Rouder, & Vandekerckhove, 2018; Kruschke & Liddell, 2018). Importantly for the aim of this paper, its use for estimating effect existence of parameters *within* a larger model remains limited, its computation being technically difficult and its interpretation not as straightforward as for “simple” tests such as t-tests or correlations.

Nevertheless, reflecting the need for such information, researchers have developed other indices based on characteristics of the posterior distribution, which represents the probability distribution of different parameter values given the observed data. This uncertainty can be summarized, for example, by presenting point-estimates of centrality (mean, median, …) and of dispersion (standard deviation, median absolute deviance, …), often accompanied with a percentage (89%, 90% or 95%) of the Highest Density Interval (HDI; referred to as the *Credible Interval* - CI). Although the Bayesian framework gives the possibility of computing many effect existence indices, no consensus has yet emerged on the ones to use, as no comparison has ever been done. This might be a rebuttal for scientists interested in adopting the Bayesian framework. Moreover, this grey area can increase the difficulty of readers or reviewers unfamiliar with the Bayesian framework to follow the assumptions and conclusions, which could in turn generate unnecessary doubt upon the entire study. While we think that such indices and their interpretation guidelines (in the form of rules of thumb) are useful in practice, we also strongly believe that such indices should be accompanied with the knowledge of the “behavior” in relationship with sample size and effect size. This knowledge is important for people to implicitly and intuitively appraise the meaning and implication of the mathematical values they report. This could, in turn, prevent the crystallization of the possible heuristics and categories derived from such indices.

Thus, based on the simulation of multiple linear regressions (one of the most widely used models), the present work aims at comparing several indices of effect existence solely derived from the posterior distribution, provide visual representations of the “behavior” of such indices in relationship with sample size, noise, priors and also the frequentist *p*-value (an index which, beyond its many flaws, is well known and could be used as a reference for Bayesian neophytes), and draw recommendations for Bayesian statistics reporting.

For all of the simulated models, we computed the following indices:

* ***p*-value**: Based on the frequentist regression, this index represents the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be greater than or equal to the actual observed results (Wasserstein, Lazar, & others, 2016).
* ***p*-direction**: the “Probability of Direction” (pd) corresponds to the probability (expressed in percentage) that an effect is strictly positive or negative (consistently with the median’s sign).
* ***p*-MAP**: the MAP-based *p*-value is related to the odds that a parameter has against the null hypothesis (Mills, 2017; Mills & Parent, 2014).
* ***p*-ROPE**: the ROPE-based *p*-value that represents the maximum percentage of HDI that does not contain (positive values) or is entirely contained (negative values) in the negligible values space defined by the ROPE.
* **ROPE (89%)**: this index refers to the percentage of the 89% HDI that lies within the ROPE, which is an interval considered as more stable than the frequentist-based 95% (McElreath, 2014, 2018).
* **ROPE (95%)**: this index refers to the percentage of the 95% HDI that lies within the ROPE. This interval corresponds to the one originally suggested by Kruschke (2014).
* **ROPE (full)**: this index refers to the percentage of the whole posterior distribution that lies within the ROPE.
* **Bayes factor**: this index indicates the degree by which the mass of the posterior distribution has shifted further away from or closer to the null value(s) (relative to the prior distribution), thus indicating if the null value has become less or more likely given the observed data. Here, the null was defined as a point null (0), and the Bayes factor was computed using the Savage-Dickey density ratio is computed, which is also an approximation of a Bayes factor comparing the marginal likelihoods of the model against a model in which the tested parameter has been restricted to the point null.

The Region of Practical Equivalence (ROPE) was defined as ranging from -0.1 to 0.1.

# Methods

The simulation aimed at modulating the following characteristics:

* **Model type**: linear or logistic.
* **“True” effect** (original regression coefficient from which data is drawn): Can be 1 or 0 (no effect).
* **Sample size**: From 20 to 100 by steps of 10.
* **Error**: Gaussian noise applied to the predictor with SD uniformly spread between 0.33 and 6.66 (with 1000 different values).

We generated a dataset for each combination of these characteristics, resulting in a total of 2 \* 2 \* 9 \* 1000 = 36000 Bayesian and frequentist models. The code used for generation is available [here](https://easystats.github.io/circus/articles/bayesian_indices.html) (please note that it takes usually several days/weeks to complete).

# Results

## Effect Detection

### Sensitivity to Noise

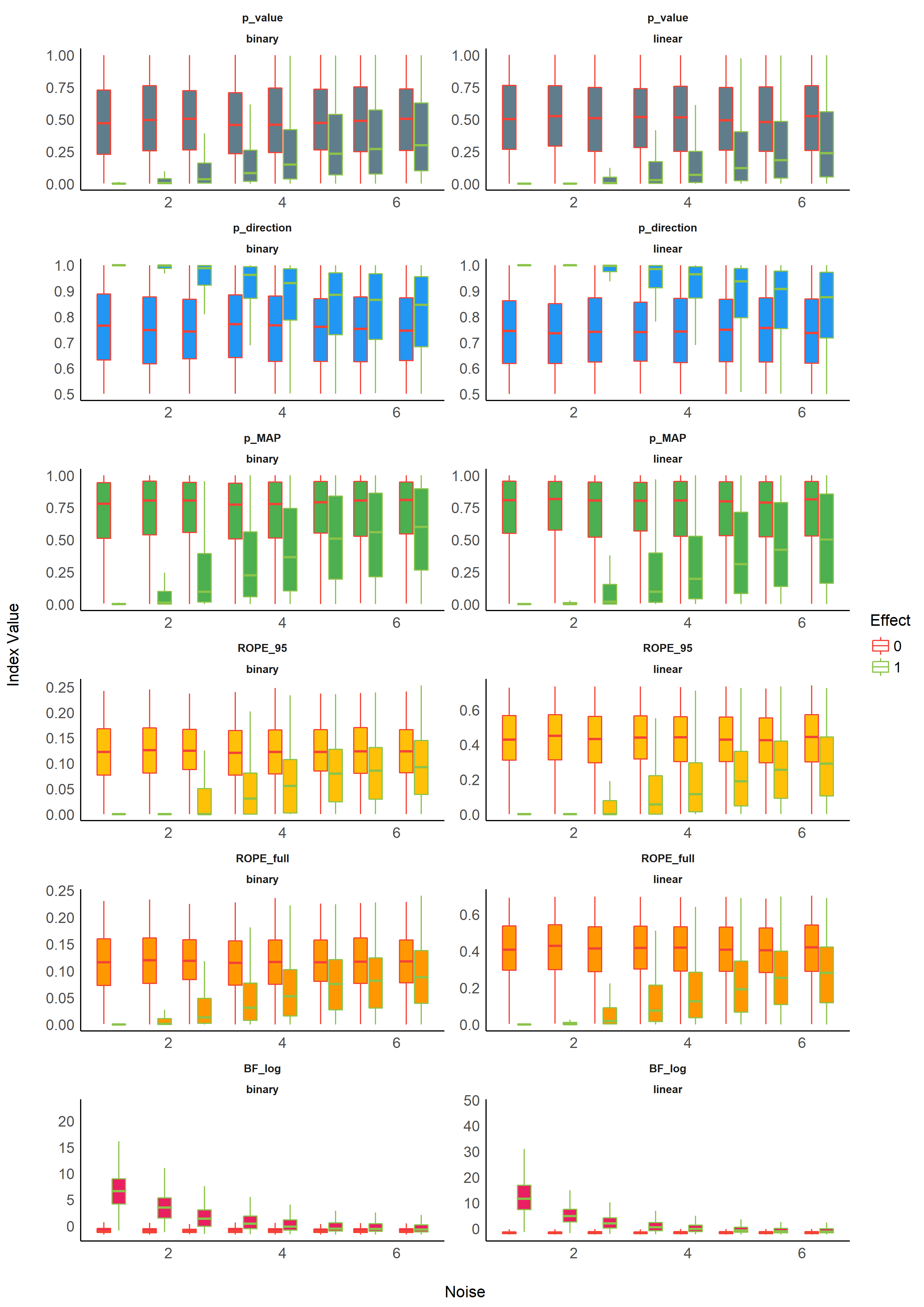


Figure 1.

### Sensitivity to Sample Size

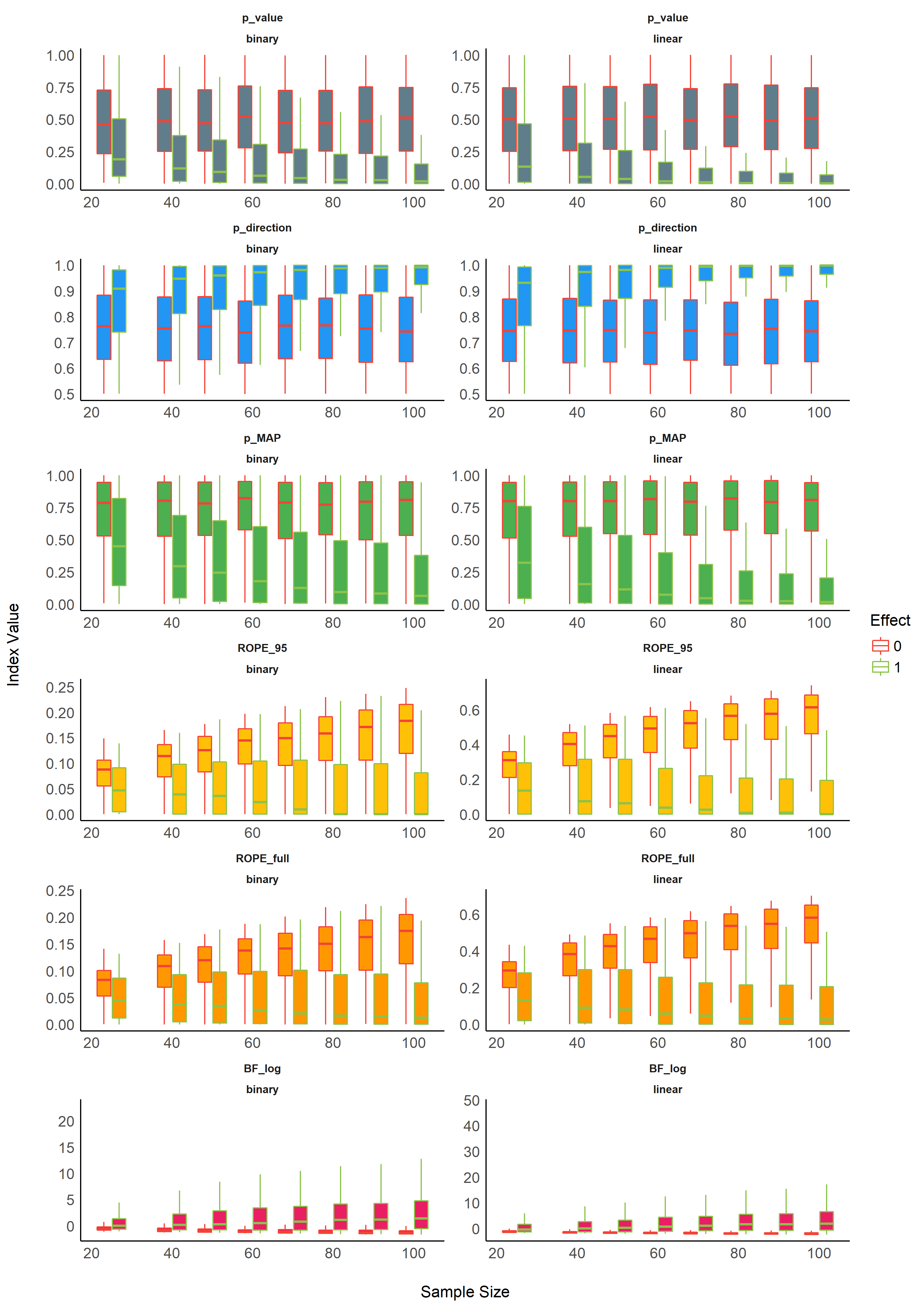


Figure 2.

### Statistical Modelling

We fitted a (frequentist) multiple linear regression to predict the presence or absence of effect with the different indices as well as their interaction with noise and sample size.

## Relationship with the frequentist *p*-value

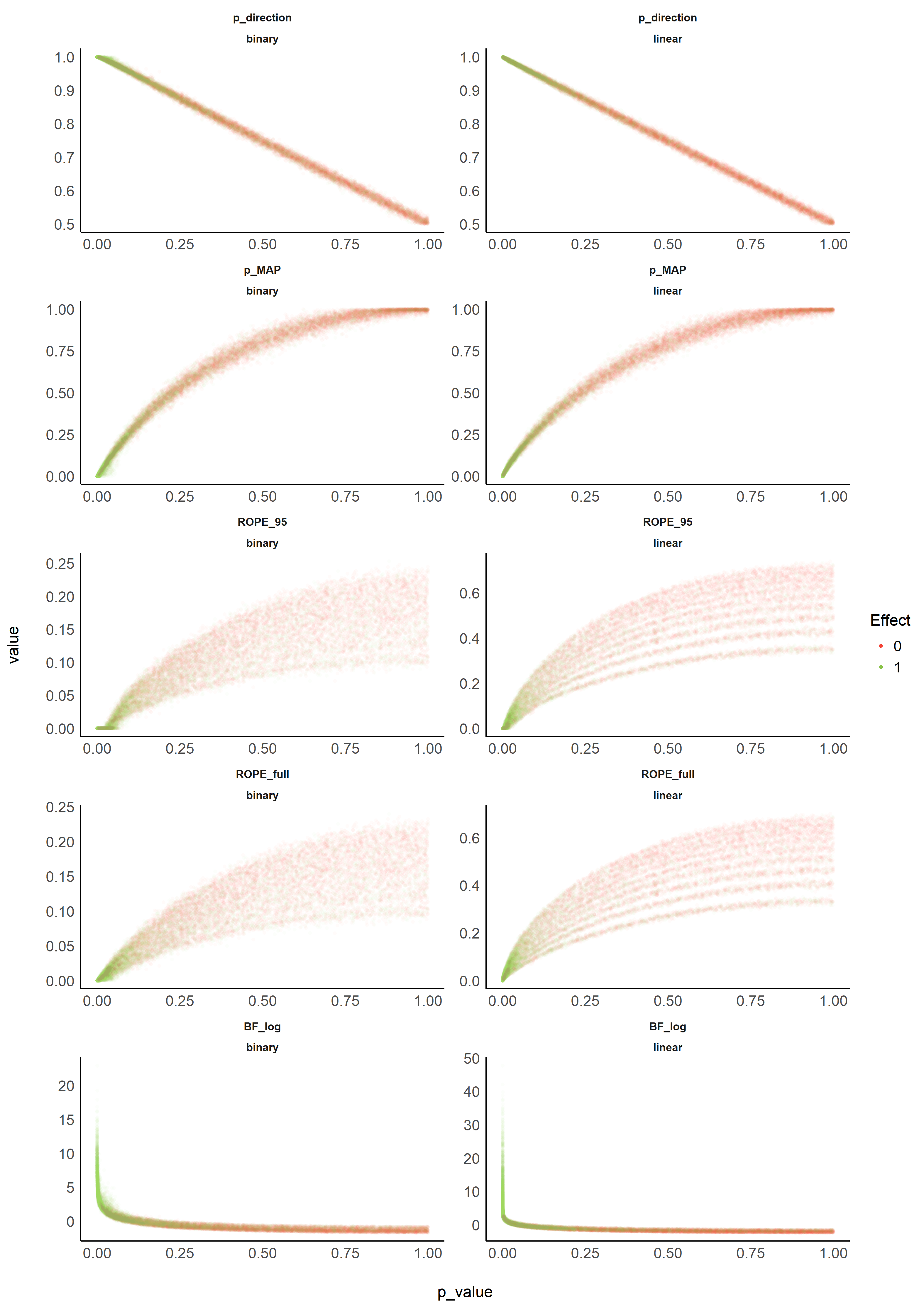
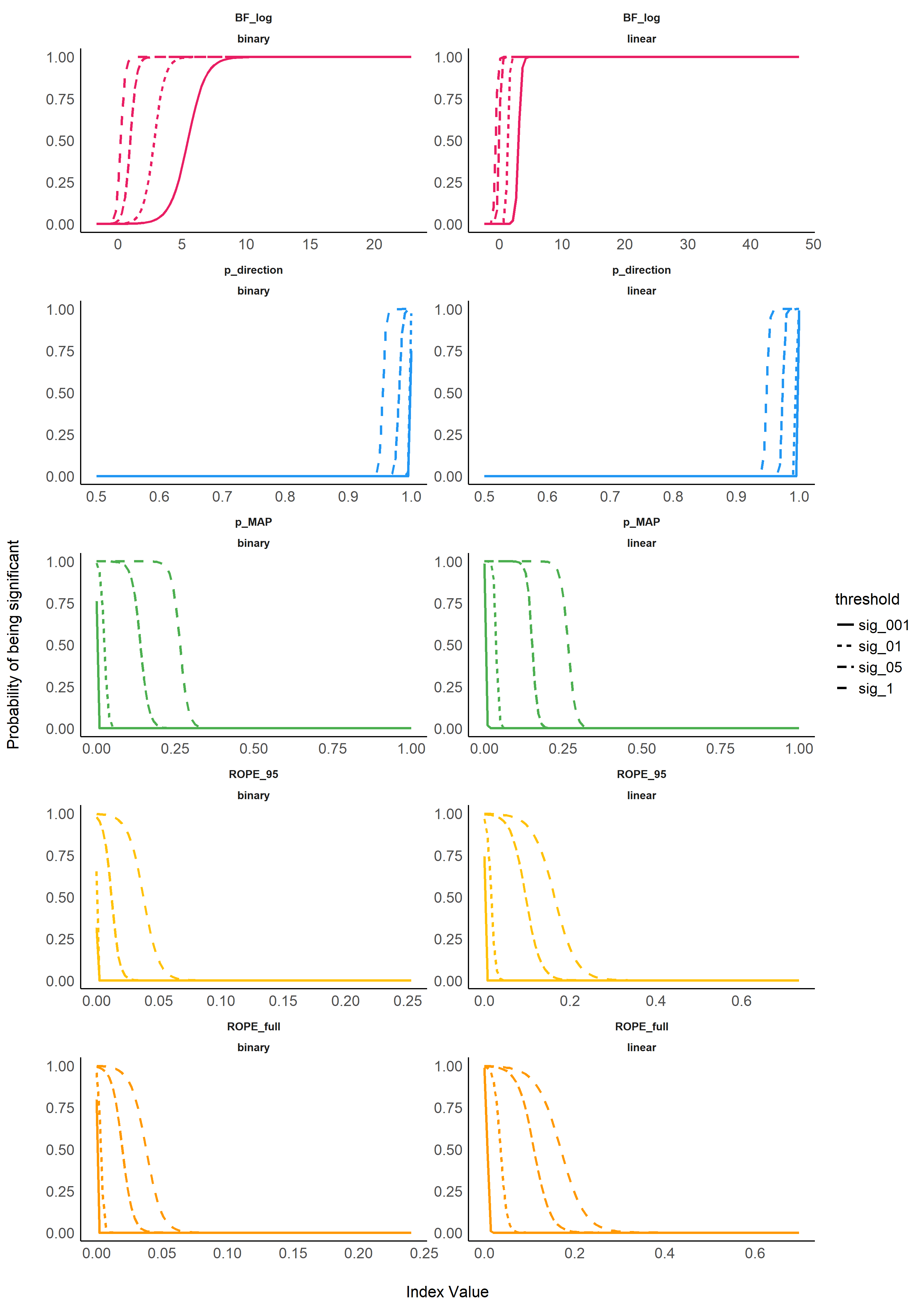
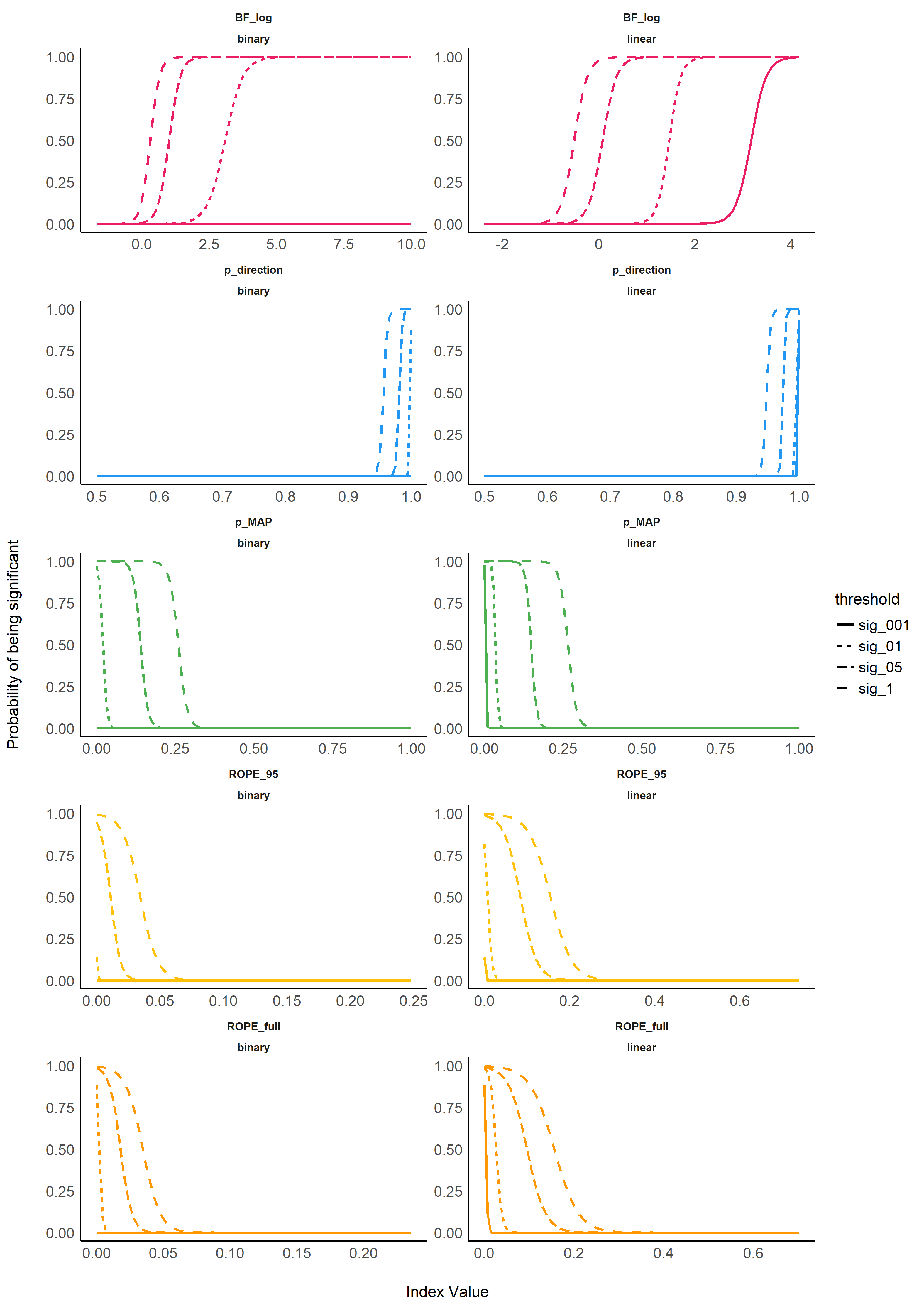


Figure 3.

## Relationship with frequentist *p*-based arbitrary clusters



## Relationship with Equivalence test

## Relationship between ROPE (full), p (direction) and Bayes Factor

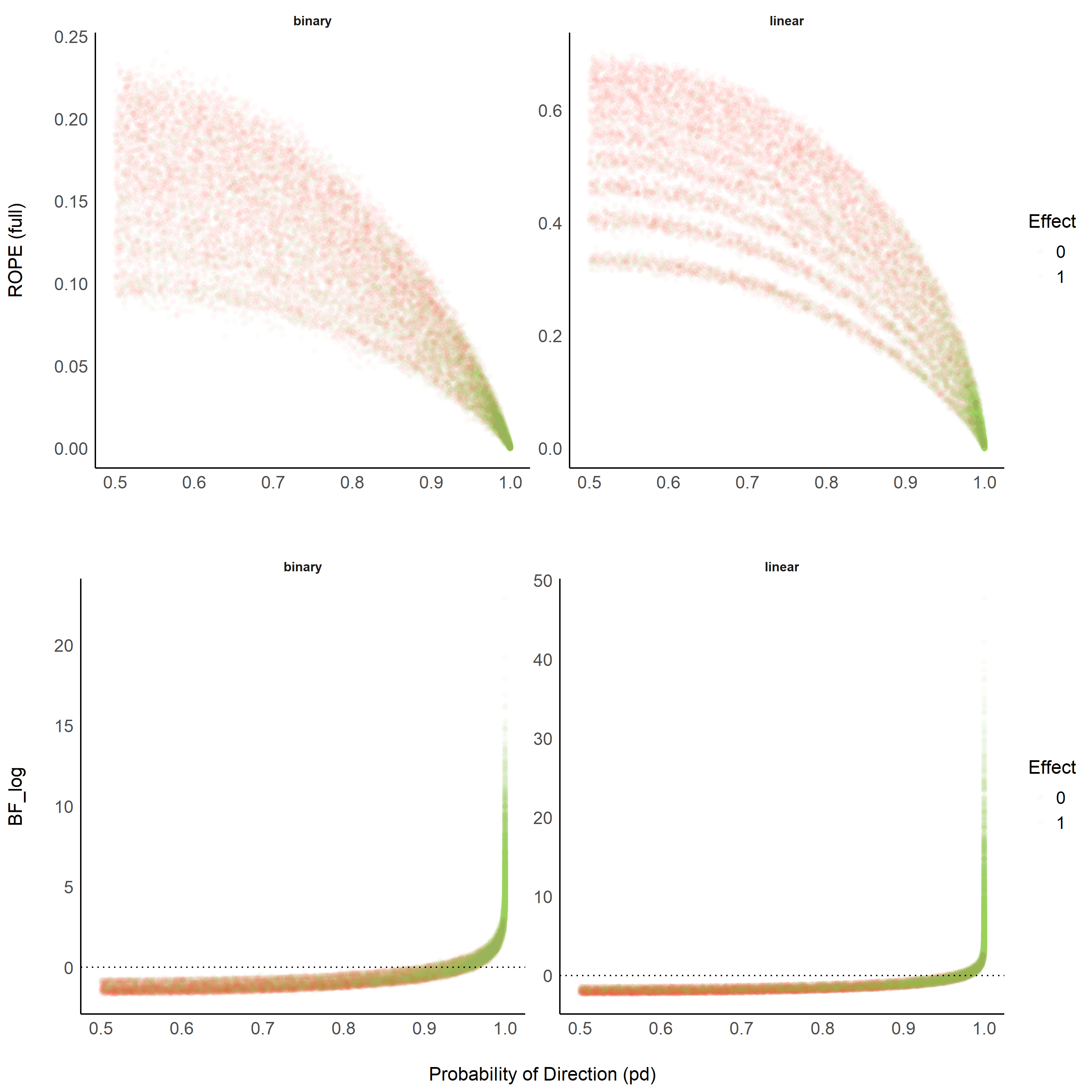


Figure 6.

> TableGrob (2 x 1) "arrange": 2 grobs  
> z cells name grob  
> 1 1 (1-1,1-1) arrange gtable[layout]  
> 2 2 (2-2,1-1) arrange gtable[layout]

# Discussion

Based on the simulation of multiple linear regressions, the present work aimed at comparing several indices of effect existence solely derived from the posterior distribution, provide visual representations of the “behavior” of such indices in relationship with sample size, noise, priors and the frequentist *p*-value.

While this comparison with a frequentist index may seem counterintuitive or wrong (as the Bayesian thinking is intrinsically different from the frequentist framework), we believe that this comparison is interesting for didactic reasons. The frequentist *p*-value “speaks” to many and can thus be seen as a reference and a way to facilitate the shift toward the Bayesian framework. This does not preclude, however, that a change in the general paradigm of effect existence seeking in necessary, and that Bayesian indices are fundamentally different from the frequentist *p*, rather than mere approximations or equivalents. Critically, we strongly agree on the distinction and possible dissociation between an effect’s existence and meaningfulness (Lakens, Scheel, & Isager, 2018). Nevertheless, we believe that assessing whether an effect is “meaningful” is highly dependent on the literature, priors, novelty, context or field, and that it cannot be assessed based solely on a statistical index (even though some of the indices, such as the ROPE-related ones, attempt at bridging existence with meaningfulness). Thus, researchers should rely on statistics to assess effect existence (as well as size and direction estimation), and systematically, but contextually, discuss its meaning and importance within a larger perspective.

Conclusions can be found in the [guidelines section](https://easystats.github.io/bayestestR/articles/guidelines.html).

# Conclusion

# References

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 1–7.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … others. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of ’playing the game’ it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17.

Cohen, J. (2016). The earth is round (p<. 05). In *What if there were no significance tests?* (pp. 69–82). Routledge.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, 2515245918773087.

Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PloS One*, *11*(2), e0149794.

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138.

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2017). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using warp-iii bridge sampling. *Psychometrika*, 1–24.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752.

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*(1), 178–206.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys’s default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487.

McElreath, R. (2014). *Rethinking: Statistical rethinking book package. R package version 1.391*.

McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

Mills, J. A. (2017). Objective bayesian precise hypothesis testing. *University of Cincinnati [Original Version: 2007]*.

Mills, J. A., & Parent, O. (2014). Bayesian mcmc estimation. In *Handbook of regional science* (pp. 1571–1595). Springer.

Piironen, J., & Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 2951–2959.

Szucs, D., & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *BioRxiv*, 071530.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … others. (2018). Bayesian inference for psychology. Part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for bayesian hypothesis testing in psychological science. *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, 123–138.

Wasserstein, R. L., Lazar, N. A., & others. (2016). The asa’s statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.