Indices of Effect Existence and Significance in the Bayesian Framework

# Abstract

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 particularly well established for statistical modelling. This facet, on which psychological science 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 extremum of this dimension lies deep learning models, used for their strong predictive power, sometimes at the expense of Human readability (these models has been often referred 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 simpler models (for instance, based on 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. Altough in reality, the 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 transition to structural modelling: “given this model of the world, how are the effects (*i.e.*, model parameters) influencing the outcome”. Nevertheless, for this last part, scientists often rely on an index of effect “significance”.

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 and logistic regressions (arguably some 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.

# Methods

## Data Simulation

The simulation of datasets suited for linear or logistic regression, we started by simulating a dependent, normally distributed *x* variable (with mean 0 and SD 1) of a given sample size. Then, the corresponding *y* variable was added, having a perfect correlation (in the case of data for linear regressions) or as a binary variable perfect separated by *x* (in the case of no effect, the *y* variable was created independent of *x*). Finally, a Gaussion noise was added (the error).

The simulation aimed at modulating the following characteristics: *data type* (*i.e.*, used for linear or logistic regression), *Sample size* (from 20 to 100 by steps of 10), *“true” effect* (original regression coefficient from which data is drawn prior to noise addition, 1 - presence of effect or 0 - absence of effect) and *noise* (Gaussian noise applied to the predictor with SD uniformly spread between 0.666 and 6.66, with 1000 different values). We generated a dataset for each combination of these characteristics, resulting in a total of 36000 (2 \* 2 \* 9 \* 1000) datasets. The code used for generation is available on github (<https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian/data>). Please note that it takes usually several days/weeks for the generation to complete.

## Indices

For each of these datasets, Bayesian and frequentist regressions were fitted to predict *y* with *x* as a unique predictor. The Bayesian models had default, mildly informative priors (normal distribution with mean 0 and SD 1) over the parameter of interest, and were fitted using MCMC (4 chains of 2000 iterations, half of which used for warm-up). For all of the simulated models, we computed seven indices related to the effect of *x*. These are availabe and described in details in the *bayestestR* R package **(BAYESTESTR CITATION)**.

### Frequentist *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 effect would be greater than or equal to the actual observed results (Wasserstein, Lazar, & others, 2016).

### Probability of Direction (*pd*)

The Probability of Direction (*pd*) varies between 50% and 100% and can be interpreted as the probability that a parameter (described by its posterior distribution) is strictly positive or negative (whichever is the most probable). It is mathematically defined as the proportion of the posterior distribution that is of the median’s sign **(BAYESTESTR CITATION)**.

### MAP-based *p*-value

The *MAP-based p-value* (also refered to as the “Bayesian *p*-value”) is related to the odds that a parameter has against the null hypothesis (Mills, 2017; Mills & Parent, 2014). It is mathematically defined as the density value at 0 divided by the density at the Maximum A Posteriori (MAP), *i.e.*, the equivalent of the mode for continuous distributions.

### ROPE (95%) and ROPE (full)

The *ROPE (95%)* refers to the percentage of the 95% HDI that lies within the ROPE. As suggested by Kruschke (2014), the Region of Practical Equivalence (ROPE) was defined as ranging from -0.1 to 0.1 for linear regressions and its equivalent, -0.18 to -0.18, for logistic models (based on the formula to convert log odds ratios to standardized differences; Cohen, 1988). The *ROPE (full)* refers to the percentage of the whole posterior distribution that lies within the ROPE.

### Bayes factor (vs. 0) and Bayes factor (vs. ROPE)

The Bayes Factor (*BF*) 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 hypothesis has become less or more likely given the observed data. We created two indices corresponding to two definitions for the null. In the case of testing against a point null (0), 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. We also computed the BF of the posterior distribution against the range of negligible values (the ROPE), …

## Data Anaysis

The aim of this study is two-fold: 1) compare indices of effect existence and significance between them, 2) provide visual guides to provide an intuitive understanding of the numeric values in relation with a known frame of reference (the frequentist NHST framework). Thus, we will start by 1) presenting the relationship between these indices and main sources of variance, such as effect existence, sample size and noise. 2) Compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds (.05, .01, .001). Taken together, these results will help us to outline numeric guides to ease the reporting and interpretation of the indices.

# Results

## Impact of Sample Size and Noise

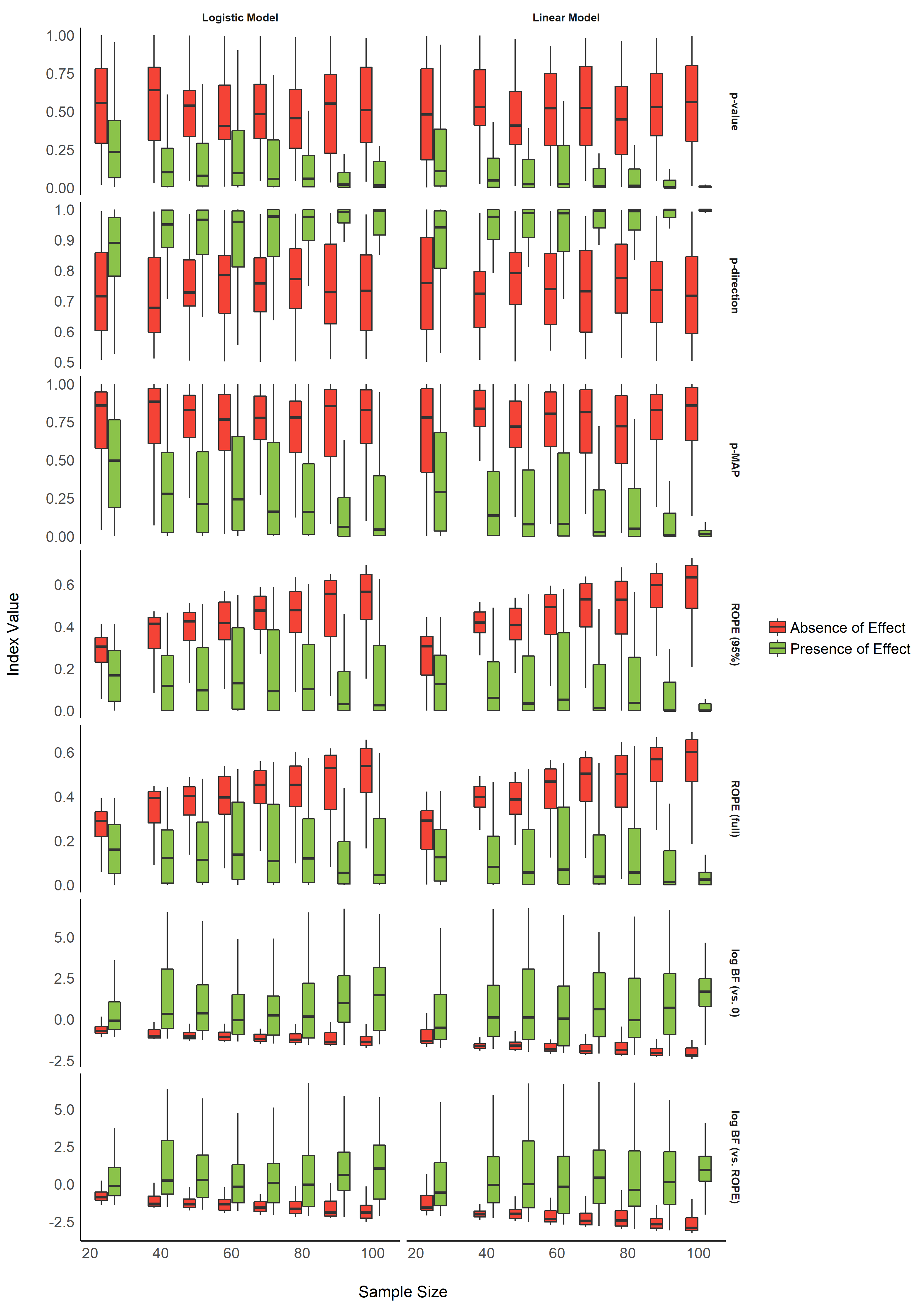


Figure 1. Impact of Sample Size.

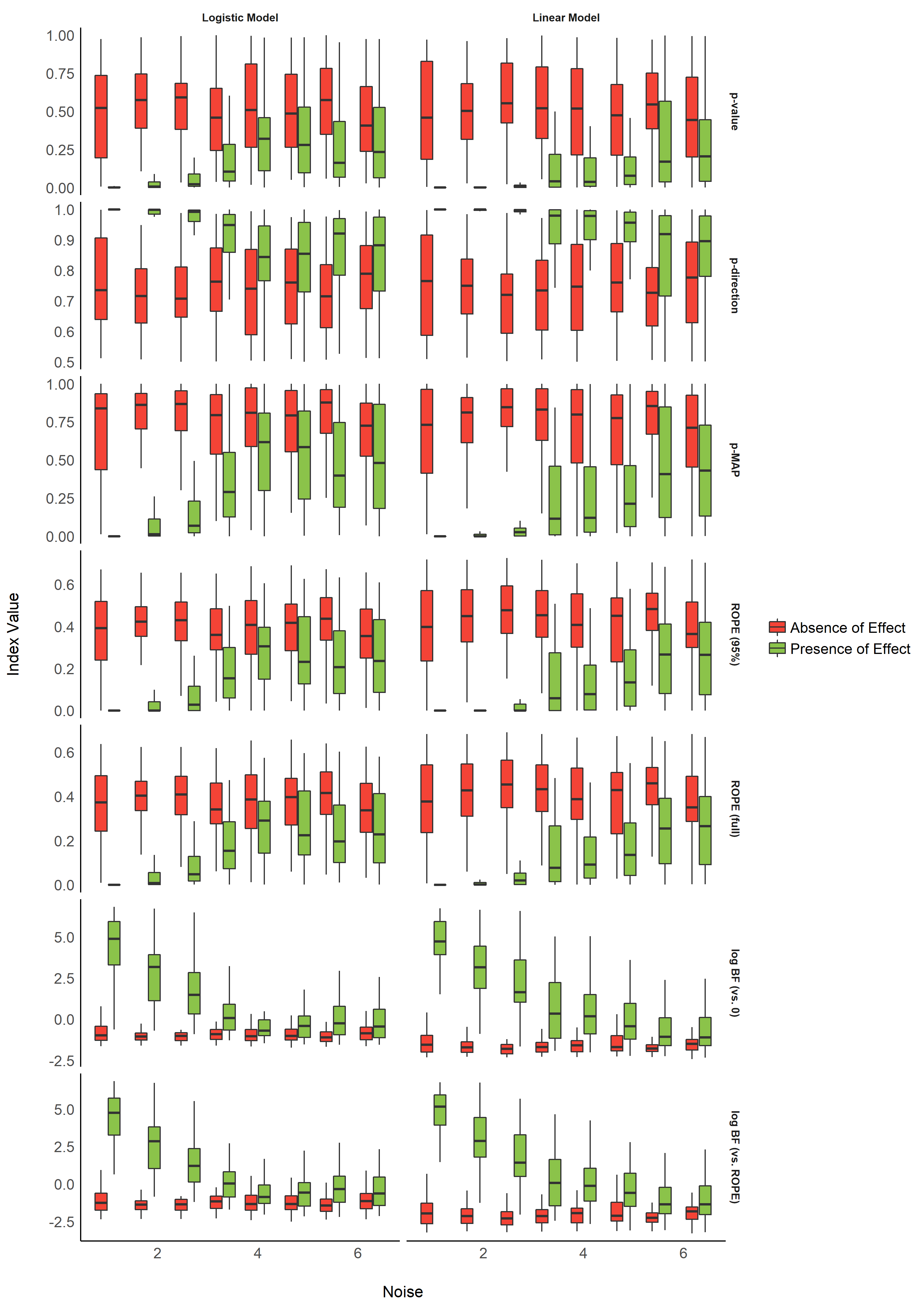


Figure 2. Impact of Noise.

Table 1. Performance comparison of the indices in predicting the presence of an effect.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model\_Type | Model | AIC | BIC | R2\_Tjur | RMSE | LOGLOSS | PCP | BF |
| linear | p\_value | 921 | 940 | 0.34 | 1.01 | 0.51 | 0.67 |  |
| linear | p\_direction | 919 | 938 | 0.34 | 1.01 | 0.51 | 0.67 | 2.3e+00 |
| linear | p\_MAP | 829 | 848 | 0.42 | 0.95 | 0.46 | 0.71 | 8.6e+19 |
| linear | ROPE\_95 | 800 | 819 | 0.44 | 0.94 | 0.44 | 0.72 | 1.5e+26 |
| linear | ROPE\_full | 795 | 814 | 0.45 | 0.94 | 0.44 | 0.72 | 1.8e+27 |
| linear | BF\_log | 742 | 761 | 0.47 | 0.90 | 0.41 | 0.74 | 6.8e+38 |
| linear | BF\_ROPE\_log | 741 | 760 | 0.47 | 0.90 | 0.41 | 0.74 | 1.0e+39 |
| binary | p\_value | 999 | 1018 | 0.27 | 1.05 | 0.55 | 0.63 |  |
| binary | p\_direction | 1000 | 1019 | 0.27 | 1.05 | 0.55 | 0.63 | 8.5e-01 |
| binary | p\_MAP | 931 | 950 | 0.33 | 1.01 | 0.51 | 0.66 | 6.4e+14 |
| binary | ROPE\_95 | 917 | 937 | 0.33 | 1.01 | 0.51 | 0.67 | 5.9e+17 |
| binary | ROPE\_full | 914 | 934 | 0.34 | 1.00 | 0.50 | 0.67 | 2.7e+18 |
| binary | BF\_log | 870 | 889 | 0.36 | 0.98 | 0.48 | 0.68 | 1.2e+28 |
| binary | BF\_ROPE\_log | 869 | 888 | 0.36 | 0.98 | 0.48 | 0.68 | 1.7e+28 |

For each index and each model type, we fitted a (frequentist) logistic regression to predict the presence or absence of effect, adjusted for noise and sample size. The comparison of the performance of these models (AIC, BIC and Tjur’s R2) revealed a consistent pattern accross model type (*i.e.*, similar for linear and logistic models), suggesting that *BF (vs. ROPE)* is the best index to discriminate between the presence and the absence of an effect, followed by *BF (vs. 0)*, *ROPE (full)*, *ROPE (95\_%)*, *p-MAP*, *p-direction* and the frequentist *p*-value. The Bayes factor (computed via BIC approximation (**REFERENCE**) against the frequentist *p*-value model), used here as a measure of relative performance, supported this conclusion.

## Relationship with the frequentist *p*-value

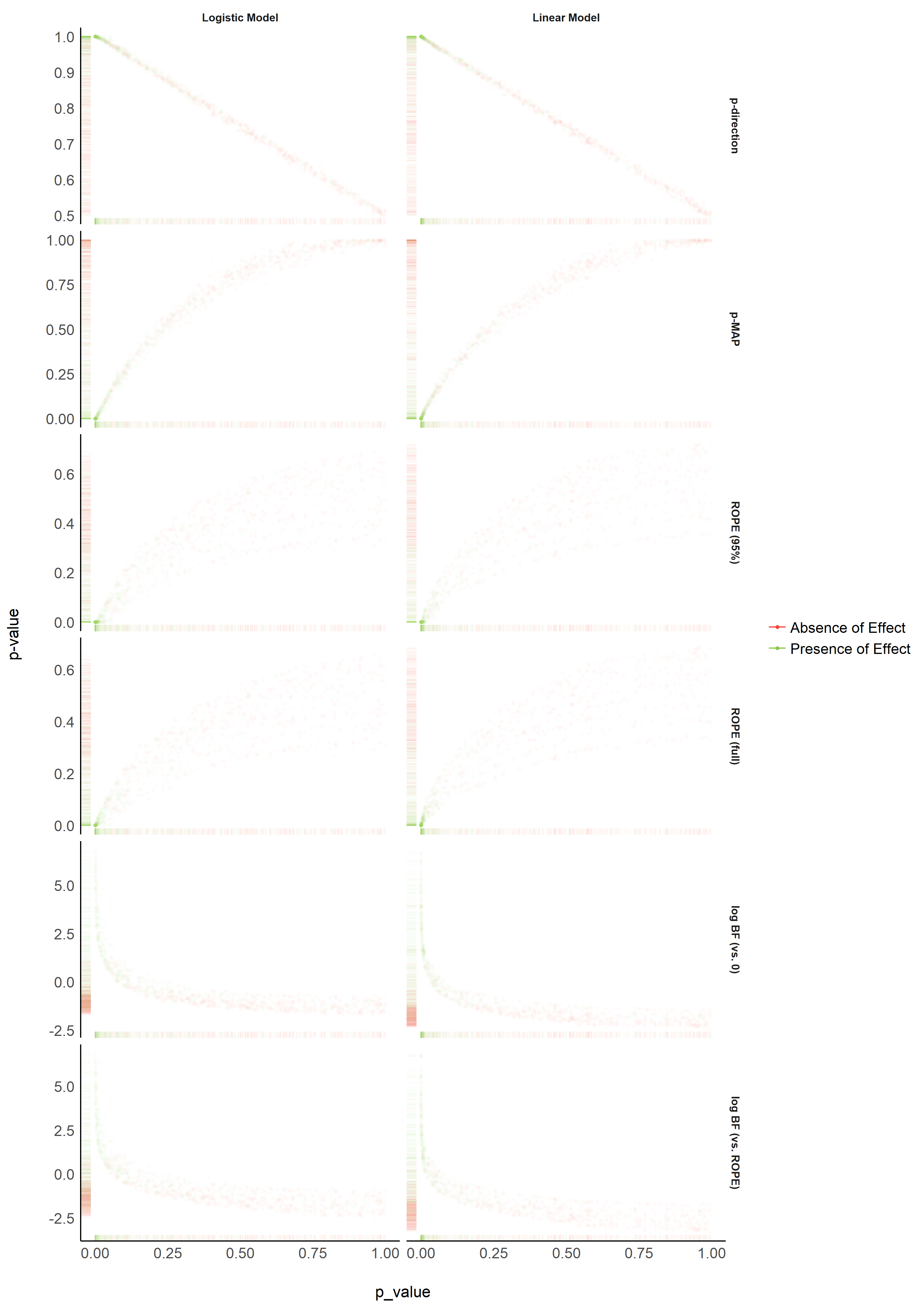


Figure 3. Relationship with the frequentist p-value.

The **Figure 3** suggests that the *pd* has a 1:1 correspondance with the frequentist *p*-value. *BF* indices still appear as having a strong relationship (altough severly non-linear) with the frequentist index, while *ROPE*-based percentages appear as being the most independent.

TODO: Add points type as sample size as ROPE seems to be directly related to groups of sample size.

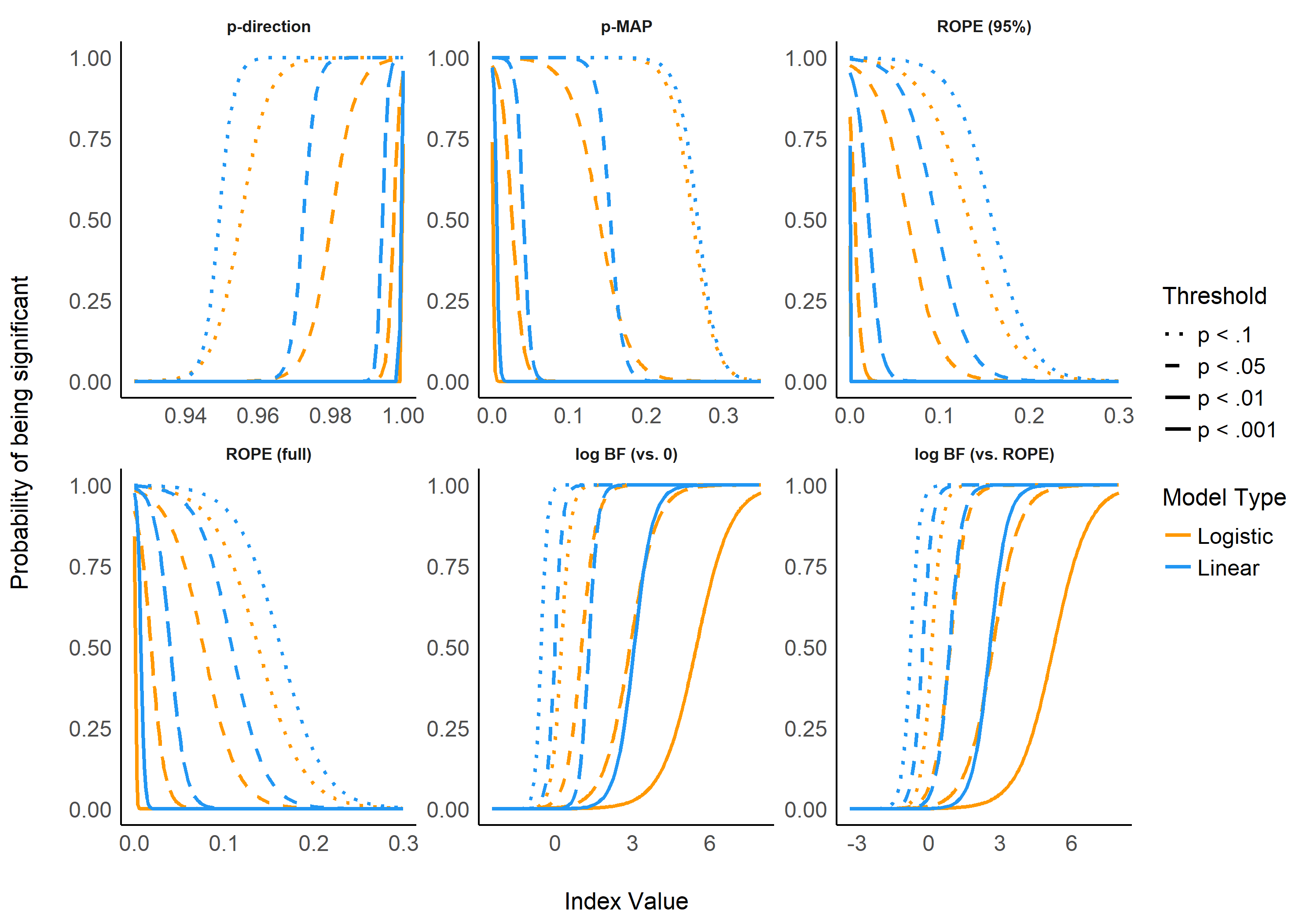


Figure 4. Relationship between Bayesian indices and frequentist arbitrary threshold of significance.

## Relationship between ROPE (full), pd and BF (vs. ROPE)

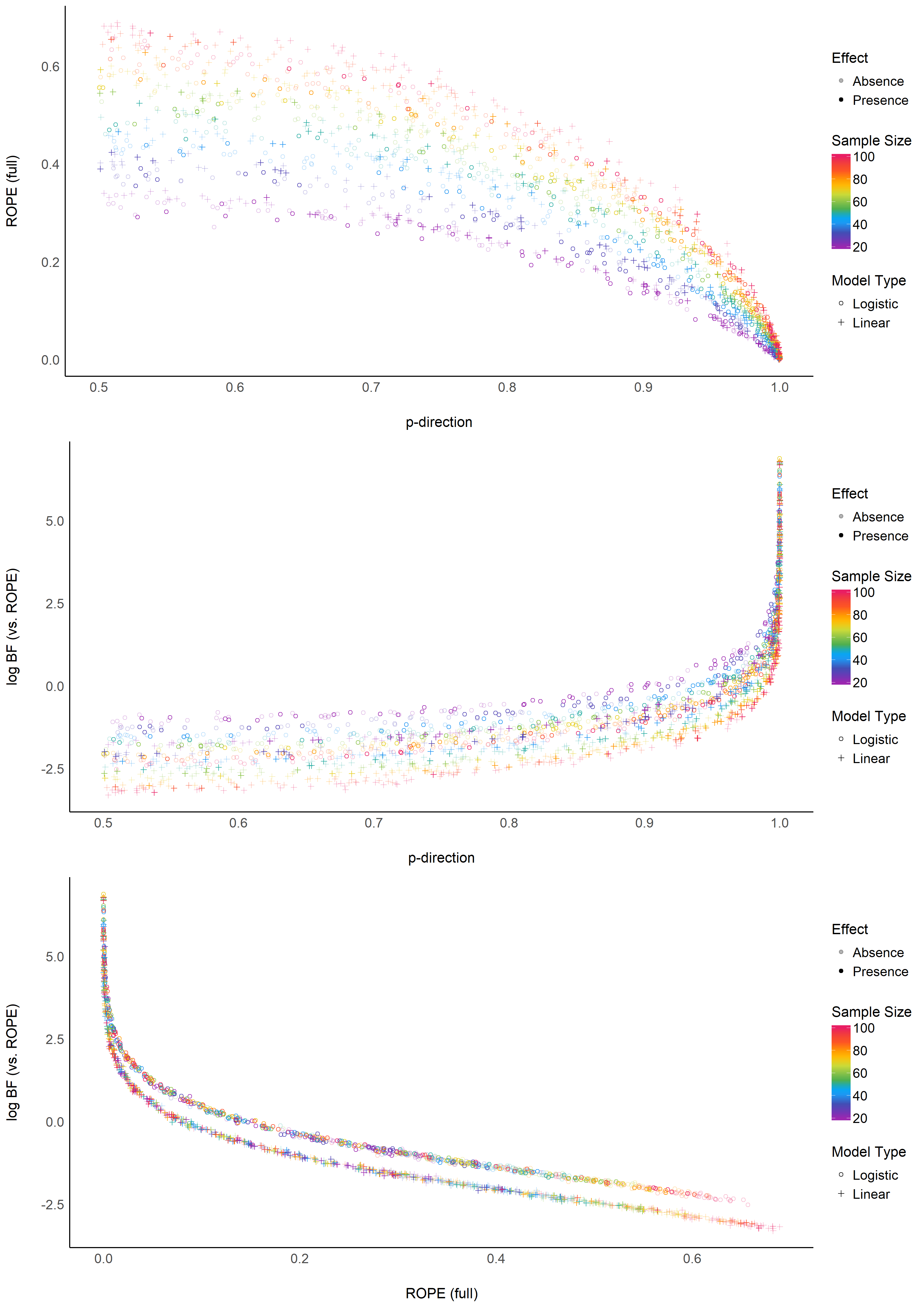


Figure 5. Relationship between three Bayesian indices.

> TableGrob (3 x 1) "arrange": 3 grobs  
> z cells name grob  
> 1 1 (1-1,1-1) arrange gtable[layout]  
> 2 2 (2-2,1-1) arrange gtable[layout]  
> 3 3 (3-3,1-1) arrange gtable[layout]

The **Figure 5** suggests that the relationship between the *ROPE (full)* and the *pd* might be strongly affected by the sample size, and the relationship between *BF (vs. ROPE)* and the *pd* might be subject to differences accross model types. Moreover, the *ROPE (full)* and the *BF (vs. ROPE)* seem very closely related within the same model type.

# 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.

# Conclusion and Limitations

Before being able to draw a definitive conclusion about the qualities of these indices, further studies need to investigate the robustness of these indices to sampling characteristics (*e.g.*, samplign algorithm, number of iterations, chains, warm-up) and the impact of prior specification.

One of the limitation the ROPE-based indices is that they require an explicit definition of the ROPE range. Importantly, as this range is fixed to the scale of the outcome response, these indices are thus sensitive to changes in the scale of the predictors.

Thus, in order to assess the existence and significance of effects within a regession model, we recommend to report, at minimum, the *pd* as an index of effect existence and the *BF (vs. ROPE)* as an index of significance. The former for its simplicity of interpretation, its robustness and its numeric proximity to the well-known frequentist *p*-value, and the latter for its ability to discriminate between presence and absence of effect, and the information it provides related to relative evidence of effect size.

# 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. (1988). *Statistical power analysis for the social sciences*.

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.

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.