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; Halsey, 2019; Marasini, Quatto, & Ripamonti, 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 (“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.*, the model parameters, influencing the outcome”), which conclusions often rely on some index of effect “significance”.

It is important to note that the importance and focus on significance assessment might differ accross fields. For instance, in applied physics or engineering, the goal of a study is often to precisely *estimate* (quantify) a given effect and its size (the question being, for example, wether a given coefficient is closer to 7 or to 9, or whether a difference is moderate or big). The Bayesian framework has demonstrated its superiorty in addressing this type of questions, its probabilistic framework leading to accurate estimations and allowing to quantify the inherent uncertainty associated with them. On the contrary, in other fields, such as for instance experimental psychology, the implicit aim has often been focused on *significance*: “are these two conditions different?”, “is there a correlation between these two variables?”. In other words, being able to conclude if a given parameter affects substantially the outcome with little care to its precise quantification.

Despite a general agreement on its defects and recurring waves of attacks occuring for decades (Gardner & Altman, 1986; anderson2000null; finch2004reform; fidler2004editors; Cohen, 2016), the tenacity and resilience of the *p*-value as an index of significance is remarkable, being still widely used and thaught (altough some journals have taken a radical step by banning them, *e.g.*, Gill (2018); expecting to shift the perspective to effect size estimation). This endurance might be informative on how such indices, and the accompanying heuristics to interpret them (*e.g.*, *.05*, *.01* and *.001*), are useful and necessary for researchers to gain an intuitive (although possibly simplified) understanding of the interactions and structure of their data. Moreover, this utility is salient in contexts where decisions have to be taken (*e.g.*, in medical settings). In spite of the statistical perspective, in which the absence of any hard mathematical thresholds or discrete categories naturally leads to the desire for a nuanced and subtle view, decisionners might often seek a threshold value to rationalize their acts. This practical need for significance assessment might be one of the mechanism fostering the wide adoption of the frequentist *p*-value and its arbitrary interpretation clusters. 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).

While significance testing (and its inherent categorical interpretation heuristics) might have its place alongside effect estimation as a complementary perspective, it does not preclude the fact that drastic improvements are needed. For instance, one possible advance could focus on improving the mathematical understanding (*e.g.*, through a new simpler index) of the values (as opposed to the obscure mathematical definition of the *p*-value that contributes to its common misintepretation). Another improvement could be found in providing an intuitive (*e.g.*, visual) understanding of the behaviour of the indices in relationship with main sources of variance, such as sample size, noise or effect presence. Such better overall understanding of the indices would hopefully act as a barrier against their mindless reporting by allowing the users to nuance the interpretations and conclusions that they draw.

So what does the Bayesian framework offer as alternatives for the *p*-value? Bayesian testing indices could be roughly grouped into three overlapping categories: Bayes factors, posterior indices and ROPE-based indices. Bayes factors reflect the predictive performance of one against another model (*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 and allowing to quantify the evidence in favour of the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). Nonetheless, its use for parameters description in complex models is still a matter of debate, being highly dependent on priors specification (Etz, Haaf, Rouder, & Vandekerckhove, 2018; Kruschke & Liddell, 2018). On the contrary, “posterior indices” reflect objective characteristics of the posterior distribution, such as for instance the proportion of strictly positive values. While the simplicity of their computation and interpretation is an asset, it could also be a limiting the information that they provide. Importantly, Bayes factors and indices derived solely from the posterior distribution are both the “natural, direct, and unavoidable consequence of Bayes’ rule” (Rouder, Haaf, & Vandekerckhove, 2018, p. 106). Finally, ROPE-based indices are related to the redefinition of the point-null hypothesis (Kruschke, 2014; Lakens, 2017) as a range of values (the Region of Practical Equivalence - ROPE), usually equally spread around 0 (*e.g.*, [-0.1; 0.1]) and considered as negligible (*i.e.*, too small to be of practical relevance). It is interesting to note that this perspective unites Bayesian indices with the focus on effect size (involving a discrete separation between at least two categories), which finds an echo in the recent statistical recommendations (Ellis & Steyn, 2003; Simonsohn, Nelson, & Simmons, 2014; Sullivan & Feinn, 2012).

Curiously, despite the richness provided by the Bayesian framework and the availability of multiple indices, no consensus has yet emerged on the ones to use. On the contrary, the litterature continues to bloom in a raging debate, often polarized between proponents of the Bayes factor as the supreme index and its detractors (*e.g.*, Spanos, 2013; Robert, 2014; robert2016expected; Wagenmakers, Lee, Rouder, & Morey, 2019), with strong theoretical arguments being developped on both sides. Unfortunately, no practical and direct comparison between these indices 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 of significance and their interpretation guidelines (in the form of rules of thumb) are useful in practice, we also strongly believe that they should be accompanied with the understanding of their “behaviour” in relationship with major sources of variance, such as sample size and noise. 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 “significance”, provide visual representations of the “behaviour” of such indices in relationship with sample size, noise and effect presence, as well as 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 finally draw recommendations for Bayesian statistics reporting.

# Methods

## Data Simulation

The simulation of datasets suited for linear or logistic regression, we started by simulating an independent, 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*) used here is based on prior and posterior distributions of a single parameter. In this context, the Baye factor 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), a Savage-Dickey density ratio was 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 (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). We also computed the BF of the posterior distribution against the range of negligible values (same as for the ROPE), by comparing the prior and posterior odds of the parameter falling within vs. outside the ROPE (see *Non-overlapping Hypotheses* in Morey & Rouder, 2011).

## 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.

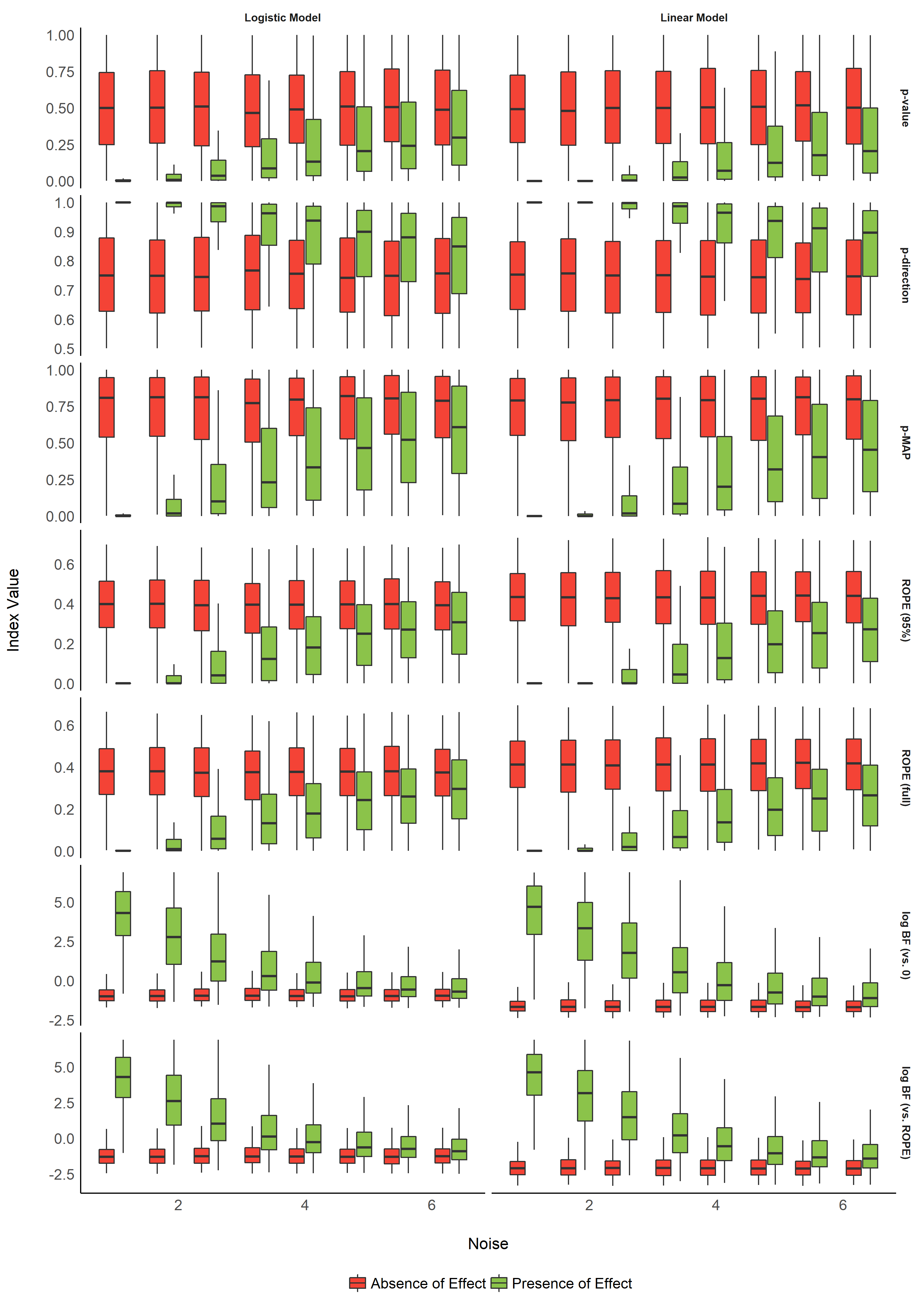


Figure 2. Impact of Noise.

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 | BF\_log |
| linear | p\_value | 18500 | 18531 | 0.33 | 1.01 | 0.51 | 0.66 |  |  |
| linear | p\_direction | 18461 | 18492 | 0.33 | 1.01 | 0.51 | 0.66 | 3.32e+08 | 19.62 |
| linear | p\_MAP | 16850 | 16882 | 0.40 | 0.97 | 0.47 | 0.70 | Inf | Inf |
| linear | ROPE\_95 | 16358 | 16389 | 0.42 | 0.95 | 0.45 | 0.71 | Inf | Inf |
| linear | ROPE\_full | 16234 | 16265 | 0.42 | 0.95 | 0.45 | 0.71 | Inf | Inf |
| linear | BF\_log | 15032 | 15063 | 0.45 | 0.91 | 0.42 | 0.73 | Inf | Inf |
| linear | BF\_ROPE\_log | 14984 | 15015 | 0.45 | 0.91 | 0.42 | 0.73 | Inf | Inf |
| binary | p\_value | 20486 | 20517 | 0.23 | 1.07 | 0.57 | 0.62 |  |  |
| binary | p\_direction | 20501 | 20532 | 0.23 | 1.07 | 0.57 | 0.62 | 4.78e-04 | -7.65 |
| binary | p\_MAP | 19200 | 19231 | 0.29 | 1.03 | 0.53 | 0.65 | 1.75e+279 | 642.98 |
| binary | ROPE\_95 | 19018 | 19049 | 0.30 | 1.03 | 0.53 | 0.65 | Inf | Inf |
| binary | ROPE\_full | 18953 | 18985 | 0.30 | 1.03 | 0.53 | 0.65 | Inf | Inf |
| binary | BF\_log | 17975 | 18007 | 0.33 | 1.00 | 0.50 | 0.66 | Inf | Inf |
| binary | BF\_ROPE\_log | 17934 | 17965 | 0.33 | 1.00 | 0.50 | 0.67 | Inf | Inf |

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 (against the frequentist *p*-value model, computed via BIC approximation Wagenmakers, 2007), used here as a measure of relative performance, supported this conclusion.

## Relationship with the frequentist *p*-value

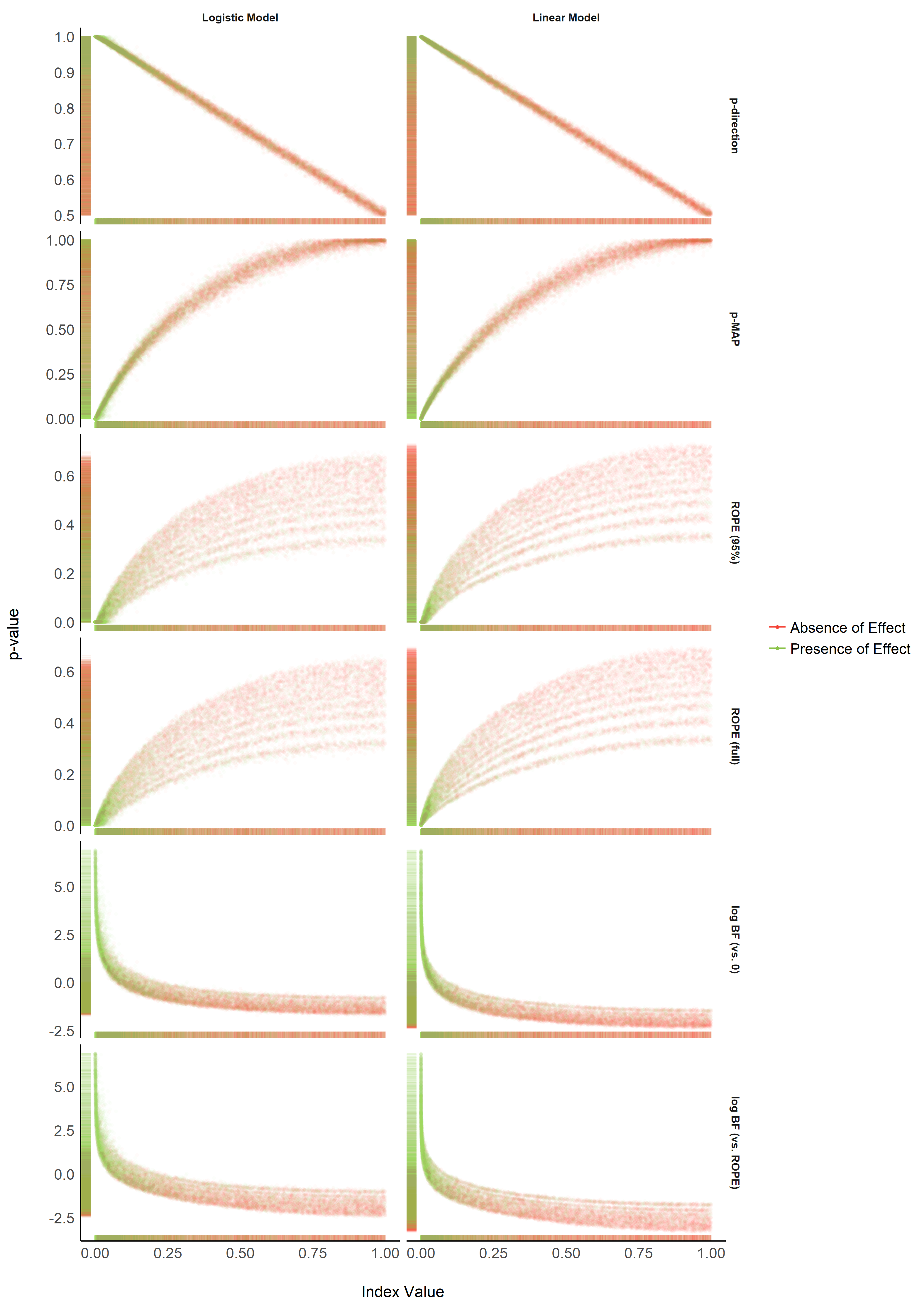


Figure 3. Relationship with the frequentist p-value.

**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, which is to be expected since the *BF* is consistant (i.e., it approaches the “true” bound as sample size increases) both when the null is true and when the alternative is ture (as can be seen in the marginal distibutions, marked by colored dashes in the plot margins), whereas the *p*-value is only consistant when the alternative is true, but has a uniform distribution [0-1] when the null is true (Rouder & Morey, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). *ROPE*-based percentages appear to be only weakly related to *p*-values.

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.

Table 2. Summary of Bayesian Indices of Effect Existence and Significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Interpretation | Definition | Strengths | Limitations |
| Probability of Direction (pd) | Probability that an effect is of the same sign as the median’s. | Proportion of the posterior distribution of the same sign than the median’s. | Straightforward computation and interpretation. Objective property of the posterior distribution. 1:1 correspondance with the frequentist p-value. | Limited information favouring the null hypothesis. |
| MAP-based p-value | Relative odds of the presence of an effect against 0. | Density value at 0 divided by the density value at the mode of the posterior distribution. | Straightforward computation. Objective property of the posterior distribution | Limited information favouring the null hypothesis. Relates on density approximation. Indirect relationship between mathematical definition and interpretation. |
| ROPE (95%) | Probability that the credible effect values are not negligible. | Proportion of the 95% CI inside of a range of values defined as the ROPE. | Provides information related to the practical relevance of the effects. | A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. Not sensitive to highly significant effects. |
| ROPE (full) | Probability that the effect possible values are not negligible. | Proportion of the posterior distribution inside of a range of values defined as the ROPE. | Provides information related to the practical relevance of the effects. | A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. |
| Bayes factor (vs. 0) | ??? | ??? | ??? | ??? |
| Bayes factor (vs. ROPE) | ??? | ??? | ??? | A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. |

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

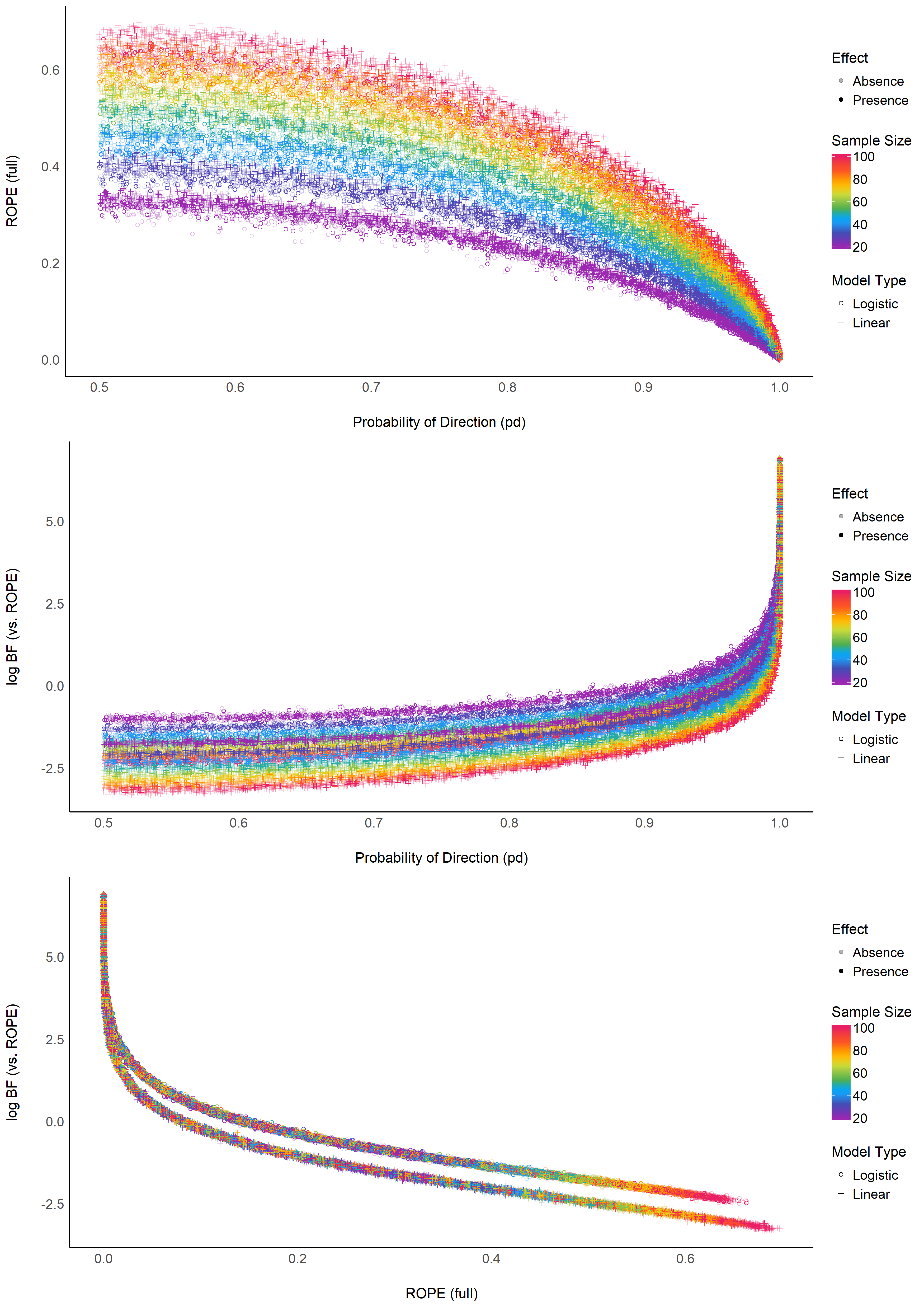


Figure 5. Relationship between three Bayesian indices.

> 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. These results reflect *ROPE (full)* and *BF (vs. ROPE)*’s consitancy both when the null is true and when the alternative is true, where the *pd*, being equvilant to the *p*-value, is only consistant when the null is true.

# 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.

: 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.

# 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 (Kass & Raftery, 1995; Kruschke, 2011), which might be concidered a feature of the Bayesian framework (Vanpaemel, 2010).

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 (De Santis, 2007), and the information it provides related to relative evidence of the size of the effect.

# Supplementary Materials

The full R code used for data generation, data processing, figures creation and manuscript compiling is available on Github at <https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian>.

# Acknowledgments

This study was made possible by the developpment of the **bayestestR** package, itself part of the [*easystats*](https://github.com/easystats/easystats) ecosystem, an open-source and collaborative project created to facilitate the usage of R. Thus, we would like to thank the [council of masters](https://github.com/orgs/easystats/people) of easystats, all other padawan contributors, as well as the users.

# 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.

De Santis, F. (2007). Alternative bayes factors: Sample size determination and discriminatory power assessment. *Test*, *16*(3), 504–522.

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.

Ellis, S., & Steyn, H. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): Research note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, *12*(4), 51–53.

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.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746–750.

Gill, J. (2018). Comments from the new editor. *Political Analysis*, *26*(1), 1–2. <https://doi.org/10.1017/pan.2017.41>

Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters*, *15*(5), 20190174.

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

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. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

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. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

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.

Marasini, D., Quatto, P., & Ripamonti, E. (2016). The use of p-values in applied research: Interpretation and new trends. *Statistica*, *76*(4), 315–325.

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.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406.

Robert, C. P. (2014). On the jeffreys-lindley paradox. *Philosophy of Science*, *81*(2), 216–232.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

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

Spanos, A. (2013). Who should be afraid of the jeffreys-lindley paradox? *Philosophy of Science*, *80*(1), 73–93.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282.

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

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lee, M., Rouder, J., & Morey, R. (2019, August). Another statistical paradox. Retrieved from <http://www.ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf>

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

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.