Towards a pragmatic use of statistics in ecology

Leonardo Castilho[*1] and Paulo Inácio Prado [†2]

[1]Graduate Program of Ecology, University of Brasília, Brazil
[2]LAGE at the Department of Ecology, Biosciences Institute, University of São Paulo , Brazil

July 10, 2018

Abstract

Although null hypothesis testing (NHT) is the primary method for analyzing data in many natural sciences, it has been increasingly criticized. Recently, a new method based on information theory (IT) has become popular and is held by many to be superior for many reasons, not least because is enables researchers to properly assess the strength of the evidence that data provide for competing hypotheses. Many studies have compared IT and NHT in the context of model selection and stepwise regression, but a systematic comparison of the most simple but realistic uses of statistics by ecologists is still lacking. We used computer simulations to compare how both methods perform

[*]**Corresponding Author:**
Universidade de Brasília, Faculdade de Planaltina
Área Universitária, 01, Vila Nossa Senhora de Fátima
Planaltina, Brasília - DF, Brasil, CEP 73345-010
leonardobcastilho@gmail.com
[†]Departamento de Ecologia, Instituto de Biociências, Universidade de São Paulo
Rua do Matão, Trav. 14, 321, Cidade Universitária
São Paulo - SP, Brasil, CEP 05508-090
prado@ib.usp.br

in four basic test designs (t-test, ANOVA, correlation tests, and multiple linear regression). Performance was measured by the proportion of simulated samples for which each method provided the correct conclusion (power), the proportion of detected effects with a wrong sign (S-error), and the mean ratio of the estimated effect to the true effect (M-error). We also checked if the p-value from significance tests correlated to a measure of strength of evidence, the Akaike weight. In most cases both methods performed equally well. The concordance is explained by the monotonic relationship between p-values and evidence weights in simple designs, which agree with analytic results. Our results show that researchers can agree on the conclusions drawn from a data set even when they are using different statistical approaches. By focusing on the practical consequences of inferences, such a pragmatic view of statistics can promote insightful dialogue among researchers on how to find a common ground from different pieces of evidence. A less dogmatic view of statistical inference can also help to broaden the debate about the role of statistics in science to the entire path that leads from a research hypothesis to a statistical hypothesis.

keywords: AIC, likelihood, model-based inference, M-error, S-error, null hypothesis testing, p-value, power, statistical inference, statistical misuse.

Introduction

Null hypothesis testing (NHT) has been the primary statistical method for drawing conclusions from data in natural sciences since, at least, the mid-1920s (Huberty, 1993). The purpose of NHT was originally to protect researchers from taking noise as a true effect (Mayo and Spanos, 2011; Gelman and Carlin, 2014). The probability of making such a mistake is gauged by the p-value calculated from a model of absence of effects (the null hypothesis). Accordingly, a model-driven definition of p-values was recently provided by the American Statistical Society: *"the probability under a specified statistical model that a statistical summary of the data (...) would be equal to or more extreme than its observed value"* (Wasserstein and Lazar, 2016). This is a statement on how compatible the data at hand is to a statistical hypothesis or model, but p-values are frequently misinterpreted as evidence about the models themselves (Cohen, 1994; Royall, 2000). In this sense, the most common misunderstandings are taking p-values as the probability that the null

hypothesis is true, as how improbable the alternative hypothesis is, or even as a measure of the effect strength (Cohen, 1994; Greenland et al., 2016; Wasserstein and Lazar, 2016). Thus, in an very broad sense, NHT is an error-control procedure that has been widely misused to express the support data provides for a given hypothesis or model. Although this criticism is not new (see Cohen, 1994), it has been recently used to challenge not only NHT, but the body of scientific knowledge that has been acquired using NHT (Ioannidis, 2005; Nuzzo, 2014).

More recently the information theoretic (IT) approach has been vigorously championed as a response to these problems. Recent reviews have presented IT as the proper approach to assess the support data gives to competing models, and popularized the criticisms to the NHT among biologists (Royall, 2000; Burnham and Anderson, 2002; Johnson and Omland, 2004). Within the IT framework, one ought to elaborate multiple competing hypotheses about the problem at hand and propose a statistical model to express each hypothesis. A measure of the relative information loss resulting of each of the competing model is then used to identify which models are the best approximations of the data at hand. The central concept is likelihood, which is any function proportional to the probability that a model assigns to the data. The likelihood function expresses the degree to which the data supports to each model, and inference in IT is used to find the best supported model (Edwards, 1972; Burnham and Anderson, 2002). Although this approach is more commonly used as framework of model selection, Burnham and Anderson (2002) advocate that its has a much broader range of uses, encompassing all types of NHT analyses done today. The authors state that NHT is a poor method of analyzing data and strongly emphasize the use of the IT approach for every type of analysis in the field of ecology.

These ideas elicited an intense discussion on the advantages of substituting traditional NHT by the IT approach in recent years. Many authors have compared both methods using real and simulated data (Whittingham et al., 2006; Glatting et al., 2007; Murtaugh, 2009; Freckleton, 2011; Lukacs et al., 2010), and raised philosophical and theoretical issues (Johnson and Omland, 2004; Steidl, 2006; Garamszegi et al., 2009); some have also identified theoretical or usage problems arising from the IT approach (Anderson and Burnham, 2002; Guthery et al., 2005; Lukacs et al., 2007; Mundry, 2011; Galipaud et al., 2014), while others highlighted the importance of maintaining both approaches in practice (Richards, 2005; Stephens et al., 2005, 2007).

Most of these works, however, have focused comparing the IT approach

to stepwise regression, a traditional technique for model selection in the NHT approach. Nevertheless, there is a long history of studies on the shortcomings of stepwise regression (Quinn and Keough 2002 but see Blanchet et al. 2008 for alternatives). Many traditional problems in NHT, particularly in stepwise regression, also appear when using the IT approach (Hegyi and Garamszegi, 2011; Mundry, 2011). Proponents of the new framework based on the IT approach, however, do not advocate the use of the method only as a substitute for traditional stepwise regression, or for more complex modeling situations. Instead, they view NHT as a whole as a poor method with much less inferential power than the IT approach (Anderson, 2008). To compare the IT approach to a technique already shown to be poor, as has been done for stepwise regression, is therefore not the most productive way to evaluate the advantages of the proposed new framework, if there are any. Furthermore, model selection is the home territory of IT approaches, while a much less appreciated topic is the use of the p-value to reject null hypotheses in standard NHT designs such as t-tests, ANOVA or linear regression. In these paradigmatic cases p-values are functions of the IT measures of evidence (Edwards, 1972; Murtaugh, 2014), suggesting that both approaches would lead to the same conclusions. However, as both approaches rely on asymptotic theories we need to check their congruence with the sample sizes usual for each knowledge area. The analytic correspondence of the IT and NHT approaches reveal important differences in asymptotic convergence in some simple cases (Figure A-1). For more complicated designs in realistic situations, computer simulations offer a straightforward way to make such comparisons.

Meanwhile, despite all the criticisms it has attracted, NHT is still widely used and taught in many research disciplines, including ecology (Stanton-Geddes et al., 2014; Touchon and McCoy, 2016; Wasserstein and Lazar, 2016). By progressing while using a technique that is the subject of many criticisms, researchers might be demonstrating that they can achieve their goals without worrying too much about philosophical controversies (Mayo and Cox, 2006). Practicing scientists might therefore feel they can agree on the conclusions drawn from a data set even if they use different statistical approaches. One obvious reason for such a pragmatic agreement is the equivalence of the conclusions using the NHT and IT approaches. The aim of this study is to test such agreement using computer simulations to run pragmatic comparisons of the NHT and the IT methodologies in standard, realistic designs for ecological studies.Pragmatic criteria assign equivalence to any outcome of equal practical consequence, despite differences in the causes (Hookway,

2016). As with any phenomenological approach, pragmatic conclusions are context-dependent and so the context must be clearly stated. Therefore, our main question is whether statistical approaches that differ in theory can lead to the same conclusion under the realistic conditions often seen in the field or in the laboratory and in cases in which both approaches are possible. Specifically, we asked if there is any difference in the conclusions drawn from data traditionally analyzed with t-tests, ANOVA, correlation tests, or linear regression when analyzed with the IT approach. We also assessed whether the use of p-values to express the strength of evidence of the conclusions led to incorrect evaluations of the support provided by the data.

Hereafter we will call a rightful or correct conclusion a result from a statistical analysis that accords with the true mean difference or relationship between variables. The probability of detecting such effects (the test power), is the usual means of gauging how frequently significance tests produce accurate conclusions. Nevertheless, a significant effect can still lead to a wrong conclusion because the estimated effect can have the opposite sign or an inflated magnitude in the sample. Gelman and Carlin (2014) defined these errors as *type-S* and *type-M* and showed that their rates increase as the test power decreases. We thus combined test power, type-S and type-M errors to evaluate the performance of the IT and NHT approaches in providing accurate conclusions regarding statistical effects .

Finally, we did not address the issue of the biological relevance of an effect that was correctly detected, because we assume this task is beyond the purpose of statistics and should be left to researchers. In many situations, procedures such as model averaging and effect size statistics can also be used to enhance the predictive power of models and to support the process of drawing conclusions and making decisions, but assessing how such *post hoc* procedures could improve a statistical method is beyond the scope of this article. Here, we deal with the initial values guiding drawing conclusions (*i.e.*: p-values and AICcs), as any statistical procedure done at a later stage would be conditional on those values. Those interested in how model averaging can enhance predictive power and how this relates to more traditional techniques are directed to Freckleton (2011).

Methods

We compared NHT and IT approaches for four standard and common analysis designs in ecology: (i) unpaired t-test design; (ii) single-factor ANOVA design; (iii) correlation design between two variables; and (iv) multiple linear model design. For each one of these designs, we sampled values from the distributions assumed by each design (univariate Gaussian for t-tests and ANOVA, and bi-variate Gaussian for correlation tests and multiple linear regressions, details below). We then performed NHT and IT procedures with the simulated samples and compared the results of each with regard to the probability of achieving a correct conclusion, and the magnitude of M-errors and S-errors (*sensu* Gelman and Carlin, 2014, , details below).

In all cases the simulated samples were defined by three parameters: the standard deviation of the sampled Gaussian distributions, the size of the samples, and the true effect size. The true effect is the value of the statistic of interest (e.g., the t-value or the correlation coefficient) that would be observed in an infinitely large sample (Gelman and Carlin, 2014). In our simulations the true effect was defined by the parameters of the sampled Gaussian distributions (e.g., the difference between the means of the two sampled Gaussian for t-test). We standardized the true effects on sample standard errors to make effects comparable across studies and designs (Lipsey and Wilson, 2001). The expressions for these standardized true effect size (henceforth used interchangeably with "effect size" or simply "effect") for each analysis design are provided below in the descriptions of the simulations of each design.

We used Latin hypercube sampling to build 2,000 combinations of parameters and sample sizes from uncorrelated uniform distributions (Chalom and Prado, 2016). The sample sizes ranged from 10 to 100 and effect sizes and standard deviations ranged from 0.1 to 8. Thus, our combinations are hypercube samplings of parameter spaces that cover typical sample sizes of studies in ecology, and low to medium effect sizes within a wide range of variation of data distributions. For each of these combinations we repeated the simulation 10,000 times. We also repeated the same procedures to run 10,000 simulations with 2,000 unique combinations of standard deviations and sample sizes for the case of zero effect size, in order to simulate a situation when the null hypothesis was true.

For every analysis simulation we extracted the proportion of the simulations that yielded a correct conclusion. In the NHT approach, such measure-

ments were the proportion of analyses that resulted in a p value lower than 0.05 if H0 was false, and higher than 0.05 otherwise. For the IT approach, we fit by maximum likelihood (i.e. by minimizing the sum of squares of residuals) those linear Gaussian models that express the null and alternative hypothesis for each design, as detailed below. We then considered a correct conclusion if the model that expressed the correct hypothesis was selected. To decide which model select, however, we took into consideration the bias of AIC to select models with uninformative parameters (Teräsvirta and Mellin, 1986). This problem arises when the true model is included in the selection procedure, along with models that provide the same fit but have additional uninformative parameters (Aho et al., 2014). As this is the case in our simulations for ANOVA and linear regression (see below and in Appendices), we identified the model with fewer parameters that was among the models with $\Delta AIC < 2$ chosen using the IT approach (Arnold, 2010).

To estimate the S-error rate and M-error size (Gelman and Carlin, 2014), from each approach we used the subset of simulations in which some effect was detected by NHT (that is, in which the null hypothesis was rejected) or by IT (that is, in which the selected model included parameters related to some effect or difference among means). We estimated type-S error rate as the proportion of this subset in which the detected effect had the opposite sign of the true effect. The expected type-M error was estimated as the mean ratio between the estimated effects and the true effect value, as in the subset of simulations defined above.

In the appendices we have provided the functions in R (R Development Core Team, 2016) that we created to run the simulations and the R scripts of all simulations and analyses.

t-test designs

We simulated an unpaired t-test design by drawing samples from two Gaussian distributions that differed in their means by a certain amount, but had the same standard deviation. One of the distributions had a mean of zero. The effect size was the true t-value, which in this case is:

$$E_t = \frac{\mu}{\sigma}\sqrt{\frac{2}{N}} \tag{1}$$

where μ is the distribution mean which is allowed to be different from zero, and σ and N are the common standard deviations of both distributions

and common samples sizes, respectively.

For the NHT approach, we calculated the t-value estimated from the samples and its corresponding p-value. For the IT approach, we calculated the Akaike Information Criterion corrected for small samples (AICc, Burnham and Anderson, 2002) of the two Gaussian linear models that express the null hypothesis and the alternative hypothesis. We recorded as correct conclusions of NHT the simulations in accordance with the simulated situation, that is, the simulations in which $p < 0.05$ when $E_t \neq 0$ and the simulations in which $p > 0.05$ when $E_t = 0$. Accordingly, we recorded as correct conclusions of IT the simulations in which the model that expressed the wrong statistical hypothesis had $\Delta AICc > 2$.

ANOVA designs

We simulated three samples from Gaussian distributions to represent measures obtained from three experimental groups. All distributions had the same standard deviations, but the true mean of one of the distributions differed by a certain amount from the mean of the other two distributions, which was set to zero. We expressed the true effect size in this case as an extension of the true t-value:

$$E_{ANOVA} = \frac{\mu}{\sigma}\sqrt{\frac{3}{N}} \tag{2}$$

For the NHT approach, we used the F-test to test the null hypothesis. If the null hypothesis was rejected, a post-hoc Tukey's test was used. For the simulations when the true difference was not zero, the conclusion was considered correct only if the three p-values of the Tukey's test agreed with the simulation. No Tukey's test was used when the difference between means was set to zero, as any significant F-test in such situation invariably leads to a wrong conclusion. In those situations, a non-significant F-test was considered a correct conclusion and a significant one was considered a wrong conclusion.

For the IT approach we fit five linear Gaussian models to express all possible statistical hypotheses regarding the differences among the three experimental groups. Of these models, one had a single parameter representing the means of all the groups expressing the null hypothesis; three had two parameters for means, which allowed one group mean to be different from the other two; and one had a single parameter signifying that each group mean expressed the hypothesis that all group means differed. The values of AICc

for each model were then calculated and we took as a correct conclusion the simulations in which the selected model agreed with the simulated situation.

Correlation designs

For the correlation design, samples of paired variables were taken from a bivariate normal distribution with correlation parameter ranging from zero to positive values. The true effect expression in this case was the correlation parameter expressed as a t-value (Lipsey and Wilson, 2001):

$$E_r = \rho\sqrt{\frac{N-2}{1-\rho^2}} \tag{3}$$

where ρ is the correlation of the bivariate Gaussian distribution from which the samples were drawn.

For the NHT approach, the p-value of the Pearson correlation coefficient of the two variables were calculated. For the IT approach, two models were fit. The first corresponded to the null hypothesis that the paired values come from a bivariate normal distribution with correlation parameter set to zero. The alternative hypothesis was represented by a model of a bivariate normal distribution with the correlation as a free parameter.

Multiple linear model designs

The multiple linear model designs had three variables: the response variable (Y), and two uncorrelated predictor variables (X_1 and X_2). The response variable was a linear function of X_1 plus an error sampled from a Gaussian distribution with zero mean and standard deviation σ:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad , \qquad \epsilon \sim N(0, \sigma)$$

The true effect size can thus be expressed as the standardized linear coefficient of X_1:

$$E_\beta = \frac{\beta_1}{\sigma}\sqrt{N} \tag{4}$$

For the NHT approach we fitted a multiple linear regression including the additive effect of X_1 and X_2 and calculated the p-values of the Walt statistics to test the effect of each predictor. The probability of correct conclusions

was estimated by the proportion of simulations that yielded a p-value for the X_1 corresponding to the correct hypothesis and a non-significant p-value for the X_2 variable. For the IT approach, four models were fit. The first was an intercept-only model where the expected value of the response Y is constant. This model corresponds to the null hypothesis of absence of effect of X_1 and X_2. The other models included only the effect of X_1, only the effect of X_2, or the effects of both predictor variables. The values of AICc for each model were then calculated and we took as a correct conclusion the simulations in which the selected model was in accordance with the simulated situation. To check the effect of collinearity we repeated the simulations above forcing a correlation of 0.5 between X_1 and X_2. As the results did not show any important difference we included this additional analysis in the appendices (See section A.2).

Measuring evidence through p-values

We also explored the relationship between the p-value and the Akaike weight (w), which is proposed as a true measure of strength of evidence (Burnham and Anderson, 2002). Ultimately, we wanted to check if there is a pragmatic disadvantage in considering lower p-values as "less evidence of the null hypothesis". A monotonic positive relationship between the p-values and the evidence weights for the model that express the null hypothesis (w_{H0}) would imply no pragmatic disadvantage. To check the relationship between the p-value and w_{H0}, we recorded both values for each simulation, for all four designs.

Results

Significance, power, S-errors and M-errors

When the null hypothesis was correct, the NHT approach achieved the nominal probability of type-I error ($\alpha = 0.05$) for the t-test, ANOVA, and correlation designs and a value close to $\alpha = 0.1$ for the linear regression. The IT approach performed slightly better in the t-tests, correlation and linear regression (Table 1).

When the null hypothesis was wrong the average proportion of correct conclusions in the simulations was used to estimate the test power β. For all

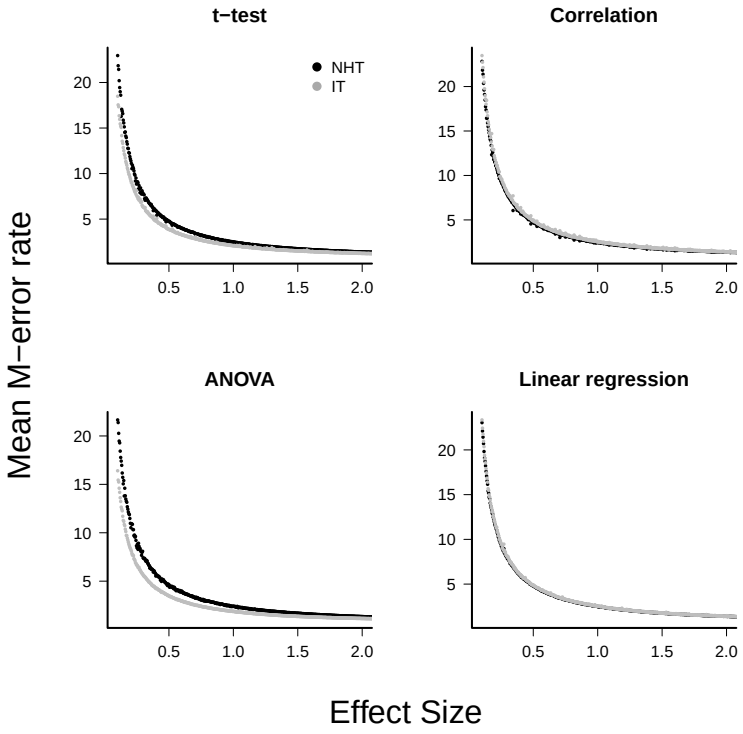

Figure 1: The mean type-M error (exaggeration rate) of null hypothesis tests (NHT, black) and information-based model selection (IT, grey) for each testing design, as a function of effect size. Each point represents the simulations of a test instance from which an effect was detected. Each test instance used a different combination of effect size, standard deviation of the values and sample size and was simulated 10,000 times. The M-error is the absolute ratio between the estimated effect size and the true effect size (Gelman and Carlin, 2014), which was estimated from the mean of this ratio for each test instance.

cases where the NHT approach was used the power was less than $\beta = 0.2$ for effect sizes below one (that is, effects less than one standard error), and achieved $\beta = 0.8$ for effect sizes of about 2.8 (Figure 2), as expected for the Gaussian distribution (Gelman and Carlin, 2014). The IT approach produced larger estimated power for small effect sizes in the t-test and ANOVA designs, but in all cases the power of both approaches converged to $\beta \approx 1$ as the effect

Table 1: Proportions of type-I error in the simulations, for the Null Hypothesis Tests (NHT) and Information-based model selection (IT).

	NHT	IT
t-test	0.050	0.044
Correlation	0.050	0.012
ANOVA	0.050	0.112
Regression	0.097	0.091

size approaches 4.0 standard error units (Figure 2). The IT approach only achieved this convergence with the additional parsimony criteria to discard models with uninformative parameters (see Supplementary Information A.2).

The exaggeration rate or type-M error was at least 2.0 when effect sizes were about 1.5 standard error units for all test designs and approaches (Figure 1) and M-errors increased steeply as the effect size decreased. Thus when an effect below 1.5 was detected the true value was exaggerated at least twofold. The IT approach had a slightly lower type-M error than NHT for the t-test and ANOVA designs for effect sizes below 2. Type-S error decreased more abruptly than the M-error with the increase of effect size (Figure 3). In our simulations the probability that the detected effect is of the wrong sign was of some concern (larger than 0.1) for effect sizes well below unit, but the IT approach had slightly larger S-errors at this range than the NHT for the t-test and ANOVA designs ((Figure 3). In all test designs and both the IT and the NHT approaches S and M errors vanished for effect sizes greater than 1.5 and 2.0, respectively. Collinearity in the linear regression design did not change none of the patterns described above (Supplementary Information A.2).

p-value as a measure of strength of evidence

The relationship between p and evidence weight w was positive and monotonic as expected. Simulations with larger effect sizes resulted in a small p-value and small evidence weight for the null hypothesis. Within the range of $0 < p < 0.1$, there was little variation around the trend, despite the wide range of parameters sampled by the hypercube and used in the simulations (Figure 4).

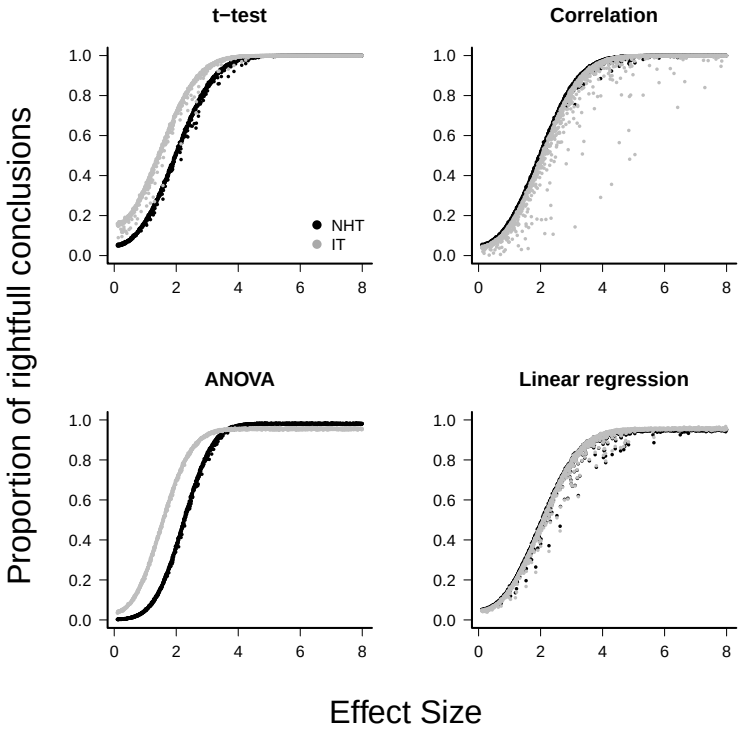

Figure 2: The power of null hypothesis tests (NHT, black) and the information-based model selection (IT, grey) for each testing design, as a function of effect size. Each point is the proportion of the 10,000 simulations of a test instance from which the effect was detected. Each test instance used a different combination of effect size, standard deviation of the values, and sample size.

Discussion

Comparing the performance of the two approaches

For all designs we have simulated the response of the power, the S-error and M-error to the standard effect size were very similar in the NHT and IT approaches. The increase in power with effect size for NHT is well known and expected from the consistency of estimators of the Gaussian distribution (*e.g.* Edwards, 1972) behind these tests. The lack of consistency of *AIC* when alternative models with uninformative parameters are considered has

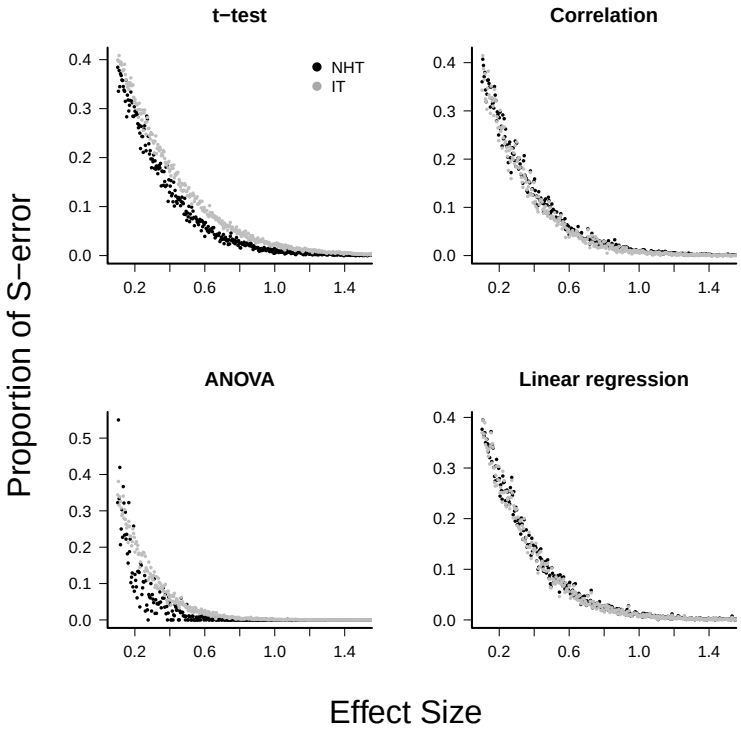

Figure 3: Mean type-S error of Null Hypothesis Tests (NHT, black) and information-based model selection (IT, grey) for each testing design, in function of effect size. Each point represents the simulations of a test instance from which an effect was detected. Each test instance used a different combination of effect size, standard deviation of the values and sample size and was simulated 10,000 times. The S-error is the probability of detecting an effect of an opposite sign of the true effect (Gelman and Carlin, 2014). For each test instance we estimated S-errors from the proportion of simulations that detected an effect of the opposite sign.

been highlighted recently (Aho et al., 2014), but this was easily circumvented with the additional parsimony criteria proposed by Arnold (2010). The relationship between S and M errors to test power (and thus to effect size) has, to date, received far less attention. We have shown that this relationship for all four of the Gaussian designs simulated agrees with those predicted by approximating the distribution of effects to a t-distribution (Gelman and

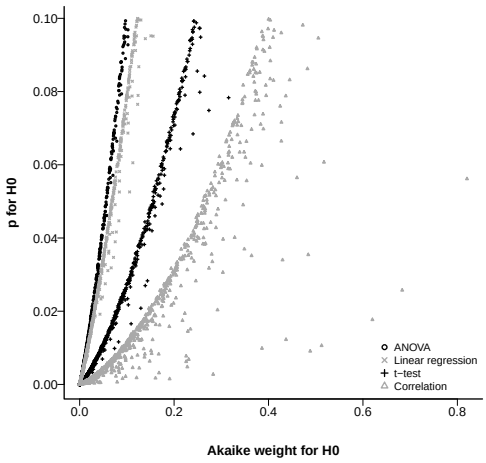

Figure 4: Relationship between the Akaike weight of the null model and the p-value of the null hypothesis found in simulations of each analysis design.

Carlin, 2014). As the standard effect size increases (and consenquently the power) both S and M errors decrease steeply, but S errors are a concern for effect sizes below one standard error, which in our simulations correspond to a test power between 0.05 (ANOVA, NHT) to 0.34 (t-test, IT). Accordingly, the exaggeration rate (M error) of twice or more occurred when the effect size is below 1.5 standard error units, which corresponds to a power value between 0.16 (ANOVA, NHT) and 0.53 (t-test, IT). These results also showed that the greater power and lower M-error mean rate of IT at small effect sizes for the t-test and ANOVA come at the cost of an increased ratio of S-error.

In summary, the performance of IT and NHT were very similar and converged quickly as standard effect sizes increased, which is caused by an increase in raw effect sizes, sample sizes or a decrease in standard errors. All these factors increase power, and is largely recognized that different inference criteria lead to the same conclusions as power increases (Gelman and Carlin, 2014; Ioannidis, 2005; Button et al., 2013). Focusing on how to obtain the best estimates of the effects can thus be a more effective contribution to scientific advancement than to dispute the value of weak inferences obtained with different statistical approaches (e.g. Gelman and Loken, 2014). Effect estimates can be improved by increasing sample sizes or controlling error

sources. Our results suggest that test designs should target a standard effect size of 2.8, which correspond to a test power above 0.8.

Moreover, statistical inference relies on the full sequence of events and decisions that led to the conclusions presented, which is not just the statistics used nor their power in a given sampling or experimental design (Gelman and Loken, 2014; Greenland et al., 2016). The choice of different inference approaches is only a part of this problem but it has dominated the debate. It might be the time to broaden our concerns to address the whole path that leads from a research hypothesis to a statistical hypothesis to be evaluated (Gelman, 2013).

p-value as a measure of strength of evidence

One of the strongest arguments recently given by the major proponents of the IT approach against the more traditional NHT is that the p-value is not a measure of strength of evidence in favor of the null hypothesis. Nevertheless, there is a widespread interpretation of p-values as "more significant" (*i.e.* less supportive of the null hypothesis) the lower they get. Furthermore, although an ecologist would hardly discard the null hypothesis based on a p-value higher than 0.1, values between 0.1 and 0.05 are usually interpreted as moderate evidence against the null hypothesis (Murtaugh, 2014). The relationship between the p-value and other statistics taken from the IT approach has been demonstrated before for the case of nested models in which the sample size is large enough to apply the log-likelihood ratio test (LRT) (Murtaugh, 2014; Greenland et al., 2016). We extended this conclusion for the simple designs we evaluated without the assumptions of LRT. The relationship between p-value and Akaike weights is monotonic positive and was poorly affected by variations of the simulations, specially at the borderline of significance.

Other authors have already pointed out that for many simple cases there is a monotonic relationship between p-values and likelihood ratios and thus to evidence weights (Edwards, 1972; Royall, 2000), by translating standard significance tests into alternative models with different parameter values (*e.g.* Figure A-2). Therefore we argue that the interpretation of p-values as measures of evidence, although conceptually wrong (Edwards, 1972; Cohen, 1994; Royall, 2000), can be empirically useful at least for the standard significance test designs.

Concluding remarks

We compared null hypothesis testing and information-theoretic approaches in situations commonly found by ecologists, considering sample sizes and correlation degrees often reported in ecological studies and only focusing on the important practical issues of both methods. The few differences between IT and NHT showed a trade-off between M and S errors and vanished as the effect size increases. We also showed that, at the borderline of significance in standard procedures with Gaussian errors, p-values can be used as a very good approximation of a measure of evidence to the null hypothesis when compared to the alternative.

The recent statement that NHT is always an outdated method for analyzing data is not supported by our findings. The basic NHT designs we analyzed have been the basis of data analyses for generations of ecologists, and still prove to be valuable in the context they were created (Gelman, 2013; Stanton-Geddes et al., 2014). They would only be outdated if the contexts of t-tests, ANOVA, correlations, and regressions designs did not exist anymore, which is, to date, obviously not true. Many criticisms to NHT are valid, but the new IT approach has also been correctly criticized and some of those criticisms are even the same as the ones used to justify NHT as an outdated technique (Arnold, 2010; Freckleton, 2011; Hegyi and Garamszegi, 2011; Richards et al., 2011) Besides, for those uncomfortable with the NHT technique, the IT technique is not the only alternative. Several alternative methods, all with their own pros and cons, have been proposed (Hobbs and Hilborn, 2006; Garamszegi et al., 2009).

As in any simulation study, the generality of the differences found in the performance of NHT and IT cannot be afforded beyond the parameter space that we have explored. Nevertheless, the simulations show that insisting on the absolute supremacy of a given approach is pointless, at least from a pragmatic perspective that seeks agreements in the findings despite the statistics used. We have shown a simple instance of agreement in which two statistical approaches in many situations lead to the same conclusions. In this case we elucidated the causes of the few divergences found, as well as the simple mathematical relationships that explain the concordances. Whether agreements are also possible by other means and in more complex designs remains to be evaluated. Nevertheless, by focusing on the consequences of a given result, a pragmatic view of statistics has a greater potential to find a common ground from different pieces of evidence and to promote a more

insightful dialogue between researchers.

Acknowledgements: Our thanks to Eduardo dos Santos Alves and Tadeu Siqueira for their keen suggestions to previous versions of this paper. PIP has research grant 2013/19250-7 of São Paulo Research Foundation (FAPESP) and scientific productivity grant from CNPq.

References

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology, **95**:631–636.

Anderson, D. R. 2008. Information Theory and Entropy. Springer, New York.

Anderson, D. R. and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. The Journal of Wildlife Management, pages 912–918.

Arnold, T. W. 2010. Uninformative parameters and model selection using akaike's information criterion. The Journal of Wildlife Management, **74**:1175–1178.

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. Ecology, **89**:2623–2632.

Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multimodel Inference - A Practical-Theoretic Approach. Springer-Verlag.

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, **14**:365–376.

Chalom, A. and P. I. Prado. 2016. pse: Parameter space exploration with Latin Hypercubes. R package version 0.4.6. `https://cran.r-project.org/package=pse`.

Cohen, J. 1994. The earth is round ($p < .05$). Am. Psychol., **49**:997–1003.

Edwards, A. W. F. 1972. Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference. Cambridge University Press, Cambridge.

Freckleton, R. P. 2011. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behavioral Ecology and Sociobiology, **65**:91–101.

Galipaud, M., M. A. Gillingham, M. David, and F.-X. Dechaume-Moncharmont. 2014. Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. Methods in Ecology and Evolution, **5**:983–991.

Garamszegi, L. Z., S. Calhim, N. Dochtermann, G. Hegyi, P. L. Hurd, C. Jørgensen, N. Kutsukake, M. J. Lajeunesse, K. A. Pollard, H. Schielzeth, et al. 2009. Changing philosophies and tools for statistical inferences in behavioral ecology. Behavioral Ecology, **20**:1363–1375.

Gelman, A. 2013. Commentary: P values and statistical practice. Epidemiology, **24**:69–72.

Gelman, A. and J. Carlin. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, **9**:641–651.

Gelman, A. and E. Loken. 2014. The statistical Crisis in science. American Scientist, **102**:460–465.

Geweke, J. and R. Meese. 1981. Estimating regression models of finite but unknown order. International Economic Review, pages 55–70.

Glatting, G., P. Kletting, S. N. Reske, K. Hohl, and C. Ring. 2007. Choosing the optimal fit function: comparison of the akaike information criterion and the f-test. Medical physics, **34**:4285–4292.

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, **31**:337–350.

Guthery, F. S., L. A. Brennan, M. J. Peterson, and J. J. Lusk. 2005. Information theory in wildlife science: critique and viewpoint. Journal of Wildlife Management, **69**:457–465.

Hegyi, G. and L. Z. Garamszegi. 2011. Using information theory as a substitute for stepwise regression in ecology and behavior. Behavioral Ecology and Sociobiology, **65**:69–76.

Hobbs, N. T. and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications, **16**:5–19.

Hookway, C. 2016. Pragmatism. In E. N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, summer 2016 edition.

Huberty, C. J. 1993. Historical origins of statistical testing practices: The treatment of fisher versus neyman-pearson views in textbooks. The Journal of Experimental Education, **61**:317–333.

Ioannidis, J. P. 2005. Why most published research findings are false. PLoS Med, **2**:e124.

Johnson, J. and K. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution, **19**:101–108.

Lipsey, M. W. and D. B. Wilson. 2001. Practical meta-analysis, volume 49 of *Applied Social Research Methods Series*. Sage publications, Thousand Oaks, CA.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and freedman's paradox. Annals of the Institute of Statistical Mathematics, **62**:117–125.

Lukacs, P. M., W. L. Thompson, W. L. Kendall, W. R. Gould, P. F. Doherty, K. P. Burnham, and D. R. Anderson. 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology, **44**:456–460.

Mayo, D. and A. Spanos. 2011. Error statistics. In M. R. F. Prasanta S. Bandyopadhyay, editor, Philosophy of statistics, pages 152–198. Number 7 in Handbook of Philosophy of Science, Elsevier.

Mayo, D. G. and D. R. Cox. 2006. Frequentist statistics as a theory of inductive inference. In J. Rojo, editor, Optimality: The Second Erich L. Lehmann Symposium, pages 77–97. Institute of Mathematical Statistics, Beachwood, Ohio.

Mundry, R. 2011. Issues in information theory-based statistical inference-a commentary from a frequentist's perspective. Behavioral Ecology and Sociobiology, **65**:57–68.

Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecology Letters, **12**:1061–1068.

Murtaugh, P. A. 2014. In defense of P values. Ecology, **95**:611–617.

Nuzzo, R. 2014. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. Nature, **506**:150–152.

Quinn, G. P. and M. J. Keough. 2002. Multiple and complex regression. In G. P. Quinn and M. J. Keough, editors, Experimental Design and Data Analysis for Biologists, pages 111–154. Cambridge University Press, Cambridge, 1st edition.

R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology, **86**:2805–2814.

Richards, S. A., M. J. Whittingham, and P. A. Stephens. 2011. Model selection and model averaging in behavioural ecology: the utility of the it-aic framework. Behavioral Ecology and Sociobiology, **65**:77–89.

Royall, R. 2000. Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London.

Stanton-Geddes, J., C. G. De Freitas, and C. De Sales Dambros. 2014. In defense of P values: Comment on the statistical methods actually used by ecologists. Ecology, **95**:637–642.

Steidl, R. J. 2006. Model selection, hypothesis testing, and risks of condemning analytical tools. Journal of Wildlife Management, **70**:1497–1498.

Stephens, P. A., S. W. Buskirk, G. D. Hayward, and C. Martínez. 2007. A call for statistical pluralism answered. Journal of Applied Ecology, **44**:461–463.

Stephens, P. A., S. W. Buskirk, G. D. Hayward, and C. Martinez Del Rio. 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology, **42**:4–12.

Teräsvirta, T. and I. Mellin. 1986. Model selection criteria and model selection tests in regression models. Scandinavian Journal of Statistics, pages 159–171.

Touchon, J. C. and M. W. McCoy. 2016. The mismatch between current statistical practice and doctoral training in ecology. Ecosphere, **7**.

Wasserstein, R. L. and N. A. Lazar. 2016. The ASA's Statement on p -Values: Context, Process, and Purpose. The American Statistician, **70**:129–133.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of animal ecology, **75**:1182–1189.

Appendices

A.1 Analytical relationship between NHT and IT approaches

Murtaugh (2014) used the Log-likelihood ratio tests (LRT) to identify a mathematical relationship between p-values and Akaike weights. This demonstration is valid under the assumptions of LRT, namely that (i) the sample size is large enough to allow the distribution of deviances (minus twice likelihood ratios) under the null hypothesis to be approximated to a Chi-square distribution and (ii) the simpler model to be compared is a particular case of the more complex one (nested models).

Edwards (1972) and Royall (2000) show many other instances of mathematical correspondence between inferences based on the Null Hypothesis Testing (NHT) and Information Theoretical (IT) approaches that do not rely on LRT. Here we show a simple example taken from Edwards (1972) to illustrate that the inferences done with both approaches tend to match as sample size increases, as found by Murtaugh (2014). Nevertheless, the convergence rates may differ, and thus for small to moderate sample sizes IT and NHT can lead to different inferences.

For many NHT standard tests, a correspondent "support test" can be defined as the degree of support the data provides for two alternative models (Edwards, 1972). The simplest case is a t-test for the null hypothesis that the sample comes from a Gaussian distribution with the mean fixed at a particular value. An alternative model is that the true mean of the distribution equals its maximum likelihood estimate (MLE), which is the sample mean. In both cases the standard deviation is set to its MLE, which is estimated from the sample. From an IT perspective the additional support that the alternative model has compared to the null model is the log-likelihood ratio. The minimum change in support required to reject the null model and choose the alternative model is a cutoff value of the log-likelihood ratio that can be expressed as function of t-statistic as:

$$\mid t_c \mid \; = \; \sqrt{(n-1)(e^{2L/n}-1)} \qquad \text{(A-1)}$$

where n is the sample size, L is the cutoff likelihood ratio and t_c is the critical t-value (Edwards, 1972). Figure A-1 shows the value of t_c calculated from equation A-1 and from the t-distribution as a function of sample size.

Both tests statistics converge to the value of two as sample size increases, as expected with a Gaussian model. This convergence is slower for the t-test, which thus is more conservative in rejecting the null model under small sample sizes. The conclusions of the two tests are the same for sample sizes larger than 30, when the Gaussian distribution becomes a good approximation of the distribution of the t-statistics.

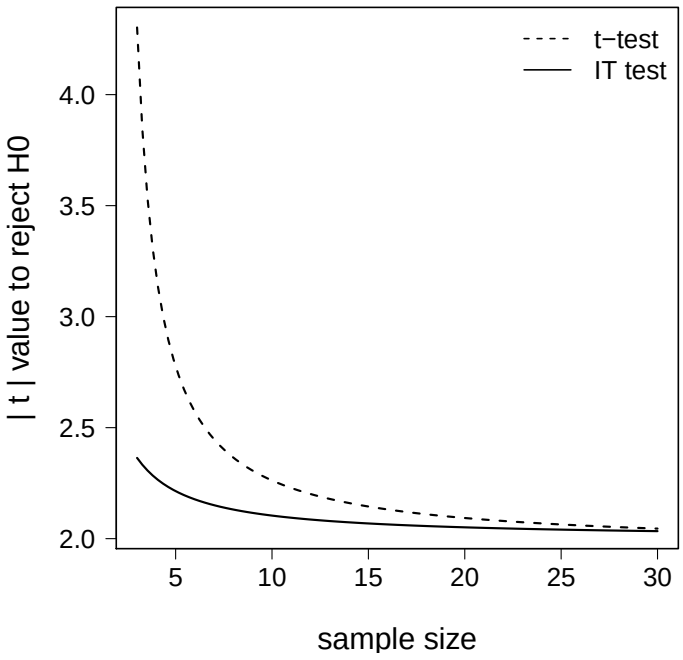

Figure A-1: The critical value of the t-statistic and its IT correspondent as a function of sample size. In both cases the critical value t_c is the threshold value to reject the null hypothesis in a one-sample t-test design. The broken line shows the critical value calculated from the t-distribution for a significance level of $\alpha = 0.05$. The continuous line shows the critical value calculated from equation A-1 for a log-likelihood ratio of two. After Edwards (1972).

The geometry of the relationship between p-value and support

Figure A-2 shows the geometry behind the proof by Edwards (1972) that the p-value and log-likelihood ratios are monotonically related in the t-test, as in many other standard significance tests. The t-distribution is a model for the t-statistic calculated from samples taken from the same Gaussian distribution. However, there are an infinite number of t-distributions for samples taken from Gaussian distributions that differ in some amount in their means. Among those alternatives will be the one that is best supported by the data. The lower the p-value of t under the null hypothesis, the higher the probability that the alternative, best supported t-distribution, assigns to this same value.

A.2 Adjusted IT criteria for uniformative models: effect on power, M-errors and S-errors

The t-test and correlation designs can be translated into two alternative models, which correspond to the null and alternative hypotheses in the NHT approach, as we detailed in the Methods section. Nevertheless, to translate ANOVA and linear regression designs to the IT approach we must fit more than one alternative model. For instance, in our linear regression example with two putative predictor variables four additive models are possible:

- $E[Y] = a_0$
- $E[Y] = a_0 + a_1 X_1$
- $E[Y] = a_0 + a_2 X_2$
- $E[Y] = a_0 + a_1 X_1 + a_2 X_2$

The first model corresponds to the null hypothesis of no effect and the second model corresponds to the correct alternative hypothesis that only predictor X_1 has an effect on the expected value of the response variable ($E[Y]$). The fourth model also has the effect of X_2, which we set to zero. Random sampling variation will give this model an estimated value of the effect close to zero in each simulated fit. In this case the fourth model is equivalent to the correct model plus an uninformative parameter a_2. Nevertheless, the small estimated value of a_2 can sufficiently improve the fit to the observed

data to make the AIC of this model with an uninformative parameter lower than the true model. The probability of this misleading selection converges to a value larger than zero as sample size increases (Geweke and Meese, 1981; Teräsvirta and Mellin, 1986), which means that AIC is not asymptotically consistent (although AIC is asymptotically efficient, see Aho et al., 2014, for a full discussion). This problem arises when the true model is surely among the competing ones and the purpose of model selection is to pick it (Aho et al., 2014), as is the case in simple significance test designs like those we simulated. To circumvent this problem we used the additional parsimony criterion proposed by Arnold (2010): to select the model with fewer parameters which was among the models with $\Delta AIC < 2$.

The figures below replicates the comparisons of power, S-errors and M-errors between NHT and IT approaches shown in figures 2 – 3 but also with the IT criterion without the parsimony correction described above. We also included the results for a simulation of the linear regression design with a correlation of 0.5 between the uninformative predictor (X_2) and the true predictor (X_1).

The power of the unadjusted IT criterion converges to the value 0.84 as effect size increases (Figure A-3), because the probability of the selection of the model with an additional uninformative parameter converged to 0.16 in our simulations. This finding is in line with the theoretical upper bound of power of the AIC model selection for two nested models (Teräsvirta and Mellin, 1986), as is the case for our simulations of ANOVA and linear regression. These results also support the proposition of Arnold (2010) that the power of AIC is bounded to 5/6 when there is a model with a 'spurious variable'.

A.3 R codes for the simulations

Functions in R to perform the simulations with any combination of parameters and the R scripts of the simulations are available at `https://github.com/piklprado/NHTxIT`.

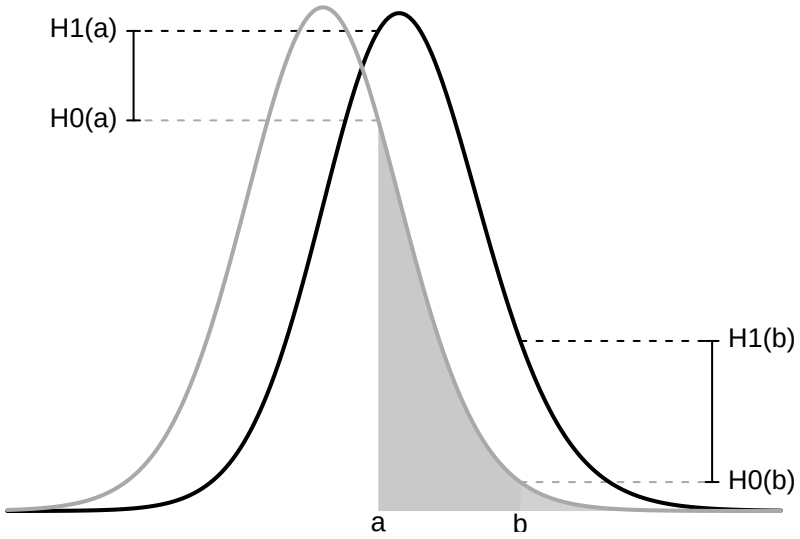

Figure A-2: The relationship between the p-value and the likelihood ratio in a t-test. The curve in grey is the standard t-distribution , which gives the probability density of a given t-value under the null hypothesis of no difference among population means. The curve in black is a non-central t-distribution, which gives the probability density of t-values under an alternative hypothesis that the population means differ to some amount. Points a and b are values of the t-statistic for two hypothetical testing situations. The grey areas are the p-values for $t = a$ (dark + light grey area) and $t = b$ (light grey area) under the null hypothesis. For each value of t, the likelihood ratio is the ratio between the probability density given by the two distributions ($\mathcal{L}_a = H_1(a)/H_0(a)$; $\mathcal{L}_b = H_1(b)/H_0(b)$). The comparison of situations a and b shows that the lower the p-value the higher the likelihood ratio H_1/H_0, and thus the stronger the support of H_1 over H_0. Akaike evidence weights w are proportional to likelihood ratios, and in such simple designs such as t-test w will increase monotonically as the p-value decreases. After Edwards (1972).

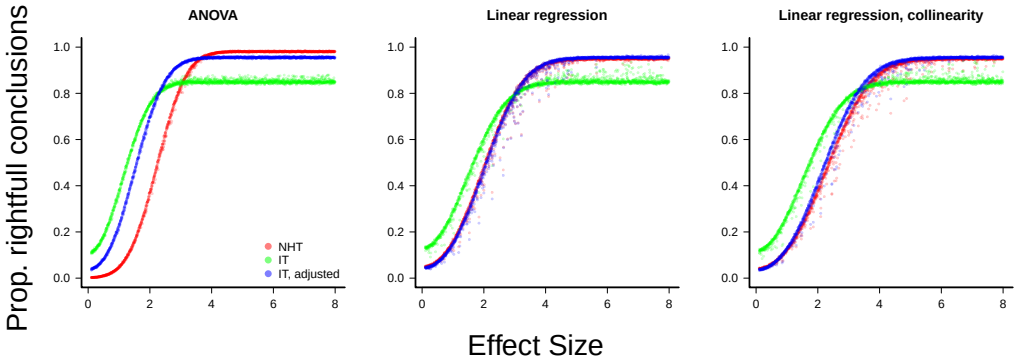

Figure A-3: The power of null hypothesis tests (NHT, red) and information-based model selection with and without the parsimony adjusting for uninformative parameters (green and blue, respectively see Aho et al., 2014, and the text above), as a function of effect size, for ANOVA and linear regression designs. Each point is the proportion of the 10,000 simulations of a test instance from which the effect was detected. Each test instance used a different combination of effect size, standard deviation of the values, and sample size. Linear regression with collinearity was simulated with a correlation of 0.5 between the two predictor variables.

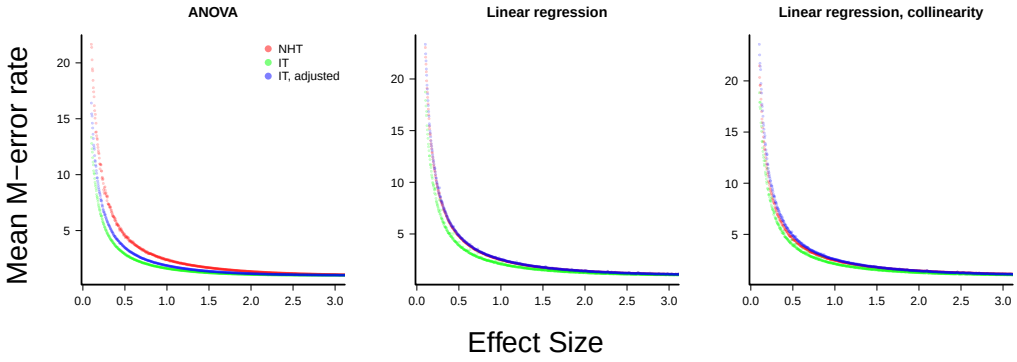

Figure A-4: The mean type-M error (exaggeration rate) of null hypothesis tests (NHT, red) and information-based model selection with and without the parsimony adjusting for uninformative parameters (IT, green and blue, see Aho et al., 2014, and the text above), as a function of effect size, for ANOVA and linear regression designs. Each point represents the simulations of a test instance from a which an effect was detected. Each test instance used a different combination of effect size, standard deviation of the values and sample size and was simulated 10,000 times. The M-error is the absolute ratio between the effect size estimated and the true effect size (Gelman and Carlin, 2014), which was estimated from the mean of this ratio for each test instance. Linear regression with collinearity was simulated with a correlation of 0.5 between the two predictor variables.

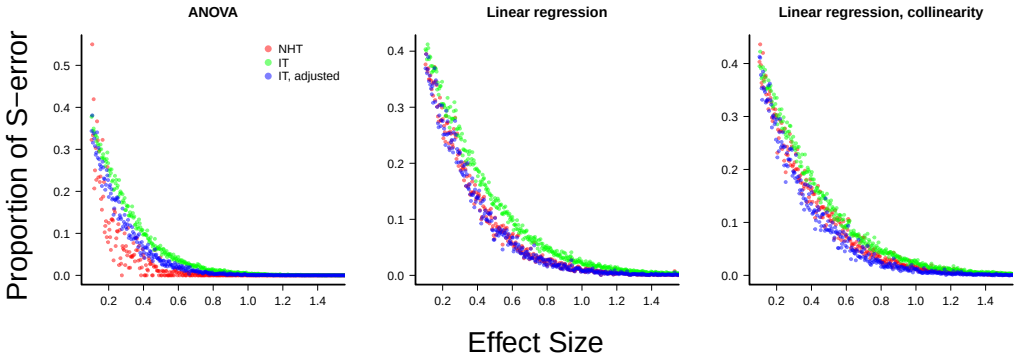

Figure A-5: The mean type-S error of null hypothesis tests (NHT, red) and information-based model selection with and without the parsimony adjusting for uniformative parameters (IT, green and blue, see Aho et al., 2014, and the text above), as a function of effect size, for ANOVA and linear regression designs. Each point represents the simulations of a test instance from a which an effect was detected. Each test instance used a different combination of effect size, standard deviation of the values and sample size and was simulated 10,000 times. The S-error is the probability of detecting an effect of an opposite sign of the true effect (Gelman and Carlin, 2014). For each test instance we estimated S-errors from the proportion of simulations that detected an effect of the opposite sign. Linear regression with collinearity was simulated with a correlation of 0.5 between the two predictor variables.

References

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology, **95**:631–636

Arnold, T. W. 2010. Uninformative parameters and model selection using akaike's information criterion. The Journal of Wildlife Management, **74**:1175–1178

Edwards, A. W. F. 1972. Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference. Cambridge University Press, Cambridge

Geweke, J. and R. Meese. 1981. Estimating regression models of finite but unknown order. International Economic Review, pages 55–70

Murtaugh, P. A. 2014. In defense of P values. Ecology, **95**:611–617

Royall, R. 2000. Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London

Teräsvirta, T. and I. Mellin. 1986. Model selection criteria and model selection tests in regression models. Scandinavian Journal of Statistics, pages 159–171