Department for Ecology and Nature Conservation
Faculty for Biology and Preclinical Medicine
University of Regensburg

Module: Theoretical Ecology
SS 2023
Supervision: Florian Hartig

A comparison of statistical methods to establish no effect

Author: Isabelle Halbhuber

Abstract

The true absence of a statistical effect is not necessarily established by a null hypothesis test result that is non-significant (p-value > significance level). Therefore, various statistical methods have been developed or used to detect a true 'no effect'. Nevertheless, non-significant results can still be interpreted by post-hoc power analyses, such as calculating the minimum detectable difference (MDD), which indicates whether the experiment could have detected a relevant effect. Apart from MDD, which has limitations for post-hoc interpretation of non-significant results, other statistical methods may be more appropriate for this purpose. However, the determination of the most reliable (e.g., low error rate) statistical test to decide whether a non-significant test result should be treated as a true negative remains vague. The choice of the correct statistical test, and hence the detection of a true null effect, is particularly important in the risk assessment of chemicals. The main impetus for risk assessment, e.g., by the European Food Safety Agency (EFSA), is to provide safe and reliable no-effect concentrations of chemicals while minimizing financial investment and ethical implications (e.g., animal testing). However, it is currently unclear which test method provides the most reliable results while requiring the least resources. To answer this question, I compared the MDD with Confidence Intervals (CI), Bayes factors (BF) and Equivalence test (EQUIV) in their ability to distinguish between true and false negatives using false trust and false mistrust rates, in this work. Additionally, I analyzed how the false trust rate of EQUIV and MDD are impacted by varying sample sizes. The latter is motivated by the recent switch of EFSA from the MDD to EQUIV in their testing protocol. Based on the R script of Mair et al. (2020), I developed two simulations, the first compares the false trust and false mistrust rates of all statistical methods, and the second investigates changes in the false trust rate of EQUIV and MDD with varying sample size. To provide a comprehensive understanding, I explain the concept of MDD and CIs, the Bayesian approach using BF, and the principles of EQUIV. My main findings are that CIs and EQUIV are equivalent and that CIs, EQUIV and BF outperform MDD in identifying true negatives among non-significant results. The reason for this is that MDD, unlike CI, EQUIV and BF, does not consider the estimated effect size in its calculation. This also explains why the old EFSA guidelines, which used MDD, required fewer animals to be tested than EQUIV. I found that EQUIVs achieve a 70% false trust rate with a sample size of 30, while MDDs reach a 70% false trust rate with a sample size of 6. EQUIV requires a larger sample size in risk assessment but ensures more reliable results compared to MDD. In conclusion, I recommend the use of CIs for deciding whether to treat a non-significant test result as a true negative. CIs are more robust to changes in sample size than BF and, unlike EQUIV, are a common method that most scientists have in their statistical toolbox. This makes CIs the most appropriate method for establishing no effect.

1. Introduction

To establish the true absence of an effect, it is necessary to understand the interpretation of non-significant (p-value > significance level) results. A non-significant result of a null hypothesis significance test (NHST) can mean that there is either evidence for the null hypothesis, indicating that no difference between parameters of groups exists, or that the data is insensitive to find any differences (Dienes, 2014). The insensitivity of the data depends on effect size and sample size. In a large sample, smallest effects of no importance can be found to be significant, whereas in a small sample, even large effects of relevance may be found non-significant (Faller, 2004). The sample size determines both the dispersion of the data around the mean, and therefore the variance, and the ability of a test to detect a significant effect when it is present, and thus the statistical power (Satterthwaite, 1941, Fay, 2007). The power is the complementary event of the Type II error, which is equivalent to a false-negative result. A false-negative result occurs when the null hypothesis is not rejected, although there is an effect that results in differences between the groups' parameters (Kaur, 2017). In addition to the Type II error, there is another error that can interfere with hypothesis testing, the Type I error. This error corresponds to a false positive, which results when the null hypothesis is incorrectly rejected even though there is no effect (Newman, 2008). To understand non-significant results, it is therefore necessary to account for possible Type I and Type II errors.

In a similar way to the concept of error rates, a false trust rate (FTR) is the rate at which a statistical method suggests trust in a non-significant result despite the presence of a real effect, and a false mistrust rate (FMR) is the rate at which a statistical method falsely suggests mistrust of a non-significant result (Mair, 2020). These FTRs and FMRs illustrate the precision of post-hoc interpretation of non-significant results and differ between statistical methods for detecting true no effects.

Detecting effects of potentially harmful chemicals is crucial because chemicals are an integral part of modern industry. Since industrial revolution, chemicals have been used extensively in a wide range of industries such as textiles, food, cosmetics, or agricultural production (Heaton, 1993). Agricultural crops are protected from damage using pesticides, which control pests like insects, weeds, or fungi through their chemical activity (Mahmood 2016). The chemical effects on non-target organisms have been studied over years for certain pesticides and lead to increasing evidence that the use of pesticides is associated with serious health and environmental risks (Williams, C. M. 1967, Nesheim, 1978, Ware, 1980, van der Werf, 1996, Fenik, 2011, Carvalho, 2017, Aguilar-Marcelino, 2023). The environmental and health damage caused by direct or indirect exposure to chemicals leads to habitat changes and affects biodiversity (Isenring, R., 2010). Since the adverse ecological side-effects of authorized chemicals became known, public awareness and concern have led policymakers to commission scientists who monitor the potentially harmful effects of chemicals.

Scientists identify the fate and effects of potentially toxic chemicals in ecosystems and advise policymakers on appropriate regulation of chemicals in the environment (Kendall, R., 2001). An example of proper risk assessment is the European Union`s REACH regulation. ‘Registration, Evaluation, Authorization and Restriction of Chemicals’ (REACH) was introduced in 2006 and aims to protect human health and the environment from the risks posed by potentially harmful effects of chemicals (Williams, 2011). In addition to REACH, which regulates the authorization and evaluation of industrial chemicals to be placed on the European market, there is the European Food Safety Authority (EFSA). EFSA is an impartial agency of the European Union that provides scientific advice on environmental risk assessment, legislation, and regulations to protect human health and the environment from food-related risks (see https://www.efsa.europa.eu/en/about/about-efsa, Oltmanns, 2019). The legislation and scientific advice on the fate and effects of potentially toxic chemicals in the environment rely on meaningful statistical approaches.

The determination of the most reliable (e.g., low error rates) statistical test to decide whether a non-significant test result should be treated as a true negative remains vague. Historically, ecotoxicological risk assessment studies have used the no-observed-effect level (NOEL) or no-effect concentration (NEC) to establish a threshold at which adverse effects occur. The highest dose that does not cause a toxic effect (NOEL) is determined by significant differences between a treatment and a control group using hypothesis testing (e.g. Wang, 1988, Srivastava, 1999, EFSA, 2016, EFSA 2017, EFSA, 2019). Because the name NOEL is considered misleading, scientists refer to the highest dose that does not cause a toxic effect as the No Observed Effect Concentration (NOEC) (Warne, 2008). The use of the NOEC is controversial: Skalski (1981) argued that NOECs violate the principle of negative inference. Hoekstra (1993) criticized that the use of NOECs is associated with a severe power problem. To overcome the problem of too low power, a post-hoc power analysis for the concentration-response relationship determined by NOEC can be performed using minimum detectable differences (MDD) (Brock, 2015, EFSA, 2016, EFSA 2017, EFSA, 2019). The MDD indicates whether the experiment could have detected a relevant effect, but it has limitations when used for post-hoc interpretation of non-significant results by threshold comparisons (Mair, 2020). In addition to MDD, other statistical methods could be used to decide whether to treat non-significant results as true negatives: The Confidence Interval (CI) is an interval estimate for a certain parameter of interest (Le, 2003) and the location of its boundaries relative to a predetermined threshold can be used to predict weather an effect is ‘acceptable low’ or not (Lakens, 2017). For establishing an 'acceptable low' or 'equivalent to zero' effect concentration, equivalence tests, which test whether the differences between two groups are less than a given threshold, are suitable (Engel, 2021, Lakens, 2017). In addition, the hypothesis comparative approach, the Bayesian Factors, are used to decide whether there is evidence for the null hypothesis or the alternative hypothesis and therefore to decide whether an effect is present or absent. These statistical tests can be prone to various sources of error and therefore may have inconsistent validity.

I compared MDD, EQIV, CI and BF for their reliability in detecting true negatives among non-significant results of a NHST, by comparing their rate to falsely mistrust a significant result, despite there is no effect (FMR) and their rate to falsely trust a non-significant result, despite there is an effect (FTR). My aim was to identify the best method for distinguishing between true negatives and false negatives to reliably establish no effect. Furthermore, I investigated the consequences that come along with the recent switch to the use of Equivalence tests instead of MDDs in risk assessment of European Food Safety Authority (EFSA).

2. Review of statistical approaches to establish no effect

2.1 NHST

The hypotheses. In science in general, and here with a focus on ecotoxicology, hypothesis testing is a statistical method used to determine whether observed differences are random or systematic (Schönbrodt, 2017). Systematic differences between two observed groups indicate that a variable, e.g. chemical exposure, has an effect on a particular parameter, e.g. the average survival rate of each two groups. One of the groups is the control group, which is not exposed to the chemical but is matched to the same conditions as the treatment group, which is exposed to the chemical. NHSTs are designed to test a previously formulated null hypothesis, which states that there is no difference or no relationship between distributions or parameters (Krueger, 2001). If differences in the parameters in both directions are of interest, a two-sided test is performed, and if differences in only one direction are of interest, a one-sided test is performed. When $Parameter_{treatment}$ is the distribution of data from the group treated with chemicals and $Parameter_{control}$ is the distribution of data from the control group, then the null hypothesis (H0) and alternative hypothesis (H1) are:

$$H_0: Parameter_{treatment} - \ Parameter_{control} = \ 0 \ (two - sided)$$
$$H_1: Parameter_{treatment} - \ Parameter_{control} = ! \ 0$$
$$H_0: Parameter_{treatment} - \ Parameter_{control} \leq \ 0 \ (one - sided)$$
$$H_1: Parameter_{treatment} - \ Parameter_{control} > \ 0$$

The alternative hypothesis H1 is complementary to the null hypothesis and refers to the same parameters as H0. The null hypothesis is the basis for statistical inference concluded by the principle of falsification (Le, 2003).

The significance level. To falsify or verify the null hypothesis, the Null Hypothesis Significance Test uses a significance level α, which is a predetermined threshold to determine the level of evidence required to falsify a null hypothesis (Gill, 1999). The significance level indicates a certain chance that the null hypothesis will be falsely rejected, resulting in a false positive also called Type I error (Banerjee, 2009).

The error types. In hypothesis testing, two types of error can occur: A Type I error and a Type II error. When a type I error occurs, the H0 is erroneously rejected, and when a type II error occurs, the H0 is erroneously not rejected, thus Type I and Type II error behave contrary (Akobeng, 2016). A reduction in Type I error would, under the same conditions, lead to a corresponding increase in Type II error. Both errors depend on the sample size: the larger the sample size, the lower the risk of either Type I error or Type II error occurring (Harmon, 2005). As mentioned above, type I error usually has a probability equal to the significance level α and Type II error has a probability denoted by β (Pollard, 1987).

The power. (1 - β) is the so-called statistical power. The power is the complementary event of the Type II error, which is equivalent to a false-negative result (Kaur, 2017). Therefore, high statistical power indicates a high probability of rejecting the null hypothesis when it is false (Källén, 2011).

The power-approach-paradox. A power analysis uses the relationships between statistical power, sample size, α and effect size. Based on three of these four parameters, the analysis can estimate the appropriate sample size a priori or calculate the power and effect size post hoc for a given experiment (Nakagawa, 2004). Post hoc power calculations are a widely debated method for interpreting non-significant results (Mair, 2020). It has been shown that for non-significant results (p-value > α), post-hoc power calculations result in low observed power (< 50%), even for large sample sizes (Goodman, 1994). This calculated power does not inform about the probability of rejecting the null hypothesis when it's false. In contrast to this, rather uninformative post hoc calculation of power, the post hoc calculation of the effect size is of greater informative value. Thereby, the resulting minimum detectable effect (MDE) provides insight into the theoretical power of the experiment to detect a particular effect, whereas the interpretation of whether to accept the non-significant result is counterintuitive (Mair, 2020). There are two counterintuitive phenomena denoted as the „power-approach-paradox": 1) Intuitively, in a set of the two experiments with identical variances but different estimated effect sizes, the experiment with the lower estimated effect size provides more evidence of a true negative effect (less Type II error). This is paradox, as both experiments have the same MDE, which is independent of the estimated effect size. 2) In a second set of two experiments with identical estimated effect sizes but different variances, one would intuitively say that the experiment with the smaller variance seems to provide more evidence for a small effect, because small variances imply less scattering around the mean value. However, determining trust in non-significant results based on the MDE value would lead to the opposite conclusion, because a smaller variance increases power and thus leads to a smaller MDE (Hoenig, 2001, Mair, 2020, Satterthwaite, 1941).

The p-value. In NHST the p-value is often used to decide whether to reject or not to reject H0: If the p-value is smaller than or equal to the significance level α, the H0 is rejected and if the p-value is larger than the significance level α, H0 is not rejected (Martínez-Abraín, 2008). Greenland (2016) explained that a p-value less than, greater than, or equal to α only means that a deviation from the hypothesized prediction would be as largen as or larger than the observed one in at most α of the cases if the deviation were caused by chance alone. Therefore, the p-value can be defined as the probability that the values of the test statistic are as extreme as or more extreme than those observed when the null hypothesis is true (Le, 2003).

The NHST is a useful tool to assume weather the null hypothesis is true, but it fails to find out how far away the true effect is from zero if the null hypothesis was not rejected (Lakens, 2017). Therefore, one needs to decide whether to trust or to mistrust non-significant results by comparing their ability to discriminate between true absence of effects and false negatives (type II errors) (Mair, 2020).

2.2 Post-hoc power analysis

The MDD. The MDD is independent of the estimated effect size but determines the minimum effect size, that would lead to significant results when comparing the means of two different groups in a set of experiments (Duquesne, 2020). It considers the theoretical power of an experiment and can therefore be used to determine an appropriate sample size (Brock, 2015) and like the minimum detectable effect size (MDE), when comparing MDD to certain thresholds, it can be used to interpret non-significant. This interpretation has some shortcomings, which are summarized by the 'power-approach-paradox' (see section 2.1). The MDD can be calculated as follows (Mair, 2020):

$$1)\ MDD = t_{critical} \times s \times \sqrt{\frac{1}{n_{control}} + \frac{1}{n_{treatment}}}$$

Where $t_{critical}$ is the critical value of the test statistic, s is the residual standard deviation and $n_{control}$ and $n_{treatment}$ are the number of observations in each group. I calculated the MDD in R Studio using the above-mentioned function (1).

2.3 Equivalence analysis

EFSA (2010) defined equivalence as the absence of differences in a certain parameter between a treated group and an untreated control group. The difference between these groups is regarded as 'acceptable' or 'equivalent to zero' if it lays within the prespecified, limiting equivalence thresholds -Δ and Δ, which are prespecified by a regulatory agency (Lakens, 2017). If Δ (>0) is the prespecified acceptable difference between two distributions and therefore the limiting upper equivalence threshold, the null hypothesis the alternative hypothesis are formulated as follows (Tango, 1998):

$$H_0^{\Delta}: Parameter_{treatment} - \Delta - Parameter_{control} > 0$$
$$H_1^{\Delta}: Parameter_{treatment} - \Delta - Parameter_{control} \leq 0$$

There are different procedures to test this null hypothesis (Hsu, 1994). In my study I will focus on the following α-controlled tests: *Confidence Intervals* and *Equivalence tests* (one-sided t-tests):

Equivalence tests. The one-sided t-test rejects the null hypothesis with a significance level α and an estimate of the standard deviation s with associated degrees of freedom, if,

$$Parameter_{treatment} - Parameter_{control} \leq \Delta - t_{1-\alpha^{s}}.$$

Significant results lead to the conclusion that the effect is equivalent to zero and the corresponding chemical concentration is accepted (Müller-Cohrs, 1990). When testing for equivalence, either one or two one-sided tests can be performed, depending on whether one is interested in detecting only an increase or a decrease or both (EFSA, 2010). I was interested in finding out weather the treatment group has greater parameters than the control parameters or not. Therefore, I tested for equivalence in one direction using the t.test() function in RStudio.

Confidence Intervals. The CI is an interval estimate, which measures the accuracy of the estimation of a parameter by providing an interval within which the parameter is assumed to lie. The interval of values that includes the 'true' value is estimated by a given statistic with a given probability (Nakagawa, 2007). Most used CIs are the 90%, 95% and 99% CI (Simundic, 2008). In a 95% CI for example will the real effect be larger than the upper and smaller than the lower CI bound in 5 % of the experiments. The CI informs about the hypothetical values that could not be rejected in a set of samples and is therefore used to represent statistical uncertainty of parameter estimation (Smithson, 2003). Hsu (1994) proposed that a usual way to declare equivalence is either to reject H_0^Δ or to show that a confidence interval CI(X), based on the observation vector X, is:

$$CI(X) \cap (-\Delta, \Delta) = \emptyset.$$

I looked at the upper limits of the CIs because I was interested in whether the interval estimate of the effect was too large to be accepted. It is acceptable if,

$$upper\ boundary\ of\ CI(X) \leq \Delta.$$

Using the t.test() function in RStudio, I extracted upper boundaries of 95% CIs in RStudio.

Equivalence Tests and Confidence Intervals have previously been found to be functionally identical when the CI is completely contained within the equivalence interval $[-\Delta, \Delta]$ (Schuirmann,1987, Tango 1998). This finding suggests that the use of Equivalence tests in general should be questioned and the use of confidence intervals should be considered, as these are part of the researcher’s standard statistical toolbox.

2.4 Hypothesis comparative analyses

Bayesian Factor analysis. In contrast to NHST, which focuses on the null hypothesis, Bayesian hypothesis testing quantifies the relative probability of alternative hypotheses and null hypotheses (Keysers, 2020). The Bayesian approach is a probabilistic inference method used to adjust the estimates of a parameter (prior) based on observed data, resulting in a new probability distribution (posterior) for the parameter of interest (Eddy, 2004). The posterior probability distribution P(H|D) can be calculated using the Bayes-Theorem (Bolstad, 2016 and Stöcklin): When P(D|H) is the likelihood, that gives the probability of the data under the assumption that the hypothesis is not rejected and P(H1) is the prior, which is the assumption for the probability that the hypothesis is not rejected, then:

$$\text{2) for H0:} \quad P(H_0|D) = \frac{P(D|H_0) \cdot P(H_0)}{P(D)}$$

$$and \text{ 3) for H1:} \quad P(H_1|D) = \frac{P(D|H_1) \cdot P(H_1)}{P(D)}$$

The Bayes factor compares the two hypotheses and decides then whether the result is evidence for the null hypothesis or the alternative hypothesis. The Bayes-Factor (BF) is the factor by which the prior is changed to the posterior and can be derived as follows:

$$4)\ \frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \cdot \frac{P(H_1)}{P(H_0)} = BF \cdot \frac{P(H_1)}{P(H_0)}$$

I computed the Bayes factors using the ‘BayesFactor’ package (from Richard D. Morey and Jeffrey N. Rouder) in RStudio.

3. Comparison of statistical Methods

My aim was to find out which statistical method is most appropriate to distinguish between true negative and false negative tests, and how sample size affects the performance of MDD and equivalence tests in finding true negatives. To address these two questions, I generated two different simulations: the FMR/FTR-simulation and the sample-size-simulation. Both simulations imply different statistical methods (such as BF, CI, EQUIV, MDD) to assess the trustworthiness of non-significant (p-value > α) results derived from a one-sided t-test. The results of the FMR/FTR-simulation were used to calculate the FTR and FMR for each statistical method. The rates were used to draw a Pareto plot (Figure 2), which provides a visual assessment of how well a test discriminates between false positives and false negatives. The sample-size-simulation requires EQUIV and MDD to analyze datasets with different sample sizes. Thereby, I wanted to investigate the effects of different sample sizes on the FTR to clarify the consequences of EFSA switching from using MDD to EQUIV.

3.1 Simulations

I generated data sets displaying a control and treatment group having the same standard deviation, the same length and being normal distributed. The only difference between the two groups lays in their means: in half of the datasets the treatment group has a higher mean than the control group, indicating a larger true effect size of the treatment group compared to the control group, while in the other half the mean is the same (see Fig. 2: 'Data set'). The generated data sets have ether the same sample size (25) or have differing sample sizes (2:70, step size = 2). I analyzed the data sets with simulations, that imply different statistical methods (BF, CI, EQUIV, MDD) to test whether to ‘trust’ or to ‘mistrust’ non-significant (p-value > α) results from a one-sided t-test (see Fig. 2: 'Simulation'). If the t-test p-value is greater than α, the 'effect detected' variable is set to 'false', and if the result of the appropriate statistical methods meets the appropriate assumptions (BF <= threshold, upper CI <= threshold, EQUIV p-value <= α, MDD <= threshold), then the 'no effect trusted' variable is set to 'true' and the effect can be considered 'acceptable low'. Accepting an effect depends on how high the threshold is set. A threshold must be chosen to translate the MDD, the CI, the EQUIV and the BF into a trust/mistrust decision. This allows the error rates of this decision (FTR and FMR) to be calculated for the different methods and thresholds. A set of thresholds (0:1.1, step size = 0.1) or a pre-calculated threshold is implemented in the respective FMR/FTR-simulation (Fig. 1: marked with purple color) and sample-size-simulation (Fig. 1: marked with blue color). For the sample-size-simulation, I calculated the threshold as in the statistical considerations of European Food Safety Authority (EFSA, 2010) (equations 5 and 6). Where 'lsd' stands for the least significant difference (equivalent to MDD), 'df' for degrees of freedom, 'i' for the model, 'a' for the significance level, 'XY' for the two groups and 'sed' for the standard error.

$$5\ )\ threshold = \exp\left(mean_{treatment} - mean_{control} - lsd\right)$$

$$6)\ lsd = \ t(df; i; a) * \ sed(XY; i)$$

Alternatively, the threshold can be calculated according to the EFSA 2023 test protocol, which states that the threshold should correspond to a 10% reduction in the treated test compared to the control. I used the results of the FMR/FTR-simulation to calculate the FTR and the FMR for each statistical method. The calculation of the two rates is based on the calculations of Mair et al. (2020):

$$7)\ FTR\ = \frac{\sum \boldsymbol{Trust\ no\ effect}\left(\boldsymbol{no\ effect\ detected}\ \ \&(\boldsymbol{true\ effect} > 0)\right)}{\sum(\boldsymbol{true\ effect} > 0)}$$

$$8)\ FMR\ =\ 1 - \left[\frac{\sum \boldsymbol{Trust\ no\ effect}\left(\boldsymbol{no\ effect\ detected}\ \ \&(\boldsymbol{true\ effect} = 0)\right)}{\sum(\boldsymbol{true\ effect} = 0)}\right]$$

I plotted the FMR against the FTR (Section 4, Fig. 2) for visual comparison. To examine the dependence of FTR on sample size, I calculated the average 'no effect trusted' for each dataset of a certain length from the results of the sample-size-simulation. I plotted the FTR against the length of the corresponding data set (Section 4, Fig. 3) for visual comparison.

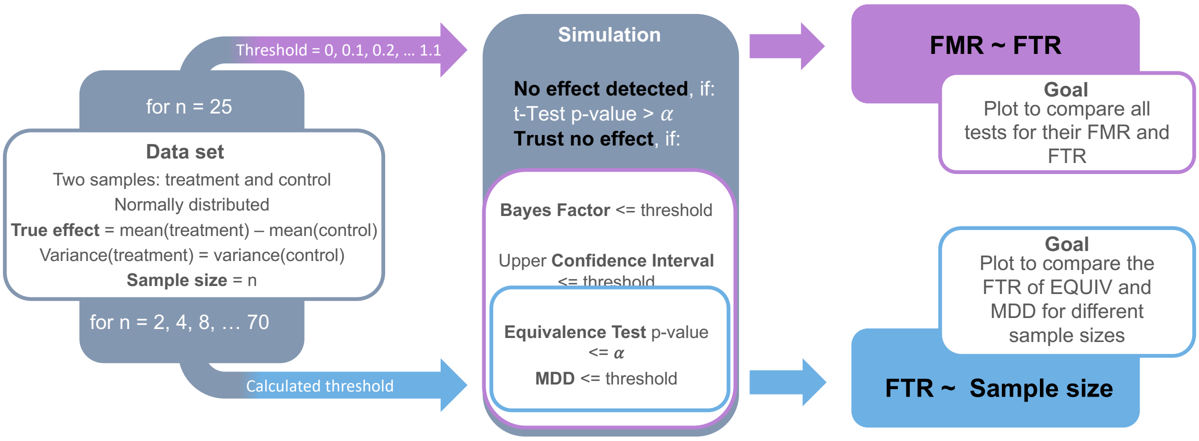
Threshold = 0, 0.1, 0.2, … 1.1
for n = 25
Data set
Two samples: treatment and control
Normally distributed
True effect = mean(treatment) – mean(control)
Variance(treatment) = variance(control)
Sample size = n
for n = 2, 4, 8, … 70
Calculated threshold
Simulation
No effect detected, if:
t-Test p-value > α
Trust no effect, if:
Bayes Factor <= threshold
Upper Confidence Interval <= threshold
Equivalence Test p-value <= α
MDD <= threshold
FMR ~ FTR
Goal
Plot to compare all tests for their FMR and FTR
Goal
Plot to compare the FTR of EQUIV and MDD for different sample sizes
FTR ~ Sample size

Figure 1: Schematic illustration of the R code workflow. The data sets display control and treatment groups with matching standard deviations, lengths, and normal distributions. Half of the data sets have the treatment group with a higher mean, indicating a larger true effect, while the other half have the same mean compared to the control group. The sample sizes of the generated data sets are either identical (n = 25) or different (n = 2:70, step size = 2) for FMR/FTR-simulation (purple) and sample-size-simulation (blue) respectively. Different thresholds are used in the two simulations. Both simulations use various statistical methods (BF, CI, EQUIV, MDD) to assess the trustworthiness of non-significant (p-value > α) results from a one-sided t-test. If the t-test p-value exceeds α, the 'effect detected' variable is set to 'false', and if the results of the relevant statistical methods meet certain assumptions (Bayes factor <= threshold, upper confidence interval <= threshold, equivalence test p-value <= α, MDD <= threshold), then the 'no effect trusted' variable is set to 'true'. The results of each simulation are plotted, and the resulting plots provide a visual comparison of the statistical methods.

4. Results

What is the most appropriate method for distinguishing true negatives from false negatives?

When comparing the performance of the four statistical methods, I found that CI (Fig. 3, green line) consistently performed exactly like EQUIV (Fig. 3, green line), regardless of the threshold or variance chosen. With increasing variance, the ability to discriminate between false trust and false mistrust of non-significant results decreased. The patterns of BF compared to CI and EQUIV were almost indistinguishable, especially at a standard deviation of 0.5 (Fig. 2). The variation in BF with respect to EQUIV and CI reflects the need to use different thresholds to achieve equivalent FTR and FMR. The threshold that indicates the highest probability of distinguishing false negatives from true negatives depends on the standard deviation. For a standard deviation of 0.5, the best threshold is 0.3 for CI and EQUIV and 0.2 for BF. At a standard deviation of 1, the best threshold is 0.3 for CI and EQUIV and 0.1 for BF. With a standard deviation of 1.5, the preferred threshold is 0.4 for CI and EQUIV and 0 for BF. It is noticeable that BF has lower thresholds than EQUIV and CI. In all cases, BF, CI and EQUIV outperformed MDD. MDD appears to behave approximately randomly (see Appendix, Fig. 4: 'Random'), which would imply that MDD cannot systematically discriminate between trust and mistrust, and therefore MDD would not help to interpret non-significant results.

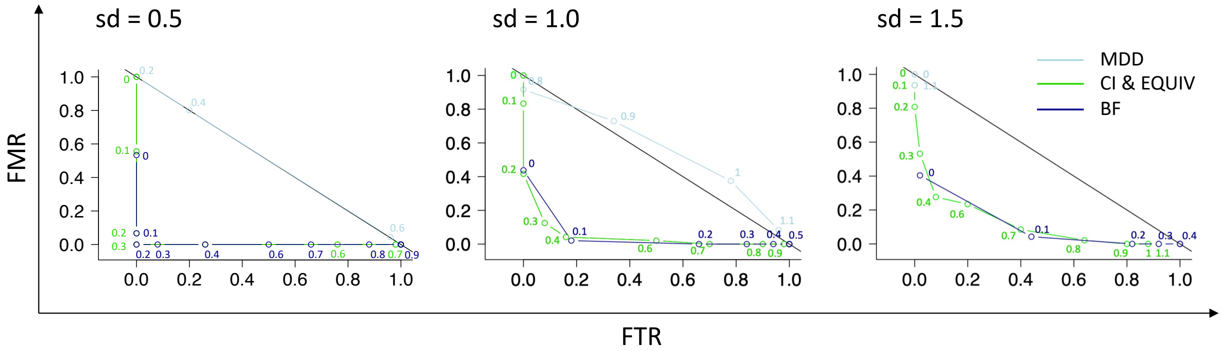

Figure 2: Application of different thresholds to see the tendency of statistical methods to falsely trust or mistrust non-significant results. A threshold (circle with corresponding value) must be chosen to translate the Minimum Detectable Difference (MDD), the Confidence Interval (CI), the Equivalence Test (EQUIV) and the Bayes Factor (BF) into a trust/mistrust decision. This decision allows error rates (False Trust Rate (FTR) and False Mistrust Rate (FMR)) to be calculated for the different methods and thresholds. Each of the three plots shows the FTR/FMR results for each statistical method on data sets with different standard deviations. With increasing standard deviation, the ability of CI (green line), EQUIV (green line) and BF (dark blue line) to discriminate between trust and mistrust of non-significant results decreased. CI consistently performed the same as EQUIV, regardless of the threshold or variance chosen. BF, CI and EQUIV outperformed MDD in all cases, showing lower FTR and FMR compared to MDD (light blue line), which behaved in all scenarios approximately random. The patterns of BF compared to CI and EQUIV were almost indistinguishable, especially at a standard deviation of 0.5. Here, the variation in BF with respect to EQUIV and CI reflects the need to use different thresholds to achieve equivalent FTR and FMR.

Old EFSA guidelines compared to new EFSA guidelines

The goal of ecotoxicological experiments is to derive a NOEC concentration or a no-observed ecologically adverse effect concentration (NOEAEC). NOECs from ecotoxicological tests that are statistically significant (p-value <= α) should be evaluated. The evaluation of NOEC is possible if the statistical power is high enough ensured by an appropriate number of species/taxa concentration–response relationships. For post hoc evaluation of statistical power, EFSA had considered MDD to be a valid concept. The calculation of the MDD allows the reporting of the actual effect which could be determined in the experiment for a given endpoint. MDD was used in European Food Safety Authority (EFSA) risk assessment in the last decade (Europäische Kommission, 2013, EFSA, 2016, EFSA 2017, EFSA, 2019). Now European Food Safety Authority (EFSA) switched to using Equivalence Tests in its ecotoxicological risk assessment: In the current risk assessment of plant protection products on bees, the derivation of a single effect estimate (NOEC) is no longer relevant, as the risk assessment is based on dose–response relationships with the goal to consider the predicted level of effect triggered by different exposure levels. For estimating the levels of risk EQUIVs are conducted to statistically analyze higher tier semi-field or field studies with the goal to identify relevant hazard parameters or 'effect endpoints'. Effect endpoints are the combination of the chosen dose-response model and the values of its parameters and give a link betwenn exposure in the field and effects in an experiment. European Food Safety Authority (EFSA) recommends to apply a one-sided equivalence test ($\alpha = 0.2$) for each endpoint, with an equivalence limit corresponding to a 10% reduction in the treated test compared to the control, to prove that there are no adverse effects (EFSA, 2023).

Can the Equivalence Test compete with MDD for small sample sizes?

The change from the use of MDD to the use of Equivalence tests in EFSA risk assessment raises questions of applicability. In practice, statistical methods use the data from animal experiments in which the number of animals tested, is to be kept as small as possible. My sample-size-simulation (section 3) shows how experimental sample size affects the acceptance of effects by the two different methods. To simulate the equivalence test, I used the threshold derived from EFSA's 2023. Figure 3 plots the FTR of MDD and EQUIV against the according sample size. MDD starts with an acceptance rate of 60% at a sample size of 2 and rapidly rises to a rate of 100% at a sample size of 10. From a sample size of around 14 it has a constant 100% chance of acceptance. EQUIV doesn't reach the 100% chance of acceptance in none of the data set lengths. From a sample size of 20, the FTR remains between around 65% and 35% for EQUIV. It begins with an acceptance rate of around 25% at a sample size of 2 and grows slightly to a maximum rate of approximately 70% at a sample size of 38. Comparing the 70% acceptance rates of the two methods, MDD requires a sample size of approximately 8 and EQUIV requires a sample size of 38, which is almost five times as much.

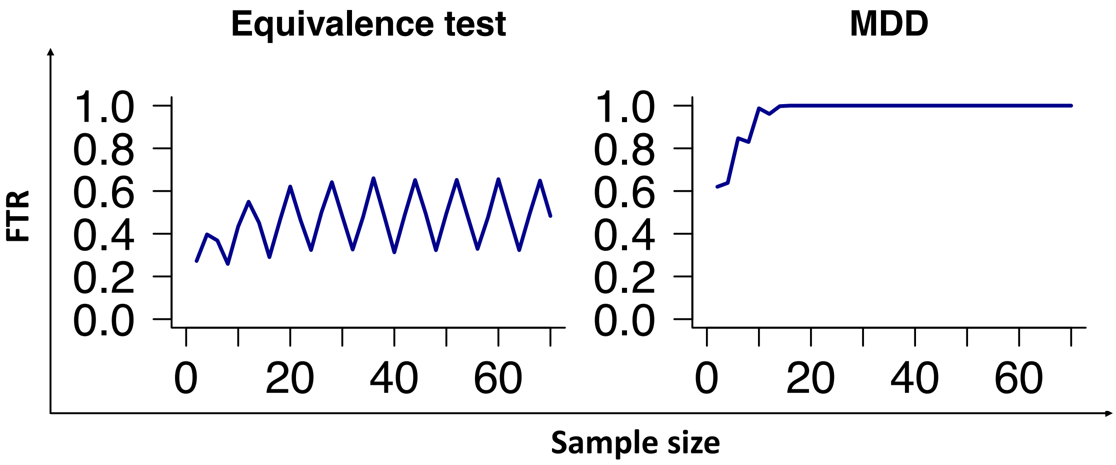

Figure 3: Acceptance Rate of no-effect concentrations in different experimental sample sizes. The y-axis shows the decision rate to falsely trust a non-significant result. It is calculated by averaging the 'Trust no effect' variable for each data set length over the total number of simulation iterations. MDD starts with an acceptance rate of 60% at a sample size of 2 and rapidly rises to a rate of 100% at a sample size of 10. From a sample size of around 14 it has a constant 100% chance of acceptance. EQUIV doesn't reach the 100% chance of acceptance in none of the data set lengths. It begins with an acceptance rate of around 25% at a sample size of 2 and grows slightly to a maximum rate of approximately 70% at a sample size of 38.

5. Discussion and conclusion

In this study, I wanted to find out which statistical method is best suited to discriminate between true negative and false negative tests, and how sample size affects the performance of MDD and equivalence tests in accepting no-effect concentrations.

My main results are that the ability of CI, EQUIV and BF to discriminate between false and true negatives among nonsignificant results decreases with increasing variance. This is because statistical power decreases as the variance of the data increases (McClelland, 2000). A decrease in power is associated with higher type II errors, resulting in higher FMR and FTR (Kaur, 2017). However, the lower FMR and FTR indicate a higher probability of detecting true negatives among non-significant results. For MDD, one would intuitively expect that lower variance would provide more evidence for the absence of an effect, and therefore lead to higher FTR and FMR, but the opposite is the case (Hoenig, 2001). This contradicts the patterns of my MDD curves, which are almost random. My findings about the performance of the MDD disagree those of Mair et al. (2020): They found that MDD determines whether a non-significant result indicates a true lack of effect. This contradiction could be due to something going wrong with my simulation of the MDD. To find the error in my simulation, I tested the use of different variances and thresholds; the behavior of the MDD curve did not change. In addition, instead of calculating the MDD (see Section 2, MDD), I used the MDD function from Mair et al. (2020) but the performance did not change here either. In addition, I found that CI and EQUIV perform equal in identifying true effects when they are present, which is consistent with other researchers who have found that CI and EQUIV are functionally equivalent (Schuirmann, 1987, Tango, 1998). I showed that BF, CI and EQUIV outperform MDD in all cases, having lower FTR and FMR. The reason for this is that MDD is, unlike CI, BF and EQUIV independent of the estimated effect size (Mair, 2021).

Furthermore, my findings show that the new EFSA guidelines require almost five times the sample size to achieve 70% acceptance when using EQUIV than when using MDD. Intuitively, as sample size and power of MDD increase, the FTR should decrease due to the controlled Type II error (less false negatives), but here the FTR increases, leading to a higher number of false negatives. This is since MDD decreases with increasing power (Hoenig, 2001), leading to higher acceptance rates than EQUIV. Intuitively, we would expect to see more evidence of an effect in EQUIV (p-value < significance level) as the sample size increases (Keysers, 2020), leading to a lower FTR. Low p-values did not become more common and the FTR did not change much as the sample size increased. Since EQUIV and BF performed equally well in interpreting non-significant results for a given sample size, it is questionable how the FTR rate of the Bayes factor changes with increasing sample sizes. Unlike EQUIV, the Bayesian t-test provides increasing evidence of no effect with increasing sample size (Keysers, 2020) leading to different FTR.

The threshold that indicates the highest probability of distinguishing false negatives from true negatives depends on the standard deviation. For a standard deviation of 0.5, the best threshold is 0.3 for CI and EQUIV and 0.2 for BF. At a standard deviation of 1, the best threshold is 0.3 for CI and EQUIV and 0.1 for BF.

A couple of questions have arisen from my study: what thresholds should be used to maintain the balance between type I and type II errors and to account for the increase in sample size and power? Furthermore, it is unclear which thresholds should be preferred, those identified as the best thresholds by EQUIV and CI or those selected as the best thresholds by BF. BFs could be investigated in more detail in a further study and an associated guidance on their use in chemical risk assessment could help scientists to better understand the Bayesian approach.

In conclusion, I agree with Colegrave et al. (2003) and Mair et al. (2020) who stated that the best way to interpret non-significant results are CIs. CIs encourage to think about the range of effect sizes that are supported by the data and CIs have reduced error rates (compared to MDD), ensuring reliable results. CIs are more robust to changes in sample size than BF and, unlike EQUIV and BF, are a common method that most scientists have in their statistical toolbox. This makes CIs the most appropriate method for establishing no effect.

Bibliography

Fay, M. P., Halloran, M. E., & Follmann, D. A. (2007). Accounting for variability in sample size estimation with applications to nonadherence and estimation of variance and effect size. *Biometrics*, *63*(2), 465–474.

Aguilar-Marcelino, L., Al-Ani, L. K. T., Wong-Villarreal, A., & Sotelo-Leyva, C. (2023). Persistence of pesticides residues with chemical food preservatives in fruits and vegetables. In *Current Developments in Biotechnology and Bioengineering* (pp. 99–118). Elsevier.

Akobeng, A. K. (2016). Understanding type I and type II errors, statistical power and sample size. *Acta Paediatrica*, *105*(6), 605–609.

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, *18*(2), 127.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Brock, T. C. M., Hammers-Wirtz, M., Hommen, U., Preuss, T. G., Ratte, H. T., Roessink, I., Strauss, T., & Van den Brink, P. J. (2015). The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environmental Science and Pollution Research*, *22*, 1160–1174.

Carvalho, F. P. (2017). Pesticides, environment, and food safety. *Food and Energy Security*, *6*(2), 48–60.

Colegrave, N., & Ruxton, G. D. (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology*, *14*(3), 446–447. https://doi.org/10.1093/beheco/14.3.446

Committee, E. S. (2016). Recovery in environmental risk assessments at EFSA. *EFSA Journal*, *14*(2), 4313.

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

Duquesne, S., Alalouni, U., Gräff, T., Frische, T., Pieper, S., Egerer, S., Gergs, R., & Wogram, J. (2020). Better define beta–optimizing MDD (minimum detectable difference) when interpreting treatment-related effects of pesticides in semi-field and field studies. *Environmental Science and Pollution Research*, *27*, 8814–8821.

Eddy, S. R. (2004). What is Bayesian statistics? *Nature Biotechnology*, *22*(9), 1177–1178.

EFSA. (2010). Scientific opinion on statistical considerations for the safety evaluation of GMOs. *EFSA Journal*, *8*(1), 1250–1311.

EFSA, E. P. on G. M. (2010). Statistical considerations for the safety evaluation of GMOs. *EFSA Journal*, *8*(2), 1250.

EFSA. (2016). *Peer review of the pesticide risk assessment of the active substance flurtamone*. https://doi.org/10.2903/j.efsa.2016.4498

EFSA. (2017). *Updated peer review of the pesticide risk assessment of the active substance flurtamone*. https://doi.org/10.2903/j.efsa.2017.4976

EFSA. (2019). *Outcome of the Pesticides Peer Review Meeting on general recurring issues in ecotoxicology*. https://doi.org/10.2903/sp.efsa.2019.EN-1673

EFSA, E. F. S., Adriaanse, P., Arce, A., Focks, A., Ingels, B., Jölli, D., Lambin, S., Rundlöf, M., Süßenbach, D., & Del Aguila, M. (2023). Revised guidance on the risk assessment of plant protection products on bees (Apis mellifera, Bombus spp. And solitary bees). *EFSA Journal*, *21*(5), e07989.

Engel, J., & van der Voet, H. (2021). Equivalence tests for safety assessment of genetically modified crops using plant composition data. *Food and Chemical Toxicology*, *156*, 112517.

Europäische Kommission. (2013). *Verordnung (EU) Nr. 284/2013 der Kommission vom 1. März 2013 zur Festlegung der Datenanforderungen für Pflanzenschutzmittel gemäß der Verordnung (EG) Nr. 1107/2009 des*

Europäischen Parlaments und des Rates über das Inverkehrbringen von Pflanzenschutzmitteln.

Faller, H. (2004). Signifikanz, Effektstärke und Konfidenzintervall. *Die Rehabilitation*, *43*(03), 174–178.

Fenik, J., Tankiewicz, M., & Biziuk, M. (2011). Properties and determination of pesticides in fruits and vegetables. *TrAC Trends in Analytical Chemistry*, *30*(6), 814–826.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*(3), 647–674.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*(3), 200–206.

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

Harmon, L. J., & Losos, J. B. (2005). The effect of intraspecific sample size on type I and type II error rates in comparative studies. *Evolution*, *59*(12), 2705–2710.

Heaton, C. A. (1993). *The chemical industry*. Springer Science & Business Media.

Hickey, G. (2010). *Ecotoxicological risk assessment: Developments in PNEC estimation*. Durham University.

Hoekstra, J. A., & Van Ewijk, P. H. (1993). Alternatives for the no-observed-effect level. *Environmental Toxicology and Chemistry: An International Journal*, *12*(1), 187–194.

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Hsu, J. C., Hwang, J. G., Liu, H.-K., & Ruberg, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, *81*(1), 103–114.

McClelland, G. H. (2000). *Increasing statistical power without increasing sample size.*

Isenring, R. (2010). Pesticides and the loss of biodiversity. *Pesticide Action Network Europe, London*, *26*.

Källèn, A. (2011). *Understanding biostatistics,* 5-13. John Wiley & Sons.

Kaur, P., & Stoltzfus, J. (2017). Type I, II, and III statistical errors: A brief overview. *International Journal of Academic Medicine*, *3*(2), 268–270.

Kendall, R. J., Anderson, T. A., Baker, R. J., Bens, C. M., Carr, J. A., Chiodo, L. A., Cobb III, G. P., Dickerson, R. L., Dixon, K. R., & Frame, L. T. (2001). Ecotoxicology. *USDA National Wildlife Research Center-Staff Publications*,516.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16.

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

Le, C. T. (2003). *Introductory biostatistics*, 188-208. John Wiley & Sons.

Mahmood, I., Imadi, S. R., Shazadi, K., Gul, A., & Hakeem, K. R. (2016). Effects of pesticides on environment. *Plant, Soil and Microbes: Volume 1: Implications in Crop Science*, 253–269.

Mair, M. M., Kattwinkel, M., Jakoby, O., & Hartig, F. (2020). The minimum detectable difference (MDD) concept for establishing trust in nonsignificant results: A critical review. *Environmental Toxicology and Chemistry*, *39*(11), 2109–2123.

Martínez-Abraín, A. (2008). Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica*, *34*(1), 9–11.

Müller-Cohrs, J. (1990). The power of the Anderson-Hauck test and the double t-test. *Biometrical Journal*, *32*(3), 259–266.

Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, *7*, 103–108.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605.

Nesheim, O. N., & Criswell, J. T. (1978). *Toxicity of pesticides*. Oklahoma State University, Cooperative Extension Service.

Newman, M. C. (2008). “What exactly are you inferring?” A closer look at hypothesis testing. *Environmental Toxicology and Chemistry: An International Journal*, *27*(5), 1013–1019.

Oltmanns, J., Bohlen, M.-L., Escher, S., Schwarz, M., & Licht, O. (2019). Applying a tested procedure for the identification of potential emerging chemical risks in the food chain to the substances registered under REACH-REACH 2: External scientific report. OC/EFSA/SCER/2016/01-CT 1. *EFSA Supporting Publications*, *16*(3), 1597E.

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*(1), 159.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Simundic, A.-M. (2008). Confidence interval. *Biochemia Medica*, *18*(2), 154–161.

Skalski, J. R. (1981). *Statistical inconsistencies in the use of no-observed-effect levels in toxicity testing*. ASTM International.

Smithson, M. (2003). *Confidence intervals* (Issue 140). Sage.

Srivastava, M. K., & Raizada, R. B. (1999). Assessment of the no-observed-effect level (NOEL) of quinalphos in pregnant rats. *Food and Chemical Toxicology*, *37*(6), 649–653.

Stöcklin, M. (n.d.). *Einführung in die Bayessche Statistik*. https://mmi.psychologie.unibas.ch//r-toolbox/Skripte/Bayes%20Einfuehrung.pdf

Swindlehurst, R. J., Johnston, P. A., Tröndle, S., Stringer, R. L., Stephenson, A. D., & Stone, I. M. (1995). Regulation of toxic chemicals in the Mediterranean: The need for an adequate strategy. *Science of the Total Environment*, *171*(1–3), 243–264.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, *17*(8), 891–908.

van der Werf, H. M. (1996). Assessing the impact of pesticides on the environment. *Agriculture, Ecosystems & Environment*, *60*(2–3), 81–96.

Wang, G. M. (1988). Regulatory decision making and the need for and the use of exposure data on pesticides determined to be teratogenic in test animals. *Teratogenesis, Carcinogenesis, and Mutagenesis*, *8*(2), 117–126.

Ware, G. W. (1980). Effects of pesticides on nontarget organisms. *Residue Reviews: Residues of Pesticides and Other Contaminants in the Total Environment*, 173–201.

Warne, M. S. J., & Van Dam, R. (2008). NOEC and LOEC data should no longer be generated or used. *Australasian Journal of Ecotoxicology*, *14*(1), 1–5.

Williams, C. M. (1967). Third-generation pesticides. *Scientific American*, *217*(1), 13–17.

Williams, E. S., Berninger, J. P., & Brooks, B. W. (2011). Application of chemical toxicity distributions to ecotoxicology data requirements under REACH. *Environmental Toxicology and Chemistry*, *30*(8), 1943–1954.

Appendix

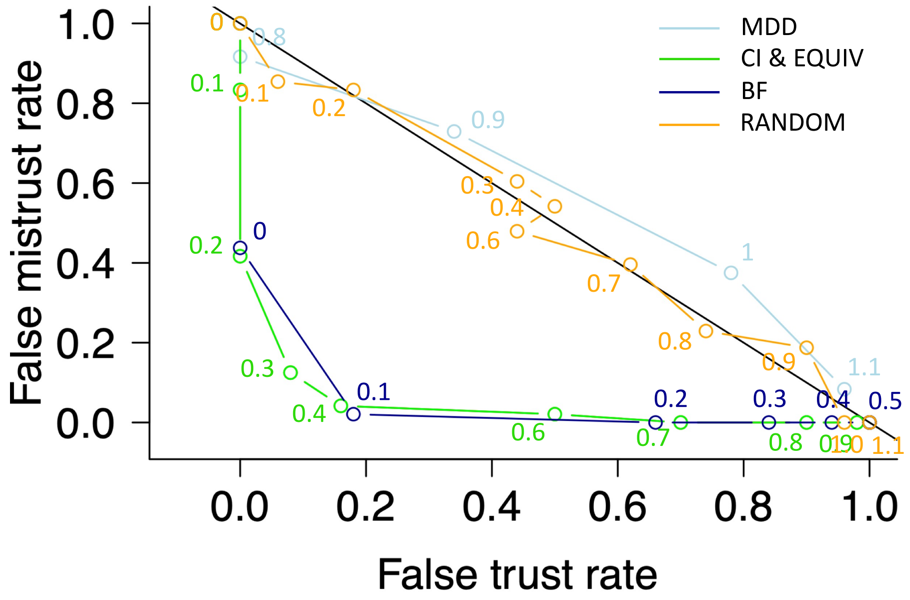

Figure 3: Application of different thresholds to see the tendency of statistical methods to falsely trust or mistrust non-significant results. A threshold (circle with corresponding value) must be chosen to translate random values (Random), the Minimum Detectable Difference (MDD), the Confidence Interval (CI), the Equivalence Test (EQUIV) and the Bayes Factor (BF) into a trust/mistrust decision. This decision allows error rates (False Trust Rate (FTR) and False Mistrust Rate (FMR)) to be calculated for the different methods and thresholds. Each of the three plots shows the FTR/FMR results for each statistical method on data sets with different standard deviations, representing the variation in the samples. With increasing standard deviation, the ability of CI (green line), EQUIV (green line) and BF (dark blue line) to discriminate between trust and mistrust of non-significant results decreases. MDD (light blue line) behaves in all scenarios approximately random (orange line), indicating that MDD does not help to show any tendency in which one would falsely trust or mistrust non-significant results. CI consistently performed the same as EQUIV, regardless of the threshold or variance chosen (both are exactly on the same line, with the same thresholds). The patterns of BF compared to CI & EQUIV were almost indistinguishable, especially at a standard deviation of 0.5 (see Fig. 2). Here, the only difference was the need to use different thresholds for CI & EQIV and BF to achieve equivalent false trust and false mistrust rates.