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

Module: Theoretical Ecology
SS 2023
Chair: Peter Poschlod
Supervision: Florian Hartig

A comparison of statistical methods to establish no effect

Author: Isabelle Halbhuber
Email: Isabelle.Halbhuber@stud.uni-regensburg.de

Abstract

Various statistical methods are designed or used to establish ‘no effect’. The true absence of an effect is not necessarily confirmed by a non-significant (p-value > 0.05) result of traditional null hypothesis significance test (NHST). Extended statistical methods like the minimum detectable difference (MDD), Confidence Intervals (CI), Bayes factors (BF) or Equivalence test (EQUIV) are a tool to establish true no-effect concentrations of potentially harmful chemicals. Chemical exposure needs to be monitored because, if left undetected, the exposure can lead to health and ecosystem damage. To prevent this environmental damage, potentially toxic chemicals are regulated through legislation by policymakers and the expertise of scientists. To scientifically identify potentially toxic chemicals, their fate in ecosystems and their potential harmful effects, statistical approaches based on NHST or Bayesian models can be conducted. These approaches can be used to decide whether a non-significant result is a false negative or indicates the true absence of an effect. However, the determination of the most appropriate statistical method remains an important issue in establishing guidelines for risk assessment. In this study, I compare the above-mentioned statistical methods for their False Mistrust and False Trust Rates (FMR and FTR) to identify the least error-prone statistical method. I explain the theory behind the statistical methods and perform simulations based on the R script by Magdalena M. Mair (2020) to provide an understanding. Finally, I will look at the European Food Safety Authority's (EFSA) switch to the use of Equivalence tests and the consequences for practical application. My main results are that Confidence Intervals and Equivalence tests are performing the same in identifying true effects if they are present and that the new EFSA guidelines with the use of Equivalence tests require more animal testing than in the past. In conclusion, I support the findings of Mair et al. (2020) in recommending the use of CIs for the interpretation of non-significant results.

1. Introduction

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 such as 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, 2010). Since the adverse ecological side-effects of authorized chemicals became known, public awareness and concern have led policymakers to commission ecotoxicologists who monitor the potentially harmful effects of chemicals.

Ecotoxicologists 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. A standard approach to meeting the ecotoxicological risk assessment requirements of REACH is to compare the Predicted Exposure Concentration (PEC) with the Predicted No Effect Concentration (PNEC), resulting in the Risk Characterization Ratio (RCR). The RCR must be low for the chemical substrate to be authorized (Williams, 2011, Hickey, 2010). 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 management, 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 most precise statistical method for ecotoxicological risk assessment is still widely debated. 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 control group using hypothesis testing (e.g. Wang, 1988, Srivastava, 1999). 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, yet the method is still used in ecotoxicological studies (e.g., Pikula, 2023). To overcome the problem of too low power when using NOEC, a statistical analysis for the concentration-response relationship determined by NOEC can be performed using minimum detectable differences (MDD) (Brock, 2015). Duquesne (2020) defined MDD as follows: 'MDD is a measure of the difference between the means of a treatment and a control that must be present to detect a statistically significant effect.' The minimum detectable effect (MDE) is also a post-hoc power analysis and leads to same results as the MDD (Mair, 2020). 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 'equivalent to zero' or not (Lakens, 2017). To establish an 'acceptable low' or 'equivalent to zero' effect concentration, which tests whether the differences between two groups are smaller than a predetermined threshold or not the Equivalence test can be conducted (EQUIV) (Engel, 2021, Lakens, 2017). In addition to post hoc power analyses and threshold comparative analyses, there is also a hypothesis comparative approach, Bayesian Factors, to decide whether there is evidence for the null hypothesis or the alternative hypothesis.

I contrast the abovementioned, statistical methods for the use in ecotoxicological risk assessment and I will investigate 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. Statistical significance tests 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 (HO) 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, a true H0 is erroneously rejected, and when a type II error occurs, a false H0 is erroneously accepted, 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 alpha and type II error has a probability denoted by β (Pollard, 1987).

The power. (1 - β) is the so-called statistical power. A high statistical power indicates a high probability of rejecting the null hypothesis when it is false (Källén, 2011).

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 that is less than, more than or equal to α only means that a deviation from the hypothesis prediction would be as large or larger than the observed one in at most α of the cases if the deviation had only been caused by chance. 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 (Mair, 2020). Trust or mistrust could be determined by ether adding equivalence boundaries or by the conduction of other statistical test like the Bayesian factor analysis or the MDD.

2.2 Post-hoc power analysis

The MDD. As mentioned above, the MDD is 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 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. The MDD considers the power of an experiment and can therefore be used to determine an appropriate sample size (Brock, 2015). I calculated the MDD in R Studio using the above-mentioned function (1).

2.3 Threshold comparative analyses

Equivalence Tests. EFSA (2010) defines equivalence as the absence of differences in a certain parameter between a treated group and an untreated control group. The difference between the groups is regarded as 'acceptable' if it lays within the prespecified, limiting equivalence thresholds (Lakens, 2017). If Δ (>0) is the prespecified acceptable difference between two distributions (the limiting equivalence threshold), the null hypothesis for a one-sided equivalence t-Test and the alternative hypothesis are formulated as follows (Tango, 1998):

$$H_0: Parameter_{treatment} - \Delta - Parameter_{control} \geq 0$$
$$H_1: Parameter_{treatment} - \Delta - Parameter_{control} < 0$$

If effects are significant, then the effect is regarded as equivalent to zero and the corresponding chemical concentration is accepted. 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 used the t.test() function in RStudio to test for equivalence in one direction.

Confidence Intervals. The *confidence interval* 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). I extracted upper CIs from the t.test() function in RStudio.

2.4 Hypothesis comparative analyses

Bayesian Factor analysis. 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, that is the assumption for the probability that the hypothesis is not rejected, then:

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

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

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. To solve these two questions, I generated two different simulations: the False Mistrust Rate/False Trust Rate (FMR/FTR)-simulation and the sample-size-simulation. Both simulations imply various statistical methods (like BF, CI, EQUIV, MDD) to assess the trustworthiness of non-significant (p-value > 0.05) results derived from a one-sided t-test.

Results of the FMR/FTR-simulation were used to calculate type I and type II error rates. Since the error rates and the error control of the above-mentioned statistical methods are different. A Pareto plot (Fig. 2) gives a visual assessment of how well a statistical method discriminates between false positives and false negatives.

The sample-size-simulation implies EQUIV and MDD to analyze data sets with different sample sizes. Considering that European Food Safety Authority (EFSA) has switched from MDD to the use of equivalence tests in their risk assessment, I was interested in finding out how the acceptance of effects is affected by the experimental sample size.

3.1 Simulation

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 is 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 (4:100, step size = 4).

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 > 0.05) results from a one-sided t-test (see Fig. 2: 'Simulation'). If the t-test p-value is greater than 0.05, the 'effect detected' variable is set to 'false', and if the result of the appropriate statistical methods meets

the appropriate assumptions (Bayes factor <= threshold, upper confidence interval <= threshold, equivalence test p-value <= 0.05, 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 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.

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

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

I used the results of the FMR/FTR-simulation to calculate the False Trust Rate (FTR) and the False Mistrust Rate (FMR) for each statistical method. The calculation of the two rates is based on the calculations of Mair et al. (2020) (equations 7 and 8). I plotted the FMR against the FTR (Section 4, Fig. 2) for visual comparison.

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

$$8)\ FMR = \frac{\Sigma\left(\boldsymbol{no}\ effect\ detected\ \&(true\ effect = 0)\right)}{\Sigma(true\ effect = 0)}$$

To examine the dependence of acceptance 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. The result variable 'no effect trusted' is dichotomous, and thus '1' stands for accepted and '0' for not accepted. I plotted the average 'no effect trusted' variable against the length of the corresponding data set (Section 4, Fig. 3) for visual comparison.

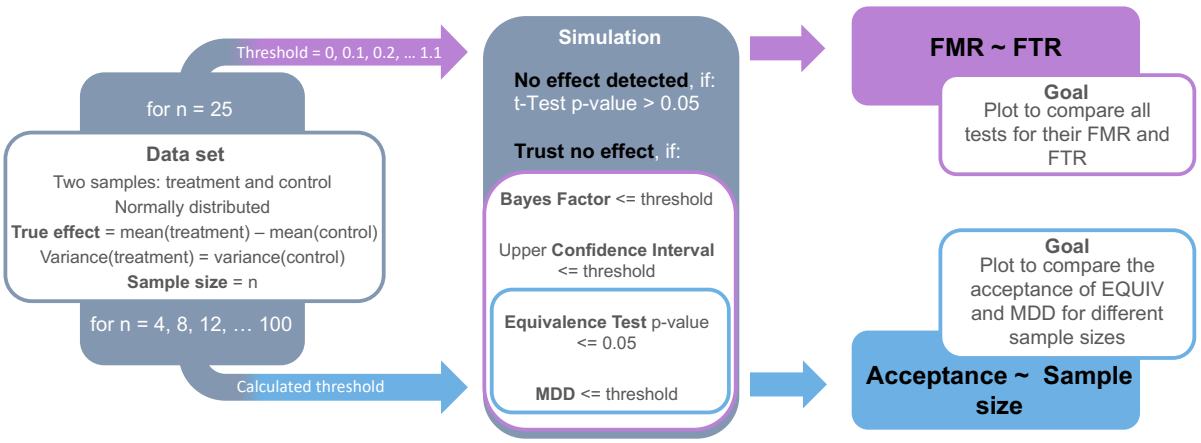

Figure 1: Schematic illustration of the R code workflow. The data sets display control and treatment groups with matching standard deviations, lengths, and normal distributions. Half of the data sets have the treatment group with a higher mean, indicating a larger true effect, while the other half have the same mean compared to the control group. The sample sizes of the generated data sets are either identical (n = 25) or different (n = 4:100, step size = 4) 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 > 0.05) results from an one-sided t-test. If the t-test p-value exceeds 0.05, 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 <= 0.05, 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 and Discussion

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 consistently performed exactly like EQUIV, regardless of the threshold or variance chosen. With increasing variance, the ability to discriminate between false trust and false mistrust of non-significant results of CI (Fig. 3, green line), EQUIV (Fig. 3, green line) and BF (Fig 3, dark blue line) decreased. This decrease is because statistical power decreases as the variance of the data increases (McClelland, 2000). A reduction in power is associated with higher type II error, leading to higher FMR/FTR rates (see Section 2, 'Power'). The patterns of BF compared to CI & EQUIV were almost indistinguishable, especially at a standard deviation of 0.5 (see Fig. 2). The difference between them is the need to use different thresholds for CI, EQIV and BF to achieve equivalent false trust and false mistrust rates. BF, CI & EQUIV outperformed MDD in all cases, showing much lower FTR and FMR than MDD. MDD is therefore less powerful for detecting true negatives among nonsignificant results than BF, CI & EQUIV. The reason for this might be that MDD is defined independently of the estimated effect size (Mair, 2021). MDD appears to behave approximately random (see Appendix, Fig.4: Random), which would imply that MDD cannot systematically discriminate between trust and mistrust, meaning that MDD

does not help to show any tendency in which one would falsely trust or mistrust non-significant results. Therefore, the use of MDD in general would be questionable. My findings on the performance of MDD contradict 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 stimulation of the MDD. To find the error in my simulation, I tested the use of different variances, thresholds, and effect sizes; the behavior of the MDD curve in the Pareto plot 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 behavior did not change here either.

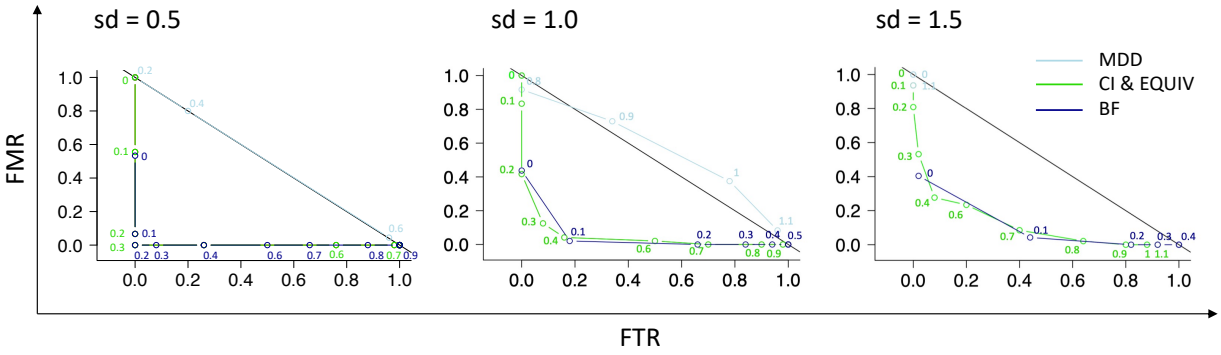

Figure 2: Application of different thresholds to see the tendency of statistical methods to falsely trust or mistrust non-significant results. A threshold (circle with corresponding value) must be chosen to translate the Minimum Detectable Difference (MDD), the Confidence Interval (CI), the Equivalence 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 (both are on the same line, with the same thresholds). BF, CI & 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 & EQUIV were almost indistinguishable, especially at a standard deviation of 0.5. Here, the only difference was the need to use different thresholds for CI & EQIV and BF to achieve equivalent FMR and FTR.

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 <= 0.05) 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 has 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 Equivalence tests 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 (α = 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 (equation 5) derived from EFSA's 2010 statistical considerations. Figure 3 plots the decision to accept or reject no-effect (y-axis, 0.0 = not accepted and 1.0 = accepted) against sample size. The MDD accepts no-effect concentrations from a sample size of 8 on, whereas the Equivalence test starts to accept no-effect concentrations from a sample size of 24 and becomes robust from a sample size of 60. The graphs show that the new EFSA guidelines, which require the use of equivalence tests, result in a higher number of animals tested to establish no-effect concentrations. Compared to other statistical tests MDD has been shown to have relatively high power. The power of MDD is even increasing with decreasing sample size (Mair, 2020). This is an important feature, because ecotoxicological risk assessment aims to minimise the number of higher tier experiments.

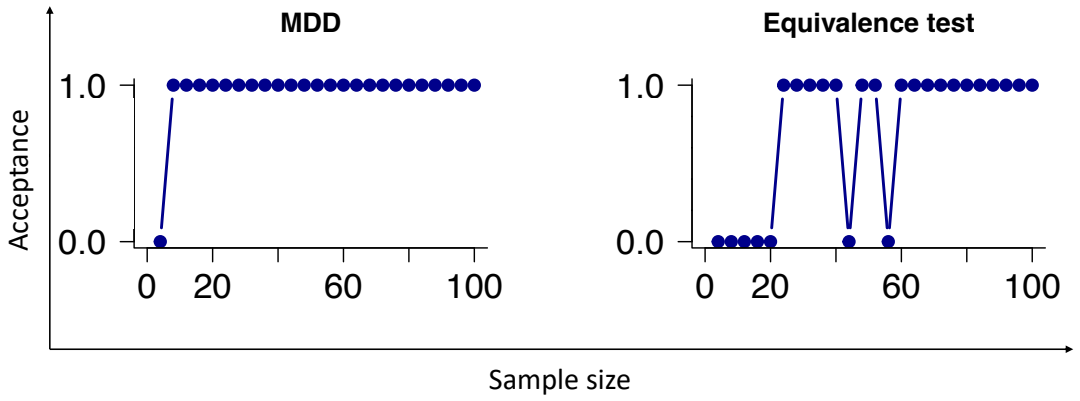

Figure 3: Acceptance of no-effect concentrations in different experimental sample sizes. The y-axis displays the decision to accept or to reject no-effect: 0.0 == not accepted and 1.0 == accepted. The Minimum Detectable Difference (MDD) accepts no-effect concentrations from a sample size of 8 on, whereas the Equivalence test starts to accept no-effect concentrations from a sample size of 24 and becomes robust from a sample size of 60 on.

5. Summary and conclusion

In this study I wanted 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 accepting no effect concentrations. My main results are that Confidence Intervals and Equivalence tests are performing the same in identifying true effects if they are present and that the new European Food Safety Authority (EFSA) guidelines with the use of Equivalence tests require more animal testing than in the past. In conclusion, I support the findings of Mair et al. (2020) in recommending the use of CIs for the interpretation of non-significant results. The reduced error rates of CIs (and Equivalence tests) ensure reliable results and unlike equivalence tests, CIs are a common method that most scientists have in their statistical toolbox.

Bibliography

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.

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

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*.

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.

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.

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.

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.

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.

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.

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.

Pikula, K., Johari, S. A., Santos-Oliveira, R., & Golokhvast, K. (2023). The Comparative Toxic Impact Assessment of Carbon Nanotubes, Fullerene, Graphene, and Graphene Oxide on Marine Microalgae Porphyridium purpureum. *Toxics*, *11*(6), 491.

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

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.

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

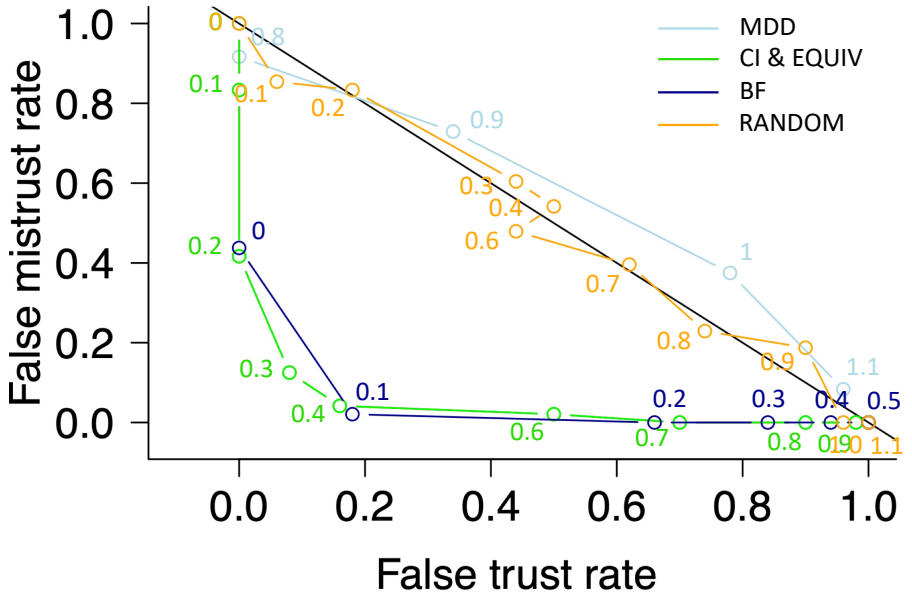

Figure 4: Application of different thresholds to see the tendency of statistical methods to falsely trust or mistrust non-significant results. A threshold (circle with corresponding value) must be chosen to translate random values (Random), the Minimum Detectable Difference (MDD), the Confidence Interval (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