Would you agree if N is three? On statistical inference for small N.

Eleni Psarou[1], Christini Katsanevaki[1,2], Eric Maris[3,*], Pascal Fries[1,2,3,4,*]

[1] Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, 60528 Frankfurt, Germany

[2] International Max Planck Research School for Neural Circuits, 60438 Frankfurt, Germany

[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, 6525 EN Nijmegen, the Netherlands

[4] Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

* These authors contributed equally to this study.

Abstract

Non-human primate studies traditionally use two or three animals. We previously used standard statistics to argue for using either one animal, for an inference about that sample, or five or more animals, for a useful inference about the population. A recently proposed framework argued for testing three animals and accepting the outcome found in the majority as the outcome that is most representative for the population. The proposal tests this framework under various assumptions about the true probability of the representative outcome in the population, i.e. its typicality. On this basis, it argues that the framework is valid across a wide range of typicalities. Here, we show (1) that the error rate of the framework depends strongly on the typicality of the representative outcome, (2) that an acceptable error rate requires this typicality to be very high (87% for a single type of outlier), which actually renders empirical testing beyond a single animal obsolete, (3) that moving from one to three animals decreases error rates mainly for typicality values of 70-90%, and much less for both lower and higher values. Furthermore, we use conjunction analysis to demonstrate that two out of three animals with a given outcome only allow to infer a lower bound to typicality of 9%, which is of limited value. Thus, the use of two or three animals does not allow a useful inference about the population, and if this option is nevertheless chosen, the inferred lower bound of typicality should be reported.

Introduction

We have recently argued that a sample of two or few (less than five) subjects allows useful inferences only about that sample but not about the population (Fries and Maris, 2022). On this basis, we recommended to either use a sample of one and make an inference about that sample, or to use a sample of five or more animals and make a useful inference about the population. We discouraged using the traditional approach in non-human primate (NHP) research of using two or three animals, because this doubles or triples the number of animals, while the inference remains limited to the sample.

Subsequently, a framework was suggested that attempts to draw an inference about the representativeness of an outcome in the population by studying a small number of animals. This framework (1) suggests to "assume that, in each animal, an experiment can lead to a number of qualitatively distinct outcomes", (2) suggests to "assume a prior distribution across all possible outcomes", calling "the most likely outcome the 'representative' outcome and other 'outliers'", (3) considers outlier proportions in the range of 10 to 20%, (4) considers an outcome as representative when the outcome is present in the majority of the tested animals, (5) concludes that requiring two

out of three subjects to show an effect strikes an efficient balance between the proportion of correct conclusions and inconclusive outcomes, (6) claims that this conclusion holds "across a wide range of prior distributions" (Laurens, 2022). Here, we critically discuss the proposed framework and conclude that it has serious shortcomings. We note that the term "prior distribution" should not be confused with a prior distribution in a Bayesian framework, and that the framework of Laurens actually does not allow to specify a prior distribution in the Bayesian sense. Most importantly, we show that the framework will only produce an acceptable inference for a narrow range of outcome distributions. We also present a way to estimate a lower bound to the representativeness of an outcome, known as typicality, for a range of M animals tested and N animals showing the outcome. We recommend that this lower bound of typicality is reported in studies that try to draw an inference about the population based on a small number of animals.

Discussion

The concept of typicality.

The probability of a given test outcome in a population is referred to as the typicality of that outcome in that population. The concept of typicality is central to a technique that has been called conjunction analysis (Friston et al., 1999). A conjunction analysis first tests for a given outcome (e.g. the presence of an effect or a trait) in each subject of a limited sample drawn from a population. It then uses the proportion of the sample with a given outcome to draw an inference about the proportion of the population that would give the same outcome. This proportion of the population is called typicality γ. The true value of γ cannot be known, but a useful lower bound to typicality, γ_c, can be estimated if the false-positive and false-negative rates of the employed tests are specified (Friston et al., 1999), and this is explained in more detail below.

The N-oo-M framework.

The framework proposed by Laurens (2022) starts by testing each investigated subject for a given outcome, and counting in a sample of M subjects the number N of subjects showing that outcome. The framework refers to this as N-out-of-M, or N-oo-M, and refers to e.g. 2 out of 3 tested subjects showing a given outcome as 2-oo-3. After counting the number of animals with a given outcome, the framework aims at drawing a binary inference about which outcome is representative of the population. The framework defines an outcome as representative if it is present in the majority of the tested subjects, i.e. at least two subjects should show the effect in the 2-oo-3 case.

Importantly, the N-oo-M framework suggests to "assume a prior distribution across all possible outcomes", calling "the most likely outcome the 'representative' outcome and other 'outliers'". Note that assuming "a prior distribution across all possible outcomes" is not an assumption in the usual sense, because an outcome always has some probability/typicality, and this just follows from the fact that it is a random variable. Conjunction analysis considers one such probability as the parameter of interest and derives a lower bound for this probability/typicality.

The use of the term "prior distribution" may cause confusion among readers that are familiar with Bayesian inference. The framework of Laurens (2022) is not Bayesian, because if it were Bayesian, one would have to specify a prior probability distribution for the parameters, and here this would be typicality γ (which in turn specifies the distribution of the outcomes, with probability γ for the representative outcome, and $1-\gamma$ for the outlier). Because this is a probability, the prior distribution would have a support over the interval [0,1], and usually this is the beta distribution. Combining this prior with the information in the data produces a posterior distribution (via Bayes' rule) that is a reweighting over the interval [0,1]: segments that were a priori likely/unlikely to contain the true

typicality value can be down/upweighted according to the information in the data. Crucially, if this prior were not a distribution over the interval [0,1] but a fixed value (e.g., $\gamma = 0.8$), there would be no space for down/upweighting. If γ is a fixed value and it is known, then no data are required for estimating it. If, on the other hand, γ is a fixed unknown value, as in frequentist inference, then data can be used to estimate its value (e.g., by means of a confidence interval). The framework of Laurens (2022) is neither Bayesian nor frequentist; its goal is to estimate the representative outcome (i.e., the event that corresponds to $\gamma > 0.5$).

The probability of incorrect conclusions depends on the assumed probability of outliers.

If the N-oo-M framework correctly identifies the representative outcome, it defines this as "correct conclusion". Here, we refer to the probability of correct conclusions as π, to the probability of incorrect conclusions (excluding cases considered "inconclusive" by Laurens (2022)) as δ, and the probability of outliers as ω. The N-oo-M framework suggests specifying the probability of outliers as a "prior" (in the sense of Laurens (2022), see above) and claims that "the N-out-of-M model leads to a similar conclusion across a wide range of prior distributions." However, Laurens (2022) considers ω-values only in a relatively narrow range of 10% to 20%, and allows for percentages to accumulate over different outlier types, including experimental errors. We argue that both, the distributions of outlier probabilities over outlier types, and the total outlier probability, considered by Laurens (2022) are arbitrary.

Fig. 1C of Laurens (2022) shows that with a single type of outlier and with ω = 10%, 2-oo-3 reaches δ = 2.8%. However, our Fig. 1A shows that for larger values of ω, the δ increases steeply (light blue curve in our Fig. 1A). Already for an ω value of 20%, the δ rises to 10.4%, which is more than twice the generally accepted error rate of 5%. Laurens (2022) actually rejected 1-oo-1 at ω = 10%, because of δ = 10%.

Laurens (2022) does actually consider an example with a total ω = 20%, yet this ω is distributed over three types of outliers occurring at 10%, 5% and 5%, respectively (Fig. 2A of Laurens (2022)), and this specific distribution leads to a δ just below 5%, namely at 4.3%. However, we note that this scenario additionally entails a 6.1% possibility of inconclusive cases. Even more worrisome is the fact that this still favorable outcome depends on the precise distribution of ω over several outlier types. If a total ω of 20% would result from only two types of outliers of 15% and 5%, then δ would be 6.8% and thereby higher than the accepted error rate of 5%.

A high and immutable prior on the typicality of representatives renders experiments obsolete.

Considering all possible combinations of several outlier types is intractable, so we focus in the following on the simple case of one outlier type, yet the reasoning holds for more than one outlier type with correspondingly adjusted numerical results. The blue line in Fig. 1A shows that δ starts exceeding the generally accepted error rate of 5% for a probability of outliers (ω-value) above 13%, corresponding to a typicality of representatives below 87%. Thus, the 2-oo-3 framework only produces an acceptable error rate for a typicality of representatives of 87% or higher. Here, we point out that if the typicality of representatives has to be >87%, this makes it meaningless to collect empirical data beyond a single animal (one single animal being needed to find out which outcome is actually typical). Remember that the framework aims at a binary decision about which outcome is representative for the population. Yet, for a typicality of representatives >87%, this decision is obsolete.

2-oo-3 versus 1-oo-1.

The core of the argument presented in favor of the proposed 2-oo-3 framework is a reduction of δ for 2-oo-3 compared to 1-oo-1. The δ values for 1-oo-1 are shown in Fig. 1A as red line, for 2-oo-3 as blue line, and their difference is shown in Fig. 1B. Here, we point out that the reductions of δ that are obtained by moving from 1-oo-1 to 2-oo-3 are a function of ω. They peak for ω values around 10-30%, close to the ones chosen by Laurens (2022), but they strongly diminish for both larger and smaller ω values.

Two to three animals allow only limited inferences about the population.

The N-oo-M framework can be evaluated through the perspective of conjunction analysis. Conjunction analysis makes no assumption about the typicality of an outcome, but instead it makes an inference about the typicality. More specifically, conjunction analysis uses the proportion of the sample that shows a given outcome to infer the lower bound of typicality for that outcome (Friston et al., 1999). This inference is directly related to the statistics of the binomial distribution, which is defined as follows:

$$P(N = i) = \binom{M}{i} p^i (1-p)^{M-i} \quad [1]$$

A binomial distribution with parameters *M* and *p* is the discrete probability distribution of *N* successes in a sequence of *M* independent experiments, with a success probability *p* (adapted from Wikipedia contributors (2023)).

When *p* is not known, but *M* and *N* are known, we can calculate a confidence interval for p (Clopper and Pearson, 1934). Fig. 2A shows the two-sided 95% confidence intervals for the success rate p as a function of the ratio N/M of animals showing the effect, for *M* = 1, 2, 3, 4, 5, 10 tested animals (for *M*≥10, see Fig. 4 of Clopper and Pearson (1934)). This figure nicely illustrates that increasing numbers of animals lead to decreasing widths of the confidence intervals, and an infinite number of animals would let the confidence interval shrink to the diagonal.

We now have to consider that the outcomes of statistical tests in individual subjects are imperfect and are characterized by (1) a false-positive rate, α, of the individual tests, and (2) a true-positive rate, β, of the individual tests, also referred to as sensitivity. Thus, the probability *p* of a significant statistical test is not identical to the true typicality γ in the population, but is a monotonic function of γ, parameterized by α and β:

$$p = (\gamma * \beta) + ((1-\gamma) * \alpha) \quad [2]$$

, leading to

$$p = \alpha + (\beta - \alpha)\gamma \quad [3]$$

, and then to

$$\gamma = \frac{p - \alpha}{\beta - \alpha} \quad [4]$$

In this equation, α is the false-positive rate of the tests, which is typically set to 0.05. Thus, a small proportion of significant tests are false positives, slightly reducing the estimated typicality. The sensitivity β cannot be determined, because it is the proportion of significant tests that are truly

positive, and we have no access to this truth. If we would assume that sensitivity was less than one, a nonsignificant test could reflect a false-negative test (and therefore essentially be discarded), allowing for any typicality to be consistent with the outcomes of the tests in the sample. This can be concluded directly from the formula: If we chose β to be less than one, the resulting lower bound γ_c would increase without limit, which is obviously meaningless. Therefore, we need to make the conservative assumption that sensitivity is one, as has been argued before (Friston et al., 1999). This paragraph is adapted from Fries and Maris (2022).

By plugging the standard assumptions for α and β, and the 95% confidence intervals for the success rate p (see Fig. 2A), into equation [4], we obtain the two-sided 95% confidence intervals for the typicality, expressed as a function of the ratio *N*/*M* (Fig. 2B).

Because we want to be conservative, from these confidence intervals, we need to use the lower bound for typicality, which we refer to as γ_c. Fig. 3A and Table 1 show γ_c as a function of *N* given different values of *M*, and reveal that γ_c for 2-oo-2 is merely 11%, and γ_c for 2-oo-3 is merely 5%. We have previously proposed that 50% is the lowest useful value for typicality, because it corresponds to the expected presence of an effect in a simple majority of the population (Fries and Maris, 2022).

We so far only used two-sided confidence intervals. However, even when using one-sided confidence intervals (Friston et al., 1999; Fries and Maris, 2022), γ_c for 2-oo-2 is merely 18%, and γ_c for 2-oo-3 is merely 9%, thus far below 50% (Fig. 3B and Table 2).

Table 1 and Table 2 report the precise γ_c values for all possible outcomes up to N=M=10, to aid the reporting of γ_c values in studies that opt for this approach.

Investigators might choose a lower false-positive rate α for their individual tests, and this will increase the resulting γ_c. However, even strong reductions of α leave γ_c far below 50%, for both 2-oo-2 and 2-oo-3 (Fig. 3C).

It might be argued that the assumption of $\beta = 1$ is necessary on theoretical grounds (see above), but that this assumption will also be wrong, because the true sensitivity will most likely never be perfect. Therefore, we explored the influence of lowering the assumption for β. As mentioned above, reducing β to arbitrarily low values will increase γ_c without limit and is therefore meaningless. Nevertheless, we considered reductions of β to a value of 0.5, which means that the test in individual subjects is so insensitive that it misses half of the subjects with an effect (or trait). Even with such strong reductions of β, and with α already reduced to 0.01, the estimate of γ_c remains below 50% (Fig. 3D), both for the 2-oo-2 and for the 2-oo-3 case.

Thus, the 2-oo-3 framework leads to γ_c values that correspond to inferences about very limited proportions of the population. Useful inferences based on 2-oo-2 or 2-oo-3 remain limited to the sample of investigated animals (Fries and Maris, 2022). Such an inference about the sample of animals is also reached with 1-oo-1. Therefore, compared to 1-oo-1, the proposal of 2-oo-2 or 2-oo-3 does not provide a gain in the quality of the inference, while at the same time doubling or even tripling the number of animals used.

Conclusion

In summary, the framework proposed by Laurens (2022) has been an unconventional and welcome addition to the discussion about statistical inferences based on small numbers of animals. However, our analysis revealed that it has serious shortcomings and limitations. If studies nevertheless choose to report the individual outcomes of two or three animals, they should also report the corresponding lower bound of typicality (see Tables 1 and 2) to avoid the common misconception that the inclusion

of a second or third animal would allow a general inference about the population. We maintain the previous conclusion that a useful inference about the population requires at least five animals (Fries and Maris, 2022). This number is currently not realized in typical NHP experiments. Therefore, any useful inference will remain limited to the investigated sample, and this will hold for a sample of three or two animals, or even a single animal. Consequently, we argue that the minimum required number for the publication of a typical NHP study should be one animal, to minimize the use of animals in research.

Table 1. The lower bound (γ_c, expressed as percentage) of the two-sided 95% confidence interval for the typicality (γ), as a function of the number of animals showing an effect (N) out of the number of tested animals (M), with α= 0.05 and β=1.

M \ N	0	1	2	3	4	5	6	7	8	9	10
1	0	0									
2	0	0	11.4								
3	0	0	4.7	25.5							
4	0	0	1.9	15.2	36.6						
5	0	0	0.3	10.2	24.6	45.1					
6	0	0	0	7.2	18.2	32.5	51.7				
7	0	0	0	5.2	14.1	25.3	39.1	56.9			
8	0	0	0	3.7	11.3	20.5	31.5	44.6	61.1		
9	0	0	0	2.6	9.2	17.1	26.2	36.8	49.2	64.6	
10	0	0	0	1.8	7.5	14.4	22.4	31.3	41.5	53.2	67.5

Table 2. The lower bound (γ_c, expressed as percentage) of the one-sided 95% confidence interval for the typicality (γ), as a function of the number of animals showing an effect (N) out of the number of tested animals (M), with α= 0.05 and β=1.

M \ N	0	1	2	3	4	5	6	7	8	9	10
1	0	0									
2	0	0	18.3								
3	0	0	9.0	33.5							
4	0	0	5.0	20.9	44.5						
5	0	0	2.8	14.7	30.8	52.6					
6	0	0	1.4	10.9	23.3	38.8	58.6				
7	0	0	0.4	8.3	18.5	30.7	45.2	63.4			
8	0	0	0	6.4	15.0	25.2	36.9	50.5	67.1		
9	0	0	0	5.0	12.5	21.2	31.0	42.1	54.8	70.2	
10	0	0	0	3.9	10.5	18.2	26.7	36.1	46.6	58.5	72.8

Acknowledgements

We thank Jean Laurens for helpful discussions.

Declaration of Interests

P.F. has a patent on thin-film electrodes and is member of the Advisory Board of CorTec GmbH (Freiburg, Germany). The other authors declare to have no competing interests.

Figure legends

Figure 1: (A) The percentage of incorrect conclusions (δ) as a function of the percentage of the assumed typicality of representatives (bottom x-axis) and the assumed probability of outliers (ω, top x-axis), for a single type of outliers, and separately for 2-oo-3 (blue) and 1-oo-1 (red). (B) The difference in δ for 2-oo-3 versus 1-oo-1. The grey shading indicates the range of outlier proportions considered in Laurens (2022).

Figure 2: (A) Two-sided 95% confidence intervals (CI) for the success rate (*p*, expressed in percentage) as a function of the ratio of animals with the trait, following Clopper and Pearson (1934). The color legend specifies the different numbers *M* of tested animals. For each *M*, two lines are plotted, corresponding to the upper and the lower limit of the two-sided 95% confidence interval. (B) Same as (A), but for the typicality γ.

Figure 3: (A) The lower bound (γ_c, expressed as percentage) of the two-sided 95% confidence interval for the typicality (γ), as a function of *N*, the number of animals showing an effect (i.e. an individually significant test), and *M*, the number of animals tested. The false-positive rate α was set to 0.05, and the sensitivity β was assumed to be 1, as explained in the main text. (B) Same as (A), but using one-sided 95% confidence intervals. (C) γ_c as a function of the false-positive rate α, with the sensitivity β fixed at a value of 1, and using one-sided 95% confidence intervals. (D) γ_c as a function of the sensitivity β, with the false-positive rate α fixed at a value 0.01, and using one-sided 95% confidence intervals. Higher values of α, like the standard value of 0.05, would lead to even lower γ_c. This plot of γ_c as a function of decreasing β is merely to illustrate the effect, while we maintain that β needs to be assumed to be 1, as previously discussed (Friston et al., 1999; Fries and Maris, 2022). For (C) and (D): If we had assumed two-sided 95% confidence intervals, the values of γ_c would be even lower.

References

Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404-413.

Fries P, Maris E (2022) What to Do If N Is Two? Journal of cognitive neuroscience 34:1114-1118.

Friston KJ, Holmes AP, Worsley KJ (1999) How many subjects constitute a study? NeuroImage 10:1-5.

Laurens J (2022) The statistical power of three monkeys. bioRxiv:2022.2005.2010.491373.

Wikipedia contributors (2023) Binomial distribution. In: Wikipedia, The Free Encyclopedia.

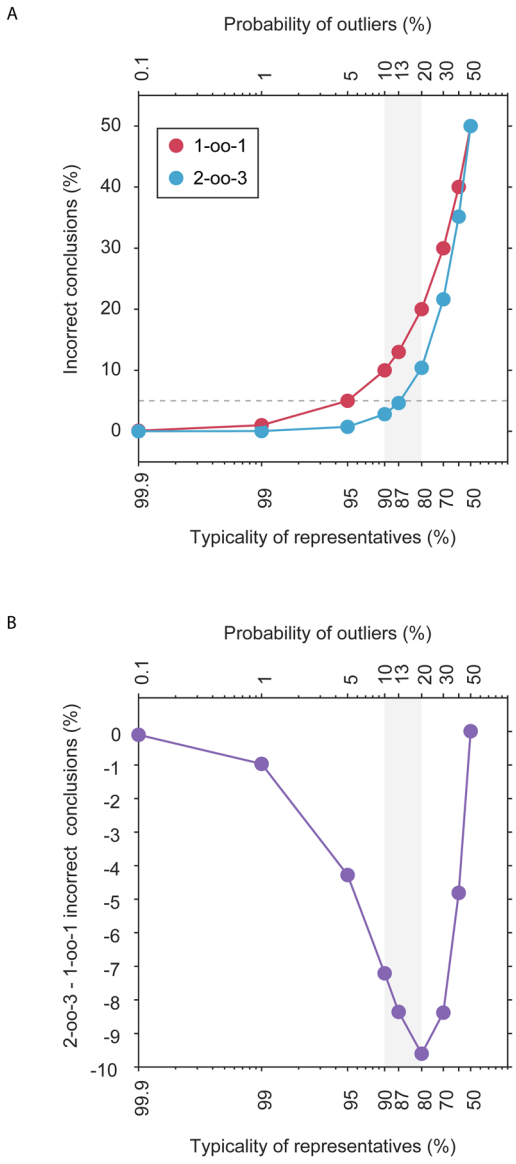

Figure 1: (A) The percentage of incorrect conclusions (δ) as a function of the percentage of the assumed typicality of representatives (bottom x-axis) and the assumed probability of outliers (ω, top x-axis), for a single type of outliers, and separately for 2-oo-3 (blue) and 1-oo-1 (red). (B) The difference in δ for 2-oo-3 versus 1-oo-1. The grey shading indicates the range of outlier proportions considered in Laurens (2022).

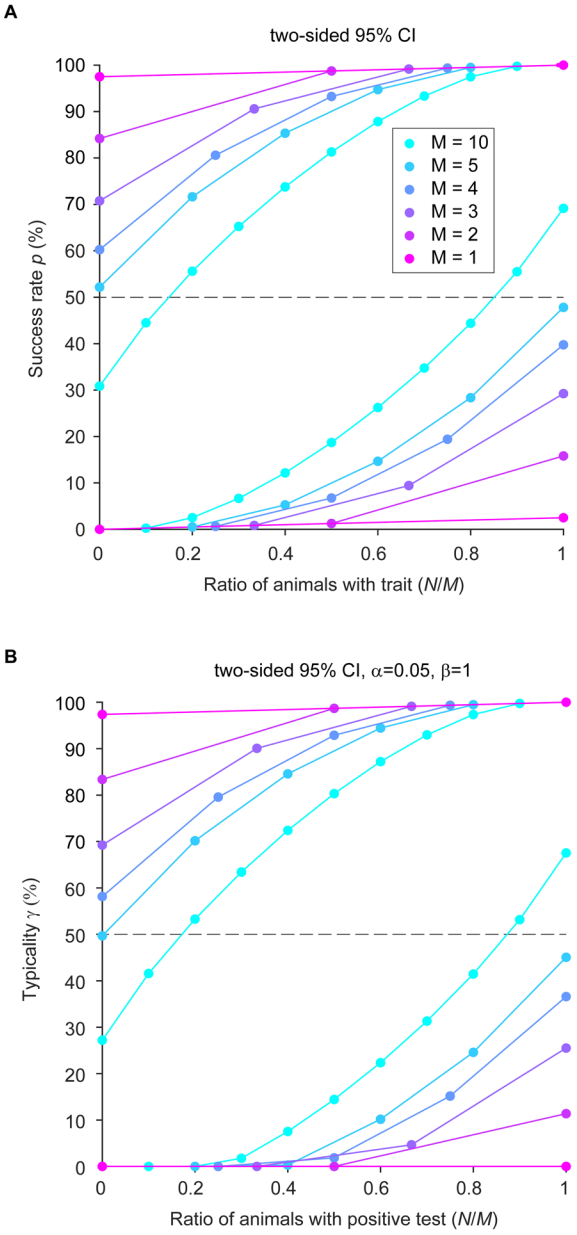

Figure 2: (A) Two-sided 95% confidence intervals (CI) for the success rate (p, expressed in percentage) as a function of the ratio of animals with the trait, following Clopper and Pearson (1934). The color legend specifies the different numbers M of tested animals. For each M, two lines are plotted, corresponding to the upper and the lower limit of the two-sided 95% confidence interval. (B) Same as (A), but for the typicality γ.

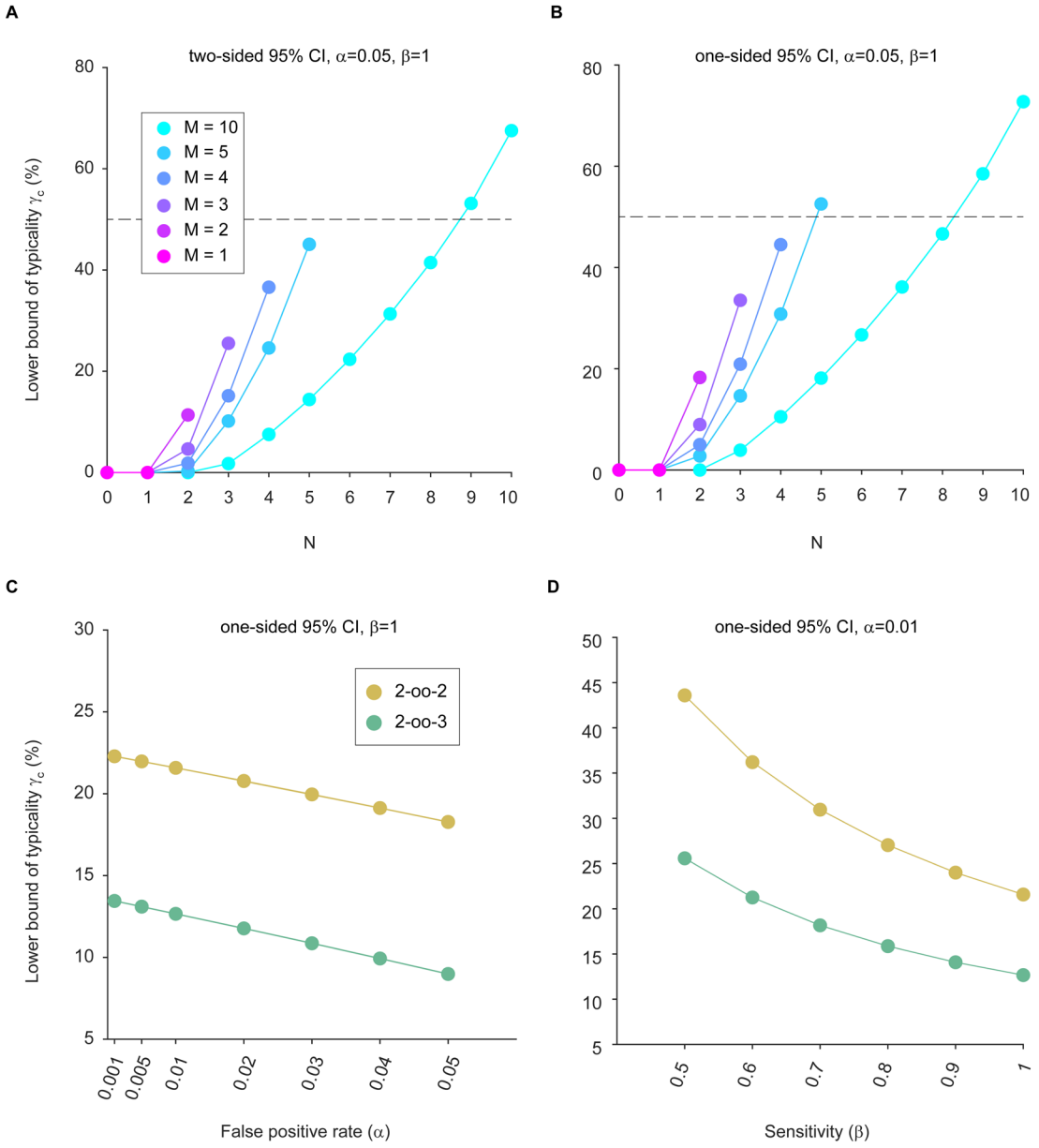

Figure 3: (A) The lower bound (γ_c, expressed as percentage) of the two-sided 95% confidence interval for the typicality (γ), as a function of *N*, the number of animals showing an effect (i.e. an individually significant test), and *M*, the number of animals tested. The false-positive rate α was set to 0.05, and the sensitivity β was assumed to be 1, as explained in the main text. (B) Same as (A), but using one-sided 95% confidence intervals. (C) γ_c as a function of the false-positive rate α, with the sensitivity β fixed at a value of 1, and using one-sided 95% confidence intervals. (D) γ_c as a function of the sensitivity β, with the false-positive rate α fixed at a value 0.01, and using one-sided 95% confidence intervals. Higher values of α, like the standard value of 0.05, would lead to even lower γ_c. This plot of γ_c as a function of decreasing β is merely to illustrate the effect, while we maintain that β needs to be assumed to be 1, as previously discussed (Friston et al., 1999; Fries and Maris, 2022). For (C) and (D): If we had assumed two-sided 95% confidence intervals, the values of γ_c would be even lower.