What to Do If N Is Two?

Pascal Fries\textsuperscript{1,2} and Eric Maris\textsuperscript{2}

Abstract

The field of in vivo neurophysiology currently uses statistical standards that are based on tradition rather than formal analysis. Typically, data from two (or few) animals are pooled for one statistical test, or a significant test in a first animal is replicated in one (or few) further animals. The use of more than one animal is widely believed to allow an inference on the population. Here, we explain that a useful inference on the population would require larger numbers and a different statistical approach. The field should consider to perform studies at that standard, potentially through coordinated multicenter efforts, for selected questions of exceptional importance. Yet, for many questions, this is ethically and/or economically not justifiable. We explain why in those studies with two (or few) animals, any useful inference is limited to the sample of investigated animals, irrespective of whether it is based on few animals, two animals, or a single animal.

INTRODUCTION

In awake monkey neurophysiology, there is the saying: “One monkey is an anecdote, two monkeys are perpetual truth.” This captures the current publication standard in the field of non-human-primate research. Yet, the situation is not unique to this field. Rather, in vivo neurophysiology studies generally and also parts of psychophysics tend to report data from very few participants per study. In fact, studies with a low number of participants appear in every field where the number of participants is restricted by either ethical or economical considerations. Such studies were also common in the early days of human brain mapping, when this field relied heavily on the use of radiotracers for PET. Partly because fMRI became widely available, the field considered the role of sample size several decades ago, made a fundamental distinction between random- and fixed effects analyses, and developed the notion of conjunction analysis. We aim to open up these fundamental statistical concepts to any field of neuroscience that typically uses two (or few) animals or participants. We do not aim here to present novel or even optimal statistical techniques. Rather, we focus on standard statistical approaches and explain what can and what cannot be inferred from them when the sample size is two or few. The insights are particularly relevant for any study using animals, because the number of used animals ought to be reduced as far as possible for ethical reasons. We will therefore mainly refer to (the number of) animals, yet the arguments apply equally to any field of science that bases statistical tests on few participants. In the following, we take “few” to mean “less than five,” although this exact number does not play a crucial role in our arguments.

Statistical tests are often reported for each of few animals, and the qualitative similarity between animals is mentioned, or the data from few animals are pooled. In both cases, the fact that more than one animal is involved is usually considered crucial. The rationale for requesting at least a second animal is that it provides a replication. In addition, the rationale for a replication is that this allows further-reaching inferences. Any scientific study on animals ideally aims for an inference on the population from which these animals are sampled, rather than an inference that is limited to the studied sample of animals. However, does a replication in two to few animals permit this population-level inference? The main messages of this article are 1) that replication across few animals does not allow for a formally well-founded and useful inference on the population of animals; 2) that such an inference would require more animals to gain acceptable sensitivity; 3) that the simple combination of data across animals (i.e., treating the data as if they were generated by a single animal) is often the best one can do; and 4) that the resulting inference is limited to the studied sample and therefore does not qualitatively differ between few, two, or even a single animal.

Fixed Effects versus Random Effects Tests

The argument starts with the basic rationale by which statistical tests draw inferences. This rationale leads us to the important difference between fixed- and random effects tests, which is directly relevant to whether one can infer on a population of animals. Let us consider a study that has recorded, in one animal, the firing rates of...
100 neurons, each in two experimental conditions. The study compares the two conditions and finds an effect: The average firing rate (with averaging over the neurons) is higher in Condition B than A, and a properly controlled statistical test gives a $p$ value of less than .05. This allows the following inference: This observation on 100 neurons would have a probability of less than 0.05, if the average effect calculated on all neurons that could be recorded from this animal would be zero, that is, if the null hypothesis were true. Therefore, we make the inference that the effect is most likely present, when calculated on all neurons.

The same would hold if the unit of observation (UO, the entity to which the observations pertain; Maris, 2012) is not the neuron, but, for example, the trial. Let us consider a study that uses optogenetics to stimulate a certain brain area in one animal. The study compares RTs in a detection task between a condition with and one without optogenetic stimulation. For both conditions, 200 trials are obtained, and RTs are compared across trials. In this case, the inference is on all trials that could be obtained in the same way in this one animal. Along the same line, if RTs are averaged per session and statistics are performed across sessions, the inference is on all sessions that could be obtained in the same way in this one animal.

Crucially, the inference is always on a population of neurons (or trials or sessions), and for this (the type of elements in this population), it does not matter whether they are pooled over one, two, few, or even many animals. If neurons are obtained from one animal, the inference is on the population of neurons from that specific animal; and if data are pooled over, for example, three animals, the inference is on the population of neurons from those specific three animals. In this sense, the inference remains identical, irrespective of whether many animals, few animals, two animals, or even a single animal are used. Crucially, when pooling over multiple animals, the inference does not pertain to a population of animals.

For example, if 60 neurons were recorded from Animal 1 and 40 from Animal 2, a common approach would be to pool neurons over animals and to perform a single $t$ test across these 100 neurons. The population to which one infers on the basis of this single $t$ test is a population of neurons of which 60% is obtained from Animal 1, and 40% from Animal 2. As in the situation with a single animal, the inference is to a population of neurons, and not to a population of animals.

Of course, it is also possible to report separate tests for the individual animals, and this allows for a different type of analysis, which we will treat in the section Conjunction analysis as a formal replication approach.

We now consider inference to a population of animals, which requires that data are obtained from multiple animals. Specifically, one can then average the relevant quantities (like spike rate) within each animal (over neurons, trials, sessions) and perform a statistical test across animals. In this statistical test, the animals have the same role as the neurons, trials, or sessions in the statistical tests that we considered previously. In all cases (neurons, trials, sessions, and animals), the same mathematical logic applies. However, there is a crucial step when moving from a sample obtained within one or a few animals to a sample of animals. Testing using a sample of neurons, trials, or sessions sampled from one or a few animals merely allows an inference on these animals, and this is independent of whether sampling is done over few animals, two animals, or a single animal. Testing using a sample of animals allows to make an inference on the population of animals.

If the UOs are neurons, trials, or sessions, an effect is typically called a fixed effect, because the inference is bound to the fixed sample of animals from which the UOs were sampled. If the UOs are animals, an effect is typically called a random effect, because the inference pertains to some population of animals from which a sample was randomly drawn.

A Random Effects Test Requires Many Animals to Reach Significance

Given that most in vivo neurophysiological studies involve more than one animal, a random effects statistical test can, in principle, be performed across this small number of animals. Unfortunately, a random effects statistical test using the data of few animals typically has a very low sensitivity (statistical power, the probability of rejecting the null hypothesis when it is in fact false). To understand why this is the case, it is useful to investigate the factors that determine the outcome of a statistical test. For concreteness, we will now focus on the paired- or dependent-samples $t$ test, although our argument is equally valid for many other tests. The paired-samples $t$-statistic (denoted by $T$) depends on three factors: (1) the average between-conditions difference, or effect size, ($M$) in the sample; (2) the standard deviation ($s$) of the effect sizes across the sample of UOs (i.e., effect sizes calculated per UO); and (3) the number of UOs ($N$). Formally,

$$T = \frac{\sqrt{N} \times M}{s}$$

For a typical in vivo neurophysiology study, $N$ is fixed to a small number, such that $T$ depends solely on the ratio $M/s$, also called “Cohen’s $d$ in the sample.” Even if $M$ is large, $T$ will be insignificant if $s$ is too large. Unfortunately, when $N$ is small, only a very narrow range of raw data values results in $s$-values that are small enough for observing a significant $T$. Consider a study comparing two conditions in each of two animals (and testing a two-sided hypothesis, i.e., making no assumption about the sign of the effect). In Condition A, both animals show an average firing rate of 10 Hz. In Condition B, the first animal shows a firing rate of 20 Hz, that is, a 100% increase. For $T$ to be significant at the 0.05 level, the second animal needs to show an
increase between 85% and 115%. Larger deviations in either direction would increase s so much that T would become insignificant.

The minimum number of animals for a random effects test can be determined if 1) we assume that there is an effect with an effect size given as Cohen’s d in the population, and 2) we specify a desired sensitivity, typically 0.8. Then, the minimum number of animals follows from standard calculations (Cohen, 1992). For a sample Cohen’s d of one, that is, a large effect size, the t test would only reach significance for a sample of at least five animals. A similar reasoning applies if only the signs of the effects of the individual animals are relevant, in which case we can use the sign test. A sign test can only be significant (p < .05) with at least five animals for a one-sided, and with at least six animals for a two-sided, hypothesis.

**Conjunction Analysis as a Formal Replication Approach**

A random effects statistical test pertains to an average effect in some population. An alternative way to make an inference about a population starts from the intuition that multiple individually significant fixed effects tests provide more evidence than a single significant fixed effects test, corresponding to the current tradition of in vivo neurophysiology: A significant test in a first animal is often replicated in a second (or few) animals, and if the test is significant in each of them, this is taken as evidence for the effect being typical in the population of animals. This intuition has been turned into a formal statistical method, which is referred to as conjunction analysis (Friston, Holmes, & Worsley, 1999).

A conjunction analysis takes an individually significant fixed effects test as evidence for the presence of an effect in that participant. A conjunction analysis only makes sense if the direction of this effect (A > B or B > A) is the same for all participants and specified a priori. Given the number of participants that each show the effect, conjunction analysis estimates the expected proportion of the population showing the effect, which is called typicality γ. Specifically, the estimated typicality depends on the following parameters: 1) the number of participants N that each show the effect; 2) the false-positive rate α of the individual fixed effects tests; 3) the true-positive rate (sensitivity) β of the individual fixed effects tests.

For explaining how the typicality γ is estimated, we start from the situation in which all individual fixed effects tests in a sample of N animals are significant; the procedure is similar for other numbers of significant fixed effects tests. Now, the probability that all individual fixed effects tests are significant is the following:

\[ p(\text{all } N \text{ tests significant}) = [\alpha \times (1 - \gamma) + \beta \times \gamma]^N \]

Note that, in this equation, the number N is a random variable; if we would have observed less than N significant fixed effects tests (denoted by n), then we would have started from an expression for the probability p (more than n out of N tests significant). We cannot estimate γ directly, but we can estimate a useful lower bound to γ, called γc. This lower bound defines a one-sided confidence interval for γ, which contains its true value with a controllable probability. Here, we use the default 95% confidence interval, which corresponds to a statistical test about γ in the following way: The one-sided 95% confidence interval contains the collection of γ-values that cannot be rejected as the null hypothesis of a (random effects) statistical test at significance level 0.05. We can find this confidence interval by solving the following equation for γc:

\[ 0.05 = [\alpha + (\beta - \gamma_c)]^N \]

In this equation, α is the significance level (false-positive rate) of the fixed effects tests, which is typically set to .05. Thus, a small proportion of significant tests are false positives, slightly reducing the estimated typicality. β, the sensitivity (true-positive rate) of the fixed effects tests, cannot be determined, because it is the proportion of significant fixed effects 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 fixed effects test could reflect a false-negative test (and therefore essentially be discarded), allowing for any typicality to be consistent with the outcomes of the fixed effects tests. This can be concluded directly from the formula: If we chose β to be less than one, the resulting lower bound γc 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).

Using α = .05 and β = 1, and numerical methods to solve the formula, Figure 1 shows N, the number of participants with a significant fixed effects test, as a function of γc, the lower bound to typicality. This allows several important insights: 1.) With an individually significant effect in two participants, one can merely infer that the probability to find the effect in the population is at least 0.18; for three participants, this value increases to 0.34. 2.) We propose that 0.5 is the lowest useful value for typicality, because it corresponds to the expected presence of an effect in at least a simple majority of the population. For a typicality of 0.5, the number of required participants is five. 3.) For larger desired values of typicality, the number of required participants rises steeply: A typicality of 0.7 requires nine participants, and a typicality of 0.9 requires 30 participants.

It is important to note that the required number of participants hardly changes if we use a lower nominal false-positive rate α for the individual fixed effects tests (which is under the researcher’s control) and keep p (all N tests significant) at .05. The number strongly increases, if we lower both α and p (all N tests significant) by the same amount.
Conjunction analysis provides a formal evaluation of the inferences that can be based on replication across few animals. It reveals that a replication in a second or third animal does enable an inference on the population that goes beyond the inference of a single-animal fixed effects analysis. However, it also reveals that this additional inference is very limited. Scientists would have to agree that there is added value in knowing that there is a 0.18 or 0.34 probability of finding the effect in the population, because this is the inference that can actually be drawn from replication in one, respectively two, additional animals. This inference falls far short of the assumption that a second or third animal allows to make a general inference on the population. The responsible scientist will need to take this into account when deciding on the number of animals used in a study. Notably, the formal analysis shows that it is highly questionable when reviewers require an individually significant effect in a second or third animal, with the explicit or implicit argument that this is the international standard.

Conjunction analysis provides a formal evaluation of the inferences that can be based on replication across few animals. It reveals that a replication in a second or third animal does enable an inference on the population that goes beyond the inference of a single-animal fixed effects analysis. However, it also reveals that this additional inference is very limited. Scientists would have to agree that there is added value in knowing that there is a 0.18 or 0.34 probability of finding the effect in the population, because this is the inference that can actually be drawn from replication in one, respectively two, additional animals. This inference falls far short of the assumption that a second or third animal allows to make a general inference on the population. The responsible scientist will need to take this into account when deciding on the number of animals used in a study. Notably, the formal analysis shows that it is highly questionable when reviewers require an individually significant effect in a second or third animal, with the explicit or implicit argument that this is the international standard.

Replication for Other Reasons than Population Inference

Our considerations are concerned with using replications for making an inference on the population. For long, this has been the only motivation for replications, but more recently, scientists have argued for the value of replications in a way that does not require formal probability calculations (Kass et al., 2016):

“Every good analyst examines the data at great length, looking for patterns of many types and searching for predicted and unpredicted results. This process often involves dozens of procedures, including many alternative visualizations and a host of numerical slices through the data. Eventually, some particular features of the data are deemed interesting and important, and these are often the results reported in the resulting publication. When statistical inferences, such as p values, follow extensive looks at the data, they no longer have their usual interpretation. [...] The only truly reliable solution to the problem posed by data snooping is to record the statistical inference procedures that produced the key results, together with the features of the data to which they were applied, and then to replicate the same analysis using new data [emphasis added].”

In the terminology of this article, this use of replication could be denoted as a “solidification of the existence proof.” In this article, all inferential procedures, both fixed and random effects, are assumed to be solid. Note that this type of solidification can also be achieved within a single animal by splitting the within-animal replications (e.g., neurons, trials, sessions) in two or more partitions and replicating the same analysis on the different partitions. However, note that a t test is efficient, that is, most sensitive, only if it uses all available data, rather than data partitions; this follows from the Neyman-Pearson lemma.

Conclusions and Recommendations

Our analysis leads to the following conclusions:

- There are no formal arguments against a fixed effects statistical test on the combined data of all animals. However, crucially, scientists should be aware that such a test does not warrant an inference on the population of animals. The inference through a fixed effects test remains limited to the studied sample of animals, irrespective of whether the conclusions are based on a single animal, two animals, few animals, or even many animals.
- Few animals, in fact as few as two animals, are mathematically sufficient to perform a random effects test across those animals. However, this test would suffer from extremely low sensitivity, that is, an extremely high false-negative rate.
- A conjunction analysis provides a formal evaluation of the inferences that can be based on replication across few animals. It reveals that a replication in few animals does enable an inference on the population that goes beyond the inference of a single-animal fixed effects analysis. However, it also reveals that this additional inference is too limited to be of interest for the
scientific community. Therefore, the tradition to provide (or request) individually significant results in each of few animals is problematic. If not combined with an explicit mentioning of its severe limitations, that is, the obtained low typicality, it likely misleads many readers.

In summary, we make the following recommendations for in vivo neurophysiology:

- If there is not a preexisting data set:
  - Investigators should first obtain and analyze data from a single animal and publish them. Editors and reviewers, who so far routinely accepted two-animal articles (or few-animal articles), should routinely accept single-animal articles, because the inference is in both cases on the sample and not on the population of animals.
  - For most respective studies, this new standard will reduce the number of animals by at least 50%, which is a substantial contribution to animal welfare.
  - A minority of studies will require multiple animals to obtain sufficient data for acceptable sensitivity, even for an inference on the sample of animals, for example, studies that intracellulary record and label one neuron per animal.
  - After such a study with inference on the sample of animals, investigators can decide about a subsequent study, testing for an inference on the population of animals.
    - A random effects test allows an inference on the average effect in the population, that is, the most common goal. In most cases, this will require at least five animals.
    - A conjunction analysis allows the estimation of typicality. We propose that the minimal useful typicality is 0.5, and the corresponding required number of animals is five.
  - If there is a preexisting data set:
    - If there are data from more than five (or six, for a two-sided sign test) animals, investigators should consider a random effects test, for inferring the average effect, or a conjunction analysis, for inferring typicality.
    - If there are data from less than five animals, a fixed effects test on the pooled data from all animals should be performed and reported. Reviewers and editors should not require tests per animal, or even significance per animal.

In brief, provided one carefully qualifies the nature of one’s inference, fixed effects analyses of data from a single animal are perfectly valid to establish an existence proof for the effect in question. Providing this existence proof is orthogonal to the assertion that this feature or effect is conserved over a population. In many instances, random effects analyses (and related conjunction analyses) may not be necessary. The canonical example here is the “talking dog.” If you were to decide about publication, what $N$ would you require before accepting a report of talking dogs?

Declaration of interest: P.F. has a patent on thin-film electrodes and is beneficiary of a respective license contract with Blackrock Microsystems LLC (Salt Lake City, UT), member of the Advisory Board of CorTec GmbH (Freiburg, Germany), and managing director of Brain Science GmbH (Frankfurt am Main, Germany).

Reprint requests should be sent to Pascal Fries, Fries Lab, Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Deutschordenstr. 46, Frankfurt am Main 60528, Germany, or via email: pascal.fries@esi-frankfurt.de.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the Journal of Cognitive Neuroscience ( JoCN ) during this period were $M(an)/M = .407, W(oman)/M = .32, M/W = .115$, and $W/W = .159$, the comparable proportions for the articles that these authorship teams cited were $M/M = .549, W/M = .257, M/W = .109$, and $W/W = .085$ (Postle and Fulvio, JoCN, 34:1, pp. 1–3). Consequently, JoCN encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article’s gender citation balance.

REFERENCES

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159. https://doi.org/10.1037/0033-2909.112.1.155, PubMed: 19565683

Friston, K. J., Holmes, A. P., & Worsley, K. J. (1999). How many subjects constitute a study? Neuroimage, 10, 1–5. https://doi.org/10.1006/nimg.1999.0439, PubMed: 10385576

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., & Reid, N. (2016). Ten simple rules for effective statistical practice. PLoS Computational Biology, 12, e1004961. https://doi.org/10.1371/journal.pcbi.1004961, PubMed: 27281180

Maris, E. (2012). Statistical testing in electrophysiological studies. Psychophysiology, 49, 549–565. https://doi.org/10.1111/j.1469-8986.2011.01320.x, PubMed: 22176204