

Four simple ways to increase power without increasing the sample size

Stanley E Lazic

Quantitative Biology, Discovery Sciences, AstraZeneca, IMED Biotech Unit, Cambridge, CB4 0WG, UK

stan.lazic@cantab.net

Abstract

Underpowered experiments have three problems: the probability of a false positive result is higher, true effects are harder to detect, and the true effects that are detected tend to have inflated effect sizes. Many biology experiments are underpowered and recent calls to change the traditional 0.05 significance threshold to a more stringent value of 0.005 will further reduce the power of the average experiment. Increasing power by increasing the sample size is often the only option considered, but more samples increases costs, makes the experiment harder to conduct, and is contrary to the 3Rs principles for animal research. We show how the design of an experiment and some analytical decisions can have a surprisingly large effect on power.

Introduction

Statistical power is the probability of detecting a true effect, and therefore when experiments have low power, a true effect or association is hard to find.^{1,2} A less appreciated consequence of lower power is that a statistically significant result is more likely to be a false positive than a true effect.²⁻⁴ The probability that a significant result is a false positive increases because a p-value less than 0.05 means either (1) an effect is present, or (2) no effect is present but an extreme (unlikely) result occurred by chance. We never know which of these options is correct, but with low power, true effects are harder to detect, and so the "effect present" option becomes less probable and the "chance occurrence" option becomes more probable. A final problem with low power is that small effects will only be statistically significant when the effect size is overestimated (or if the within-group variance is underestimated).⁵⁻⁷ In other words, the correct qualitative conclusion might be reached (there is an effect) but the magnitude or strength of the effect is overestimated.⁶ Thus, lower power has a triple-negative effect on statistical inference.

Statisticians and scientists have recently been arguing for a more stringent significance threshold of 0.005 because the traditional 0.05 threshold provides little evidence against the null hypothesis.⁸⁻¹² If journals start requiring a lower threshold

for significance, the power of all experiments will be further reduced, exacerbating the above problems. By way of example, suppose we are conducting a two group experiment with independent samples in each group. To achieve 80% power to detect an effect size of 1.25 units, with a within-group standard deviation of 1 and a significance threshold of 0.05, we require 11 samples per group, or 22 in total. If the significance threshold is reduced to 0.005, 38 total samples are required—a 71% increase in sample size. If only 22 samples are used with the more stringent 0.005 significance threshold, the power of the experiment is only 44%.

To increase power, many researchers only consider increasing the sample size (N). Indeed, standard power and sample size calculations in textbooks and review articles suggest that the only option to increase power is to increase the number of samples. The design of the experiment is taken as given, the probability of making a false positive decision is set to $\alpha=0.05$, and the power is usually fixed to 80% or 90%. The effect size is the smallest effect that is biologically or clinically meaningful and that the researcher would like to detect, and the within group standard deviation is derived from pilot studies, published results, or an educated guess. (In practice, the effect size and standard deviation are often adjusted to give the sample size that investigators were planning to use all along^{13,14}). This only leaves N to be adjusted to meet the desired power. (We leave aside the assumed requirement that the sample size must be fixed in advance. Data can be collected until a strong conclusion is reached or until the allocated time and resources are expended. Controlling the false positive rate becomes a concern, as does the analysis of such an experiment, but both issues are handled naturally with Bayesian methods.^{15,16})

Increasing the sample size makes the experiment more expensive, harder to conduct, and has ethical implications for animal experiments. Also, it is often not possible to increase N while holding everything else constant. A larger experiment may need to be conducted in smaller batches, perhaps half of the samples are run on two separate days, or conducted by two researchers instead of one. This changes the design of the experiment because *Day* and *Researcher* are new variables that could influence the outcome and were not included in the power calculation. Similarly, data from a small experiment may be collected over a short period of time (e.g. one or two hours), making circadian effects negligible. A larger experiment may need to collect data over a longer period, and now circadian effects may become more pronounced. The design of the experiment now needs to change to accommodate the circadian effects; for example, by using time as a blocking factor.¹⁷ This is again a different experiment to the one used for the power calculation.

Simple options are available to increase power—often dramatically—while keeping the sample size fixed. Or coming from the other direction, certain design and

analytic options can be avoided to prevent loss of power. Four options are described below that apply to most biological experiments, and other options are described in Lazic.¹⁷ The R code for these examples is provided as supplementary material.

Trade-offs are inevitably required when planning an experiment, and defining a key question or hypothesis enables the experiment to be designed to maximise the chance of success, but at the cost of being unable to address other questions. Hence, the first step is to clearly define the key question that the experiment will answer. This may sound trite, but consider an experiment testing a new compound in an animal model of a disease. The key questions could be:

1. Is the compound active (does it work)?
 2. Is the compound active at a specific dose?
 3. What is the minimum effective dose?
 4. What is the dose that gives half of the maximum response (ED50)?
 5. Is one dose better than another?
- No design is optimal for answering all of these questions; some designs are better suited to answer some questions and other designs are better for other questions. Once the question or hypothesis is defined, the four points below can be used to plan the experiment.

1. Use fewer factor levels for continuous predictors

- In experiments with continuous predictors such as dose, concentration, time, pH, temperature, pressure, illumination level, and so on, how do we decide on the minimum and maximum levels, the number of points in between, and the specific levels where observations will be made or samples allocated? Some choices are easy, for example, when a minimum value of zero serves as a control condition, and the maximum value is the upper limit that we are interested in testing or learning about. But what about the points between the min and max? The places where observations are made (min, max, plus intermediate points) are called the *design points*.¹⁸

- To illustrate how the number of design points affects the power, we compare four designs with 2 to 5 design points (experimental groups) and a fixed sample size of 20. Assume that the dose of a compound is the factor of interest, which ranges from min = 0 to max = 100. 10,000 data sets were simulated from the true model (Fig. 1A), which has a maximum response of 40 at Dose = 0, and a minimum response of 27 at Dose = 100. The variability of the data is shown by the scatter of points around the dose-response line (standard deviation = 9). Data were simulated under each design and analysed with a one-way ANOVA, testing the general hypothesis "is there any difference between the experimental groups"? The power for each design is then calculated as the proportion of significant results from the ANOVA (overall F-test).

Despite the same sample size, the power of these experiments differs greatly (Fig. 1B, "ANOVA line"). The power of Design 1 with two groups is 84% and steadily decreases to 40% with Design 4.

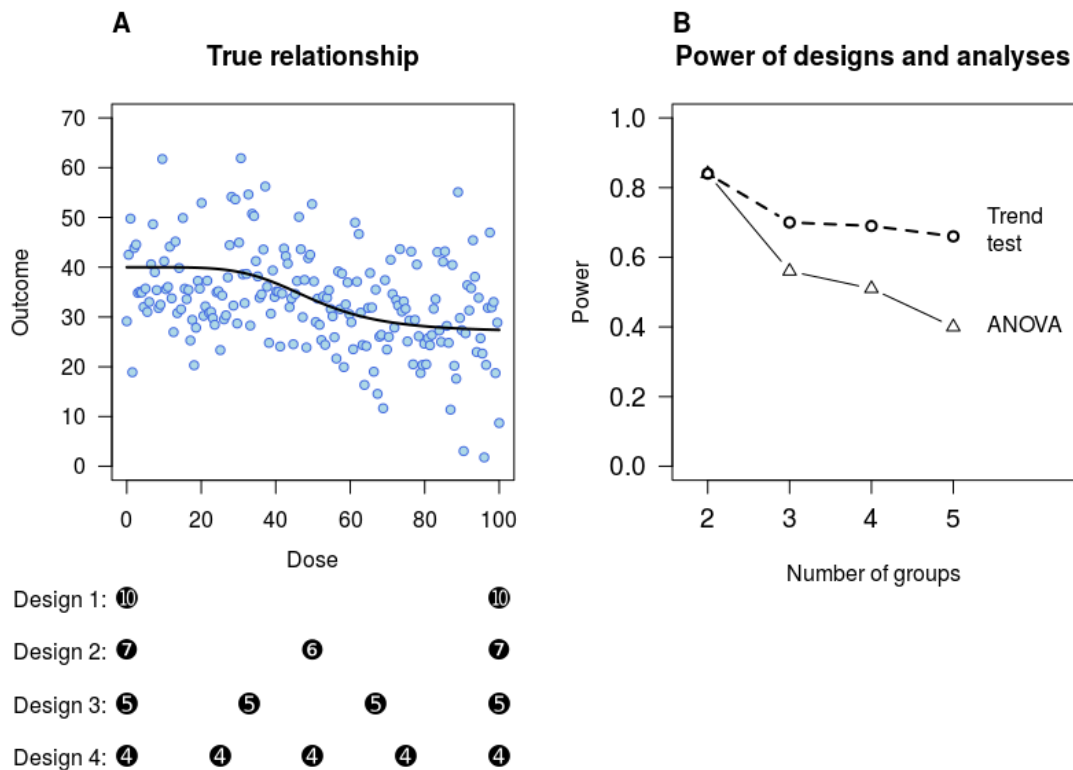


Figure 1. Effect of increasing the number of groups. A fictitious example of the relationship between an outcome and the dose of a compound (A). Four designs with the same sample size but with observations taken at different doses are indicated with black circles. For example, Design 1 has 10 samples at doses 0 and 100. Increasing the number of groups from 2 to 5 decreases the power for the overall ANOVA from 84% to 40% (B). Testing a more specific hypothesis for a trend has improved power compared with an ANOVA analysis, but still loses power with more groups.

It is clear that using two groups maximises power. What if the true relationship is not sigmoidal but a "U", inverted-"U", or some other complex shape? A design with two groups would be unable to detect such relationships, but if a linear or monotonic relationship is expected, then one additional design point can allow departures from the assumed relationship to be detected (e.g. Design 2 in Figure 1).¹⁸ Trade-offs are always necessary when designing experiments, although adding an additional group allows a more complex relationship to be detected, it lowers the probability of detecting a linear relationship if that is indeed the correct one. If the aim of the study is to fit a 4 parameter logistic model (the black line in Fig. 1A), then Design 4 is better, illustrating how the aim of the experiment or the question asked influences

the design. Berger and Wong discuss a wider range of designs and how they affect power for different types of relationships.¹⁸

A related point to maximise power is to ensure that the predictor variable covers a wide enough range; for example, the compound would appear to have no effect if the maximum dose is 30.

2. Use a focused hypothesis test

A second way to increase power is to test a specific hypothesis instead of a general "are there any differences?" hypothesis. For the simulated example above, the ANOVA analysis can detect any pattern of differences between groups, but the trade-off is that it has less power to detect any specific pattern. If we expect the outcome to either steadily increase or decrease as the dose increases, then a focused test of this pattern is more powerful (Fig. 1B, "Trend test" line).

With two groups the power of the trend test is identical to the ANOVA analysis (both are equivalent to a t-test), but with five groups (Design 4) the power of the trend test is 66%, compared with 40% for the ANOVA analysis. Testing for a trend is available in most statistical packages and usually requires that the predictor variable—dose in this case—is treated as an *ordered* categorical factor. The output from such an analysis will usually include the test for a linear trend (linear contrast). Alternatively, treating dose as a continuous variable and analysing the data with a regression analysis instead of an ANOVA is another option that has high power.¹⁹

3. Don't dichotomise or bin continuous variables

Dichotomising or binning refers to taking a continuous variable and reducing it to two groups (e.g. Low/High) based on a threshold such as the median. Sometimes the continuous variable is reduced to more than two groups such as Low/Medium/High, and both outcome and predictor variables can be dichotomised. This practice is common, despite many papers warning against it. Dichotomising variables reduces power, can bias estimates, and can increase false positives.¹⁹⁻³³

To illustrate this point, Figure 2 shows data for 50 samples that have a correlation of 0.4. These are naturally analysed with a Pearson correlation or linear regression (solid line), both of which give $p = 0.002$ for the association. If the variables are dichotomised at their median (dashed lines) the number of data points falling in each quadrant can be counted (numbers in grey boxes), forming a 2×2 table. This doubled-dichotomised data is commonly analysed with a χ^2 test or Fisher's Exact test, and the χ^2 test gives $p = 0.024$ —almost ten times larger than the regression analysis. To estimate the reduction power when using the χ^2 test, 5000 data sets like the one in Figure 2 were simulated with $N = 50$ and a correlation of 0.4. Analysing the continuous values with a regression analysis had 84% power, while binning and using the χ^2 test reduced power to 34%.

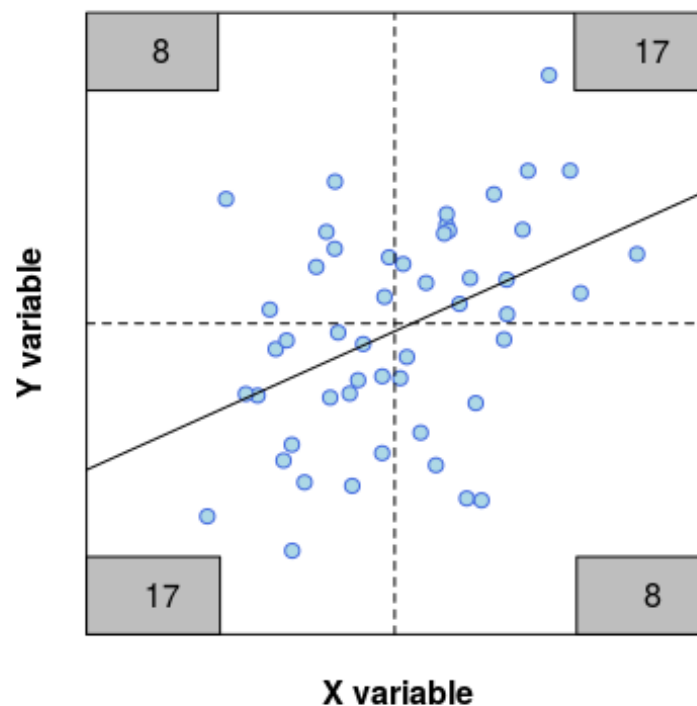


Figure 2. Analysis of continuous and binned data. A regression analysis (black line) gives $p = 0.002$. Binning the data with a bivariate median-split (dashed lines) and analysing the number of samples in each quadrant (grey boxes) with a χ^2 test gives $p = 0.024$. The power of the χ^2 test is only 34%, compared with 84% for the regression analysis.

4. Cross your factors, don't nest them

The levels of an experimental factor are either (1) set by the researcher, (2) a property of the samples, or (3) a technical aspect of how the experiment is conducted. For example, the dose of a drug that an animal receives is controlled by the researcher, while the sex of the animal is a property of the animal. If the experiment is conducted over multiple days or by more than one researcher, then Day and Researcher are technical factors. *Factor arrangement* refers to how multiple factors relate to each other, and there are three possibilities.

When two factors are *completely confounded*, levels of one factor always co-occur with the same levels of the other factor; for example, if all the control animals are run on the first day and all the treated animals are run on the second day. Confounding a treatment effect that we are interested in testing with an uninteresting technical effect is never a good idea because it is impossible to attribute any differences between treatment groups to the effect of the treatment—differences may have arisen from the day-to-day variation. To conclude anything

about a treatment effect we would have to assume that the Day effect is zero, and therefore this arrangement should be avoided.

The second possibility is a *crossed* or *factorial* arrangement, which occurs when all levels of one factor co-occur with all levels of another factor, and is the most common arrangement in experimental biology. The final possibility is a *nested* arrangement, where levels of one factor are grouped or nested under the levels of another factor. Figure 3 shows the difference between crossed and nested arrangements. Suppose that we have 16 mice (assume all male for simplicity) from 4 litters (A-D), and we want to test the effect of a compound at a single dose. Assume that each mouse can be randomly and individually assigned to one of the two treatment groups. Although this is a simple two-group design, we might expect differences between litters and want to take this into account in the design.

When all animals from a litter are in the same condition, the factor Litter is said to be nested under the factor Treatment (Fig. 3, left). When animals from a litter are spread across both treatment groups, the Treatment and Litter factors are crossed (Fig. 3, right).

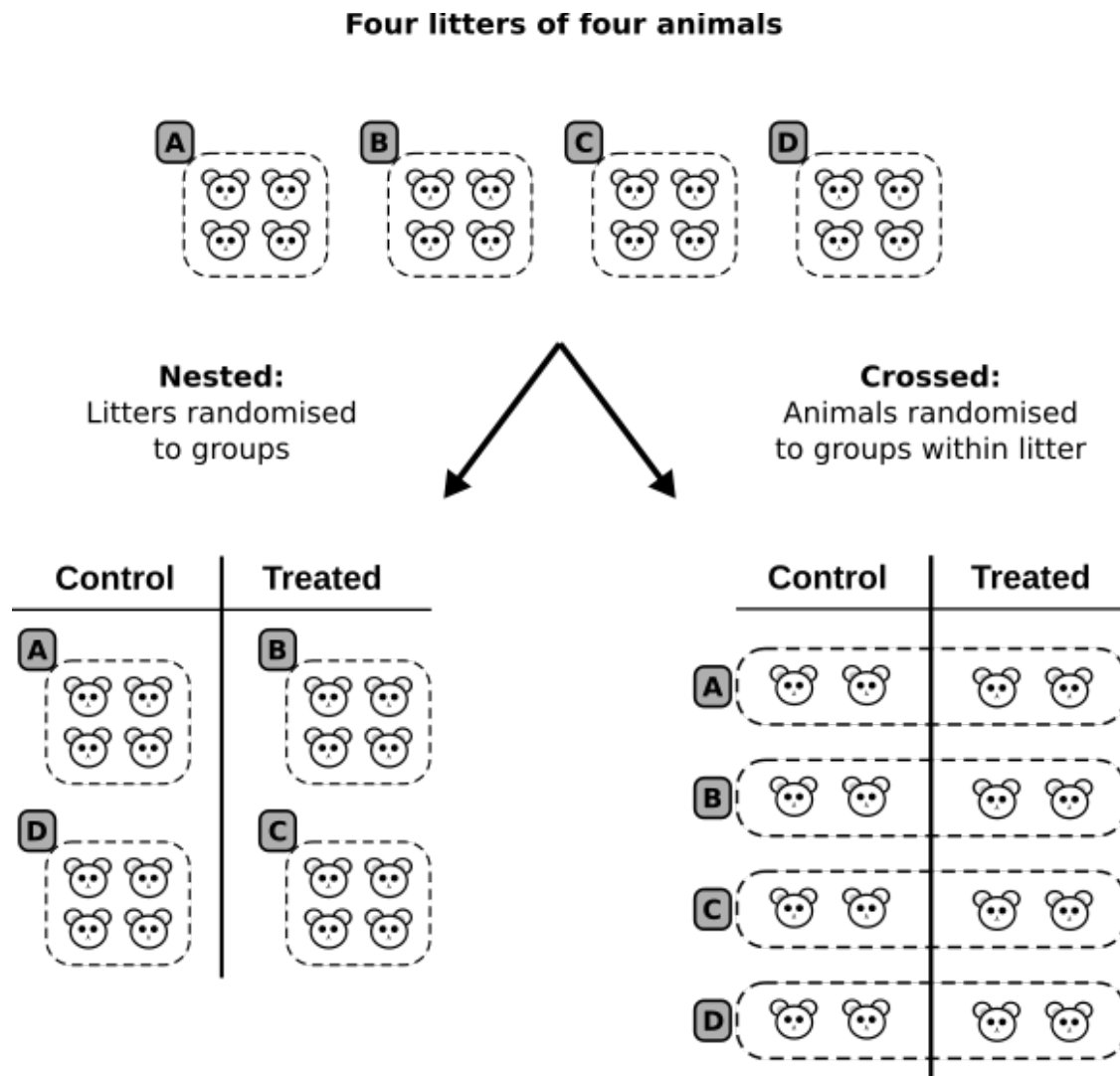


Figure 3. Crossed versus nested designs. An experiment with 16 mice from four litters (A-D) can be randomised to treatment groups either by litter, leading to a nested design (left), or within litter, leading to a crossed design (right).

The main point here is that the nested arrangement is much less powerful than the crossed one. Example data is shown in Figure 4 to illustrate the difference in power between the two designs. The litter means are drawn from a normal distribution with a standard deviation of 3, and the values for the individual animals are drawn from a normal distribution with a standard deviation of 0.5. Thus, the litter-to-litter variation is large relative to the variation between animals within a litter. Figure 4A shows the data before the application of a treatment, and an immediate danger can be seen with a nested design: if litters A and B end up in one group and C and D in the other, large differences between groups exist at baseline. Hence, nested designs can also lead to more false positive findings.^{34,35}

Figure 4B shows one possible randomisation of a nested design, where litters A and D end up in the control group and B and C in the treated group. The effect of the treatment in this example is to decrease the outcome by 2 units (note how the control group values of black O's and green x's are identical in Fig. 4A and B).

- Analysing the data with a t-test gives a p-value of 0.14. However, since the litters were randomised to the treatment conditions and not the individual animals, the litter is a more appropriate experimental unit, that is, the sample size is 4 litters, not 16 animals.^{17,36} One way to conduct this analysis is to calculate the litter average and use these values for a standard statistical test, which gives $p = 0.57$. Figure 4C shows one possible randomisation of the crossed design, with the same effect size of -2 units. Analysis of this design, which includes litter as a variable in the model gives $p = 0.0007$ for the effect of the treatment.

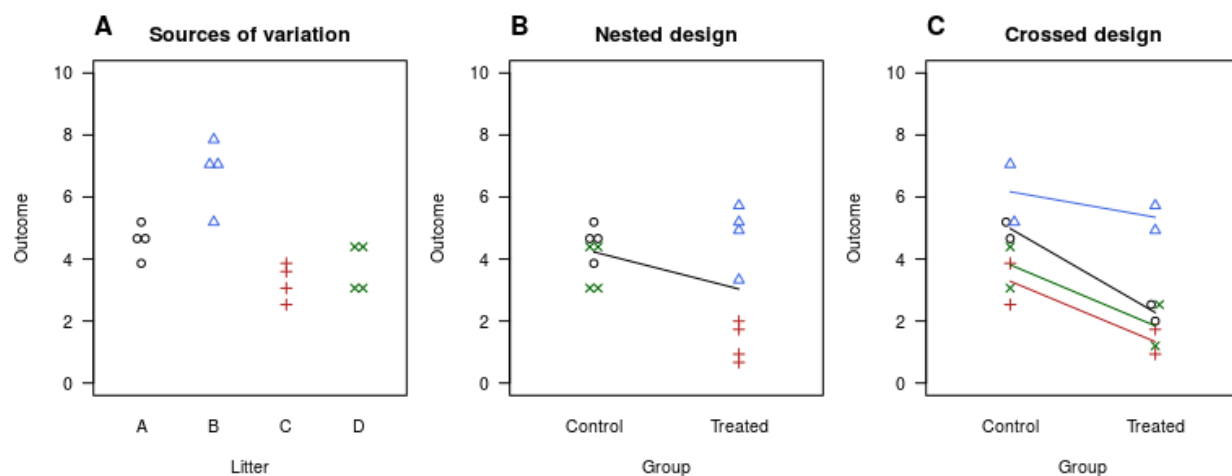


Figure 4. Simulated data for a nested and crossed design. Natural variation between and within litters (A). The nested design compares the effect of the treatment against the high variation between litters, leading to a large p-value (0.14 or 0.57, depending on the analysis; B). The crossed design compares the treatment effect to the smaller within-litter variation, leading to a much smaller p-value of 0.0007 (C).

- To calculate the power of the two designs and both analyses of the nested design, 5000 data sets were generated with the above characteristics. The nested design treating the animals as the experimental unit has 55% power, and treating the litter as the experimental unit (the more appropriate analysis) has only 7% power. The power of the crossed design is over 99%. These large differences in power exist because the nested design compares the effect size of -2 units against the high litter-to-litter variation (standard deviation = 3), whereas the crossed design compares the effect size against the much smaller variation of animals within a litter (standard deviation = 0.5). With the crossed design the test for a treatment effect is performed within the litters and hence large differences between litters are irrelevant for

testing treatment effects. Litter is used as a "blocking variable" to remove the unwanted litter-to-litter variation.^{34,37}

The difference in power between nested and crossed designs become less pronounced as litter effects get smaller, animal variation within a litter gets larger, or both. The nested design should be avoided because of low power, more false positives, and ambiguity in defining the sample size. Unfortunately, nested designs are common because they are often easier to conduct and manage; for example, it's easier to randomise litters to treatment groups as littermates can be housed together in the same cage.

Other technical variables or properties of subjects such as batches, microtitre plates, cages or other housing structure, body weight, day, and experimenter, can also be nested or crossed with the treatment effects of interest. These factors need to be carefully arranged to ensure high power and no confounding. The points discussed above can combine to reduce power even further; for example, using a nested design, dichotomising a variable, and testing a general hypothesis will have a dramatic loss of power compared with a crossed design, without dichotomisation, and testing a focused hypothesis.

Conclusion

Low power continues to undermine many biology experiments, but a few simple alterations to a design or analysis can dramatically increase the information obtained without increasing the sample size. In the interest of minimising animal usage and reducing waste in biomedical research,^{38,39} researchers should aim to maximise power by designing confirmatory experiments around key questions, use focused hypothesis tests, and avoid dichotomising and nesting that ultimately reduce power and provide no other benefits.

References

1. Maxwell, S. E. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychol Methods* **9**, 147-163 (2004).
2. Button, K. S. *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365-376 (2013).
3. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med* **2**, e124 (2005).
4. Nuzzo, R. Scientific method: Statistical errors. *Nature* **506**, 150-152 (2014).
5. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640-648 (2008).

6. Gelman, A. & Carlin, J. Beyond power calculations: assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science* **9**, 641-651 (2014).
7. McShane, B. B. & Bockenholt, U. You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science* **9**, 612-625 (2014).
8. Berger, J. O. & Sellke, T. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* **82**, 112-122 (1987).
9. Sellke, T., Bayarri, M. J. & Berger, J. O. Calibration of p values for testing precise null hypotheses. *The American Statistician* **55**, 62-71 (2001).
10. Johnson, V. E. Revised standards for statistical evidence. *PNAS* **110**, 19313-19317 (2013).
11. Singh Chawla, D. Big names in statistics want to shake up much-maligned P value. *Nature* **548**, 16-17 (2017).
12. Benjamin, D. J. *et al.* Redefine statistical significance. *Nature Human Behaviour* (2017). doi:[doi:10.1038/s41562-017-0189-z](https://doi.org/10.1038/s41562-017-0189-z)
13. Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* **365**, 1348-1353 (2005).
14. Bacchetti, P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med* **8**, 17 (2010).
15. Schonbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M. & Perugini, M. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychol Methods* **22**, 322-339 (2017).
16. Schonbrodt, F. D. & Wagenmakers, E. J. Bayes factor design analysis: Planning for compelling evidence. *Psychon Bull Rev* (2017).
17. Lazic, S. E. *Experimental design for laboratory biologists: Maximising information and improving reproducibility*. (Cambridge University Press, 2016).
18. Berger, M. P. F. & Wong, W. K. *An introduction to optimal designs for social and biomedical research*. (Wiley, 2009).
19. Lazic, S. E. Why we should use simpler models if the data allow this: Relevance for ANOVA designs in experimental biology. *BMC Physiol* **8**, 16 (2008).

20. Barnwell-Menard, J.-L., Li, Q. & Cohen, A. A. Effects of categorization method, regression type, and variable distribution on the inflation of Type-I error rate when categorizing a confounding variable. *Stat Med* **34**, 936-949 (2015).
21. Bennette, C. & Vickers, A. Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol* **12**, 21 (2012).
22. Chen, H., Cohen, P. & Chen, S. Biased odds ratios from dichotomization of age. *Stat Med* **26**, 3487-3497 (2007).
23. Cohen, J. The cost of dichotomization. *Applied Psychological Measurement* **7**, 249-253 (1983).
24. Fedorov, V., Mannino, F. & Zhang, R. Consequences of dichotomization. *Pharm Stat* **8**, 50-61 (2009).
25. Irwin, J. & McClelland, G. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research* **40**, 366-371 (2003).
26. Kenny, P. W. & Montanari, C. A. Inflation of correlation in the pursuit of drug-likeness. *J Comput Aided Mol Des* **27**, 1-13 (2013).
27. Maxwell, S. & Delaney, H. Bivariate median splits and spurious statistical significance. *Quantitative Methods in Psychology* **113**, 181-190 (1993).
28. Owen, S. V. & Froman, R. D. Why carve up your continuous data? *Res Nurs Health* **28**, 496-503 (2005).
29. Royston, P., Altman, D. G. & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* **25**, 127-141 (2006).
30. Taylor, J. & Yu, M. Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis* **83**, 248-263 (2002).
31. Walraven, C. van & Hart, R. G. Leave 'em alone - why continuous variables should be analyzed as such. *Neuroepidemiology* **30**, 138-139 (2008).
32. Wainer, H., Gessaroli, M. & Verdi, M. Finding what is not there through the unfortunate binning of results: The mendel effect. *CHANCE* **19**, 49-52 (2006).
33. Kuss, O. The danger of dichotomizing continuous variables: A visualization. *Teaching Statistics* **35**, 78-79 (2013).
34. Lazic, S. E. & Essioux, L. Improving basic and translational science by accounting for litter-to-litter variation in animal models. *BMC Neurosci* **14**, 37 (2013).

35. Lazic, S. E. Comment on 'stress in puberty unmasks latent neuropathological consequences of prenatal immune activation in mice'. *Science* **340**, 811; discussion 811 (2013).
36. Lazic, S. E., Clarke-Williams, C. J. & Munafo, M. R. What exactly is 'n' in cell culture and animal experiments? *bioRxiv* (2017). doi:[10.1101/183962](https://doi.org/10.1101/183962)
37. Festing, M. F. W. Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *ILAR J* **55**, 472-476 (2014).
38. Macleod, M. R. *et al.* Biomedical research: Increasing value, reducing waste. *Lancet* **383**, 101-104 (2014).
39. Ioannidis, J. P. A. *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166-175 (2014).