The Four Horsemen of neglecting experimental design

Elise JACQUEMET [1], Emeline PERTHAME [1], Stevenn VOLANT [1], Thomas OBADIA [1], Hugo VARET [1], François LAURENT [1], Pascal CAMPAGNE [1]

[1] Institut Pasteur, Université de Paris, Hub de Bioinformatique et Biostatistique, F-75015 Paris, France

Introduction

Reproducibility of experimental results has become a central concern in biological sciences. Increasing attention is brought on how to carefully design experiments and choose appropriate statistical tools. Controlled experiments may be hard to conduct as they are subjected to many technical constraints, especially in animal experiments. Regular issues are such that (i) experimental outcomes may not be properly designed, statistically speaking; (ii) basic statistical tools may not be suited to properly analyse data that are partly shaped by technical constraints. Both issues are not rare in studies made at Pasteur and their impact on results are seldom considered. A downside of neglecting experimental design is an uncontrolled inflation of false-positives that may occur well above the typical 5% control threshold. Furthermore, a lack of power insidiously leads to overestimating the magnitude of experimental effects, all the more when they are coupled with confounding factors, due to ill-defined designs. Here we present four problems we often encounter in our daily life of data analysts. We illustrate why & how they may drastically alter the validity of some scientific findings.

Carrying out underpowered studies

Intrepid biologist speaking: *I work on a contagious disease that require mice to be kept in an isolator. This will be a very expensive experiment. I have no other choice than working with reduced sample sizes. This should be fine since I expect large effects...*

One-dimensional

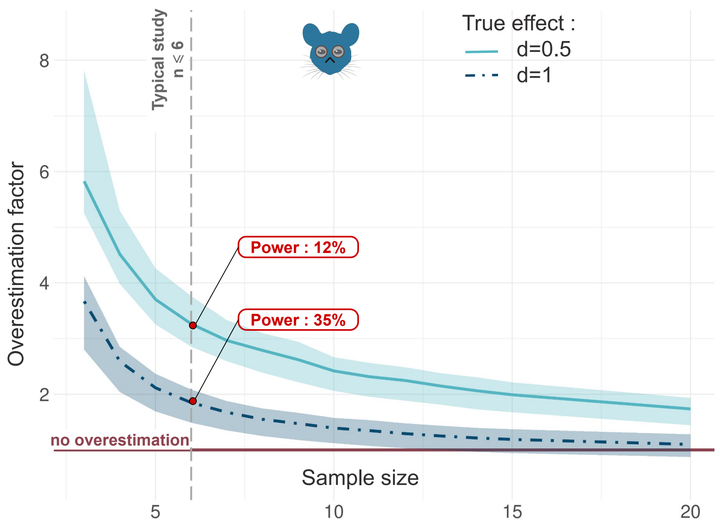


Fig 1 - Overestimation of experimental effects in reduced sample size due to a lack of power. Data were simulated to compare mean values obtained in two experimental groups with a t-test (using fixed values of *d*, i.e. the standardised difference of means among groups, and various sample sizes). Standardised difference among groups was recorded only when the *p*-value < 0.05 (10,000 simulated comparisons for each parameter combination). The curves and envelopes represent the average and interquartile ranges of the *Overestimation factor* (i.e., the ratio between estimated and true effect).

Multi-dimensional

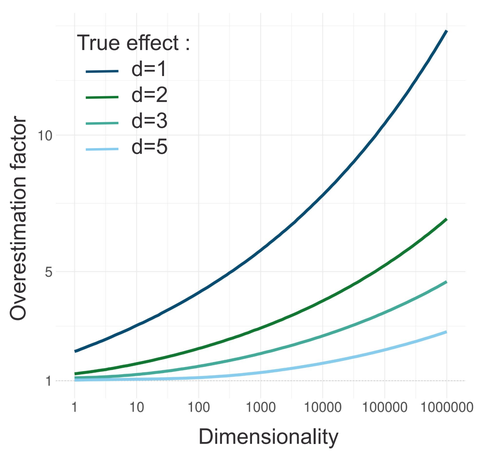
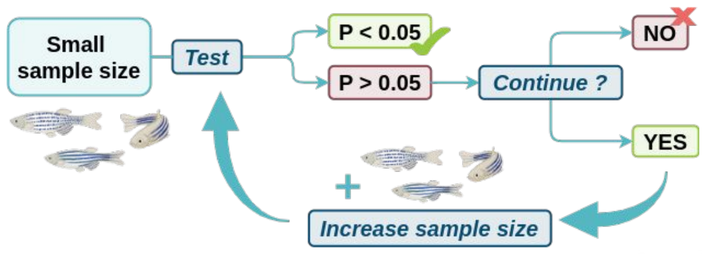


Fig 2 - Expected rise of overestimated experimental effects when analysing each variable separately in highly dimensional data. While the number of simulated variables (dimensionality) is varied, the experimental set-up is similar to that of Fig. 1 (2 groups, t-test, $n = 5$). Here, the rise of overestimated effects results from procedures of *p*-value adjustment to account for multiple testing.

Sequentially recruiting samples

Intrepid biologist speaking: *I want to investigate the experimental effect of two treatments in zebrafish, but I have no guess about the strength of this effect. To minimize the number of fish used in my study, I will make a first experiment with a small sample size and I will repeat this experiment until I find a significant difference.*

Flawed experimental process



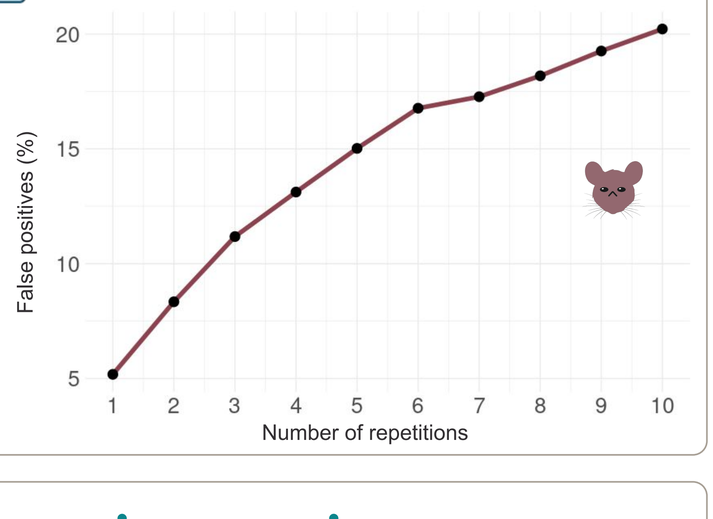


Fig 3 - Inflation of false-positives when repeating an experiment, due to sequential recruiting. Data were simulated to compare mean values obtained in two experimental groups with a t-test, under H0. A repetition consists in making the same experiment again with a constant sample size (adding 3 individuals per group). The figure reads as follow: e.g. after recruiting samples with 5 sequential repetitions, the rate of false positives reaches 15% instead of remaining at 5% as mistakenly assumed.

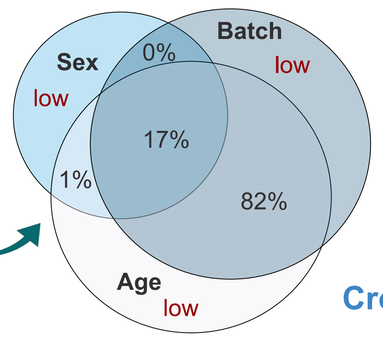
Confounding experimental effects

Intrepid biologist speaking: *For convenience reasons I will prepare libraries and proceed to sequencing samples as they come, by experimental group. I should be able to catch the effects of both Age and Sex anyhow.*

Table 1 - Example of experimental design with confounded variables: Age, Sex and Sequencing batch (inspired from real data - labels and variables were changed for the sake of anonymity).

Individuals	Age group	Sex	Sequencing batch
1	2	Male	A
2	2	Male	A
3	2	Male	A
4	3	Female	A
5	3	Female	A
6	3	Female	A
7	1	Male	B
8	1	Male	B
9	1	Male	B

Flawed design



Crossed design

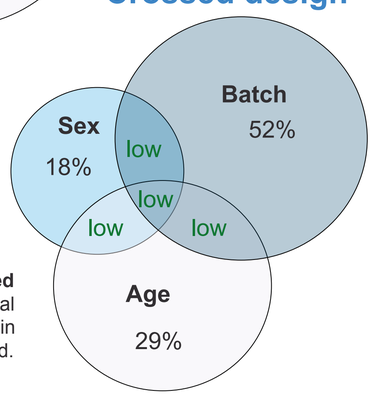


Fig 4 - Partition of variance. Venn diagrams representing the amount of variation explained by 3 variables: age, sex and sequencing batch. Data were simulated by attributing identical additive effects of variables across two experimental designs: (i) a flawed design, as presented in Table 1 and (ii) a proper blocked design where all experimental and technical factors are crossed. Partition of variance was obtained by using linear models (such as ANOVA).

Overlooking nestedness in experiments

Intrepid biologist speaking: *I work on an infectious disease using a mouse model. I made an experiment with 2 groups of 18 mice. To avoid contamination among individuals, I can't mix the two groups in the same cages and I will proceed with batches (i.e., in each group: 3 cages of 6 mice). Then I will analyse data with usual statistical tests (e.g., Student's test, etc.).*

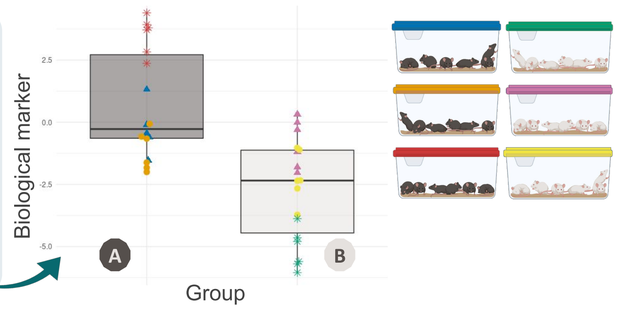


Fig 5 - Simulated distribution of data structured by cage effects (different dot colours represent different cages).

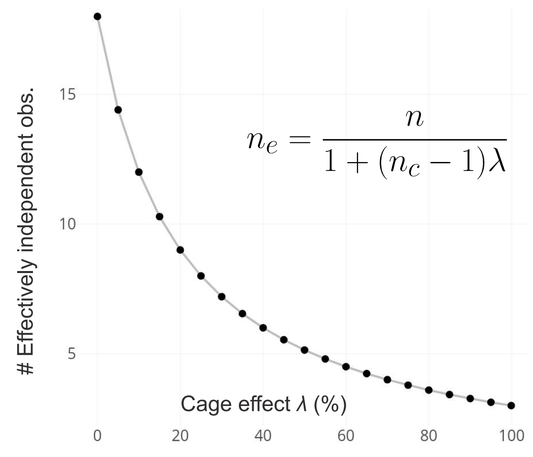


Fig 6 - Hierarchical (cage) effects λ decrease the effective number of independent observations (*ne*). The cage effect λ is the fraction of variance bound to consistent differences among experimental units (cages). Consequence is an overestimated test-statistic value, leading to an uncontrolled inflation of false positives (see Fig. 7). Parameters: n=18, n_c=6 animals per cage.

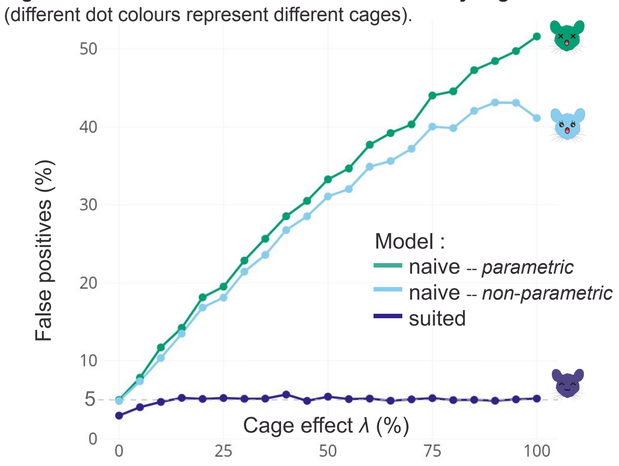
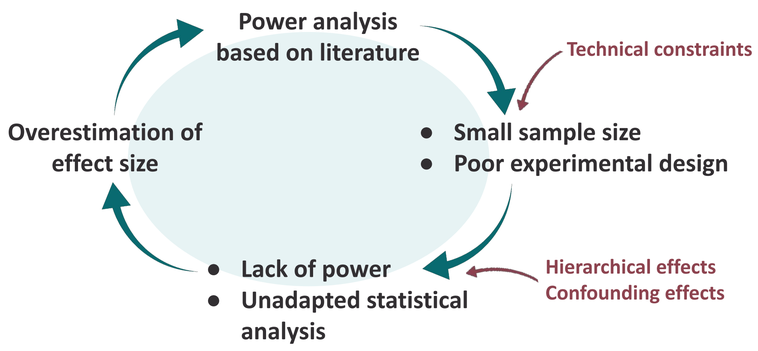


Fig 7 - Inflation of false-positives due to neglecting hierarchical effects in the analysis. In the simulations, no differences exist among groups (H0): 3 models were fitted to compare two experimental groups (10,000 simulations per dot). *False positives*: percent comparisons misinterpreting the cage effect as a group effect.

A non-reproducibility loop

Such issues regarding experimental design and consistent data analysis are not isolated aspects. This may worsen the problem, e.g., sequentially recruiting samples while not taking hierarchical effects into account. Unfortunately it is not uncommon. For illustration sake, we sketched the kind of loop we are aiming to break, guess why (!)



Support & tools provided by the Hub