POINTS OF VIEW

Designing comparative experiments

Good experimental designs limit the impact of variability and reduce sample-size requirements.

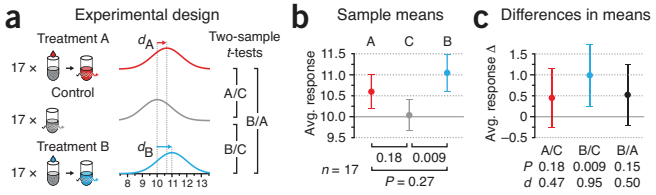

Figure 1 | Design and reporting of a single-factor experiment with three levels using a two-sample *t*-test. (**a**) Two treated samples (A and B) with $n = 17$ are compared to a control (C) with $n = 17$ and to each other using two-sample *t*-tests. (**b**) Simulated means and *P* values for samples in **a**. Values are drawn from normal populations with $\sigma = 1$ and mean response of 10 (C), 10.6 (A) and 11 (B). (**c**) The preferred reporting method of results shown in **b**, illustrating difference in means with CIs, *P* values and effect size, *d*. All error bars show 95% CI.

In a typical experiment, the effect of different conditions on a biological system is compared. Experimental design is used to identify data-collection schemes that achieve sensitivity and specificity requirements despite biological and technical variability, while keeping time and resource costs low. In the next series of columns we will use statistical concepts introduced so far and discuss design, analysis and reporting in common experimental scenarios.

In experimental design, the researcher-controlled independent variables whose effects are being studied (e.g., growth medium, drug and exposure to light) are called factors. A level is a subdivision of the factor and measures the type (if categorical) or amount (if continuous) of the factor. The goal of the design is to determine the effect and interplay of the factors on the response variable (e.g., cell size). An experiment that considers all combinations of *N* factors, each with n_i levels, is a factorial design of type $n_1 \times n_2 \times \ldots \times n_N$. For example, a 3 × 4 design has two factors with three and four levels each and examines all 12 combinations of factor levels. We will review statistical methods in the context of a simple experiment to introduce concepts that apply to more complex designs.

Suppose that we wish to measure the cellular response to two different treatments, A and B, measured by fluorescence of an aliquot of cells. This is a single factor (treatment) design with three levels (untreated, A and B). We will assume that the fluorescence (in arbitrary units) of an aliquot of untreated cells has a normal distribution with $\mu = 10$ and that real effect sizes of treatments A and B are $d_A = 0.6$ and $d_B = 1$ (A increases response by 6% to 10.6 and B by 10% to 11). To simulate variability owing to biological variation and measurement uncertainty (e.g., in the number of cells in an aliquot), we will use $\sigma = 1$ for the distributions. For all tests and calculations we use $\alpha = 0.05$.

We start by assigning samples of cell aliquots to each level (**Fig. 1a**). To improve the precision (and power) in measuring the mean of the response, more than one aliquot is needed[1]. One sample will be a control (considered a level) to establish the baseline response, and capture biological and technical variability. The other two samples will be used to measure response to each treatment. Before we can carry out the experiment, we need to decide on the sample size.

We can fall back to our discussion about power[1] to suggest *n*. How large an effect size (*d*) do we wish to detect and at what sensitivity? Arbitrarily small effects can be detected with large enough sample size, but this makes for a very expensive experiment. We will need to balance our decision based on what we consider to be a biologically meaningful response and the resources at our disposal. If we are satisfied with an 80% chance (the lowest power we should accept) of detecting a 10% change in response, which corresponds to the real effect of treatment B ($d_B = 1$), the two-sample *t*-test requires $n = 17$. At this *n* value, the power to detect $d_A = 0.6$ is 40%. Power calculations are easily computed with software; typically inputs are the difference in means ($\Delta\mu$), standard deviation estimate (σ), α and the number of tails (we recommend always using two-tailed calculations).

Based on the design in **Figure 1a**, we show the simulated samples means and their 95% confidence interval (CI) in **Figure 1b**. The 95% CI captures the mean of the population 95% of the time; we recommend using it to report precision. Our results show a significant difference between B and control (referred to as B/C, $P = 0.009$) but not for A/C ($P = 0.18$). Paradoxically, testing B/A does not return a significant outcome ($P = 0.15$). Whenever we perform more than one test we should adjust the *P* values[2]. As we only have three tests, the adjusted B/C *P* value is still significant, $P' = 3P = 0.028$. Although commonly used, the format used in **Figure 1b** is inappropriate for reporting our results: sample means, their uncertainty and *P* values alone do not present the full picture.

A more complete presentation of the results (**Fig. 1c**) combines the magnitude with uncertainty (as CI) in the difference in means. The effect size, *d*, defined as the difference in means in units of pooled standard deviation, expresses this combination of measurement and precision in a single value. Data in **Figure 1c** also explain better that the difference between a significant result (B/C, $P = 0.009$) and a nonsignificant result (A/C, $P = 0.18$) is not always significant (B/A, $P = 0.15$)[3]. Significance itself is a hard boundary at $P = \alpha$, and two arbitrarily close results may straddle it. Thus, neither significance itself nor differences in significance status should ever be used to conclude anything about the magnitude of the underlying differences, which may be very small and not biologically relevant.

CIs explicitly show how close we are to making a positive inference and help assess the benefit of collecting more data. For example, the CIs of A/C and B/C closely overlap, which suggests that at our sample size we cannot reliably distinguish between the response to A and B (**Fig. 1c**). Furthermore, given that the CI of A/C just barely crosses zero, it is possible that A has a real effect that our test failed to detect. More information about our ability to detect an effect can be obtained from a *post hoc* power analysis, which assumes that the observed effect is the same as the real effect (normally unknown), and uses the observed difference in means and pooled variance. For A/C, the difference in means is 0.48 and the pooled s.d. (s_p) = 1.03, which yields a *post hoc* power of 27%; we have little power to detect this difference. Other than increasing sample size, how could we improve our chances of detecting the effect of A?

Our ability to detect the effect of A is limited by variability in the difference between A and C, which has two random components. If

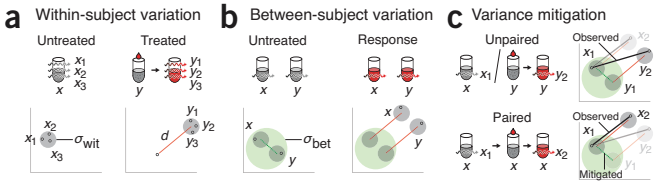

Figure 2 | Sources of variability, conceptualized as circles with measurements (x_i, y_i) from different aliquots (x,y) randomly sampled within them. (**a**) Limits of measurement and technical precision contribute to σ_{wit} (gray circle) observed when the same aliquot is measured more than once. This variability is assumed to be the same in the untreated and treated condition, with effect d on aliquot x and y. (**b**) Biological variation gives rise to σ_{bet} (green circle). (**c**) Paired design uses the same aliquot for both measurements, mitigating between-subject variation.

we measure the same aliquot twice, we expect variability owing to technical variation inherent in our laboratory equipment and variability of the sample over time (**Fig. 2a**). This is called within-subject variation, σ_{wit}. If we measure two different aliquots with the same factor level, we also expect biological variation, called between-subject variation, σ_{bet}, in addition to the technical variation (**Fig. 2b**). Typically there is more biological than technical variability ($\sigma_{bet} > \sigma_{wit}$). In an unpaired design, the use of different aliquots adds both σ_{wit} and σ_{bet} to the measured difference (**Fig. 2c**). In a paired design, which uses the paired t-test[4], the same aliquot is used and the impact of biological variation (σ_{bet}) is mitigated (**Fig. 2c**). If differences in aliquots (σ_{bet}) are appreciable, variance is markedly reduced (to within-subject variation) and the paired test has higher power.

The link between σ_{bet} and σ_{wit} can be illustrated by an experiment to evaluate a weight-loss diet in which a control group eats normally and a treatment group follows the diet. A comparison of the mean weight after a month is confounded by the initial weights of the subjects in each group. If instead we focus on the change in weight, we remove much of the subject variability owing to the initial weight.

If we write the total variance as $\sigma^2 = \sigma_{wit}^2 + \sigma_{bet}^2$, then the variance of the observed quantity in **Figure 2c** is $2\sigma^2$ for the unpaired design but $2\sigma^2(1 - \rho)$ for the paired design, where $\rho = \sigma_{bet}^2/\sigma^2$ is the correlation coefficient (intraclass correlation). The relative difference is captured by ρ of two measurements on the same aliquot, which must be included because the measurements are no longer independent. If we ignore ρ in our analysis, we will overestimate the variance and obtain overly conservative P values and CIs. In the case where there is no additional variation between aliquots, there is no benefit to using the same aliquot: measurements on the same aliquot are uncorrelated ($\rho = 0$) and variance of the paired test is the same as the variance of the unpaired. In contrast, if there is no variation in measurements on the same aliquot except for the treatment effect ($\sigma_{wit} = 0$), we have perfect correlation ($\rho = 1$). Now, the difference measurement derived from the same aliquot removes all the noise; in fact, a single pair of aliquots suffices for an exact inference. Practically, both sources of variation are present, and it is their relative size—reflected in ρ—that determines the benefit of using the paired t-test.

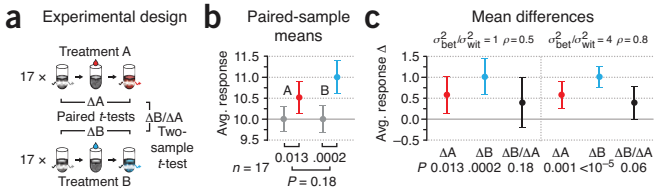

Figure 3 | Design and reporting for a paired, single-factor experiment. (**a**) The same $n = 17$ sample is used to measure the difference between treatment and background ($\Delta A = A_{after} - A_{before}$, $\Delta B = B_{after} - B_{before}$), analyzed with the paired t-test. Two-sample t-test is used to compare the difference between responses (ΔB versus ΔA). (**b**) Simulated sample means and P values for measurements and comparisons in **a**. (**c**) Mean difference, CIs and P values for two variance scenarios, $\sigma_{bet}^2/\sigma_{wit}^2$ of 1 and 4, corresponding to ρ of 0.5 and 0.8. Total variance was fixed: $\sigma_{bet}^2 + \sigma_{wit}^2 = 1$. All error bars show 95% CI.

We can see the improved sensitivity of the paired design (**Fig. 3a**) in decreased P values for the effects of A and B (**Fig. 3b** versus **Fig. 1b**). With the between-subject variance mitigated, we now detect an effect for A ($P = 0.013$) and an even lower P value for B ($P = 0.0002$) (**Fig. 3b**). Testing the difference between ΔA and ΔB requires the two-sample t-test because we are testing different aliquots, and this still does not produce a significant result ($P = 0.18$). When reporting paired-test results, sample means (**Fig. 3b**) should never be shown; instead, the mean difference and confidence interval should be shown (**Fig. 3c**). The reason for this comes from our discussion above: the benefit of pairing comes from reduced variance because $\rho > 0$, something that cannot be gleaned from **Figure 3b**. We illustrate this in **Figure 3c** with two different sample simulations with same sample mean and variance but different correlation, achieved by changing the relative amount of σ_{bet}^2 and σ_{wit}^2. When the component of biological variance is increased, ρ is increased from 0.5 to 0.8, total variance in difference in means drops and the test becomes more sensitive, reflected by the narrower CIs. We are now more certain that A has a real effect and have more reason to believe that the effects of A and B are different, evidenced by the lower P value for $\Delta B/\Delta A$ from the two-sample t-test (0.06 versus 0.18; **Fig. 3c**). As before, P values should be adjusted with multiple-test correction.

The paired design is a more efficient experiment. Fewer aliquots are needed: 34 instead of 51, although now 68 fluorescence measurements need to be taken instead of 51. If we assume $\sigma_{wit} = \sigma_{bet}$ ($\rho = 0.5$; **Fig. 3c**), we can expect the paired design to have a power of 97%. This power increase is highly contingent on the value of ρ. If σ_{wit} is appreciably larger than σ_{bet} (i.e., ρ is small), the power of the paired test can be lower than for the two-sample variant. This is because total variance remains relatively unchanged ($2\sigma^2(1 - \rho) \approx 2\sigma^2$) while the critical value of the test statistic can be markedly larger (particularly for small samples) because the number of degrees of freedom is now $n - 1$ instead of $2(n - 1)$. If the ratio of σ_{bet}^2 to σ_{wit}^2 is 1:4 ($\rho = 0.2$), the paired test power drops from 97% to 86%.

To analyze experimental designs that have more than two levels, or additional factors, a method called analysis of variance is used. This generalizes the t-test for comparing three or more levels while maintaining better power than comparing all sets of two levels. Experiments with two or more levels will be our next topic.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M.I. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
2. Krzywinski, M.I. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
3. Gelman, A. & Stern, H. *Am. Stat.* **60**, 328–331 (2006).
4. Krzywinski, M.I. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.