POINTS OF SIGNIFICANCE

Analysis of variance and blocking

Good experimental designs mitigate experimental error and the impact of factors not under study.

Reproducible measurement of treatment effects requires studies that can reliably distinguish between systematic treatment effects and noise resulting from biological variation and measurement error. Estimation and testing of the effects of multiple treatments, usually including appropriate replication, can be done using analysis of variance (ANOVA). ANOVA is used to assess statistical significance of differences among observed treatment means based on whether their variance is larger than expected because of random variation; if so, systematic treatment effects are inferred. We introduce ANOVA with an experiment in which three treatments are compared and show how sensitivity can be increased by isolating biological variability through blocking.

Last month, we discussed a one-factor three-level experimental design that limited interference from biological variation by using the same sample to establish both baseline and treatment values[1]. There we used the *t*-test, which is not suitable when the number of factors or levels increases, in large part due to its loss of power as a result of multiple-testing correction. The two-sample *t*-test is a specific case of ANOVA, but the latter can achieve better power and naturally account for sources of error. ANOVA has the same requirements as the *t*-test: independent and randomly selected samples from approximately normal distributions with equal variance that is not under the influence of the treatments[2].

Here we continue with the three-treatment example[1] and analyze it with one-way (single-factor) ANOVA. As before, we simulated samples for $k = 3$ treatments each with $n = 6$ values (**Fig. 1a**). The ANOVA null hypothesis is that all samples are from the same distribution and have equal means. Under this null, between-group variation of sample means and within-group variation of sample values are predictably related. Their ratio can be used as a test statistic, F, which will be larger than expected in the presence of treatment effects. Although it appears that we are testing equality of variances, we are actually testing whether all the treatment effects are zero.

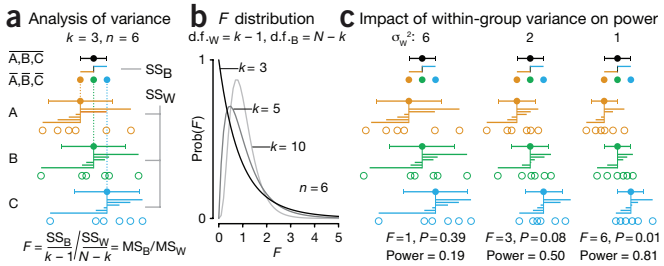

Figure 1 | ANOVA is used to determine significance using the ratio of variance estimates from sample means and sample values. (**a**) Between- and within-group variance is calculated from SS_B, the between treatment sum of squares, and SS_W, the within treatment sum of squares.. Deviations are shown as horizontal lines extending from grand and sample means. The test statistic, F, is the ratio mean squares MS_B and MS_W, which are SS_B and SS_W weighted by d.f. (**b**) Distribution of F, which becomes approximately normal as k and N increase, shown for $k = 3$, 5 and 10 samples each of size $n = 6$. $N = kn$ is the total number of sample values. (**c**) ANOVA analysis of sample sets with decreasing within-group variance ($\sigma_w^2 = 6,2,1$). $MS_B = 6$ in each case. Error bars, s.d.

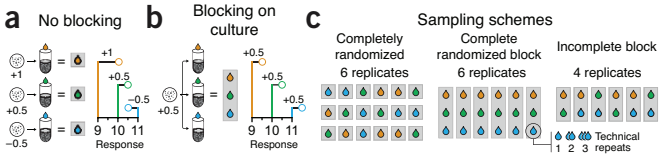

Figure 2 | Blocking improves sensitivity by isolating variation in samples that is independent from treatment effects. (**a**) Measurements from treatment aliquots derived from different cell cultures are differentially offset (e.g., 1, 0.5, −0.5) because of differences in cultures. (**b**) When aliquots are derived from the same culture, measurements are uniformly offset (e.g., 0.5). (**c**) Incorporating blocking in data collection schemes. Repeats within blocks are considered technical replicates. In an incomplete block design, a block cannot accommodate all treatments.

ANOVA calculations are summarized in an ANOVA table, which we provide for **Figures 1**, **3** and **4** (**Supplementary Tables 1–3**) along with an interactive spreadsheet (**Supplementary Table 4**). The sums of squares (SS) column shows sums of squared deviations of various quantities from their means. This sum is performed over each data point—each sample mean deviation (**Fig. 1a**) contributes to SS_B six times. The degrees of freedom (d.f.) column shows the number of independent deviations in the sums of squares; the deviations are not all independent because deviations of a quantity from its own mean must sum to zero. The mean square (MS) is SS/d.f. The F statistic, $F = MS_B/MS_W$, is used to test for systematic differences among treatment means. Under the null, F is distributed according to the F distribution for $k - 1$ and $N - k$ d.f. (**Fig. 1b**). When we reject the null, we conclude that not all sample means are the same; additional tests are required to identify which treatment means are different. The ratio $\eta^2 = SS_B/(SS_B + SS_W)$ is the coefficient of variation (also called R^2) and measures the fraction of the total variation resulting from differences among treatment means.

We previously introduced the idea that variance can be partitioned: within-group variance, σ_{wit}^2, was interpreted as experimental error and between-group variance, σ_{bet}^2, as biological variation[1]. In one-way ANOVA, the relevant quantities are MS_W and MS_B. MS_W corresponds to variance in the sample after other sources of variation have been accounted for and represents experimental error (σ_{wit}^2). If some sources of error are not accounted for (e.g., biological variation), MS_W will be inflated. MS_B is another estimate for MS_W, additionally inflated by average squared deviation of treatment means from the

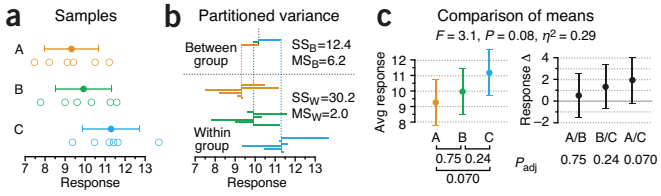

Figure 3 | Application of one-factor ANOVA to comparison of three samples. (**a**) Three samples drawn from normal distributions with $\sigma_{wit}^2 = 2$ and treatment means $\mu_A = 9$, $\mu_B = 10$ and $\mu_C = 11$. (**b**) Depiction of deviations with corresponding SS and MS values. (**c**) Sample means and their differences. P values for paired sample comparison are adjusted for multiple comparison using Tukey's method. Error bars, 95% CI.

© 2014 Nature America, Inc. All rights reserved.

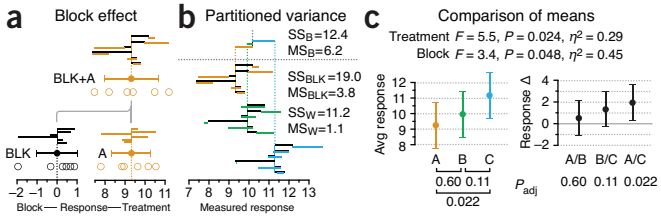

Figure 4 | Including blocking isolates biological variation from the estimate of within-group variance and improves power. (**a**) Blocking is simulated by augmenting each sample ($\sigma_{wit}^2 = 1$) with a fixed random component ($\mu_{blk} = 0$, $\sigma_{blk}^2 = 1$). (**b**) Variance is partitioned to treatment, block (black lines) and within-group. (**c**) Summary statistics for treatment and block effects in the same format as **Figure 3c**. In the presence of a sufficiently large blocking effect, MS_W is lowered and treatment test statistic $F = MS_B/MS_W$ is increased. Smaller error bars on sample mean differences reflect reduced MS_W.

grand mean, θ^2, times sample size if the null hypothesis is not true ($\sigma_{wit}^2 + n\theta^2$). Thus, the noisier the data (σ_{wit}^2), the more difficult it is to tease out σ_{treat}^2 and detect real effects, just like in the *t*-test, the power of which could be increased by decreasing sample variance[2]. To demonstrate this, we simulated three different sample sets in **Figure 1c** with $MS_B = 6$ and different MS_W values, for a scenario with fixed treatment effects ($\sigma_{treat}^2 = 1$), but progressively reduced experimental error ($\sigma_{wit}^2 = 6,2,1$). As noise within samples drops, a larger fraction variation is allocated to MS_B, and the power of the test improves. This suggests that it is beneficial to decrease MS_W. We can do this through a process called blocking to identify and isolate likely sources of sample variability.

Suppose that our samples in **Figure 1a** were generated by measuring the response to treatment of an aliquot of cells—a fixed volume of cells from a culture (**Fig. 2a**). Assume that it is not possible to derive all required aliquots from a single culture or that it is necessary to use multiple cultures to ensure that the results generalize. It is likely that aliquots from different cultures will respond differently owing to variation in cell concentration, growth rates, medium composition, among others. These so-called nuisance variables confound the real treatment effects: the baseline for each measurement unpredictably varies (**Fig. 2a**). We can mitigate this by using the same cell culture to create three aliquots, one for each treatment, to propagate these differences equally among measurements (**Fig. 2b**). Although measurements between cultures still would be shifted, the relative differences between treatments within the same culture remain the same. This process is called blocking, and its purpose is to remove as much variability as possible to make differences between treatments more evident. For example, the paired *t*-test implements blocking by using the same subject or biological sample.

Without blocking, cultures, aliquots and treatments are not matched—a completely randomized design (**Fig. 2c**)—which makes differences in cultures impossible to isolate. For blocking, we systematically assign treatments to cultures, such as in a randomized complete block design, in which each culture provides a replicate of each treatment (**Fig. 2c**). Each block is subjected to each of the treatments exactly once, and we can optionally collect technical repeats (repeating data collection from the measurement apparatus or multiple aliquots from the same culture) to minimize the impact of fluctuations in our measuring apparatus; these values would be averaged. In the case where a block cannot support all treatments (e.g., a culture yields only two aliquots), we would use combinations of treatment pairs with the requirement that each pair is measured equally often—a balanced incomplete block design. Let us look at how blocking can increase ANOVA sensitivity using the scenario from **Figure 1**.

We will start with three samples ($n = 6$) (**Fig. 3a**) that measure the effects of treatments A, B and C on aliquots of cells in a completely randomized scheme. We simulated the samples with $\sigma_{wit}^2 = 2$ to represent experimental error. Using ANOVA, we partition the variation (**Fig. 3b**) and find the mean squares for the components ($MS_B = 6.2$, $MS_W = 2.0$; **Supplementary Table 2**). MS_W reflects the value $\sigma_{wit}^2 = 2$ in the sample simulation, and it turns out that this variance is too high to yield a significant *F*; we find $F = 3.1$ ($P = 0.08$; **Fig. 3c**). Because we did not find a significant difference using ANOVA, we do not expect to obtain significant *P* values from two-sample *t*-tests applied pairwise to the samples. Indeed, when adjusted for multiple-test correction these P_{adj} values are all greater than 0.05 (**Fig. 3c**).

To illustrate blocking, we simulate samples to have the same values as in **Figure 3a** but with half of the variance due to differences in cultures. These differences in cultures (block effect) are simulated as normal with mean $\mu_{blk} = 0$ and variance $\sigma_{blk}^2 = 1$ (**Fig. 4a**), and are added to each of the sample values using the complete randomized block design (**Fig. 2c**). The variance within a sample is thus evenly split between the block effect and the remaining experimental error, which we presumably cannot partition further. The contribution of the block effect to the deviations is shown in **Figure 4b**, now a substantial component of the variance in each sample, unlike in **Figure 3b**, where blocking was not accounted for.

Having isolated variation owing to cell-culture differences, we increased sensitivity in detecting a treatment effect because our estimate of within-group variance is lower. Now $MS_W = 1.1$ and $F = 5.5$, which is significant at $P = 0.024$ and allows us to conclude that the treatment means are not all the same (**Fig. 4c**). By doing a *post hoc* pairwise comparison with the two-sample *t*-test, we can conclude that treatments A and C are different at an adjusted $P = 0.022$ (95% confidence interval (CI), 0.30–3.66) (**Fig. 4c**). We can calculate the *F* statistic for the blocking variable using $F = MS_{blk}/MS_W = 3.4$ to determine whether blocking had a significant effect. Mathematically, the blocking variable has the same role in the analysis as an experimental factor. Note that just because the blocking variable soaks up some of the variation we are not guaranteed greater sensitivity; in fact, because we estimate the block effect as well as the treatment effect, the within-group d.f. in the analysis is lower (e.g., changes from 15 to 10 in our case); our test may lose power if the blocks do not account for sufficient sample-to-sample variation.

Blocking increased the efficiency of our experiment. Without it, we would need nearly twice as large samples ($n = 11$) to reach the same power. The benefits of blocking should be weighed against any increase in associated costs and the decrease in d.f.: in some cases it may be more sensible to simply collect more data.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth.3005).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 597–598 (2014).
2. Krzywinski, M. & 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.

© 2014 Nature America, Inc. All rights reserved.