POINTS OF SIGNIFICANCE

Sources of variation

To generalize conclusions to a population, we must sample its variation.

Variability is inevitable in experiments owing to both biological and technical effects. Whereas technical variability should be tightly controlled to enhance the internal validity of the results, some types of biological variability need to be maintained to allow generalization of the results to the population of interest. Experimental control, randomization, blocking and replication are the tools that allow replicable and meaningful results to be obtained in the face of variability.

In previous columns we have given examples of how variation limits our ability to detect effects by reducing the power of tests. This month we go into more detail about variability and how it affects our ability to replicate the experimental results (internal validity) and generalize from our experiment to the population (external validity).

Let's start with an idealized experiment, which we will then expand upon. Suppose that we are able to culture a single murine cell under tightly controlled conditions so that the response of different aliquots of the culture is identical. Also, suppose that our measuring device is so accurate that the difference between measurements of an aliquot is below the detection limit. If measurement does not disrupt the cell culture, we require only a single aliquot: we measure the baseline response, apply the treatment and measure the treatment response. No replication is needed because differences between the measurements can only be due to the treatment.

This idealized system has perfect internal validity—the response variable solely reflects the treatment effect, and repeating the experiment on another aliquot from the same cell culture will give identical results. However, the system lacks external validity—it tells us about only a specific cell from a specific mouse. We know that cells vary within a single tissue, and that tissues vary from mouse to mouse, but we cannot use this ideal system to make inferences about other cell cultures or other mice because we have no way of determining how much variability to expect. To do so requires that we sample the biological variation across relevant experimental variables (**Fig. 1**).

A well-designed experiment is a compromise between internal and external validity. Our goal is to observe a reproducible effect that can be due only to the treatment (avoiding confounding and bias) while simultaneously measuring the variability required to estimate how much we expect the effect to differ if the measurements are repeated with similar but not identical samples (replicates).

When administering the treatment *in vivo*, we can never control the many sources of biological variability in the mice sufficiently to achieve identical measurements for different animals. However, with careful design, we can reduce the impact of this variability on our measurements by controlling some of these factors.

Genotype and gender are examples of sources of variability that are under complete experimental control. We can eliminate the source entirely by selecting a single level or select several levels so that the effects can be determined. For gender we can observe all the possible levels, so we can treat gender as a fixed factor in our experiment. Genotype can be a fixed effect (specific genotypes of interest, such as a mutant and its background wild type) or a random (noise) effect (several wild-type strains representing the wild-type population). Only by deliberately introducing variability can we make general statements about treatment effect—and then only across factors that were varied.

Other sources of variability, such as diet, temperature and other housing effects, are under partial experimental control. Noise factors that cannot be controlled, or are unknown, can be handled by random assignment[1] (to avoid bias), replication[2] (to increase precision) and blocking[3] (to isolate noise).

When dealing with variation, two principles apply: the precision with which we can characterize a sample (e.g., s.e.m.) and the manner in which variances from different sources combine together[4]. The s.e.m. of a random sample is $\sigma/\sqrt{n}$, where σ is the s.d. of the population (also written as $\mathrm{Var}(\overline{X}) = \mathrm{Var}(X)/n$). With sufficient replication (large n), our precision in measuring the mean as measured by the s.e.m. can be made arbitrarily small (**Fig. 2a**). When multiple independent sources of variation are present, the variance of the measurement is the sum of individual variances.

These two principles can be combined to obtain the variation of the mean in a nested replication scenario[2] (**Supplementary Fig. 1**). Suppose that variances due to mouse, cell and measurement are M, C and ε (Var() is omitted for brevity). The variance of the measurement of a single cell will be $M + C + \varepsilon$, the sum of the individual variances. If we measure the same cell n_ε times, the variance of the average measurement will be $M + C + \varepsilon/n_\varepsilon$. If we measure n_C cells, each n_ε times, the variance will be $M + C/n_C + \varepsilon/(n_C \times n_\varepsilon)$. Finally, if we repeat the procedure for n_M mice, the variance will be reduced to $M/n_M + C/(n_M \times n_C) + \varepsilon/(n_M \times n_C \times n_\varepsilon)$. In general, the variance of each source is divided by the number of times that source is independently sampled. This is illustrated in **Figure 2b** for $M = 1$, $C = 4$ and $\varepsilon = 0.25$. As we have already seen[2], the number of replicates at each layer (n_M, n_C, n_ε) can be controlled to optimally reduce variation (increase power) within practical constraints (cost). For example, to reduce the total variance to 25% of the total $M + C + \varepsilon$, we can sample using $n_M = 4$, $n_C = 1$ or $n_M = n_C = 3$ (**Fig. 2b**). Sampling a single mouse allows us to reduce variance only to M, but it would not allow us to estimate the variation at the mouse layer and therefore would not allow for inference about the population of mice. For our example, technical variation is much smaller than biological variation, and technical replicates are of little value—variance is reduced by only 5% for $n_M = n_C = 1$ and $n_\varepsilon = 10$ (**Fig. 2b**, gray trace) and can be reduced only to $M + C$.

When measurements themselves are an average of a large number of contributing factors, biological variability of the components can be underestimated. For example, measuring two samples from the

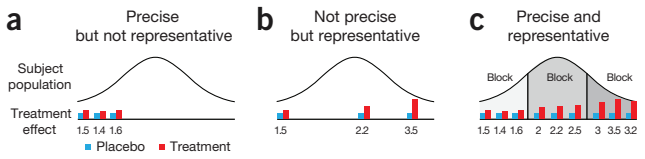

Figure 1 | Internal and external validity relate respectively to how precise and representative the results are of the population of interest. (**a**) Sampling only a part of the population may create precise measurements, but generalizing to the rest of the population can result in bias. (**b**) Better representation can be achieved by sampling across the population, but this can result in highly variable measurements. (**c**) Identifying blocks of similar subjects within the population increases the precision (within block) and captures population variability (between blocks).

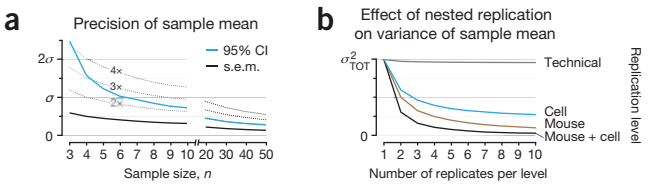

Figure 2 | In the presence of variability, the precision in sample mean can be improved by increasing the sample size, or the number of replicates in a nested design. (**a**) Increasing the sample size, n, improves the precision in the mean by $1/\sqrt{n}$ as measured by the s.e.m. The 95% CI is a more intuitive measure of precision: the range of values that are not significantly different at $\alpha = 0.05$ from the observed mean. The 95% confidence interval (CI) shrinks as $t^*/\sqrt{n}$, where t^* is the critical value of the Student's t-distribution at two-tailed $\alpha = 0.05$ and $n - 1$ degrees of freedom. t^* decreases from 4.3 ($n = 3$) to 2.0 ($n = 50$). Dotted lines represent constant multiples of the s.e.m. (**b**) For a nested design with mouse, cell and technical variances of $M = 1$, $C = 4$, $\varepsilon = 0.25$ ($\sigma^2_{TOT} = 5.25$), the variance of the mean decreases with the number of replicates at each layer.

same homogenized tissue, gives us the average of all cells. There is essentially no biological variation in these measurements because n in the s.e.m. term is very large—the only variability that we are likely to find is due to measurement error. We must not confuse the reproducibility of the tissue average with response of individual cells, which can be quite variable.

Blocking[3] on a noise variable allows us to remove a noise effect by taking a difference of two measurements that share the same value of the noise (e.g., same sample before and after treatment). Blocking enhances external validity—within the block, variability is controlled as tightly as possible for internal validity. The blocks themselves are chosen to cover the range of variability needed to estimate the response variability in the population of interest (**Fig. 1c**). This is the approach taken by the paired t-test, in which the block is a subject. For another example, a heterogeneous tissue could not be homogenized and a block would be defined by a spatial boundary between different cells. Neglecting to account for this would disregard the block boundaries in **Figure 1c** and would reduce sensitivity.

There can also be multiple sources of technical variability, such as reagents, measurement platforms and personnel. The same principles apply as for biological inference, measures of technical variability are seldom of interest—the usual objective is to minimize it. Blocking may still be used to eliminate known sources of noise—for example, collaborating labs may each do one complete replicate of an experiment to provide sufficient replication while eliminating any variability due to lab effects in the treatment comparisons.

Consider an experiment that assesses the effect of a drug on the livers of male mice of a specific genotype, at both the animal and cell layers. If the drug is administered *in vivo*, the animal is euthanized and the response measured on many cells, animals exposed to the drug cannot be their own controls. So, we expect variability at both the mouse layer and the cell (within mouse) layer. As well, we expect variability due to cell culture and maternal effects.

In the simplest experiment, we have a nested design, with mice selected at random for the treatment and the control. After dissection, cells are sampled from each liver, and their response to the drug is measured. The total variation of the measurement is the sum of variances of each effect, weighted by the number of times the effect was independently sampled (**Fig. 2b**). Using the same variances as above and $(n_M, n_C, n_\varepsilon) = (10, 5, 3)$ we find $\mathrm{Var}(\bar{X}) = 1/10 + 4/50 + 0.25/150 = 0.18$. The variance of the difference in the means of two measurements (e.g., reference and drug) will be twice this, 0.36, and our power to detect an effect of $d = 1.5$ is 0.65 (**Supplementary Note**).

Suppose that we discover that the mouse variation, $M = 1$, has significant components from maternal and cell culture effects, given by variances M_{MAT} and M_{CELL}. In this context, we can partition $M = M_{MAT} + M_{CELL} + M_0$, where M_0 is the unique variance not attributable to maternal or cell culture effects. We can attempt to control maternal effects by using sibling pairs (a block) and subjecting one mouse from each pair to the drug and one to the control. As the pairs have the same mother, the maternal effects cancel. Similarly, variance due to cell culture effects can be minimized by concurrently euthanizing each sibling pair (another block) and jointly preparing the cell cultures.

Having blocked these two effects, although M_{MAT} and M_{CELL} still contribute to the variance for both control and drug, we have effectively removed them from the variance of the difference in means. If these effects account for half of the mouse variance, $M_{MAT} + M_{CELL} = M/2 = 0.5$ (using $M = 1$ as above), blocking reduces the variance in the difference by $2(M_{MAT} + M_{CELL})/10$ from 0.36 to 0.26 and increases our power to 0.79 (**Supplementary Note**).

We can use the concept of effective sample size, $n = \mathrm{Var}(X)/\mathrm{Var}(\bar{X})$, to demonstrate the effect of this blocking. In the nested replication design, n is typically smaller than the total number of measurements ($n_M \times n_C \times n_\varepsilon$) because we do not independently sample each source of variation in each measurement[2] (it is largest for $n_C = n_\varepsilon = 1$). As a result, replication at the cell and technical layers decreases $\mathrm{Var}(\bar{X})$ proportionally more slowly than replication at the topmost mouse layer. When both maternal and cell culture effects are included, $\mathrm{Var}(X) = M + C + \varepsilon = 5.25$ and the effective sample size is $n = 5.25/0.36 = 15$. When maternal and cell effects are blocked, $\mathrm{Var}(X)$ remains the same, but now $\mathrm{Var}(\bar{X})$ is reduced to 0.26 and $n = 5.25/0.26 = 20$.

Given the choice, we should always block at the top layer because the noise in this layer is independently sampled the fewest times. We can use the effective sample size n to illustrate this. Blocking at mouse layer decreased M from 1 to 0.5 (by 50%) and increased n from 15 to 20 (power from 0.65 to 0.79). In contrast, a proportional reduction in C from 4 to 2 increases n to 19 (power to 0.76), whereas a reduction in ε has essentially no effect on n.

We need to distinguish between sources of variation that are nuisance factors in our goal to measure mean biological effects from those that are required to assess how much effects vary in the population. Whereas the former should be minimized to optimize the power of the experiment, the latter need to be sampled and quantified so that we can both generalize our conclusions and robustly determine the uncertainty in our estimates.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3224).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Krzywinski, M. & Altman, *Nat. Methods* **11**, 597–598 (2014).
2. Blainey, P., Krzywinski, M. & Altman, *Nat. Methods* **11**, 879–880 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).

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