VIEWPOINT

Reproducibility

Determining how many cells to average for statistical testing of microscopy experiments

Adam Zweifach[1]

From a statistical standpoint, individual cells are typically not independent experimental replicates. To test for differences in mean, cells from each experimental sample can be averaged and each sample's average treated as an *n* of 1. Here, I outline how to determine how many cells to average per sample.

Individual cells are usually not independent experimental replicates

Microscopy is a core tool in cell research that can generate single-cell data. However, single cells (or objects within them like organelles) are rarely independent experimental replicates. Reasons for this include that the entire sample is usually treated at once and it is treatment that defines an experimental replicate, cells in a sample can influence one another via gaseous, soluble, or contact-mediated signals, and all the cells within the same dish experienced the same history (passage number, position in incubator, etc.) (Vaux et al., 2012). Failure to recognize that cells are not independent replicates when testing for differences in means with *t* tests or ANOVA causes pseudoreplication (Lazic, 2010; Lord et al., 2020; Eisner, 2021), an extremely common and serious error that artificially decreases P values and can create false positive effects.

There are two common statistically appropriate strategies for testing for differences in means when measurements from individual cells are available but are not independent replicates. The simple approach is to average values from a number of cells in each sample and treat each sample's average as an *n* of 1 in statistical tests. A more sophisticated approach is to treat individual objects as nested within their replicate and sample and use appropriate multilevel statistical models to calculate P values (Aarts et al., 2014; Dowding and Haufe, 2018). However, recent work casts doubt on the supposed benefits of the nested approach (McNabb and Murayama, 2021), and implementing it correctly requires specialized statistical knowledge and software that many researchers may not possess. When modeling is done incorrectly or does not converge properly it can also result in low P values that generate errors like those caused by pseudoreplication. Blainey et al. explored how best to distribute different numbers of replicates to different levels in multilevel modeling of a gene sequencing experiment (Blainey et al., 2014). The present tutorial is limited in scope to the simple averaging approach. Estimating how many cells from each replicate to include in a multilevel statistical model treatment is not addressed.

Averaging is associated with a particular kind of noise that decreases statistical power

If researchers choose to use the simple averaging approach, they must decide how many cells to average. Averaging is associated with a particular kind of variability or error—averaging error—since a different value will be generated each time a different subset of cells is averaged. The size of averaging error depends both on how the parameter is distributed in the sample and on the number of cells averaged. The standard error of the mean, estimated as the sample standard deviation (SD) divided by the square root of the number averaged, is a measure of averaging noise. If the population distribution is narrow, only a few cells need to be averaged to achieve low averaging noise, but if it is broad many cells must be averaged. Multiple sources of variability affect every experiment, and total experimental variability is the square root of the sum of the squares of all of them. If averaging noise is large, total overall experimental variability will tend to be high and experimental power, the likelihood of detecting effects by getting a P value below the cutoff chosen for significance (usually 0.05), will likely be low. When power is low, real effects will be missed and any effects that are found will be more likely to be false positives (Button et al., 2013; Colquhoun, 2014). Thus, achieving high statistical power is very important. It is easier and cheaper to analyze more cells than to perform additional replicates. However, because of the other sources of variability there are limits to what averaging more cells can achieve, and averaging more than necessary can waste time and effort for no real benefit.

We can estimate the size of averaging noise but not other sources of noise

To calculate the number of cells to average, we would have to determine both the size of averaging error and the aggregate size of all

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA.

Correspondence to Adam Zweifach: adam.zweifach@uconn.edu.

Rockefeller University Press

https://doi.org/10.1083/jcb.202401074

the other sources of variability. There are methods that can be used to do this (Searle, 1995), but while the size of averaging error can be estimated reasonably accurately using these methods, simulations I conducted suggest it may require more experimental replicates than most researchers are likely to acquire to estimate the size of the other errors effectively. One can think about the size of errors in terms of percent coefficient of variation (% CV), which is 100 × SD/mean, as this offers easy comparison to effect sizes when they are also expressed as a percentage. For the discussion that follows, % CV is used to discuss both noise sources and effect sizes. Researchers can normalize results from their experiments so that errors can be expressed as % CV by calculating the average of all the control samples, then dividing all experimental values by that average and multiplying them by 100. (An Excel sheet included as Data S1 that calculates the number of cells to average normalizes data to % CV as part of its function.) I further assume that samples are spatially uniform, and that the parameter of interest is a property of whole cells such as their size, speed, numbers of a particular organelle, levels of a particular protein or probe, etc., at a single time point. More complex situations could be analyzed but are beyond the scope of this work.

Thinking about experiments and power suggests that a good target for averaging noise is ∼2.5–5% CV

While we may not be able to estimate all the noise components needed to calculate how many cells to average, if we think about the size of effects and errors in common experiments and make some reasonable assumptions, having an estimate of between-cell SD turns out to be sufficient to allow us to set a target for averaging error. We can distinguish two basic types of cell-based experiments based on whether time-dependent biological variability occurring between repeats of the experiment is shared by samples. This time-dependent variability can be as large as 20–50% CV in both primary and immortalized cells (Molloy et al., 2003) and is likely to be the most significant source of variation other than averaging noise when a technique like microscopy is used (Zweifach, 2024). In the first type of experiment, samples come from different sources such as different tissues or primary cell isolations, different immortal cell lines, or different stable transfectants of the same line. These will not share time-dependent biological variability because they are independent uncorrelated samples. Like all noise, time-dependent biological variability from two sources sums as the square root of each squared. We can calculate the minimum size of the difference between two experimental groups that can be detected 80% of the time (i.e., with 80% power, the minimum recommended by most statisticians) using Student's *t* test, the appropriate test for this experiment (Zweifach, 2024), over a range of averaging noise in the presence of different levels of other experimental errors, including time-dependent biological variability. I did this for three replicates, the most common value reported in the literature, and found that if total non-averaging errors including time-dependent biological variability are <5% CV, reducing averaging error to 5% or less would allow effects 25% or smaller to be detected with 80% power (Fig. 1 A, lower three traces) and reducing averaging noise more would further decrease the size of the effect that could be detected. However, small effects may not be biologically important, and each twofold reduction in averaging noise requires averaging four times as many cells. If the sum of non-averaging errors is larger—which seems more likely—reducing averaging error below ∼5% does relatively little to further reduce the minimum detectable effect size (Fig. 1 A, upper three traces). Power can be increased by performing additional replicates, but even with larger sample sizes there is little practical benefit to reducing averaging error below ∼5% (see Fig. S1).

In the second kind of experiment, measurements are made on cells taken from a single source that is split into aliquots that either serve as a control or are treated in some way, such as with drugs or genetic reagents. This is a matched (or blocked, or paired) experimental design. In this experiment, samples do share time-dependent biological variability, but its effects on statistical tests can essentially be eliminated by using a test (like a paired *t* test) that takes this into account, provided averaging error and any other variation that affects samples individually is low enough (Zweifach, 2024). If the sum of errors that are neither time-dependent biological variability nor averaging error is 5% or smaller, reducing averaging error to 2.5% will ensure minimum detectable effect sizes <25% (Fig. 1 B, lower three traces). Again, though, if independent errors are larger, little is gained by reducing averaging error further. Taken together, it seems that a reasonable target for averaging noise in microscopy experiments is in the range of 2.5–5% CV.

Estimating how many cells to average to reach the target

The first step in determining how many cells to average to reach a 2.5–5% CV target is acquiring and analyzing some data so that averaging noise can be estimated. Every experimental workflow is different, and if acquiring and analyzing data is difficult and time-consuming researchers will probably want to estimate the number of cells to average from relatively small data sets, perhaps obtained from a pilot analysis of a single sample. In other cases, analysis might be easy enough that researchers can use larger data sets, perhaps after all data have been collected but before they have all been analyzed. As will become clear below, both strategies can be used, but larger samples give better estimates. Whatever researchers do, data should first be normalized as described above. If data from a single sample is to be used, sample SD can then be calculated and the number of cells to average will be:

$$Number\ to\ average = \lceil \left(\frac{SD}{target}\right)^2 \lceil$$

where the outer brackets indicate that the value should be rounded up to the nearest whole number. If data from multiple samples is used, researchers should subtract each sample's mean from all of that sample's values, then calculate the SD for the entire data set and use that value in the equation above. Subtracting the means removes the effect of differences between them on SD. The Excel spreadsheet included as Data S1 will perform the calculation for up to 16 cells from six replicates, which should be more than sufficient (see below).

To determine how well the procedure estimates between-cell SD, I used R software (R Core Team, 2020) to simulate data under a variety of conditions: different SD, normal or log normal distribution of the parameter within samples, different combinations of other experimental errors, and whether or not samples in a replicate shared time-dependent biological variability. For each condition, I generated 1,000 data sets with three replicates in each of two experimental

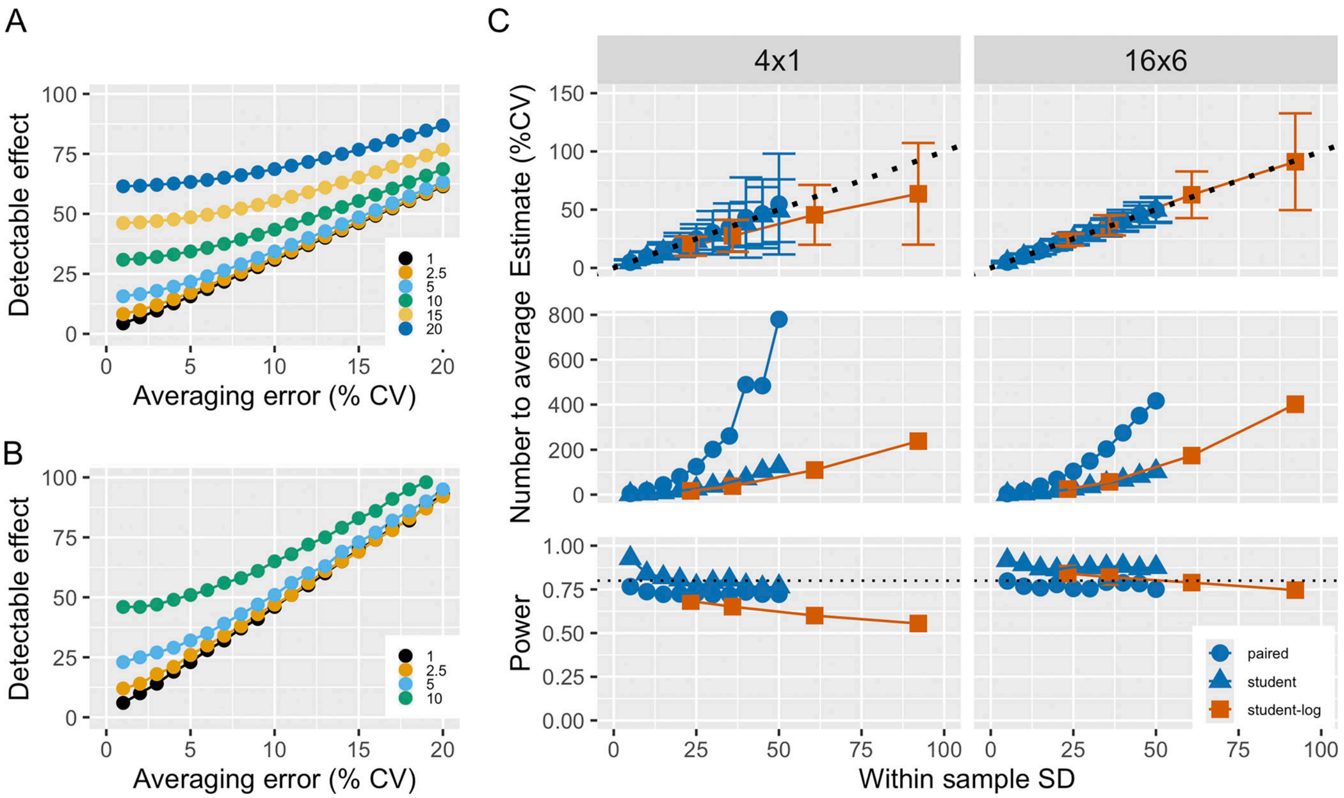

Figure 1. **Determining how many cells to average to achieve reasonable statistical power. (A)** Minimum effect detectible with 80% power as a function of averaging error when independent samples are analyzed with Student's *t* test. Other errors were (in % CV, from bottom trace to top trace): 1 (black), 2.5 (orange), 5 (light blue), 10 (green), 15 (yellow), and 20 (dark blue). *N* was 3 replicates. **(B)** Minimum effect detectible with 80% power as a function of averaging error when samples in replicates share error of 20% CV and are analyzed with paired *t* tests. Other errors were (in % CV, from bottom trace to top trace): 1 (black), 2.5 (orange), 5 (light blue), and 10 (green). *N* was 3 replicates. **(C)** Top row: Plots of estimated within-sample SD as a function of actual within-sample SD when four cells from a single condition are used (left) or 16 cells from each of six replicates are used (right) for the estimate. When six replicates were used, three were simulated with mean of 100 (control) and three with mean of 125 (treated). Blue traces (circles and triangles) were simulated with normally distributed values and red (squares) with log-normally distributed values. Error bars are SD. Blue triangles represent data in which samples shared between-replicate variability of 20% CV, while blue circles represent data simulated without shared variability. The dashed line has an intercept of zero and slope of 1. Middle row: Estimates of the number of cells to average per sample to reach a target of 5% CV for samples that do not share variability (blue triangles for normal data, red squares for log-normal) or 2.5% CV for data simulated to share between-replicate variability (blue circles). Conditions were chosen from A and B to correspond to an effect detectible with 80% power of ∼25 % CV. Bottom row: Power (fraction of P values <0.05) under the conditions described above. The dashed line indicates the expected value of 0.8 for all conditions.

groups. Each simulated sample had at least 10,000 "cells." I used the procedure that is implemented in the Excel spreadsheet to estimate between-cell SD using either four cells from one sample or 16 cells from each of the six samples. When values in samples were normally distributed, the procedure estimated averaging noise reasonably well up to a population SD of 50% CV even if only four cells from a single sample were used (Fig. 1 C, top row, left). 50% CV is a practical maximum since 95% of a normally distributed population is within 2 SD of the mean and in most cases biological values cannot be negative; for SD to be higher, data must have some right skew or be log-normally distributed. Estimates were more accurate when larger samples were used (Fig. 1 C, top row, compare the error bars on the right to those on the left). When multiple samples were used and variance differed in samples, the resulting estimate was an average of the SD in all the samples (not shown), which serves well for defining how many cells to average. If the parameter of interest was log-normally distributed in the population, estimates of between-cell SD were poor with only four cells but better when more cells were used. If estimates of between-cell SD obtained from a small number of cells are >50% CV, researchers should probably acquire more data and repeat the process.

The routines allowed me not only to estimate population SD but also to calculate the number of cells per sample to average to achieve 80% power to detect a difference of 25% between "mean" and "treated" groups using the correct test for the experiment (Fig. 1 C, middle row). For each data set, I calculated the number of cells per sample to average, then averaged that number of cells from the simulated data sets, and finally performed the appropriate statistical test using the averaged data. As expected, in most cases I got 700–900 P values out of 1,000 < 0.05 regardless of within-sample variation (Fig. 1 C, bottom row). This confirms that the procedure works as intended, rendering statistical test results independent of within-sample variation. While this work focuses on two experimental conditions analyzed with *t* tests, I expect similar behavior when three or more conditions are analyzed with ANOVA (or two-way ANOVA with one factor treatment and the other replicate, which is a better choice [Zweifach, 2024]) followed by post hoc tests.

Estimates of the number of cells per sample to average are what statisticians call point estimates. Averaged over many repetitions, using point estimates will lead to the desired statistical power. However, any single point estimate of SD might be either too high or too low, and an estimate of SD that is too low would lead researchers to average fewer cells than needed, making power lower than expected. The Excel calculator sheet that is included calculates both point estimates and estimates based on the upper bound of the 95% confidence interval for estimated SD (upper bound estimates). When data from many cells is entered and/or the SD within the samples is relatively small, the point estimate and the upper bound estimate will be similar. However, when data from only a few cells is used, or when the SD in samples is very large, the point estimate and the upper bound can be quite different. In such a case, researchers should consider using the upper bound estimate to be sure of averaging enough cells.

An example may help to make things clear. Imagine a simple experiment: dishes of cells taken from a common culture are treated either with vehicle or a drug, stained with an antibody, and the numbers of a particular kind of labeled structure are counted. After one replicate (of three planned) has been completed, 10 cells from each dish are imaged and analyzed, generating the values that prepopulate the Excel calculator sheet. For the cells in the spreadsheet, the point estimate of SD is 17.8% CV, but the 95% confidence interval extends to 26.0% CV. Because the final values on which statistical testing will be conducted are the average number of puncta per cell, we can ignore that individual cell data are counts and may not be distributed normally in cells in the population and use a *t* test to assess effects of the treatment. Because the cells were taken from the same culture and share time-dependent biological variability, a paired *t* test is likely to have higher power provided enough cells are averaged, so 2.5% CV should be entered as the averaging noise target in the Excel sheet. Using the point estimate of SD, we calculate that 51 cells should be averaged, but if the upper bound estimate is correct we should average 108 cells. As long it is not too difficult or time consuming to analyze data, it would be better to average 108 cells as it is more likely to result in the desired power. Had the experiment been conducted on two cell independent cell lines that did not share time-dependent biological variability, Student's *t* test would be appropriate, and the target for averaging error would be 5% CV. Entering this in the calculator sheet results in estimated numbers of cells to average of 13 for the point estimate and 27 for the upper bound.

Conclusion

Having a simple way to calculate how many cells to average should help researchers conduct more efficient and more powerful microscopy experiments. It will also hopefully decrease the incidence of pseudoreplication and thus help promote more robust and reproducible science.

Online supplemental material

Fig. S1 shows power at different sample sizes. Data S1 is an Excel sheet for estimating how many cells to average. Data S2 is a guide to using the Excel sheet. Data S3 shows the R code used in this work. Data S4 shows the derivation of the equation for the number of cells to average.

Acknowledgments

I would like to thank Dr. Haim Bar for helpful initial discussions, and Dr. Kenneth G. Campellone, Dr. David A. Knecht, and Nadine Lebek for comments on an early draft.

Author contributions: A. Zweifach: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft, Writing—review & editing.

Disclosures: The author declares no competing financial interests.

References

Aarts, E., et al. 2014. *Nat. Neurosci.* https://doi.org/10.1038/nn.3648

Blainey, P., et al. 2014. *Nat. Methods.* https://doi.org/10.1038/nmeth.3091

Button, K.S., et al. 2013. *Nat. Rev. Neurosci.* https://doi.org/10.1038/nrn3475

Colquhoun, D. 2014. *R. Soc. Open Sci.* https://doi.org/10.1098/rsos.140216

Dowding, I., and S. Haufe. 2018. *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2018.00103

Eisner, D.A. 2021. *J. Gen. Physiol.* https://doi.org/10.1085/jgp.202012826

Lazic, S.E. 2010. *BMC Neurosci.* https://doi.org/10.1186/1471-2202-11-5

Lord, S.J., et al. 2020. *J. Cell Biol.* https://doi.org/10.1083/jcb.202001064

McNabb, C.B., and K. Murayama. 2021. *Curr. Res. Neurobiol.* https://doi.org/10.1016/j.crneur.2021.100024

Molloy, M.P., et al. 2003. *Proteomics.* https://doi.org/10.1002/pmic.200300534

R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Searle, S.R. 1995. *Metrika.* https://doi.org/10.1007/BF01894301

Vaux, D.L., et al. 2012. *EMBO Rep.* https://doi.org/10.1038/embor.2012.36

Zweifach, A. 2024. *Mol. Biol. Cell.* https://doi.org/10.1091/mbc.E23-05-0159

Supplemental material

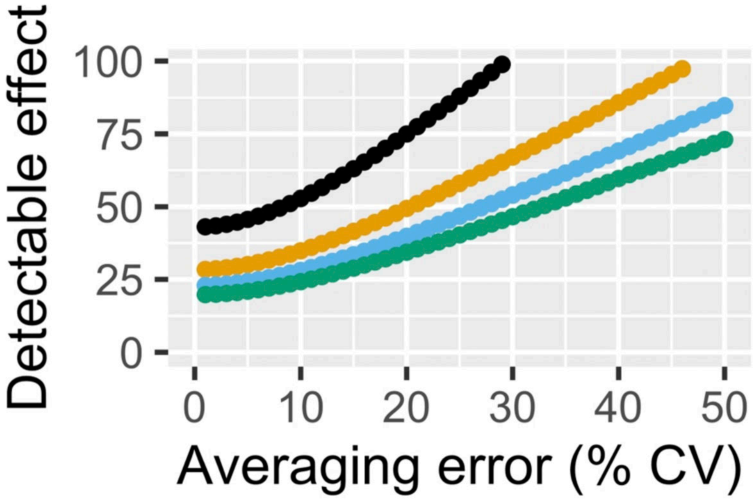

Figure S1. **Minimum detectible effect as a function of averaging error when independent variation is 20% CV and *n* is (from top) 3 (black), 5 (orange), 7 (light blue), or 9 (green).**

Provided online are Data S1, Data S2, Data S3, and Data S4. Data S1 is an Excel sheet for estimating how many cells to average. Data S2 is a guide to using the Excel sheet. Data S3 shows the R code used in this work. Data S4 shows the derivation of the equation for the number of cells to average.