If your $P$ value looks too good to be true, it probably is: Communicating reproducibility and variability in cell biology

Samuel J. Lord, Katrina B. Velle, R. Dyche Mullins, Lillian K. Fritz-Laylin

* Corresponding authors: dyche.mullins@ucsf.edu & lfritzlaylin@umass.edu

ABSTRACT: Cell biology is littered with erroneously tiny $P$ values, often the result of evaluating individual cells as independent samples. Because readers expect low $P$ values and small error bars to imply that an observed difference would persist if the experiment were to be duplicated, the sample size ($N$) used for statistical tests should actually be the number of times an experiment was repeated. $P$ values calculated using the number of cells do not reflect the reproducibility of the result and are thus highly misleading. To help authors avoid this mistake, we provide examples and practical tutorials to create figures that communicate both the cell-level variability and the experimental reproducibility.

SUMMARY SENTENCE: The sample size for a typical cell biology experiment is the number of times the experiment was repeated, not the number of cells measured.

$P = 0.00000000000000001$?! Think again

Error bars and $P$ values are often used to assure readers of a real and persistent difference between populations or treatments. $P$ values are based on the difference between population means (or other summary metric) as well as the number of measurements used to determine that difference. In general, increasing the number of measurements decreases the resulting $P$ value. To convey anything about experimental reproducibility, $P$ values and standard error of the mean should be calculated using independent measurements of a population of interest, typically observations from separate experiments (Lazic, 2010; Aarts et al., 2014; Naegle et al., 2015; Lazic et al., 2018). Limited time and resources usually constrain cell biologists to repeat any particular experiment only a handful of times, so a typical sample size $N$ is something like 4. Too often, however, authors mistakenly assign $N$ as the number of cells, or even the number of observed subcellular structures. This number may be on the order of hundreds or thousands, resulting in vanishingly small $P$ values and tiny error bars that are not useful for determining the reproducibility of an experiment.

Cell biology is a noisy discipline, and our measurements have many sources of variance. For example, quantitative experiments are difficult to replicate exactly from day-to-day and lab-to-lab. Well-designed studies embrace both cell-to-cell and sample-to-sample variation (Altman and Krzywinski, 2015). Unlike measuring the gravitational constant or the speed of light, repeatedly quantifying a biological parameter rarely converges on a single "true" value. Calculating standard error from thousands of cells conceals this expected variability. We have written this tutorial to help cell biologists calculate meaningful $P$ values and plot data to highlight both experimental robustness and cell-to-cell variability. Specifically, we propose the use of distribution-reproducibility “SuperPlots” that display the distribution of the entire dataset, and report statistics (such as means, error bars, and $P$ values) that address the reproducibility of the findings.

This paper is not an overview of statistical analyses; many have described the troubles with $P$ values (Gardner and Altman, 1986; Sullivan and Feinn, 2012; Greenland et al., 2016) and there are several excellent practical guides to statistics for cell biologists (Lamb et al., 2008; Pollard et al., 2019). In this paper we specifically address ways to communicate

---

1 We freely admit that our past selves are not innocent of the mistakes described in this manuscript.
reproducibility when performing statistical tests and plotting data.

**Biological Replicates vs Technical Replicates**

Before we dig into why the total number of observed cells is usually the wrong number to use as the sample size, we should clarify what we mean by experimental replication. A biological replicate is an independent, repeated experiment; technical replicates are repeated measurements of the same sample. For example, if I use a ruler and measure the length of 200 hairs on my head, I would have many technical replicates and a precise estimate of my own hair length; measuring the hair from 200 different people would give me many biological replicates and a more accurate estimate of the average human hair length. In other words, biological replicates address the reproducibility of an experimental result, and technical replicates can explore the distribution within each sample. An individual cell measurement is typically closer to a technical replicate (Lazic, 2010; Lazic et al., 2018).

Deciding what makes for a good biological replicate can be challenging (Aarts et al., 2014; Blainey et al., 2014; Naegle et al., 2015). For example, is it acceptable to run multiple experiments from just one thawed aliquot of cells, or do I need to borrow an aliquot from another lab? Is it necessary to generate multiple knockout strains? Is it sufficient to test in one cell line, or do I need to use multiple cell types or even cells from multiple species? Can I perform all experiments in one lab, or should I include results from a lab on the other side of the country (Lithgow et al., 2017)? There’s no single right answer: each researcher must balance practicality with robust experimental design. At a minimum, researchers must perform an experiment multiple times if they want to know whether the results are robust.²

**What Hypothesis is Being Tested?**

Here’s a simple question to help you clarify what your sample size N should be (Naegle et al., 2015; Lazic et al., 2018; Pollard et al., 2019): What population are you trying to sample? A P value is often used to indicate that a measured difference between populations (treatments, strains, knockdowns, etc.) is “significant,” or unlikely to be due only to random chance.³ What those two populations are depends on how you sample them. For example, if you want to know if a particular treatment changes the speed of crawling cells, you could split a flask of cells into two wells of a 96-well plate, dose one well with a drug of interest and one with a placebo, and then track individual cells in each of the two wells. If you use each cell as a sample, the two populations you end up comparing are the cells in those two particular wells. By repeating the experiment multiple times from new flasks, and using each experiment as a sample, you evaluate the effect of the treatment on any arbitrary flask of cells. Multiple observations within one well increases the precision for estimating the mean for that one sample, but doesn’t reveal a truth about all cells in all wells (just like measuring many hairs on my own head doesn’t give me insight into the average haircut).

If you only care about cell-to-cell variability within a particular sample, then maybe N really is the number of cells you observed. Making inferences beyond that sample, however, would be questionable, because the natural variability of individual cells can be overshadowed by systematic differences between biological replicates. Whether caused by passage number, confluency, or location in the incubator, cells often vary from flask-to-flask and day-to-day. Entire flasks of cells might even be described as “unhappy.” Accordingly, cells from experimental and control samples (e.g. tubes, flasks, wells, or coverslips) may differ from each other, regardless of the intended experimental treatment. When authors report the sample size as the number of cells, the resulting statistical analysis cannot help the reader evaluate whether any observed differences are due to the intended treatment or simple sample-to-sample variability. We are not prescribing any specific definition of N, we are simply encouraging researchers to consider what main source of variability they hope to overcome when designing experiments and statistical analyses (Altman and Krzywinski, 2015) (see Table 1).

² This raises the question of how many cells one should look at in each sample. Is it better to look at many cells in a few biological replicates or spend less time measuring individual cells and redirect that effort to repeating the experiment additional times? Multiple analyses have found that increasing the number of biological replicates has a larger influence on the statistical power than imaging many more cells in each sample (Aarts et al., 2014; Blainey et al., 2014).

³ Despite warnings from statisticians (Greenland et al., 2016).
To test the hypothesis that two treatments or populations are different, the treatment must be applied or the populations sampled multiple times. In a drug trial for a new painkiller pill, one of my knees cannot be randomly assigned to the treatment group and the other to placebo, so researchers cannot count each of my knees as a separate N. (If the trial is testing steroid injections, then under certain statistical models, each knee could be considered a separate sample [Aarts et al., 2014].) Similarly, neighboring cells within one flask or well treated with a drug are not separate tests of the hypothesis, because the treatment was only applied once. But if individual cells are microinjected with a drug or otherwise randomly assigned to a different treatment, then each cell really can be a separate test of a hypothesis.

How to Calculate P Values from Cell-Level Observations

Cell biologists often observe hundreds of cells per experiment and repeat an experiment multiple times. To leverage that work into robust statistics, one needs to take into account the hierarchy of the data. Simply pooling all the cell-level data from multiple biological replicates squanders the available information about run-to-run variability (Figure 1). There is ample literature about the analysis of this type of hierarchical data (Gelman and Hill, 2006; Galbraith et al., 2010), which takes into account both the variance within a sample as well as the clustering across multiple experimental runs (Aarts et al., 2014), or that propagate the error up the chain, such as a nested ANOVA (Krzywinski et al., 2014).

A simple approach—which permits conventional t-test or ANOVA calculations—is to average the cell-level data for each biological replicate and compare the subsequent sample-level means (Altman and Bland, 1997; Galbraith et al., 2010; Lazic, 2010). For example, if you have three biological replicates of control and treated samples, and you measure the cell diameter of 200 cells in each sample, first calculate the mean of those 200 measurements for each sample, then run a t-test on those sample means (three control, three treated). This simple approach might fail to detect small but real differences between groups, where more advanced techniques may prove to be more powerful (Galbraith et al., 2010; Aarts et al., 2014).

Communicating Variability with SuperPlots

After analyzing hundreds of cells across multiple rounds of experimentation (and calculating useful P values), it would be nice to incorporate both the cell-level variability and experimental repeatability of all that work into a single diagram. Figure 1 shows plots that can be improved with correct thinking about biological replicates.

If we summarized all hair styles with a single average hair length, it would not describe asymmetrical “hipster” haircuts nor capture the large range of hair lengths. Similarly, bar graphs are problematic because they obscure the distribution of cell-level data as well as the sample-to-sample repeatability (Weissgerber et al., 2015). Although it’s tempting to place the blame on the use of bar graphs, the problem is deeper. In Figure 1, the plots on the left have small error bars and comically tiny P values, which should raise red flags given how difficult it would be to replicate a cell biology experiment with identical results and/or to repeat it hundreds of times, which such miniscule P values imply. While beeswarm, column scatter, box-and-whisker, and violin plots are great at conveying information about the range and distribution of the underlying data, plotting the entire dataset does not make it appropriate to treat repeated measurements on the same sample as separate experiments.

Therefore, we suggest authors incorporate information from technical and biological replicates by creating “SuperPlots,” which show distribution and reproducibility by superimposing summary statistics on a graph of the entire dataset (Figure 1, right columns). Not only does a SuperPlot convey more information than a conventional bar graph or beeswarm plot, it also makes it clear that statistical analyses (e.g. error bars and P values) are correctly calculated across biological replicates, not individual cells—even when each cell is represented on the plot. For example, the mean from each experiment could be listed in the caption, or plotted as a larger dot on top of the many smaller dots that denote individual measurements.

When possible, it is best to link samples by run, for instance, by color-coding the dots by experiment or drawing a line linking paired measurements together (Figure S1D). The benefit of these linkages is

---

4 In this metaphor, a bar graph would imply that everyone has a bowl cut.
to convey the repeatability of the work: readers learn more if they know that one experiment exhibited high readings across the board than if they have to guess the trend in each sample (Figure 1 B vs C, right columns). An additional benefit to linking data is that it eliminates the need to normalize data in order to directly compare different experimental runs. Often, multiple experiments might all exhibit the same trend, but different absolute values (Figure 1B). By encoding the biological replicate into the data, such trends can be revealed without normalizing to a control group. P values can then be calculated using statistical tests that take into account linkages among samples (e.g. a paired or ratio t test).5

An impressive amount of information can be crammed into color-coded beeswarm SuperPlots (see Figure 1, rightmost plots; Figure 2), where each cell-level datapoint divulges which biological replicate it came from (Galbraith et al., 2010; Weissgerber et al., 2017). This helps convey to the reader whether each experimental round gave similar results or if one run biases the conclusion (Figure 1C). The summary statistics and P values in beeswarm SuperPlots (calculated using the means from each biological replicate, not the value from each individual cell) are overlaid on the color-coded scatter.

Error Bars that Communicate Reproducibility

The choice of error bars on a SuperPlot depends on what you hope to communicate: descriptive error bars characterize the distribution of measurements (e.g. standard deviation), while inferential error bars evaluate how likely it is that the same result would occur if the experiment were to be repeated (e.g. standard error of the mean or confidence intervals) (Cumming et al., 2007). To convey how repeatable an experiment is, it is appropriate to choose inferential error bars calculated using the number of biological replicates as the sample size. However, calculating standard error of the mean by inputting data from all cells individually fails in two ways: first, the natural variability we expect from biology would be better summarized with a descriptive measure, like standard deviation; and second, the inflated N produces error bars are artificially small (because the calculation for standard error of the mean includes \( \sqrt{N} \) in the denominator) and communicate nothing about the repeatability of the experiment. Just like measuring my own hair many times gives me a precise estimate of my own hair length, but cannot be used to infer the average haircut.

The problems that arise from calculating error bars using cell number as the sample size are illustrated by comparing the left-hand panels of Figure 1A–B: when each cell measurement is erroneously treated as an independent replicate, the standard error of the mean is equally tiny in plots A and B, despite high variability between experimental replicates in B. In contrast, the SuperPlots in the right-hand columns show error bars that were calculated using biological replicates. Note that the standard error of the mean grows when the results vary day-to-day (Figure 1B, right columns). In cases where displaying every data point is not practical, authors should consider some way of representing the cell-to-cell variability as well as the run-to-run repeatability. This could be error bars that represent the standard deviation of the entire dataset, but with P values calculated from biological replicates.

However an author chooses to display their data (Figure S1), it is critical to list the number of biological replicates (or “independent experiments”) in the figure or caption. For SuperPlots, the caption should explain that the means and the statistical tests (e.g. t-tests, ANOVAs, and post tests) were calculated on the average observed value for each biological replicate.

Conclusion

Despite our emphasis on biological replicates, a scientist can gather a lot of information by observing a single cell. In fact, some of the greatest discoveries had an N of one (Marshall et al., 1985). The field will squander opportunities for conceptual breakthroughs if it leaves no room for fleeting observations and singular results (Abercrombie et al., 1970; Tilney and Portnoy, 1989). Such striking isolated observations, however, should not be garnished with meaningless P values.

On a practical note, if you calculate a P value smaller than Planck’s constant, take a moment to consider: What variability does your P value represent? How many independent experiments have you performed, and does this match with your N? (See Table 1 for practical examples of this analysis.) Reviewers

5 In fact, not taking into account linkages can make the t-test too conservative, yielding false negatives (Galbraith et al., 2010).
and editors should gently remind authors to state the $N$ in terms of biological replicates and calculate $P$ values based on that, rather than the total number of cells observed in each sample. We encourage authors and editors to focus not on reporting satisfying yet superficial statistical tests such as $P$ values, but instead on presenting the data in a manner that conveys both the variability and the reproducibility of the work.

Bibliography

Aarts, E., Verhage, M., Veenhuijzen, J. V., Dolan, C. V., and van der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. Nat. Neurosci. 17, 491–496.

Abercrombie, M., Heaysman, J. E., and Pegrum, S. M. (1970). The locomotion of fibroblasts in culture. II. "Ruffling". Exp. Cell Res. 60, 437–444.

Altman, D. G., and Bland, J. M. (1997). Statistics notes. Units of analysis. BMJ 314, 1874.

Altman, N., and Krzywinski, M. (2015). Points of significance: Sources of variation. Nat. Methods 12, 5–6.

Blainey, P., Krzywinski, M., and Altman, N. (2014). Points of significance: replication. Nat. Methods 11, 879–880.

Cumming, G., Fidler, F., and Vaux, D. L. (2007). Error bars in experimental biology. J. Cell Biol. 177, 7–11.

Galbraith, S., Daniel, J. A., and Vissel, B. (2010). A study of clustered data and approaches to its analysis. J. Neurosci. 30, 10601–10608.

Gardner, M. J., and Altman, D. G. (1986). Confidence intervals rather than $P$ values: estimation rather than hypothesis testing. Br. Med. J. (Clin. Res. Ed) 292, 746–750.

Gelman, A., and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge: Cambridge University Press.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, $P$ values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31, 337–350.

Krzywinski, M., Altman, N., and Blainey, P. (2014). Points of significance: nested designs. For studies with hierarchical noise sources, use a nested analysis of variance approach. Nat. Methods 11, 977–978.

Lamb, T. J., Graham, A. L., and Petrie, A. (2008). T testing the immune system. Immunity 28, 288–292.

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci. 11, 5.

Lazic, S. E., Clarke-Williams, C. J., and Munafò, M. R. (2018). What exactly is "N" in cell culture and animal experiments? PLoS Biol. 16, e2005282.

Lithgow, G. J., Driscoll, M., and Phillips, P. (2017). A long journey to reproducible results. Nature 548, 387–388.

Marshall, B. J., Armstrong, J. A., McGeachie, D. B., and Glancy, R. J. (1985). Attempt to fulfil Koch’s postulates for pyloric Campylobacter. Med. J. Aust. 142, 436–439.

Naegle, K., Gough, N. R., and Yaffe, M. B. (2015). Criteria for biological reproducibility: what does "n" mean? Sci. Signal. 8, fs7.

Pollard, D. A., Pollard, T. D., and Pollard, K. S. (2019). Empowering statistical methods for cellular and molecular biologists. Mol. Biol. Cell 30, 1359–1368.

Sullivan, G. M., and Feinn, R. (2012). Using Effect Size—or Why the $P$ Value Is Not Enough. J Grad Med Educ 4, 279–282.

Tilney, L. G., and Portnoy, D. A. (1989). Actin filaments and the growth, movement, and spread of the intracellular bacterial parasite, Listeria monocytogenes. J. Cell Biol. 109, 1597–1608.

Velle, K. B., and Campellone, K. G. (2018). Enteropathogenic E. coli relies on collaboration between the formin mDia1 and the Arp2/3 complex for actin pedestal biogenesis and maintenance. PLoS Pathog. 14, e1007485.

Weissgerber, T. L., Milic, N. M., Winham, S. J., and Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biol. 13, e1002128.

Weissgerber, T. L., Savic, M., Winham, S. J., Stanisavljevic, D., Garovic, V. D., and Milic, N. M. (2017). Data visualization, bar naked: A free tool for creating interactive graphics. J. Biol. Chem. 292, 20592–20598.
Figure 1. The importance of displaying reproducibility. The problematic figures in the “Nope” and “Still Nope” columns on the left treat \( N \) as the number of cells, resulting in tiny error bars and comically small \( P \) values. Also, these plots conceal any systematic error run-to-run, convolving it with cell-to-cell variability. To correct that, in the “Good” and “Even Better” columns, “SuperPlots” superimpose summary statistics from independent experiments (AKA biological replicates) on top of data from all cells, and \( P \) values were correctly calculated using \( N \) of three, not 300. In this case, the cell-level values were first pooled for each biological replicate and the mean calculated of each pool; those three means were then used to calculate the average (horizontal bar), standard error of the mean (error bars), and \( P \) value. Furthermore, each biological replicate is color-coded: the averages from one experimental run are yellow dots, another independent experiment is represented by gray triangles, and a third experiment is shown as blue squares. This helps convey that the trend is observed within each experimental run, as well as for the dataset as a whole. The beeswarm SuperPlots in the rightmost column represent each cell with a dot that is color coded according to the biological replicate it came from. Data in A, B, and C have similar ultimate averages, but more variability from run-to-run in B and C. Specifically, in B the measured speeds from the gray experiment are consistently lower and those from the blue are higher, presumably because of air temperature, passage number, media age, or some
other unanticipated variable. C is the same cell-level data, but imagines a case where the run-to-run variability is random. Note that the plots that treat each cell as its own *N* (left columns) utterly fail to distinguish the three scenarios, claiming a significant difference after drug treatment, even when that is the result of random fluctuations. The *P* values represent an unpaired two-tailed t test for the left columns and a paired two-tailed t test for the right columns.
Figure 2: Tutorial for Making SuperPlots. We describe how to make SuperPlots in GraphPad Prism 8 graphing software (also see Supplemental Movie 1). If using other graphing software, one may create a separate, different colored plot for each replicate, then overlay those plots in software like Adobe Illustrator. (A) When adding data to the table, leave a blank row between replicates. (B) Create a new graph of this existing
data; under type of graph select “Column” and “Individual values,” and select “No line or error bar.” (C) After formatting the universal features of plot from B (e.g. symbol size, font, axes), go back to the data table and highlight the data values that correspond to one of the replicates. Under the “Change” menu, select “Format Points” and change the color, shape, etc. of the subset of points that correspond to that replicate. (D) Repeat for the other replicates to produce a graph with each trial color coded. (E-F) To display summary statistics, take the average of the technical replicates in each biological replicate (so you will have one value for each condition from each biological replicate), and enter those averages into another data table and graph. Use this data sheet that contains only the averages to run statistical tests. (G) To make a plot that combines the full dataset with the correct summary statistics, format this graph and overlay it with the above scatter SuperPlots (In Prism, this can be done on a “Layout.”). This process could be tweaking to display other overlayed, color-coded plots (e.g. violin).
Supplemental Figure 1: Other Plotting Examples. Bar plots can be enhanced even without using beeswarm plots. (A) The bar plot that calculates $P$ and error bars using the number of cells as $N$. (B) But a bar graph can be corrected by plotting the mean of the replicates. (C) Showing each replicate reveals more than a simple bar graph. (D-E) Linking each pair by the replicate conveys important information about the trend in each experiment. (F) A SuperPlot shows not only information about each replicate and the trends, but also superimposes the distribution of the cell-level data, here using a violin plot.
Table 1. How the choice of $N$ influences conclusions

| Experiment | If $N =$ number of observations | If $N =$ number of experiments | Potential Outcomes & problems |
|------------|---------------------------------|--------------------------------|------------------------------|
| A researcher measures the speeds of 20 crawling cells per condition, and repeats the experiment on 3 different days. | | | Here, the researcher would come to a different conclusion based on what they consider “$N$.” In this example, data were collected on cells prior to treatment, so they are all untreated. Therefore, differences are due to unpredictable sample-to-sample fluctuations. This example highlights that these chance differences are amplified when each cell is considered its own experiment. When comparing actual treatments, using $N = 60$ could lead to erroneously small $P$-values. |
| $N =$ 60 cells Test: ANOVA + Tukey post test $P =$ 0.044 0.029 | $N =$ 3 experiments Test: ANOVA + Tukey post test $P =$ 0.49 0.38 | | |

| A researcher measures the pixel intensity of actin staining in cell protrusions. The researcher measures 10 protrusions per cell for 5 cells, and repeats the experiment 3 times. | | | In this example, there are three choices for $N$: the number of protrusions, the number of cells, or the number of experiments. In this case (Velle and Campellone, 2018), the $P$-value is less than 0.05 regardless of which $N$ is used for statistical analysis. This example illustrates that while a miniscule $P$-value could raise red flags, the underlying conclusions may still be appropriate. While using $N = 150$ may seem preposterous, it’s understandable that a researcher measuring >1,800 protrusions by hand would begin to think of each cell as a separate experiment. |
| $N =$ 150 protrusions Test: unpaired t-test $P =$ 8.9 x $10^{-33}$ | $N =$ 3 experiments Test: paired t-test $P =$ 0.041 | | |

| Speed (μm/min) | | |
|----------------|-------------------|
| Treatment A | Treatment B | Treatment C |

| Relative Fluorescence Intensity | | |
|-------------------------------|-------------------|
| Control | Drug |

| Relative Fluorescence Intensity | | |
|-------------------------------|-------------------|
| Control | Drug |
| Test Scenario | Description |
|---------------|-------------|
| **Researchers compare nucleus size within tumors vs the surrounding cells.** | They look at 100 transformed and normal cells in 5 different tissue samples. They compare the nucleus size in transformed and normal cells, paired by tissue sample. Now they have five biological replicates, each encompassing 100 technical replicates. They perform a paired test and get a more reasonable *P* value. By using an *N* of 500, it is likely that the *P* value will be artificially smaller. By averaging the technical replicates and using an *N* of 5, the researchers can confirm if the observed difference across patient samples is larger than the natural variation within any single tissue sample. |
| **In an in vitro experiment, researchers compare the rates of filament growth of actin orthologs.** | They measure the growth rate of thousands of filaments from each organism. They then calculate the statistics using the number of filaments as *N*, resulting in a tiny *P* value. Counting each filament as a separate sample provides an estimate of the inherent variability of filament growth rate within that one sample. But using that to then compare two orthologs on different coverslips is not appropriate, because there is nothing controlling for different handling of the two proteins during imaging. Repeating the experiment with different stocks of protein takes into account random and systematic errors including con- |
Researchers use laser ablation to cut individual spindle fibers, then observe the downstream effects in those same individual cells. In other cells in the same sample, they apply the same laser dose, but not directed at spindle fibers.

| N = 30 |
|--------|

They perform the treatment or control on 30 different cells, approaching each cell as an $N$ and calculating a $P$ value to compare the two treatments.

Because each cell can be randomly assigned to the control or treated group, it is appropriate in this case that cell is its own sample and calculate a $P$ value from one day’s experiment using $N$ as the number of cells. The researchers may conclude that cells treated with laser ablation differ significantly from those that don’t have their spindle fibers cut.

But that may only be true for that cell strain at that particular passage number and at a specific temperature of the scope room. Observing similar results over multiple days or even with different cell types makes this claim robust.