Basic Statistics in Cell Biology

David L. Vaux

The Walter and Eliza Hall Institute, Parkville, Victoria 3052, Australia

Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia; email: vaux@wehi.edu.au

Annu. Rev. Cell Dev. Biol. 2014. 30:23–37

First published online as a Review in Advance on July 2, 2014

The *Annual Review of Cell and Developmental Biology* is online at cellbio.annualreviews.org

This article's doi: 10.1146/annurev-cellbio-100913-013303

Keywords

p-value, standard deviation, standard error, replicate, error bar

Abstract

The physicist Ernest Rutherford said, "If your experiment needs statistics, you ought to have done a better experiment." Although this aphorism remains true for much of today's research in cell biology, a basic understanding of statistics can be useful to cell biologists to help in monitoring the conduct of their experiments, in interpreting the results, in presenting them in publications, and when critically evaluating research by others. However, training in statistics is often focused on the sophisticated needs of clinical researchers, psychologists, and epidemiologists, whose conclusions depend wholly on statistics, rather than the practical needs of cell biologists, whose experiments often provide evidence that is not statistical in nature. This review describes some of the basic statistical principles that may be of use to experimental biologists, but it does not cover the sophisticated statistics needed for papers that contain evidence of no other kind.

Contents

INTRODUCTION

Science is knowledge gained through repeated observation or experimentation. By convention, when scientific research reveals some new knowledge, it is published in a journal as an article that not only contains the conclusions but also includes the evidence that supports them.

We publish papers to publicize this new knowledge, to make a claim to its discovery, and to convince the readers of the paper that it is true. The evidence supporting the conclusions might be in the form of tables of data, graphs, or images.

In 2005, the statistician John Iaonnidis (2005) published a paper proving that most claimed research findings are false. His paper focused mainly on medical research papers that formed conclusions solely on the basis of results that were statistically significant as assessed by a *p*-value of less than 0.05.

Although Iaonnidis's claim was often dismissed as an exaggeration, in 2011 and 2012, papers from pharmaceutical companies Bayer and Amgen reported that they were unable to replicate findings reported by academic labs 65% and 89% of the time (Begley & Ellis 2012, Prinz et al. 2011).

These numbers are shockingly high, but in the absence of evidence to the contrary, they cannot be dismissed. As roughly one million new publications are listed in PubMed each year, whether Iaonnidis is right, and the conclusions of more than 500,000 are false, or the pharmaceutical companies are right, and as many as 900,000 are irreproducible, there is clearly something going wrong with how biomedical research is conducted. A large part of this problem may be due to misuse of statistics. Even though many papers in experimental biology do not involve sophisticated statistics, an understanding of basic statistics may help experimenters reduce the number of papers with false conclusions, help readers identify those papers that are sound, and help interested colleagues understand the reasons for reproducibility or lack of reproducibility in a given study.

THREE USES FOR STATISTICS IN CELL BIOLOGY

Just like a microscope, a computer, or a flow cytometer, statistics is a tool that can be used to provide or process information that can be used in generating scientific knowledge. In experimental cell biology, statistics can be used (*a*) before results are published to monitor the performance of an experiment, (*b*) to help analyze and draw inferences from the data, and (*c*) in publications to help readers understand and interpret the experiments.

Although all experiments must be repeated, or include multiple independent observations, in experimental cell biology much of the evidence can be nonstatistical in nature. For example, some experimental results involving Western blots, histology, coimmunoprecipitations, PCR, flow cytometry, or electron microscopy can be convincing without the need for statistical analysis. However, with advances in computers and big-data technologies, such as proteomics and next-generation sequencing, the impact of statistics in cell biology is increasing.

Cell biology papers typically contain multiple figures that each show the results of several experiments, and these provide mechanistically different lines of evidence supporting the conclusions. In this way, cell biology papers are distinct from papers in epidemiology, psychology, or clinical studies, which often describe the results of a single experiment and do not pick apart the underlying biological mechanisms. Although this second group of papers depends entirely on statistical evidence, most papers in experimental cell biology and molecular biology do not. Nevertheless, an understanding of basic statistics can be useful to cell biologists when they are conducting experiments, analyzing the results, and critically evaluating the results in papers published by others.

Statistics to Monitor the Conduct of Experiments—Descriptive Statistics

The results of experiments in cell biology will vary not only because biological systems are complex but also owing to the limits of the measuring instruments, sampling error, and observer error. In addition, there may be variation in the results that may reflect accidents such as mislabeling or equipment failure. Basic descriptive statistics can be used to help monitor experiments to alert cell biologists to problems with their performance. The cell biologist must be able to distinguish expected variation, such as that caused by sampling error or biological variation, from spurious results owing to equipment failure or errors of data entry.

If, for example, you were performing a colony assay using bone marrow cells cultured in soft agar, you might set up triplicate plates for each of the conditions. As triplicates are intended to be as similar to each other as you can make them, you would expect each of the triplicate plates to grow a similar number of colonies. If one of the cultures in a triplicate had ten times more or ten times less colonies than the other two, you would be prompted to find out why. Perhaps the bone marrow cells were not dispersed properly, and one of the triplicate cultures got a clump of cells; perhaps the colonies are not bone marrow–derived white blood cells but are contaminating yeast colonies; perhaps there was a blockage in the pipette when the cells were dispensed.

In some cases, it will be obvious that something has gone wrong, and a little investigation will reveal the cause. At other times, it might be unclear whether the variation between the replicates is within normal expectations. To make this judgment requires an understanding of the sources of variation in the data and how the results are likely to be distributed.

Descriptive statistics can be used to show how broadly or narrowly the data are distributed and can make it easier to discriminate between the expected amount of variation and something going wrong during an experiment. The most commonly used descriptive statistics are the range and the standard deviation (SD).

Range. The range is the interval from the lowest sample value to the highest sample value (**Figure 1**). If you were measuring the white blood cell counts from several identically treated mice, or doing replicate counts on a single sample, the range would show how the counts were distributed: If they were all very similar, the range would be narrow; if there was a great deal of variability, the range would be wide.

Although range is an easy descriptive statistic to understand and to calculate, it is not often used because it is not a robust descriptive statistic. Consider if you had the white blood cell counts from only two mice. Then the range would be the interval from one count to the other. If you did a count on one more mouse, and its white blood cell count was between the other two, the range interval would not change in size. However, if the white blood cell count was less than or more than the other two, the range would get bigger. The more and more mice you examined, the bigger and bigger the range interval would get. In other words, the range gets bigger as the number of samples increases.

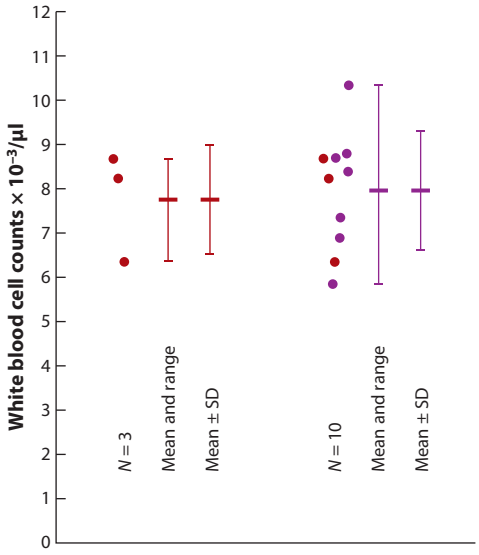


Figure 1

Range and standard deviation (SD) error bars can describe how the data are distributed. White blood cell counts (*red dots*) from three mice (*left*), showing the mean and range and mean ± SD. On the right are the first three samples, plus samples from a further seven mice (*purple dots*). As *N* increases, the range becomes larger. The sample SD does not change in a consistent way as *N* increases, but it becomes a more and more accurate estimate of the population SD.

Standard deviation. A much more commonly used descriptive statistic is the SD. SDs are calculated by taking the square root of the average of the squared differences between the samples and the sample mean, so they can be thought of as the typical difference between the sample values and the sample mean. It is a robust statistic because in a normal distribution, the mean ± the SD interval encompasses approximately two-thirds of the data points, regardless of how many samples there are. The SD of several samples is always the best estimate of the SD of the entire population, just as the mean of the samples is always the best estimate of the mean of the entire population.

The smaller the SD, the closer most of the sample values will be to the mean. Note, however, that unless you are measuring the whole population, the summed squares of the (sample − mean) values are divided by $N - 1$, rather than N, because this gives a more accurate estimate of the population SD. Also, when N is very small (e.g., only three), the sample SD tends to underestimate the SD of the population.

Although the mean and SD of the samples will always provide the best estimate of mean and SD of the population, your confidence in the accuracy of the estimate will be greater the more samples you have. If you have only two or three samples, the sample mean and SD could both be very, very, different from the population mean and population SD.

Sources of variation. Suppose you were performing a cell death assay, and you had a set of three replicate 4-mL cultures that, if you counted every cell in all three wells, had 1,000,000 cells, of which 500,000 were alive and 500,000 were dead. There would be some differences in the number of cells per well, as well as in the percentage of live and dead cells, owing to variability in pipetting accuracy, and there would also be slight differences owing to variation in the distribution of heat and humidity in the incubator. In addition, when you went to count the number of live and dead

cells, you would also see variation in the results because of sampling error, the variation caused by the random likelihood of a dead or a live cell entering the pipette.

Binomial distribution. If you were to analyze each of the wells by counting 100 cells from each in a haemocytometer, using trypan blue exclusion to determine the percentage of cell death, there would be some random variation in your measurements depending on whether a live or a dead cell happened to be sucked into the pipette tip, or whether it came to rest over the haemocytometer grid (**Figure 2**). Because each cell counted could be either alive or dead, the counts should fall into a binomial distribution. The expected variance of a binomial distribution is N (the number of cells counted; in this case, 100) $\times$ p (the fraction of cells that are alive; in this case 0.5) $\times$ $(1 - p)$ (the fraction of cells that are dead); i.e., 100 $\times$ 0.5 $\times$ 0.5, or 25. The SD is the square root of the variance, in this case, 5 (or 5% of 100 cells). Therefore, if you did counts on many wells, approximately two-thirds of the time your counts would be between 45% and 55% live cells (mean $\pm$ 1 SD), and approximately 95% of the time your counts would be between 40% and 60% live cells (mean $\pm$ 2 SD). This variability is termed sampling error, and it can be seen in **Figure 2**. If one of the counts showed zero live cells, you would be prompted to check to see if there was something wrong with that culture.

In **Figure 2**, you can also see variability in the sample SDs, which are expected to be 5%. Some of the SDs of the triplicates are smaller than this, and some are larger. In a real, rather than a simulated, experiment, you would expect the SDs to be even larger, because the variation in pipetting would be added to the variation owing to sampling. If the SD in a set of triplicate cultures were very large, you should try to determine why. However, if the SDs of all of the triplicates

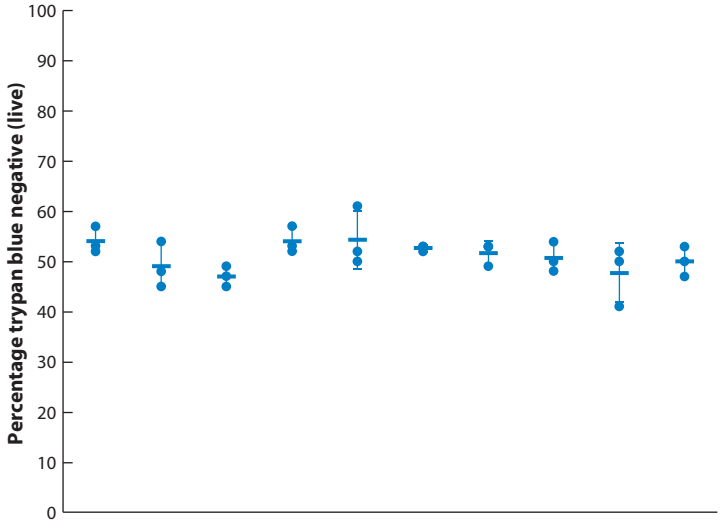


Figure 2

Percentages of live (trypan blue negative) cells from triplicate samples (*blue dots*) with the mean $\pm$ standard deviation (SD). In each of these 10 sets of triplicates, 100 cells were sampled from the same stock suspension of cells that were 50% viable. Note how the sample mean and standard deviation vary. The expected mean is 50%, and the expected SD is 5%. (These data were created by performing a random coin toss 100 times for each point. In a real experiment, the expected SDs would be larger, owing to variation in, for example, pipetting accuracy or observer error, in addition to the sampling error.)

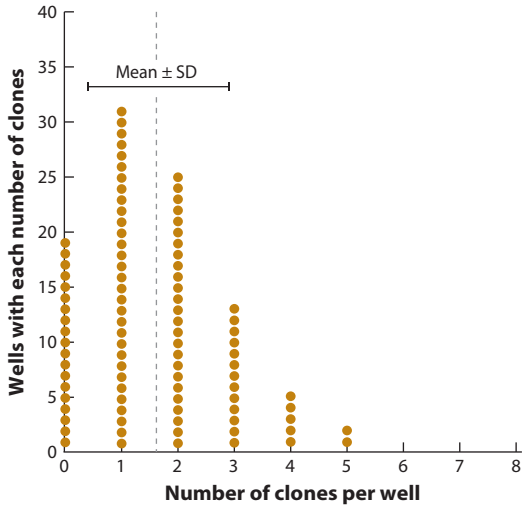


Figure 3

Number of hybridoma clones per well in a 96-well plate. In this case, 19 wells grew no hybridomas, and no well had more than 5 hybridomas; the average number of hybridomas per well was 1.6, and the SD was 1.2. The number of hybridomas per well should fit a Poisson distribution.

were the same, or all were much smaller than 5%, you should question those who performed the experiment and who provided you with the results, as they might be biased.

If you did the same experiment, but this time analyzed cell viability by measuring uptake of propidium iodide on a flow cytometer, counting 10,000 cells from each well, the expected mean viability would still be 50%, but the SD of the three wells would be expected to be $\sqrt{[N \times p \times (1-p)]} = \sqrt{(10{,}000 \times 0.5 \times 0.5)} = \pm 50$ cells, or $50\% \pm 0.5\%$ viable cells. By sampling 100 times more cells, the SDs of the triplicate samples should be ten times smaller.

Poisson distribution. Another discrete distribution that frequently arises in cell biology is the Poisson distribution, the distribution of the number of cells in a microscope field, the number of colonies on a plate, or the number of hybridoma clones in each well at limit dilution. Poisson distributions involve integer numbers and do not have values less than zero, and large numbers can occur, at least in theory. Plotted out, a Poisson distribution looks like a binomial distribution that has been squashed up against the y axis and has a long tail to the right (see **Figure 3**). For example, if you performed a colony-forming assay by plating out bone marrow cells in soft agar in the presence of CSFs, some plates might grow no colonies and some might have one or two, or very many. A fact worth remembering about Poisson distributions is that the expected SD is the square root of the mean. If someone in your lab presented results showing numbers of cells per microscope field, and all the fields had one or two cells but none had zero or three, the SD would appear too small for a Poisson distribution, and the results might be biased. You should ask more questions about how the experiment was done.

Normal distribution. When p is close to 0.5, as N increases, a binomial distribution adopts a bell shape that approximates a normal distribution. Unlike the binomial distribution, which is used for discrete values, the normal distribution is used for continuous variables, e.g., for the weights of mice, or the length of their tails. In a normal distribution, the top of the bell is the mean, and the SDs are the points of inflection on the sides of the curve (**Figure 4**).

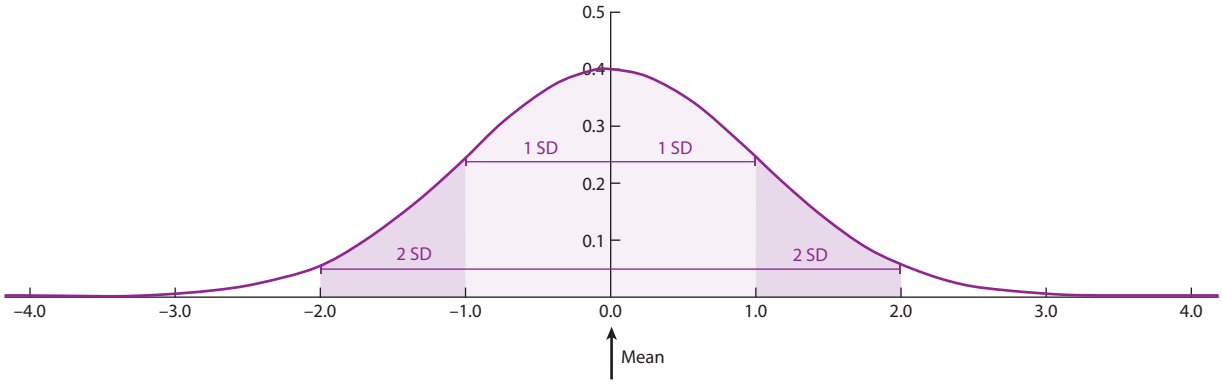


Figure 4

The normal distribution curve. The mean is adjusted to zero, and the standard deviation (SD) = 1. The tails extend to infinity on both the left and the right. The interval mean ± 1 SD covers approximately two-thirds of the data (*light purple*), and the interval mean ± 2 SD covers ~95% (*darker purple*). According to the central limit theorem, because many biological parameters are the result of a large number of variables, they very often fall on a normal distribution curve.

Using Statistics to Make Inferences

Once cell biologists are satisfied that they have performed the experiment well, and once they are convinced that the results are reliable, they can get to the next step of analyzing the data, to draw inferences from it, and to test the hypotheses they are interested in. Often they want to answer questions such as, does this cell, gene, or molecule do something? Can I be sure that what it appears to be doing is real, and not a spurious effect owing to coincidence? If it does have an effect, what is the size of the effect, and what is its biological importance?

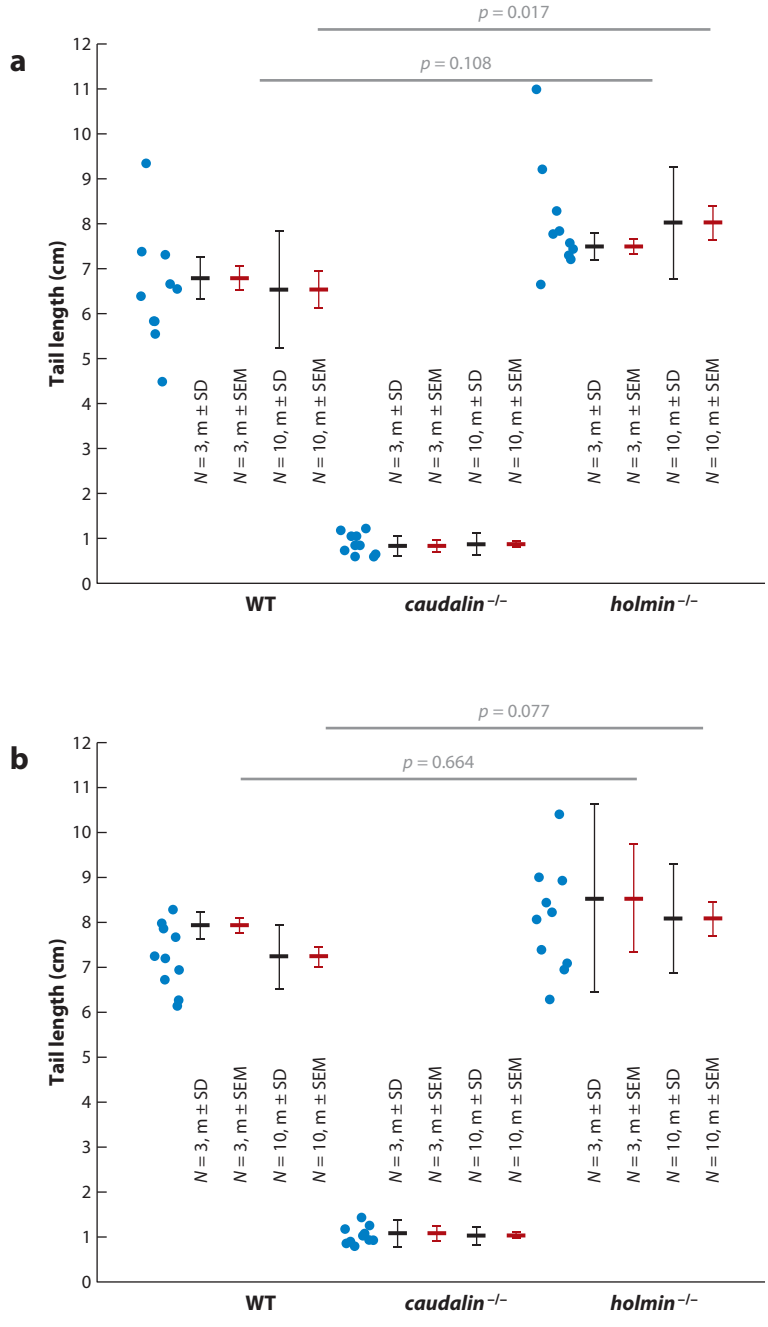
Sometimes the answers to these questions are obvious. For example, if you were trying to determine whether a mouse was transmitting a transgene that expressed green fluorescent protein, and you had a picture of a litter of eight mice from a normal and a green fluorescent protein transgenic parent, and three of them glowed bright green under ultraviolet light, you could conclude, with a very high degree of confidence, that the transgene had been transmitted, without doing any statistical tests at all. If you still had some doubts, you could take tissues from one of the green and one of the nongreen littermates, extract the DNA, and look for the sequence of the transgene, or you could extract the protein and perform a Western blot.

Where results are not obvious, inferential statistics can help you decide whether preliminary results are worth repeating or following up with confirmatory experiments. The most commonly used inferential statistics are standard error of the mean (SEM) and confidence intervals (CIs).

Suppose you were trying to see whether a gene (*caudalin*) had a role in determining tail size in mice. You have ten wild-type (WT) mice and ten *caudalin* homozygous gene knockout mice, all of the same sex and age. **Figure 5*a*** shows the lengths of the tails. In this case, there is a clear difference between the lengths of the tails where the only difference between the sets of mice is the presence or absence of the *caudalin* gene. It would be reasonable to infer that Caudalin was necessary for tail development, and there would be no need to do any statistical analysis (or show any in a published figure). These results would best be presented by plotting the data points, rather than showing the results as a bar graph, as this allows readers to readily understand how many mice were examined and see how much the tail lengths varied.

Now suppose you knocked out a different gene, *holmin*, and measured the lengths of the tails of ten male WT mice and ten *holmin* knockouts of the same age, and you got the results shown in

Figure 5*a*. Here it is not immediately obvious whether the differences in tail lengths between the WT and *holmin* knockouts could just be due to chance, or whether differences of this size would be unlikely if Holmin had no effect on tail length. To make this kind of judgment, you could use an inferential statistic, such as the SEM, or increase the number of mice you sampled, but a much better approach would be to carry out a different experiment, for example, crossing the *caudalin*$^{-/-}$



and *holmin*$^{-/-}$ mice to see if there is an epistatic effect, or measuring the size of other organs, or looking at the levels of Holmin protein in the *caudalin*$^{-/-}$ mice.

SEM. The SEM is the sample SD divided by the square root of *N*, the number of independent samples. If you took the same number of samples from a population multiple times and calculated multiple sample means, these means would fall in a distribution with its own SD. That is why SEM is sometimes referred to as the SD of the mean.

SEM is referred to as an inferential statistic because it can help you make inferences from the data. For example, if the data are normally distributed, you can infer that the population mean will be within the interval: sample mean $\pm$ 2 $\times$ SEM, approximately two-thirds of the time when *N* is 3, and approximately 95% of the time when *N* is 10 or greater. Note that this does not tell you where the population mean is, it only gives you an indication about where it is likely to be.

As the number of samples, *N*, increases, the SEMs get smaller (because they are the SD divided by the square root of *N*), and the interval, sample mean $\pm$ SEM, gets narrower and narrower, giving you a more and more precise estimate of where the population mean might lie. You can see this in **Figure 5** if you compare the size of the SEMs (*red*) for $N = 3$ and $N = 10$. As the sample size (*N*) increases, the sample SD will not change consistently but will converge on the population SD, whereas the SEM will tend to get smaller and smaller and give a more precise indication of where the population mean is likely to be.

SEMs can also be used to compare samples from two separate groups, e.g., control and drug-treated mice, or WT and gene-deleted mice, when trying to judge whether the drug or gene had some effect on the parameter being measured. Often, the researcher's first question is, Am I seeing an effect of the drug or gene deletion, or could the differences I have observed just be due to chance? Put another way, this question is, What would be the likelihood of seeing differences of the size I have observed if I forgot to administer the drug, or if the mice were mistyped and were all actually WT?

In this case, you are using the means and SEMs of the control and experimental samples to infer where the population means lie. If the estimated population means of the control and experimental groups lie in nonoverlapping intervals, you can infer that the samples are likely to come from two different populations. The more separation there is between the control samples' mean $\pm$ SEM

Figure 5

The effects of gene deletion on tail length in mice. (*a*) The lengths of the tails of 10 wild-type (WT), *caudalin*$^{-/-}$, and *holmin*$^{-/-}$ mice were measured (*blue dots*). The standard deviation (SD) (*black*) and standard error of the mean (SEM) (*red*) of the first three of these mice are shown ($N = 3$), and then to the right, the SD and SEM of all 10 mice ($N = 10$). *p*-values have been calculated comparing the WT and *holmin*$^{-/-}$ mice for $N = 3$ and $N = 10$, using an unpaired two-tailed t-test. If only the first three mice were considered, the differences between the tail lengths of the WT and *holmin*$^{-/-}$ mice would not be statistically significant ($p = 0.108$), but if all 10 mice were included, the differences would be statistically significant ($p = 0.017$). (*b*) The result of repeating the simulation using the same mean values and SDs as in part *a*. In this case, the differences in tail length between the WT and *holmin*$^{-/-}$ mice does not reach statistical significance, even when 10 mice of each type are considered. A lab that carried out the experiment shown in *a* might come to a different conclusion from a lab that carried out the experiment shown in *b*, even though they were both sampling from the same populations of mice. It is clear from both experiments *a* and *b* that whether groups of 10 or 3 mice are considered, deletion of *caudalin* does have an effect on mouse tail length, and this is clear from plotting the data points; statistics are not necessary to demonstrate this. Even if you did find that deletion of *holmin* had a statistically significant effect (as in part *a*), you should not stop there but perform further experiments that confirm the effect in an independent way, and then consider the size of the effect, as well as its biological importance.

interval and the experimental samples' mean ± SEM interval, the more confident you can be in inferring that the drug or gene deletion (or whatever other experimental condition you are testing) is having an effect.

Once you are confident that the differences you have seen are statistically significant (i.e., would be very unlikely to have arisen if you were sampling from the same population), then you should consider the size of the effect and whether it has any biological importance. You should also plan what sort of experiment to perform to confirm (or disprove) your conclusion in some other, independent way. The ability to confirm inferences by independent, mechanistically different experiments is what distinguishes experimental cell biology from other biomedical research fields, such as epidemiology or clinical research, which often rely wholly on statistical evidence.

When presenting your data (i.e., a figure showing the results of your initial experiment and, ideally, a figure giving mechanistically independent confirmatory evidence), it is best to just plot the data points and not show any error bars or statistics, especially if N is less than 10. If you do wish to draw conclusions from your data using statistical arguments, then it might be appropriate to show an inferential error bar, such as SEM or CI. You could, if you wished, show SDs, but this would require the reader to do the calculations to estimate statistical significance. Of course, the most important things are to mention what N is in the figure legend and, if you do show them, describe what the error bars are.

Confidence intervals. CIs are an inferential statistic closely related to SEMs, and they can be used to make inferences about data in much the same way. Formally, a 95% CI is defined as the interval that, if generated 100 times by taking groups of samples from a population, would be likely to encompass the population mean 95 times. CIs are calculated by multiplying the SEM by the t-value, a statistic used by Gosset when testing batches of beer at the Guinness Brewery. When N is only 3, t is approximately 4, and when N is 10 or more, t is approximately 2. So, as a rule of thumb, 95% CI intervals are roughly mean ± 4SEM (when $N = 3$) and mean ± 2SEM (when $N =$ 10 or greater). Although they are commonly used in papers in clinical and psychology journals, CIs rarely appear in cell biology papers, so they will not be discussed further here, but those who want further detail can find it in Cumming's (2012) book, *Understanding the New Statistics*.

***p*-values.** p-values were developed by Ronald Fisher (1928) to help distinguish statistically significant results (i.e., those that look promising enough to be worth verifying) from nonsignificant ones (i.e., those could plausibly have arisen if you were just testing groups of controls). The p stands for probability. It is the probability of obtaining the difference you observed (or larger) if you were actually sampling from a single population. Put another way, it is the probability of getting the result or larger if the null hypothesis were true (e.g., that the gene you altered or the drug you tested had no effect). p-values are widely misunderstood and misused (Goodman 2008); however, because cell biology research seldom relies wholly on statistical evidence, p-values are rarely critical to the overall conclusions in cell biology papers. In this sense, much of experimental cell and molecular biology remains like physics in Rutherford's day. Statistics (and p-values) have their main use in the design and conduct of experiments, before a manuscript is written, rather than appearing in publications to provide the evidence upon which the conclusions rest.

Suppose you had a suspension of cells that were 50% live and 50% dead (as in **Figure 2**). If you used trypan blue exclusion to measure the viability by counting multiple samples of 100 cells, you would expect, on average, to count the same number of live cells as dead cells. Of course, it would be very unlikely to always count exactly 50 live and 50 dead cells in each aliquot; you would sometimes count 52 live and 48 dead, and sometimes 49 live and 51 dead. In fact, as this is a binomial distribution, approximately two-thirds of the time your live cell counts would be between 45 and 55, and approximately 95% of the time your counts would be between 40 and 60.

Given that the cell suspension had exactly 50% live cells, the probability of any individual count having 40 or less live cells is ~2.5%.

If you had an identical cell suspension, except that you had added a cytotoxin that you were testing, you could also do viability counts on these cells. If you counted 100 dead cells and zero live cells, it would be reasonable to assume that the cytotoxin was active, because if it had no effect, the probability of getting 100 dead cells from a 50% live, 50% dead population of cells would be ~1 in 10^{-32}. What if you counted 57 dead cells and 43 live cells? Could you confidently conclude that the cytotoxin was active? If you assumed the null hypothesis, that is that the compound is inert, and the viability of the population was still 50%, the probability of getting a dead cell count of 57 or more is greater than 5%, so you could not confidently rule out the possibility that the cytotoxin was inactive.

When you are calculating *p*-values, you are doing much the same thing. You calculate the probability of getting the results you have obtained (or more extreme ones) if the null hypothesis were true, i.e., that you were drawing all your samples from the same population. Suppose you treated one group of mice with a drug and another group of mice with a placebo, and you observed a difference in the means of their responses. The *p*-value is the probability of observing a difference of the size you did or larger if you had given all of the mice the placebo.

In **Figure 5**, *p*-values have been calculated for the differences in average tail length between the WT and *holmin*$^{-/-}$ mice, for the first three mice and then for all ten mice of each type. For both **Figure 5*a*** and **5*b***, if only three mice were compared, the *p*-values would be greater than 0.05. You would interpret this to mean that with these data you could not reject the null hypothesis, as differences of this size could arise by chance even if the *holmin* gene had no effect on tail length. However, you could not conclude that the *holmin* gene did not have an effect on tail length either.

If you still wished to test your hunch that the *holmin* gene affected tail length, you could test it further by examining more mice, or by doing some other type of experiment. Now consider the *p*-values when groups of ten mice are examined. In **Figure 5*a***, the *p*-value is 0.017, indicating a significant difference. If you were to find this in your experiments, you should not stop there. Firstly, it is important to note the size of the effect. In this case, deleting the *holmin* gene apparently increased tail length by approximately 15%, and the biological importance of this finding, if any, must be considered. Secondly, in describing testing for statistical significance, Fisher (1926, p. 504) wrote, "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance." This implies that the properly designed experiment would have to be confirmed multiple times and almost always have given a *p*-value of <0.05. In other words, if you find a result that is statistically significant, you still have to confirm your observation. Statistically significant results from one experiment should be considered in the context of the results of other experiments, what has previously been published, and the plausibility of the proposed model. Even then, the conclusions should be confirmed by an additional experiment that verifies them in a mechanistically different way.

In experimental cell biology, readers are rarely interested in a single observation and usually want to know the size of the effect, as well as evidence of mechanism or causation. If you saw a statistically significant effect of a drug on an enzyme, for example, it would be more convincing to follow up with experiments determining the binding affinity of the drug or showing that a point mutant form of the enzyme that could not bind the drug was no longer inhibited.

Figure 5*b* shows the results of repeating the same simulation as in **Figure 5*a***, using the same means, SDs, and distributions. Unlike in **Figure 5*a***, in **5*b*** the difference in tail lengths between the

10 WT and 10 *holmin*$^{-/-}$ mice is not statistically significant ($p = 0.077$). Together, **Figure 5*a*** and ***b*** show that the 0.05 significance level is arbitrary, and the same experiment can give statistically significant results ($p < 0.05$) or nonsignificant results ($p > 0.05$). It also illustrates the folly of relying on tests for statistical significance when N is as small as 10, and that when N is only 3, basing conclusions on statistical grounds is reckless.

Statistics in Figures

When you are preparing figures from your data for publication, your goal should be to present them in such a way that the readers can make their own interpretations of the data, and if they were sufficient to convince you, they should be sufficient to convince them. It is therefore important to include as much of the data as possible and not to be biased by presenting only data that fits a preconceived idea. If you have repeated an experiment several times, you should include all the results, and certainly not just the results of a single representative experiment. Of course, for data that are not numerical, such as Western blots or photomicrographs, it is fine to show a representative image. However, if you wish to use them to make quantitative claims, e.g., you have performed densitometry on several Western blots, you must show all of the independent results.

Sometimes it can be difficult to determine whether data are replicate (and cannot be used to test your hypothesis) or whether they are independent repeats (and can be used to test your hypothesis) (Vaux et al. 2012). If the distinction is not immediately clear, two questions to keep in mind are the hypothesis you are testing and the population you are sampling from.

Suppose an archaeologist was trying to determine the age of two layers of sediment by radiometric dating. She might take one sample from the upper layer and one from the lower layer, divide each of them in three, and run the first set of three through the counter, followed by the second set. In this case, although she would have produced six counts in total, these were two sets of triplicate counts, and she performed only one comparison, so $N = 1$, and she could not generalize about each layer of sediment. However, the triplicate counts might be useful in showing how reliable the dating equipment was.

If, however, she took three independent samples from different parts of the upper layer and three independent samples from across the lower sediment layer, and she performed radiometric dating on them, then she would have information from two sets of three independent samples, and she could draw inferences that could be generalized across those sedimentary layers.

When you have small numbers of results to graph, it is better to simply plot the data points than to create bar graphs showing the means and error bars. **Figure 6** illustrates some different ways data could be presented. You should choose the way that you think is easiest to understand.

Looking critically at others' papers. When looking at the figures in a paper, it should be clear from the figure and the legend what experiment was performed. Except for nonquantitative representative results (such as photomicrographs, blots, or flow cytometry plots), the number of independent samples (N) should be stated. If error bars are shown, the legend should describe what they represent. If the figures in a paper do not include such essential information, move on to another paper. If the authors show p-values or error bars or draw inferences from replicates, rather than independent data, be highly skeptical.

Make your own estimate of statistical significance. Even if p-values are provided, try to judge from the data or the error bars whether the differences that the authors claim to be statistically significant are indeed so. To estimate statistical significance, you need inferential error bars, so if a figure shows SDs, convert them into SEMs by dividing by the square root of N, the number of independent samples.

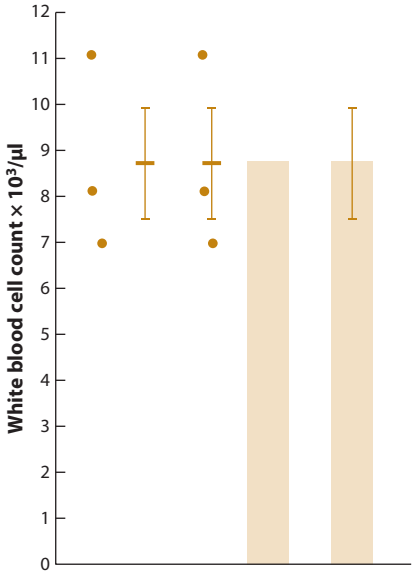


Figure 6

Three white blood cell (WBC) counts can be graphed in several ways. The clearest and simplest way is just to plot the data points. Many people plot the mean ± standard error of the mean (SEM), but this loses information and requires reading the legend to see if the error bars are SEM, SD, or something else. The middle option shows both the data points and the mean ± SEM. The fourth option, a column, shows what the mean WBC count was but provides no further information. The column with SEM error bars on the right provides more information, but it is still less clear than if the actual values were plotted. The figure legend should state whether the WBC counts are independent samples or replicates. If they are replicate counts, statistics and error bars should not be shown, because readers might wrongly assume the size of the error bars was relevant to the hypothesis being tested, rather than merely reflecting the reproducibility of the pipetting and counting. If the samples are independent, and N is only 3, I recommend simply plotting the three data points, as on the left.

As a rule of thumb, if the lower SEM bar from one set of samples overlaps the upper SEM bar from another set of samples, p is >0.05, and the differences cannot be statistically significant. If there is a gap separating twice the lower SEM bar from one set of samples (2SEM) from twice the upper SEM bar from the other set of samples, so that the gap between the means is greater than 4SEM, p is <0.05, and the differences must be statistically significant (**Figure 7**).

While you are looking at the error bars, think about how plausible they are. Be suspicious if the SD are all small {especially in discrete distributions if they are consistently less than expected from sampling error, i.e., $<\sqrt{[N \times p \times (1-p)]}$ for binomial distributions, $<\sqrt{(\text{mean})}$ for Poisson distributions}.

It was by analysis of sampling errors that John Maddox, James Randi, and Walter Stewart showed that the claims by Jacques Benveniste (Davenas et al. 1988) in support of homeopathy were "a delusion" (Maddox et al. 1988). Benveniste's paper showed the mean ± SEM of basophils counted in a haemocytometer in triplicates. In this case, you would expect the counts to fit a Poisson distribution, with the expected SEMs being the square root (mean/3), but the SEMs reported were consistently far smaller than that. This provided very strong evidence that the reported counts were biased.

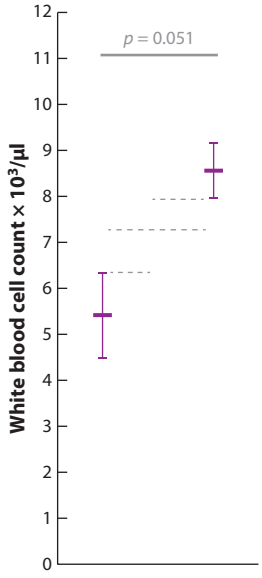


Figure 7

Estimating statistical significance from error bars. Graphs show mean ± standard error of the mean (SEM) where $N = 3$ independent counts for each of two groups of three mice. The gap between the top of the SEM bars on the left and the bottom of the SEM bars on the right is approximately double the size of the SEMs (so that there are approximately four SEMs between the two means). This suggests p is close to 0.05. Note that when $N = 3$, estimates of the means, SEMs, and p-values is extremely imprecise. If $p > 0.05$, differences of similar size or greater could have arisen even if all the blood samples came from the same group of mice. Even if $p < 0.05$, no conclusions should be made unless there is supporting evidence in the other figures.

SUMMARY POINTS

1. Statistics can be useful in cell biology to (*a*) monitor the conduct of experiments to make sure they were performed well, (*b*) analyze data when deciding which experiments are worth following up, and (*c*) present and interpret results in publications.
2. If you have statistically significant results, consider the effect size and biological significance, and try to confirm them by mechanistically different experiments. If your results are not statistically significant, try to design a better experiment, rather than just increasing the sample numbers.
3. If statistics (error bars, p-values) are shown in figures, they must be described in the figure legends, and N, the number of independent data points, must be stated. If you are reading a paper that shows error bars but does not say what they are in the figure legends or Materials and Methods, throw it away; it has not been carefully read by the authors, reviewers, or editors.
4. Consider whether the data distribution and error bars look plausible, taking into account sampling error, type of experiment, and instruments used. In binomial and Poisson distributions, if the SDs are consistently less than expected, be suspicious. Do your estimates of statistical significance correspond with the authors' conclusions and p-values (if stated)?

5. Statistics from replicates should be used only to monitor the conduct of experiments and should not be used to compare experimental groups or draw conclusions, which require independent data.
6. If you are unclear whether data are replicate results or independent results, think about what hypothesis is being tested and what population is being sampled.
7. If *N* is less than 10, plot the data points, and consider not showing any statistics.
8. Representative results should be shown only for nonnumerical data (e.g., photomicrographs, flow cytometry plots, Western blots) and not for numerical data or graphs, which should include the results of all relevant experiments. If there is quantitation of Western blot bands with error bars, the legends should make clear how many blots were measured and how the results were processed.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was funded by National Health and Medical Research Council (NHMRC) Grant 1016701 and Fellowship 1020136 and was made possible through Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institutes Infrastructure Support Scheme.

LITERATURE CITED

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. *Nature* 483:531–33

Cumming G. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge

Davenas E, Beauvais F, Amara J, Oberbaum M, Robinzon B, et al. 1988. Human basophil degranulation triggered by very dilute antiserum against IgE. *Nature* 333:816–18

Fisher RA. 1926. The arrangement of field experiments. *J. Minist. Agric. G. B.* 33:503–13

Fisher RA. 1928. *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd

Goodman S. 2008. A dirty dozen: twelve p-value misconceptions. *Semin. Hematol.* 45:135–40

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2:p.e124

Maddox J, Randi J, Stewart WW. 1988. "High-dilution" experiments a delusion. *Nature* 334:87–91

Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10:712

Vaux DL, Fidler F, Cumming G. 2012. Replicates and repeats—What is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep.* 13:291–96