NATURE | COMMENT

Research methods: Know when your numbers are significant

David L. Vaux

Nature **492**, 180–181 (13 December 2012) doi:10.1038/492180a
Published online 12 December 2012

Experimental biologists, their reviewers and their publishers must grasp basic statistics, urges David L. Vaux, or sloppy science will continue to grow.

Subject terms: Cell biology Peer review Publishing Mathematics and computing

ILLUSTRATION BY PETE ELLIS/DRAWGOOD.COM

The incidence of papers in cell and molecular biology that have basic statistical mistakes is alarming. I see figures with error bars that do not say what they describe, and error bars and *P* values for single, 'representative' experiments. So, as an increasingly weary reviewer of many a biology publication, I'm going to spell out again [1] the basics that every experimental biologist should know.

Simply put, statistics and error bars should be used only for independent data, and not for identical replicates within a single experiment. Because science represents the knowledge gained from repeated observations or experiments, these have to be performed more than once — or must use multiple independent samples — for us to have confidence that the

results are not just a fluke, a coincidence or a mistake. To show only the result of a single experiment, even if it is a representative one, and then misuse statistics to justify that decision, erodes the integrity of the scientific literature.

It is eight years since *Nature* adopted a policy of insisting that papers containing figures with error bars describe what the error bars represent[2]. Nevertheless, it is still common to find papers in most biology journals — *Nature* included — that contain this and other basic statistical errors. In my opinion, the fact that these scientifically sloppy papers continue to be published means that the authors, reviewers and editors cannot comprehend the statistics, that they have not read the paper carefully, or both.

Why does this happen? Most cell and molecular biologists are taught some statistics during their high-school or undergraduate years, but the principles seem to be forgotten somewhere between graduation and starting in the lab. Often, the type of statistics they learnt is not relevant to the kinds of experiment they are now doing. And, once in the lab, people generally just do what everyone else does, without always understanding why.

Even if experimental biologists do not need to use statistical evidence for their own experiments, they should have an understanding of the basics so that they can interpret others' work critically. They don't all need to understand complex statistics, or to hire professional statisticians, but there would be fewer sloppy papers if every author, reviewer and editor understood statistical concepts such as standard deviation, standard error of the mean (s.e.m.), sampling error and the difference between replicate and independent data (see 'Statistics glossary').

Table 1: Statistics glossary: Some common statistical concepts and their uses in analysing experimental results.

Back to basics

In the life sciences there are typically two types of publication: those that use large data sets and rely mostly or wholly on statistical evidence (for example, epidemiology, psychology, clinical trials and genome-wide association studies), and those that do not — such as much cell and molecular biology, biochemistry and classical genetics.

For papers with large data sets that rely purely on statistical evidence, recommendations exist for computing sample size, reporting on outlying results and other issues[3, 4]. But these guidelines do not serve authors of the other category of papers. Cell and molecular biologists have the luxury of being able to probe their experimental systems in multiple, independent ways and can therefore often get by with *N*s of three, without the need for sophisticated statistics.

The first figure in a typical paper in cell or molecular biology, for example, might show the difference in phenotype between three wild-type and three gene-deleted mice. The second figure might compare the levels of proteins in cells derived from the mice, looking at both the deleted protein and one of its substrates, or the effects of treating wild-type cells with an inhibitor of the protein encoded by the deleted gene. If the evidence from these experiments is consistent, and gives support to a coherent model, it would be unnecessary to analyse 30 mice of each type, or to repeat the Western blots of protein levels 30 independent times. Watson and Crick's paper on the structure of DNA[5] does not contain statistics, graphs with error bars or large *N*s.

Understanding the rudiments of statistics would stop experimental biologists from calculating a *P* value and a s.e.m. from triplicates from one representative experiment, and might stop the reviewers and editors from letting these pass unquestioned. If the results from one representative experiment are shown, then $N = 1$ and statistics do not apply. Besides, it is always better to include a full data set, rather than withholding results that are not representative. When *N* is only 2 or 3, it would be more transparent to just plot the independent data points, and let the readers interpret the data for themselves, rather than showing possibly misleading *P* values or error bars and drawing statistical inferences.

If the data in an experiment are equivocal, or the effect size is small, it is much better to come up with an extra, mechanistically different, experiment to test the hypothesis, than to repeat the same experiment until *P* is less than 0.05.

If statistics are shown, it should be for a good reason. Descriptive statistics, such as range or standard deviations, are only necessary when there are too many data points to visualize easily. Inferential statistics (an s.e.m., confidence interval or *P* value) should be shown only if they make it easier to interpret the results, and they should not detract from other key considerations such as the magnitude of the effects or their biological significance.

Figure legends should state the number of independent data points and, for experiments in which replicates were performed, only the mean of the replicates should be shown as a single independent data point. For replicates, no statistics should be shown, because they give only an indication of the fidelity with which the replicates were created: they might indicate how good the pipetting was, but they have no bearing on the hypothesis being tested[6].

All experimental biologists and all those who review their papers should know what sort of sampling errors are to be expected in common experiments, such as determining the percentages of live and dead cells or counting the number of colonies on a plate or cells in a microscope field. Otherwise, they will not be able to judge their own data critically, or anyone else's.

"Experimental biologists should know what sort of sampling errors are to be expected."

Repeat after me

How can the understanding and use of elementary statistics be improved? Young researchers need to be taught the practicalities of using statistics at the point at which they obtain the results of their very first experiments.

To encourage established researchers to use statistics properly, journals should publish guidelines for authors, reviewers and editors on the use and presentation of data and statistics that are relevant to the fields they cover. All journals should follow the lead of the *Journal of Cell Biology*[7] and make a final check of all figures in accepted papers before publication. They should refuse to publish papers that contain fundamental errors, and readily publish corrections for published papers that fall short. This requires engaging reviewers who are statistically literate and editors who can verify the process. Numerical data should be made available either as part of the paper or as linked, computer-interpretable files so that readers can perform or confirm statistical analyses themselves.

When William Strunk Jr, a professor of English, was faced with a flood of errors in spelling, grammar and English usage, he wrote a short, practical guide that became *The Elements of Style* (also known as Strunk and White)[8]. Perhaps experimental biologists need a similar booklet on statistics.

References

1. Cumming, G., Fidler, F. & Vaux, D. L. *J. Cell Biol.* **177**, 7–11 (2007).
2. Vaux, D. L. *Nature* **428**, 799 (2004).
3. Landis, S. C. *et al. Nature* **490**, 187–191 (2012).
4. Nakagawa, S. & Cuthill, I. C. *Biol. Rev. Camb. Philos. Soc.* **82**, 591–605 (2007).
5. Watson, J. D. & Crick, F. H. *Nature* **171**, 737–738 (1953).
6. Vaux, D. L., Fidler, F. & Cumming, G. *EMBO Rep.* **13**, 291–296 (2012).
7. Rossner, M. *The Scientist* **20**, 24–25 (2006).
8. Strunk, W. Jr & White, E. B. *The Elements of Style* (5th edn) (Allyn & Bacon, 2009).

Download references

More info

Author information

Affiliations

David L. Vaux is professor of cell biology at the Walter and Eliza Hall Institute of Medical Research and at the University of Melbourne, Parkville, Victoria 3052, Australia.

Corresponding author

Correspondence to: David L. Vaux

Comments

2012-12-12 06:21 AM

John Smith said: For in vivo experiments, this is of course a straightforward issue, but for in vitro cell culture work all this talk about "n" is meaningless and misleading. By definition, there is NO biological replicate in a cell culture system, so whether one repeats an experiment on two separate days or not, you are still testing for pipetting accuracy and, in a sense, you are simply introducing variability for the sake of introducing variability. The variability is not biological in any way; it is simply technical which, when it comes down to it, is really no different that using well-replicates in a single plate.

2012-12-12 06:26 AM

John Smith said: ....and by the way, I'm in no way saying that one should not bother reproducing an experiment. I'm just making the point that statistically, and in theory, you are quantifying technical variability and nothing else.

2012-12-13 12:01 PM

H T said: Researchers are not suffering from a lack of understanding of elementary statistics, but many lack understanding of the scientific method and the incentive to do so. The scientific method is about formulating and then falsifying hypotheses, but who in their right minds would do that nowadays? In the past, young children associated scientists with imaginary Einsteins coming up with wonderful theories and unambiguous experiments to 'prove' them. They grew up to become grant reviewers and university administrators now expecting the same from real-life scientists.

BTW, simple statistical analysis remain valuable as a guard against experimental errors, which include failure to reproduce results at a technical level. They may not be valuable for interpreting the significance of a paper, but I hope to see them in the right places, such as within logbooks.

2012-12-13 12:10 PM

H T said: In response to this commentary piece, may I quote the title of this week's editorial – "Words are not enough"? If anyone can turn the tide of poor statistics in research, editors of major journals like Nature appear to be our best bet.

2012-12-14 06:59 AM

More info

Hendrik Luuk said: David L. Vaux?s comment on the alarmingly high incidence of statistical mistakes in cell and molecular biology research papers makes an important point while also raising concerns about the credibility of some of its content. For example, the author makes a vague suggestion that it is fine to analyze samples of size three if several independent methods are involved and the results are consistent with the proposed mechanistic model. However, for the Wilcoxon-Mann-Whitney test, a method of choice for two-sample comparisons when N is small, the lowest possible two-tailed p-value when N=3 is 0.1, which, by convention, is not statistically significant. Furthermore, the apparent coherency between a proposed model and experimental results can be reinforced by choosing not to report the experiments failing to support the model.

I am also concerned about the ambiguity of the definitions of independent data and replicate data provided in the glossary. It seems to me that there are a lot of experiments of the same kind, which are somehow linked but, yet, not ?linked as much as possible? meaning that they represent neither independent nor replicate data. Consider the following examples. A behavioral test is run on knock-out and wild-type mice on a single day. Here, the data from individuals is usually considered as independent since the mice belong to a larger breeding colony. On the other hand, the experimental conditions on that particular day (the litter of the mice tested, time of day, season, whether the experimenter had played with a cat the day before etc.) are highly similar raising concerns whether the results can be reproduced a month later. Now consider an experiment where a batch of mouse embryonic fibroblasts (a heterogenous pool extracted from many embryos) is tested in a biochemical assay. Do the readings from different wells of identically treated fibroblasts represent independent or replicate measurements? From the genetic and environmental viewpoints, the similarity of the cell colonies in different wells and the similarity of littermates of the same genotype (assuming an inbred strain) might be roughly of the same order. On the other hand, the experiment as a whole is only performed once (e.g. on a single day), suggesting that we are dealing with replicate data.

In sum, consistent and well-articulated criteria for producing credible research are urgently needed.

2012-12-14 07:19 AM

Philippe Pognonec said: "John Smith" must be joking, right? No other variations than experimental when replicating experiments in cell culture. Ah ah ah! I do appreciate this "representative" comment, for sure...

2012-12-14 10:30 AM

Steven McKinney said: " Even if experimental biologists do not need to use statistical evidence for their own experiments ..... "

If there is one phrase that encapsulates the problems with this article, this has to be it.

Dr. Vaux really should consult with a statistician, and take some refresher courses in experimental design and the statistical analysis thereof. Terry Speed is right there at WEHI.

Nature has in the past few months published a series of considered comments and essays regarding the problem of poor reproducibility of results from life science researchers. This article does not add to them, but rather illustrates why they are so relevant at present.

"They don't all need to understand complex statistics, or to hire professional statisticians ..... "

If you do not understand statistics, you should indeed hire a professional statistician. Lack of statistical discipline is at the heart of this problem currently plaguing the life sciences. More info

"For papers with large data sets that rely purely on statistical evidence, recommendations exist for computing sample size, reporting on outlying results and other issues. But these guidelines do not serve authors of the other category of papers."

I couldn't disagree more. There is no "other category". No discipline in the quantitative sciences can ignore the properties of measurements subject to random variation, least of all those that seek to keep sample sizes small to avoid excessive harm to living organisms such as knock-out mice. Running experiments too small in size yields poor results as has by now been so well documented, and represents unethical conduct when the experimental units are living creatures such as higher order mammals. When poorly researched findings end up forming the basis for human medical interventions, people (healthy and sick) suffer unnecessarily.

2012-12-17 07:35 AM

John Smith said: @Philippe Pognonec: Well, let's see. I repeat a transfection, say, on two different days at exactly the same time of day. I use the same cell passage from the same mother aliquot, the same batch of media, the same batch of siRNA, identical cell numbers, the same incubator, same hood, exactly the same transfection method, same antibodies, same Westerns, same qRT-PCR primers/probes, same mastermix blah blah blah.

Where exactly is the biological variability? it can't be the cells; they are identical at the same passage (and for many cells +/- several passages). So, you tell me: what are you testing in this set-up? It's certainly not biological.

2012-12-19 03:58 AM

Axel Berger said: Dear Professor Vaux,

writing for nature you address an international audience of scientists not of your subject. As a non-native speaker and a non-biologist myself I find your valuable comment hard to understand in places and I believe the latter impediment to be the relevant one here. I don't quite get your distinction between replicate and independent data. Whenever a scientist reads about a result in a journal and replicates it in his own lab or with new patients in his own clinic, that's what the scientific method and replication is all about. But your "replicate" seems not to be the same as everybody else's "replication". Trying to understand what you may mean and having done undergraduate physics, graduated in engineering, and now dabbling in undergraduate archaeology this is what I can come up with:

When a sample is sent to a radiocarbon lab to be dated and after treatment and homogenization is split into two parts and both measured, that's replicate data. When two different samples are taken from the same stratum, both sent in and both dated, that's independent.

Even if I should happen to be right and thus having understood your meaning correctly, this does not make your wording less ambiguous and more understandable to workers in other subjects, for whom your worthwhile comments are just as relevant. I therefore ask you, and much more so the editors at nature, to keep specialist insider wording out of articles and make them comprehensible to the wider audience they're aimed at.

Thank you
Yours truly

Axel Berger

More info

2013-01-08 08:06 AM

Stanley Lazic said: @Axel Berger: You are quite right that some of the terms used in the article are not consistent between disciplines (or even within). In many fields, *replicate* refers to independent data, which can lead to confusion (some definitions are given in this article).

@John Smith: You are correct that biological variability is zero in many cell culture experiments (e.g. with cell lines), which does not allow one to generalise to other samples (e.g. people, animals) in the way that most studies do. That is why it is necessary to perform multiple repetitions of the whole experimental procedure, because at least one can then generalise to "other experiments conducted under similar circumstances". So instead of demonstrating that the intervention applies across a range of heterogeneous biological material (i.e. independent samples), one demonstrates that the intervention works multiple times on the same homogeneous biological material. That is why it is best to make the experiments as independent as possible (so make a new batch of media, etc., but there will likely be practical limitations), and why taking more measurements in a single experiment doesn't provide one with greater certainly about treatment effects.

Subscribe to comments

Related links

Related links in Nature Research

Methods: Face up to false positives

Drug development: Raise standards for preclinical cancer research

Must try harder

Nature ISSN 0028-0836 EISSN 1476-4687