Reporting p Values

Wolfgang Huber[1,*]
[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstrasse 1, Heidelberg 69117, Germany
*Correspondence: whuber@embl.de
https://doi.org/10.1016/j.cels.2019.03.001

Cell Systems Editorial Board member Wolfgang Huber demonstrates how he thinks about small p values.

In Douglas Adams' science fiction series *The Hitchhiker's Guide to the Galaxy*, the "infinite improbability drive" exploits the probabilistic laws of physics to enable instantaneous travel across enormous distances. According to quantum theory, the location of an object is described by a wave function that usually peaks at one particular point in space but that also stretches out across the whole universe so that, with some small probability, the object can instantaneously materialize anywhere. This probability is exceedingly small, but in *The Hitchhiker's Guide to the Galaxy*, the drive funnels it into spaceship propulsion and various other plot twists.

Absurdly small probabilities occur in statistical tests. Just as the wave function stretches to the boundaries of the universe, many statistical distributions extend all the way to infinity. An example is shown in Figure 1. Clearly, the two groups are different, but is a probability as small as $p = 1 \times 10^{-106}$ meaningful outside of science fiction? Is the result more significant than one with $p = 10^{-30}$ or less than one with $p = 10^{-200}$?

In practice, we should not get carried away. As scientists, we should not report a p value that is smaller than the probability that there was an unforeseen mishap in the experiment or the processing of the data. Statistical tests and p values are designed to take into account various sources of noise and error in the data and to quantify the associated probabilities faithfully; but they cannot take into account *everything*. What is outside of scope depends on the test: it can range from mundane problems like a sample swap to unfortunate calibration errors in the measurement instrument to cosmic rays messing up bits in computer memory and aliens visiting the lab and secretly modifying your lab book. Such things tend to be unlikely—that is, they have very small, or even very, very, very small probabilities; but hardly as small as $p = 1 \times 10^{-106}$. Thus, our stated confidence in a result must be a combination of what the statistical test says (the p value) and everything else that could go wrong. Compared to a p value in the 10^{-6}s, the latter may be negligible; compared to one in the 10^{-106}s, it most likely is not.

An example that is familiar from popular culture and notorious among statisticians is match probabilities from DNA evidence in criminal cases. When there is a match between the suspect's DNA and crime scene DNA, and someone says that the probability that a random person would match is 10^{-16}, who would not consider the case solved? But how big is the reference database of random people? It can hardly be bigger than the world population (around 10^{10}), so how can we reliably estimate a probability that implies choosing among a million times that number of people? In practice, the database will cover different genetic backgrounds differently, so a defendant from a rare genetic background may be at a disadvantage. Moreover, the probabilities of sample swaps, contamination, or tampering with evidence are non-zero but do not enter the statistical calculations.

Common statistical tests are good at modeling common sources of errors, and they deal reasonably with moderate probabilities. However, they tend to fail for rare events: artifacts, mistakes, mishaps. If these happen, they tend to have a catastrophic impact on the outcome, but we have no reliable ways of quantifying them. By reporting extremely small p values, we convey a false certainty.

When the probabilities spewed out by statistical software come into the regions of the age of universe (in whichever units) or the number of protons in the known universe (around 10^{80}), we signal our appreciation of the limits of our probability models by not reporting such a number at face value, but by saying "it's below detection limit." For instance, many tests in the statistics environment R report p values that are smaller than 2.2×10^{-16} simply as "$p < 2.2 \times 10^{-16}$".

This is good scientific practice. Just as we do not report measurements readings with many more digits than the instrument precision, we do not report close-to-zero probabilities with ridiculously small numbers and rather use such an interval notation.

Where is the cutoff? This, of course, depends. (There is a whole subfield of statistics, extreme value distributions, concerned with legitimate uses of very small probabilities.)

Our above argument against literally reporting very small p values is a pragmatic one. Besides that, there are four more theoretical points. First, statistical formulae or the software that computes them are often not very accurate at the extreme ends of the distributions. Approximations that work well for moderate probabilities break down there. Usually

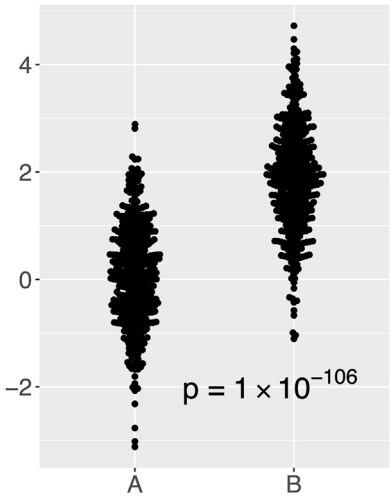

Figure 1. Two-Groups Comparison with a t Test

we do not mind, as we are concerned with moderate probabilities; but if the numbers become extremely small, we'd better take them with a grain of salt.

Second, the introduction of the p value into scientific practice has been a choice of mathematical convenience, not usefulness. Everyone knows that the p value is the probability of these or more extreme data under the null, but that is a fairly abstract concept. It tends to be mathematically tractable and, thus, easy to deploy—but arguably, it is the right answer to the wrong question. What scientists usually want is something like, "what is the probability that I'll later be proven wrong if I publish this?"; in technical terms, that is what the false discovery rate (FDR) measures. Unfortunately, in the classical statistical framework, this quantity is hard to get to. Indeed, a small p value neither implies a small FDR nor vice versa. However, there is a break in the clouds: modern statistical approaches from the fields of multiple testing, of Bayesian inference, or those based on simulations can produce these more useful measures.

Third, there is the issue of statistical significance versus effect size (and, more generally, scientific significance). Especially in "omics" analyses, where tests can be aggregated over thousands of measurements, it is easy to find significant effects (small p values) even if the effect size is hardly noticeable. Such situations also open the door for confounding, i.e., spurious associations that are really caused by a third, perhaps unobserved variable. When the level of replication in our data is so high that very small p values are even possible, and we see one, we should gracefully acknowledge its presence and then focus away from it and look at effect size and causality instead.

Fourth, more philosophically, a null hypothesis that can be rejected with such a small p value—and, more generally, something that we can conclude with such certainty—is probably not very interesting. It is too easy. We probably knew it before, or it is self-evident. Genuinely surprising, non-obvious discoveries usually come at the cost of a little less certainty.

ACKNOWLEDGMENTS

The author acknowledges fruitful discussions with Susan Holmes.