Advances in Health Sciences Education (2019) 24:637–642
https://doi.org/10.1007/s10459-019-09915-3

EDITORIAL

Statistics 101

Geoff Norman[1]

Over the last few weeks, I have encountered several situations that demanded a serious look at the logic of statistics, a subject we all (well, most) anguished through in our near or distant past. Here are 3 apparently disparate anecdotes:

1. We recently submitted a paper to an anatomy journal. The paper showed a large and highly significant difference in learning resulting from different presentation formats. One reviewer was highly critical because

   "There is not [sic] power analysis presented. Should be included as part of experimental design."
2. Last week I rejected two papers in a 1 h period for precisely the same reason. Both involved a survey with multiple items. Authors correlated responses to each item to other measures such as personality, demographics, attitude to social situations. Typically there were 20–40 items in the questionnaire, resulting in somewhere between 100 and 200 tests.

   Within 24 h, a colleague sent me a recently published paper in which the authors looked at differences on a survey related to a number of demographic variables, based on an item by item analysis on a test of about 40 items. They examined a total of about 100 contrasts.
3. I read a report of the keynote address at the 59th annual meeting of the Psychonomic Society in 2018 delivered by Hal Paschler, a highly regarded scientist working in science of learning. Paschler was addressing the issue of non-replication in psychology, an issue that I have recently written about (Norman 2017). The basic problem is that many highly cited findings in cognitive and social psychology cannot be found when the study is repeated. His solution was to build greater statistical power into the studies by increasing the number of trials per subject (increasing sample size).

What do these three anecdotes have in common, aside from an emphasis on statistics? My claim is simple. All are at least partially wrong. And they are wrong because the authors or reviewers did not understand the elementary logic of statistical hypothesis-testing. That's why I decided to turn this editorial into a primer of basic statistical logic. If published and peer-reviewed papers, as well as invited presentations by senior researchers, are getting it wrong, then it's time to set the record straight.

✉ Geoff Norman
norman@mcmaster.ca

[1] McMaster University, Hamilton, ON, Canada

Regrettably, to defend my claim, I must resort to pedantry, For some of you, this may be a subject you could recite in your sleep. To decide if you are one of them, I have devised a simple pretest. It goes like this:

> One of your graduate students shows you the findings from her latest study. It is a simple two group comparison – the details do not matter. She compared the two group means using an independent sample t test, which turned out to be just significant –> p = .0498.
> After due celebration at the local pub, you are haunted by all the recent publicity about non-replication. So the next day you insist that she repeat the study just to be sure.
> She does it exactly like she did before. Nothing has changed in the design.
>
> QUESTION: On purely statistical grounds, what is the probability that you will arrive at the same conclusion in the second study; that is, you will reject the null hypothesis a second time?

You will find the correct answer in the footnote[1] on the next page, just to keep you from peeking. For those who got it right, for the right reasons (a legitimate logical argument as to how it comes out that way), skip the next section. For the remainder, read on.

A primer of statistical logic

It's time for a review. The basic logic of statistical inference is now about 100 years old (RIP Sir Ronald), and has withstood repeated challenges (Cohen 2016), but lives on. We've added effect sizes and confidence intervals and odds ratios, but like it or not, you still will have trouble publishing anything quantitative without the magical "$p < .05$".

So let's be clear on what it is and is not telling us. To make it easy, let's take the simplest case in the book—comparing a sample to a population. Suppose, for example, we have come up with a pediatric analog to "Luminosity" (which does NOT work, by the way). Instead of reducing cognitive decline with aging, we're going to work at the other end, and devise an online intervention, Lubricosity—designed to make fluid intelligence flow just a bit better, and raise IQ of kids (that doesn't work either, but let's pretend it does for now).

However we begin by doing the opposite and setting up a "null hypothesis", which in contracted form is:

$$H_0 : \text{Population mean(lubricosity)} = \text{Population mean} = 100.$$

The basic logic is simple: *If* our study kids are a random sample from the population of all 12 year olds, and *if* the treatment doesn't work, and *if* we did the study a zillion times and displayed all the sample mean IQ's, they would be normally distributed around 100, the mean IQ in the untreated population.

So we randomly sample 100 6 year olds, enrol them in the program for 3 months, then measure their IQ. Critically, because we re looking at the distribution of *means computed from sample size 100,* the width of the normal distribution of means, the *standard error of the mean,* would be the standard deviation of the original scores (15 for IQ) divided by the square root of the sample size ($\sqrt{100} = 10$), or 1.5. The data would look like Fig. 1a), where we have also converted the IQ to a Z score, expressing everything in standard error units.

The next step in the logic is to declare that, if the sample mean of the study is sufficiently unlikely to have arisen by chance if it came from this (null hypothesis) distribution, we will r*eject the null hypothesis* and declare that the observed difference is statistically

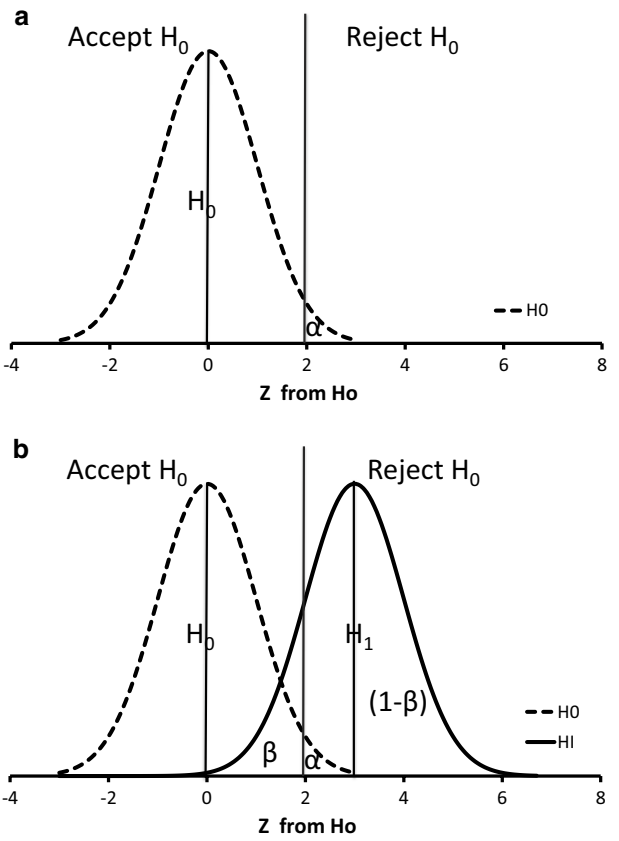

Fig. 1 **a** The distribution of means under the null hypothesis. The width of the curve is the standard error of the mean (SEM). The critical value corresponds to a *p* value of .05 (two-tailed). **b** Addition of the distribution of means under that alternative hypothesis, centred on the observed difference of 4.5, or 3.0 SEMs

significant. And "unlikely" is always defined the same way; a probability of occurrence of less than 5 in 100.

This then establishes a critical value out on the distribution beyond which the tail probability is $<.05$, which is conventionally called "alpha". For the simple z test, this arises at a Z of 1.96, as shown in the figure. In turn, this defines two zones: one to the left of the critical value, where we fail to reject (accept) H_0, and one to the right, where we reject H_0. (Again, because this is a two-tailed test, there is a similar zone on the left side of the graph, but we'll ignore this). See Fig. 1a).

Now comes the critical part. If we reject H_0, we then logically declare it comes from a different distribution, the H_1 distribution, somewhere to the right of the critical value. Now, if we were doing all this hypothetically before we had the data, H_1 could be centered almost anywhere (which is why sample size calculations always come up with the right number!) But once the study is over, we have information about where the H_1 distribution might be—exactly where we observed it. So we assume that the second population of studies that "worked" has a distribution centered on our "best guess" of the new population mean, the observed sample mean. We also assume the distribution has the same standard deviation as the untreated population ("homoscedasticity", if you want to sound intellectual).

[1] The probability that you will reject the null hypothesis a second time is 0.50

Let's assume we found that the study mean was 104.5, 3 standard errors above H_0 mean. Then the curve looks like Fig. 1b):

Now the important bit is the area of the H_1 curve to the *left* of the critical value. That is beta, the likelihood that you would *not* declare a significant difference, under the alternative hypothesis that there was a difference of 4.5 IQ points. In this case, it's 0.15. And (1-beta) is the likelihood of detecting a difference if there was one, which is called "power". This is (1–.15) = .85

To be very clear, what this means is that even though this study found a difference of Z = 3.0, corresponding to a *p* value of 0.0001, the chance of replicating the finding of a significant difference is still only 85%.

And that brings us to the question posed at the beginning of the primer. If we computed a *p*-value of exactly .05, this corresponds to a sample mean at Z = 1.96 on the H_o distribution. That means the H_1 distribution is centred right on the critical value. Half of the distribution lies to the left of the critical value and half to the right. The likelihood of replicating the original finding of a significant difference is only 50%!

In Fig. 2, I've plotted the likelihood of replication as a function of the calculated *p* value (for "significant" results). It goes from .50 to .97.

Now let's return to the 3 scenarios posed at the beginning:

1. "There is not [sic] power analysis presented. Should be included as part of experimental design."

   As we described above, the power of a statistical test is the probability of finding a significant difference **when there is one present** (i.e. when the data come from the H_1 distribution). In the study we reported all the critical differences had *p* values ranging from 0.001 to 0.0001. The probability of finding a significant difference was 1.0, because we **did** find a significant difference. The power calculation adds nothing.

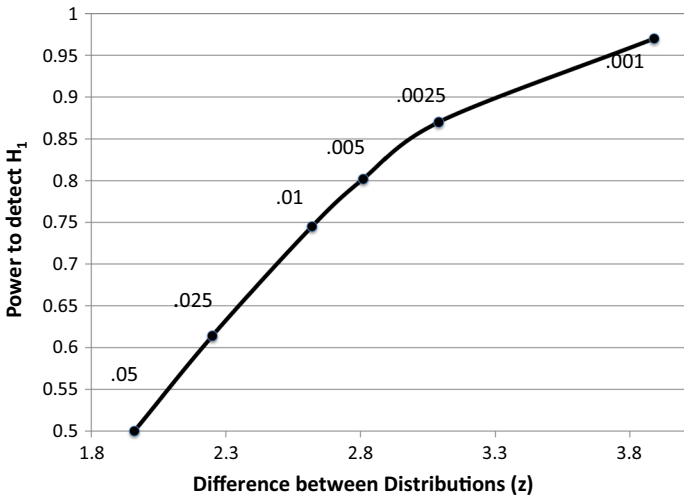

Fig. 2 Relationship between the alpha level associated with an observed difference and the power to detect the difference

Power calculations are useful when a difference is expected but was not found, to estimate the likelihood of finding a difference of some presumed magnitude. They have no value when a difference was detected.

2. Pashler's solution to non-replication was to build greater statistical power into the studies by increasing the number of trials per subject (increasing sample size).

   As it turns out, whether increased sample size will or will not reduce problems of non-replication depends on your belief that the original finding of a significant effect was true or not.

   It would appear that most of the literature on replication and non-replication holds the view that the original finding is a false positive; the effect is really not there. (John et al. 2012; Masicampo and Lalande 2012; Simmons et al. 2011). If this is the case, then the conceptual logic we have described demonstrates that the likelihood of a false positive is always 0.05, because alpha is set at .05 from the outset. No amount of increase in sample size changes that. (In making this claim, we are deliberately ignoring the many potential investigator biases discussed in some of the references in the bibliography, and are simply looking at the theoretical probability).

   So what does increased sample size achieve? Going back to the derivation, as sample size increases, the standard error of the mean decreases, so the two curves move further apart on the original scale. Power increases as the overlap decreases, but this only impacts on the likelihood of detecting a true difference. In other words, increased power will permit detection of smaller and smaller effects if they are real, but does not change the likelihood that an effect will be falsely declared significant.

3. Correlations at item level from a multi-item survey.

   As an extension of the effect of study design on alpha, there is insufficient attention to the effect of multiple tests on the overall alpha level. Analysis of differences observed in a multi-item inventory is rarely sensible, for at least two reasons. First, as the number of tests increases, the likelihood of observing a significant difference increases. As a simple example, using an alpha of .05, the likelihood of finding at least one significant difference (false positive) with 5 tests is .23; with 10 tests, .40; with 20, .65 and with 50, .92. It is not possible to distinguish false from real positives. Moreover, with several hundred participants, which is not uncommon in surveys, even tiny correlations will emerge as statistically significant. One solution is a Bonferroni correction, dividing the conventional alpha by the number of proposed tests. In the published example, the approximately 40 significant results in the paper drops to 3 with a Bonferroni correction.

   There is a second reason why such an approach is not useful. The whole point of a multi-item inventory is recognition that a single item is not sufficiently reliable to yield credible results. Indeed a factor analysis is directed at providing guidance for creating subscales to identify underlying dimensions, and an overall internal consistency calculation is based on the assumption that all items are measuring the same underlying dimension. So analysis should be conducted at the scale or subscale level, not the item level.

Conclusion

I expect for many readers, this whole exposition has lead to a succession of deep yawns. But the analysis illustrates a number of key principles that emerge from the logic of statistical inference. One is that is that the likelihood of replicating a true positive study is always

less than 100%, and in many typical situations, is substantially less. This has nothing at all to do with researcher bias, sloppy methods, deliberate falsification, or anything else. It's predestined by the logic of statistical inference. At least some of the instances of non-replication derive from the underlying theory, and are unavoidable. Further, the analysis exposed some common, but inadequate statistical practices, which are based on inadequate understanding of the underlying logic.

While there is an extensive literature directed at understanding failure to replicate based on various methodological biases (Francis 2013; Schulz et al. 1995), there is, I think inadequate recognition of the fact that non-replication is an architectural feature of Fisherian statistical inference.

References

Cohen, J. (2016). The earth is round (p < . 05). In *What if there were no significance tests?* (pp. 69–82). Routledge.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153–169.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below.05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271–2279.

Norman, G. (2017). Generalization and the qualitative–quantitative debate. *Advances in Health Sciences Education*, *22*(5), 1051–1055.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA, 273*(5), 408–412.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.