A short overview of testing and reporting significance

A. Rusanen
INAR/Physics, University of Helsinki
Helsinki, Finland

ABSTRACT

This work tries to succinctly summarize the pitfalls of p-values in data science and to present common alternative approaches. It was done for a short seminar course in the University of Helsinki.

KEYWORDS

Hypothesis testing, significance, inference

1 INTRODUCTION

Statistical hypothesis test and p-values are valuable tools. They are however very easy to misuse and combined with publication bias in some fields this has led to a replication crisis [4, 7]. There has been some radical reactions to this, with some journals declaring blanket bans on p-values and confidence intervals [12]. Other journals have made statements, which essentially boil down to: You can do statistical tests and they are sometimes useful, but they should not be substituted for thinking [12]. Similar sentiments have been echoed from the researcher community [1].

The rise and fall of p-values is essentially an unintended consequence of misinterpreting what statistical significance means [6, 9]. In addition to the false belief in the absolute nature of statistical significance, there are various human, research cultural and financial factors contributing to a publication bias and essentially data fabrication [4, 7, 9]. There have been calls for more research and funding focus on replication and less on the hunt for novelty in order to reduce these factors, see for example [4]. Another suggestion discussed in [4] and [9] is pre-registration of confirmatory studies and then publishing the results regardless of the statistical significance, in order to reduce publication bias.

In this work I will try to summarize some of more useful concepts for natural sciences and to explain the common pitfalls that lead to misuse.

2 HYPOTHESIS TESTING

As any textbook on basic statistics will tell you, a statistical hypothesis test is essentially a statistical model including (see for example [11]):

- A null hypothesis, H_0, and its complement H_1.
- Several assumptions about the data and how it was generated.

Based on this model, one calculates a test statistic, and based on the test statistic, one gets a probability of the model generating as or more extreme deviation in the test statistic. This is called the p-value. Unlike some journals, students and authors seem to think, this is not an absolute measure of certainty in your results and probably should not be the defining factor in whether you publish your study or not [6]. Also, no single statistical test is appropriate in all situations.

Things are further complicated by the fact that the usefulness and validity of your p-value depends on what your null hypothesis is and if the assumptions you made are valid. Very often the tests null hypothesis is simply that the value in question is zero, that is, your test only comments on if there seems to be any effect at all. There is a common misconception that it guarantees anything about the accuracy of your computed value [6]. Greenland et al. [6] also notes that violations of the model assumptions are both common and make interpreting p-values hard. In more detail, if your assumptions are false, your test is likely to be wildly optimistic, since the data is then highly unlikely under the statistical model behind the test.

While software libraries have made statistical tests easy to do and automate, usually ensuring that the assumptions are correct is left to the user. As [6] notes the assumptions are in general hard to verify, because you are usually relying on another statistical model when testing for deviations from the assumptions.

2.1 Terminology

Some terminology needs to be introduced, in order to make subsequent discussion easier. The notation here is from [2], but is common in statistical literature:

- `Critical threshold`, α: If the p-value is below this, you reject H_0. This sets your accepted false positive rate, assuming your data matches the statistical model, including H_0.
- `False positive`: Rejecting H_0, when it is true. Also known as Type I error.
- `False negative`: Failing to reject H_0, when H_1 is true. Also known as Type II error.
- `Power`: 1 - false negative rate. Essentially a measure of how often the test is significant, if H_1 is true. Usually higher with larger samples and with larger effect sizes. Calculating power usually requires making assumptions [5].
- `Base rate`: How many of the initial H_1 can be expected to be true. A measure of how unlikely your hypotheses are in general.

If we discount, the human factors for a moment, [4] summarizes the idea behind the first claim in [7] that most results are false as follows: The more unlikely it is to find something, the larger fraction of published results with significance are false positives. A measure that they use for this is, false positive report rate:

$$\alpha(1-\pi)/[\alpha(1-\pi)+(1-\beta)\pi] \tag{1}$$

Where α is your significance limit, π the base rate of actually true hypotheses and β is the power of the test. Equation 1 calculated with various values of power and base rate can be seen in Figure 1. Another related probability, positive predictive value, is used in [7]. As Ioannidis argues, things get considerably worse when we start including the effects of bias and repeated hypothesis testing [7].

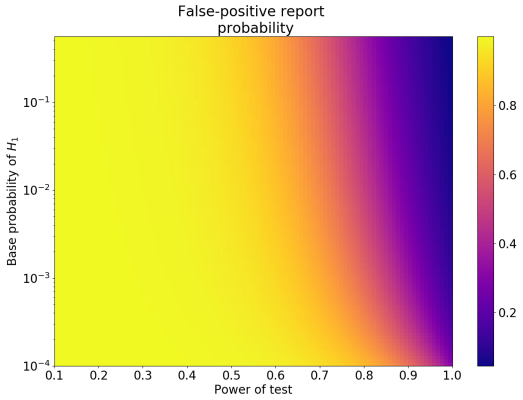

Figure 1: False-positive report rate calculated from equation 1 with α=0.05 . It is the probability of a positive result being a false positive, without accounting for human and model errors. As can be seen in studies with low pre-study probability of H_1, almost all reported positives are likely to be false. The crucial point here is that α is not the probability that you are wrong, given a positive result. It's not even close. The figure shows that with very high powers the problem can be avoided, but calculating the actual power of your study is not simple [5].

3 MULTIPLE HYPOTHESIS TESTING

So how do we keep the error rate low when we ourselves are doing multiple hypothesis testing? Essentially, doing multiple hypothesis testing is a balancing act on controlling false positive and false negative rates on the tests. Some statistical models already include this sort of control by default, such as the popular ANOVA tests. In general, there are two quantities one might try to control [10]:

- `per-family error rate, PFER`: The expected amount of Type I errors.
- `Familywise error rate, FWER`: The probability of at least one Type I error.

Although it can be noted that several others have also been used in the literature [2].

3.1 Single step procedures

All p-values get the same adjustment, these procedures are conservative so they reject results eagerly [2]. Most commonly know are the Bonferroni $p_{j,new} = min(m*p_j, 1)$ and Šidák $p_{j,new} = 1-(1-p_j)^m$, where m is the number of tests, which aim to minimize FWER [2]. There are other procedures, but they usually have to make more liberal assumptions about the hypotheses being tested [2].

3.2 Stepwise procedures

These methods consider the tests in succession, in addition to just the number of tests, thus gaining some more power [2]. They essentially work by ordering the tests by p-value and subtracting the index of the test from m, the ordering determines whether this is a step-down, where the p-values are in ascending order, or a step-up where they are in descending order [2].

3.3 Resampling

In many cases the you don't know what the joint distribution of the test statistic looks like, in these situations you can use resampling strategies, such as bootstrapping [3] to try and adjust the p-values [2].

3.4 Confidence intervals

Another related concept to p-values. While they avoid the singular value focus of a p-value, the problem with confidence intervals is that they are a similar construct with a variety of approximations and models behind them [6]. Here are two common ways to calculate confidence intervals [6]:

- `From the statistical model`: Essentially taking all values for H_0 for which your test would be significant and calculating a lower and upper bound. In general case hard to calculate.
- `Assume model`: Give confidence limits based on a an assumed gaussian or other distribution for your point estimate, may use bootstrap method [3] which is essentially resampling to estimate these parameters.

In essence confidence intervals trade one value for two limits, that one can misuse all the same. Even with simple confidence intervals, a variety of very similar misconceptions to those in p-values have risen among authors [6]. One of these is that directly comparing confidence intervals does something meaningful [6].

4 BAYES FACTORS

Bayes factors are essentially comparing the probabilities of data, given your two models [8]:

$$BF = \frac{p(D|H_1)}{p(D|H_2)} \tag{2}$$

Where D is your data and H_n are your two hypotheses. Bayes factors allow you to compare evidence between two different models and can work well when the underlying models are discrete or simple, otherwise they may be just very difficult to calculate and require defining prior distributions which may be unknown [8]. While there are some guidelines for the strength of evidence [8], trying to draw similar hard decision lines that led to the misuse of p-values should be avoided [1].

5 SUMMARY

Statistical tests, p-values and confidence intervals are useful tools when doing statistical analysis. In general, I would go with the advice from [1] and [12]: A total ban of a particular statistic doesn't achieve anything and that unfortunately we just must put in the effort of interpreting what has been actually done, instead of relying on mental shortcuts, which are often not even remotely true [6]. This can cause some uncertainty, but it is important to note that the certainty provided by p-value cutoff was illusory at best. In applying statistical methods, some key points are:

- `p-value`: The cutoff is arbitrary; it doesn't imply that your results are (un)publishable. P-hacking, that is unethically adjusting data so your test is significant, is mostly a human problem, to which pure statistics has no answers.

- `Confidence intervals`: Are sometimes useful to have for point estimates, but you really need to specify what you mean by them.
- `Multiple hypothesis testing`: Use controls on your acceptance criteria. If you can, have a separate validation data set or a properly planned validation study, since no statistical trickery saves you from mistakes made during the data collection.
- `Bayes factors` These allow you to imply different levels of confidence assuming you can calculate the probabilities. It is a tool that can help you choose between two different models.

REFERENCES

[1] Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. *Nature* 567 (2019), 305–307. https://doi.org/10.1038/d41586-019-00857-9

[2] Sandrine Dudoit, Juliet P. Shaffer, and Jennifer C. Boldrick. 2003. Multiple hypothesis testing in microarray experiments. *Statist. Sci.* 7, 1 (2003), 71–103.

[3] Bradley Efron. 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Amer. Statist. Assoc.* 78, 382 (1983), 316–331.

[4] Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H. Parker. 2017. Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews* 92, 4 (2017), 1941–1968. https://doi.org/10.1111/brv.12315

[5] Andrew Gelman and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651. https://doi.org/10.1177/1745691614551642

[6] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 4 (01 Apr 2016), 337–350. https://doi.org/10.1007/s10654-016-0149-3

[7] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2, 8 (8 2005), 696–701. https://doi.org/10.1371/journal.pmed.0020124

[8] Robert E. Kass and Adrian E. Raftery. 1995. Bayes Factors. *J. Amer. Statist. Assoc.* 90, 430 (1995), 773–795.

[9] Regina Nuzzo. 2014. Scientific method: Statistical errors. *Nature* 506 (2014), 150–152. https://doi.org/10.1038/506150a

[10] Juliet P. Shaffer. 1995. Multiple hypothesis testing: A review. *Annual review of Psychology* 46 (1995), 561–584.

[11] Larry Wasserman. 2004. *All of statistics : a concise course in statistical inference.* Springer, New York.

[12] Ronald L. Wasserstein and Nicole A. Lazar. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70, 2 (2016), 129–133. https://doi.org/10.1080/00031305.2016.1154108