sciendo

Kairos. Journal of Philosophy & Science 24, 2020
Centre for Philosophy of Sciences of the University of Lisbon

On how statistics is used and abused to find truth in Science

Tiago A. Marques
tiago.marques@st-andrews.ac.uk
Centre for Research into Ecological and Environmental Modelling,
The Observatory, University of St Andrews,
St Andrews, KY 16 9LZ, Scotland,
+44 1334 461842, +44 1334 461800
Centro de Estatística e Aplicações,
Departamento de Biologia Animal,
Faculdade de Ciências, Universidade de Lisboa, Portugal

DOI 10.2478/kjps-2020-0013

"Truth is neither absolute nor timeless.
But the pursuit of truth remains at the heart of the scientific endeavour",
Michela Massimi
<https://aeon.co/essays/its-time-for-a-robust-philosophical-defence-of-truth-in-science>

The notions of truth and falsity are fundamental in our everyday lives. We certainly want to be able to distinguish these two different, in fact opposing, concepts, and we hope to be able judge upon them when faced with a given fact. If Donald or Jair tell me that a drug is good for me, should I believe it, or should I not believe it. In other words, are Donald or Jair telling me the truth, or are Donald or

Jair lying to me? In fact, Donald and Jair might be agents of disinformation, if they are conscientiously telling me a lie, or they may be misinformed if they are just repeating a statement they believe is true but in fact is false. In that case, they are not lying, even if they are saying something which is not true.

The aim of this paper is not to discuss this kind of hard, theoretical problems but to call attention to the central role of the concepts of true and false in the specific scientific domain of Statistics. Unfortunately, most practitioners of statistics are not really interested in the details of the procedures they use, nor is the key aspect that these methods have assumptions, and the failure of these assumptions can have severe consequences. Often, practitioners of statistics want to be able to make true statements. Yet, unfortunately, one might add, what they often do is to base these statements on data that does not contain reliable information for the required inferences. In that sense, we can perhaps think of statistics as a possible way to guide us in the process of distinguishing falsity and truth in Science. But the distinction is not always easy. In fact, it being easy is probably the exception, not the rule. Not good news for those who use statistics without thinking carefully about the process. The notions of true and false in science are indeed far from straightforward. Some statements are easily identified as true or false. But others are tricky, for a multitude of reasons. For anything but the simplest statements, the notion of true or false might depend on a number of unknowns. Even if a statement is simple whether it is true or not might not be obvious.

As an example, I suspect that most humans, if not all, would agree that the statement "If you are reading this, yesterday your hearth pumped blood through your body" is true. Nonetheless, the statement "If you are reading this, tomorrow your hearth will pump blood through your body" might not be true. I hope it is of course, at least for most readers, but there is a non-zero probability that this sentence will not be true. And in that sense, the sentence is not true, because it is not always true. We do not know as the question gets raised, but interestingly, 48 hours after the sentence is used, that sentence, unambiguously, will have been be either true or false. There is a probability that the sentence is true or false, but we do not know what that probability is beforehand. We can estimate it, but even then the way to estimate that probability might depend on a number of assumptions, perhaps in the form of an epidemiological model, we are willing to make about the processes involved. The probability that the statement is true will be far higher if you are healthy than if you eat fast food every day, smoke two packs a day and have not made any kind of sports in the last 30 years. Unfortunately, all your good health conditions will not help you if you are run over by a car tonight. Even if we often prefer to ignore it, luck does play an instrumental role in what becomes true or false. Sometimes you cannot say whether something is true or false, you can at best provide a probability that something is true or false. We are surrounded by events of probability 0 that occurred. In other words, we live in a world where what was supposed to never occur, does occur all the time; where statements that were almost for sure false, then appear to be true. If you believe in evolution (and if you do not, do the rest of us a favour and go read about it from reliable sources!) each one of us is the result of our ancestors having reproduced, each and every time, since the dawn of times. That is amazing in itself. Imagine that the probability of a human reproducing during his lifetime is 0.5. I am not sure what the real probability is, but certainly lower than that, just think about all the humans that never even reach the age of reproduction. Now, consider the last 500000 years, a plausible duration for the time we humans have been *Homo sapiens*, and a generation time of about 25 years. All just guesses here, but the actual numbers are not that important, and I am only using them for the sake of argument. Then just in the last 500000 years your ancestors must have reproduced 500000/25 times, i.e., 20000 times. Now track yourself back to your ancestor in the tribe of the first *sapiens*. Note that $0.5^{20000} \approx 0$, and it would have been even smaller had we not ignored the many millions of generations before we were humans, all the way back to some funny small unicellular animal that, against all odds, got lucky once! Therefore, none of us should be here, the probability that "Your descendants will successfully reproduce for the next 20000 generations" is zero, and so that sentence will almost for sure always be false. And yet, all of us are here proving it to have been true. One of the lessons of this apparent inconsistency is that conditional probabilities are hard to grasp, and often counter-intuitive.

As I hope to be able to show you, the notions of truth and falsity, or at least quite related concepts, are pervasive and actually fundamental for the understanding of statistics and its everyday use in our lives. And ironically again, the influence of statistics in our everyday lives is inversely proportional to what most people would be willing to admit or conceive. "Statistics is more important that Economics or Medicine" is a bold claim, one that most people would be fast at labelling as false. But Statistics would live well without the other two, while modern Economics or Medicine depend heavily on Statistics. In that sense, the sentence is probably true. Nowadays, increasingly so via artificial intelligence and machine learning - jargon terms that are mostly useful to impress those that do not really understand what these mean, but usually are just statistics in disguise – most of our actions are being controlled by a statistical model, from the way our GPS works to tell us how much time we will be late to our destination, to how much food we should be eating or how much money we should be saving, to how we should be acting upon the COVID19 pandemic. That is the hard truth. And that is an horrible feeling, because we must face the fact that we live in a world that is not black and white, but in the shades of gray lie, some might say, its dangers, and others, its beauty.

The last few years have brought us strange phenomena including the notion of alternative facts – one of the many strange ideas that became a new normal, and we can date these back to the science wars in the 1990's (see e.g. Mermin 2008) - which is a fancy way of saying the truth depends from where you stand, it is not an absolute property of a statement. Until recently, people would in general know what to believe in, and trust in scientists was to some extent indisputable. In the 20th century we started seeing people denying, supposedly based on *alternative sciences*, what is based on mainstream science. The earth is flat, vaccines are bad for you and should be avoided, cigarettes do not promote cancer, global warming is not real. These are all examples that I will not discuss, but that some believe in them represents the failure of a society, perhaps even the failure of a civilization. These are all false statements, widely recognized as such by the scientific community and the vast majority of humans. Nonetheless, still, a non-negligible portion of the humans on earth believes them to be true, typically grounded on some obscure conspiracy theory.

That is also the case of the pandemic we have seen rising in the last few months - the COVID-19. The news are filled with constant information about it, the number of cases, the number of exposed, the number of dead. The cure, the treatment, vacines. Together with the pandemic we saw a relatively new phenomena coined an infodemic. Dr. Tedros Ghebreyesus, WHO Director-General, used the term at the Munich Security Conference on Feb 15: “We’re not just fighting an epidemic; we’re fighting an infodemic” (Zarocostas, 2020) An infodemic occurs when there is so much information about a topic available that the average person can no longer distinguish what is true or false. However, beyond all that disinformation process, we still have Science, that is, we hope that Science evolves towards producing true knowledge. And while I believe that is fair to say it happens over the long run, the processes is not flawless. Truth is obtained when all the previous truths that were believed to be so are proven false. Hopefully, if the scientists are serious, follow due procedure, and observe a reasonable code of ethics, the progress over the long term is ensured. Every time an hypothesis is proven wrong we move a little bit forward in our quest for the true new knowledge. Interestingly, in the modern world, that procedure depends heavily on statistics, and is deeply rooted in the statistical concepts of null hypothesis significance testing, statistical significance and P-values. These terms, despite their central role in science in the XXth century, have been, for the last few decades, under attack by the same community that started by proposing them (e.g. Wasserstein & Lazar, 2016). In fact, it is perhaps surprising that P-values are the subject of "43 innovative and thought-provoking papers from forward-looking statisticians" in The American Statistician in 2019 (Wasserstein et al., 2019). At the very least, this should make us all think that there might something more than meets the eye, if a simple concept that the rest of us uses on a daily basis is still the focus of such heated discussions and debates.

Tools that are routinely used to allow researchers to distinguish between true and false claims apparently are not deemed to do so. Why is that the case? When set under such a setting we are facing a decision process under uncertainty, and under uncertainty there is always a probability that we will make the wrong decision. Figure 1 presents the typical setting under a null hypothesis significance testing scenario. The researcher defines an hypothesis, called a null hypothesis, that typically represents a null effect and that the researcher would like to falsify. Clearly, we can be correct, i.e. we can be led to believe in something that is true, under two scenarios. Either our null hypothesis is true, and we do not reject it, or our null hypothesis is false, and we reject the null hypothesis. Unfortunately, we can also be wrong under two scenarios. Either our null hypothesis is true, and we reject it, a type I error under statistical jargon, or our null hypothesis is false, and we fail to reject it, a type II error. We call the rejection of the null hypothesis a strong decision, in that we were able to collect enough information to disprove the null. The non-rejection of the null hypothesis is a weak decision, in that if the sample is small, you never have enough information to reject the null hypothesis, even it was false. Under those circumstances we say that our procedure lacks statistical power, i.e., the ability to reject the null when the null is false. The null hypothesis test proceeds with the collection of data, in the form of a random sample, and the computation of a test statistic based on said random sample. The test statistic fundamental property is that we know its distribution under the null hypothesis, i.e. assuming the null hypothesis was true. In other words, if the null hypothesis were true and we collected thousands of samples, and computed the test statistic for each of those, we would know the exact shape of the distribution these would present. And that is the trick. We collect a single sample, we compute the test statistic, and we compare it with what would be expected if the null were true. For a given significance level, which represents the risk of making a type I error we are willing to incur, we can define a threshold value beyond which we would reject the null hypothesis. If the value we observe for the test statistic is inconsistent with what we would expect, we reject the null. If it is consistent with what one might expect, we do reject it. A very important detail, strictly you should never accept the null or the alternative hypothesis, the statement is about whether you reject, or do not reject, the null; if you do not reject it, it could be just lack of power. The natural problem is that, just by chance alone, an extreme value of the test statistic might be observed even if the null is true, and a value that would be expected under the null being true could be observed even if the null is false. Figure 2 illustrates the distribution of the test statistic of a t-test based on a sample of size 30. The t-test is a parametric test that can be used to test if the mean of a distribution from which an available random sample was drawn, is equal to 0, or if two samples could have come from populations with the same mean, and is perhaps one of the most widely used statistical tests. The distributions shown are 10000 realizations of the test statistic when the null hypothesis is true (the true population was Gaussian, mean was 0, standard deviation 1), and additional 10000 when the null hypothesis is false (true population was Gaussian, but the mean was 0.7, standard deviation 1). While we can see that in most cases the correct decision is made (and this happens because in the simulated example the test presents a reasonable power), there are instances in which we still reach the wrong conclusion. And that happens even given that we did all by the book (well, most of all of it, see below). So, even the correct scientific procedure might lead one to believe in something that is false.

The recent provocative statement by Bishop (2020) comes to mind: "Just as lab scientists are not allowed to handle dangerous substances without safety training, researchers should not be allowed anywhere near a P value or similar measure of statistical probability until they have demonstrated that they understand what it means". (However, let's not forget, scientists do need to make decisions.). On our daily lives we have all been exposed to Hollywood movies where the exact analogous situation occurs. The accused is presumed innocent until proven guilty. And he should be considered innocent until proven guilty beyond reasonable doubt. Sending an innocent to jail is the equivalent of a type I error. And letting a guilty man go free is akin to type II error. As in courts, in null hypothesis significance testing, a type I error is considered worst than a type II error, and so we guard against it. We do so by setting a significance level. That is the equivalent to the level of evidence we require to be beyond reasonable doubt. As in the legal system, where jails are full of innocents and lots of guilty men walk free, it so happens in science with hypothesis and theories. The goal of science is also to make sure that over time less and less wrong/guilty ideas walk around freely making additional victims.

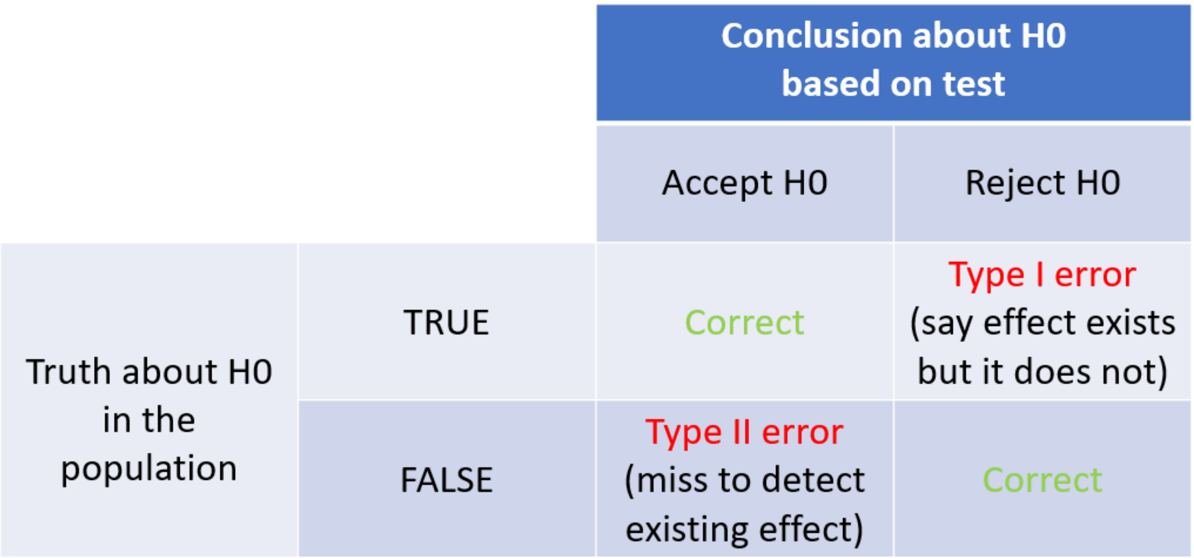
There is much debate nowadays about all the epistemological problems that are inherent to this procedure, but one that is worth keeping in mind is that no single statistical test should be taken as strong evidence *per se*. Since under the null hypothesis the distribution of the the P-value is uniform, even an extremely unlikely P-value value can be observed if the null is true. Hence, the only way for us to be sure that a given null hypothesis is truly false is through the test of time. If over time, over repeated experiments, we keep finding data inconsistent with the null hypothesis, i.e. extreme values for the test statistic, low P-values, then it is quite likely that the null hypothesis is false. If on the other hand, when we repeat the experiment, we no longer manage to find the effect, i.e. we are no longer able to reject the null hypothesis, we must consider the fact that our original result was a fluke, a spurious observation.

This poses complex problems for the way we do science, which have led to bold claims, namely that of Ioannidis (2005): “Most Published Research Findings Are False”. The discussion is ongoing, most scientific journals in most scientific areas have devoted space to the issue, and opinions divide themselves regarding the extent of the problem (e.g. Baker 2016, Fanelli 2018). But the arguments about why we are potentially made believe in facts that later are proven wrong is easy to follow. Consider our current situation, where we are facing a global pandemic. Every day we hear all sorts of claims in the news, from reasonable, on face value at least, or at least innocuous statements that Vitamin D provides protection against COVID, to perhaps some more dubious assertions that Hydroxychloroquine is a good treatment for COVID-19 (but see e.g. Horby et al. 2020), to ludicrous claims that eating cucumber is the solution (Fonseca et al. 2020). How do these make the headlines, and sometimes are even supported by sound scientific work. Note here I am providing an illustrative example, because all the claims above are not really based on reliable science and there is not enough time into the pandemic to find true falsities – it takes time to find these. But there might be claims out there currently made on reliable science that, in a few months or years, will be proved wrong. How does it happen?

Naturally, teams of researchers in most, if not all, countries in the world are devoting an extraordinary amount of time and money to study COVID-19. Therefore, they are looking for potential solutions, trying to identify treatments and identify risk and protection factors. So there are literally hundreds of teams conducting trials, looking at retrospective analysis, trying to identify real effects. Not surprisingly, given what I have described above, even in the absence of a real effect, just by using null hypothesis statistical testing, some teams will get spurious results. And that alone would be troublesome. Unfortunately, that effect is exacerbated by the fact that results are easier to publish if they report on the presence of effects. Put it in another way, it is hard to get negative results into high profile journals. The same in the media, such inherent bias for positive results is inevitable. Can you imagine the headline "Cucumber does not protect against COVID"? No one would care about it, no one would re-tweet it. Now consider "Cucumber provides protection against COVID". That is the kind of stuff that goes viral on the internet and hence it easily gets out of hand, even if the claim originated from a pre-print that clearly would not endure the test of a review by peers.

Science has its own strategies to ensure the quality of the knowledge that gets produced, and one of them is peer review. The gold standard under that context would arguably be double blind peer review, where reviewers do not know the authors they are evaluating, avoiding preconceived expectations about the work they are reviewing. Similarly, the authors do not know the reviewers, allowing the reviewers to provide feedback knowing they will not be hunted down personally even if they provide feedback that the authors disagree with. Unfortunately, for reasons that are not the scope of the current work, peer review is only the best of a number of non-ideal solutions. There are several reviews of issues associated with peer review (e.g. Kelly et al. 2014), but a specific problematic aspect is key to mention here. The peer review process takes time, and this creates a delay between the time the research is ready for being useful to others, and when others actually get to see it in print. This waiting time might be unacceptably long when the topic is urgent, as is happening during the current COVID-19 pandemic. And therefore, not surprising, many researchers have opted by producing pre-prints, a different type of publication form, popularized by the availability of online pre-prints repositories. While these have the advantage of making research available almost immediately, and have seen bursts of support, being widely regarded as positive in the academic community, I would dare to say that pre-prints should not be considered as reliable sources, except for researchers. Why? Because while a researcher in a given topic presumably is capable to distinguish between good work and something that should not be taken seriously, media outlets cannot do it themselves. And hence the likelihood of false news (or fake news!) being produced based on pre-prints increases enormously when compared to news based on peer reviewed papers. While this seems a rather obvious and hard to dispute argument in favour of peer review, peer review is not flawless, far from it. Certainly counter-examples can be made, with great papers coming out initially as pre-prints and horrible papers being published even in top peer reviewed journals. A famous example of the latter is the unfortunate analysis by Tatem et al. 2004 in *Nature*, suggesting that in less than 200 years, woman would run faster than men. It was no long after that said paper was criticized by a group of 16-18 year old students in the same high profile outlet, showing that such truth was simply false (wrong!). It was the consequence of committing several capital sins in statistics, including extrapolation of a linear model beyond the range of the data for a process that would certainly not be linear. In other words, the authors committed the capital sin in science of not turning the brain on before turning the computer on.

It is hard to define what science is. But I quite like an interesting definition found on the NASA web site for kids (https://spaceplace.nasa.gov/science/en/, accessed 8th August 2020). "Science is not just a tidy package of knowledge. Science is not just a step-by-step approach to discovery. Science is more like a mystery inviting anyone who is interested to become a detective and join in the fun". To which I would add that, in that detective search the goal is to find the truth. The falsities around the truth are the equivalent to the darkness that keeps us from seeing it. Each time you disprove an hypothesis, you reject it, you establish it as false, truth becomes more visible. And while that process is far from straightforward, over the long run it is undeniable the trend is positive. The truth is out there. Collectively as a whole, every day that passes we enlighten ourselves and we know more about the world around us.



		Conclusion about H0 based on test	
		Accept H0	Reject H0
Truth about H0 in the population	TRUE	Correct	Type I error (say effect exists but it does not)
	FALSE	Type II error (miss to detect existing effect)	Correct

Figure 1 – Possible scenarios when deciding about whether to reject or not reject a null hypothesis under a null hypothesis significance test. Under uncertainty, we hope to get decisions that are correct, but we risk to reach false conclusions. The statistical test is designed to minimize the probability of errors, but there is a natural trade-off between type I and type II errors.

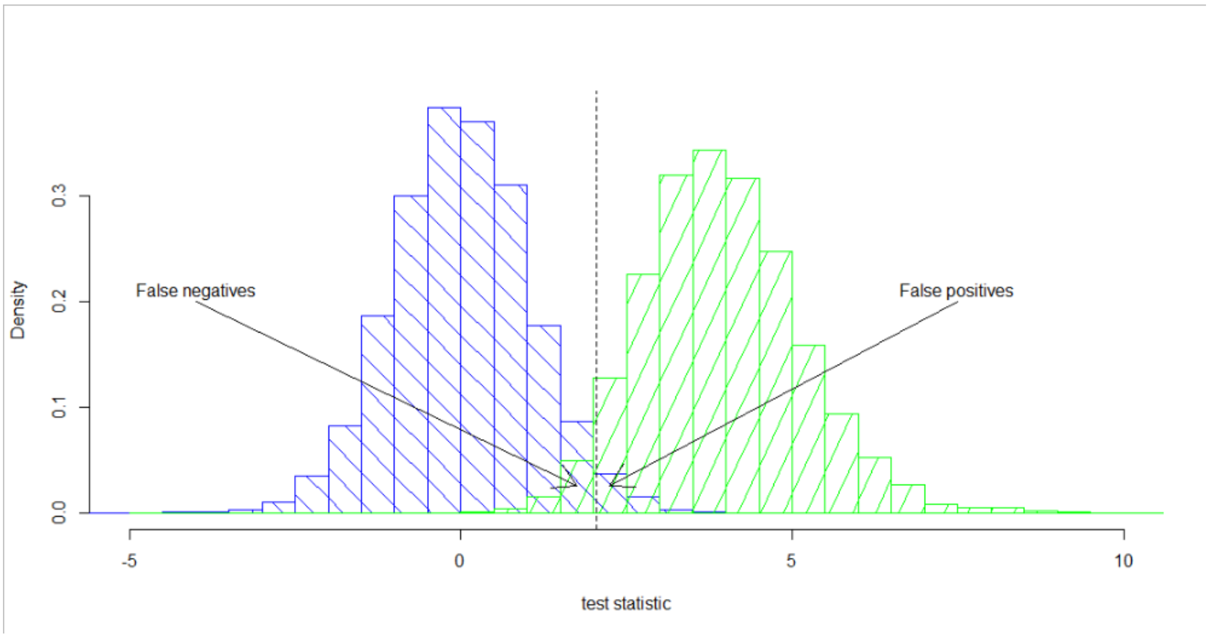


Figure 2 – Ten thousand realizations of a t-test when the null hypothesis is true (blue) and when it is false (green). The vertical dashed line represents the value beyond which we reject the null hypothesis. In other words, a correct decision is reached for a blue value for the left of the vertical dashed line, and for a green value

to the right of the dashed line. Green values below the dashed line and blue values above the dashed line are errors, false negatives and false positives, respectively. In a statistical test we typically set the significance value, which leads to a given dashed line, and we do so to control the number of false positives (or type I errors). As a consequence, and given the sample size, the variance of the process and the effect size, the quantity of false negatives is also set up.

References

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility *Nature* 533: 452-454

Bishop, D. 2020. How scientists can stop fooling themselves over statistics *Nature* 584: 9-9

Fanelli, D. 2018. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences* 115: 2628-2631

Fonseca, S. C.; Rivas, I.; Romaguera, D.; Quijal-Zamorano, M.; Czarlewski, W.; Vidal, A.; Fonseca, J. A.; Ballester, J.; Anto, J. M.; Basagana, X.; Cunha, L. M. & Bousquet, J. 2020. Association between consumption of vegetables and COVID-19 mortality at a country level in Europe. MedRXiv https://doi.org/10.1101/2020.07.17.20155846

Horby, P.; Mafham, M.; Linsell, L.; Bell, J. L.; Staplin, N.; Emberson, J. R.; Wiselka, M.; Ustianowski, A.; Elmahi, E.; Prudon, B.; Whitehouse, A.; Felton, T.; Williams, J.; Faccenda, J.; Underwood, J.; Baillie, J. K.; Chappell, L.; Faust, S. N.; Jaki, T.; Jeffery, K.; Lim, W. S.; Montgomery, A.; Rowan, K.; Tarning, J.; Watson, J. A.; White, N. J.; Juszczak, E.; Haynes, R. & Landray, M. J. 2020. Effect of Hydroxychloroquine in Hospitalized Patients with COVID-19: Preliminary results from a multi-centre, randomized, controlled trial. MedRXiv https://doi.org/10.1101/2020.07.15.20151852

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2: e124

Kelly, J., Sadeghieh, T., & Adeli, K. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *EJIFCC 25*: 227–243

Mermin, N. D. 2008. Science wars revisited. *Nature* 454: 276-277

Tatem, A. J.; Guerra, C. A.; Atkinson, P. M. & Hay, S. I. 2004. Momentous sprint at the 2156 Olympics? *Nature* 431: 525-525

Wasserstein, R. L. & Lazar, N. A. 2020. The ASA's statement on pvalues: context, process, and purpose. *The American Statistician 70*: 129-133

Wasserstein, R. L.; Schirm, A. L. & Lazar, N. A. 2019. Moving to a world beyond “p<0.05". *The American Statistician*, 73: 1-19

Zarocostas, J. 2020. How to fight an infodemic. *The Lancet* 395: 676

Acknowledgements

I would like to thank the comments by Charles Paxton as well as Kairos editorial team that greatly improved the manuscript. I also thank partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020).