<https://www.marugroup.net/insights/blog/danger-of-relying-on-statistical-significance> (accessed on 20/11/2021)

**The Danger of Relying on “Statistical Significance”**

By Andrew Grenville, Chief Research Officer | June 3, 2019

“Is that statistically significant?”

This question sounds scientific and  well-meaning, but is too often misguided and dangerous according to [leading statisticians](https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108?scroll=top&needAccess=true) [1]. The popular misconception that being *statistically significant* is the be-all and end-all of truth has many unintended consequences. The biggest problem is that often what appears to be fact is simply statistical noise, and what may well be fact is sadly ignored.

The problem is not so much with the statistical tools, but rather in how they are applied. You’ve probably seen data tables in which every column is compared to every other column, usually at a 95% confidence interval. “Significant” differences are noted with a letter or number. Each table can have 100 or more tests. Researchers often scour the tables looking for “significant” differences and then try to explain the differences. This is where the trouble begins.

Most people are familiar with the idea that one in twenty of these tests will yield a false positive—indicating there is a difference when none exists. But the reality is the error rate is actually much higher, particularly when there are multiple comparisons involved and the tools are used in ways they were never intended.

“Most scientists would look at…(a) p-value of 0.01 and say that there was just a 1% chance of his result being a false alarm. But they would be wrong”, [Regina Nuzzo wrote in Nature](https://www.nature.com/news/scientific-method-statistical-errors-1.14700#/cause) [2]. “The p-value cannot say this: all it can do is summarize the data assuming a specific null hypothesis. It cannot work backward and make statements about the underlying reality. That requires another piece of information: the odds that a real effect was there in the first place.”

She goes on to say “According to one widely used calculation [3], a p-value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a p-value of 0.05 raises that chance to at least 29%.”

Replication crisis in academia

It has been the convention to use a 95% confidence interval (p ≤ 0.05) as a marker of truth not only in the world of market research but in the scientific community in general. Journals are generally not interested in publishing studies in which the results are not “significant.” And academics need to publish to get promoted or even stay employed. The [result](https://arstechnica.com/science/2009/09/for-clinical-trials-design-and-results-dont-always-match/) [4] has been a disaster for science.

There is plenty of evidence to suggest that many published papers have findings that cannot be reproduced because of a misuse of statistical testing. The journals Nature and Science are very prestigious, and academics clamor to have their papers accepted by them. But a study published in nature entitled “Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015” [5] reported they were able to replicate six in ten of them, and “the effect size of the replications is on average about 50% of the original effect size.”

Similarly, dismal measures of reproducibility and replicability have been reported in many other fields including [psychology](https://science.sciencemag.org/content/349/6251/aac4716)[6], [economics](https://science.sciencemag.org/content/351/6280/1433)[7], and [medicine](https://www.nature.com/articles/nrd3439-c1)[8]. One high profile casualty of a mindless focus on “significant findings” is American researcher and professor [Brian Wansink](https://en.wikipedia.org/wiki/Brian_Wansink)[9].

His research focused on how people make food choices and he is the author of [*Mindless Eating*](https://www.amazon.ca/Mindless-Eating-More-Than-Think/dp/0345526880)[10] and [*Slim by Design*](https://www.amazon.ca/Slim-Design-Mindless-Solutions-Everyday/dp/0062136526/ref=pd_cp_14_2?pd_rd_w=3Mz53&pf_rd_p=64a84aa4-b549-4445-9bf3-48bdeed5f24a&pf_rd_r=1KSYXSAZ3DPPT2YRE1R5&pd_rd_r=6fa712a0-70e6-11e9-b88a-83630affe59b&pd_rd_wg=c8nYE&pd_rd_i=0062136526&psc=1&refRID=1KSYXSAZ3DPPT2YRE1R5)[11].His work popularized the ideas that plate size and color influence how much you eat and that 100 calorie packages reduce the amount overweight people eat. [Many problems were later discovered](https://arstechnica.com/science/2017/04/the-peer-reviewed-saga-of-mindless-eating-mindless-research-is-bad-too/?utm_source=pocket&utm_medium=email&utm_campaign=pockethits)[12] with Wansink’s work, but what got people looking into his work was his uncritical use of statistical tests, or “[p-hacking](https://www.explainxkcd.com/wiki/index.php/882:_Significant)[13].” He was caught out after publicly encouraging graduate students and collaborators to troll through data sets looking for “statistically significant” findings, rather than following the scientific process of testing predetermined hypotheses.

According to [Tim Vanderzee](https://scholar.google.nl/citations?user=vsuJSN0AAAAJ&hl=nl)[14], a researcher who [investigated Wansink’s work](http://www.timvanderzee.com/the-wansink-dossier-an-overview/)[15], there are alleged problems with 52 of his publications, which have been cited over 4,000 times in 25 different journals and in 8 books. When evidence of his unscientific approach surfaced, his university suspended him from teaching and ultimately released him. Misuse of statistical testing can have dire effects.

As researchers we seek to inform and guide decision making. Uncritical use of stats testing can result in us misleading and misinforming instead. That’s a situation no one wants. So, what’s a researcher to do? Fortunately, the American Statistical Association (ASA) has some advice.

On the use and abuse of significance testing

The ASA put out a formal [statement](https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.XNHR245Kg2z)[1] on the use and misuse of p-values. They set forth “principles underlying the proper use and interpretation of the p-value.” They state, “The widespread use of ‘statistical significance’ (generally interpreted as “p ≤ 0.05”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process. A conclusion does not immediately become ‘true’ on one side of the divide and ‘false’ on the other.”

They suggest “Researchers should recognize that a p-value without context or other evidence provides limited information.” They counsel that researchers “should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.” They conclude that good statistical practice emphasizes “understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.” Context is truly the anecdote.

Statisticians are also warning against thinking too narrowly about what a p-value can reveal. [A comment in Nature](https://www.nature.com/articles/d41586-019-00857-9)[16] by Valentin Amrhein, Sander Greenland, Blake McShane, and more than 800 signatories points to the reduction of a p-value to a significant/not significant dichotomy as a big part of the problem. They write “we are not advocating a ban on p values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.”

“One reason to avoid such ‘dichotomania’ is that all statistics, including p-values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree. In fact, random variation alone can easily lead to large disparities in p values, far beyond falling just to either side of the 0.05 threshold. For example, even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving p < 0.05, it would not be very surprising for one to obtain p < 0.01 and the other p > 0.30.”

They suggest “[t]he trouble is human and cognitive more than it is statistical: bucketing results into ‘statistically significant’ and ‘statistically non-significant’ makes people think that the items assigned in that way are categorically different.”

That kind of categorical thinking makes it easy to miss real differences and focus on false ones. A more nuanced approach to judging what is “significant” is clearly needed.

How did we get into this mess?

If our reliance on over-simplified “significance” testing is a problem, how the heck did we end up here? Incredibly, the story involves two statisticians who despised each other and ended up with their ideas mashed together in an unholy alliance that neither approved of. And it all started with a cup of tea in 1920’s England.

A group of academics had gathered for tea. One was Dr. Blanche Bristol who, when offered a cup of tea by a colleague, turned it down. The trouble was the man poured the tea into the cup first, then added the milk. Dr. Bristol rejected it because she preferred the milk to be poured into the cup first, and the tea afterward. The man who had poured the tea suggested she surely could not tell the difference. She insisted she could. The man, Dr. Ronald Aylmer Fisher, proposed a test, which he famously described in his book [*The Design of Experiments*](https://www.phil.vt.edu/dmayo/PhilStatistics/b%20Fisher%20design%20of%20experiments.pdf)[17]. He would prepare eight cups of tea; four with the tea poured first and four with the milk poured first. She had to guess which was which.

He proposed the null hypothesis that she would be unable to do that correctly. Fisher calculated that her chance of guessing all cups correctly was 1/70. He was provisionally willing to concede her ability (rejecting the null hypothesis) in this case only. She, reportedly, got them all correct. The null hypothesis was rejected. This was the beginning of significance testing.

Meanwhile, two statisticians, Jerzy Neyman and Ergon Pearson, were working on hypothesis testing – selecting among competing hypotheses based on the experimental evidence alone. Neyman suggested that hypothesis testing was an improvement on significance testing. That did not sit well with Fisher, who already disliked Neyman because he had worked with Pearson’s father with whom Fisher had a long-running disagreement. They battled over which way was better until Fisher’s death.

But in the meantime, something funny happened. Gerd Gigerenzer sums it up rather nicely in his (wryly caustic) paper [Statistical Rituals: The Replication Delusion and How We Got There](https://journals.sagepub.com/doi/10.1177/2515245918771329)[18]: “Early textbook writers struggled to create a supposedly objective method of statistical inference that would distinguish a cause from a chance in a mechanical way, eliminating judgment. The result was a shotgun wedding between some of Fisher’s ideas and those of his intellectual opponents, the Polish statistician Jerzy Neyman (1894–1981) and the British statistician Egon S. Pearson (1895–1980). The essence of this hybrid theory is the null ritual.”

He describes what he calls the “null ritual” this way:

“1. Set up a null hypothesis of ‘no mean difference’ or ‘zero correlation.’ Do not specify the predictions of your own research hypothesis.

2. Use 5% as a convention for rejecting the null hypothesis. If the test is significant, accept your research hypothesis. Report the test result as p < .05, p < .01, or p < .001, whichever level is met by the obtained p-value.

3. Always perform this procedure.”

“The null ritual does not exist in statistics proper”, Gigerenzer continues. “This point is not always understood; even its critics sometimes confuse it with Fisher’s theory of null-hypothesis testing and call it ‘null-hypothesis significance testing.’ In fact, the ritual is an incoherent mishmash of ideas from Fisher on the one hand and Neyman and Pearson on the other, spiked with a characteristically novel contribution: the elimination of researchers’ judgment.”

The way forward

We find ourselves in a situation where uncritical use of a bastardized test has lead to a “replication crisis” in science and the misuse and abuse of the notion of significance. We need to rethink how we use significance testing.

Firstly, we need to take a finding of “significance” with a grain of salt. The error rate is larger than is commonly assumed. Secondly, we need to stay away from data fishing; trolling for “significant” differences. Thirdly, we need to understand that something which might not pass the significance test might be meaningful. Fourthly, we need to consider context. As the ASA recommends, it is essential to take into account “understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.”

So next time someone asks: “Is that statistically significant?” we should consider if that is the right question. A better question would be “are the differences meaningful?”

Asking this would steer us away from the trap of binary thinking about a pair of numbers, and guide us toward contextualizing a finding in a world rich with information.

**References:**

[1] The ASA Statement on p-Values: Context, Process, and Purpose, <https://doi.org/10.1080/00031305.2016.1154108>

[2] Nuzzo, Regina (2014), Scientific method: Statistical errors, Nature, volume 506, 150–152.

[3] Goodman, S. N. (2001), Of P-Values and Bayes: A Modest Proposal, Epidemiology, volume 12, 295–297.

[4] [Anti-“publication bias” efforts not panning out for science | Ars Technica](https://arstechnica.com/science/2009/09/for-clinical-trials-design-and-results-dont-always-match/)

[5] Camerer, Colin F. et al. (2018), Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015, Nature Human Behaviour volume 2, pages637–644.

[6] Evaluating replicability of laboratory experiments in economics <https://www.science.org/doi/10.1126/science.aac4716>

[7] Evaluating replicability of laboratory experiments in economics <https://www.science.org/doi/10.1126/science.aaf0918>

[8] Believe it or not: how much can we rely on published data on potential drug targets? <https://www.nature.com/articles/nrd3439-c1>

[9] <https://en.wikipedia.org/wiki/Brian_Wansink> (accessed on 29/11/2021).

[10] <https://www.amazon.ca/Mindless-Eating-More-Than-Think/dp/0345526880>

[11] <https://www.amazon.ca/Slim-Design-Mindless-Solutions-Everyday/dp/0062136526/ref=pd_cp_14_2?pd_rd_w=3Mz53&pf_rd_p=64a84aa4-b549-4445-9bf3-48bdeed5f24a&pf_rd_r=1KSYXSAZ3DPPT2YRE1R5&pd_rd_r=6fa712a0-70e6-11e9-b88a-83630affe59b&pd_rd_wg=c8nYE&pd_rd_i=0062136526&psc=1&refRID=1KSYXSAZ3DPPT2YRE1R5>

[12] <https://arstechnica.com/science/2017/04/the-peer-reviewed-saga-of-mindless-eating-mindless-research-is-bad-too/?utm_source=pocket&utm_medium=email&utm_campaign=pockethits>

[13] <https://www.explainxkcd.com/wiki/index.php/882:_Significant>

[14] <https://scholar.google.nl/citations?user=vsuJSN0AAAAJ&hl=nl>

[15] <https://www.timvanderzee.com/the-wansink-dossier-an-overview/>

[16] Amrhein, Valentin et al. (2019), Scientists rise up against statistical significance, <https://www.nature.com/articles/d41586-019-00857-9>

[17] Fisher, R.A. (1935), *The Design of Experiments*, Edinburg and London, Oliver and Boyd.

[18] Gigerenzer, Gerd (2018), Statistical Rituals: The Replication Delusion and How We Got There, <https://doi.org/10.1177/2515245918771329>