IJABT (2013), 1(1):1-6

The Significance of "Statistical Significance" in Modern Medical Statistical Analyses

Jacob Levman[1,2]

1 Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
jacob.levman@eng.ox.ac.uk
2 Imaging Research, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Canada

ABSTRACT:
When a scientist performs an experiment they normally acquire a set of measurements and are expected to demonstrate that their results are "statistically significant" thus confirming whatever hypothesis they are testing. The main method for establishing statistical significance involves demonstrating that there is a low probability that the observed experimental results were the product of random chance. This is typically defined as $p < 0.05$, which implies there is less than a 5% chance that the observed results occurred randomly. This research study demonstrates to the general medical scientist that the commonly used definition for "statistical significance" can erroneously imply a significant finding. Qualitatively insignificant results can yield "statistically significant" findings at moderately large sample sizes which are very common in modern medical scientific literature. Achieving statistical significance without qualitatively significant results can lead scientists to erroneous conclusions with respect to the value of their own work. This can contribute to the advancement of medical treatments and technology that don't warrant further investigation, but persist due to a scientist's overconfidence in statistical testing.

Keywords: Statistical significance, hypothesis testing, sample size, p-value, t-test

INTRODUCTION

Establishing statistical significance is extremely common in modern medical research. This involves demonstrating that there is a low probability that a scientist's observed measurements were the result of random chance (typically referred to as a deviation from the null hypothesis). Statistical significance is typically defined as ($p<0.05$) and a statistically significant finding based on this definition implies that there is less than a 5% chance that the observed differences were the result of random chance. Statistical significance can be established using a wide variety of statistical tests which compare a scientist's measurements with randomly generated distributions to determine a p-value (from which statistical significance is established). As the number of samples in a modern experiment increases, the amount of difference needed between two distributions of measurements in order to obtain statistical significance ($p<.05$) gets smaller. The main focus of this research paper is to take a unique approach to demonstrating that as the number of samples becomes large, the amount of separation between our two groups needed to obtain 'statistical significance' becomes negligible. Thus when dealing with large sample sizes scientists have an extremely low threshold for obtaining statistical significance. In its most extreme form, a "statistically significant" effect is in fact qualitatively insignificant.

In this study we have elected to perform statistical testing using the widely accepted and established two-sample t-test [1]. It should be noted that the t-test was developed in a beer factory in 1908 over one-hundred years ago by a scientist writing under a false name (Student). This was long before the advent of computers, thus long before a scientist had the ability to perform statistical testing on groups of data with large numbers of samples. The original introduction of the t-test [1] provided look-up tables to assist in statistical calculations that allow the researcher to perform analyses on data groups with up to only 10 samples. In those days it was unreasonable for someone to manually compute p-values on hundreds or thousands of samples. In the present research environment a journal paper reviewer is likely to require many more than 10 samples from a typical scientist's experiment, thus inadvertently lowering the bar for obtaining the desired "statistical significance".

In standard modern scientific thought it is assumed that adding more samples to our experiments will make the results of our statistical analyses more reliable. This paper takes an original approach to demonstrating that as sample sizes increase, the amount of difference required between a medical scientist's control and experimental groups in order to achieve statistical significance gets smaller. This indicates that a study's statistical results aren't necessarily more reliable with increased samples when the scientist's conclusions are based on achieving statistical significance ($p<.05$). Instead, more samples can contribute to making it

extremely easy for our experiments to achieve statistical significance. This effect should make the modern medical scientist whose experiment includes many samples less confident with respect to the true significance of their studies that achieve statistical significance ($p<.05$).

Although this study's conclusions affect many methods for obtaining a p-value from a statistical metric, the focus of this paper's experiments is the two sample t-test, one of the most widely used statistical methods for obtaining a p-value [1]. After reading this study it should be clear to the modern medical scientist that the t-test was developed for another era and alternative techniques would benefit the modern scientist. As Dr. Rozeboom wrote in 1960 "The statistical folkways of a more primitive past continue to dominate the local scene" [2]. Rozeboom was writing about problems with statistical testing over 50 years after the t-test was first created. It is now 50 years after Rozeboom wrote this commentary and his words are still as relevant as ever. There have been many examples of critiques of how statistical significance testing and null hypothesis testing is performed [2-20], yet despite the many shortcomings highlighted, performing hypothesis testing based on a p-value threshold ($p<.05$) is still one of the most common statistical techniques used in modern medical science.

Common scientific thought has it that if we increase our number of samples, then the computed statistical p-value becomes more and more reliable. However, as we add more and more samples, the amount of separation needed between our groups to achieve statistical significance gets smaller. This is because the p-value computations are based on random data. Once the number of samples becomes very large, the amount of overlap observed between large randomly generated distributions will always be large, leading to very little separation required between the two distributions to achieve a p-value below 0.05. Or put another way, we have a threshold for statistical significance that is so low that (as long as we have an adequate number of samples) all we need is to have two noisy signals that are marginally dissimilar from each other in order to achieve "statistical significance" ($p < 0.05$).

This low threshold for achieving statistical significance has the potential to greatly affect a scientist's approach to their experiments. As scientists, our career prospects (and thus our prestige and personal finances) are heavily dependent on our accumulation of peer-reviewed journal papers. This personal motivation biases us towards getting our research accepted for publication. Since it is extremely common for a journal paper reviewer to require that our experimental results be tested for statistical significance, we are generally biased towards finding statistical significance in our experiments in order to accumulate journal publications and to succeed in our careers. The word 'significant' is qualitative and subjective. Whether something is 'significant' is in the eye of the beholder. When we add the word 'statistics', we add a word with strong quantitative implications to the very qualitative word 'significant'. This lends an appearance of credibility and certainty to any experiment that achieves a p-value below 0.05, simply because this is the widely accepted threshold for achieving 'statistical significance'.

Since statistical significance is based on random distributions, performing hypothesis testing on the p-value calculation ($p<.05$) is like asking the question: did our experiment do better than 95% of randomness? Since the vast majority of medical scientists are likely to have constructed their experiments in a somewhat logical manner, they are generally liable to do at least a little better than random chance. Thus scientists are highly likely to find statistical significance in their own experiments, especially if they perform their experiments with many samples. This study is designed to illustrate that achieving a statistically significant ($p<.05$) difference between two groups (experimental and control) at moderately large sample sizes is possible with experimental data that scientists would qualitatively describe as insignificant.

MATERIALS AND METHODS

A randomized simulation trial was performed generating 30000 random distribution pairs of a variable number of samples (2 to 1000) with each group distributed normally. This was performed with Matlab (Mathworks, Natick, MA, USA). The relationship between the number of samples included in the groups being compared by the two sample t-test and the amount of separation needed between the two groups in order achieve a barely statistically significant result ($p<0.05$) was tracked. This was performed by comparing two randomly generated normally distributed groups and computing p-values for each of the 30000 pairs analyzed. The pair of distributions whose p-value is highest while still being below 0.05 was selected for further analysis. Thus this paper looks at example distributions that are barely statistically significant - as the term is used ($p<.05$). This allows us to focus on statistically significantly different distribution pairs that are very close to the standard threshold for statistical significance, thus allowing us to analyze how much difference is required between our experimental and control groups to achieve statistical significance at varying sample sizes. The distance between the two randomly generated distributions' mean values as measured in standard

deviations is tracked as the number of samples in our study is varied.

P-values are computed from lookup tables which are created from randomly generated distributions of data. A secondary aspect of this research study's methods is to visually illustrate how much difference is required between two groups of numbers in order to achieve the standard definition of statistical significance ($p<.05$) at a variety of sample sizes. This was accomplished by generating large amounts of normal (Gaussian) random distributions and comparing them with the two-sample t-test. For this section of the analysis, 1000 pairs of random distributions were created at each example sample size provided. Of all the randomly generated cases, the pair that exhibit the highest p-value below 0.05 is selected for presentation as a visual example of how much separation is needed between two groups of data in order to achieve 'statistical significance' at the given sample size. Random distributions were generated across a wide variety of group sample sizes that accommodate presentation in image form, where the image's dimensions are expressed as a factor of 2 (4, 16, 64, 256, 1024, 4096, 16384, 65536 and 262144 samples in the example distributions presented). The variance in these noise pairs demonstrates how the amount of separation between two barely statistically significantly different groups changes as the number of samples is varied.

All statistical significance testing was performed using one of the most common statistical tests available, the two-sample t-test. This was selected so that our statistical testing method matches the type of distributions being randomly generated (Gaussian noise / normal distributions). In addition, for each sample size setting, the number of randomly created distributions that have a p-value below 0.05 are enumerated in order to confirm this experiment's results (we should expect about 5% of the randomly generated cases to be statistically significant $p<.05$). All random normal (Gaussian) distributions were created using the mathematical and statistical package Matlab (Mathworks, Natick, MA, USA). Statistical testing was performed with the established two-sample t-test provided in Matlab.

RESULTS

Figure 1 provides a plot comparing variations in the amount of separation required between our two groups (as measured in standard deviations) and how those values vary with increasing sample size. Pairs of randomly generated distributions with p-values just below 0.05 are included as visual aids to further help illustrate the effects described in this study. Figure 2 demonstrates pairs of randomly generated statistically significantly different normal distributions in each row with 4 samples (2x2, top row), 16 samples (4x4, middle row) and 64 samples (8x8, bottom row). Figure 3 demonstrates randomly generated pairs of statistically significantly different normal distributions in each row with 256 samples (16x16, top row), 1024 samples (32x32, middle row) and 4096 samples (64x64, bottom row). Figure 4 demonstrates randomly generated pairs of statistically significantly different normal distributions in each row with 16384 samples (128x128, top row), 65536 samples (256x256, middle row) and 262144 (512x512, bottom row). Table 1 presents the p-values of each of the pairs selected for viewing in figures 1, 2 and 3 as computed by Matlab's two-sample t-test. Table 1 also presents the total number of randomly generated distributions which achieved a statistically significant difference as the term is typically defined ($p<.05$), using the popular and well established two-sample t-test. Since the experiment involves creating 1000 randomized distributions we expect to find approximately 50 (5%) of those samples being statistically significant ($p<.05$). The results from each trial were confirmed to be close to 50 samples achieving statistical significance out of each 1000 randomly created distribution pairs.

When examining each noise image pair, a scientist can interpret the two visual image distributions as being very close to the threshold for obtaining statistical significance at the given number of samples. If the sample pairs provided in figures 2 through 4 were slightly less different from each other then the pairs would not be statistically significant ($p<.05$). Note that the difference between two statistically significantly different distributions gets smaller as the number of samples increases. All image pairs in each row of figures 2 through 4 represent statistically significantly different data as is typically defined ($p<.05$).

DISCUSSION

Figure 1 demonstrates that as the number of samples in our statistical analysis increases, the amount of separation (as measured in standard deviations) required between our experimental and control groups gets smaller. On the right side of the plot it is clear that extremely little separation is needed between the two groups being compared in order achieve statistically significant results ($p<.05$). With group sizes of just 100 samples (or more) statistical significance ($p<.05$) is obtained with distributions whose means differ by less than 2% of their standard deviations. Thus statistical testing (in this case with the two-sample t-test) can yield statistically significant results when comparing two distributions that are only marginally dissimilar.

Figures 2 through 4 provide example images of randomly generated data where each row of each figure represents a statistically significant finding

when comparing the left and right image. It can be seen from the results that at a p-value just below 0.05, the two randomly generated groups of 4 samples each are substantially different from each other as the image on the right is clearly darker overall than the image on the left (see figure 2 top line). At 16 and 64 samples, it is apparent that the random image on the left is darker than the one on the right although it is clear that the 64 sample images are substantially more similar to each other than the images with 16 or 4 samples. All of the statistically significant pairs presented in Figure 2 appear qualitatively significantly different from each other. Once the size of the images has been increased to just 256 samples it becomes challenging to see a significant difference between the two distributions, even though the results displayed are statistically significant ($p = 0.0493$) as the term is traditionally used (see figure 3 top line). When comparing two distributions with more than 256 samples, the distributions appear qualitatively insignificantly different from each other despite having obtained "statistical significance" (see figures 3 and 4).

Data was also included demonstrating that approximately 5% of the randomly generated samples created for figures 2 through 4 in this experiment achieve statistical significance (as the term is typically defined $p < .05$). This is presented in the final column of Table 1. This information is provided to simply demonstrate that the experiment is matching expectations – that about 5% or about 50 out of 1000 randomly created distributions have a p-value below 0.05.

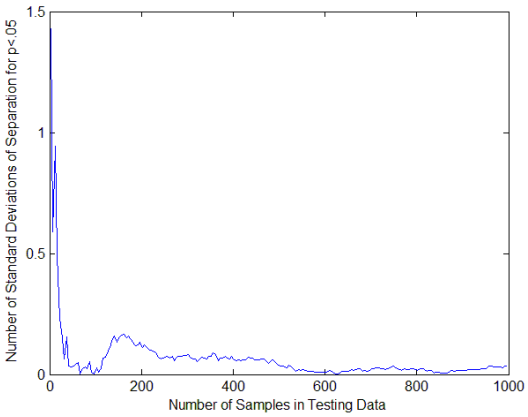

Fig. 1 The tradeoff between how much separation is required between two groups (as measured in standard deviations) in order to achieve statistical significance ($p<.05$) at varying sample sizes.

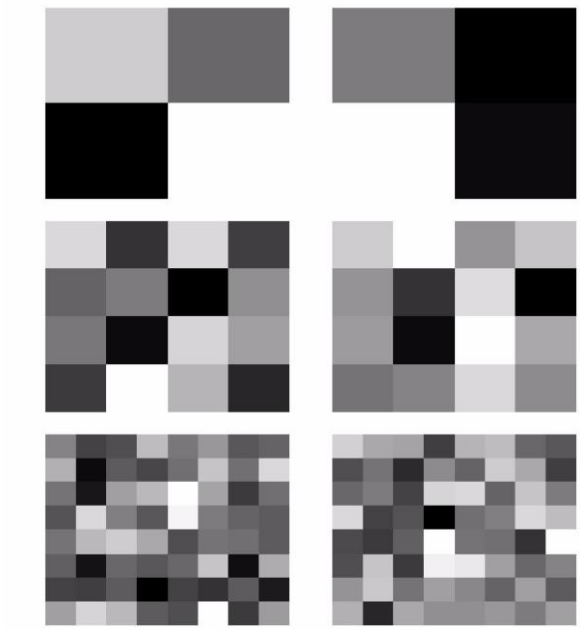

Fig. 2 Randomly generated pairs of statistically significantly different ($p<.05$) distributions with 4 samples (top row), 16 samples (middle row) and 64 samples (bottom row).

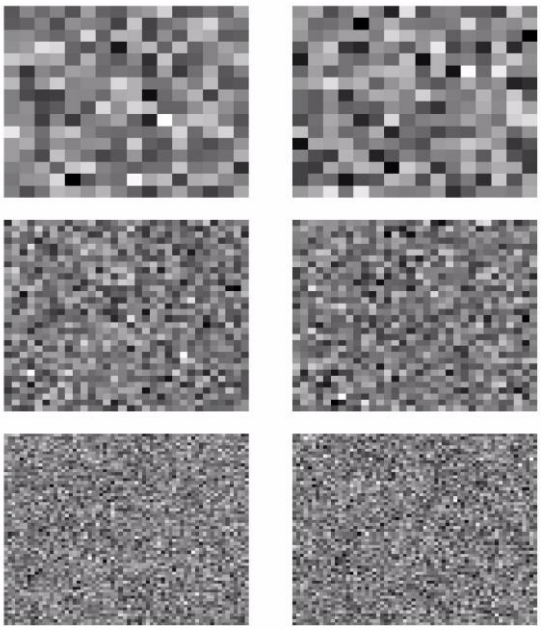

Fig. 3 Randomly generated pairs of statistically significantly different ($p<.05$) distributions with 256 samples (top row), 1024 samples (middle row) and 4096 samples (bottom row).

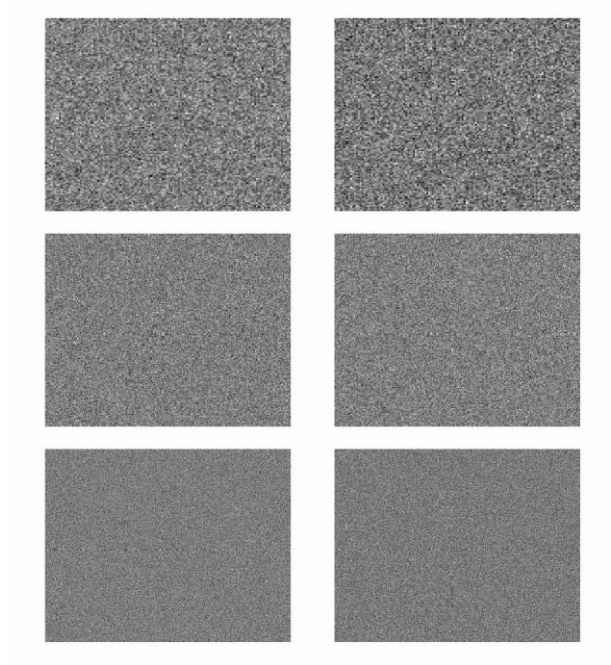

Fig. 4 Randomly generated pairs of statistically significantly different (p<.05) distributions with 16384 samples (top row), 65536 samples (middle row) and 262144 samples (bottom row).

Table 1. P-values and Associated Data for the Randomly Generated Distributions of Figures 1, 2 and 3

Random Distribution Size	P-value of Pair Presented in Figures Above	Number of Random Cases with p < 0.05
2x2=4	0.0499	51/1000
4x4=16	0.0459	44/1000
8x8=64	0.0488	48/1000
16x16=256	0.0493	49/1000
32x32=1024	0.0498	54/1000
64x64=4096	0.0499	56/1000
128x128=16384	0.0483	46/1000
256x256=65536	0.0485	42/1000
512x512=262144	0.0496	42/1000

P-value computations from single sample statistical tests are intuitive: a new single sample can be compared against the pre-existing group and the 0.05 p-value threshold causes only those samples that fall on the outskirts of the distribution to be considered statistically significantly different. The same does not hold once we move to two-sample statistical testing. If we generate two random groups of data with each group containing many samples, then it is inevitable that the two groups will overlap each other substantially. Even in the 5% of cases where the two large random distributions are most dissimilar, we will still find highly overlapping distributions as demonstrated in this paper's results. This has the effect of setting the bar for obtaining statistical significance in our experiments extremely low (especially when the number of samples is large) and may have led researchers to conclude a significant effect from their experimental results when in fact the effect observed is much smaller or possibly even non-existent. Achieving statistical significance (p<.05) merely demonstrates that the experimental results outperformed 95% of the randomly generated noise from which the p-value is computed. Ascribing 'significance' to any experiment is a subjective task which should be evaluated by whomever is interested in examining the experiment, not by a single p-value computation.

This study's findings are potentially of broad interest to scientists in general and especially to medical scientists who should avoid coming to conclusions regarding the significance of their research based on their experiment's computed p-values.

Figure 3 (top line) demonstrates that achieving statistical significance (p<.05) on groups with only 256 samples only confirms the existence of an extremely marginal effect. Scientific studies based on at least a couple hundred samples in each group are extremely common in the biomedical literature. The results presented in this paper are of particular importance in medical research where over-confidence in a study's statistically significant results can lead to the advancement of drugs, medical interventions and procedures whose results don't warrant further investigation due to only marginal (though statistically significant) differences between the control and experimental population.

Establishing statistical significance is often a prerequisite for publication of a scientific study. Scientists who find statistical significance in experiments containing many samples haven't actually demonstrated that their findings are qualitatively significant at all (unless they've included well separated confidence intervals). It is doubtful that anyone would qualitatively describe the pairs of results presented in figures 3 and 4 as significantly different from each other even though they meet the normal criteria for statistical significance (p<.05).

Establishing statistical significance with a p-value below 0.05 provides us with an answer to the question "did we outperform 95% of randomness?" But outperforming randomness is a very low bar to set for ourselves, thus ensuring that scientists who work with reasonably large sample sizes will be able to go on

finding statistically significant ($p<.05$) results (almost) wherever they look for them.

ACKNOWLEDGEMENTS

This work was supported by the Canadian Breast Cancer Foundation.

REFERENCES

[1] Student S. (1908), The Probable Error of a Mean. Biometrika, 6(1):1-25.

[2] Rozeboom W.W., (1960) The fallacy of the null hypothesis significance test. Psychological Bulletin 57:416-428.

[3] Wilhelmus K., (2004) Beyond the *P*: I: Problems with probability. Journal of Cataract & Refractive Surgery 30(9):2005–2006.

[4] Siegried T., (2010) Odds are, it's wrong: Science fails to face the shortcomings of statistics. Science News 177(7):26-29.

[5] Cohen J. (1994) The Earth is Round ($p < .05$). American Psychologist 49(12):997-1003.

[6] Bakan D. (1966) The test of significance in psychological research. Psychological Bulletin 66:1-29.

[7] Morrison D.E. and Henkel R.E. (1970) The significance test controversy. Chicago: Aldine Publishing Company.

[8] Falk R. and Greenbaum C.W. (1995) Significance Tests Die Hard. Theory & Psychology 5(1):75-98.

[9] Goodman S. (1999) Toward Evidence-Based Medical Statistics 1: The P Value Fallacy. Annals of Internal Medicine 130(12):995-1004.

[10] Wagenmakers E. (2007) A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review 14:779-804.

[11] Dixon P. (2003) The p-value fallacy and how to avoid it. Canadian Journal of Experimental Psychology 57(3):189-202.

[12] Panagiotakos D. (2008) The Value of p-value in Biomedical Research. Open Cardiovasc Med J 2:97-99.

[13] Goodman S. (1993) p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. American Journal of Epidemiology 137(5):485-496.

[14] Sainani K., (2010) Misleading Comparisons: The Fallacy of Comparing Statistical Significance. Physical Medicine and Rehabilitation 2:559-562.

[15] Taylor K. and Frideres J. (1972) Issues Versus Controversies: Substantive and Statistical Significance. American Sociological Review 37:464-472.

[16] Gliner J., Leech N. and Morgan G. (2002) Problems With Null Hypothesis Significance Testing (NHST): What Do The Textbooks Say?. The Journal of Experimental Education 71(1):83-92.

[17] Pocock S., Hughes M. and Lee R. (1987) Statistical Problems in the Reporting of Clinical Trials. New England Journal of Medicine 317:426-432.

[18] Kaye D. (1986) Is Proof of Statsitical Significance Relevant?. Washington Law Review 61:1333.

[19] Nurminen M. (1997) Statistical significance – a misconstrued notion in medical research. Scandinavian Journal of Work Environment and Health 23(3):232-235.

[20] Hopewell S., Loudon K., Clarke M.J., Oxman A.D. and Dickersin K. (2009) Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews, Issue 1.