Null Hypothesis Testing (I) — 5% Significance Level

In early 2015, the psychology journal *Basic and Applied Social Psychology* announced its ban on the use of "Null Hypothesis Significance Testing Procedure" (NHSTP), because of its invalidity.[1] Authors are now required to remove all vestiges of NHSTP, and any statement of "significant" differences. It is the first academic journal to take action against NHSTP, but the challenge of NHSTP began about 50 years ago.

The issue was reported in the scientific magazines *Nature*[2] and *Scientific American*.[3] The responses varied in the community of statistics and the community of research. There was heated debate because the ecosystem of the research community was being challenged.

Colquhoun[4] concluded that "if you use $p = 0.05$ to suggest that you have made a discovery, you will be wrong at least 30% of the time". He proved his argument from several perspectives. He suggested that if we wanted to keep the false discovery rate below 5%, we need to insist "three-sigma rules" or $p \leq 0.001$ and never use the word "significant". Imagine! Our knowledge pool will be contaminated rapidly by false discovery. Psychiatrists, and other medical professionals, should not tolerate such an enormous false-positive rate.

In this article, the magic figure $p \leq 0.05$ will be discussed. The uses, limitations, and interpretation of NHSTP will be further discussed in a later essay.

Evidence-based medicine (EBM) was first described in the late 1980s and early 1990s. It advocates that decision-making in medical practice should be based on evidence derived from well-designed and conducted research. Most of this research is quantitative with conclusions based on the result of NHSTP. In almost all applications of NHSTPs, the magic figure $p \leq 0.05$ has been used as the cut-off point to decide whether a conclusion is "significant" or not. You may ask: is $p \leq 0.05$ an evidence-based figure? If it is not, how can EBM claim to be evidence-based? Some of our colleagues already question $p \leq 0.05$ as an appropriate cut-off. It is difficult to imagine a scenario where 1 in 20 studies in medical research yield false discoveries.

To answer this question, we can revisit the thoughts of one of the founders of NHSTP, Sir Ronald Aylmer Fisher. In his book, *Statistical Methods for Research Workers*,[5] in a discussion of "normal distribution", he stated: "The rapidity with which the probability falls off as the deviation increases is well shown in these tables (I & II). A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials, while Table II shows that to exceed the standard deviation six-fold would need nearly a thousand million trials. The value for which $p = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant."

Fisher defined deviation exceeding twice (but not one or thrice) the standard deviation (of normal distribution), which equals almost $p = 0.05$, as the value considered to be significant or not, because it is convenient. He provided no other rationale. Therefore, at the time of defining $p = 0.05$ as the level of significance, there was no logical reason or evidence to support such a decision. Unfortunately, although Fisher made some inconsistent comments about the choice of significance level later on in his book, we have adopted the value $p = 0.05$ as the level of significance for at least the last 80 years.

In an earlier edition published in 1925, *Statistical Methods for Research Workers*,[6] Fisher implied $p = 0.05$ as the level of significance. At the end of Chapter VII "Interclass Correlations and the Analysis of Variance", he presented Table VI for the value of Z for F distribution. For some distributions, like *t* distribution or Chi-square distribution, the *t* or Chi-square value depends on a single value of degree of freedom only. Then, the tables (2-dimensional on a page) can be presented with different levels of significance, say 0.1, 0.05, 0.01, 0.005, and so on. For F distribution, the F value depends on 2 values of degree of freedom from 2 population distributions. As a result, in a single table, only one level of significance can be chosen. In the first edition of the book (1925),[6] Fisher presented a table only for F distribution with $p = 0.05$. The reader would be led to believe that Fisher preferred the significance level to be set at $p = 0.05$, although another table with $p = 0.01$ was added in the fifth edition (1934).[5] Once again, there was no reason to select $p = 0.05$.

The aim of writing *Statistical Methods for Research Workers* was to enhance the statistical knowledge of the researcher, particularly the biologist. We are able to tolerate a rate of false positivity at a 0.05 level when we are discussing factors such as the yield of a crop in agricultural study, or the birth and death rate of endangered species in ecology. Can we tolerate a rate of false discoveries of 1 in 20 in medical research, when we are talking about "life and death" issues and huge expenditure on medical services?

Don't forget! The rate of false positives can be cumulative because the conclusions of any current study are based on the conclusions of previous studies. Say, in the first "generation" of research, the false-positive rate is

0.05. If it remains 0.05 in the second, the third, the fourth… generations, the cumulative false-positive rate will have risen to greater than 50% after 13 "generations". By this time, the probability of obtaining a "true" knowledge is less than half. The problem of accumulating false positives will occur more rapidly in this age of acquiring new information because the duration of each generation is shorter.

Gelman and Loken[7] described it as "The Statistical Crisis in Science". If it is, it should be the "First Statistical Crisis" in the 120-year history of inferential statistics. Compared with the Three Crises in Mathematics in 5000 years' history of Mathematics, the "First Statistical Crisis" has emerged earlier than expected. Because the crisis involves the whole community of research (not only medical research), its impact will be much greater than any one of the mathematical crises.

In the meantime, before this "Statistical Disaster" (not crisis) emerges, suggestions of Colquhoun[4] can be adopted to minimise the influence of the crisis:

1. To use "three-sigma rules" ($p = 0.0027$) instead of the "two-sigma rule" ($p = 0.05$); or
2. To insist on $p \leq 0.001$; and
3. Never use the word "significant".

Unfortunately, for a lower p value, a larger sample size is required. As a result, increased cost and resources are unavoidable. The decision about whether to adjust the significance level below 0.05 requires careful consideration. It must balance the cost of larger studies against the cost to society of a high false-positive rate that may result in inappropriate social policy, unnecessary treatment, or further redundant follow-up studies.

In this article, we have focused on a 5% significance level. Although there are other methods that can further increase the power of a study and lower the false-positive rate, such as meta-analysis, those methods are not discussed here.

The action taken by the journal *Basic and Applied Social Psychology*[1] might be too extreme and statistical techniques will consequently regress to being merely descriptive. Nonetheless, if no action is taken, the pool of medical knowledge may be further contaminated. As a responsible and influential psychiatry journal in Hong Kong, China, other countries of Asia and the rest of the world, please consider appropriate action to protect our knowledge pool.

The term EBM is not convincing if the cornerstone of quantitative research, $p \leq 0.05$, is not itself evidence-based. It is to be hoped that adoption of Colquhoun's suggestion is a short-term measure until other methods of NHSTP are developed or current techniques are optimised.

Dr Kai-Choi Wong, FHKAM (Psychiatry), GradStat
e-mail: kc@drkcwong.com
Specialist in Psychiatry
Graduate Statistician (Royal Statistical Society)
Hong Kong SAR, China

References

1. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych 2015;37:1-2.
2. Woolston C. Psychology journal bans P values. Nature 2015;519:9.
3. Nuzzo R. Scientists perturbed by loss of stat tools to sift research fudge from fact. Available from: http://www.scientificamerican.com/article/scientists-perturbed-by-loss-of-stat-tools-to-sift-research-fudge-from-fact/. Accessed 16 Apr 2015.
4. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014;1:140216.
5. Fisher RA. Statistical methods for research workers. 5th ed. Edinburgh: Oliver and Boyd; 1934: 44-5, 232-5.
6. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925: 210.
7. Gelman A, Loken E. The statistical crisis in science. Data-dependent analysis — a "garden of forking paths" — explains why many statistically significant comparisons don't hold up. Am Sci 2014;102:460-5.