Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

The Origin of $p < .05$: How Did the Concept of Statistical Significance Arise?

Frederick L. Coolidge[1,*]

[1] Psychology Department, University of Colorado, Colorado Springs, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918, USA

ARTICLE INFO

Article history:
Received 30 September 2024
Received in revised form 14 February 2025
Accepted 10 March 2025
Available online 28 March 2025

Keywords:
Statistical significance; $p < .05$; null hypothesis; 5% level of significance; Ronald A. Fisher

ABSTRACT

The present article describes the origins of statistical significance and the concept of p < .05. Their origins can be traced to a Cambridge University trained mathematician, Ronald A. Fisher (1890-1962), who joined an English agricultural research center upon the recommendation of William S. Gosset (1876-1937), an Oxford University mathematician, who worked as the head brewer for Guinness Brewery in Ireland. In Fisher's 1935 classic book, The Design of Experiments, he reported that at teatime at Rothamsted (probably in the early 1920s), another researcher Muriel Bristol, PhD claimed that the milk has been added to the tea, and she preferred the opposite. Fisher protested that such a discrimination was impossible. It was suggested that she be offered eight cups of tea, where four cups had tea added to the milk and four cups had milk added to the tea (presumably not subsequently stirred). Fisher deduced that the probability of getting all eight cups right by chance alone was 1 out of 70 or 1.4%. Because Fisher doubted that she could accurately identify the cups any better than chance, he thus established the concept of the null hypothesis. Fisher further stated that it would be "convenient" for experimenters to take 5% as the standard level of significance and to ignore results that did not reach this standard. The article continues to discuss the origin of Gossett's *t* sampling distribution, which may also be traced to the tradition of afternoon tea breaks.

1. Introduction

It may be common knowledge among students of psychology and others that $p < .05$ indicates statistical significance, but what were its origins? Scientists typically begin an experimental study to determine the influence of one variable upon another with the null hypothesis, that is, there is no causative relationship between two variables, or there is no relationship between them beyond chance differences. Next, based on a statistical analysis of the data, the choice is to retain the null hypothesis (or fail to reject it) or reject the null hypothesis in favor of an alternative hypothesis, that is, there is a causative relationship. When the null hypothesis is rejected, the conventional notation of statistical significance is attached, i.e., $p < .05$. It also may be interpreted as the probability (p) of

* *Corresponding author*
E-mail address: fcoolidg@uccs.edu

https://doi.org/10.37934/ard.126.1.131134

rejecting the null hypothesis when the null hypothesis is true. But where did less $p < .05$ (i.e., less than 5 chances out of 100) actually originate?

1.1 Teatime at Rothamsted

The origin of the null hypothesis and statistical significance can be traced to an agricultural research center, Rothamsted, England. In 1919, a young Cambridge University trained mathematician, Ronald A. Fisher (1890-1962), joined Rothamsted to analyze data gathered from years past about crop yields like wheat, barley, etc. as a function of weather and various kinds of fertilizers [1]. Fisher had been recommended for the job by William S. Gosset (1876-1937), an Oxford University mathematician, who worked as the head brewer for Guinness Brewery. Gosset, of course, is known for his creation of the *t* sampling distribution and the *t* test for independent samples.

In 1935, Fisher published one of his two classic texts, *The Design of Experiments*. In it, he reported that at teatime at Rothamsted (probably in the early 1920s), Muriel Bristol (PhD in algae studies), and William Roach (her soon-to-be husband) were having tea when she exclaimed that the milk has been added to the tea, and she preferred the opposite. Fisher protested that such a discrimination was impossible. Roach suggested that they offer her eight cups of tea, where four cups had tea added to the milk and four cups had milk added to the tea (presumably not subsequently stirred). Bristol was aware that the cups had been randomly presented and was aware of the nature of the eight cups (half tea to milk, half milk to tea). Fisher noted, "Her task was to divide the 8 cups into two sets of 4 …" (p. 14). Fisher also noted, "There are 70 ways of choosing a group of 4 objects out of 8" using "… an argument familiar to students of 'permutations and combinations…'" (p. 14). Fisher deduced that the probability of getting all eight cups right was 1 out of 70 or 1.4%.

Because Fisher doubted that she could accurately identify the cups any better than chance, he was firmly establishing the concept of the null hypothesis. He also noted in that classic text that the null hypothesis is never 'proven,' but it is 'possible' to offer evidence so that it might be disproven. Fisher further reasoned that the null hypothesis might need to be balanced between having 'no possible result' being acceptable (to reject the null hypothesis) and having some maximum number of trials be ascribed to chance ("… in an average of 5 trials out of 100") [2] yet still rejecting the null hypothesis. And he stated in the very next sentence:

"It is usual and convenient for experimenters to take 5% as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard…" (p. 15-16). It is interesting to note that Fisher had already established that Dr. Bristol's probability of getting all eight cups right was 1.4%, but the 5% standard subsequently became the conventional level of statistical significance.

Fisher [3] had also previously heralded this criterion in a paper to a society investigating psychical phenomena when he wrote:

"It is common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary but convenient, level of significance for the practical investigator …" (p. 390).

Most curiously, Fisher did not report the outcome of the tea experiment! However, in a book about the history of 20th century statistics, Salsburg [4] reported that one of Fisher's colleagues, H. Fairfield Smith, claimed that Dr. Bristol correctly identified all of the cups of tea! It is also important to note that Fisher went on to make major contributions to modern statistics like analysis of variance, the *F* distribution, and the foundations of multivariate statistics.

1.2 Teatime Again Plays a Role in the Origins of Statistical Significance!

Gosset [5] had published his first paper in 1907, *On the error of counting with a haemacytometer*, in the prestigious journal *Biometrika*, among whose cofounders were Francis Galton and Karl Pearson. Gosset used a haemacytometer to count yeast cells in the beer brewing process. Interestingly, Gosset published his paper under a pseudonym, Student. A previous Guinness employee had published one of Guinness' recipes for making beer, so they had a strict publishing policy of not using the Guinness name, one's own name, or any beer recipes. In his 1908 paper, *On the probable error of the mean*, also published under the name Student, Gosset [6] presented the basis for the *z* distribution. In 1912, Fisher wrote to Gosset and demonstrated that Gosset's formula about the error of the mean could be improved by dividing by $n - 1$ instead of n, thus also establishing the concept of degrees of freedom. Fisher also derived a geometric proof of Gosset's *z* distribution and told him that he had solved his formula in n – dimensional space. Perhaps as a measure of Gosset's humbleness, he replied to Karl Pearson, "I couldn't understand his stuff. I don't feel at home in more than three dimensions even if I could understand it otherwise." Fisher also derived the *t* sampling distribution from Gosset's *z* distribution by using $t = z\sqrt{n - 1}$ [7]. In a review of Gosset and Fisher's 25 years of letters and their publications and friendship, Eisenhart [8] speculated that the shift from the *z* distribution to the *t* distribution was clearly Fisher's work, but the choice of the letter *t* was Gosset's. As for Gosset's pseudonym choice, he liked working in blank notebooks, on scraps of paper, and on the backs of envelopes. On one of his early notebooks, there was a publisher's imprint on the cover that read, *The Student's Science Notebook*, and thus there is the speculation that his pseudonym came from the cover [9]. Although, as stated earlier, Fisher derived the *t* sampling distribution, Gosset suggested the use of the symbol *t*. Further, the leading hypothesis for the use of the *t* symbol is that Gosset chose the letter *t* from the practice of 'teatime' both in England and Ireland. And thus, teatime may have played an inadvertent and advertent role in the origins of $p < .05$, statistical significance testing, the null hypothesis, and in the name for the *t* sampling distribution.

1.3 Coda

Although Dr. Bristol's correct answer appears to defy science, it is possible that adding the milk to the hot tea may be more likely to curdle the milk than adding the tea to the milk (she preferred the latter).

Acknowledgements
Portions of this paper appeared previously as a blog in *Psychology Today Online*. The blog was not peer-reviewed, and the author owns the copyright.

Conflict of Interest Statement and Ethical Review
The author states that there is no conflict of interest. There were no participants or data collection involved so there was no need for an institutional review. It is a historical review paper.

References
[1] Box, Joan Fisher. R. A. Fisher, the life of a scientist. Wiley, 1978.
[2] Fisher, Ronald Aylmer. Design of experiments. Oliver and Boyd, 1935.
[3] Fisher, Ronald. Aylmer. "The statistical method in psychical research." Proceedings of the Society for Psychical Research 79 (1929): 386-391.

[4] Salsburg, David. The lady tasting tea: How statistics revolutionized science in the twentieth century. Freeman, 2001.
[5] Student. [William. Sealy Gosset] "On the error of counting with a haemacytometer." Biometrika 5 (1907): 351-360. https://doi.org/10.1093/biomet/5.3.351
[6] Student. [William Sealy Gosset]. "On the probable error of a mean." Biometrika 6 (1908): 300-310. https://doi.org/10.2307/2331554
[7] Zabell, Sandy L. "On Student's 1908 article "The Probable Error of a Mean."" Journal of the American Statistical Association 103 (2008): 1-7. https://doi.org/10.1198/016214508000000030
[8] Eisenhart, Churchill. "On the transition from "Student's" *z* to "Student's" *t*." The American Statistician 33 (1979): 6-10. https://doi.org/10.2307/2683058
[9] Ziliak, Stephen T. "Guinnessometrics: The economic foundation of "Student's" *t*." Journal of Economic Perspective 22 (2008): 199-216. https://doi.org/10.1257/jep.22.4.199