

Principles and Practice of Clinical Research

A Global Journal in Clinical Research



PPCR

ISSN: 2378-1890

P-value's historical background put to a simulation test

MA. Almeida-Santos^{1*}

¹Division of Postgraduation (Master Degree and Doctorate) in Health and Environment. Tiradentes University, Aracaju, Brazil.

*Corresponding authors: Division of Postgraduation (Master Degree and Doctorate) in Health and Environment. Tiradentes University, Aracaju, Brazil. Avenida Murilo Dantas, 300, CEP 49032-490, Aracaju, Sergipe, Brazil. Email: marcosalmeida2010@yahoo.com.br

Received August 16, 2015; accepted August 27, 2015; published September 16, 2015.

Abstract:

Background and Aim: Fisher's "tea experiment", among other important aspects, may be considered the cradle of the modern usage of p -value equal to 0.05 as a resourceful estimation of statistical significance. We aim to shed light on the matter by presenting its historical background as well as a computer simulation of the estimated probabilities.

Methods: The main concepts concerning the p -value and its interpretation are discussed. We also present a statistical simulation of a "modified tea experiment". A binomial distribution probability test is applied in two different strategies. We estimate the probabilities of guessing the correct answer under different scenarios. The commands as well as the results are presented in two mainstream statistical computer packages: R and Stata. We compare the simulation with the "standard" threshold of statistical significance, generally accepted in clinical research as a p -value equal to 0.05 or below.

Conclusion: The presentation of a historical background on a par with the computer simulations are helpful to shed light on Fisher's tea experiment. The combination of understandable information within a short statistical expression took eventually the allure of simplicity and, to some extent, may explain the prestige and overall usage of the p -value.

Keywords: Statistics, binomial distribution, statistical programs.

DOI: <http://dx.doi.org/10.21801/ppcrj.2015.12.6>

INTRODUCTION

Weird as it may, tea tasting and statistics have something in common. Perhaps few of us would even dare to imagine that the much-praised p -value equal to 0.05 (the alpha) had in fact a very trivial cradle. It happened out of plain serendipity, and not as a consequence of painstaking labor, cutting-edge research, let alone the result of mind-boggling calculations.

As matter of fact, in spite of Fisher being taken as the statistician who "introduced the p -value in the 1920s", or the researcher responsible for the "canonization of the 5% level as a criterion for statistical significance", there is evidence of quite similar values (from 1.5% to 7%) back in the nineteenth century².

To avoid misunderstandings, it is appropriate to start by presenting some definitions of p -value and differentiating it from the alpha level. Basically, the p -value is linked to the result of a given statistical estimation, issued when the researcher tests the

probability of the alternative hypothesis being accepted or, in other words, the null hypothesis being rejected. The alpha is the level below which the p -value would be taken as "significant" from a statistical point of view. Under a slightly different perspective, the alpha reflects the type I error. To be clear, if $\alpha = 0.05$, whenever a test provides a lower p -value, we may say the null hypothesis can be rejected, logically under a calculated "risk" of 5% for taking as true a false positive result. This may also be interpreted as a very low probability that the result was due to chance alone, so tiny that, scientifically speaking and in practical terms, there would not be much concern if we disregard this possibility¹.

As elsewhere explained, "the p -value is compared to the predetermined significance level alpha to decide whether the null hypothesis should be rejected"⁴. There is a cautionary note, however. P -values shall not be taken as a way to quantify the "size of the effect". Rather, they just "measure the strength of evidence for an effect"⁵.

Criticism are also found galore, mostly related to its misuse as well as an alleged overstatement of its properties. In a recent review, it was stated that “p-values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume”¹.

METHODS

Historical background

With regards to the “birth” of the p-value, there is an intriguing “story”, already turned into history. We can find several narratives about the subject, but two texts are referential.

The first, a benchmark description of the randomization process, entitled “The Design of Experiments” by Ronald Fisher, originally published in 1935,⁶ where the author preferred to stick to the methodology, rather than the anecdotal details. The second, a historical perspective, entitled “The Lady Tasting Tea”.⁷

Fisher’s report lacks details about the lady as well as the results of the estimations. However, we may take the text, albeit just roughly a dozen pages, as a pioneer description of a properly to-be-done trial. In the chapter entitled “the principles of experimentation, illustrated by a psycho- physical experiment”, several subchapters would catch the attention of any accomplished researcher: “statement of the experiment”, “the test of significance”, “the null hypothesis”, “the effectiveness of randomization”, etc. As a matter of fact, several terms have so far become common ground in whatsoever field of quantitative research.

The “tea experiment” is an emblematic case of “creative inspiration” in the world of statistics. The episode involved Ronald Aylmer Fisher, a brilliant statistician and mathematician, among other talents. And an English gentleman surely he was.

Therefore, the traditional 5 o’clock teatime is a hit or miss, and even more so in those old days, around the 1930s. Then, people talked about trivialities; ladies and gentlemen gathered together around a table with tea; they ate cookies, maybe pieces of cake; and some fellows had milk, for a few of them would rather drink tea blended with a few drops of milk. Naturally, some may prefer it “pure”, that is, without milk, and some may be meticulous on the quality of the milk. What is more, rituals may slightly differ: one can pour milk first and then tea, or the opposite.

But, on that very special day, in Cambridge, there was a “situation”: a lady simply refused to drink tea, on account of the fact that she “knew” it was poured the “wrong way”, i.e., not in the sequence according to her

preferences. Such a statement has probably entailed dismay and raised eyebrows among the audience, and an atmosphere of mistrust or benign neglect started to spread out. During leisure times everybody tries to present his/her sporting attitude or even make jokes: “she must be kidding, don’t you think so”?

Then, on the spur of the moment, a man decided to “make an experiment”. Indeed, a trial, the “primeval” randomized experiment. And that man was Ronald Fisher, twentieth century’s “patriarch” of the methodic approach of current trials. In a question of minutes, people got probably excited and engrossed in the same task: helping with the arrangements so as to put that lady’s tasting skills to a test. The “experiment” was very down-to-earth: eight teacups to be tasted, the information that four cups would have milk poured first and four cups would have tea poured first, and the “service” in random order.

There has been much discussion over the original experiment but, in general, the most appropriate solution is considered a permutation test, which would give a probability of around 1.4% ($p = 0.014$)⁸. By the way, Fisher chose this approach and reached the same results. In fact, he was based on the principle of “combinations”, that is, not an ordered sequence as we would expect in a permutation analysis⁵. To be clearer, had he employed a permutation analysis with 8 trials and 4 unbalanced arrays, that would give 1680 possibilities and a probability of 0.00059524 (i.e., 1/1680). But the lady’s challenge was simpler than that, because it would only matter whether she was capable to present right combinations of 4 “milk first” tea cups and 4 “tea first” tea cups. That would leave 70 combinations, what makes the probability of selecting the right array equal to 0.01428571 (i.e., 1/70).

Computer simulations and discussion over the results

Now, we could imagine for minute that the original experiment had the teacups presented in four batches of two. In each batch, one teacup had milk poured first whereas the other teacup had tea poured first.

Let’s conceive the “study question” this way: can the lady identify if milk has been poured before or after the tea? Or, generally speaking, can a human being identify if milk

has been poured before or after the tea?

Null hypothesis: that lady – particularly – or a human being – in general – cannot tell the difference between both sequences.

Alternative hypothesis: it is possible (and not due to a mere coincidence, i.e., “chance”, or, technically speaking,

“random error”) for this lady – or a human being – to tell the difference between both sequences.

Interesting enough, we could call this a true to type “N of 1 trial” as well, once batches of paired teacups could be presented to a single individual under a multiple crossover pattern. We may now hazard a guess on the probability (p-value) of “hitting the targets” overall, that is, guessing correctly the sequences of four tests “not by chance”. Nowadays, instead of doing the estimations “by hand” (or with the aid of an electronic calculator) we can perform the estimations using a statistical computer package.

Basically, since the answer must only be correct or incorrect (therefore, a binary variable), we need to apply a test that checks the probabilities under a binomial distribution.

In Stata, there are two options to perform this test. Either from the drop-down menu, where we browse through the windows and click on the following titles: Statistics > Summaries, tables, and tests > Classical tests of hypotheses > Binomial probability test calculator. Or we may choose to type the commands directly in the command window:

```
. bitesti k n p
```

In the R statistical software, we shall type in the console:

```
>dbinom(n, k, p)
```

Apart from the differences concerning the main command (bitesti and dbinom), there are a few aspects to be aware of when dealing with one of these statistical packages. Contrary to what we must do in R, in Stata there is no comma between parameters, neither parenthesis. That said, “n” stands for the number of correct observations; “k” stands for the number of trials; and “p” stands for the probability of success in each trial. Finally, “n” is the first item in R, but the second in Stata.

For example, here employing both types of software (R and Stata): considering we have a “fair coin”, that is, there is a 50% chance of having “tails” (as well as “heads”), the commands for a binomial distribution related to tossing a coin ten times and guessing correctly in 8 trials are:

```
>dbinom(8, 10, 0.5)
```

```
. bitesti 10 8 0.5, detail
```

We may have the p-value as the sole result in R. Stata also provides other possibilities (for example, if the number of trials are ≤ 8) but the answer we seek is clearly presented under the expression $\text{Pr}(k=8)$, and we shall have it by including detail in the options. Therefore, under both statistical packages we would get the same result: a p-value equal to 0.04394531

With these parameters in mind, we might well foresee the probabilities of guessing right in the “modified

tea experiment”. We decided to perform the estimations under two different conditions. In the first one, we considered the lady was given eight tea cups, and they are independently presented (Table 1).

However, this option does not mirror Fisher’s experiment quite well. We must consider this: under such study design, if the individual wrongly selects a cup of a batch, will this decision influence the results concerning the remaining cup?

As a matter of fact, when cups are presented in pairs, once the first cup of each batch is guessed (wrong or right), the remaining cup has forcefully to be taken as the opposite alternative. On account of this particularity, a binomial probability test reflecting the “modified tea experiment” should ideally encompass only four “independent” batches, therefore taking in consideration uniquely the result of the first guess in each batch (Table 2).

This argument, sound as it seems to be, relieves the pressure on the lady’s skills, so to speak. In short, instead of a round p-value = 0.004 for guessing all cups right, there is only need of a p-value around 0.06 for the (still, rather demanding) task. If we speculate whether Dr. Muriel Bristol Roach (the lady’s name) missed just one batch, that would give a p-value = 0.25.

There are further aspects worth mentioning. Among them, the symmetry of the distribution, only when the probability of correct guesses is 50%. In case the probability differs from 50%, the distribution assumes a skewed pattern (Figure 1). Also worth noticing, the highest p-values stay close to the estimated percentage of corrected guesses.

There is a crucial point to underline: no matter we use the design of eight batches or the one with four trials, we are bound to fail to spot the “benchmark” parameter, i.e., a p-value equal to 0.05. Even in case we speculate the lady guessed right in all four batches, we would get a p-value = 0.0625.

Such a value under most study designs would be interpreted as “nonsignificant”. Add to it that, if we recall the results from the true-to-type tea experiment ($p = 0.014$), we are also very far from reaching the ubiquitous probability of type I error of 5%.

In order to get much closer to the benchmark “alpha” value, we might perform a binomial test of probability under 10 trials and 8 correct guesses ($p = 0.04394531$). On the same verge, six correct guesses out of seven trials would entail an even closer value ($p = 0.0546875$).

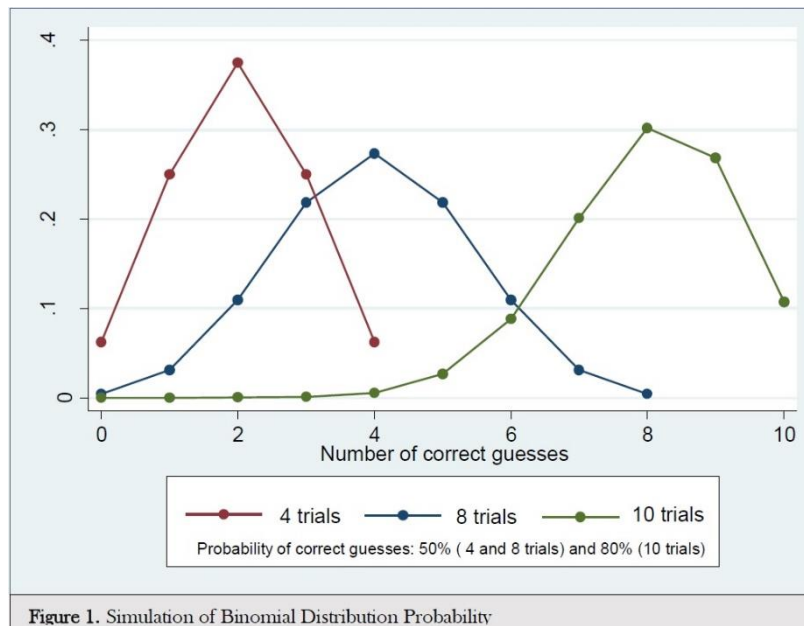
Now, we may have unveiled the gist of this “story”. The pristine p-value calculated for that emblematic experiment was not 5% at all.

Table 1. Binomial probability test related to 8 trials with independent guesses

Number of correct guesses	Number of trials	Estimated p-value
8	8	0.00390625
7	8	0.03125000
6	8	0.10937500
5	8	0.21875000
4	8	0.27343750
3	8	0.21875000
2	8	0.10937500
1	8	0.03125000
0	8	0.00390625

Table 2. Binomial test related to 4 trials with independent guesses

Number of correct guesses	Number of trials	Estimated p-value
4	4	0.0625
3	4	0.2500
2	4	0.3750
1	4	0.2500
0	4	0.0625



This notwithstanding, Fisher – in spite of having underlined several caveats of the discriminatory usage of p- values – showed some sort of “sympathy” for the threshold of p equal to 0.05: “It is usual and convenient for

experimenters to take 5 per cent as a standard level of significance [...]”.⁶ However, we must match this sentence of the renowned statistician with a previous one, written a couple of lines before: “it is open to the experimenter to

be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result”.

In summary, according to Fisher’s manuscript, a statement concerning the probability of type I error – the alpha – must be made before the experiment. Moreover, the threshold for statistical significance is somewhat “discretionary”.

CLOSING REMARKS AND CONCLUSION

When contextualized with the historical background, computer simulations can help to clarify interesting aspects related to Fisher’s tea experiment. Finally, may the reader be curious about the true results of the “tea experiment”, it seems, according to an eye-witness, as reported by David Salsburg,⁷ the lady was absolutely right: she gave the right answers concerning every single cup. Indeed, a remarkable lady.

In spite of criticism and caveats, the p-value still remains as a straightforward resource to test the null hypothesis. Its allure stems from the apparent simplicity of the presentation of the results. On the other hand, its pitfall lies on the fact that the p-value is not aimed at measuring the “effect size”. For that specific matter, depending on the applied statistical analysis, there are several tests available, such as Cramer’s V, Cohen’s d, Omega-squared, Odds ratio and Relative risk, among others.¹⁰

Hopefully this inspirational experience points out the need to be less dogmatic and more reflexive when interpreting p-values. It seems Sir Ronald Fisher didn’t forcefully determine the “right” or “ideal” p-value. Thence, with regards to statistical significance, shouldn’t we acknowledge the p-value of 0.05, fundamentally and under most scenarios, like a (reasonable) “gentlemen’s agreement”?

Conflict of interest and financial disclosure

The authors followed the International Committee or Journal of Medical Journals Editors (ICMJE) form for disclosure of potential conflicts of interest. All listed authors concur with the submission of the manuscript, the final version has been approved by all authors. The authors have no financial or personal conflicts of interest.

REFERENCES

- 1 Nuzzo R. Scientific method: statistical errors. *Nature*. 2014; 506:150- 152.
- 2 Stigler S. Fisher and the 5% Level. *Chance*. 2008; 21(4):12.
- 3 Woodward M. *Epidemiology: study design and data analysis*. 2nd ed. New York: Chapman and Hall/CRC Press; 2014. p.34.
- 4 Pagano M. *Principles of biostatistics*. 2nd ed. Belmont: Kimberlle Gauvreau; 2001. p. 234.

- 5 Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. *Regression Methods in Biostatistics*. 2nd edition. New Yourk: Springer; 2012. p.5.
- 6 Fisher R. *The design of experiments*. New York: Hafner Publishing Company; 1971.
- 7 Salsburg D. *The lady tasting tea: how statistics revolutionized science in the twentieth century*. New York: Henry Holt and Company; 2001.
- 8 Yates F. Test of significance for 2 X 2 contingency tables. *J.R. Statist. Soc.* 1984;147: 426-463.
- 9 Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. *Med. Care*. 2011; 49(8):761-8.
- 10 Sullivan GM, Feinn R. Using effect size – or why the p value is not enough. *J. Grad. Med Educ*. 2012; 4(3): 279-2829. Carozzi NE, Grech J, Tulsy DS. Memory functioning in individuals with traumatic brain injury: an examination of the Wechsler Memory Scale-Fourth Edition (WMS- IV). *Journal of clinical and experimental neuropsychology*. 2013;35(9):906-14.
- 11 Guangyong Zou: A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology* 2004,159:702-706.