Article

p-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 28 April 2022; *Accepted 12 May 2022*; *Published online 19 May 2022*; *Published 1 September 2022*

Abstract
The *p*-value is at the heart of statistical significance tests, a very important issue related to the role of statistical inference in advancing scientific discovery. Over the past few decades, *p*-value based statistical significance tests have been widely used in most statistics-related research papers, textbooks, and all statistical software around the world. Numerous scientists in various disciplines hold the *p*-value as the gold standard for statistical significance. However, in recent years, the *p*-value based statistical significance tests have been questioned unprecedentedly, mainly because the paradigm of significance tests is wrong, *p*-value is too sensitive, *p*-value is a dichotomous subjective index, and statistical significance is related to sample size, etc. Scientific research can only be falsified, not confirmed. *p*-value based statistical significance tests are one of the sources of false conclusions and research reproducibility crisis. For this reason, many statisticians advocate to abandon *p*-value based statistical significance tests and replace them with effect size, Bayesian methods, meta-analysis, etc. Scientific inference that combines statistical testing and multiple types of evidence is the basis for producing reliable conclusions. Reliable scientific inference requires appropriate experimental design, sampling design, and sample size; it also requires full control of the research process. For complex and time-varying problems, the network or systematic methods should be used instead of the reductionist methods to obtain and analyze data. To change the scientific research paradigm, the paradigm of multiple repeated experiments and multi-sample testing should be adopted, and multiple parties should verify each other to improve the authenticity and reproducibility of the results. In addition to writing, publishing and adopting new statistical monographs and textbooks, the most urgent task is to revise and distribute various statistical software in the new versions based on the new statistics for further use. Before the popularization of new statistics, what we can do is to improve data quality, strict *p*-value levels of statistical significance tests, use more reasonable analysis methods or testing standards, and combine statistical analysis and mechanism analysis, etc.

Keywords *p*-values; statistical significance tests; reproducibility; Bayesian methods; effect size; Bootstrap.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Concepts of the *p*-value and statistical significance tests

Statistical significance hypothesis testing (including student's *t*-test, *F*-test, etc.), usually referred to as significance testing or hypothesis testing, was proposed in the 1920s, and is one of the most important statistical inference methods of the frequentist school. Hypothesis testing has always been a standard part of statistics textbooks. Researchers widely use statistical significance as a certificate of scientific discovery. Whether a research result is statistically significant is judged by the *p*-value obtained from hypothesis testing, and usually using, for example, p=0.05 as the threshold dichotomy. That is, p<0.05 means statistically significant, and p>0.05 means not statistically significant (Rosner, 2006; Huang, 2021a, b). In other words, when p<0.05, i.e., the probability due to chance is less than 5%, the results are considered to be statistically significant (Bergstrom and West, 2021).

p-value is the occurrence probability of a sample observation or more extreme results when the null hypothesis is true. *p*-value can be defined as the probability of set of extreme values of a sample statistic (For example, the the difference between the sample means of two contrasting samples) under a given statistical model. It can be expressed as: $p = \Pr(D|H)$, where D represents the set of extreme values of a certain sample statistic, and H represents a given statistical model/statistical assumption (Bergstrom and West, 2021). *p*-value, also known as the significance level (i.e., type I error), is the probability that the type I error occurs, i.e., the probability of a false positive. If *p*-value is very small, indicating that the probability of the occurrence of the null hypothesis is very small, and if it occurs, according to the principle of small probability, we have a reason to reject the null hypothesis. The smaller the *p*-value, the stronger the reason for us to reject the null hypothesis. In short, the smaller the *p*-value, the more significant the result. But whether the testing result is significant, moderately significant, or highly significant, we need to judge it according to the *p*-value and the actual problem. When we test a statistical hypothesis H_0, the *p*-value is the occurrence probability of a sample result or a more extreme result when H_0 is true, or the probability of obtaining a sample that is at least as extreme as the actually measured sample. The smaller the *p*-value , the more you should reject the null hypothesis, that is, the more significant the result. The false negative here should be understood like this, for example, the null hypothesis (H_0) is that the drug has no effect on cancer cells, and as a result we have a probability of 0.1 to conclude that cancer cells have no death, thus the null hypothesis is supported. This is a negative result (no rejection of the null hypothesis), and it is also an error result. So it is a false negative error (Shao, 2018).

p-value is based on the random variable, and the values of the random variable follows a certain probability distribution. Mathematically, *p*-value can be described as follows. There is a random variable X, and its value is denoted as x. The value of the random variable X being less than x, is a function of x and can be expressed as:

$$F(x) = P(X < x)$$

$F(x)$ is called the distribution function. If the distribution function, $F(x)$, is known, then for any x, the probability $P(X<x)$ that the value of the random variable X is less than x can be obtained. Continuous random variable distributions, are commonly normal distribution, t distribution, F distribution, χ^2 distribution, exponential distribution, power law distribution, etc. For different types of continuous random variables, the form of the probability density function $f(x)$ is different, satisfying $\int_{-\infty}^{+\infty} f(t)dt = 1$, and

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$$

The probability that a continuous random variable X takes a value greater than x, is:

$$p = 1 - F(x) = 1 - P(X < x) = 1 - \int_{-\infty}^{x} f(t)dt$$

The probability that a continuous random variable X takes a value within (x_1, x_2) is:

$$p = 1 - P(x_1 < X < x_2) = 1 - \int_{x1}^{x2} f(t)dt$$

At this time, the confidence interval of the values of the random variable X is (x_1, x_2), The corresponding confidence level is $1 - p$. The probability p, is called p-value, or significance level (Fig. 1).

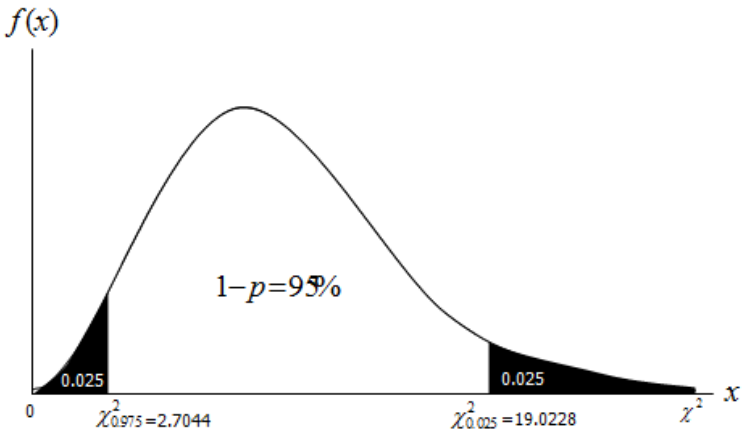
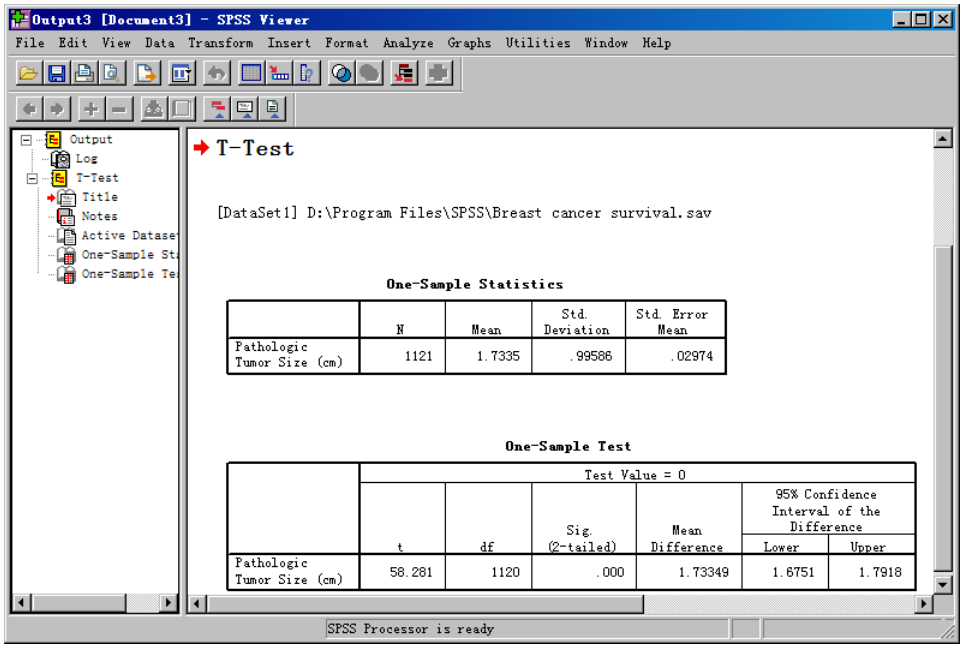


Fig. 1 Illustration of χ^2 distribution and p-value based statistical significanc.



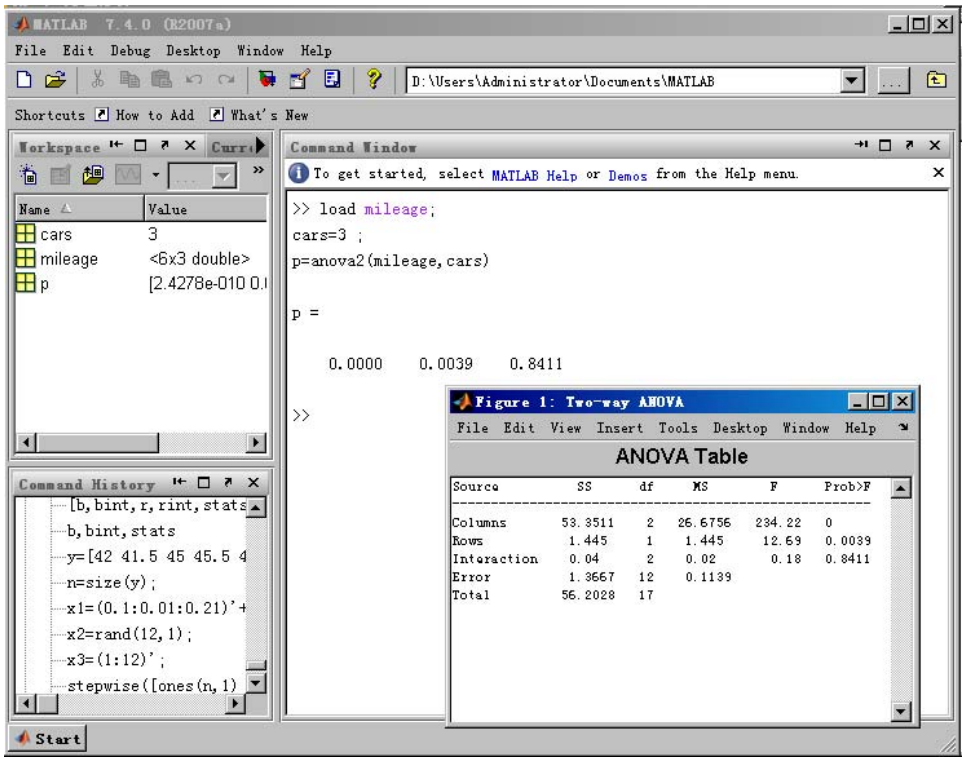


Fig. 2 *p*-value based statistical significance tests are being overwhelmingly used in science. Upper: a *t*-test summary in SPSS; Lower: an ANOVA table in Matlab.

The *p*-value is at the heart of the statistical significance tests, a very important concept about the role of statistical inference in driving new scientific discoveries. Over the past few decades, *p*-values based statistical significance tests have been used in most statistics-related research papers, monographs, textbooks, and all statistical software worldwide, and countless scientists in various disciplines have touted the *p*-value as the gold standard for statistical significance (Sun, 2016; Fig. 2).

2 Misconception, Misuses and Critiques

2.1 History of the statistical significance tests

Statistical significance tests began with a cup of milk tea incident in England in the 1920s (Xie, 2022b; Grenville, 2019). Several scientists had afternoon tea together, one was Dr. Bristol, and the other was Dr. Fisher. Fisher handed Bristol a cup of brewed milk tea, but the latter refused, because Fisher poured the tea into the cup first and then the milk, but Dr. Bristol likes to pour the milk into the cup first and then mix the tea. Fisher thought that Bristol must not be able to tell the difference between the two but Bristol insisted that she could tell the difference. Fisher then proposed a testing, a significance testing, to make inferences. Laterly, Fisher wrote this whole thing (called The Lady Tasting Tea) into his well-known book, *The Design of Experiments* (Fisher , 1935) and the testing story became famous throughout the world. For this testing, Fisher prepared eight teacups: four poured tea first and then milk, and four poured milk first and then tea. Bristol had to judge which cup of milk tea is brewed by which method (Xie, 2022b; Grenville, 2019).

Fisher proposed a null hypothesis to assume that she could not make a correct judgment. Fisher calculated that, assuming the above null hypothesis holds, the probability that Bristol can correctly guess the brewing method of all 8 cups of milk tea is 1/70. Fisher was willing to conditionally acknowledge her ability to judge

correctly (i.e. reject the null hypothesis) only for this testing. It is said that none of Bristol's judgments were wrong, and the null hypothesis was rejected. This is the origin of the significance testing (Xie, 2022b; Grenville, 2019).

2.2 Misuses and critiques of *p*-value based statistical significance tests

Modern statistics have gone through 120 years since Pearson (1904) put forward the 'chi-square (χ^2) test'. Unfortunately, the statistical community has still not been able to give satisfactory answers of the most basic problems in statistical theory and data analysis, namely 'The definition and interpretation of probability in real life' and 'statistical hypothesis testing' (Xie, 2022c).

For a long time, the statistical significance tests are a widely used but controversial statistical inference method in scientific research. In the past ten years, the theory of statistical significance tests has been greatly challenged (Huang, 2021a, b).

The problem is not really with the statistical analysis methodss themselves, but with how these methodss are used. We often see data tables where each column in the table is analyzed and compared with other columns. Those results that are significantly different are identified by a letter or number. Each table may contain multiple testing results. Often researchers try to interpret results based on these significant differences, and errors can thus arise (Xie, 2022a; Grenville, 2019).

As early as 1951, Yates had evaluated the role of statistical hypothesis testing: the emphasis on significance testing and the separate consideration of the results of each experiment has had the undesirable effect, which makes researchers often performing significance testing on data from an experiment as the ultimate goal to see if the result is statisticall significant or not (Yates, 1951; Xie, 2022d).

The ultimate purpose of statistical testing is to demonstrate scientific significance. However, one of the main flaws of the statistical significance tests paradigm is that statistical significance is used to represent scientific significance. Although many scientists know that statistical significance is not the same as scientific significance, and statistical inference is not the same as scientific inference, statistical significance is still often misused and misunderstood by researchers. The *p*-value, at the center of the controversy of statistical significance tests, is a precise and sensitive indicator. To abandon or ban the *p*-value indicator, most people may not agree. The *p*-value is criticized because it is too precise and too sensitive. In particular, the *p*-value is greatly related to the sample size and can be used by researchers to yielded statistically significant results by expanding the sample size (Sun, 2016).

Research by McShane and Gal (2017) shows that even researchers who are themselves statisticians are prone to misuse and misunderstanding *p*-values, leading to similar errors (Huang, 2021a, b). Although *p*-values are widely used, but few people really understand what *p*-value means. In 2002, German researchers conducted a survey in psychology researchers and students and presented them with six statements about *p*-value. All students could not correctly understand the meaning of *p*-value (Haller and Krauss, 2002). Even the teachers who taught statisticall methodology, 80% of them could not correctly understand *p*-value. This shows that researchers are very easy to misunderstand *p*-value. This result is basically consistent with an earlier survey (Oakes, 1986). If the meaning of a statistic is so easily misunderstood that even most methodology teachers cannot understand it correctly, the value of its existence and application is questionable (Grenville, 2019).

Nuzzo (2014) pointed out that when most scientists see a result with the p=0.01, they will say that their result has only a 1% chance of being a false positive result, but this understanding is obviously wrong. The *p*-value itself cannot provide this information. The *p*-value is simply a numerical value of a summary statistic about the data, assuming a null hypothesis holds. It cannot be used to make reverse inferences and draw conclusions about whether the hypothesis is true. Making such a reverse inference requires another piece of

information, the probability that a true effect exists beforehand (Nuzzo, 2014; Xie, 2022a; Grenville, 2019). According to a widely used formula (Goodman, 2001), the p=0.01 corresponds to a probability of a false positive result of at least 11%, depending on the probability that the true effect pre-exists; the p=0.05 increases the probability of a false positive result to at least 29% (Nuzzo, 2014). It is known that 1/20 of the testing results are false positive results, that is, the testing results are positive when the real difference does not exist. The true situation may be that the error rate of this false positive is actually much higher than 1/20, which is especially true for multiple comparisons (Xie, 2022a; Grenville, 2019).

p-value is not only easy to be misunderstood, but also often to be abused. In the statistical significance tests paradigm, p-value is used as a measure of the statistical significance of research results. p<0.05 is generally considered to be statistically significant for research results. Academic journals usually prefer to publish results that are statistically significant, and for a long time, p-values such as p< 0.05 have almost become the passport for publishing papers. Obviously, the importance of p<0.05 is self-evident, which makes researchers consciously and unconsciously tend to pursue p<0.05, and take some measures to make p<0.05 during data collection and processing. This practice is called 'p-hacking'. For example, for the t-test of two sample means, as long as the sample size is increased, p<0.05 will always be satisfied, and the difference between the two sample means (i.e., the effect size) will hardly change with the sample size (Grenville, 2019).

The researcher will consider that an important scientific discovery has been made based on the p-value of being less than a certain threshold (for example, 0.05 or 0.01) in significance testing. It also shows that scientific research is both long-term and phased. Qualified scientific researchers should be aware that it is impossible for any one study, or even a series of studies, to reach a definitive conclusion. However, for each research it is possible to draw staged conclusions. It is especially important that the scientific research process also needs to be standardized and easy to implement. Therefore, following the procedure of statistical significance tests to conduct scientific research is recognized still by many researchers. It is a normative and feasible research route. As for some researchers to manipulate the statistical significance tests, it is purely academic misconduct (Li, 2022).

There is no necessary relationship between the p-value and the actual value. For example, the research proves that a drug has the effect of lowering blood sugar, but it does not mean that this effect has the value of treating diabetes, because it may be that the lowering effect (such as it can only be reduced from 15 to 13, but it has therapeutic value when it is reduced to below 8) is negligible for patients (Shao, 2018).

Amrhein et al. (2019) pointed out that downgrading the p-value to a dichotomous indicator that distinguishes between significance/non-significance is the main cause of the problem (Ziliak and McCloskey, 2008). They emphasize that this is not an argument for disabling the p-value, confidence intervals, or any other statistics, but only advocate that they (continuous statistics) cannot be processed/interpreted in discrete ways, including the binary interpretation of statistical significance, and the categorical interpretation of other statistics such as Bayesian factors (Xie, 2022a; Grenville, 2019). One reason to avoid such a dichotomy is that all statistics, including p-values and confidence intervals, varies hugely. In fact, just random fluctuations can easily cause huge changes in the corresponding p-values, and the range of which is far beyond the threshold of 0.05. Even if researchers were able to replicate the same research trial exactly and the study did have a true effect, further assuming 80% power (probability to achieve p<0.05), they would end up with the p-values that one is less than 0.01 and the other is greater than 0.30 (Xie, 2022a; Grenville, 2019). The problem is that human perception of cognition is not consistent with the meaning of statistical analysis results, which are classified as 'statistically significant' and 'not statistically significant', make people think that the classified results are qualitatively different. This kind of thinking of qualitatively differentiating results makes

it easy to ignore the real differences and focus on the false differences. We need a more precise method of research analysis to determine what is a 'significant' outcome (Xie, 2022a; Grenville, 2019).

In 2016, the American Statistical Association issued a formal statement on the correct and incorrect application of *p*-values (Wasserstein and Lazar, 2016). This statement clearly stated that p-values cannot be used as a basis for judging the validity of hypotheses and the importance of results (Sun, 2016). The statement states that the application of statistical significance (often in the form of $p \leq 0.05$) as a license to confirm a scientific discovery (or the truth it represents) has led to serious misrepresentation of the scientific process. A research result does not become 'true' just because it falls on one side of a two-division area, and 'false' on the other side. At the same time, the underlying principles to correctly use and interpretate *p*-values were proposed (Xie, 2022a; Grenville, 2019). Thereafter, Benjamini et al. (2019) further stated that: (1) *p*-value do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone; (2) A *p*-value, or statistical significance, does not measure the effect size or the importance of a result, and (3) by itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

In March 2019, *The American Statistician* published a special issue of statistical significance with 43 articles. The editor-in-chief of this special issue made it clear that the concept of 'statistical significance' should be completely discarded, and we should not state 'statistically significant'. They thought that no *p*-value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into 'statistically significant' and 'not statistically significant' is taken as an imprimatur of authority on these characteristics.

In March 2019, more than 800 scientists jointly called for abandoning the whole concept of 'statistical significance' in *Nature* (Grenville, 2019).

To date, all debates in the *p*-value based statistical significance tests have essentially fallen into three main categories:

(1) Reservation: For example, Li (2021a, 2022), etc., believe that the statistical significance tests and the corresponding *p*-value are infallible. As for the criticisms and accusations of some researchers, it is just that many researchers have a problem with the correct use of statistical significance tests. The *p*-value is just a tool, and is misunderstood and abused. Really it is the user's problem, not the *p*-value itself. Li (2022) pointed out that *p*-value itself is not 'correct' ' or 'wrong'. Because the physical meaning of *p*-value itself is not very clear, it is easy to be misunderstood. As a comparison, the physical meaning of the statistic, sample mean, is very clear and will not be misunderstood. Because *p*-value represents the probability value, the probability value itself is unitless, that is, dimensionless. Thus, this is not a disadvantage of *p*-value, but an advantage. Different sample sizes have different statistical power. Or, different sample sizes results in different data quality, different errors, and different results in statistical significance tests. So the methods of statistical significance tests cannot be rejected (Li, 2022; Xie, 2022b).

(2) Improvement: They acknowledge the limitations of significance testing and *p*-values, and propose further improvements, for example, increasing the popular threshold of statistical significance from 0.05 to 0.005, etc., as an imperfect short-term solution (Grenville, 2019; Xie, 2022c).

(3) Abandonment: It is believed that significance testing and *p*-value are invalid or even harmful, and need to be banned and abandoned (Xie, 2022a-e). For example, Ziliak and McCloskey (2008) completely rejected significance testing. Another example, *Basic and Applied Social Psychology* officially banned the use of significance testing and confidence intervals in 2015 (Trafimow and Marks, 2015).

I personally hold that the *p*-value based statistical significance tests should be carried out more strictly, and the multi-party cross-validation of the research results should be combined with statistical analysis and mechanism analysis, etc.

2.3 Mistake: Fisher's Significance Testing + Neyman/Pearson Testing of Hypothesis = Null Hypothesis Significance Test (NHST)

In the 1930s, two statisticians, Jerzy Neyman and Ergon Pearson, studied hypothesis testing: that is, among competing hypotheses, decisions are made based on experimental data only (Neyman and Pearson, 1933). Neyman believes that hypothesis testing is an improvement on the significance test, and Fisher reject Neyman's viewpoint. Neyman and Fisher fought over which test method was better until Fisher died (Grenville, 2019).

Fisher's 'Significance Testing' is a concept that is substantially different from the 'Testing of Hypothesis' laterly proposed by Neyman/Pearson. Fisher's 'Significance Testing' does not require the so-called 'Alternative Hypothesis'. Statistical power (statistical power), confidence interval, type I error/type II error and other concepts, also only belong to Neyman/Pearson's 'Testing of Hypothesis'. The paradigms of statistical inferences based by 'Significance Testing' and 'Testing of Hypothesis', are completely different. Fisher's 'Significance Testing' uses the induction (from special cases to general) to infer the most probable conclusion, while Neyman/Pearson's 'Testing of Hypothesis' draws the conclusion by using deduction (mathematical logic deduction). Fisher's 'Significance Testing' emphasizes that staged conclusions are drawn from a set of sample data, while Neyman/Pearson's 'Testing of Hypothesis' emphasizes the upper bounds governing the eventual Type I and Type II errors of a sampling process (the same sampling event is repeated multiple times). Neyman pointed out that the statistical power is lower than the specified significance level, e.g., α=0.05, β>0.95, the results of Fisher's 'Significance Testing' are mathematically worse than useless. For example, if allowed to set in the example of The Lady Tasting Tea, H_0: The probability that Bristol can make a correct judgment = 0.5; H_1: The probability that Bristol can make a correct judgment = 0.501, which is likely to cause the statistical power to be lower than the significance level. On the other hand, Fisher accused 'Testing of Hypothesis' is not a true statistical testing, but a decision rule (Xie, 2022b; Salsburg, 2002).

Fisher and Neyman/Pearson's debate on the theory and method of statistical testing reflects a deeper problem that logic and probabilistic arguments are incompatible (Xie, 2022b). In other words, from a mathematical point of view, Fisher's 'Significance Testing' is not logically rigorous, but Fisher treats or uses the results of the analysis of a certain set of sample data as evidence to judge the validity of the hypothesis concerned. On the other hand, although Neyman/Pearson's 'Testing of Hypothesis' is logically rigorous, but loses its original meaning of statistical testing. The analysis results of any set of sample data cannot be used as evidence to judge the authenticity of the relevant hypothesis. Researchers can only decide to accept the null hypothesis or reject it according to the results. Therefore, as far as the analysis results of any set of sample data are concerned, neither Fisher's 'Significance Testing' nor Neyman/Pearson's 'Testing of Hypothesis' can theoretically enable us to confirm the authenticity of scientific hypothsis. In view of this, starting in the 1940s, the authors of statistical textbooks (Snedecor and Cochran, 1991) tried their best to construct a seemingly objective statistical inference method, wishing to combine Fisher's 'Significance Testing' with Neyman/Pearson's 'Testing of Hypothesis' and finally construct the 'Null Hypothesis Significance Test' (NHST). Almost all standard statistics textbooks use such a mixture method that has never been theoretically proved to be the correct (Li, 2022). Of course, some textbooks have the NHST paradigm taken from more of Fisher's 'Significance Testing' and some from more of Neyman/Pearson's 'Testing of Hypothesis'. These NHST statistical testing paradigms have a common feature. They all emphasize that the analysis results (such as *p*-values or 95% confidence intervals) of a certain set of sample data can be obtained

by only following a few mechanical steps, and then they are regarded as or used as evidence to judge the truth or falsehood of the relevant hypothesis (Xie, 2022b).

In addition, hypothesis testing theory is based on sampling distribution theory, and the basic condition is random sampling. To achieve random sampling, it is necessary to define the sampling population first. In reality, it is often difficult to define the sampling population and conduct random sampling. In order to minimize the influence of confounding factors, an experimental design is required, otherwise, the study is just an observational experiment, not a controlled experiment. For observational experiments, the basic requirements of NHST conditions cannot be met. Although all of standard statistics textbooks stress the basic conditions required by these hypothesis testing paradigms, most studies analyze actual data and check whether these conditions are met, performing NHST as long as there is a set of sample data. Unfortunately, the NHST, a mechanical ritual-like hypothesis testing paradigm, is very consistent with people's white or black thinking mode, and also in line with the wishful thinking of statistics textbook authors to replace scientific experiments with statistical testing. NHST has being popular since the 1950s and 1960s (Xie, 2022b).

To this end, Gigerenzer (2018) concludes in his article that authors of early statistics textbooks struggled to develop a seemingly objective method of statistical inference that would mechanically separate true causes from randomly varying phenomena , without the user's additional thinking and judgment. As a result, Fisher's method and Neyman/Pearson method are forcibly kneaded together. The core of this hybridization theory is the Null Ritual of Neyman/Pearson (Gigerenzer, 2018; Grenville, 2019):

(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 $p<0.05$, $p<0.01$, or $p<0.001$, whichever level is met by the obtained p-value.

(3) Always perform this procedure.

Clearly, Null Ritual is wrong. This is not always understood, and even critics of Null Ritual sometimes confuse it with Fisher's null hypothesis testing theory and call it NHST (Grenville, 2019).

An example of one-way ANOVA can be used to illustrate the irrationality of NHST (Crawley, 2012; Wasserstein and Lazar, 2016; Li, 2022; Xie, 2022b). Suppose a crop is given three fertilizers, i.e., three treatments A, B, C; 10 data for each treatment (sample size is 30); the experiment is a completely randomized design. The results of data analysis (95% confidence interval) are: treatment A (7.8, 12.0), treatment B (9.4, 13.6), treatment C (12.2, 16.4). According to the NHST standard, because the confidence intervals do not overlap, the average output of treatment C is higher than that of treatment A; but because the confidence intervals overlap, it is impossible to judge whether treatment B has a higher average output than treatment A. If each treatment takes 5 data (sample size is 15), the data analysis results (95% confidence interval) become: treatment A (6.6, 13.2), treatment B (8.2, 14.8), treatment C (11.0, 17.6). Obviously, according to the NHST standard, only because the sample size becomes smaller, all comparisons are not statistically significant. Similarly, if each treatment takes 100 data (sample size is 100), the data analysis results (95% confidence interval) becomes: treatment A (9.2, 10.6), treatment B (10.8, 12.2), treatment C (13.6, 15.0), then there is a statistically significant difference between two treatments. The conclusion is that, even under the most ideal experimental design and correct application conditions, the results of NHST directly depend on the sample size. Another problem is that even we follow the most appropriate experimental design to collect data, there are still confounding factors that cannot be excluded. For example, when the same experiment is conducted by region and year, the results vary and the conclusions about statistical significance also vary (Crawley,

2012; Wasserstein and Lazar, 2016; Li, 2022; Xie, 2022b).

The more fundamental problem is that the sampling population is uncertain, and the obtained sample data cannot satisfy the basic conditions of sampling distribution theory. Therefore, the generalization of statistical anaysis results is not valid based on sampling distribution due to the violation of those fundamental assumption conditions. Both *p*-value and confidence interval are continuous variables, dividing them into 'statistically significant' or 'not statistically significant' will produce a contradiction that cannot be justified. From the classic example of statistical analysis, the irrationality of NHST can be clearly seen (Hahn and Meeker, 1993; Xie, 2022b, 2022d).

2.4 *p*-value manipulations and academic misconducts

2.4.1 *p*-value hacking, base rate fallacy and publication bias

Bergstrom and West (2021) have summarized several common misconducts in science, *p*-value manipulations, base rate fallacy and publication bias.

(1) *p*-value hacking

In current publishing paradigm, whether a paper has a chance to be published is affected by the *p*-value it reports. However, those papers that do get published are a biased sample of the full experiment. In the literature, statistically significant results are often overrepresented, and statistically insignificant results are underrepresented. Experimental data that do not yield significant results end up being thrown into the filing cabinet by scientists, the so-called file drawer effect (Bergstrom and West, 2021).

Goodhart's law points out that if an indicator becomes a target, it is no longer a good indicator. In a sense, *p*-value has this characteristic. Because the *p*-value lower than 0.05 is a good indicator for paper publication, so it is no longer a good measure of statistical support. If *p*-value is irrelevant to whether or not a scientific paper is published, then *p*-value will still be a valid measure of statistical support for rebutting the null hypothesis. *p*-values have fallen out of use as journals clearly prefer papers with *p*-values below 0.05 (Bergstrom and West, 2021).

(2) Base rate fallacy

Disease A is known to be rare. Someone who has a blood test for fear of contracting disease A and the test result is positive, but there is a 5% chance of a false positive. The probability of that person having disease A is intuitively 95%, but the actual situation proves to be wrong. For someone without disease A, the probability of testing negative is of course 95%. But the probability of testing positive for disease A is low because disease A is rare. Bergstrom and West (2021) point out that in an area where disease A is endemic, only 1 in 1000 people is infected. Assuming that 10000 people are tested, then one can expect about 10 true positives and about $0.05 \times 10000 = 500$ false positives. In testing positive cases fewer than 1 in 50 of those are actually infected. Therefore, even if they test positive, the probability of getting infected will not exceed 2% (Bergstrom and West, 2021). Treating less than 2% probability of being infected as having a 95% chance of being infected is a common error. We sometimes call this the base rate fallacy because the base rate of the disease in a population is ignored when interpreting testing results (Bergstrom and West, 2021).

However, if disease B is known to be common, e.g., with an incidence of 20%, then the base rate fallacy is not a big problem. For example, 5% of people who test without disease B will be positive. Assuming that 10000 people are tested, there will be about 2000 true positive results, and of the remaining 8000 people, this probability is about 5%, or about 400 people will get false positive results. Therefore, of the people who test positive , about 5/6 of the people are actually infected with disease B (Bergstrom and West, 2021).

(3) Publication bias

Ioannidis (2005) has drawn an analogy between scientific research and the interpretation of medical test results. He argues that due to publication bias, most negative findings are not published, so we see mostly

positive results in the literature (Fanelli, 2012; Fig. 3). If the impossible hypothesis is tested, then most positive results should be false positives, as in the disease A example abobe. If there are no other risk factors, positive testing results are mostly false positives (Bergstrom and West, 2021). Many experiments published in excellent journals cannot be replicated. If many of the positive results of these experiments are false positives, it is exactly what we would expect (Bergstrom and West, 2021).

About how to test for publication bias, the FDA's Eric Turner has found a way to solve this problem. US law states that any research team conducting clinical trials (trials that use people as experimental subjects to test the results of a treatment) must register with the FDA, submit documents and explain what the trial is going to test, how the trial will be conducted, and how the results will be measured. Once the trial is complete, the team also needs to report the trial results to the FDA. However, they are not required to be published in a scientific journal (Bergstrom and West, 2021). This system facilitated Turner and his colleagues to count published and unpublished trials in a particular field of study. Turner listed 74 evaluations of 12 different antidepressants clinical trials of efficacy, of which 51 trials have published results, including 48 positive results (the drug is effective) and 3 negative results. Looking at these published literature, any researcher would think that these antidepressants are usually is valid. But after investigating all the trials initially registered, the FDA found that the situation was not what had been expected. Of the 74 trials, 38 yielded positive results, 12 yielded equivocal results, and 24 yielded negative results. As a result, it is possible for us to draw a more pessimistic conclusion: it seems that only a subset of antidepressants can work in some cases (Bergstrom and West, 2021). Thus clinical trials of a success rate of 51% only were ultimately published in 94% of papers claiming success. One reason is that almost all positive results are published, while less than 1/2 of equivocal or negative results are published. Another and more important reason is that of the 14 published equivocal or negative results, 11 were redefined as positive results (Bergstrom and West, 2021).

Professional journals generally hold a negative attitude to publishing insignificant research results, and scientists must rely on publishing articles to gain promotion or even keep their positions, and the consequences are disastrous for the scientific career (Timmer, 2009; Xie, 2022a; Grenville, 2019).

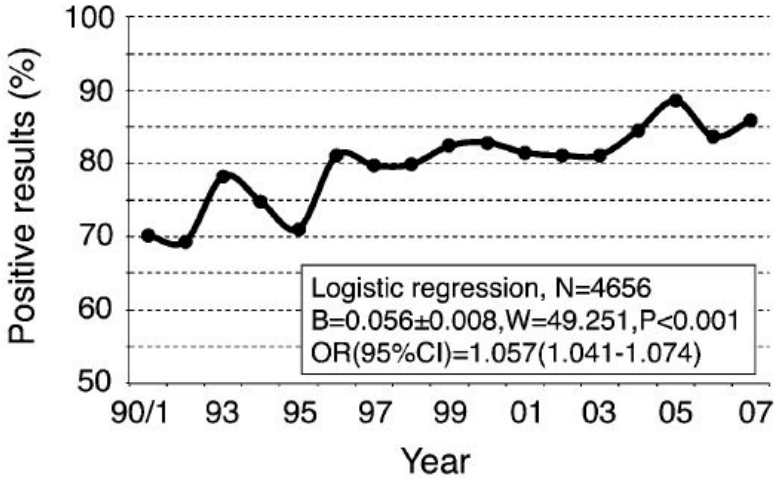


Fig. 3 The proportion of papers reporting a positive result in the sample has been fluctuatingly rising since 1990. On average, the rates of reporting positive results have increased by around 6% every year, showing a statistically highly significant trend (Source: Fanelli, 2012).

2.4.2 Academic misconduct

The Dutch National Survey on Research Integrity (NSRI) has conducted the largest survey of academic misconducts to date, with more than 64000 researchers invited to participate in an anonymous survey (Vrieze, 2021). The results of this survey showed that, 53% of PhD students admitted to having regularly engaged in one of 11 questionable research practices in the past three years, compared to 49% of associate and full professors. 8.3% admitted to more serious research misconducts committed in the past 3 years: 1 in 12 people admited to falsifying research results at least one time. Life sciences and medicine are still the hardest hit areas for academic misconducts (EVEE, 2022).

As mentioned earlier, the excessive pursuit of positive results has distorted the yardstick that science itself respects objective facts. A 2012 study showed that from 1990 to 2007, the proportion of positive results in published papers increased by more than 22%. In 2007, more than 85% of papers claimed to have found positive results, but Fanelli, the author of the study, believes that the scientific objectivity of published papers is declining (Fanelli, 2012; Fig. 3).

In 2020, *Nature* published an article titled "Fraud, bias, negligence and hype in the lab — a rogues' gallery" featuring the case of ever renowned nutritional psychologist Brian Wansink. The most famous expert on dietary behavior in the world, was appointed executive director of the USDA Center for Nutrition Policy and Promotion, and directed the National Dietary Guidelines. However, many of Wansink's papers have been exposed to the existence of fraud (falsified data), bias (academic bias), negligence (ignoring mistakes), hype (malicious hype), etc. (Wansink, 2014; O'Grady, 2017; The Skeptical Scientist, 2021). The key point that draws attention to his research results is his unprincipled abuse of statistical tests, or 'p-hacking' (XKCD, 2021). Whenever a study was carried out with a negative result, Wansink passed some perverse solution, such as dismantling data to obtain an ideal *p*-value for an otherwise insignificant result. He openly encouraged graduate students and his collaborators to confirm ' statistical significance' by deliberately searching through the data (Xie, 2022a; Grenville, 2019). The problems identified above were present in 52 publications that Wansink participated in, including articles in 25 different journals and 8 books, which were cited more than 4000 times. During the two years, a total of 18 papers were retracted, with a maximum of 6 papers being retracted in one day. When the evidence of these unscientific research results of Wansink was exposed, Cornell University removed him from research and teaching activities. Consequences of excessive pursuit of positivity have been fulfilled in Wansink. However, under the existing evaluation system, there are still many people who are doing so (Fidler, 2020; Ritchie, 2020).

The fundamental reason behind this phenomenon is that it is difficult for the scientific community to accept negative results. Fortunately, the scientific community's view on negative results is also gradually changing, and at least some databases and peer-reviewed journals are now willing to accept negative results for publishing.

In fact, the significance of a negative result is far greater than we imagined: on the one hand, a negative result has a strong warning effect, which can indicate to peers that the road is not available, and on the other hand, a negative result may also inspire future generations to create an opportunity to overturn existing theories (Mehta, 2019).

2.5 The reproducibility crisis in science

Hypothesis testing and *p*-value have been strongly questioned and severely challenged in recent years. The background is that many new scientific research results and discoveries are considered to be false positives and cannot be confirmed by repeated experiments, leading to the so-called reproducibility crisis. Many scientists believe that the application or misuse of significance tests and *p*-values is one of the main causes of false positives and reproducibility crisis (Baker, 2016b; Huang, 2021a, b). Of course, it is obviously inappropriate to

blame the reproducibility crisis entirely on the significance tests and *p*-value, but the reproducibility crisis reflects the methodological limitations and defects of the significance tests, that is, the statistical significance tests paradigm itself has its own problem. In other words, the reproducibility crisis as an abnormal or counterexample under the statistical significance tests paradigm is the source of the crisis in the statistical significance tests paradigm (Grenville, 2019; Bergstrom and West, 2021).

A large amount of evidence shows that many published articles cannot be replicated due to the wrong application of statistical tests. *Nature* and *Science* are two of the most authoritative professional journals, and scientists strive to make their articles recognized by these two journals. However, a research article published in *Nature* (Camerer et al., 2018) stated that they were only able to replicate the results of 60% of the published articles, and the results obtained by repeating the validation study showed that the effect size was on average about 50% of the effect size reported in the original article (Camerer et al., 2018; Grenville, 2019; Xie, 2022a).

Similarly, ugly replicability and reproducibility of analysis data results have also been exposed in other research fields, including psychology (Open Science Collaboration. 2015; Fig. 4), economics (Camerer et al., 2016), and medicine (Prinz et al., 2011). A case of unprincipled reliance on significant findings is Brian Wansink (Wansink, 2010).

Although many cancer studies claim to have achieved a significance of 0.05, many results are not replicable. Both *Nature* and *Science* have devoted discussions on how to understand *p*-value and how to define the significance of a study (Servick, 2017; Nuzzo, 2014; Shao, 2018). This is because more and more scientists and statisticians believe that the misuse or abuse of significance tests is one of the main reasons for the reproducibility crisis, and that reproducibility crisis reflects a methodological flaw in significance tests, that is, the significance tests paradigm itself is problematic. Therefore, some scientists and statisticians have called for a complete abandonment of the concept of statistical significance and significance tests. For example, a 2019 editorial in *The American Statistician* declared that it is time to stop using the term 'statistical significance' altogether (Wasserstein et al., 2019). As noted, the American Statistical Society issued a statement in 2016 stating that the misuse of *p*-value has become an important reason why many studies cannot be replicated. Some guiding principles for the use of *p*-value are proposed, and it is clearly stated that misuse of *p*-value cannot be used as a basis for judging the authenticity of hypotheses and the importance of research results (Sun, 2016).

In 2005, a paper published in PLOS Medicine, "Why most published research findings are false", first sparked widespread discussions about the reproducibility of scientific findings (Ioannidis, 2005; Wu, 2022; Zhang, 2022), which once again caused a shock in the scientific community. In 2021, a project started in 2013 and costing a total of 2 million US dollars to verify the top research in the field of cancer biology-Reproducibility Project: Cancer Biology, RPCB, published the complete replicated results in eLife (Errington et al., 2021; Wu, 2022; Zhang, 2022).

The RPCB project aims to replicate 193 experiments in more than 50 high-impact top journals such as *Nature*, *Science*, *Cell* and other papers under the same conditions and in as similar ways as possible. Over a period of 8 years, it took an average of 197 weeks to replicate all experiments in each study and cost $53000 each. Ultimately, the project replicated only 50 experiments out of 23 papers. Based on conclusions (yes or no) and given 5 criteria of statistical significance, only 47% of the 23 papers could get a conclusion that was completely consistent with the original experiment. If the effect size was considered, then only 25% of the repeated experiments had the effect size in the 95% confidence interval of original result (Errington et al., 2021; Wu, 2022; Zhang, 2022).

In addition, according to a survey of 1576 people sponsored by *Nature*, more than 70% of the researchers said they had been unable to replicate the experiments of other groups; more than 50% of the researchers said

they could not replicate their own experiments; 52 % of the respondents believe that there is a major crisis of experimental reproducibility. Among the respondents, most scientists indicated that they have experienced the failure of repeated experiments, and the biological field ranked second (Baker, 2016a; Fig. 5).

In 2021, *Nature* dedicated an editorial, emphasizing that researchers, research funders and publishers must take reproducible research efforts more seriously (Nature Editorial, 2021). Errington even proposes to elevate reproducibility to the same level as research novelty: reproducibility is an important feature of scientific research. However, contemporary research culture tends to emphasize features such as novelty, while placing repproducibility on a secondary level (Zhang, 2022; Errington et al., 2021).

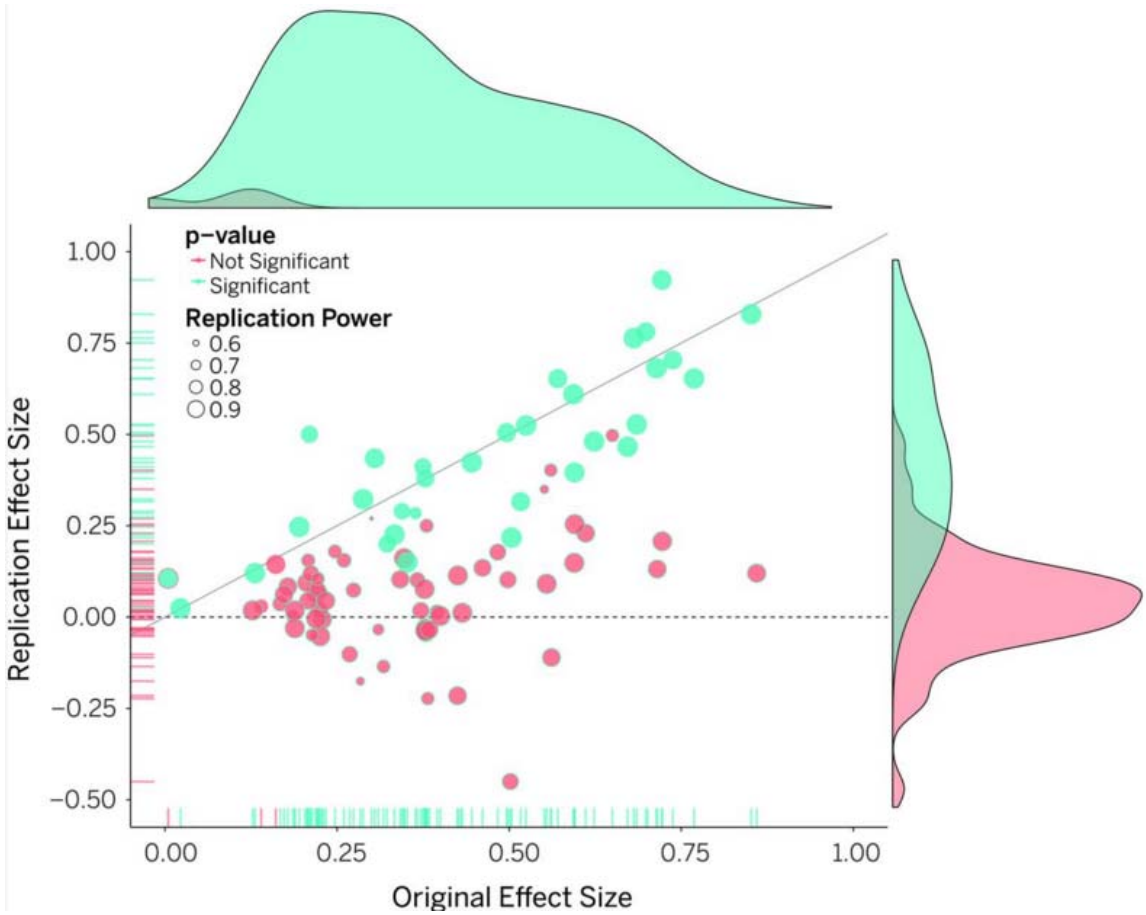


Fig. 4 Correlation coefficients of original study effect size vs replication effect size. Diagonal line represents that replication effect size equals to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects (Source: Open Science Collaboration, 2015).

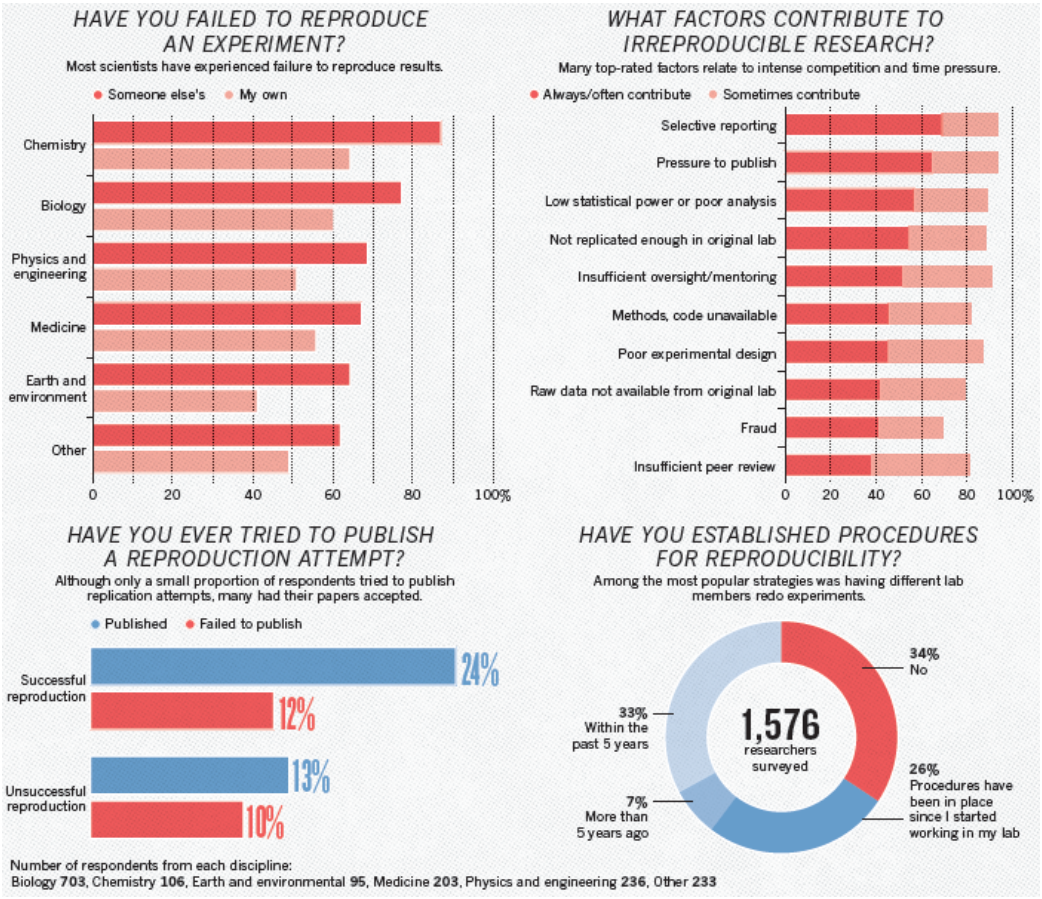


Fig. 5 Most researchers indicated that they had the experience of repeated experiment failure (Source: Baker, 2016a).

3 Actions

The unprincipled use of a hybrid *p*-value based statistical significance tests has led to the massive misuse and abuse of the concept of statistical significance, and a crisis of large-scale irreproducibility of research findings. In light of this, in 2016 American Statistical Association (ASA) issued a formal statement for the correct and incorrect application of *p*-values (Wasserstein and Lazar, 2016). It warns against the misuse of statistical significance and *p*-values, and explicitly states that *p*-values cannot be misused as a judgment of hypothesis truth or result importance. Based on this, some guiding principles for the use of *p*-values are proposed (Sun, 2016). This statement suggests that researchers must recognize that, without relevant scientific connotation-based or other evidence, the information provided by a *p*-value only is limited. The statement advises researchers that many factors related to scientific content need to be taken into account in order to draw scientific inferences. These factors include the design of the study protocol, the quality of the observational data, the external evidence (obtained outside the study) of the observed phenomenon, and the reasonableness of the assumptions underlying the data analysis. This statement asserts that good statistical analysis emphasizes understanding of the observed phenomenon, and analysis results should have scientific connotation explanations, complete reports and reasonable logic, as well as quantitative explanations represented by data analysis results. No single statistical indicator can replace scientific reasoning analysis. The content of the research question is the real connotation (Xie , 2022a; Grenville, 2019). According to Ron Wasserstein, executive director of the society, this is the first time in 177 years that the ASA has issued a cautionary statement on this most fundamental statistical principle, mainly because the ASA is very concerned about harmful outcomes from misuse of *p*-values. ASA recommends that scientists avoid drawing

scientific conclusions and formulating policies based on *p*-value alone. Scientific reports should not only describe the final statistical data resulting from the data analysis, but should also provide all statistical test calculations, otherwise the reliability of statistical analysis results is difficult to judge (Sun, 2016).

Further, the ASA organized a two-day symposium on statistical inference in October 2017, the results of which constituted the 43 articles published in the special issue of *The American Statistician* (TAS) Volume 73 in March 2019 on statistical significance (Hubbard et al., 2019; Tong, 2019). The album of TAS volume 73 begins with an "Editorial: Moving to a World Beyond "p<0.05"" (Wasserstein et al., 2019), in which a summary of the views and proposals of the 43 articles published in this special issue is made. The editorial clearly proposes to completely abandon the concept of 'statistical significance' - the concept of the cornerstone of the NHST, and not to mention 'statistically significant', (Hubbard et al., 2019; Tong, 2019; Xie, 2022b, c).

In March 2019, more than 800 scientists jointly called for abandoning the whole concept of statistical significance in *Nature* (Grenville, 2019; Amrhein et al., 2019; Huang, 2021a, b). Amrhein et al. (2019) suggested that hypothesis testing should 'retire'. Clearly, the era of *p*-value dominance is over (Halsey, 2019), and academia will enter the post p<0.05 era (Wasserstein et al., 2019).

Researchers must be aware of the fact that in many research fields, it is rare that an experiment is the real key to the experiment, and it is more common to need to conduct multiple experiments on the same scientific research question. These experimental results are aggregated to obtain a comprehensive result that conforms to the scientific truth. For example, in agricultural field experiments, the effects of experimental treatments will generally vary with soil and meteorological conditions. The general applicability of the research results makes it absolutely necessary to replicate the same experiment in different regions and years. In this case, a series of moderately accurate experiments is more valuable than a single but very high accurate experiment. (Yates, 1951; Xie, 2022d).

Xie (2022c) pointed out that statistical analysis cannot remove the inherent uncertainty of the data itself. The application of the statistical analysis paradigm of the NHST makes statistics often regarded as an alchemy that can remove the inherent uncertainty of the data itself: Input data with uncertainty, and output results of measured successufully by statistical significance tests on the experimental data (whether the experimental treatment variable is really effective). Accepting the uncertainty principle, first requires we seek better measurement variables, better experimental designs, and larger sample sizes. Accepting uncertainty also drives us to be more realistic. Accepting the uncertainty principle and accepting dichotomies measured by statistical significance are incompatible with each other. To accept the uncertainty principle, naturally prompts us to pay attention to the reproducibility of research results, and to seek the verification of the evidence formed by the synthesis of the results of other similar independent research experiments. Consider using statistics including *p*-values and various other statistics as tools/indicators for statistical analysis to obtain answers to questions, seek and report probability distributions of estimated effect sizes, and provide sufficient detail to warrant the reported results to be reproducible (Xie, 2022c). We need to understand and clearly tell everyone the limitations of our research results, and recognize the gap between the complexity of the real-life phenomena we study and the statistical models that we have established. The so-called only correct statistical model does not exist. We should recognize that scientific inference/reasoning is a broader concept than statistical inference. In conclusion, we should acknowledge that scientific inference/reasoning and statistical inference are both difficult and complex, recognizing that knowledge discovery cannot depend on implementation of simplified and mechanical rules and procedures (Xie, 2022c).

In order to realize the reform of statistical analysis that completely abandons the concept of statistical significance in any aspect or level, the existing organizational system related to this must first be reformed.

There must be a mechanism to support and affirm the original scientific research results to carry out replicated experimental research. Professional journals should first and more pay attention to the quality of research design, data and methods, and complete transparency of reporting before considering the validity and significance of its findings when evaluating whether submitted articles can be published. Finally, a major reform change is needed in statistics education to accommodate the post $p<0.05$ era (Xie, 2022c).

Applying statistical analysis from a scientific point of view (e.g. reproducibility, universality, the mechanism/mechanism behind the phenomenon of things) is better than simply following a stylized mechanical black-and-white analytical paradigm. Analyzing conclusions is much more difficult. Therefore, this reform must be carried out in different fields and different groups of people, which have different difficulties, and the time required will be different also (Xie, 2022c).

Dealing with reproducibility and uncertainty is at the heart of statistical science. Research findings are reproducible if they can be validated in further research with new data. Leaving aside the possibility of fraud, the important sources of poor reproducibility include poor study design and implementation, insufficient data, lack of understanding of model selection, inadequate description of analytical and computational procedures, and selective choice of reported results. Selective reporting can lead to distorted perceptions of the evidence even some of the persuasive results are highlighted in the reporting. In some cases, this problem can be mitigated by adjusting for multiplicity. Controlling and explaining uncertainty should begin with the design of the research and observation process, and continue through each stage of the analysis until the results are reported. Even in well-designed and well-implemented studies, inherent uncertainty remains and statistical analysis should appropriately account for this uncertainty (Benjamini et al., 2021).

p-value based statistical significance theory as the mainstream paradigm of statistics has a history of nearly 100 years, but now it has encountered unprecedented severe challenges. The debate on whether to reserve, improve, or abandon these mainstream paradigms is still ongoing. However, the challenges are also opportunity: they prompt us to re-examine these mainstream paradigms and seek to create new paradigms (such as developing more reasonable statistical inference methods). Especially when compiling new statistical textbooks, we need to consider whether to reserve these concepts or theories. If they are reserved, the author should point out their limitations and flaws, rather than some erroneous or misleading knowledge being taught to students as 'truth' as in today's textbooks (Huang, 2021a, b).

The scientific question: "Are these differences meaningful?" is clearly more appropriate than "Is that result statistically significant?" (Xie, 2022a; Grenville, 2019). Asking this question prompts us to interpret an analysis results while avoids the dichotomy of thinking in the results, and can guide us to interpret the analysis results from the perspective of disciplinary content. Interpreting data analysis results from the perspective of disciplinary content is much more informative (Xie, 2022a; Grenville, 2019). We need to think about how to apply the analysis method of significance tests (Xie, 2022a; Grenville, 2019). Xie (2022a) further proposes that: (1) A so-called significant finding needs to be cautiously treated with skeptical attitudes, the rate of false positives is actually higher than the commonly believed; (2) we need to move away from fishing for data analysis and trawling for significant differences; (3) we need to recognize that some statistical test results of insignificance can also be scientifically meaningful, and (4) we need to take into account the scientific implications of the study (Xie, 2022a; Grenville, 2019).

Halsey (2019) argues that a "power" vacuum has emerged at the end of the era of *p*-value dominance, and discusses several statistics that may replace *p*-value to fill the "power" vacuum, including confidence intervals, Bayes factors, Akaike information Criterion (AIC). For example, Huang (2019) recently proposed a statistic with a potential alternative to *p*-value based on the law of conservation of energy: Signal content index (SCI). Statistics reform is also known as statistical reform, mainly reform of statistical inference. The

main point of scientists who advocate statistical reform is to abandon the significance tests and implement estimation of effect size (Huang, 2021a, b).

In fact, the calls for statistical reform have been around for a long time. Fidler and Cumming (2007) write that convincing arguments for reforming statistical practice have been presented in many disciplines, some for decades, but achieving reform has proven difficult. In 2014, Cumming, the main advocate of statistical reform, proposed the concept of new statistics, which includes estimation of effect size, confidence interval, meta-analysis, etc. (Cumming, 2013). Cumming showed that these new methods are not actually new, but widespread use of these methods is new to many researchers (Huang, 2021a b).

In recent years, the search for better statistics to replace *p*-value is becoming a very hot research topic in statistics (Huang, 2021a, b).

4 Solutions

4.1 Strict *p*-value's criterion

The *p*-value is a valid statistic that reflects the uncertainty inherent in quantitative results. In fact, *p*-values and significance tests are among the most studied and well-understood statistical methods in the statistical literature. They are an important tool to apply and advance science achieved through appropriate uses.

As early as July 2018, Benjamin et al. (2018) hoped to improve the standard of significance in scientific research, thereby improving the reproducibility of results. They recommended to replace the currently commonly used *p*-value criterion with a stricter *p*-value of 0.005 of 0.05. They believe that similar changes have been successful in other areas, such as sequencing data analysis, where more stringent *p*-values have been adopted. They believe that this approach can significantly reduce false positives in study results. The views of Benjamin et al. (2018) are divided (Servick, 2017). Even some proponents disagree on whether and how much significance level should be set in absolute terms. Among them, there is a view that such an initiative may exacerbate *p*-value hacking, which is to publish only positive results and hide others. Others worry that increasing *p*-value requires more samples whic requires more research funding. For example, in the general case (normal distribution), increasing the *p*-value from 0.05 to 0.005 may require 70% more samples (Shao, 2018).

Many biological and medical studies may not aim to prove that a conclusion is true, but to improve the posterior probability of a conclusion being true by summarizing and inferring quantitative data. A deterministic study can make us believe that a certain conclusion is true. The probability that the conclusion is true is close to 1. For problems with multiple explanations, whether these explanations are true or not are probabilistic events. Then the progress of research is to reduce the information entropy corresponding to these probabilistic events until we have a definite theory. From this point of view, a low *p*-value is beneficial. It can also be estimated from a realistic point of view, for example, someone calculates that 53% of preclinical studies are not reproducible (and about 28 billion US dollars of wasted funding)(Kaiser, 2015). In addition, when conducting experimental design and interpretation of results, the scientific judgment of the researcher (also the estimation of the prior risk) is also very important. When the prior probability of H_1 is weak, that is, when the results are very surprising, the stricter *p*-value tests are often required (Shao, 2018).

Benjamini et al. (2021) pointed out that many of the controversies surrounding statistical significance can be eliminated by better understanding the uncertainty, variability, multiplicity and reproducibility of the study subjects. When properly applied and interpreted, the use of *p*-values and significance is an important tool that should not be discarded. Thresholds are helpful when decisions need to be made. Although *p*-values themselves provide valuable information, matching *p*-values to significance levels may useful. *p*-values and statistical significance should be understood as assessments of measurements or effects related to sampling

change, not necessarily as measures of actual significance. If it is deemed necessary to consider thresholds as part of decision-making, then thresholds should be clearly defined according to research objectives, and the consequences of wrong decisions should be considered. As a matter of convention, threshold (significance level) criteria should vary by subject and analytical objectives (Shao, 2018; Benjamini et al., 2021).

Li (2022) argues that the reason for abandoning *p*-value and significance cannot be a strong basis, and does not agree with simply abandoning the statistical significance tests.

Xie (2022d) believes that *p*-value can measure the degree of inconsistency between sample data and a given statistical model. It is a continuous statistic, so as long as we do not use it as a discrete statistic, that's no problem. For example, if a *t*-test gets p=0.06, it can be concluded that the null hypothesis does not match the observed sample data to a large extent, but it is impossible to determine the reason for the inconsistency. It may be that the null hypothesis does not hold, or the null hypothesis is true but the sample is not representative, or it may be that a certain hypothesis does not meet the requirements (for example, the sample data are not independent of each other), and the reason needs to be further found through other analysis. Two different statistical models are fitted to the same set of data, and it can be confirmed that the model with the better *p* fits the data better.

In conclusion, *p*-values and significance tests, when properly applied and interpreted, add to the rigor of conclusions drawn from data. Analyzing data and summarizing results is often more complex than is commonly recognized. Although all scientific methods have limitations, however, the correct application of statistical methods is crucial for interpreting the results of data analysis and improving the reproducibility of scientific results (Benjamini et al., 2021).

4.2 Perform meta-analysis

Meta-analysis, i.e., looking at multiple studies on the same trial, is one of the most effective methods. With meta-analysis, it is possible to know whether the published literature is likely to be representative of all trials and whether they reflect some problematic behavior, such as *p*-value hacking, publication bias. Meta-analysis has become a hot area of statistical research (Bergstrom and West, 2021).

4.3 Using Bayesian methods

In 2013, the 250th anniversary of Bayes' Theorem, Efron published a commemorative article titled "A 250-year argument: belief, behavior, and bootstrap" (Efron, 2013). Efron mentioned that for the past two and a half centuries, Bayesian and frequentist schools have been competing with each other. The 20th century was dominated by frequentist schools, especially in applications, but the 21st century has seen a strong revival of Bayesian schools (Efron, 2013; Grenville, 2019; Huang, 2022).

Although both the Bayesian school and the frequentist school use the concept of probability, the two schools have different interpretations of probability. The frequentist school believes that probability is the frequency with which an event occurs in a large number of repeated trials. It is objective. For the frequentist school, it is meaningless to discuss the probability of not being able to repeat the experiment. The Bayesian school believes that probability is the degree of belief that an event occurs, which is subjective. Therefore, For an event that occurs one time only, probability (belief degree) can also be used (Efron, 2013; Grenville, 2019; Huang, 2022).

Frequentist school' statistical inference methods include *p*-value based statistical significance tests (such as *t*-test, *F*-test), parameters' point estimates, confidence intervals, etc. Because of the emphasis on the objectivity of inference, these methods usually only rely on data and subjective prior knowledge are not allowed to be used (except for necessary assumptions). Bayesian methods involve Bayesian factors, posterior probability distributions, and credible intervals, etc. (Benjamini et al., 2021). Bayesian school's statistical inference allows the use of subjective prior knowledge. New information (data) is combined with prior

knowledge to generate updated knowledge, namely (Efron, 2013; Grenville, 2019; Huang, 2022): updated knowledge = prior knowledge + new information (data).

Bayesian school's statistical inference is in line with people's cognitive processes. People always revise or update previous knowledge or beliefs based on newly acquired information. If you have a strong belief in a proposition (e.g., with 100% prior belief degree to represent), then regardless of the new information (current probability) about the proposition, Bayes Theorem states that your posterior belief degree is 100% of the same as your prior belief degree, that is, your belief is not affected by the new information. On the other hand, if you have no prejudice against a proposition (e.g., indicated by a prior belief of 50%), then if new information about the proposition gives a current probability of P%, Bayes Theorem says that your posterior belief degree of is also P%, that is, you completely accept new information. In real life, people will always be more or less influenced by new information to revise their previous thoughts (beliefs) (Shao, 2018).

Bayesian statistics has become mainstream in recent decades. For example, most of the papers on measurement of uncertainty published in *Metrologia* in the past ten years are based on Bayesian statistics.

The basis of the Bayesian school is Bayes Theorem (Bayes rule). Bayes Theorem is a methamatical rule, which expresses the interrelationships between the conditional, marginal, and joint propbability distributions of random variables, as defined in the following formula (Upton and Cook, 2008; Zhang, 2016, 2018; Pandey et al., 2022)

$$\Pr(B|A) = \Pr(A, B) / \Pr(A) = \Pr(A|B) \times \Pr(B) / \Pr(A)$$

where A and B are two random variables, $\Pr(A)$ and $\Pr(B)$ are the marginal probability distributions of A and B, respectively, $\Pr(B|A)$ is the conditional probability distribution of B given A, $\Pr(A|B)$ is the conditional probability distribution of A given B, and $\Pr(A, B)$ is the joint probability distribution of A and B.

Obviously, Bayes Theorem is expressed in terms of conditional probability (Huang, 2022), and it can also be expressed as: posterior probability $\propto$ prior probability $\times$ current probability.

Another form of Bayes rule states that the probability that a postulate A will be true is positively proportional to the multiplication of the postulate's prior probability and the conditional probability of information, I, being observed given H is true. Known a discrete sample space S, and suppose $A_i \in S$, i=1, 2, …, where $\cup A_i = S$, and $A_i \cap A_j = \phi$, $i \neq j$. Bayes rule is expressed as (Zhang, 2016, 2018):

$$\Pr(A_i / I) = \Pr(I / A_i) \Pr(A_i) / \sum \Pr(I / A_j) \Pr(A_j)$$

For the continuous sample space, Bayes rule is:

$$\Pr(A / I) = \Pr(I / A) \Pr(A) / \int \Pr(I / A) \Pr(A)\, dA$$

Bayes rule can be used in the selection of models (Mackey, 1992; Zhang, 2018). Suppose there are several models, and known the data or information I, the posterior probability of model A_i, is given by Bayes rule:

$$\Pr(A_i / I) = \Pr(I / A_i) \Pr(A_i) / \Pr(I)$$

where $\Pr(A_i)$: the prior probability of the model A_i; $\Pr(I / A_i)$: the evidence of the model A_i, which is expressed as:

$$\Pr(I / A_i) = \int \Pr(I / w, A_i) \Pr(w / A_i)\, dw$$

where $w = (w_1, w_2, \ldots, w_n)$ is a weight vector.

Based on Bayes factor and prior odds, let's examine the relationship between false positive rate and power given *p*-value and prior odds (Fig. 6). The Bayes factor is the probability of getting the measured x_{obs} data under the H_1 hypothesis divided by the probability of getting the measured data under the null hypothesis. The smaller the α (significance level), the larger the Bayes factor. Bayes factor and prior risk are linked together by the Bayesian formula. By measuring the data, the ratio of the H_1 hypothesis and the H_0 hypothesis is considered to be satisfied (Benjamin et al., 2018; Shao, 2018):

$$\Pr(H_1|x_{obs}) / \Pr(H_0|x_{obs}) = [f(x_{obs}|H_1) / f(x_{obs}|H_0)] \times [\Pr(H_1) / \Pr(H_0)] \equiv \text{BF} \times (\text{prior odds})$$

The Bayes factor can be regarded as the information about the H_1 hypothesis and the H_0 hypothesis obtained from the data. The prior risk is related to the specific problem of the researcher and the scientific consensus. The prior odds can be expressed as (1-ϕ) / ϕ, where ϕ is $\Pr(H_0)$, the prior probability that the null hypothesis holds (Shao, 2018). Essentially, *p*-value is just a measure of whether the result may be randomly generated under the existing experimental conditions, and its corresponding causal chain is that the smaller the *p*-value, the less likely the result will be randomly generated, and the larger the Bayes factor. But this does not mean that $\Pr(H_1|x_{obs}) / \Pr(H_0|x_{obs})$ is larger. Because according to the above formula , it also depends on the ratio of the probability of H_1 and H_0, which is the prior risk. Take an example, suppose the probability of H_0 is 1 and the probability of H_1 is 0, that is, a proposition cannot be true. But there are 50 labs do the same experiment, and there is a high probability that one lab will get a result with a significance of 0.05 to show that the proposition is true (the Bayes factor is still large) (Prinz et al., 2011). Obviously, such a result does not make any sense (Shao, 2018).

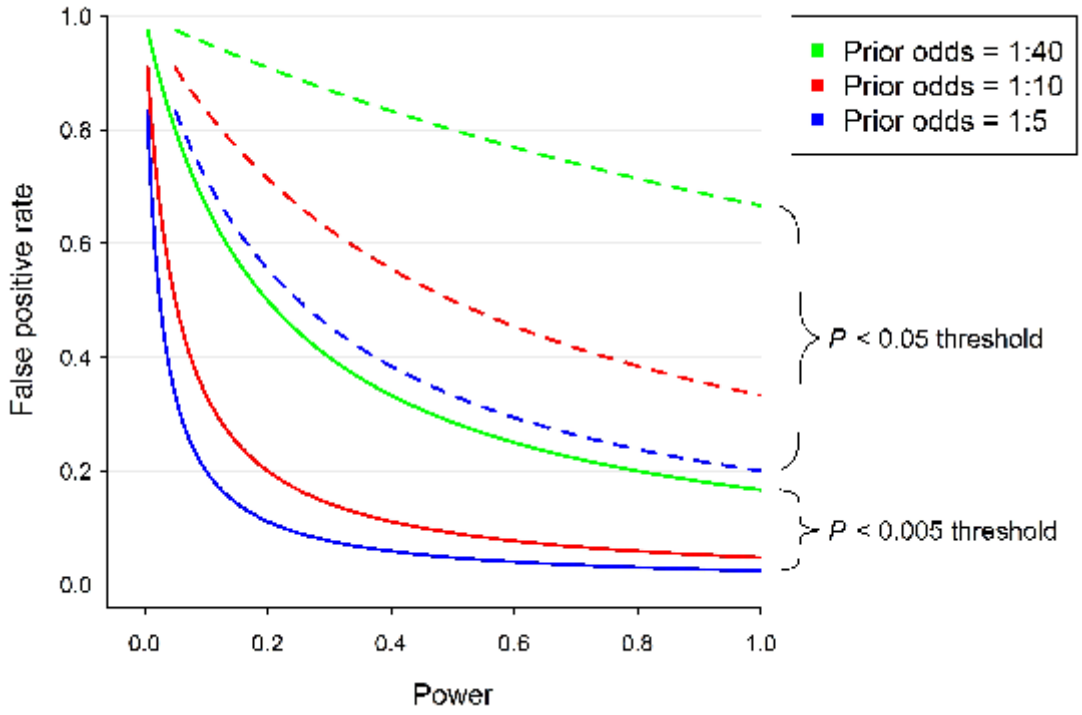


Fig. 6 The relationship between false positive rate and power given *p*-value and prior odds (Source: Benjamin et al., 2018).

Fig. 6 is based on the following formula:

$$\text{false positive rate} \approx \alpha\phi / (\alpha\phi + (1 - \beta)(1 - \phi))$$

The larger the ϕ, the greater the probability of false positive results. The stronger the significance, the smaller the α, the lower the false positive. The greater the statistical power (power), the smaller the β, the lower the false positive. You can find in this figure α is also the effect of the significance level. It is found that when $\alpha = 0.05$, if the prior risk is 1:10, then the probability of false positive results is at least greater than 33%. When we take a p value of 0.005, the probability of false positive results is at least 33%. The probability is less than 10% in many cases (Shao, 2018).

The false positive problem can be illustrated with a graph from Sellke et al. (2001) (Fig. 7). In this figure, the black is the probability that the H_1 hypothesis is true, and the yellow is the probability that the H_0 hypothesis is true. The first row is the prior risk, where on the left is prior odds = 1:19, in the middle is prior odds = 1:1, and on the right is prior odds = 9:1 (H_1 is more likely). The second row is the posterior probability of H_0 and H_1 at a significance level of 0.01 or 0.05. It can be found that on the right, when the probability of H_1 is relatively large, the two *p*-value results are not much different; on the left, when the probability of H_1 is about 5%, The posterior probability of H_1 at p=0.01 is about 2 times larger than that in the case of p=0.05. That is to say, in this case, the probability improvement brought by a smaller *p*-value is larger (2 times *vs* 6 times). For the probabilities of H_1 and H_0 are not much different, it can be found that the difference of changes from p=0.01 and p=0.05 is not very significant (1.4 times *vs* 1.8 times) (Sellke et al., 2001; Nuzzo, 2014; Shao, 2018).

The statistical power here is the probability of discovering a fact through experiments and other means. Statistical power is generally expressed as 1 - β, where β is the probability of type II error. For example, a drug can kill cancer cell, if the statistical power is 0.9, it means that there is a 0.9 of probability that it can be found by experimentation to kill cancer cells. Then in this case, there is a 0.1 of probability that even though the drug is effective, the cancer cells are not killed. The 0.1 here is the probability of type II error, which is the probability of false negative error (Camerer et al., 2018). Statistical power is related to experimental reliability and should ideally be close to 1, but in reality, for more complex experiments, it is often not so high (Shao, 2018).

There are many excellent examples of using Bayesian methods to solve statistical inference problems. For example, Pandey et al. (2022) successfully applied Bayesian Network for statistical inference in their research (Pandey et al., 2022; Xie, 2022e; Fig. 8). The Bayesian Network (BN) method is based on Bayes Theorem. It can be used for determining the complex interrelationships between the variables. A BN model is a graphical representation (i.e., a directed acyclic graph) of a joint probability distribution of a set of random variables in which each variable is represented by a node and the dependency relationship is represented by a link for two associated variables (Pearl, 1988; Kjærulff and Madsen, 2008; Pandey et al., 2022; Xie, 2022e).

Shao (2018) believes that although the Bayesian method is effective for some problems in some fields, it is of little significance for practical applications, and the Bayesian method complicates the originally simple problem.

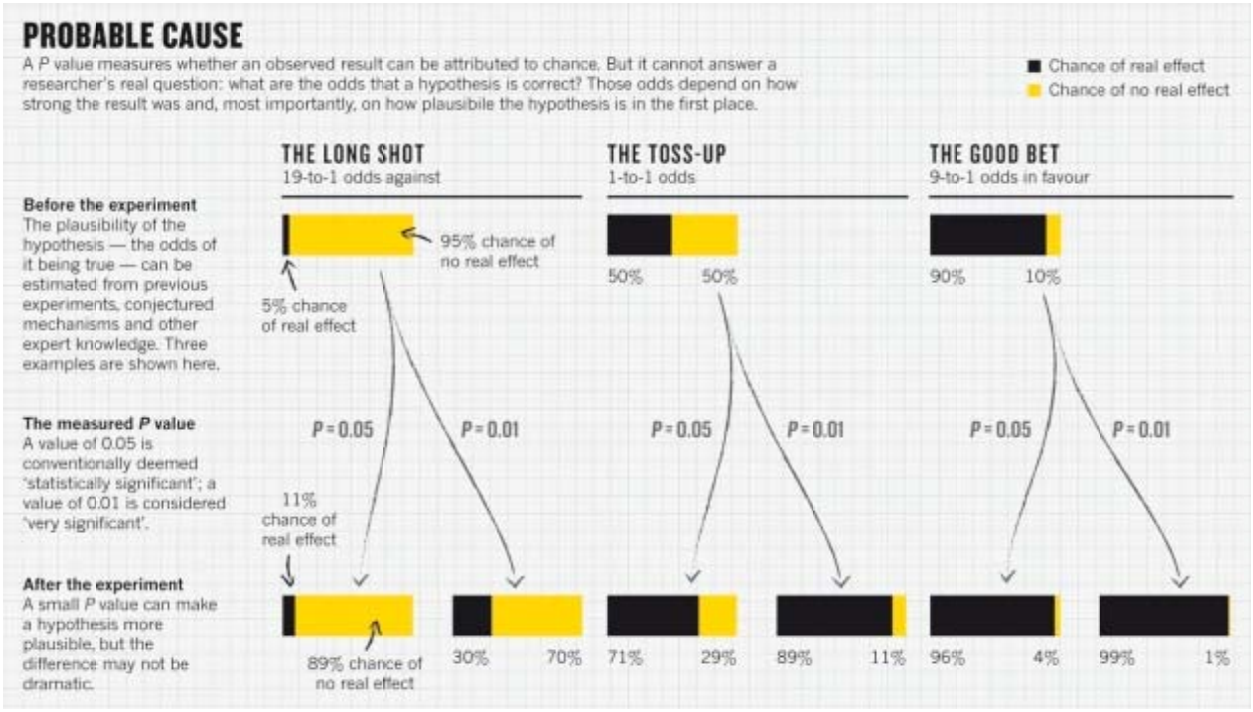


Fig. 7 *p*-value and false positive rate (Source: Sellke et al., 2001).

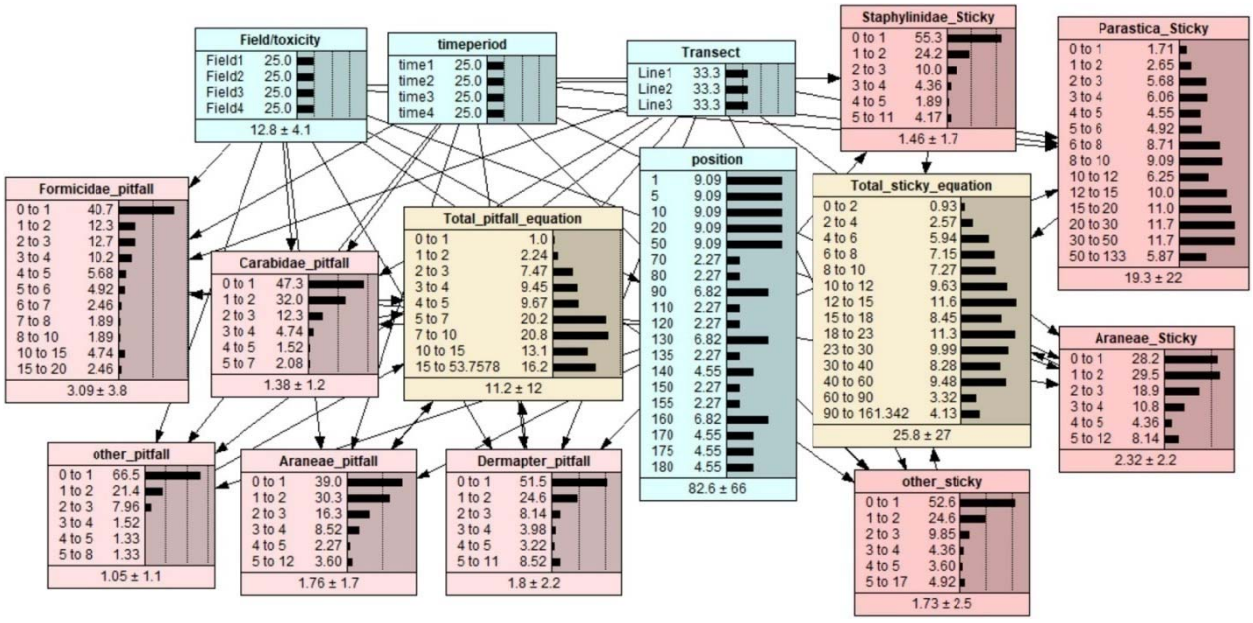


Fig. 8 A Bayesian Network (Sources: Pandey et al., 2022; Xie, 2022e).

Today, the distance between Bayesian and frequentist schools is getting smaller and smaller. Especially with the rise of objective Bayesian, Bayesian and frequentist schools may even move towards union (Efron, 2013), for example, Zhang (2021a, b, c) jointly used Bayesian method and frequency method to make casuality inference for nominal variables, Boolean variables and linearly correlated variables.

4.4 Using effect size (ES)

In order to solve the problem that the *p*-value is too sensitive and the dichotomy is used to determine the statistical significance, the effect size estimation paradigm can be used. Contrary to the significance tests paradigm, the effect size estimation paradigm guides scientists to pay attention to scientific significance, and directly conduct scientific research based on effect size inference (Huang, 2021a, b).

Effect size (ES) is an indicator used to measure the size of the effect. It can quantify the degree of association between variables, compare the changes before and after itself, compare the differences between groups, etc. Effect size is not just a certain indicator, different statistics tests, corresponding to different effect size indicators. At present, more than 100 kinds of effect size indicators have been applied in the field of statistics, mainly including several categories:

(1) Effect size used to measure correlation: Pearson correlation *r*, coefficient of determination r^2, ω^2, etc.;

(2) Effect size used to measure the difference: Cohen's *d*, Hedges' *g*, etc.;

(3) Effect sizes used to measure categories: Cohen's *w*, Odds ratio (OR), Relative risk (RR), etc.

Effect size has basic properties such as measurement unit independence, sample size independence, and monotonicity. In particular, unlike statistical significance *p*-value, effect size is not affected by sample size. Effect size can solve the problem that *p*-value cannot describe the degree of correlation and difference. The use of effect size can also avoid the problem of *p*-value hacking.

Differences in means, proportions explained by ANOVA, proportions explained by regression analysis, etc., can all be described by effect size. Effect size is very important for estimating treatment effects. If the effect size is too small, it means that the treatment has no practical value even if it reaches a significant level .

Several common effect sizes used for statistical tests are as follows:

(1) Comparison of difference between the two groups

$$\text{Cohen's } d = |m_1 - m_2| / s$$

where m_1 and m_2 are the mean of the two groups, respectively, and s is the combined standard deviation of the two groups:

$$s = (((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2))^{1/2}$$

where n_1 and n_2 are the sample sizes of the two groups, respectively, and s_1 and s_2 are the respective standard deviations of the two groups. Cohen's d = 0.20 is a small effect, Cohen's d = 0.5 is a medium effect, and Cohen's d = 0.80 is a high effect. The standard is just Cohen's personal experience, which should vary across disciplines, is flawed as the sole criterion for judgment.

When the sample size is small, such as when the overall sample is less than 20 or each group of samples is less than 10, Cohen's *d* will have a large deviation. In view of this, Hedges and Olkin (1985) proposed a method to calculate Cohen's *d* based on small samples, i.e.

$$\text{Hedges' } g = [|m_1 - m_2| / s] \times [1 - 3 / (4(n_1 + n_2 - 2) - 1)]$$

If the variables do not conform to the normal distribution, the data needs to be transformed to calculate the

effect size. The Box-Cox method can be used for data transformation. If the data transformation still cannot meet the requirements, rank transformation can be tried.

(2) Analysis of variance

I. One-way between-group effect size (cohen' *f*): $ES = (F / n)^{1/2}$

where, F and n are the F value and the number of groups (levels) in one-way ANOVA. ES = 0.10 is a small effect, ES = 0.25 is a medium effect, and ES = 0.40 is a high effect.

II. Partial η^2: $\eta_p^2 = SS_A / (SS_A + SS_E)$

In one-way ANOVA, SS_A is the between-group variation and SS_E is the within-group variation.

(3) Chi-Square Test for 2×2 Contingency Tables

$$\phi = (\chi^2 / n)^{1/2}$$

$\phi = 0.10$ is a small effect, $\phi = 0.30$ is a medium effect, and $\phi = 0.50$ is a high effect.

(4) Chi-Square Test for $R \times C$ Contingency Tables

$$\phi = (\chi^2 / min(C\text{-}1, R\text{-}1))^{1/2}$$

$\phi = 0.10$ is a small effect, $\phi = 0.30$ is a medium effect, and $\phi = 0.50$ is a high effect.

(5) Correlation coefficient

Pearson correlation (Pearson, 1895; Zhang and Li, 2015), point correlation (Zhang, 2017b, 2018), Spearman rank correlation (Spearman, 1904; Schoenly and Zhang, 1999; Zhang, 2015b, 2018).

(6) Interaction between two variables

I. Cohen's d: $d = 2t / df^{1/2}$

where, t is the t-value, df is the degree of freedom.

II. Cohen's η^2: $\eta^2 = Fdf_b / (Fdf_b + df_w)$

where, F is the F-value, df_b and df_w are the between-group and within-group degrees of freedom, respectively (Cohen, 1988; Li, 2021a). $\eta^2 = 0.1$ is a small effect, $\eta^2 = 0.3$ is a medium effect, and $\eta^2 = 0.5$ is a high effect.

In the general statistical significance tests, the sample size, mean M and standard deviation SD, t-value or F-value, df and p-value of each group are given. Based on this, the effect size statistics d and η^2 can be calculated. (Li, 2021a). Therefore, from the results of the statistical significance tests, the effect size can be calculated (Li, 2021a). Reporting both the p-value and the effect size can complement each other (Li, 2021a).

η^2 is one of the most commonly used effect size indicators. Li (2021a) used a literature example to illustrate the correspondence and difference between *p*-value and effect size (η^2) (Zarkadi and Schnall, 2013; Li, 2021a). The researcher will examine the influence of black and white background or gray background (priming condition) on moral judgment, and the experimental material is 6 social issues (pornography, adultery, drug use, littering, smoking, use of profanity), and asked to rate their morality on a scale of -5 (=very immoral) to +5 (=very moral). The researchers predicted that priming with black and white visual contrast would lead to more extreme morality than no priming judgment (Zarkadi and Schnall, 2013; Li, 2021a).

The researchers gave a deviation score index, that is, the distance between the results of the subjective judgment and the midpoint of the scale to evaluate the extreme situation of moral judgment. The results showed that the mean deviation score under the black and white condition (M = 2.50, SD = 0.96) was greater than the mean of the deviation scores in the grey condition (M = 2.05, SD = 0.91) (F(1,128) = 7.35, p = 0.008, η^2 = 0.05). When the researchers analyzed the 6 items separately, they showed the same general pattern of smoking (F(1,128) = 5.69, p = 0.02, η^2 = 0.04), and drug use (F(1,128) = 4.31, p = 0.04, η^2 = 0.03), adultery (F(1,128) = 8.34 , p = 0.005, η^2 = 0.06), the priming condition effect was significant. In addition, the difference in mean severity of the two priming conditions (black and white condition: M = -1.79, SD = 1.57; grey condition: M = -1.05, SD =1.32) (F(1,128) = 1.05, p = 0.31, η^2 = 0.008) is not significant (Li, 2021a). Comparing the above *p*-value and effect size η^2, it can be seen that *p*-value is much more sensitive than effect size η^2 (Li, 2021a).

Combining the experimental materials, it can be seen that the two indicators of *p*-value and effect size η^2 have complementary effects. For example, the contrasting background of black and white will polarize moral judgments, and this polarization is related to specific social issues. At the same time, the impact on adultery judgments (p = 0.005, η^2 = 0.06) versus smoking judgment (p = 0.02, η^2 = 0.04) is more obvious. It is especially important that the effect size of the black and white contrast background on the severity of moral judgments is very small (η^2 = 0.008), thus, suggesting that the impact of black and white contrast on moral judgments should not be examined with severity. Because even if the sample size is enlarged and the power is improved, and it does not make much sense (Li, 2021a). This study shows that the effect of black and white contrast on the polarization of moral judgment does not reach the standard of small effect size. This shows that the impact of black and white contrast on moral judgment is quite subtle, and although the effect is small, it does exist. Therefore, for other studies, if you want to repeat the study, you need to pay attention to improving the power (such as increasing the sample size), otherwise you may not get significant results. This suggests that reporting of both *p*-value and effect size can complement each other (Li, 2021a).

4.5 Pay attention to statistical validity

Capturing the uncertainty associated with statistical conclusions is critical. Different measures of uncertainty can complement each other and no single measure serves all purposes. Sources of variation in statistical conclusions should be included and reported in scientific articles. Where possible, those sources of variation that have not been described should also be identified (Benjamini et al., 2021).

Statistical validity refers to the degree of statistical results as the truth, which involves the appropriateness and accuracy of statistical analysis (Cook and Campell, 1979; Li, 2021b). Data analysis often has specific statistical assumptions. If these premises are violated, the statistical analysis may be inappropriate. Therefore statistical validity has more to do with the appropriateness of statistical analysis. At the same time, statistical validity also includes the accuracy of statistical analysis.

Factors affecting statistical validity include (Cook and Campbell, 1979; Li, 2021b):

(1) The reliability and validity of the measurement indicators

We need to perform statistical analysis on the data, and the data comes from specific measurement indicators. The fundamental condition for the effectiveness of statistical analysis is that these measurement indicators must have the required reliability and validity. If the analysis data is unreliable, the statistical validity of the corresponding research cannot be guaranteed.

(2) Type I error and statistical significance

Type I errors (false positives) occur when an effect is thought to exist but does not actually exist. If the null hypothesis is rejected at a certain level of significance (e.g. $p<0.001$), it is concluded that there is an effect, and the following issues need to be considered:

(i) Data sources. Is the data from an enumeration or a sample? If it is from an enumeration, it is not suitable for significance detection.

(ii) The nature of the sample. Is the data from a random sample or a non-random sample? If it comes from a non-random sample, the error of the result cannot be determined.

(iii) Nature of research. If the data are from a random sample, determine the appropriate significance level (p=0.01, p=0.001, etc).

(iv) Randomness nature. There are often two types of randomization in scientific research, one is random sampling and the other is random allocation. The principle of random sampling should be strictly followed. Random allocation is a type of random experiment. Random experiments can approximately control additional variables, but cannot control for sampling bias. For example, if the subjects are all male, then it may not be valid for females, even if the statistical analysis is significant.

(v) Properties of tests. To examine a large number of relationships, determine whether a significance test is based on ex-assumptions or data fishing? If it is a data fishing, with a significance level of p=0.05, one of the 20 relationships will be significant in probability. Post-hoc multiple comparisons need to adjust the significance level, usually a Bonferroni correction, i.e., the original significance level is divided by the number of comparisons. For example, the significance level is p=0.01, and 5 pairwise comparisons are made, then the significance is adjusted to 0.01/5 = 0.002.

(3) Type II error and statistical power

Type II error (false negative) occurs when a relationship is thought to be non-existent when it actually exists. If the researcher concludes that there is no relationship at a certain level of significance (e.g., $p<0.001$), then consider the following question:

Are the statistical procedures used by the researcher sufficiently powerful? If the power is greater than or equal to 0.80, then the conclusion that there is no such relationship is valid.

Is it just because the sample size is insufficient that the null hypothesis cannot be rejected or accepted?

(4) Interaction and nonlinearity

Whether interaction and nonlinear effects are considered. In theory, for a model: $y=f(x)$, If the function satisfies the principle of multiplication and additivity, that is: $f(\alpha x_1+\beta x_2)=\alpha f(x_1)+\beta f(x_2)$, where, $\alpha, \beta \in R$, Then the model is a linear model, otherwise it is a nonlinear model (Zhang, 2007).

(5) The correlation and causality are not clear

For correlation, it is necessary to clarify the type of correlation (Pearson correlation, Boolean correlation, nominal variable correlation, Spearman rank correlation, etc) (Pearson, 1895, 1904; Zhang, 2015a, b, 2016, 2017b, 2018). It is a direct correlation or indirect correlation (Zhang and Li, 2015; Zhang, 2016, 2018). For the causality relationship, the direction of causality must be confirmed, for example, causal direction inference based on Pearson correlation, Boolean correlation, and nominal variable correlation (Zhang, 2021a, b, c; Fig. 9).

(6) Ecological fallacy

It is assumed that the correlation at the individual level is the same as the correlation at the group level, or vice versa. Robinson (1950) have shown that the correlation at the individual level will appear higher, lower, or even in the opposite direction than the correlation at the group level.

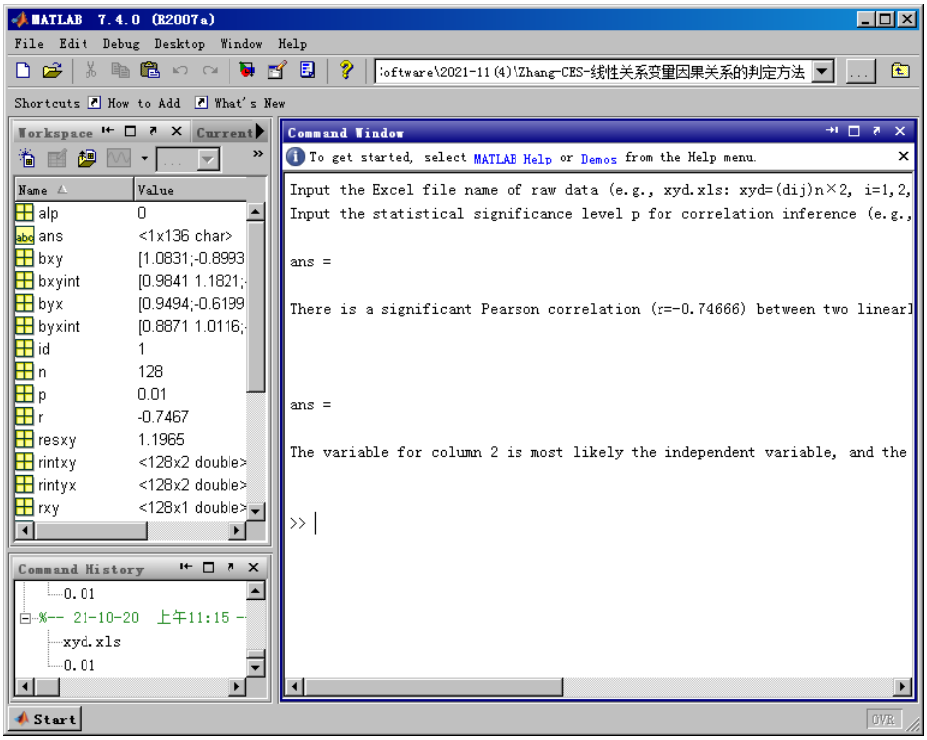


Fig. 9 A Matlab algorithm for making causal inference (Zhang, 2021b).

4.6 Using non-parametric statistics: Bootstrap methods, randomization tests

Bootstrap methods are a kind of non-parametric Monte Carlo methods, proposed by Efron of Stanford University on the basis of summarizing the previous research results (Manly, 2007; Zhang and Schoenly, 1999a,b; Figs. 10, 11). The method makes full use of the given measurement information, does not require other assumptions in the model and add new measurements, and is robust and efficient. First, Bootstrap can avoid the problem of sample reduction caused by cross-validation through resampling. Second, Bootstrap can also be used to create randomness in data (Figs. 10, 11). For example, in the well-known random forest algorithm, the first step is to use the Bootstrap method to randomly select k new bootstraped sample sets from the original training data set with replacement, and then construct k classification and regression trees. Through Bootstrap resampling and statistical calculation, it is helpful to discoverthe hidden internal mechanism, and makes the statistical inference more credible and more realistic.

Here I present several Bootstrap methods as follows:

(1) Bootstrap method for bootstrapping two paired samples in between-sample difference test

Suppose there are two paired samples of size s. The sample data are (x_i, y_i), i =1, 2,..., s. If $min\ (x_i, y_i)<0$, then let $x_i = x_i - min\ (x_i, y_i)$, $y_i = y_i - min\ (x_i, y_i)$, i=1,2,…,s. Suppose m is the maximum decimal places among all values in (x_i, y_i), let $(x_i, y_i) = (x_i, y_i)\ 10^m$, i=1,2,…,s. Through these transformations all of the values in sample data become integers which are equivalent to numbers of individuals. If no difference exists, then the distribution of individuals in the two paried samples will be a result of allocating the mixed sample values at

random into two paired samples of size equal to those of the original sample (Solow, 1993; Manly, 1997; Zhang, 2007, 2011b, 2018). Assume that the two paired samples to be tested are x and y, which contain $\sum_{k=1}^{s} x_k$ and $\sum_{k=1}^{s} y_k$ individuals respectively. The $\sum_{k=1}^{s} x_k+\sum_{k=1}^{s} y_k$ individuals of the combined sample are randomly reallocated into two randomized paired samples with $\sum_{k=1}^{s} x_k$ and $\sum_{k=1}^{s} y_k$ labeled individuals. The two randomized paired samples are thus two bootstrapped paired samples.

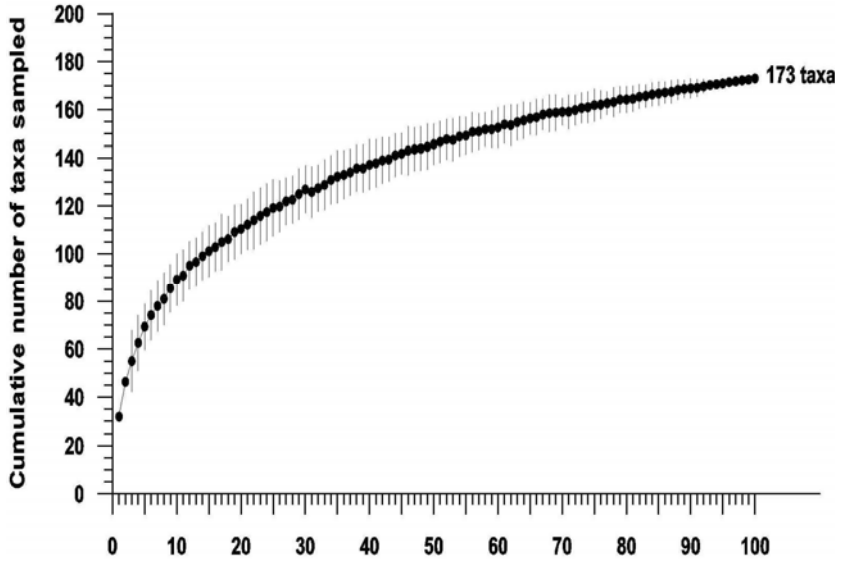


Fig. 10 Collector's curve for rice invertebrates gathered from 100 suction samples from the IRRI farm, 29 days after transplanting, dry season 1996. Each point is the mean of 100 randomizations of sample pooling order; vertical bars are means ±2 standard deviations (Zhang and Schoenly, 1999a). The Bootstrap method and randomization test were used in yielding collector's curve. Different from the single point (the last pont in the figure) from one sampling (100 samples), mechanism and trend of homogeneity and completeness of samples can be easily found from the collector's curve which was derived from Bootstrap re-samplings (Matlab algorithm: bootSamples.m).

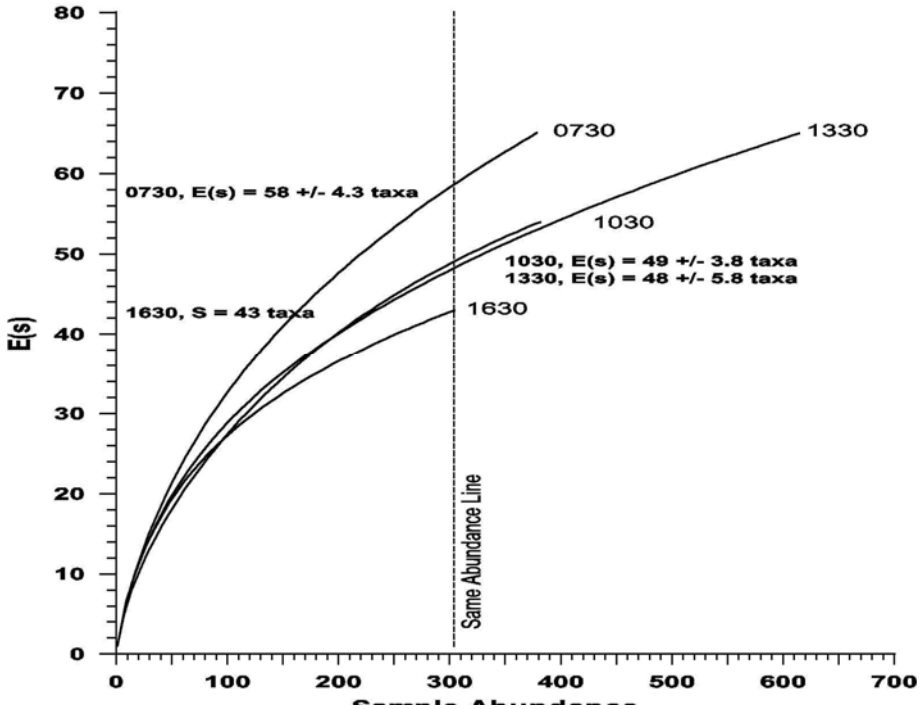


Fig. 11 Rarefaction curves for rice invertebrates suction-sampled at four times of the day (0730, 1030, 1330, 1630 h) during the vegetative stage, IRRI farm, dry season 1998 (Schoenly and Zhang, 1999) (Matlab algorithm: bootOneSamp.m).

Matlab function, bootTwoSamp.m, of the method is as the following:

```
function [xnew,ynew]=bootTwoSamp(x,y)
%x and y: two samples; xnew and ynew: two bootstraped samples.
m=max(size(x));
if (max(size(y))~=m)
error('Sample sizes do not match.');
end
dums=0;
dum=min(min(x),min(y));
if (dum<0)
x=x-dum;
y=y-dum;
dums=dum;
end
ma=-1e10;
for j=1:2
for i=1:m
ins=1;
if (j==1) dum=x(i);
else dum=y(i);
end
while (m~=0)
if ((abs(dum-floor(dum))<1) & (~(abs(dum-floor(dum))<=1e-10)))
ins=ins*10;
dum=dum*10;
if ((floor(dum+1e-10))~=(floor(dum))) break; end
else break; end
end
if (ins>ma)
ma=ins;
end
end
end
x=x.*ma.*1.0;
y=y.*ma.*1.0;
nrx=sum(x);
nrxy=sum(sum(x+y));
ar=floor(x+y);
col=sum(ar);
br(1)=ar(1);
for i=2:m
br(i)=br(i-1)+ar(i);
end
cols=randperm(nrxy);
```

```
xnew(1:m)=0;
for j=1:m
if (ar(j)==0) continue; end
if (j==1) temp=0;
else temp=br(j-1);
end
for i=1:nrx
if ((cols(i)>temp) & (cols(i)<=br(j)))
xnew(j)=xnew(j)+1;
end
end
end
ynew=ar-xnew;
xnew=xnew/ma;
ynew=ynew/ma;
if (dums<0)
xnew=xnew+dums;
ynew=ynew+dums;
end
```

The Matlab algorithm will generate two bootstrapped paired samples ($x_{new\,i}$, $y_{new\,i}$), i =1, 2,..., s, from two original paired samples (x_i, y_i), i =1, 2,..., s.

(2) Bootstrap method for bootstrapping a sample containing m categories of totally s individuals

Matlab function, bootOneSamp.m, of the method is as the following:

```
function [xnew,w]=bootOneSamp(x,n)
%x: a sample with m categories (sample size=m) of totally s individuals.
%xnew: the bootstraped sample given n individuals (n<=s).
%w: the number of categories in the bootstraped sample xnew.
m=max(size(x));
s=sum(x);
if (n>s)
error('Too many Bootstrapped individuals!');
end
ys=randperm(s);
y(1)=x(1);
for i=2:m
y(i)=y(i-1)+x(i);
end
for j=1:m
xnew(j)=0;
end
for i=1:n
for j=1:m
```

```
if (j==1) tem=0;
else tem=y(j-1);
end
if ((ys(i)>tem) & (ys(i)<=y(j)))
xnew(j)=xnew(j)+1;
end
end
end
w=sum(xnew~=0);
```

The Matlab algorithm will generate a bootstrapped sample ($x_{new\,i}$), i =1, 2,..., m, totally w categories and n individuals, from the original sample (x_i), i =1, 2,..., m, totally s individuals ($n \leq s$) (Schoenly and Zhang, 1999).

(3) Bootstrap method for bootstrapping s samples containing m categories given n required samples

Matlab function, bootSamples.m, of the method is as the following:

```
function xnew=bootSamples(x,n)
%x: a matrix (s columns, m rows) with s samples of m categories (sample size=m).
%xnew: the bootstraped sample given n samples (n<=s).
s=size(x,2);
m=size(x,1);
if (n>s)
error('Too many Bootstrapped samples!');
end
ys=randperm(s);
for i=1:n
xnew(:,i)=x(:,ys(i));
end
end
```

The Matlab algorithm will generate n bootstrapped samples, x_{new}, from the original s samples, x, where $n \leq s$ (Zhang and Schoenly, 1999a; Zhang, 2011a).

Randomization test is a kind of nonparametric test method (Solow, 1993; Schoenly and Zhang, 1999; Gentle, 2002; Manly, 2007; Zhang, 2011a, b c). The method is based on the given measurement data, through randomization simulation Measure the data any number of times to compare and calculate the relative difference between the measurement index and the simulation index. Randomization tests are mostly based on Bootstrap re-sampling data. Again, this method makes full use of the given measurement information, without model assumptions and adding new measurements , strong robustness and high efficiency (Zhang, 2011a, b c; Zhang et al., 2014). Although Bootstrap significance tests can be used, it should be noted that this interpretation of significance is different from random sampling from the population, in this case repeated sampling , that is, to extract samples from the existing data (Li, 2021b).

4.7 Using better experimental protocol and sampling protocol, and determining suitable sample size

(1) Experimental design

The so-called experiment is a practical activity used to test a hypothesis about the nature (Krebs, 1989). The design of an experiment refers to the logical structure of an experiment. Experiments can be divided into two categories:

(i) Observational experiments. This type of experiment does not require treatments, directly measures the experimental units, and should measure multiple samples.

(ii) Controlled experiments. In this type of experiment, the treatments should be set, and the treatment should be set up with multiple replicates.

In a strict sense, almost all experiments have varying degrees of uncertainty or randomness. Moreover, many experiments are controlled experiments. Controlled experiments should be repeated with replicates, and statistical analysis of the results is necessary. Number of replicates refers to how many experimental units are in each treatment. The purpose of setting up replicates is to avoid the effect of accidental events, to allow for estimated experimental error, so that statistical significance can be tested and confidence intervals can be calculated. Unfortunately, many controlled experiments, including most biological experiments have not replicates, or even have just one experimental unit only, and naturally there is no statistical analysis of the results. In these studies, only one experiment is carried out, neither under the same conditions, nor under different space-time conditions. The experimental results give general conclusions that may be wrong in nature, and cannot be replicated naturally (Ioannidis, 2005; Open Science Collaboration, 2015; Errington et al., 2021; Natural Editorial, 2021; Zhang, 2022). Such experiments can be called "gamblers' experiments", in other words, the results are true or false or right or wrong, all depends on luck.

The purpose of experimental design is to reduce errors and improve the accuracy of the results. There are four ways to improve the accuracy: (1) the experimental unit should be homogeneous; (2) the replicates should be enough; (3) try to use information provided by correlated variables, e.g., analysis of covariance; (4) use of a more efficient experimental design, e.g., a balanced design with the same number of replicates per treatment. The effect of an experiment depends on a good experimental design. Therefore, how to design an experiment that meets the statistical requirements so that the obtained results are statistically reasonable, analyzable and interpretable, is the primary problem to be solved in most experimental studies (Krebs, 1989; Zhang, 2007).

Most statistical tests assume that the measurements are independent of each other, which can be achieved through randomization, i.e., random sampling, or by randomly assigning treatments to experimental units. Through the randomization process, bias can be reduced and the accuracy of treatment estimates can be improved (Krebs et al., 1989; Zhang, 2007).

How to scatter treatments in space and time is an important issue. Different experimental design types have different methods of scattering (Krebs, 1989). Randomization and scattering are often contradictory, and by increasing the number of replicates, it is helpful to solve this problem. In experiments that are not well scattered, the replicates are not independent and thus do not meet the primary assumption of statistical inference. In this case, the repicates are called quasi-replicates. Underwood (1981) and Hurlbert (1984) point ourt that abour 48% and 78% of ecological papers, respectively, are such cases, which violates the principles of experimental design.

Common types of experimental designs, including completely randomized designs, randomized block designs, nested designs, split-plot designs, etc.

(2) Sampling design

How to determine the appropriate sampling method so that the sample is representative of the population studied is one of the important issues in observational experiments. For a detailed discussion of sampling design, see Ardilly (2005), Cochran (1977), Krebs (1989), Tille (2006), et al. Before conducting random sampling, three points need to be clarified:

(i) Defining the sampling population to be studied. The statistical population and the sampling population can be the same or different. In some cases, it is sometimes difficult to define because of the scale and boundaries of the space involved, or because the sampling population changes over time.

One way to define a sampling population is to strictly define the sampling population on a local scale and derive its statistical inferences (Krebs, 1989). The statistical population is often much larger than the sampling population, so the statistical inference must be extrapolated, A general conclusion is reached. Of course, the most reliable method is still to sample at the statistical population scale.

(ii) Determining the sampling unit. A sampling unit refers to the unseparable unit of sampling, which can be simple, such as an individual; or complex, such as a quadrat, etc. All sampling units must cover the entire sampling population and must not overlap.

(iii) Using a sampling protocol to draw samples. The purpose of making a sampling protocol is to provide the best statistical estimate with the smallest cost and the smallest confidence interval. To draw a sample, follow the principle of probability sampling, that is (Krebs, 1989; Zhang, 2007): (i) Define a group of significant samples S_i, i = 1, 2, …, and each sample contains some sampling units; (ii) assign a selection probability to each sample, and (iii) with the help of a random number table, select a sample from the sample set S_i, i = 1, 2, …, with given probability. According to the probability sampling principle above, a suitable sampling theory can be found to explain the data taken for analysis.

Common sampling protocols, including simple random sampling, stratified random sampling, multi-stage sampling, etc.

In addition, we should also test the homogeneity of samples obtained (Coleman et al., 1982; Zhang and Schoenly, 1999a; Zhang, 2011a).

In particular, it should be pointed out that it is easy to find examples with similar values but actually from different sampling populations, and only relying on statistical data analysis itself cannot distinguish the real sampling populations. For observational experiments, when there is only a single sample data, statistical inference is completely unreliable (Crawly, 2012; Xie, 2022d).

In scientific discovery research, the most important thing is to make multiple experiments or draw multiple samples under different conditions in terms of the same question/hypothesis. Only by synthesizing the results of multiple samples can the scientific truth be revealed, which is also the essence of statistical inference analysis based on sampling distribution. In the case of single sample data, statistical analysis should only do descriptive statistical analysis. The conclusion that a scientific discovery can be made at the level of statistical significance with one sample or no-replicate experiment has been the biggest mistake made by statistics education in the past few decades (Xie, 2022d). Unfortunately, in many studies, including biological research, only single sample or only one sampling unit is always used; using results from single sample or one sampling unit alone to draw generalized conclusions may be substantially wrong, and the results naturally cannot be reproduced (Ioannidis, 2005; Open Science Collaboration, 2015; Errington et al., 2021; Natural Editorial , 2021; Zhang, 2022).

(3) Determination of sample size

In sampling, after the sampling protocol used is determined, for a sample, it is necessary to determine the sample size to be used (Zhang, 2011a). A suitable sample size is that has minimum sampling cost and maximum data information. For sample size issues, see Ardilly (2005), Cochran (1977), Krebs (1989), Mace (1964), Tille (2006), etc.

In addition, we should also examine sampling completeness (Coleman et al., 1982; Zhang and Schoenly, 1999a; Zhang, 2011a).

Many sample size formulations assume that the random variable follows a normal distribution. When the

random variable is not normally distributed, according to the central limit theorem of statistics, when the sample size increases, it tends to be normally distributed, and these formulations are still available. Regarding sample size (n), we can use the following criterion (Cochran, 1977)

$$n>25\ [\sum(x_i-x')^3/(ns^3)]^2$$

If this formula is met, the sample size is considered to be large enough. Or, the sample size is determined according to the extensive experience of their respective majors.

4.8 Whole-process control of statistical analysis

In all research, a research consciousness needs to be formed, i.e., the results cannot be considered reliable and valid after passing statistical analysis only. In addition to considering the quality of the data itself (experimental design, sampling design, etc.) in statistical analysis, it is also necessary to consider (Cook and Campell, 1979; Li, 2021b):

(1) Whether the choice of statistical analysis method is correct. For example, choose multiple regression or stepwise regression, choose factor analysis or principal component analysis; when multivariate analysis of variance is required, do multiple one-way analysis of variance; Posterior comparison should followed by prior comparison; if a one-tailed test is required, a two-tailed test should be required.

(2) Whether the preconditions of statistical analysis methods are met, such as normality test, independence test, variance homogeneity test, multivariate collinearity test, regression homogeneity test, etc. To give a few examples:

(i) In most parametric statistical methods, the variables are required to obey the normal distribution, so it is necessary to test the normal distribution of the variables.

(ii) In the analysis of variance, if each treatment (level) is required to obey the normal distribution and the overall variance is the same, it is necessary to test the homogeneity of variances for the variables.

(iii) In factorial analysis and principal component analysis, if there is sufficient correlation between variables, KMO test and Bartlett test of sphericity need to be performed on the correlation of variables.

(iv) In linear model (regression) analysis, it is required that the variables obey a normal distribution, the variances of the variables are the same, the variables are independent of each other, there is no collinearity between the variables, and the residuals are independent, etc. It is necessary to test the normality, independence test, variance homogeneity test, collinearity diagnostics, residual independence Durbin-Watson test, singular value diagnosis, etc.

(3) Reasonable interpretation of statistical results. The interpretation of statistical results should be cautious, and the interpretation of results should match the characteristics of data representativeness and statistical methods.

4.9 Replacing reductionist methods of data acquisition and analysis with network methods

In some important scientific fields, the data representation is poor, because the population or system is large and complex, nonlinear, or changes in space and time, while the sampling population is defined based on reductionism (Zhang, 2017, 2019c; Wu, 2022), which is presented as partial, fragmented , linear, static, and lack of representativeness. Therefore, in addition to the aforementioned methods, in these important scientific fields, the network methods (ISNB, 2011-2012; Zhang, 2017a, 2018, 2019a, b) can be used to design experiments and analyze data, thus more comprehensive and systematic data can be obtained, and more accurate conclusions can be obtained.

4.10 Scientific inference = statistical conclusion + mechanism analysis

In any scientific research, the mechanism analysis of the research object is extremely important, at least it

ranks as important as data measurement and phenomenon measurement. Only mechanism analysis and data analysis (phenomenon observation is also a kind of data) should verify each other, and the research conclusions can thus be regarded as more rigorous scientific inferences. Pure statistical analysis from data to data can usually only be regarded as speculation, and should not be treated as deterministic inferences. Therefore, statistical significance and methods such as confidence intervals need to be used with caution. This 'confidence' itself is questionable and not completely reliable (Grenville, 2019).

5 Discussion

Hurlbert et al. pointed out that many controversies in statistics are mainly or entirely caused by poor journal quality control, poor quality statistics textbooks, poor teaching quality, unclear writing, and lack of understanding of historical literature (Hurlbert et al., 2019). They recommend that the term 'statistical significance' and all its cognates and symbolic adjectives should not be used in the scientific literature. Statistical reforms to address issues of statistical significance involve the development or revision of journal policies, the writing or revision of statistics textbooks, as well as writing or revising statistical analysis software (SPSS, SAS, Matlab, etc.) (Huang, 2021a, b), are already underway, albeit not in full swing and slowly. For example, as early as 2013, Cumming published a textbook on new statistics (Cumming, 2013).

According to Kuhn's paradigm shift theory, a new paradigm accepted by the academic community must appear to replace the old paradigm in order to achieve a paradigm shift. Will the big debate over statistical significance tests and $p<0.05$ lead to a 'paradigm shift'? it is regrettable that so far there has not been a universally accepted new paradigm that can replace the statistical significance tests paradigm. The academic community is still far from reaching a consensus on the paradigm shift. Reformers suggest effect size as a new paradigm to replace the significance tests paradigm. Reservationists strongly oppose to abandoning the significance tests paradigm (Benjamini et al., 2021; Kafdar, 2021). There is a heated debate between reformers and reservationists (Benjamini et al. al., 2021; Gelman, 2021; Kafdar, 2021; Higgs, 2021; Huang, 2021a, b). Therefore, for a long time in the future, the statistical significance tests paradigm will still exist as the mainstream. The *p*-value debate will continue (Grenville, 2019).

I believes that we need to change the paradigm of scientific research, abandon the paradigm of "one trial → publishing", and adopt the paradigm of "multiple repeated trials/multiple sample testing + multi-party validation → publishing", in order to greatly improve the authenticity and reproducibility of results. Before popularization, what we can do is to improve the quality of data in various aspects according to the aforementioned requirements, strictly the *p*-value level, adopt more reasonable analysis methods or test standards, and cross-validate multiple evidences such as statistical analysis combined with mechanism analysis, etc. etc. In addition to writing, publishing and adopting new statistical works and teaching textbooks, it is imperative to revise and distribute various statistical software in new versions based on new statistics for use by scientists. In addition, professional and applied statisticians should be encouraged to popularize new statistics (Shao, 2018; Huang, 2021a, b, 2022; Li, 2021a,b, 2022; Xie, 2022a-e).

References

Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. Nature, 567: 305-307. https://doi.org/10.1038/d41586-019-00857-9

Ardilly P. 2005. Sampling Methods: Exercises and Solutions. Springer, Netherlands. https://www.nhbs.com/sampling-methods-book

Baker M. 2016a. 1,500 scientists lift the lid on reproducibility. Nature, 533: 452-454. https://www.nature.com/articles/533452a

Baker M. 2016b. Statisticians issue warning over misuse of *P* values. Nature, 531: 151. https://doi.org/10.1038/nature.2016.19503

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, et al. 2018. Redefine statistical significance. Nature Human Behaviour, 2: 6-10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini Y, De Veaux RD, Efron B, Evans S, Glickman M, et al. 2021. The ASA President's Task Force Statement on Statistical Significence and Replicability. The Annals of Applied Statistics, https://doi.org/10.1214/21-AOAS1501

Bergstrom CT, West JD. 2021. Manipulated P-values: Mathematical nosense in scientific papers. https://www.laitimes.com/en/article/3km6i_41b77.html. Accessed 2022-4-23

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2: 637-644. https://doi.org/10.1038/s41562-018-0399-z

Camerer CF, Dreber A, Forsell E, Ho TH, et al. 2016. Evaluating replicability of laboratory experiments in economics. Science, 351(6280): 1433-1436. https://doi.org/10.1126/science.aaf0918

Cochran WG. 1977. Sampling Techniques (3rd ed). Wiley, New York, USA. https://www.wiley.com/en-us/Sampling+Techniques%2C+3rd+Edition-p-9780471162407

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Lawrence Earlbaum Associates, Hillsdale, New Jersey, USA. https://doi.org/10.4324/9780203771587

Coleman BD, Mares MA, Willig MR, et al. 1982. Randomness, area, and species richness. Ecology, 63: 1121-1133. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1937249

Cook TD, Campell DT. 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago, USA. https://www.scholars.northwestern.edu/en/publications/quasi-experimentation-design-and-analysis-issues-for-field-settin

Crawley MJ. 2012. The R Book (2nd ed). Wiley, USA. https://www.wiley.com/en-us/The+R+Book%2C+2nd+Edition-p-9780470973929

Cumming G. 2011. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge, New York, USA. https://www.routledge.com/Understanding-The-New-Statistics-Effect-Sizes-Confidence-Intervals-and/Cumming/p/book/9780415879682

Cumming G. 2013. The new statistics: Why and how. Psychological Science, 25: 7-29. https://doi.org/10.1177/0956797613504966

Efron B. 1979. Bootstrap methods: Another look at the Jackknife. Annals of Statistics, 7(1): 1-26. https://doi.org/10.1214/aos/1176344552

Efron B. 2013. A 250-year argument: belief, behavior, and bootstrap. Bulletin of the American Mathematical Society, 50: 129-146. https://doi.org/10.1090/S0273-0979-2012-01374-5

Errington TM, Mathur M, Soderberg CK, et al. 2021. Investigating the replicability of preclinical cancer biology. eLife, 10: e71601. https://elifesciences.org/articles/71601

EVEE. 2022. World's Largest Academic Integrity Survey: More than half of PhDs admit to engaging in questionable research, 1 in 12 falsified/fabricated data. https://mp.weixin.qq.com/s/-itNmspuoYXuRBFUYnyg_Q. Accessed 2022-5-6

Fanelli D. 2012. Negative results are disappearing from most displines and countries. Scientometrics, 90:

891-904. https://doi.org/10.1007/s11192-011-0494-7

Fidler F. 2020. Fraud, bias, negligence and hype in the lab — a rogues' gallery. Nature, 583: 515-516. https://doi.org/10.1038/d41586-020-02147-1

Fidler F, Cumming G. 2007. Lessons learned from statistical reform efforts in other disciplines. Psychology in the Schools, 44: 441-449. https://doi.org/10.1002/pits.20236

Fisher RA. 1935. The Design of Experiments. Oliver and Boyd, Edinburg and London, UK. http://www.medicine.mcgill.ca/epidemiology/hanley/tmp/Mean-Quantile/DesignofExperimentsCh-III.pdf

Gelman A. 2021. Thoughts on "The American Statistical Association President's Task Force Statement on Statistical Significance and Replicability". Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2021/07/12/thoughts-on-the-american-statistical-association-presidents-task-force-statement-on-statistical-significance-and-replicability/. Assessed on 2021-9-13

Gentle JE. 2002. Elements of Computational Statistics. Springer, Germany. https://link.springer.com/book/10.1007/b97337

Gigerenzer G. 2018. Statistical rituals: The replication delusion and how we got there. Advances in Methods and Practices in Psychological Science, 1(2): 198-218. https://doi.org/10.1177/2515245918771329

Goodman SN. 2001. Of *p*-values and Bayes: A modest proposal. Epidemiology, 12: 295-297. https://journals.lww.com/epidem/fulltext/2001/05000/of_p_values_and_bayes__a_modest_proposal.6.aspx

Grenville A. 2019. The danger of relying on "statistical significance". https://www.marugroup.net/insights/blog/danger-of-relying-on-statistical-significance. Accessed 2019-6-15

Hahn GJ, Meeker WQ. 1993. Assumptions for statistical inference. The American Statistician, 47(1): 1-11. https://doi.org/10.2307/2684774

Haller H, Kraus S. 2002. Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research, 7(1): 1-20. https://psycnet.apa.org/record/2002-14044-001

Halsey LG. 2019. The reign of the *p*-value is over: what alternative analyses: Could we employ to fill the power vacuum? Biology Letters. https://doi.org/10.1098/rsbl.2019.0174

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-Analysis. San Diego, Academic Press CA, USA. https://idostatistics.com/hedges-olkin-1985-statistical-methods-for-meta-analysis/

Higgs MD. 2021. Thoughts on the task force statement. Critical Inference. https://critical-inference.com/thoughts-on-the-task-force-statement/. Accessed 2021-9-13

Hubbard R, Haig BD, Parsa RA. 2019. 2019. The limited role of formal statistical inference in scientific inference. The American Statistician, 73(S1): 91-98. https://doi.org/10.1080/00031305.2018.1464947

Huang HN. 2021a. Statistics reform: challenges and opportunities. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1318043.html. Accessed 2021-12-26

Huang HN. 2021b. What are the most misunderstood and misleading concepts or theories in statistics textbooks today? ScienceNet. https://blog.sciencenet.cn/blog-3427112-1269013.html. Accessed 2021-1-26

Huang HN. 2022. Bayesian vs frequentist: a debate spanning two and a half centuries. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1331814.html

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, 54: 187-211. https://doi.org/10.2307/1942661

Hurlbert SH, Levine RA, Utts J. 2019. Coup de grace for a tough old bull: "statistically significant" expires. The American Statistician, 73(supp1): 352-357. https://doi.org/10.1080/00031305.2018.1543616

Ioannidis JPA. 2005. Why most published research findings are false. Plos Medicine,

https://doi.org/10.1371/journal.pmed.0020124

ISNB. 2011-2022. Network Biology (ISSN 2220-8879). http://www.iaees.org/publications/journals/nb/nb.asp

Kafdar K. 2021. Editorial: Statistical significance, *p*-values, and replicability. The Annals of Applied Statistics. https://doi.org/10.1214/21-AOAS1500

Kaiser J. 2015. Study claims $28 billion a year spent on irreproducible biomedical research: Economists extrapolate cost of preclinical studies that don't hold up. Science. https://doi.org/10.1126/science.aac6811

Kjærulff UB, Madsen AL. 2008. Bayesian Networks and Influence Diagrams. Springer, New York, USA. https://link.springer.com/book/10.1007/978-0-387-74101-7

Krebs CJ. 1989. Ecological Methodology. HarperCollinsPublishers, New York, USA. https://www.worldcat.org/title/ecological-methodology/oclc/709877724?referer=di&ht=edition

Li HH. 2021a. *p*-values are too sensitive, and the effect size is long and good to save. ScienceNet. http://blog.sciencenet.cn/blog-2619783-1286084.html. Accessed 2021-5-11

Li HH. 2021b. Statistical validity: a class of validity that is easily overlooked. http://blog.sciencenet.cn/blog-2619783-1292778.html. Accessed 2021-6-26

Li HH. 2022. Significance test, let's talk about it. ScienceNet. https://blog.sciencenet.cn/blog-2619783-1324953.html. Accessed 2022-2-11

Mace AE. 1964. Sample-Size Determination. Reinhold, New York, USA. https://www.biblio.com/book/sample-size-determination-mace-arthur-e/d/1424373395

Mackey DJC. 1992. Bayesian interpolation. Neural Computation, 4: 415-447. https://doi.org/10.1162/neco.1992.4.3.415

Manly BFJ. 2007. Randomization, Bootstrap and Monte Carlo Methods in Biology (3nd ed). Chapman and Hall/CRC, New York, USA. https://www.taylorfrancis.com/books/mono/10.1201/9781315273075/randomization-bootstrap-monte-carlo-methods-biology-bryan-manly

McShane BB, David Gal D. 2017. Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association, 112(519): 885-895. https://doi.org/10.1080/01621459.2017.1289846

Mehta D. 2019. highlight negative results to improve science. Nature. https://doi.org/10.1038/d41586-019-02960-3

Nature Editorial. 2021. Replicating scientific results is tough — but essential. Nature, 600: 359-360. https://doi.org/10.1038/d41586-021-03736-4

Neyman J, Pearson ES. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A, 231(694-706): 289-337. https://doi.org/10.1098/rsta.1933.0009

Nuzzo R. 2014. Scientific method: Statistical errors. Nature, 506: 150-152. https://doi.org/10.1038/506150a

Oakes M. 1986. Statistical Inference: A Commentary for the Social and Behavioural Sciences. Wiley, USA. https://onesearch.nihlibrary.ors.nih.gov/discovery/fulldisplay/alma991000107349704686/01NIH_INST:NIH

O'Grady C. 2017. “Mindless Eating” or how to send an entire life of research into question. ARS Technica. 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. Accessed 2017-6-20

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science, 349(6251). https://doi.org/10.1126/science.aac4716

Upton G, Cook I. 2008. Oxford Dictionary of Statistics (2nd ed). Oxford University Press, UK. https://doi.org/10.1017/S0025557200184025

Pandey S, Johnson AC, Xie G, Gurr GM. 2022. Pesticide regime can negate the positive influence of native vegation donor habitat on natural enemy abundance in adjacent crop fields. Frontiers in Ecology and Evolution. https://doi.org/10.3389/fevo.2022.815162

Pearl J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Elsevier, USA. https://doi.org/10.1016/C2009-0-27609-4

Pearson K. 1895. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58: 240-242. https://doi.org/10.1098/rspl.1895.0041

Pearson K. 1904. On the Theory of Contingency and its Relation to Association and Normal Correlation. Dulau and Co, London, UK. https://www.worldcat.org/title/on-the-theory-of-contingency-and-its-relation-to-association-and-normal-correlation/oclc/63332451

Prinz F, Schlange T, Khusru Asadullah K. 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery, 10: 712. https://doi.org/10.1038/nrd3439-c1

Ritchie S. 2020. Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth. Metropolitan Books, New York, USA. https://www.amazon.com/Science-Fictions-Negligence-Undermine-Search/dp/1250222699

Robinson WS. 1950. Ecological correlations and the behavior of individuals. American Sociological Review, 15: 351-357. https://doi.org/10.2307/2087176

Rosner B. 2006. Fundamentals of Biostatistics (6th ed). Thomson Brooks/Cole, California, USA. https://www.worldcat.org/title/fundamentals-of-biostatistics/oclc/58918467

Salsburg D. 2002. The Lady Tasting Tea: How Statistics Revolutionized Science in The Twentieth Century. Holt Paperbacks, USA. https://doi.org/10.1038/90908

Schoenly KG, Zhang WJ. 1999. IRRI Biodiversity Software Series. V. RARE, SPPDISS, and SPPRANK: Programs for Detecting Between-Sample Differences in Community Structure. IRRI Technical Bulletin No.5. International Rice Research Institute, Manila, Philippines. http://books.irri.org/TechnicalBulletin5_content.pdf

Sellke T, Bayarri MJ, Berger JO. 2001. Calibration of p values for testing precise null hypotheses. The American Statistician, 55(1): 62-71. https://doi.org/10.1198/000313001300339950

Servick K. 2017. It will be much harder to call new findings ‘significant' if this team gets its way: Proposal to change widely ace pted p-value threshold stirs reproducibility debate. Science. https://doi.org/10.1126/science.aan7154

Shao B. 2018. What results are significant: A brief discussion of *p*-values. ScienceNet. http://blog.sciencenet.cn/blog-927304-1093043.html. Accessed 2019-6-20

Snedecor GW, Cochran WG. 1991. Statistical Methods (8th ed). Wiley-Blackwell, USA. https://www.wiley.com/en-us/Statistical+Methods%2C+8th+Edition-p-9780813815619

Solow AR. 1993. A simple test for change in community structure. Journal of Animal Ecology, 62: 191-193. https://doi.org/10.2307/5493

Spearman C. 1904. The proof and measurement of association between two things. American Journal of Psychology, 15: 72-101. https://doi.org/10.2307/1422689

Sun XJ. 2016. Improper use of *p*-values is a nonsense. ScienceNet. http://blog.sciencenet.cn/blog-41174-961169.html. Accessed 2016-4-5

The Skeptical Scientist. 2021. The Wansink Dossier: An Overview. https://www.timvanderzee.com/the-wansink-dossier-an-overview/. Accessed 2022-3-24

Tille Y. 2006. Sampling Algorithms. Springer, Netherlands. https://link.springer.com/book/10.1007/0-387-34240-0

Timmer J. 2009. Anti-"publication bias" efforts not panning out for science: A survey of clinical trials placed in public registries indicate that by the …. ARS Technica. https://arstechnica.com/science/2009/09/for-clinical-trials-design-and-results-dont-always-match/. Accessed 2021-10-9

Tong C. 2019. Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. The American Statistician, 73(S1): 246-261. https://doi.org/10.1080/00031305.2018.1518264

Trafimow D, Marks M. 2015. Editorial. Basic and Applied Social Psychology, 37: 1-2. https://doi.org/10.1080/01973533.2015.1012991

Underwood AJ. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and Marine Biology: An Annual Review, 19: 513-605. https://www.worldcat.org/title/techniques-of-analysis-of-variance-in-experimental-marine-biology-and-ecology/oclc/704173640

Vrieze JD. 2021. Landmark research integrity survey finds questionable practices are surprisingly common. Science. https://doi.org/10.1126/science.abk3508

Wansink B. 2010. Mindless Eating: Why We Eat More Than We Think. Bantam Books, NJ, USA. https://doi.org/10.1080/22243534.2021.2006941

Wansink B. 2014. Slim by Design: Mindless Eating Solutions for Everyday Life (Illustrated ed). William Morrow, NJ, USA. https://doi.org/10.1111/obr.12249

Wasserstein RL, Lazar NA. 2016. Editorial: The ASA statement on *p*-values: Context, process, and purpose. the American Statistician, 70(2): 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein RL, Schirm AL, Lazar NA, 2019. Editorial: Moving to a world beyond "$p<0.05$". The American Statistician, 79: 1-19. https://doi.org/10.1080/00031305.2019.1583913

Wu JR. 2022. The dilemma of involution in life sciences and its solution. Chinese Bulletin of Life Sciences, 34(4): 339-344. https://mp.weixin.qq.com/s?__biz=MzA4MTQyNDEyMQ==&mid=2651003900&idx=1&sn=1c81fc2fed009cdcc19e3b5d184b2d4a&scene=21#wechat_redirect. Accessed 2022-4-26

XKCD. 2021. 882: Significant. https://www.explainxkcd.com/wiki/index.php/882:_Significant. Accessed 2021-5-10

Xie G. 2022a. An English-to-Chinese article on "statistical significance" worth reading. ScienceNet. https://blog.sciencenet.cn/blog-3503579-1322100.html. Accessed 2022-1-24

Xie G. 2022b. History and recent developments of the statistical significance problem. https://www.researchgate.net/publication/359365092_tongjixianzhexingwentidelishiyoulaijizuixinjinzhan. Accessed 2022-4-15

Xie G. 2022c. "Statistically significant" - please stop saying and using it when doing statistical data analysis! ScienceNet. https://blog.sciencenet.cn/blog-3503579-1324675.html. Accessed 2022-2-10

Xie G. 2022d. What can *p*-values do without "statistical significance"? ScienceNet. https://blog.sciencenet.cn/blog-3503579-1325287.html. Accessed 2022-2-14

Xie G. 2022e. Say goodbye to 'statistical significance': A journal article and the story behind it. https://blog.sciencenet.cn/blog-3503579-1333915.html. Accessed 2022-4-16

Yates F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46: 19-34. https://doi.org/10.2307/2280090

Zarkadi T, Schnall S. 2013. "Black and white" thinking: Visual contrast polarizes moral judgment. Journal of

Experimental Social Psychology, 49: 355-359. https://doi.org/10.1016/j.jesp.2012.11.012

Zhang WJ. 2007. Methodology on Ecology Research. Sun Yat-sen University Press, Guangzhou, China. https://books.google.com/books/about/%E7%94%9F%E6%80%81%E5%AD%A6%E7%A0%94%E7%A9%B6%E6%96%B9%E6%B3%95.html?id=btTzPQAACAAJ

Zhang WJ. 2011a. A Java program to test homogeneity of samples and examine sampling completeness. Network Biology, 1(2): 127-129. http://www.iaees.org/publications/journals/nb/articles/2011-1(2)/Java-program-to-test-homogeneity-of-samples.pdf

Zhang WJ. 2011b. A Java algorithm for non-parametric statistic comparison of network structure. Network Biology, 1(2): 130-133. http://www.iaees.org/publications/journals/nb/articles/2011-1(2)/Java-algorithm-non-parametric-statistic-comparison-network-structure.pdf

Zhang WJ. 2011c. A Java program for non-parametric statistic comparison of community structure. Computational Ecology and Software, 1(3): 183-185. http://www.iaees.org/publications/journals/ces/articles/2011-1(3)/Java-program-non-parametric-statistic-comparison-community-structure.pdf

Zhang WJ. 2015a. A hierarchical method for finding interactions: Jointly using linear correlation and rank correlation analysis. Network Biology, 5(4): 137-145. http://www.iaees.org/publications/journals/nb/articles/2015-5(4)/partial-correlation-analysis-in-finding-interactions.pdf

Zhang WJ. 2015b. Calculation and statistic test of partial correlation of general correlation measures. Selforganizology, 2(4): 65-77. http://www.iaees.org/publications/journals/selforganizology/articles/2015-2(4)/statistic-test-of-partial-correlation-of-general-correlation-measures.pdf

Zhang WJ. 2016. Selforganizology: The Science of Self-Organization. World Scientific, Singapore. https://doi.org/10.1142/9685

Zhang WJ. 2017a. Advances in Network Pharmacology of Chinese Herbal Medicines. IAEES. http://www.iaees.org/publications/journals/np/articles/cn/index.asp. Accessed 2022-4-12

Zhang WJ. 2017b. Network pharmacology of medicinal attributes and functions of Chinese herbal medicines: (II) Relational networks and pharmacological mechanisms of medicinal attributes and functions of Chinese herbal medicines. Network Pharmacology, 2(2): 38-66. http://www.iaees.org/publications/journals/np/articles/2017-2(2)/networks-and-mechanisms-of-medicinal-attributes-and-functions.pdf

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK. https://doi.org/10.1142/q0149

Zhang WJ. 2019a. The English book of Professor Zhang, Sun Yat-Sen University was published in London, UK. IAEES. http://www.iaees.org/publications/journals/nb/cn/fnb/index.asp. Accessed 2022-2-21

Zhang WJ. 2019b. Tutorial of Network Biology. ResearchGate. https://doi.org/10.13140/RG.2.2.17798.22082

Zhang WJ. 2019c. Medicine and health: the reasons for the decline in the success rate of global new drug development. Health and Environment Communications. https://mp.weixin.qq.com/s?__biz=MzUwOTA4NzYwOA==&mid=2247484043&idx=1&sn=8313d5d4d785415aa04c6d1dbccc34be&chksm=f916d89cce61518afadfab588fa42b5ba3bfdb79d0aac82d33dd415e7b802af8a73d7608b966&token=1681770050&lang=zh_CN&scene=21#wechat_redirect. Accessed

2022-3-15

Zhang WJ. 2021a. A statistical simulation method for causality inference of Boolean variables. Network Biology, 11(4): 263-273. http://www.iaees.org/publications/journals/nb/articles/2021-11(4)/a-method-for-causality-inference-of-Boolean-variables.pdf

Zhang WJ. 2021b. Causality inference of linearly correlated variables: The statistical simulation and regression method. Computational Ecology and Software, 11(4): 154-161. http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-linearly-correlated-variables.pdf

Zhang WJ. 2021c. Causality inference of nominal variables: A statistical simulation method. Computational Ecology and Software, 11(4): 142-153. http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-nominal-variables-with-statistical-simulation-method.pdf

Zhang WJ. 2022. Publish early, assess late: The reproducibility crisis in experimental science. Health and Environment Communications. https://translate.google.com/?hl=zh-CN&sl=zh-CN&tl=en&text=%E6%97%A9%E5%8F%91%E8%A1%A8%EF%BC%8C%E6%99%9A%E8%AF%84%E4%BB%B7%EF%BC%9A%E5%AE%9E%E9%AA%8C%E7%A7%91%E5%AD%A6%E7%9A%84%E5%8F%AF%E9%87%8D%E5%A4%8D%E6%80%A7%E5%8D%B1%E6%9C%BA%E9%97%AE%E9%A2%98&op=translate. Accessed 2022-3-6

Zhang WJ, Li X. 2015. Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed. Selforganizology, 2(3): 39-45. http://www.iaees.org/publications/journals/selforganizology/articles/2015-2(3)/linear-correlation-analysis-in-finding-interactions.pdf

Zhang WJ, Qi YH, Zhang ZG. 2014. Two-dimensional ordered cluster analysis of component groups in self-organization. Selforganizology, 1(2): 62-77. http://www.iaees.org/publications/journals/selforganizology/articles/2014-1(2)/two-dimensional-ordered-cluster-analysis.pdf

Zhang WJ, Schoenly KG. 1999a. IRRI Biodiversity Software Series. II. COLLECT1 and COLLECT2: Programs for Calculating Statistics of Collectors' Curves. IRRI Technical Bulletin No.2. International Rice Research Institute, Manila, Philippines. http://books.irri.org/TechnicalBulletin2_content.pdf

Zhang WJ, Schoenly KG. 1999b. IRRI Biodiversity Software Series. III. BOUNDARY: a program for detecting boundaries in ecological landscapes. IRRI Technical Bulletin No.3. International Rice Research Institute, Manila, Philippines. http://books.irri.org/TechnicalBulletin3_content.pdf

Ziliak ST, McCloskey DN. 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, USA. https://books.google.com/books/about/The_Cult_of_Statistical_Significance.html?id=iuLtAAAAMAAJ&source=kp_book_description