Article

Confidence intervals: Concepts, fallacies, criticisms, 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 30 May 2022; *Accepted 10 June 2022*; *Published online 14 June 2022*; *Published 1 September 2022*

Abstract
For a long time, confidence interval theory is the basis of statistics, and confidence interval has been regarded as an important content of statistical analysis. Almost all statistical textbooks and statistical analysis software contain the contents of confidence intervals, which are used to estimate statistical parameters or parameters of mathematical models, and are an important part of many methods such as interval estimation, analysis of variance, and regression analysis, etc. They are recommended or required by the method guidelines of many reputable journals. So far, confidence interval theory and methods have been widely used in various scientific or engineering fields including life sciences, medicine, environmental science, chemistry, physics, and psychology. However, due to the fallacies or deficiencies of the confidence interval theory and methodology, it has caused a wide range of misuses, and has been criticized more and more in recent years. Some statisticians even suggest abandoning the confidence interval theory. To avoid the problems of classical confidence interval theory, one can use Bayesian credible intervals, use uncertainty methods, calculate confidence intervals by avoiding statistic significance tests, or use the Bootstrap credible interval method proposed by me, etc. In practice, for controlled experiments, multiple replicates or treatments should be designed; for observational experiments, multiple representative samples should be drawn, and even a single sample can be used if sufficient sample size is ensured. It is necessary to implement the whole process control for every procedures from sampling to statistical analysis. Cross-comparison and validation of confidence interval analysis results with other multi-source results should be conducted to obtain the most reliable conclusions. Finally, in addition to writing, publishing and adopting new statistical works and teaching materials as soon as possible, it is imperative to revise and distribute various statistical software in new editions based on new statistics for use.

Keywords confidence interval; fallacies; Bayesian credible interval; Bootstrap credible interval; new statistics.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Concepts of Confidence Intervals

There are many types of interval estimates, and their principle basis and calculation methods are different, however all types or methods are estimates of parameters, including parameter estimates that allow sampling

uncertainty, that is, by calculating statistical parameters (population mean, median, variance, or any other unknown statistics) or parameters in mathematical models to generate a series of values to account for measurement or sampling uncertainty.

Interval estimation has long been regarded as a key aspect of statistical analysis. Of the types of interval estimation, the most popular and commonly used are confidence intervals, which are intervals that, on average, contain the true parameter value in some known proportion of repeated sampling (Morey et al., 2016).

The concept of confidence interval (CI) was first proposed by Neyman (1934, 1937), who believed that the X% confidence interval of the parameter θ was the interval (L, U) generated by the algorithm; for all possible θ values, in repeated sampling, there is an X% probability that the interval contains the true value of θ (Neyman, 1937). On repeated sampling, the interval has a fixed probability that contains the parameter θ. If the algorithm generates an interval containing θ with a probability of 0.5, it is a 50% CI; similarly, the probability of a 99% CI is 0.99. The modern definition of a confidence interval allows a probability of at least X%, not exactly X%.

In his classic paper, Neyman (1937) laid the formal basis for confidence intervals. Assuming the researcher is interested in estimating the parameter θ, Neyman recommends that the researcher perform the following three steps: (1) Conduct an experiment and collect relevant data. (2) Calculate two values, L and U. The smaller L is, the larger U is; an interval (L, U) is formed according to the specified algorithm. (3) Prove that $L<\theta<U$, that is, θ is in this interval. (L, U) is the confidence interval of the parameter θ.

The width of the confidence interval is considered an indicator of the accuracy of the estimate. Confidence intervals are considered to indicate which parameter values are reasonable; a confidence coefficient (e.g., 95%) is considered a reasonable indicator that the true parameter is included in the confidence interval. For example, Masson and Loftus (2003) stated that in the absence of any other information, the confidence interval obtained has a 95% probability of including the population mean. Cumming (2014) states that our interval can be guaranteed to include the parameter with 95% confidential degree (confidence coefficient), and the lower and upper bounds can be considered as possible lower and upper bounds of the parameter. The confidence coefficient for a confidence interval is derived from the algorithm that generated it. Therefore, it is helpful to distinguish an algorithm from a confidence interval: an X% confidence algorithm is any algorithm that generates an interval covering θ in X% of repeated sampling, and a confidence interval is a specific interval generated by such an algorithm. The confidence interval algorithm is a stochastic process for observing and fixing confidence intervals (Morey et al., 2016).

Confidence intervals and statistic significance tests, which have been infamously criticized in recent years (Benjamini et al., 2021; Bergstrom and West, 2021; Hahn and Meeker, 1993; Wasserstein and Lazar, 2016; Grenville, 2019; Hubbard et al., 2019; Tong, 2019; Wasserstein et al., 2019; Xie, 2022a, b; Zhang, 2022) have a strong relationship. All confidence intervals can be obtained by invertion of significance tests and vice versa. There is a one-to-one correspondence between confidence intervals and significance tests. However, significance tests and confidence intervals are not equivalent. For example, if the confidence interval for the difference between two means does not contain 0 but is close to 0, it indicates that the two means are not different in any practical sense (we know a priori that they are not absolutely equal, there is infinite decimal points). Confidence intervals are more informative than significance tests. If confidence intervals are not used as tests, it is widely believed that confidence intervals allow one to avoid the pitfalls of significance tests (Matloff, 2011, 2014).

At present, almost all statistical textbooks and statistical analysis software contain the content of confidence intervals (Fig. 1), which are used to estimate statistical parameters or paramters in mathematical models, etc. They are an important part of interval estimation, ANOVA, regression analysis, etc. Their use is recommended or required by the method guidelines of many reputable journals (Psychonomics Society, 2012;

Wilkinson, 1999; Morey et al., 2016). Clearly, confidence interval theory is the foundation of statistical methodology (Cumming, 2014; Loftus, 1996).

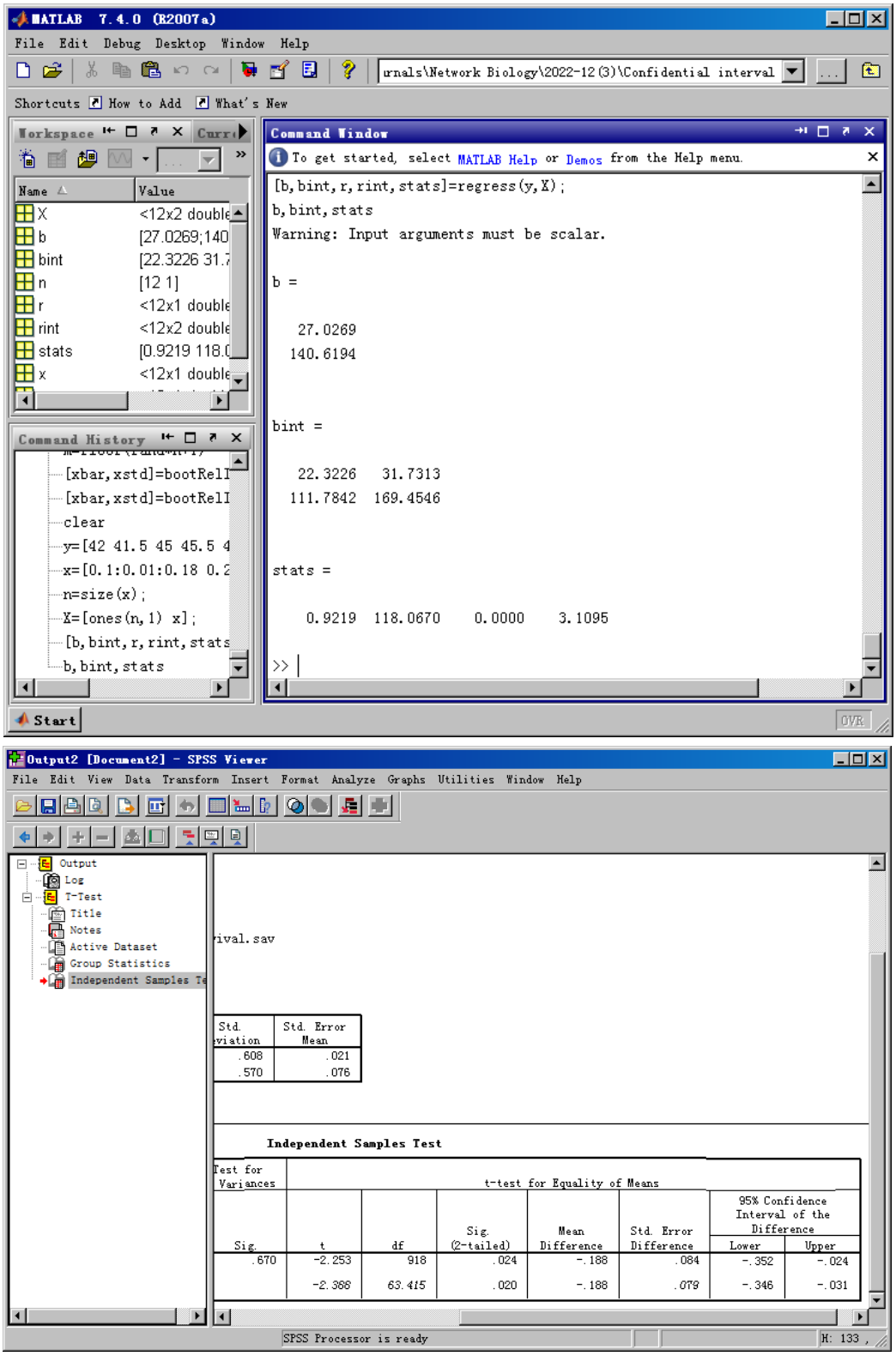

Fig. 1 Confidence intervals of the parameters of a Matlab regression model (the upper) and the confidence intervals in the SPSS independent sample test (the lower).

2 Fallacies, Criticisms and Controversies

2.1 Fallacies and criticism

A deep understanding of what types of inferences are and are not allowed by confidence interval theory is critical in deciding how to conduct relevant scientific research in the future.

Confidence interval theory originated from the famous statistician Neyman, it is a methodology to avoid data reasoning problems by proposing dichotomous statements (Neyman, 1934, 1937, 1941), after widespread dissemination, it eventually became what many people think the best way to make inferences from data (Cumming and Finch, 2005; Cumming and Fidler, 2015) and an easy way to avoid dichotomous statements (Cumming, 2014; Hoekstra et al., 2006; Wilkinson, 1999).

Proponents of confidence intervals argue that confidence intervals have three desirable properties. First, a confidence coefficient can be considered as a measure of the uncertainty that a confidence interval contains a parameter should have. Second, the confidence interval is a measure of the uncertainty of the estimate. Third, confidence intervals contain "likely" or "reasonable" values for the parameter. These all involve reasoning about parameters from observed data, i.e., they are "post-data" reasoning methods (Morey et al., 2016).

In the aforementioned three steps of Neyman's (1937) confidence interval theory for estimating the parameter θ, the result of step (3) is not a belief, a conclusion, or any inference from the data. Furthermore, it is not related to any degree of uncertainty about whether θ is actually in the interval. It's just a dichotomous statement, which means it has a certain probability of being true in the long run. The frequency assessment of the confidence interval algorithm is based on the so-called "power" of the algorithm, that is, the frequency with which wrong parameter values are excluded. On average, better intervals will be shorter and will exclude incorrect parameter values more frequently (Lehmann, 1959; Neyman, 1937; 1941; Welch, 1939). Considering a particular error parameter value $\theta=\theta_0$, different confidence interval algorithms will exclude this error parameter value with different frequencies. If the confidence interval algorithm A excludes θ_0 more frequently than the confidence interval algorithm B on average, then A is better than B in terms of that value. Sometimes we find that one algorithm excludes all wrong parameter values at a higher rate than the others, in which case the first algorithm is more powerful than the second. There might even be an "optimal" confidence interval algorithm that excludes every wrong value of θ with a higher frequency than any other possible confidence interval algorithm. This is similar to the strictest significance test. Although an optimal confidence interval algorithm does not always exist, we can always compare one algorithm to another to decide which is better (Neyman, 1952). Therefore, the confidence interval algorithm is closely related to hypothesis testing: the confidence interval algorithm controls the rate of inclusion of true values, while a better confidence interval algorithm has a greater ability to exclude false values.

It has been found, however, that confidence intervals are not without problems. Hacking (2016) points out that they are basically pre-trial rules. That is, before looking at the data, we set a rule for generating confidence intervals, and then calculate the confidence intervals from the data. This can lead to some weird situations. A classic example is a sample of size 2 drawn from a uniform distribution over (α-1, α+1). where the minimum and maximum values form a 50% confidence interval (they will include 50% of the computations regardless of the α value). But if the range > 1, the interval "must" contain α. Here we have a 50% confidence interval, which in some implementations includes the parameter with probability 1 (Matloff, 2014). According to Mayo (1981), the misunderstanding of confidence intervals seems to be rooted in people's desire for confidence intervals to provide something they cannot reasonably provide, namely a measure of the probability, belief, or degree of support that an unknown parameter value lies within a particular interval. Recent studies have shown that this misunderstanding is widespread among researchers (Hoekstra et al., 2014).

In fact, when Neyman first proposed the theory of confidence intervals in the 1930s, people were quick to

doubt the validity of confidence intervals. Almost at the same time, another statistical pioneer, Fisher, Neyman's academic rival, criticized the theory for potentially leading to conflicting inferences. Fisher argues that the confidence interval theory is a broad and very beautiful theory, but it was built at a considerable cost, and it may be worth considering the cost. The first thing to notice is the loss of uniqueness of results, and the danger of apparently contradictory inferences arising there from (Neyman, 1934; Fisher, 1935). Of course, as will be discussed later, these criticisms are valid, but in a broader sense they miss the point. Like proponents of confidence intervals, critics fail to understand that Neyman's goals differ from proponents: Neyman actually developed a theory of behavior aimed at controlling for error rates, rather than a theory of reasoning from data (Neyman, 1941).

Despite the criticism, confidence intervals subsequently grew in popularity, and the theory of confidence intervals became the mainstream paradigm through the multi-year spread of statistics textbooks as the most widely used interval estimator. Its alternatives, such as Bayesian credible intervals and Fisher benchmark intervals, have been ignored for a long time because people do not understand the differences between confidence intervals, Bayesian and benchmark theories, and the inability to interpret the resulting intervals in the same way. However, there are still occasional doubts and criticisms of it in the statistical community. In the most poignant recent example, Morey et al. (2016) elaborated on the pitfalls and limitations of confidence intervals, noting that confidence intervals are not used to infer unknown parameters, and advising the scientific community to abandon confidence interval theory (Huang, 2022a).

There are some common fallacies, both conceptually and in application, about confidence intervals. Morey et al. (2016) pointed out that there are several types of confidence interval fallacies that people often commit.
(1) Fallacy 1: The Fundamental Confidence Fallacy. If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observation interval contains the true value is also X%; alternatively, we have X% confidence that the observed interval contains the true value.

The reasoning behind The Fundamental Confidence Fallacy seems reasonable: from a given sample, we can get any possible confidence interval. If the 95% possible confidence interval contains the true value, then in the absence of any other information, it seems reasonable to say that we have obtained one of the confidence intervals that contains the true value with 95% confidential degree. This interpretation is implied by the name "confidence interval" itself: the word "confidence", in common usage, is closely related to the concepts of rationality and belief. The name "confidence interval", rather than the more accurate "coverage algorithm" etc., encourages the occurrence of The Fundamental Confidence Fallacy.

The key confusion behind The Fundamental Confidence Fallacy is the confusion between what is known before the data is observed - i.e. whatever the confidence interval is, there is a fixed chance of containing the true value - and what is known after the observed data. Frequentist confidence interval theory says nothing at all about the probability that a particular, observed confidence interval contains the true value; it is either 0 (if the interval does not contain the parameter) or 1 (if the interval does contain the true value).

Therefore, confidence intervals cannot be used to assess the certainty that a parameter is within a specific range. The following examples will illustrate that the known information before the calculation interval and the known information after the calculation interval may be different.
(2) Fallacy 2: The Precision Fallacy. The width of the confidence interval indicates how precisely we know the parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge.

Proponents of confidence intervals argue that confidence intervals are useful for assessing the precision of estimable parameters. This is considered to be one of the main reasons why confidence procedures are applied to null hypothesis significance testing (Cumming and Finch, 2005; Cumming, 2014; Fidler and Loftus, 2009;

Loftus, 1993, 1996). For example, Cumming (2014) argues that wide confidence intervals will quickly let us know if our experiments are weak and can only give imprecise estimates. Young and Lewis (1997) noted that it is important to know how accurately point estimates represent true differences between groups. The width of the confidence interval gives us information about the accuracy of the point estimate of the parameter.

Steiger (2004) pointed out that the relationship between confidence interval width and precision is not perfect and in some cases it can be severely affected for a variety of reasons. Morey et al. (2016) have exemplified that, in fact, there is no necessary relationship between the point estimation accuracy and the width of the confidence interval. Analyze data from 50 participants in an experiment. Participants were randomly divided into two groups, A and B, of 25 people each, and half of the data sets could be analyzed separately. As a result, the 95% *t* confidence interval for group A was 52 ± 2, and for group B it was 53 ± 4. The two sets of results were generally consistent, and an equal-weighted average of 52.5 of the two individual point estimates could be used as an overall estimate of the true mean. However, Group A believes that the two methods should not be weighted equally: the CI of Group A is half the width of Group B, and thus Group A is considered to have a more accurate estimate and therefore should be weighted more heavily. However, this cannot be correct, as the estimate weighted for these two means is not the same as the estimate for analyzing the full dataset, which must be 52.5. The error in Group A is to assume that the CIs directly represent post-data precision. In fact, the width of the confidence interval and the uncertainty of the estimated parameter can be negatively correlated in one case and completely uncorrelated in another.

(3) Fallacy 3: The Likelihood Fallacy. The confidence interval contains the possible values of the parameter. Values within the confidence interval are more likely to occur than values outside.

There is a third common interpretation of the confidence interval, for example, Loftus (1996) argues that it gives an indication of how well the observed pattern of the mean should be considered to reflect the underlying pattern of the population mean. This explanatory logic is used when confidence intervals are used to test a theory (Velicer et al., 2008) or to argue for an invalid or in fact invalid hypothesis (Loftus, 1996). In fact, we cannot interpret an observed confidence interval as containing the true value with some probability, nor can we interpret a confidence interval as the precision of an estimate.

Confidence interval algorithms may indeed have a fixed mean probability of containing the true value, but whether to contain a "reasonable" value on any given sample is a different question. Even a "good" confidence interval from the point of view of confidence interval theory can exclude almost all reasonable values, and can be an empty or infinitely narrow interval that excludes all possible values (Blaker and Spjøtvoll , 2000; Dufour, 1997; Steiger, 2004). This is the outcome of our decisions, independent of "reasoning" or "conclusion" (Neyman, 1941). Mayo and Spanos (2006) also point out that just because a particular value is in an interval does not mean it is reasonable to accept it; they call it the "acceptance fallacy". This fallacy is akin to accepting the null hypothesis in a significance test simply because it was not rejected.

2.2 The debate between pre-data theory and post-data theory

Morey et al. (2016) compared five confidence interval algorithms with several examples and believed that among the five algorithms considered, the interval (credible interval) of the Bayesian algorithm is the only one that can be said to have a 50% probability containing the true value when observing the data. More importantly, the ability to interpret intervals in this way comes from Bayesian theory, not from confidence interval theory. Equally important, it is necessary to specify a prior to obtain the desired interval; the interval should be interpreted in accordance with the specified prior. Of the other four algorithms, none can be shown to provide "reasonable" inferences or conclusions from the data, and they have no prior distributions that might form these intervals. From this perspective, Neyman's refusal to draw "conclusions" and "inferences" from the data naturally stems from his theory, which after all does not support such an idea. Of the five algorithms, only the

Bayesian algorithm correctly tracks the estimation accuracy and covers the true value in the expected way, the other algorithms produce intervals that deterministically contain the true value by simple logic, but are still "50%" interval. However, Welch (1939) pointed out that the Bayesian algorithm is not the best way to construct confidence limits.

The difference between frequentist and Bayesian theory stems from the different goals of the two theories. Frequentism is a pre-data theory. It looks to the future, designing algorithms with special averaging properties in repeated sampling (Neyman, 1937; Jaynes, 2003; Mayo, 1981; 1982). As mentioned above, the idea is clearly seen in Neyman's (1941) study: once the algorithm has been deduced, the inference is over. Confidence interval theory is attributed to the mean frequency of including or excluding correct and incorrect parameter values, respectively. Given the observed data, any given inference may (or may not) be plausible, but that is not Neyman's concern, and he denies any conclusions or beliefs based on the data. Bayesian theory, on the other hand, is a post-data theory: Bayesian analysis uses information from the data to determine what can be reasonably believed, based on model assumptions and prior information (Gelman, 2008; Wasserman, 2008).

Morey et al. (2016) argue that post-data inferences using intervals demonstrated by pre-data theory may lead to unreasonable and potentially arbitrary inferences (Berger and Wolpert, 1988; Wagenmakers et al., 2015). Any confidence interval algorithm that does not focus, at least in part, on the properties of the data behind it is incomplete at best.

One of the misconceptions about the relationship between Bayesian inference and frequentist inference is that they will lead to the same inference, so all confidence intervals can simply be interpreted in a Bayesian fashion. For example, in the case of normally distributed data, certain priors result in confidence intervals that are numerically identical to Bayesian confidence intervals computed using a Bayesian posterior (Jeffreys, 1998; Lindley, 1991). This may lead one to suspect that it doesn't matter whether a confidence interval algorithm or a Bayesian algorithm is used. However, Morey et al. (2016) showed that confidence intervals and credible intervals may differ significantly. The only way to be sure that a confidence interval is numerically the same as some credible interval is to prove it.

Because of the explicit use of priors, Bayesian credible intervals support the interpretation of probabilities in terms of likelihood. Bayesian algorithms provide the ability to calculate the reasonability of any given range of values. Because all of these inferences must be made from the posterior distribution, the inferences must remain consistent with each other (Lindley, 1991; Fisher, 1935). In most cases, there is no reason why likelihoods and posteriors cannot augment or even replace confidence intervals (Kruschke, 2010). The arbitrariness of confidence or confidence coefficients is completely avoided through likelihood or a posteriori.

2.3 Other debates on confidence intervals

M. Thomas argues that, just as rejecting a significance test because it reduces all results to a spurious dichotomy, it is unreasonable to reject a confidence interval because all results within the interval are considered equivalent, and vice versa. Significance tests are closely related to confidence intervals, both of which are highly condensed summaries of information about the likelihood function (or the posterior distribution if you are a Bayesian) with due regard to the sample space. We only have problems when we use these tools blindly and uncritically. Uncritical use of significance tests raises far more problems than uncritical use of confidence intervals, but it's probably just a sample size issue, and more work is over-relying on significance tests than confidence intervals. Arguments must be careful because none of our inference systems are fully satisfactory (Matloff, 2014). Frequentist confidence intervals and Bayesian (with "non-informative" priors) intervals are numerically consistent in the estimation of the mean of a normal distribution.

3 Solutions

3.1 Using Bayesian credible intervals

Bayesian credible intervals are derived from Bayesian theory, which is based on Bayesian rule. Bayesian rule expresses the interrelationship among the conditional probability distribution, marginal probability distribution, and joint probability distribution of random variables, and is defined as follows (Upton and Cook, 2008; Pandey et al., 2022; Zhang, 2016, 2018, 2022):

$$Pr(B \mid A) = \frac{Pr(A,B)}{Pr(A)} = Pr(A \mid B) \times \frac{Pr(B)}{Pr(A)}$$

Where A and B are random variables, $Pr(A)$ and $Pr(B)$ are 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, Bayesian rule is expressed in terms of conditional probability (Huang, 2022), and it can also be expressed as: posterior probability $\propto$ prior probability × current probability.

Based primarily on the Bayesian credible interval method, Morey et al. (2016) provide clear guidelines for interpreting and reporting confidence intervals. Morey et al. argue that unless the interpretation of the interval can be justified by some other theory of reasoning, confidence intervals must remain uninterpreted to avoid making arbitrary inferences or inferences that contradict the data. This even holds for good confidence intervals constructed by inverse significance tests (Steiger, 2004).

The guideline argues that any author who chooses to use confidence intervals should ensure that the intervals numerically correspond to the confidence intervals under some reasonable prior. At this point the confidence interval should be called a credible interval. The guidelines recommend against using confidence interval algorithms with unknown Bayesian properties. As noted by Casella (1992), the post-data properties of algorithms are necessary to understand what can be inferred from intervals. Any algorithm that has not yet explored Bayesian properties may have properties that make it unsuitable for post-data inference. If the confidence algorithm does not correspond to a Bayesian algorithm, the user is cautioned not to interpret the confidence interval as containing a parameter with probability X%, i.e., not by the precision of the measurement, nor by saying that it contains a value that should be taken seriously: before sampling, the interval has an X% probability of containing the true value (Hoekstra et al., 2014).

The guideline further warns against reporting confidence intervals without paying attention to the algorithm and corresponding statistics. As described, there are many different ways to construct confidence intervals, and they have different properties. Some will have better frequentist properties than others; some correspond to credible intervals, while others will not. Unfortunately, authors often report confidence intervals without paying attention to how they are constructed. Not knowing which confidence interval algorithm was used can lead to absurd inferences. Also, enough information should be provided so that anyone can calculate different confidence intervals or credible intervals.

Moving from confidence intervals to credible intervals requires a mindset shift away from a test-centric view of intervals. While every confidence interval can be interpreted as a test, a credible interval cannot be interpreted as such. As Berger (2006) states, it is "completely wrong" to assess the Bayesian confidence of a particular parameter value by checking whether it is contained within the credible interval. When testing a particular value of interest (e.g., the null hypothesis), the particular value must be assigned a non-zero probability a priori.

Cumming (2014) proposed so-called "cat's eye" intervals, which correspond to Bayesian posteriors under

"non-informative" priors for normally distributed data. In recent years, researchers interested in learning more about applied Bayesian statistics have developed many excellent resources, including estimation of posterior distributions and credible intervals, such as works by Bolstad and Curran (2016), Lee and Wagenmakers (2013), Lynch (2007), and Jackman (2009) et al.

3.2 Using Bootstrap credible intervals

The main problem in the practical application of the classical confidence interval theory is the lack of sampling information caused by a single sample and (or) a small sample size. According to the Central Limit Theorem, random variables with arbitrary distribution will tend to be normally distributed when the sample size is large enough: from any population with mean μ, a random sample of size n is drawn. When n is large enough, the sampling distribution of $\overline{X}$ approximately follows a normal distribution with mean μ. Based on the idea of the central limit theorem, I hereby propose to use the Bootstrap method to perform random resampling within the sample size of [1, n] (Zhang, 2007, 2011a-b, 2021a-c, 2022; Zhang and Schoenly, 1999a-b), calculate the specified parameters (mean, variance, proportion, etc.) of the resampling; so randomly resampling s times (resampling sample size s=1000, 10000, etc.), the distribution of the calculated specified parameters tends to in a normal distribution. In this way, the following characteristics of the normal distribution can be used to define credible intervals: about 50% of the values are within 0.68 standard deviations of the mean; about 68% of the values are within 1 standard deviation of the mean; about 95% of the values are within 1.96 standard deviations of the mean; about 99% of the values are within 2.58 standard deviations of the mean. The credible interval obtained in this way is called the Bootstrap credible interval. The Bootstrap credible interval algorithm avoids or partially avoids the problems of a single sample and small sample size. Here I give the Matlab algorithm, bootRelInt, to calculate the Bootstrap credible interval for the population mean as follows:

```
function [xbar,xstd]=bootRelInt(x,s)
%x: a sample of size n.
%s: number of bootstrap re-samplings, e.g., 10000, 20000, etc.
%xbar,xstd: estimated mean and standard deviation of total population.
n=max(size(x));
xs=zeros(1,s);
for sim=1:s
m=floor(rand*n+1);
ran=randperm(n);
for i=1:m
xnew(i)=x(ran(i));
end
xs(sim)=xs(sim)+mean(xnew);
end
xbar=mean(xs);
xstd=std(xs);
sprintf(['Estimated Mean and Standard Deviation of Total Population\n','Mean=',num2str(xbar),'; Standard deviation=',num2str(xstd),'\n\n','Credible intervals\n','About 50%% of means will be in the interval [',num2str(xbar-0.68*xstd),',',num2str(xbar+0.68*xstd),']\n','About 68%% of means will be in the interval [',num2str(xbar-xstd),',',num2str(xbar+xstd),']\n','About 95%% of means will be in the interval [',num2str(xbar-1.96*xstd),',',num2str(xbar+1.96*xstd),']\n','About 99%% of means will be in the interval [',num2str(xbar-2.58*xstd),',',num2str(xbar+2.58*xstd),']\n'])
```

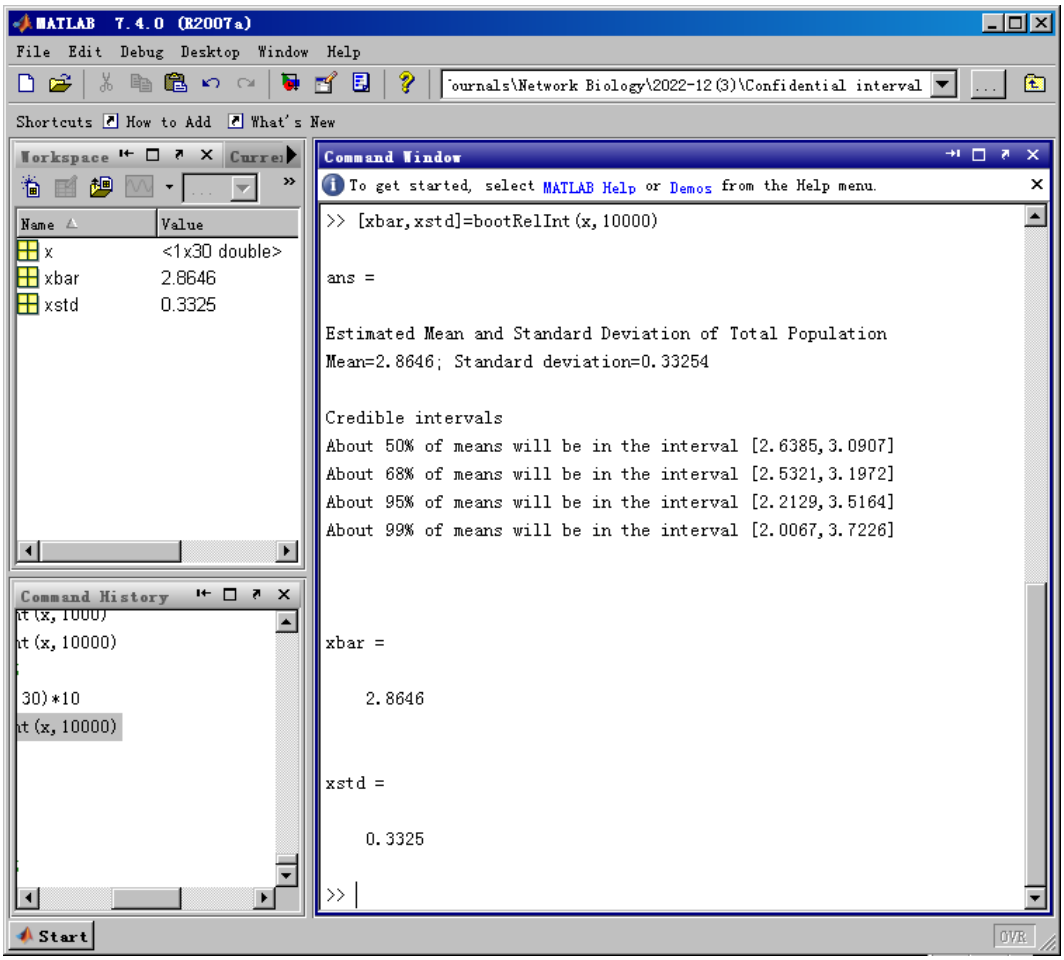

Fig. 2 An example of the Bootstrap credible interval for the population mean.

3.3 Using the uncertainty method

In 1993, the International Organization for Standardization (ISO) and seven international organizations issued the Guidelines for the Representation of Uncertainty in Measurement. The release of Guidelines for the Representation of Uncertainty in Measurement marls that it officially replaces the traditional theory of measurement error with the theory of measurement uncertainty. The statistical basis of measurement uncertainty in Guide to the Representation of Measurement Uncertainty is Neyman's confidence interval theory and the small sample theory based on *t* distribution. If the population standard deviation is unknown, the confidence interval for a confidence level of *p*% obtained from *n* repeated measurements is:

$$(\bar{x}-t_p\frac{s}{\sqrt{n}},\bar{x}+t_p\frac{s}{\sqrt{n}})$$

where $\bar{x}$ is the sample mean, *s* 为 is standard deciation of the sample, t_p is the *t*-value of confidential level *p*%. When the sample size is large enough, the sample mean is approximately normally distributed. In the 2021 International Standard ISO:24578:2021(E), the half-width of the above *t*-interval is not used as the expanded uncertainty, but an unbiased estimate of the expanded uncertainty is adopted (ISO, 2021; Huang, 2022) . Define the following half-width of confidence interval as the unbiased estimate U_p of the expanded uncertainty:

$$U_p = z_p \frac{s}{c_4\sqrt{n}}$$

where, z_p is the z-value of confidential level p%, c_4 is the correction factor of standard deviation for the sample:

$$c_4 = \sqrt{\frac{2}{n\text{-}1}}\Gamma(\frac{n}{2})/\Gamma(\frac{n\text{-}1}{2})$$

where, c_4=0.7979, 0.9213, 0.9515, 0.9650 and 0.9727 correspond to n=2, 4, 6, 8 and 10 respectively。For example, the expanded uncertainty corresponding to the 99.9% confidence level is

$$U_{99.9} = z_{99.9}\frac{s}{c_4\sqrt{n}} = 3.291\frac{0.02121}{0.7979\sqrt{2}} = 0.0619$$

This result is much better than that (9.5) of the classical confidence interval.

In addition, we may use the unified theory of measurement error and uncertainty, proposed by Huang (2018). This unified theory is entirely based on frequentist statistics. It restores the traditional classification of random and systematic errors (primary classification) and retains the Type A and B classifications (secondary classification) of the “Guide to the Expression of Uncertainty in Measurement”.

3.4 Calculating confidence intervals by avoiding significance tests or using other statistics

Proponents of confidence intervals suggest calculating confidence intervals for many other statistics, for example, standardized effect sizes Cohen's *d* (Cumming and Finch, 2001; Zhang, 2022), medians (Bonett and Price, 2002; Olive, 2008), correlations (Zhang, 2015, 2016, 2018; Zhang and Li, 2015; Zou, 2007), ordinal associations (Spearman, 1904; Schoenly and Zhang, 1999; Woods, 2007; Zhang, 2015, 2016, 2018, 2021d), etc.

Steiger and Fouladi (1997) pointed out that the advantage of confidence intervals is that the width of the interval provides a ready indication of the measurement accuracy. Steiger (2004) introduced confidence intervals by emphasizing the desire to avoid significance tests and to focus more on the precision of the estimates. Steiger argues that scientists are more interested in knowing how big the difference between the two groups is (and how precisely it can be determined), rather than whether the difference between the two groups is zero.

Confidence interval algorithms based on significance tests, even a good, robust significance test, often do not provide reasonable inferences. Steiger provides a confidence interval algorithm for the effect size ω^2 by inversing the significance test. ω^2 is the proportion of variance in ANOVA. When there are more than two levels in a one-way design, the parameter ω^2 is used as the effect size. Such confidence intervals were proposed by Steiger (2004) (see also Steiger and Fouladi, 1997), cited by Cumming (2014), implemented in software for social scientists (Kelley, 2007a, b), and evaluated only for their frequency properties (Finch and French, 2012). The issues discussed here are the same as for other relevant confidence intervals, such as confidence intervals for η^2 (Zhang, 2022), partial η^2, noncentrality parameters of the F distribution, signal-to-noise ratio f, RMSSE, and other issues discussed by Steiger (2004) .

To see how confidence intervals are constructed by inversing a significance test, consider a two-sided significance test of size α, which can be thought of as a combination of two one-sided tests of size α/2: one for each tail. When one of the one-tailed tests is rejected, the two-sided test is rejected. To establish a 68% confidence interval, we can use two one-sided tests of size (1-0.68)/2=0.16. Suppose we have a one-way design with three levels of 10 participants in each group. The effect size ω^2 in this design indicates how large F is: larger values of ω^2 tend to produce larger values of F. The F-distribution for a given effect size ω^2 is

called the non-central *F*-distribution. When ω^2=0 (i.e., no effect), the familiar central *F*-distribution is obtained. However, Steiger's confidence interval algorithm produces a "suspicious" confidence interval as long as the *p*-value of the corresponding *F*-test is $p>\alpha/2$. For 95% confidence intervals (meaning $p>0.025$), both Steiger and Fouladi (1997) advise against using confidence intervals for the purposes they (and other confidence interval proponents) suggest. This is not just a theoretical question. In a cursory review of papers citing Steiger (2004), Morey et al. (2016) found that many papers obtained and reported dubious confidence intervals but did not state them (e.g., Cumming et al., 2012; Gilroy and Pearce, 2014; Hamerman and Morewedge, 2015; Lahiri et al., 2013; Hamerman and Morewedge, 2015; Winter et al., 2014). Others do not use confidence intervals, but rely on point estimates of effect sizes and *p*-values (e.g., Hollingdale and Greitemeyer, 2014). But from the *p*-values it can be inferred that if "good practice" is followed and such confidence intervals are calculated, they will obtain intervals that cannot be explained except by the inversed *F* test according to Steiger (2004).

Steiger and Fouladi (1997) concluded that the central problem with confidence intervals was that in order to maintain correct coverage probabilities, a pre-data concern of frequentism, they sacrificed what researchers wanted for confidence intervals: a measure of post-data indexing precision. If our goal is to stay away from significance tests, we should not use unexplainable methods other than inverse significance tests.

3.5 Other suggestions

Personally, I believe that confidence intervals can be retained, as long as the confidence intervals are not absoluteized using a significance test in the application. The sample size should be large enough, or preferably the multi-sample confidence intervals should be used. For a single sample, as mentioned above, the resampling method can be used to increase the amount of sampling information. Instead of using confidence intervals as the only interpretation, the confidence interval results should be cross-compared and validated with other multi-source results to obtain the most credible conclusions.

4 Discussion

In the century since its inception, statistics has been forging ahead amid doubts, criticisms and debates. Statistics do have serious problems (Matloff, 2014). It has even been argued that statistics is an outdated field, with most of its common core from an era when so many assumptions were required, that it would be absurd to teach anyone to use critical values. In view of the characteristics of statistics and many problems in theory, especially in practice, many professional statisticians also show lack of confidence. For example, M. Thomas argues that in his 40-year career as a professional statistician, he has never done a "correct" analysis, and everything is based on assumptions that are at best approximate (Matloff, 2011, 2014). Entering the era of big data, with the massive presentation of data information and the unprecedented improvement of computing power, more and more people are mining information based on overall analysis rather than classical statistics that based on sampling, coupled with the rapid progress of artificial intelligence technology, the sense of statistical frustration and powerlessness is likely to grow.

I don't think there is a substantial problem with statistics itself. Statistics are based on sample data, and its research objects and conclusions are naturally imprecise and uncertain. We cannot approach statistical theories and methods for uncertainty with deterministic expectations and mindsets, which are the subjective source of many problems in the application of statistics. For example, dichotomizing the continous *p*-value problem for for statistical significance tests, etc. The natural design of the human brain is more suitable for deterministic analysis than uncertainty analysis. When faced with uncertain problems, people will consciously or unconsciously tend to look for deterministic solutions. What we need to adjust is to return to the essence and apply statistical theories and methods with uncertain expectations and mentality. This requires that we cannot

make statistical conclusions certain and absolute. For controlled experiments, multiple replicates or treatments should be designed; for observational experiments, multiple representative samples should be drawn, and sufficient sample size should be ensured even if a single sample is used. More reasonable statistical methods should be used. It is necessary to implement the whole process control from sampling to statistical analysis. Statistical analysis results should be cross-compared and validated with other multi-source results to obtain the most reliable conclusions (Zhang, 2022).

Most of the problems discussed in statistics in recent years have a long history, and some solutions have already been proposed. The reason why these problems have received unprecedented attention and criticism in recent years is mainly due to the increasingly serious problems of reproducibility crisis in scientific research and academic misconducts. A paper published in 2005, "Why Most Published Research Findings Are False ", sparked the first widespread discussion about the reproducibility of scientific research (Ioannidis, 2005; Wu, 2022; Zhang, 2022). In 2012, in an article published in *Nature*, the American biotechnology company repeated the experiments in 53 so-called landmark papers, but were only able to confirm the results of 11% of them, causing a shock in the scientific community. The excessive pursuit of positive results has made many new scientific research achievements and discoveries considered false positives and cannot be confirmed by repeated experiments. According to a survey 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 that they could not replicate their own experiments; 52% of the investigators believed that there were significant experimental reproducibility crisis. Most researchers indicated that they had failed repeated experiments (Baker, 2016a, b). The problems of reproducibility crisis in scientific research and academic misconducts are due in part to the misuse and abuse of statistical methods, as well as problems with experimental design and sampling design.

In 2021, *Nature* pointed out that researchers, research funders and publishers must take reproducibility more seriously (Nature Editorial, 2021). Consider the horrific sight that two-thirds or more of the scientific findings published in the past cannot be replicated. Errington proposed to elevate reproducibility to the same level as research novelty, reiterating that reproducibility is an important feature of scientific research (Errington et al., 2021; Zhang, 2022). Strictly speaking, non-reproducible research results are false or spurious, or at least moot. Not only do they waste resources, they are not beneficial to science, but more importantly, they harm science by being cited and supported. We should not rest assured that we are not required to repeat the verification, but should face the reality and act consciously to change the scientific research paradigm. Therefore, attaching importance to experimental design and sampling design, as well as the correct application of statistical methods, etc., are urgent issues that the majority of researchers must face. In addition, in addition to writing, publishing and adopting new statistical works and teaching materials as soon as possible, it is imperative to revise and distribute various statistical software in new editions based on new statistics for use.

References

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

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

Berger JO. 2006. Bayes factors. In: Encyclopedia of Statistical Sciences (2nd edition) Vol 1 (Kotz S, Balakrishnan N, Read C, Vidakovic B, Johnson NL, eds). 378-386, Hoboken, John Wiley & Sons, New Jersey, USA. https://www.wiley.com/en-us/Encyclopedia+of+Statistical+Sciences%2C+Volume+1%2C+2nd+Edition-p-9780471743910

Berger JO, Wolpert RL. 1988. The Likelihood Principle (2nd edition). Institute of Mathematical Statistics, Hayward, CA, USA. http://web.uvic.ca/~dgiles/blog/Berger_and_Wolpert.pdf

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

Blaker H, Spjøtvoll E. 2000. Paradoxes and improvements in interval estimation. The American Statistician, 544: 242-247. https://doi.org/10.1080/00031305.2000.10474555

Bolstad W, Curran JM. 2016. Introduction to Bayesian Statistics (3rd ed). Wiley, USA. https://www.wiley.com/en-us/Introduction+to+Bayesian+Statistics%2C+3rd+Edition-p-9781118593226

Bonett DG, Price RM. 2002. Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. Psychological Methods, 7: 370-383. https://doi.org/10.1037/1082-989x.7.3.370

Casella, G. 1992. Conditional inference from confidence sets. Lecture Notes-Monograph Series, 17: 1-12. https://doi.org/10.1214/lnms/1215458835

Cumming G. 2014. The new statistics: Why and how. Psychological Science, 25: 7-29. https://doi.org/10.1177/0956797613504966

Cumming G, Fidler F. 2015. Confidence intervals: Better answers to better questions. Zeitschrift f¨ur Psychologie, 217: 15-26. https://doi.org/10.1027/0044-3409.217.1.15

Cumming G, Finch S. 2001. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educational and Psychological Measurement, 61: 532-574. https://doi.org/10.1177/0013164401614002

Cumming G, Finch S. 2005. Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist, 602: 170-180. https://doi.org/10.1037/0003-066X.60.2.170

Cumming SP, Sherar LB, Gammon C, Standage M, Malina RM. 2012. Physical activity and physical self-concept in adolescence: A comparison of girls at the extremes of the biological maturation continuum. Journal of Research on Adolescence, 224: 746-757. https://doi.org/10.1111/j.1532-7795.2012.00821.x

Dufour JM. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. Econometrica, 656: 1365-1387. https://www.econometricsociety.org/publications/econometrica/1997/11/01/some-impossibility-theorems-econometrics-applications

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

Fidler F, Loftus GR. 2009. Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. Zeitschrift fur Psychologie, 2171: 27-37. https://doi.org/10.1027/0044-3409.217.1.27

Finch WH, French BF. 2012. A comparison of methods for estimating confidence intervals for omega-squared effect size. Educational and Psychological Measurement, 721: 68-77. https://doi.org/10.1177/0013164411406533

Fisher RA. 1935. The fiducial argument in statistical inference. Annals of Eugenics, 6: 391-398. https://doi.org/10.1111/j.1469-1809.1935.tb02120.x

Fisher RA. 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society Series B Methodological, 17: 69-78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Gelman A. 2008. Rejoinder. Bayesian Analysis, 3: 467-478. http://www.stat.columbia.edu/~gelman/research/published/badbayesresponsemain.pdf

Gilroy KE, Pearce JM. 2014. The role of local, distal, and global information in latent spatial learning. Journal of Experimental Psychology, 402: 212-224. https://doi.org/10.1037/xan0000017

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

Hacking I. 2016. Logic of Statistical Inference. Cambridge University Press, USA. https://doi.org/10.1017/CBO9781316534960

Hamerman EJ, Morewedge CK. 2015. Reliance on luck: Identifying which achievement goals elicit superstitious behavior. Personality and Social Psychology Bulletin, 413: 323-335. https://doi.org/10.1177/0146167214565055

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

Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. 2014. Robust misinterpretation of confidence intervals. Psychonomic Bulletin and Review, 215: 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

Hollingdale J, Greitemeyer T. 2014. The effect of online violent video games on levels of aggression. PLoS ONE, 911: e111-790. https://doi.org/10.1371/journal.pone.0111790

Huang HN. 2018. A unified theory of measurement errors and uncertainties. Measurement Science and Technology, 29(12). https://iopscience.iop.org/article/10.1088/1361-6501/aae50f/meta

Huang HN. 2022a. A Fallacy of Confidence Interval Theory for Measurement of Uncertainty Evaluation. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1340388.html. Accessed 2022-5-30

Huang HN. 2022b. Bayesian *vs* Frequentist: A Debate Spanning Two and A Half Centuries. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1331814.html. Accessed 2022-4-30

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

Ioannidis JPA. 2005. Why Most Published Research Findings Are False. Plos Medicine, https://doi.org/10.1371/journal.pmed.0020124

ISO 24578:2021(E). 2021. Hydrometry — Acoustic Doppler Profiler — Method and Application For Measurement of Flow in Open Channels From A Moving Boat (1st edition). Geneva, Switzerland. https://www.iso.org/standard/70758.html

Jackman S. 2009. Bayesian Analysis for The Social Sciences. Wiley, UK. https://www.wiley.com/en-us/Bayesian+Analysis+for+the+Social+Sciences-p-9780470011546

Jaynes E. 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, USA. https://doi.org/10.1017/CBO9780511790423

Jeffreys H. 1998. Theory of Probability (3rd edition). Oxford University Press, New York, USA. https://global.oup.com/academic/product/the-theory-of-probability-9780198503682?cc=cn&lang=en&

Kelley K. 2007a. Confidence intervals for standardized effect sizes: Theory, application, and implementation. Journal of Statistical Software, 20(8): 1-24. https://doi.org/10.18637/jss.v020.i08

Kelley K. 2007b. Methods for the behavioral, educational, and social sciences: An R package. Behavioral Research Methods, 394: 979-984. https://doi.org/10.3758/bf03192993

Kruschke JK. 2010. What to believe: Bayesian methods for data analysis. Trends in Cognitive Sciences, 147:

293-300. https://doi.org/10.1016/j.tics.2010.05.001

Lahiri DK, Maloney B, Rogers JT, Ge YW. 2013. PuF, an antimetastatic and developmental signaling protein, interacts with the Alzheimer's amyloid-beta precursor protein via a tissuespecific proximal regulatory element PRE. Bmc Genomics, 14: 68. https://doi.org/10.1186/1471-2164-14-68

Lee MD, Wagenmakers EJ. 2013. Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press, USA. https://bayesmodels.com/

Lehmann EH. 1959. Testing Statistical Hypotheses. Wiley, New York, USA. https://doi.org/10.1002/nav.3800080209

Lindley DV. 1991. Making Decisions (2nd ed). Wiley, London, UK. https://www.wiley.com/en-us/Making+Decisions%2C+2nd+Edition-p-9780471908081

Loftus GR. 1993. A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. Behavior Research Methods. Instrumentation and Computers, 25: 250-256. https://doi.org/10.3758/BF03204506

Loftus GR. 1996. Psychology will be a much better science when we change the way we analyze data. Current Directions in Psychological Science, 5: 161-171. https://doi.org/10.1111/1467-8721.ep11512376

Lynch SM. 2007. Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. Springer, New York, USA. https://doi.org/10.1007/978-0-387-71265-9

Masson MEJ, Loftus GR. 2003. Using confidence intervals for graphically based data interpretation. Canadian Journal of Experimental Psychology, 57: 203-220. https://doi.org/10.1037/h0087426

Matloff N. 2014. Why Are We Still Teaching *t*-Tests? Mad (Data) Scientist. https://matloff.wordpress.com/2014/09/15/why-are-we-still-teaching-about-t-tests/. Accessed on June 1, 2022

Matloff N. 2011. From Algorithms to Z-Scores: Probabilistic and Statistical Modeling in Computer Science. https://heather.cs.ucdavis.edu/probstatbook. Accessed 2022-6-1

Mayo DG. 1981. In defense of the Neyman-Pearson theory of confidence intervals. Philosophy of Science, 482: 269-280. https://doi.org/10.1086/288996

Mayo DG. 1982. On After-Trial Criticisms of Neyman-Pearson Theory of Statistics. Proceedings of the Biennial Meeting of the Philosophy of Science Association, PSA, USA. https://www.journals.uchicago.edu/doi/abs/10.1086/psaprocbienmeetp.1982.1.192663

Mayo DG, Spanos A. 2006. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. British Journal for The Philosophy Of Science, 57: 323-357. https://doi.org/10.1093/bjps/axl003

Morey RD, Rouder JN, Verhagen J, Wagenmakers EJ. 2014. Why hypothesis tests are essential for psychological science: A comment on Cumming. Psychological Science, 1289-1290. https://doi.org/10.1177/0956797614525969

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagemakers EJ. 2016. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev, 23: 103-123. https://doi.org/10.3758/s13423-015-0947-8

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. 1934. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 974: 558-625. https://www.stat.cmu.edu/~brian/905-2008/papers/neyman-1934-jrss.pdf

Neyman J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London. Series A. Mathematical and Physical Sciences, 236: 333-380. https://doi.org/10.1098/rsta.1937.0005

Neyman J. 1941. Fiducial argument and the theory of confidence intervals. Biometrika, 322: 128-150. https://doi.org/10.2307/2332207

Olive DJ. 2008. Applied Robust Statistics. http://lagrange.math.siu.edu/Olive/ol-bookp.htm. Accessed 2012-5-20

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

Psychonomics Society 2012. Psychonomic Society Guidelines on Statistical Issues. http://www.springer.com/psychology?SGWID=0-10126-6-1390050-0. Accessed 2022-6-1

Schoenly KG, Zhang WJ. 1999. IRRI Biodiversity Software Series. V. RARE, SPPDISS, and SPPANK: programs for detecting between-sample difference in community structure. IRRI Technical Bulletin No.5. International Rice Research Institute, Manila, Philippines. http://books.irri.org/TechnicalBulletin5_content.pdf

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

Steiger JH. 2004. Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological Methods, 92: 164-182. https://doi.org/10.1037/1082-989X.9.2.164

Steiger JH, Fouladi RT. 1997. Noncentrality interval estimation and the evaluation of statistical models. In: What If There Were No Significance Tests? (Harlow L, Mulaik S, Steiger J, eds). 221-257, Mahwah, Erlbaum, New Jersey, USA. https://www.routledge.com/What-If-There-Were-No-Significance-Tests-Classic-Edition/Harlow-Mulaik-Steiger/p/book/9781138892477

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

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

Velice, WF, Cumming G, Fava JL, Rossi JS, Prochaska JO, Johnson J. 2008. Theory testing using quantitative predictions of effect size. Applied Psychology, 574: 589-608. https://doi.org/10.1111/j.1464-0597.2008.00348.x

Wagenmakers EJ, Verhagen J, Ly A, Bakker M, Lee D, Matzke MD, Rouder JN, Morey RD. 2015. A power fallacy. Behavioral Research Methods, 47(4): 913-917. https://doi.org/10.3758/s13428-014-0517-4

Wasserman L. 2008. Comment on article by Gelman. Bayesian Analysis, 3: 463-466. https://doi.org/10.1214/08-BA318D

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

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

Welch BL. 1939. On confidence limits and sufficiency, with particular reference to parameters of location. The Annals of Mathematical Statistics, 101: 58-69. https://doi.org/10.1214/aoms/1177732246

Wilkinson L. 1999. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54: 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Winter C, Van Acker, F, Bonduelle, M, Desmyttere, S, De Schrijver, F, Nekkebroeck J. 2014. Cognitive and psychomotor development of 5-to 6-year-old singletons born after PGD: A prospective case-controlled

matched study. Human Reproduction, 299: 1968-1977. https://neuro.unboundmedicine.com/medline/citation/24993932/Cognitive_and_psychomotor_development_of_5__to_6_year_old_singletons_born_after_PGD:_a_prospective_case_controlled_matched_study_

Woods CM. 2007. Confidence intervals for gamma-family measures of ordinal association. Psychological Methods, 122: 185-204. https://doi.org/10.1037/1082-989X.12.2.185

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

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

Young, KD, Lewis RJ. 1997. What is confidence? Part 1: The use and interpretation of confidence intervals. Annals of Emergency Medicine, 303: 307-310. https://doi.org/10.1016/S0196-0644(97)70166-5

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 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. 2015. 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. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK. https://doi.org/10.1142/q0149

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. 2021d. Construction and analysis of the word network based on the Random Reading Frame (RRF) method. Network Biology, 11(3): 154-193. http://www.iaees.org/publications/journals/nb/articles/2021-11(3)/construction-and-analysis-of-word-network-from-Random-Reading-Frame.pdf

Zhang WJ. 2022. *p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3): 80-122. http://www.iaees.org/publications/journals/ces/articles/2022-12(3)/p-value-based-statistical-significance-tests.pdf

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, 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

Zou GY. 2007. Toward using confidence intervals to compare correlations. Psychological Methods, 124: 399-413. https://doi.org/10.1037/1082-989X.12.4.399