Article

Dilemma of *t*-tests: Retaining or discarding choice and solutions

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

Received 9 June 2022; *Accepted 18 June 2022*; *Published online 24 June 2022*; *Published 1 December 2022*

Abstract
The *t*-test theory has laid the foundation of modern statistics and it is one of the main contents of statistics. This theory can be found in all statistics textbooks and is at the core of almost all applied statistics courses. At the same time, almost all statistical software or tools have *t*-test content, such as Matlab, SAS, SPSS, R, etc. However, *t*-test theory has been widely criticized in recent years due to its theoretical flaws and misuse. The *t*-test is only used for the problems of normal distribution population with small sample size. Even so, its sample size cannot be too small due to problems such as *t*-transformation distortion. In terms of significance test, the *t*-test has the general defects of statistical significance tests, coupled with the inherent fallacies of confidence intervals, and the peculiar uncertainty problems of *t*-intervals, make the *t*-test methodology obviously insufficient. The *t*-test theory is faced with the retaining or discarding decision in statistics, and some statisticians have advocated and abolished the *t*-test theory from statistics textbooks. As a statistical significance test, the solutions of *t*-tests include using Bayesian methods, performing meta-analyses, using effect sizes, stressing statistical validity, using nonparametric statistics, using good experimental and sampling designs and determining appropriate sample size, the network methods are used instead of the reductionist method to obtain and analyze the data, and the statistical conclusions are combined with the mechanism analysis to draw scientific inferences, etc. As a *t*-interval uncertainty problem, its solutions include using the Bayesian credible interval method, using the Bootstrap credible interval method, inferring directly from the central limit theorem, using the unified theory of uncertainty, etc.

Keywords *t*-tests; *t*-interval; statistical significance tests; uncertainty; Bootstrap credible interval; Bayesian credible interval.

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 *t*-tests

Assuming that $X \sim N(0,1)$, $Y \sim \chi^2(n)$, where X and Y are independent of each other, then the random variable

$$T = \frac{X}{\sqrt{Y/n}}$$

is called following the *T*-distribution. The *T*-distribution was firstly proposed in 1908 by W. S. Gosset, an engineer at the Guinness brewery in Dublin, Ireland (Student, 1908). Based on the *T*-distribution, Gosset proposes a method for estimating possible errors. Due to industry confidentiality, Gosset used "Student" as the author's name when publishing the paper. Since then, the famous statistician, Fisher, has made further improvements and named it *T*-distribution (or *t*-distribution, the same below), and the corresponding test theory is called *T*-test (Student's *t*-test, or t-test, the same below) (Fisher, 1924). The *t*-test theory laid the foundation of modern statistics and marked the beginning of statistics as a science.

If the random variable *T* follows a *t*-distribution, then the probability density function, $f(t)$, of *T* is (Fig. 1)

$$f(t)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}$$

where, *n* is a positive integer (degree of freedom), and *t* is the value of the random variable *T*. The curve of probability density function $f(t)$ is symmetric about $t=0$ and takes the maximum value at $t=0$.

Assuming that $T\sim t(n)$, then $\lim_{n\to\infty} T \to N(0,1)$, in other words, when the degree of freedom *n* is large enough, the *t*-distribution approximates the standard normal distribution (Fig. 1). On the other hand, $t(1)$ is the Cauchy distribution. Mean and variance are absent from the Cauchy distribution.

Assuming that $T\sim t(n)$, then the cumulative distribution function of *T* is $F(t)=P\{T<t\}=\int_{-\infty}^{t} f(x)dx$.

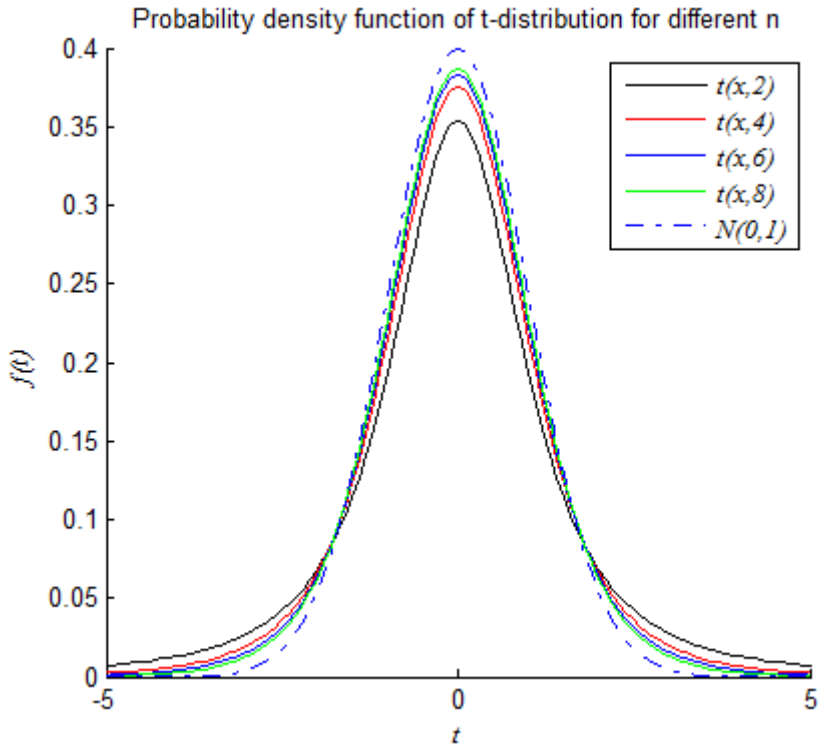

Fig. 1 Probability density functions of *t*-distribution (n=2, 4, 6, 8) and the standard normal distribution, $N(0,1)$.

The *t*-test theory is based on the *t*-distribution theory and is mainly used for the problems of small samples and normal distributions with unknown standard deviation σ (Student, 1908). The *t*-test is mainly used to calculate the probability of the difference, and to compare whether the mean difference is significant,

etc. Common *t*-tests include single population test, double population test, and paired samples test, etc. The *t*-test is also used in many other procedures.

Several representative procedures of *t*-tests are as follows:

(1) Parametric test of population mean of normal distribution (single population test)

The population variance of normal distribution is unknown, and the sample is a small sample, then the random variable that sample mean is standardized follows a *t*-distribution with *n*-1 degrees of freedom:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

The parametric test rules for the population mean (μ) are as follows:

$$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0, |t| < t_\alpha(n-1), p = F(t) \times 2 > \alpha(H_0 : True)$$

$$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0, t > t_\alpha(n-1), p = F(t) > \alpha(H_0 : True)$$

$$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0, t < -t_\alpha(n-1), p = 1 - F(t) > \alpha(H_0 : True)$$

(2) Parametric test of the difference between the means of two populations of independent normal distributions (two-population test)

$\overline{x}_1 - \overline{x}_2$ follows the *t*-distribution with n_1+n_2-2 degrees of freedom, the variances of σ_1^2 and σ_2^2 of the two normal populations (populations of normal distributions) are unknown, and $\sigma_1^2 = \sigma_2^2$, and the sample is a small sample, then we have:

$$t = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right]}}$$

The parametric test rules for the difference between the means of two populations of independent normal distributions are as follows:

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2, |t| < t_\alpha(n_1 + n_2 - 2), p = F(t) \times 2 > \alpha(H_0 : True)$$

$$H_0 : \mu_1 \leq \mu_2, H_1 : \mu_1 > \mu_2, t > t_\alpha(n_1 + n_2 - 2), p = F(t) > \alpha(H_0 : True)$$

$$H_0 : \mu_1 \geq \mu_2, H_1 : \mu_1 < \mu_2, t < -t_\alpha(n_1 + n_2 - 2), p = 1 - F(t) > \alpha(H_0 : True)$$

(3) Parametric test of the difference between the means of two correlated normal populations (paired samples test)

The variance of the difference between the means of two correlated populations of normal distributions (σ_d^2) is unknown, and the sample is a small sample, then we have:

$$t = \frac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$$

The parametric test rule for the difference between the means of two correlated normal populations is as the following:

$$H_0 : \mu_d = 0, H_0 : \mu_d \neq 0, |t| < t_\alpha(n-1), p = 2 \times F(t) > \alpha(H_0 : True)$$

In the *t*-test, a distinction needs to be made between one-tailed and two-tailed tests. The one-tailed test cutoff is smaller than the two-tailed test cutoff, so it is easier to reject the null hypothesis. It has been argued that if the differences are in a specific direction, it is only necessary to consider the one-tailed probability distribution and split the resulting *p*-value of the *t*-test in half. It has also been argued that the *p*-value of the standard two-tailed *t*-test should be reported in all cases.

One-tailed tests include left-sided test and right-sided test:

$$H_0 : \mu \geq \mu_0 ; H_1 : \mu < \mu_0$$
$$H_0 : \mu \leq \mu_0 ; H_1 : \mu > \mu_0$$

In the two-tailed test, the null hypothesis is an equality, and the alternative hypothesis is an inequality:

$$H_0 : \mu = \mu_0 ; H_1 : \mu \neq \mu_0$$

Two-tailed test, no matter whether the difference is positive or negative, given the significance level α, it must be equally distributed to the left and right sides according to the principle of normal symmetry: each side is $\alpha/2$, correspondingly, the lower critical value is $-t_{\alpha/2}$, and the upper critical value is is $t_{\alpha/2}$.

For multiple-group difference tests, we can use analysis of variance (ANOVA). ANOVA is a generalization of the *t*-test.

(4) Significance tests of correlations

For Pearson correlation (*r*) or Spearman rank correlation (*r*) (Spearman, 1904; Zhang, 2015, 2018; Zhang and Li, 2015), calculate

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

The test rule for correlation is

$$H_0 : r = 0(Correlation : True), H_1 : r \neq 0(Correlation : False), |t| < t_\alpha (n-2)(H_0 : False)$$

(5) Interval estimation of the normal population mean

If the variance of the normal population is unknown and the sample is a small sample, the sample variance (s^2) is used to replace the population variance. The normalized sample mean ($\bar{x}$) is the random variable *t*:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

which follows the *t*-distribution with *n*-1 degrees of freedom. The 1-α confidence interval for the population mean (μ) at significance level α is (Zhang, 2022a):

$$\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

and its half width

$$U_t = t_{\alpha/2} \frac{s}{\sqrt{n}}, s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

is called *t*-interval.

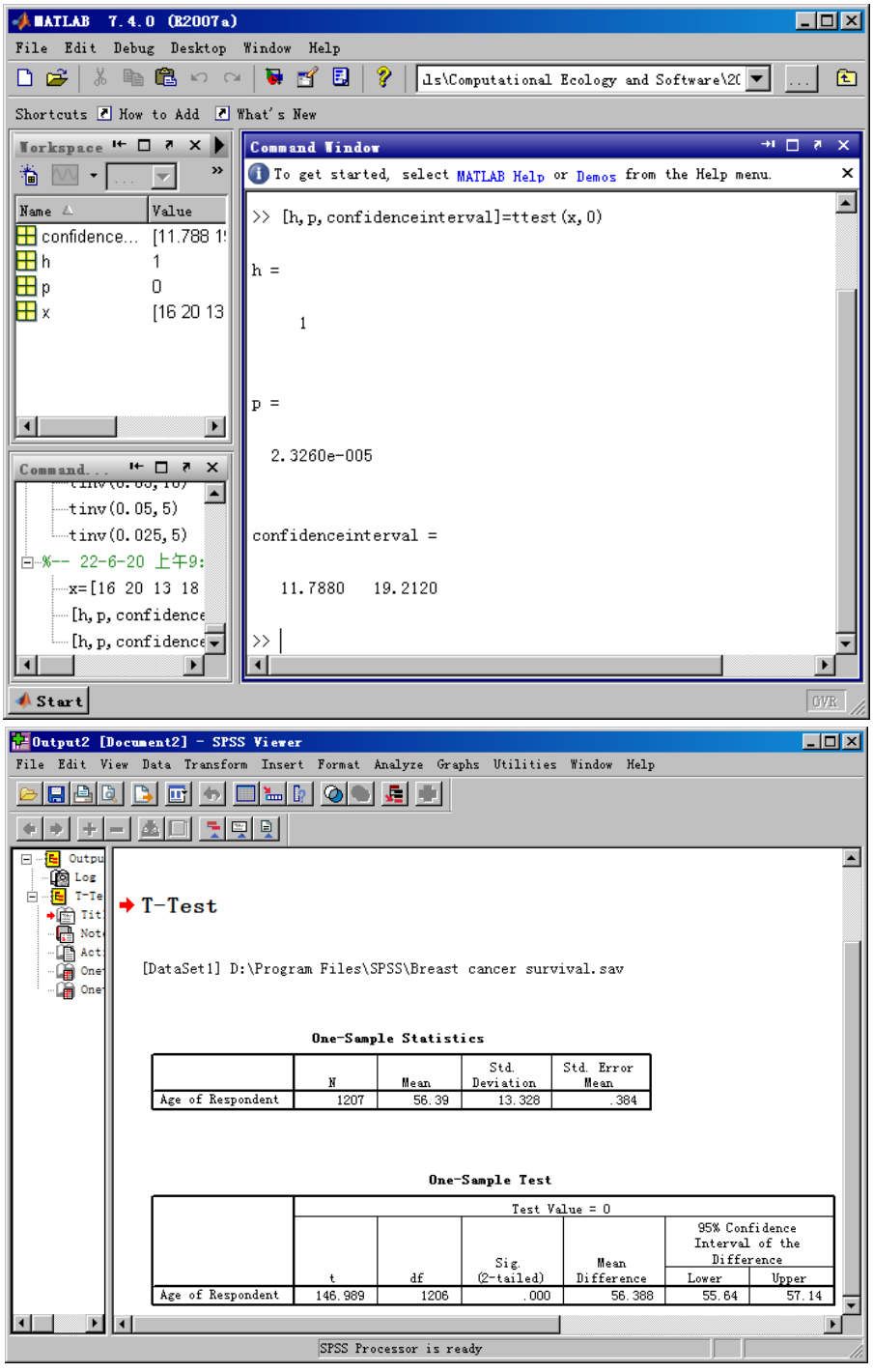

Fig. 2 The results of the parametric test calculation of the normal population mean in Matlab (top) and SPSS (bottom).

The *t*-test theory is one of the main contents of statistics, which can be found in all statistics textbooks, and is the core of almost all applied statistics courses. Almost all statistical software or tools have *t*-test contents, such as Matlab, SAS, SPSS, R, Stata, Microsoft Excel, Python, Minitab, etc. (Fig. 2). For example, in Matlab, several main functions for *t*-distribution and *t*-tests are:

trnd(n): *t*-distributed random number with *n* degrees of freedom.

tpdf(t, n): *t*-distribution probability density function $f(t)$ with *n* degrees of freedom.

tinv(α, n): The inverse cumulative distribution function of the *t*-distribution with *n* degrees of freedom the significance level *α*, that is, the *t*-value.

tcdf(t, n): *t*-distribution cumulative distribution function with *n* degrees of freedom, $F(t)=P\{T<t\}$.

ttest(x, mu, alpha, tail): Parametric test of the mean of a normal population.

ttest2(x, y, alpha, tail): Parametric test of the difference between two independent normal populations.

2 Defects and Criticisms

In the common probability distributions, the normal distribution (Gaussian distribution) comes from the measurement error and is used to describe the law of error distribution. The Rayleigh distribution comes from the wave energy and is used to describe the spectrum of wave energy. Weibull distribution is derived from fracture and failure risks and is used to describe reliability and longevity. Power-law distribution is ubiquitous in the physical world and is used to describe the node degree distribution of scale-free networks, f self-similarity in ractal structures, and more. However, the *t*-distribution does not come from the real world. It is a probability distribution with no physical meaning, and is difficult to correspond to real-world prototypes. The *t*-distribution is also mathematically flawed, and the standard deviation from *t*(1) and *t*(3) does not exist (Huang, 2021a-b). Relevant defects in the *t*-distribution also form part of the source of the *t*-test problems.

2.1 *t*-test problems

In terms of significance tests, the aforementioned *t*-test procedures (1)-(4) can be attributed to statistical significance tests based on *p*-values. In a *t*-test, the larger the *p*-value, the greater the confidence to accept the null hypothesis. For the pitfalls and problems of statistical significance tests, see my paper "*p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond" (Zhang, 2022b), and other related literature (Wasserstein and Lazar, 2016; Benjamini et al., 2018, 2021; Grenville, 2019; Bergstrom and West, 2021).

Regarding the problem of *t*-statistical significance test, Matloff, a professor of statistics at the University of California, criticized that some statistical programs in software such as R focus too much on significance tests (including *t*-tests) (Matloff, 2014). Harrell (2012) suggested in r-devel that at least the "asterisk system" in R output should be removed (different numbers of asterisks indicate different levels of statistical significance *p*-values).

The essence of statistical inference is to infer unknown population parameters based on sample statistics. No matter in theory or practical application, there is no need to artificially distinguish between large samples and small samples. In other words, a valid statistical inference method should be applicable to any sample size (Huang 2021a, b). However, the *t*-test is mainly for small sample problems. For this reason, Ilya Kipnis believes that *t*-tests are only meaningful when data are difficult to obtain. Even so, the *t*-test sample is insufficiently informative and at worst highly misleading. In this era of easy data collection, the *t*-test is of less and less significance (Matloff, 2011, 2014).

Assuming that the population is normally distributed (i.e. normal population), inferences based on the *t*-distribution are robust and their approximations are reasonable. However, as Matloff pointed out, in many cases we don't know if the population follows a normal distribution: it can be any probability distribution. In fact, all real-life random variables are bounded (as opposed to an infinitely supported normal distribution) and discrete (as opposed to a continuous normal distribution). Therefore, Matloff advocates skipping the *t*-test and inferring directly from the Central Limit Theorem (Matloff, 2011, 2014). He argues that the *t*-test is about the mean, so it is easy to prove by the Central Limit Theorem that the sampling distribution is always normal. For this reason, Matloff eliminated the content of *t*-distribution and *t*-interval in his statistics monograph (Matloff, 2011). The American statistician Huang (2018b, 2021a-b) agrees with Matloff and

gives examples to illustrate the validity of inferences based on the Central Limit Theorem.

In view of the problems arised from statistical significance tests such as *t*-test, in early 2015, the international journal, *Basic and Applied Social Psychology* (BASP) declared the null hypothesis significance test procedure (NHSTP) (*p*-value, *t*-value, etc.) is invalid (Trafimow and Marks, 2015).

2.2 *t*-interval problems

Since the population mean μ is unknown, the error cannot be known, resulting in uncertainty. Whether the confidence interval contains μ is uncertain, and it thus is an uncertainty measure. The aforementioned *t*-interval is a measure of uncertainty also. Student (1908) and Craig (1927) call this type of uncertainty possible error.

Similar to the aforementioned *t*-interval, if the standard deviation of normal population, σ is known or the sample size *n* is large enough, the uncertainty is the uncertainty based on *z*, that is, the *z*-interval:

$$U_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

According to the "Guide to the Expression of Uncertainty in Measurement" (JCGM, 2008a-b), the *t*-interval is the Type A uncertainty evaluation, and the *z*-interval is the Type B uncertainty evaluation (Huang, 2018a-b, 2022).

If the population standard deviation σ is known, the *t*-interval and *z*-interval can be calculated at the same time, and the two types of uncertainty are consistent. However, Jenkins (2007) and Huang (2012) found that the *t*-interval uncertainty has a larger bias and precision error *vs* the *z*-interval uncertainty. This incompatibility of Type A assessment with Type B assessment is called the uncertainty paradox.

In addition to the uncertainty paradox, traditional methods produce a paradox (Du and Yang, 2000), known as the Du-Yang paradox, when determining the minimum sample size required to estimate the population mean with the maximum allowable error.

The *t*-interval uncertainty also leads to Ballico's paradox, that is, the estimation of the high-precision range, based on the uncertainty of the *t*-interval, is greater than the uncertainty of the low-precision range (Huang, 2016). D'Agostini (1998) gave an example of questioning the uncertainty of the *t*-interval: two measurements were made and the difference between the measurements was found to be 0.3 mm, but in order to have 99.9% confidence in the result, the *t*-interval should be 9.5 cm wide, which is a ridiculous result.

Huang pointed out that for the ultra-small sample size (such as $n<10$), the *t*-interval uncertainty problem is caused by the *t*-transform distortion (Huang, 2018b, 2021a, 2022).

The *t*-interval uncertainty problem is rarely questioned because it only becomes apparent when the sample size is small. In addition, the small sample *t*-interval theory is the mainstream paradigm of statistics. According to Thomas Kuhn, the habitual recognition of the mainstream paradigm leads to blindness to the problem (Kuhn, 1962).

In view of the problems of *t*-interval uncertainty, the international journal, *Basic and Applied Social Psychology*, officially banned confidence intervals in 2015 (Trafimow and Marks, 2015), the ban was for *t*-intervals or sample-based intervals, not for *z*-intervals or population-based intervals.

In summary, *t*-tests are only used for normal population and small sample problems, but the sample size should not be too small; *t*-test has the general defects and problems of statistical significance test (Zhang, 2022b), along with the inherent fallacies of confidence intervals (Zhang, 2022a), and the peculiar uncertainty problem of *t*-intervals makes the *t*-test theory obviously defective. Therefore, the *t*-test theory is faced with the choice of retaining or discarding in statistics.

3 Solutions

3.1 Solutions for *t*-value based statistical significance test

The *t*-test procedures of the aforementioned (1)-(4) can be attributed to the *p*-value based statistical significance tests, and the solutions for the problems is detailed in my paper "*p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond" (Zhang, 2022b). These solutions include using Bayesian methods, performing meta-analyses, using effect sizes, stressing statistical validity, using nonparametric statistics, using good experimental and sampling designs and determining appropriate sample size, the network methods are used instead of the reductionist method to obtain and analyze the data, and the statistical conclusions are combined with the mechanism analysis to draw scientific inferences, etc. (Zhang, 2022b).

3.2 Using the Bayesian credible interval method

As an alternative to *t*-intervals, the Bayesian credible interval method can be used (Morey et al., 2016; Pandey et al., 2022; Zhang, 2022a). However, Bayesian methods do not always yield the most credible results. When dealing with informative priors, we should check the effective sample size (Morita et al., 2012). There is a problem if the prior effective sample size is not small compared to the study sample size. Essentially, inferences are driven by preconceptions rather than data. There may be more opportunities to falsify results from Bayesian analysis than from frequentist analysis, in part because people are less familiar with the technical details of Bayesian methods (Matloff, 2014). A recent guideline for Bayesian analysis is expected to be valuable to people (Kruschke, 2021).

3.3 Using the Bootstrap credible interval method

Using resampling (Coleman et al., 1982; Solow, 1993; Zhang and Schoenly, 1999a-b; Gentle, 2002; Manly, 2007; Zhang, 2011a-b, 2021a-c, 2022a-b; Chihara and Hesterberg, 2018; Lock et al., 2021) rather than the classical *t*-test and *z*-test for inference is an important research area. As an alternative to *t*-intervals, the Bootstrap credible interval method, proposed by me (Zhang, 2022a), can be used.

3.4 Abandon the *t*-test and infer directly based on the Central Limit Theorem

Matloff advocates abandoning the *t*-test and making inference directly based on the Central Limit Theorem, and the same is true for regression (Matloff, 2011, 2014; Huang, 2022). The Central Limit Theorem (CLT) is a set of theorems in probability theory about the partial sum of a series and the distribution of a random variable approximating the normal distribution. The Central Limit Theorem proves mathematically that, under appropriate conditions, the mean of a large number of mutually independent random variables converge to the standard normal distribution according to the distribution after proper standardization. In nature, some phenomena are affected by many independent random factors. If the effect of each factor is very small, the total effect can be regarded as obeying a normal distribution. This set of theorems is the theoretical basis of mathematical statistics and error analysis, and they point out the conditions under which the sum of a large number of random variables is approximately normally distributed (Kallenberg, 2002).

The de Moivre–Laplace Central Limit Theorem is the original version of the Central Limit Theorem. The de Moivre-Laplace theorem states that the limit of the binomial distribution is the normal distribution. The Lindeberg-Levy Central Limit Theorem, an extension of the de Moivre-Laplace theorem, is the central limit theorem for a series of independent and identically distributed random variables. This theorem states that the normalized sum of a series of random variables that are independent and identically distributed and have limited mathematical mean and variance approximates the standard normal distribution (Fig. 3). The Lindeberg-Feller Central Limit Theorem, an advanced form of Central Limit Theorem, is an extension of the Lindeberg-Levy theorem, which holds for the sum of independent but not necessarily identically distributed random variables. The theorem states that when certain conditions are met, the normalized sum of a series of

random variables that are independent but not necessarily identically distributed approximates the standard normal distribution (Kallenberg, 2002).

(1) Lindeberg-Levy Central Limit Theorem

Assuming that X_1, X_2, ...X_n, ..., are independent and identically distributed random variables, which have limited means and variances: $E(X_i)=\mu$, $D(X_i)=\sigma^2$ (i=1, 2, ...), then for any x, the distribution function

$$F_n(x)=P\{\frac{\sum_{i=1}^{n}\boldsymbol{X_i}-n\mu}{\sigma\sqrt{n}}\le x\}$$

satisfies

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\{\frac{\sum_{i=1}^{n}\boldsymbol{X_i}-n\mu}{\sigma\sqrt{n}}\le x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\phi(x)$$

The theorem states that when n is large, the random variable

$$Y_n=\frac{\sum_{i=1}^{n}\boldsymbol{X_i}-n\mu}{\sigma\sqrt{n}}$$

approximately follows the standard normal distribution, $N(0, 1)$。Thus, when n is large

$$\sum_{i=1}^{n}X_i=\sigma\sqrt{n}Y_n+n\mu$$

approximately follows the standard normal distribution, $N(n\mu, n\sigma^2)$。This theorem is the simplest and most commonly used form of the Central Limit Theorem. As long as n is large enough, the sum of independent and identically distributed random variables can be regarded as a normal variable.

(2) Lindeberg-Feller Central Limit Theorem

Assuming that X_1, X_2, ...X_n, ..., are independent but not but not necessarily identically distributed random variables, $E[X_i]=0$ and the variance is not limitless, and their partial sum is:

$$S_n=\sum_{i=1}^{n}X_i$$

Let

$$s_i^2=\mathrm{var}(X_i)$$

$$\sigma_n^2=\sum_{i=1}^{n}s_i^2=\mathrm{var}(S_n)$$

If for each ε>0, the series satisfies

$$\lim_{n\to\infty}\frac{1}{\sigma_n^2}\sum_{i=1}^{n}E[X_i^2;\{|X_i|>\varepsilon\sigma_n\}]=0$$

it is called to satisfy Lindeberg condition. The series that satisfies this condition approximates the standard normal distribution, i.e.

$$s_n/\sigma_n\xrightarrow{d}N(0,1)$$

This is also a necessary condition for approximating the normal distribution of the sum of independent random variables with limited variance and zero mean.

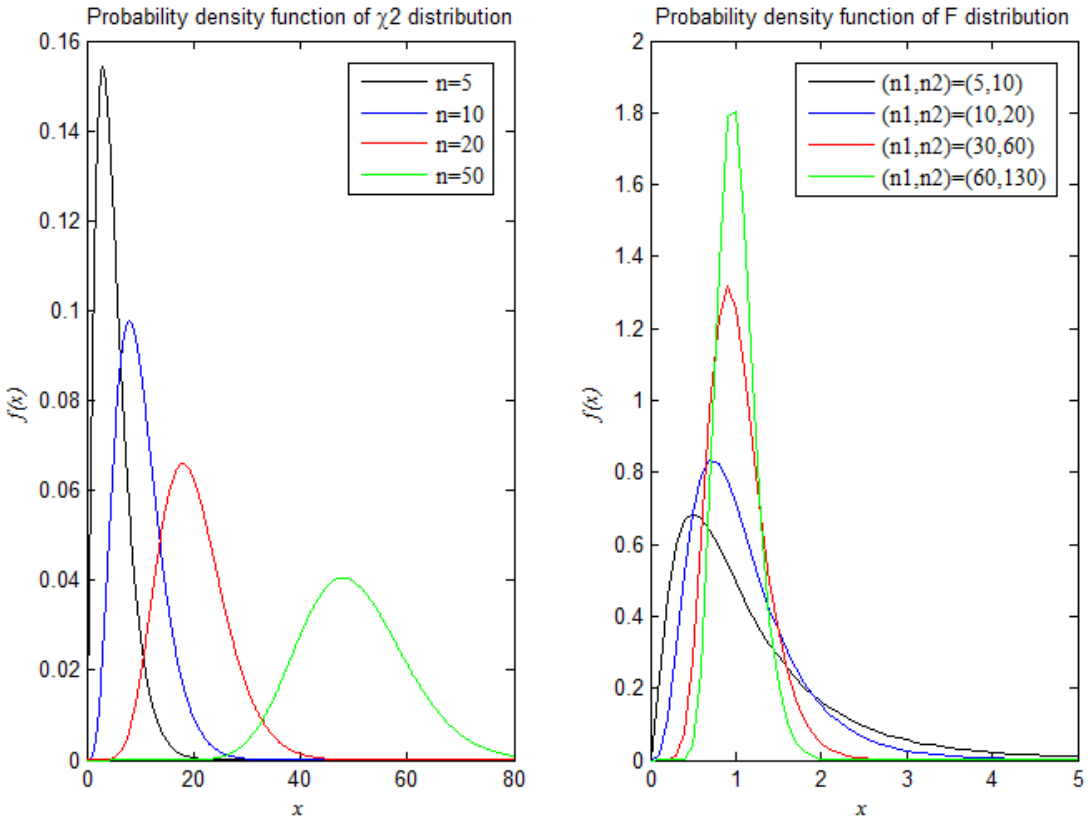

Fig. 3 $\lim_{n\to\infty}\chi^2 \to N(\mu,\sigma^2)$, $\lim_{n_1,n_2\to\infty}F \to N(\mu,\sigma^2)$.

3.5 Using the unified theory of uncertainty

In the 2021 International Standard ISO:24578:2021(E), unbiased estimates of expanded uncertainties are incorporated (ISO, 2021; Huang, 2022; Zhang, 2022a), where it defines the half-widths of the following 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 at confidence level p%, and c_4 is the bias correction factor for sample standard deviation:

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

e.g., c_4=0.7979, 0.9213, 0.9515, 0.9650, and 0.9727 correspond to n=2, 4, 6, 8, and 10, respectively.

Huang (2018a) proposed a unified theory of measurement error and uncertainty. 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" (JCGM, 2008a-b). These two new estimators significantly simplify uncertainty analysis, avoiding the difficulty and subjectivity of determining Type B uncertainty and the limitations of the Welch-Satterthwaite formula.

3.6 Other methods

In order to solve the uncertainty problem based on the *t*-value, Jenkins (2007) proposed the unbiased estimator of empirical mean based on the uncertainty of the *z*-value as an alternative to the former. Huang (2012) found through an Internet search that the unbiased estimator of theoretical mean is just the first term in a series proposed by Craig for estimating possible errors (Craig, 1927).

Matloff (2014) used Slutsky's theorem to prove the effectiveness of using s instead of σ to construct approximate confidence intervals in the *z*-interval, thus avoiding the t-interval problem.

In the Judd-Mcclelland-Culhane model comparison method (Judd et al., 1995), the regression model is the original model and all statistical problems are presented as comparisons between models. Within this framework, many statistical tests can be developed from the first principle.

According to Ilya Kipnis, it is not always necessary to assume the type of probability distribution of the data. If there is enough data, plot frequency plots, etc., run some quantile tests, e.g. 5th percentile, 2.5th percentile, etc., to decide which method to use (Matloff, 2014).

References

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

Chihara LM, Hesterberg TC. 2018. Mathematical Statistics with Resampling and R (2nd Edition). Wiley, USA. https://www.wiley.com/en-us/Mathematical+Statistics+with+Resampling+and+R%2C+2nd+Edition-p-9781119416531

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

Craig CC. 1927. Remarks on the probable error of a mean. The American Mathematical Monthly, 34: 472-476. https://doi.org/10.1080/00029890.1927.11986750

D'Agostini G. 1998. Jeffeys priors versus experienced physicist priors: arguments against objective Bayesian theory. Proceedings of the 6th Valencia International Meeting on Bayesian Statistics. Alcossebre, Spain. https://www.semanticscholar.org/paper/Jeffreys-priors-versus-experienced-physicist-priors-D'Agostini/e8f406828349ac6397ceac98c2d5f8a3318213c8

Du XL, Yang BF. 2000. A paradox in determining the minimum sample size. Beijing Statistics, 130: 36

Fisher RA. 1924. A method of scoring coincidences in tests with playing cards. Proceedings of the Society for Psychical Research, 34: 181-185. https://royalsocietypublishing.org/doi/pdf/10.1098/rsbm.1963.0006

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

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

Harrell F. 2022. Harrell Miscellaneous functions useful for data analysis. https://www.freshports.org/devel/R-cran-Hmisc/. Accessed 2022-6-10

Huang H. 2012. Comparison of uncertainty calculation models. Cal Lab Magazine – The International Journal of Metrology, 19: 24-29. https://www.callabmag.com/comparison-of-uncertainty-calculation-models/

Huang HN. 2018a. 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. 2018b. Uncertainty estimation with a small number of measurements, part I: new insights on the t-interval method and its limitations. Measurement Science and Technology, 29: 015004. https://doi.org/10.1088/1361-6501/aa96c7

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. A fallacy of confidence interval theory for measurement of uncertainty evaluation. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1340388.html. Accessed 2022-5-30

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

Jenkins JD. 2007. The Student's t-distribution uncovered. Proceedings of Measurement Science Conference. Long Beach, California, USA. https://msc-conf.com/about-us/

Joint Committee for Guides in Metrology (JCGM). 2008a. Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement (GUM 1995 with Minor Corrections) (Sevres). https://www.bipm.org/en/committees/jc/jcgm. Accessed 2022-6-15

Joint Committee for Guides in Metrology (JCGM). 2008b. JCGM 101: Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement'—Propagation of Distributions Using a Monte Carlo Method (Sevres). https://www.bipm.org/en/committees/jc/jcgm. Accessed 2022-6-15

Judd CM, McClelland GH, Culhane SE. 1995. Data analysis: Continuing issues in the everyday analysis of psychological data. Annual Review of Psychology, 46(1): 433-465. http://scholar.google.com/scholar?cluster=14481702516523332591

Kallenberg O. 2002. Foundations of Modern Probability (2ed). Springer. https://doi.org/10.1007/978-1-4757-4015-8

Kruschke JK. 2021. Bayesian analysis reporting guidelines. Nature Human Behaviour, 5: 1282-1291. https://www.nature.com/articles/s41562-021-01177-7

Kuhn T. 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, USA. https://archive.org/details/structureofscie00kuhn

Lock R, Lock PF, Morgan KL, Lock EF, Lock DF. 2021. Statistics: Unlocking the Power of Data. Wiley, USA. http://www.lock5stat.com/

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

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

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 2022-6-1

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

Morita S, Thall PF, Muller P. 2012. Prior effective sample size in conditionally independent hierarchical models. Bayesian Analysis, 7(3): 591-614. https://doi.org/10.1214/12-BA720

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

Student (William Sealy Gosset). 1908. The probable error of a mean. Biometrika, 6(1): 1-25. http://seismo.berkeley.edu/~kirchner/eps_120/Odds_n_ends/Students_original_paper.pdf

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

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

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. 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, 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. 2022a. Confidence intervals: Concepts, fallacies, criticisms, solutions and beyond. Network Biology, 12(3): 97-115. http://www.iaees.org/publications/journals/nb/articles/2022-12(3)/confidence-intervals-fallacies-criticisms-solutions.pdf

Zhang WJ. 2022b. *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, 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