Qeios

Statistics Reform: Practitioner's Perspective

Hening Huang
Teledyne RD Instruments (retired)
San Diego, CA 92127, USA

Abstract It is generally accepted that one of the main causes of the replication crisis in scientific research is some of the most commonly used statistical methods, such as null hypothesis significance testing (NHST). This has prompted many scientists and statisticians to call for statistics reform. As a practitioner in the fields of hydraulics and measurement science, the author used statistical methods extensively in many environmental engineering and hydrological survey projects. The author strongly concurs in the need for statistics reform. This paper offers a practitioner's perspective on statistics reform. In the author's view, some statistical methods are good and should withstand statistics reform; some are bad and should be abandoned and removed from statistics textbooks and computer software packages. This paper focuses on two statistical methods derived from the *t*-distribution: the two-sample *t*-test and the *t*-interval method for measurement uncertainty calculation. We show why these two methods should be abandoned. We recommend using descriptive statistic analysis as an alternative to the two-sample *t*-test and an unbiased estimation method as an alternative to the *t*-interval method for measurement uncertainty calculation. Two examples are provided to demonstrate the recommended alternatives.

Keywords Effect size, point estimation, statistical methods, statistics reform, *t*-intervals, *t*-tests, measurement uncertainty

1. Introduction

In recent years, the scientific community has become increasingly concerned about the replication crisis. It is generally accepted that one of main causes of the replication crisis is some of the most commonly used statistical methods. Specifically, the suite of null hypothesis significance testing (NHST) and its associated *p*-values, and claims of statistical significance, have come in most to blame (Nuzzo 2014). Siegfried (2010) wrote, "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." Siegfried (2014) stated, "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies." Therefore, many authors suggested retiring or abandoning statistical significance and *p*-values (e.g. Amrhein et al. 2019, McShane et al. 2018, Halsey 2019, Wasserstein and Lazar 2016, Wasserstein et al. 2019). *Basic and Applied Social Psychology* has officially banned the NHST procedures since 2015 (Trafimow and Marks 2015). Furthermore, many scientists and statisticians call for statistics reform (e.g. Wagenmakers et al. 2011, Haig 2016, Colling and Szűcs 2021). Cumming (2014) proposed 'New Statistics' as a form of statistics reform. The 'New Statistics' he proposed mainly includes (1) abandoning the NHST procedures and (2) using the estimation of effect sizes and confidence intervals. Normile et al. (2019) introduced the new statistics in the classroom. Claridge-Chang and Assam (2016) suggested replacing significance testing with estimation statistics. Very recently, a co-published editorial of 14 physiotherapy journals (Elkins et al. 2022) "… advises researchers that some physiotherapy journals that are members of the International Society of Physiotherapy Journal Editors (ISPJE) will be expecting manuscripts to use estimation methods instead of null hypothesis statistical tests." Although some authors still defend NHST

https://doi.org/10.32388/JVYEJZ

and *p*-values (e.g. Benjamini et al. 2021, Hand 2022, Lohse 2022), "A paradigm shift away from null hypothesis significance testing seems in progress (Berner and Amrhein 2022)."

As a practitioner in the fields of hydraulics and measurement science, the author has used statistical methods extensively in many environmental engineering and hydrological survey projects (citation omitted). Especially, the author processed thousands of small samples collected in streamflow measurements using acoustic Doppler current profilers (ADCPs). An ADCP streamflow measurement usually involves a few of observations (typically 4). According to statistics textbooks and *Guide to the Expression of Uncertainty in Measurement* (GUM) (JCGM 2008), the uncertainty of the sample mean of a small sample should be calculated using the *t*-interval method. However, the author (Huang 2010) was puzzled by the fact that the *t*-based uncertainty (i.e. the half-width of the *t*-interval) was unrealistic and misleading, which leads to the so-called "uncertainty paradox" (Huang 2010, 2018a) and a high false rejection rate in the quality control of ADCP streamflow measurements (Huang 2014).

The author is not the only one to question the *t*-interval method for measurement uncertainty calculation. Jenkins (2007) also discovered that the *t*-based uncertainty has large bias and precision errors. D'Agostini (1998) showed an example: "…having measuring the size of this page twice and having found a difference of 0.3 mm between the measurements… Any rational person will refuse to state that, in order to be 99.9% confidence in the result, the uncertainty interval should be 9.5 cm wide (any carpenter would laugh…). This may be the reason why, as far as I known, physicists don't use the Student distribution." Moreover, Ballico (2000) reported an important counterinstance of the *t*-interval method for uncertainty calculation during a routine calibration at the CSIRO National Measurement Laboratory (NML), Australia. A thermometer was calibrated for 1 mK range (higher precision) and 10 mK range (lower precision) and the uncertainty was estimated using the WS-*t* approach (WS standards for Welch-Satterthwaite formula, *t* stands for the *t*-interval). It was intuitive and expected that the thermometer in the higher precision range (1 mK) should have a lower uncertainty than the thermometer in the lower precision range (10 mK). However, the WS-*t* approach gave a counter-intuitive result: the estimated uncertainty for the 1 mK range was 37.39, which was *greater* than 35.07, the estimated uncertainty for the 10 mK range! This counter-intuitive result was later referred to as the Ballico paradox (Huang 2016).

Practitioners in science and industry rely on statistical methods in their work. Bad statistical methods can harm their work. Ziliak and McCloskey (2008) showed in their 322-page book that "*Statistical significance is an exceptionally damaging one."* However, the author's impression over the years is that, practitioners are often accused of misunderstanding and misusing certain statistical methods or concepts, especially *p*-values. This is not fair! In the author's opinion, if a statistical method or concept is easily and often misunderstood or misused, and even our schools fail to teach it properly, then there must be something wrong with that method or concept. Trafimow (2023) stated, "NHST is problematic anyway even without misuse." And "There is practically no way to use them [*p*-values] properly in a way that furthers scientific practice." While the debate about the NHST procedures continues, one thing is certain: the NHST procedures (including the *t*-test) have failed the test of time after about 100 years.

We understand that no statistical method is perfect, nor can any statistical method be used without limitations or conditions. However, the author's own experience and that of some other practitioners have shown that some statistical methods are good and useful, such as least squares method and point estimation, but some statistical methods are bad, such as the *t*-interval method for measurement uncertainty calculation. The coexistence of good and bad statistical methods can

confuse practitioners. In particular, many (if not most) practitioners may not be able to tell that certain statistical methods are bad, flawed, or controversial, and continue to use them. Therefore, the author strongly concurs in the need for statistics reform. This paper offers a practitioner's perspective about statistics reform. We argue that good methods should withstand statistics reform and bad methods should be abandoned and removed from statistics textbook and computer software packages. This paper focuses on two statistical methods derived from the *t*-distribution: the two-sample *t*-test and the *t*-interval method for measurement uncertainty calculation. We will show why these two methods should be abandoned and what the alternatives are.

In the following sections, Section 2 briefly reviews some examples of good statistical methods that should withstand statistics reform. Section 3 discusses why the two-sample *t*-test should be abandoned. Section 4 describes an alternative method to the two-sample *t*-test. Section 5 discusses why the *t*-interval method for measurement uncertainty calculation should be abandoned. Section 6 describes an alternative method to the *t*-interval method for measurement uncertainty calculation. Section 7 provides conclusion and recommendation.

2. Examples of good statistical methods that should withstand statistics reform

In the author's opinion, a good statistical method should have the following characteristics: (a) to have clear mathematical meaning and can be easily understood even by those without advanced statistics training; (b) to give realistic results in real-world applications; and (c) to be less controversial in the scientific community. Ideally, a good statistical method might be related to a physical principle or has physical meaning. There are many good statistical methods that are used frequently in practice, four examples of which are listed below.

- Method of least squares
- Method of maximum likelihood
- Central Limit Theorem
- Akaike information criterion (AIC)

Perhaps, the method of least squares is one of the most commonly used statistical methods in practice; it is hardly controversial in the scientific community. Importantly, the method of least squares conforms to the principle of minimum energy, one of the fundamental principals in physics. The sum of squared errors can be regarded to represent the internal energy of the system under consideration, which must approach the minimum value at equilibrium.

The method of maximum likelihood is another most commonly used statistical methods. It is also hardly controversial in the scientific community. The method of maximum likelihood is intuitive: "The likelihood supplies a natural order of preference among the possibilities under consideration (Fisher 1956)." Accordingly, the mode of a likelihood function corresponds to the mostly preferred parameter value given the data (Huang 2022). This idea does not require advanced knowledge of statistics to understand. In addition, the method of maximum likelihood is essentially consistent with the method of least squares.

The Central Limit Theorem states that, given a sufficiently large sample size, the sampling distribution of the sample mean will approximately be a normal distribution, regardless of the original distribution. Since error is the difference between the true value and the measured value (e.g. the sample mean), the Central Limit Theorem confroms to the law of error, which is one of the most important laws in statistics and measurement science.

The Akaike information criterion (AIC) builds on the concept of entropy in information theory. A minimum AIC model is the model that minimizes the information loss among a set of candidate models. The AIC is essentially consistent with the method of maximum likelihood or the method of least squares.

Of course, good statistical methods such as the four motioned above should withstand statistics reform.

3. Why should the two-sample *t*-test be abandoned?

Perhaps, the two-sample *t*-test is the most commonly used procedure among the NHST procedures. Therefore, if we abandon the NHST procedures, the two-sample *t*-test should be abandoned first. However, the reasons for abandoning the two-sample *t*-test do not appear to be explicitly discussed in the literature, other than some general discussions and debates about the problems of the NHST procedures and associated *p*-values. It is important to note that *p*-values are output of statistical methods, such as the two-sample *t*-test. Therefore, the problem with *p*-values is not just about *p*-values. The problem of *p*-values should be tracked back to the statistical method that generated the *p*-values. In this section, we address two main issues with the two-sample *t*-test: rationale and performance. We argue that these two issues should explain why the two-sample *t*-test should be abandoned.

3.1 Rationale issue: the two-sample t-test is philosophically misleading

The two-sample *t*-test is philosophically misleading. Suppose we have two datasets (groups): A and B. Group A is the result (data) from treatment A and Group B is the result (data) from treatment B. We are interested in whether treatment A is superior to treatment B, or vice versa. In the standard NHST setting for a two-sample *t*-test, we assume a null hypothesis (a "strawman"): the unknown population means of the two groups are the same, and also assume an alternative hypothesis: the two means are different. Then, we use the two-sample *t*-test generates a *p*-value. If $p<0.05$, we infer that the deference between the two group means is "statistically significant", i.e. the "strawman" is disproven. Clearly, the two-sample *t*-test does not answer our question about "whether treatment A is superior to treatment B, or vice versa?" Instead, it misleads us into considering whether treatment A is different from treatment B based on the "statistical significance" quantified by an arbitrary threshed *p*-value (e.g. 0.05). Therefore, the rationale behind the two-sample *t*-test is wrong. In practice, we know that treatment A is different from treatment B just by looking at the data or two group means. Therefore, there is no need to assume the null and alternative hypotheses. In other words, we do not need a "strawman" (the null hypothesis) and then try to disprove it; we can directly assess the practical significance of the difference between the two groups based on our domain knowledge. We can further perform a probabilistic analysis to determine the probability that treatment A is superior to treatment B (or vice versa).

3.2 Performance issues: uncertainty, inconsistency, and dependence on sample size

Even if we accept its rationale and use it for comparing the means of two groups, the two-sample *t*-test cannot provide reliable results. This can be understood by looking at the behaviors of the *p*-

value generated from the two-sample *t*-test. First, like any sample statistics, the *p*-value has uncertainty. Halsey et al. (2015) discussed the uncertainty associated with the *p*-value of two-sample *t*-tests through simulations. Their simulation results showed that "a major cause of the lack of repeatability is the wide sample-to-sample variability in the *P* value." They stated, "As we have demonstrated, however, unless statistical power is very high (and much higher than in most experiments), the *P* value should be interpreted tentatively at best. Data analysis and interpretation must incorporate the uncertainty embedded in a *P* value." Lazzeroni et al. (2016) introduced *p*-value confidence intervals for the "true population *P* value" or π value, which they defined as the value of *P* when parameter estimates equal their unknown population values. They stated, "*P* values are variable, but this variability reflects the real uncertainty inherent in statistical results."

Second, the two-sample *t*-test may give inconsistent results for essentially the same evidence. Bonovas and Daniele (2023) discussed the inconsistency of the two-sample *t*-tests in two trials of a new drug. A *p*-value of 0.11 was obtained in a single-center, randomized, double-blind, placebo-controlled trial, indicating "no difference" between the active drug and placebo, whereas a *p*-value of 0.001 was obtained in a multi-center trial, indicating "significant difference" between the active drug and placebo. However, the risk ratio was the same in both trials: 0.70, indicating that the efficacy of the experimental drug was the same in both trials.

Third, the *p*-value generated from a two-sample *t*-test depends on the sample size; it decreases with increasing sample size. Therefore, *p*-values can be easily "hacked" through "*N*-chasing" (named by Stansbury 2020), which can guarantee the "statistical significance" at any pre-specified threshold, even if the effect size (e.g. the difference between the means of two groups) is very small and has no practical meaning. Therefore, "*N*-chasing" is the most effective way of *p*-hacking, leading to false research findings. In the author's opinion, the only solution to "*N*-chasing" or *p*-hacking is to abandon the two-sample *t*-test.

4. Alternative to the two-sample *t*-test: descriptive statistic analysis

We recommend using descriptive statistic analysis as an alternative to the two-sample *t*-test. The descriptive statistic analysis focuses on the comprehensive presentation of a set of descriptive statistics including: effect size (ES), relative effect size (RES), standard uncertainty (SU), relative standard uncertainty (RSU), signal-to-noise ratio (SNR), signal content index (SCI), exceedance probability (EP), and net superiority probability (NSP). Each of these eight statistics has a clear mathematical or physical meaning and is easy to understand. The superiority of treatment A over treatment B (or vice versa) is measured by ES (or RES) and EP (or NSP). The reliability of the estimated ES (or RES) is measured by SU, RSU, SNR, and SCI. It is important o note that, we do not specify a threshold for any of these descriptive statistics. Whether the estimated effect size is of practical importance should be judged with our domain knowledge with the consideration of these eight descriptive statistics. In addition, the descriptive statistic analysis does not use any of the terminology and language used in the NHST paradigm. Terms like null-hypothesis, alternative hypothesis, p-value, statistical significance, and statistical power are gone.

4.1 Effect size (ES) and relative effect size (RES)

Effect size (ES), denoted by $|\Delta|$, is the absolute deference between the two group means (denoted by $\bar{x}_A$ and $\bar{x}_B$, respectively), written as

$$|\Delta| = \bar{x}_A - \bar{x}_B. \tag{1}$$

It is important to note that the ES $|\Delta|$ is the so-called "simple" effect size. It is the raw difference between the means of two groups, expressed in the original (physical) unit of the quantity of interest; it is not standardized like Cohen's *d*. Because the simple effect size has the original (physical) unit, it will nearly always be more meaningful than standardized effect size (Baguley 2009). Schäfer (2023) argued that in their unstandardized form, effect sizes are easy to calculate and to interpret. Standardized effect sizes, on the other hand, bear a high risk for misinterpretation. In real-world applications, our domain knowledge about a quantity of interest is related to the physical unit of that quantity. Therefore, it is easier for practitioners to assess the practical significance of effects using the original (physical) unit than the dimensionless unit of standardized effect sizes. Baguley (2009) discussed the advantages of using simple effect sizes over standardized effect sizes. He stated, "For most purposes simple (unstandardized) effect size is more robust and versatile than standardized effect size." Therefore, we do not recommend using any standardized effect size like Cohen's *d*.

Note also that $|\Delta|$ is the absolute magnitude of the effect. In practice, we are often interested in the relative magnitude of the effect, i.e. relative effect size (RES), defined as (Huang 2022)

$$\text{RES} = \frac{|\bar{x}_A - \bar{x}_B|}{\bar{x}_w}, \tag{2}$$

where $\bar{x}_w$ may be calculated as the inverse-variance weighted-average

$$\bar{x}_w = \frac{\dfrac{\bar{x}_A}{\text{Var}(\bar{x}_A)} + \dfrac{\bar{x}_{B,D}}{\text{Var}(\bar{x}_B)}}{\dfrac{1}{\text{Var}(\bar{x}_A)} + \dfrac{1}{\text{Var}(\bar{x}_B)}}, \tag{3}$$

where $\text{Var}(\bar{x}_A) = s_A^2/n_A$ and $\text{Var}(\bar{x}_B) = s_B^2/n_B$; s_A and s_B are the sample standard deviation of Group A and Group B respectively; n_A and n_B are the sample size of Group A and Group B respectively. RES is usually expressed as a percentage.

Also note that the ES $|\Delta|$ (or RES) is not a function of sample size. As such, it only emphasizes the (treatment) effect rather than confounding the effect with sample size like the two-sample *t*-test does. Increasing sample size has essentially no influence on the ES $|\Delta|$ (or RES), but its reliability increases. Because of this, unlike the *p*-value of *t*-tests, which can be easily hacked through "*N*-chasing", the ES $|\Delta|$ (or RES) cannot be hacked through "*N*-chasing".

4.2 Standard uncertainty (SU), relative standard uncertainty (RSU), signal-to-noise ratio (SNR), and signal content index (SCI)

The ES $|\Delta|$ is a point estimate. Its reliability must be quantified and assessed. The descriptive statistics SU, RSU, SNR, and SCI can be used to quantify and assess the reliability of the ES $|\Delta|$.

Let $u(\Delta)$ denote the SU associated with the ES $|\Delta|$. It is defined as the standard deviation of $\Delta = \bar{x}_A - \bar{x}_B$, written as

$$u(\Delta) = \sqrt{\mathrm{Var}(\Delta)} = \sqrt{\mathrm{Var}(\bar{x}_A) + \mathrm{Var}(\bar{x}_B)}\,. \tag{4}$$

In measurement science, SU is a measure of the precision of a measurement. If we regard the ES $|\Delta|$ as a measurement result, $u(\Delta)$ measures the precision of the estimated ES $|\Delta|$. Note that $u(\Delta)$ has the same physical unit as $|\Delta|$.

In practice, we are also interested (if applicable) in the relative standard uncertainty (RSU), defined as

$$\mathrm{RSU} = \frac{u(\Delta)}{|\Delta|}\,. \tag{5}$$

The signal-to-noise ratio (SNR) is defined as the ratio between signal energy and noise energy. It is commonly quoted for electrical signals, but can be applicable to any form of signal (Huang 2019a). For comparing the means of two groups, the ES $|\Delta|$ is the signal and the associated SU $u(\Delta)$ is the noise. Therefore, the SNR can be calculated as

$$\mathrm{SNR} = \frac{E_{\mathrm{signal}}}{E_{\mathrm{noise}}} = \frac{\Delta^2}{u^2(\Delta)} = \frac{(\bar{x}_A - \bar{x}_B)^2}{\mathrm{Var}(\bar{x}_A) + \mathrm{Var}(\bar{x}_B)}\,. \tag{6}$$

The SNR is related to a statistic called signal content index (SCI) (Huang 2019a). For comparing the means of two groups, the SCI is calculated as

$$\mathrm{SCI} = \frac{E_{\mathrm{signal}}}{E_{\mathrm{signal}} + E_{\mathrm{noise}}} = \frac{\Delta^2}{\Delta^2 + u^2(\Delta)} = \frac{\mathrm{SNR}}{1 + \mathrm{SNR}}\,. \tag{7}$$

The SCI has a clear physical meaning; it is the relative amount of signal energy contained in the measurement result (Huang 2019a).

Either the SNR or SCI can be used to measure the reliability of the estimated ES $|\Delta|$. However, since the SCI ranges between 0 and 1, its interpretation is more intuitive than the SNR. A high SCI value (e.g. close to 1) indicates that the estimated ES $|\Delta|$ is highly reliable; a low SCI value (e.g. close to 0) indicates that the estimated ES $|\Delta|$ is unreliable due to noise.

It should be noted that, unlike the ES $|\Delta|$ (or RES), which is independent of sample size, the SU $u(\Delta)$, RSU, SNR, or SCI is a function of sample size. The larger the samples, the smaller the SU and RSU, and the larger the SNR and SCI. Thus, there is a clear distinction between the ES $|\Delta|$ and its reliability measure $u(\Delta)$, RSU, SNR, or SCI. In practice, both the ES $|\Delta|$ and its reliability measures should be interpreted and assessed based on our domain knowledge.

It should also be noted that we do not use confidence interval to quantify the uncertainty (or precision) of the ES $|\Delta|$. This is because the concept of confidence interval is controversial and has been debated in the scientific community for decades (e.g. Karlen 2002, Etz 2015, Morey et al. 2016a,b, Trafimow 2018). In particular, the *t*-interval, which is a commonly used confidence interval for small samples, is problematic and should be abandoned as discussed in Section 5.

4.3 Exceedance probability (EP) and net superiority probability (NSP)

The ES $|\Delta|$ measures the true significance of the difference (on average) between two treatments A and B. In other words, it measures the superiority of treatment A over treatment B (assuming $\bar{x}_A - \bar{x}_B > 0$) at the average effect level. In practice, we are also interested in the superiority at the elemental effect level. That is, we want to compare the elements (scores) in the two groups and see how often is that the elements (scores) in Group A are superior to the elements (scores) in Group B (or vice versa).

The probability that Group A is superior to Group B at the elemental effect level is called exceedance probability (EP) defined as (Huang 2022)

$$\mathrm{EP}_{X_A \geq X_B} = \Pr(X_A \geq X_B) = \int_0^\infty p(y)dy, \tag{8}$$

where $p(y)$ is the probability density function for the quantity $Y = X_A - X_B$, X_A is the random variable associated with Group A, and X_B is the random variable associated with Group B.

The meaning of $\mathrm{EP}_{X_A \geq X_B}$ is essentially the same as that of the following statistics: CLES (common language effect size) (McGraw and Wong 1992), PS (probability of superiority) (Vargha and Delaney 2000, Grissom and Kim 2001), AUC (area under the receiver operating characteristic) or A for its nonparametric version (Delaney and Vargha 2002, Ruscio and Mullen 2012). It should be pointed out that the calculation of CLES requires the standard parametric assumptions of population normality and equal variances, while the calculation of $\mathrm{EP}_{X_A \geq X_B}$ does not require normality and homoscedasticity assumptions. In this regard, CLES is an approximation of $\mathrm{EP}_{X_A \geq X_B}$. In addition, CLES should not confuse practitioners because of its name; CLES is a probability, not an effect size.

Assume that both X_A and X_B are normally distributed with unknown mean and variance. The estimated distribution of X_A is $N(\bar{x}_A, \frac{s_A}{c_{4,nA}})$ and the estimated distribution of X_B is $N(\bar{x}_B, \frac{s_B}{c_{4,nB}})$, $c_{4,n}$ is the bias correction factor, $c_{4,n} = \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$, and $\Gamma(.)$ stands for Gamma function (Wadsworth 1989). In addition, the estimated distribution of $Y = X_A - X_B$ is

$$Y \sim N\left[(\bar{x}_A - \bar{x}_B), \sqrt{\mathrm{Var}(X_A) + \mathrm{Var}(X_B)}\right]. \tag{9}$$

Then, $\mathrm{EP}_{X_A \geq X_B}$ can be calculated as (Huang 2023a)

$$\mathrm{EP}_{X_A \geq X_B} = \Pr(Z \geq -e') = 1 - \Phi(-e') = \Phi(e'), \tag{10}$$

where *e'* is calculated as

$$e' = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{s_A}{c_{4,nA}}\right)^2 + \left(\frac{s_B}{c_{4,nB}}\right)^2}}. \tag{11}$$

On the other hand, the EP for $X_B \geq X_A$ can be calculated as

$$\mathrm{EP}_{X_B \geq X_A} = \Pr(Z \geq e') = 1 - \Phi(e') = \Phi(-e'). \tag{12}$$

Furthermore, the net superiority probability (NSP), denoted by ξ, is related to the exceedance probabilities as (Huang 2023a)

$$\xi = \mathrm{EP}_{X_A \geq X_B} - \mathrm{EP}_{X_B \geq X_A}. \tag{13}$$

Although Eq. (13) is based on the normality assumption, it is considered as a general definition of the NSP for any distribution of X_A and X_B (Huang 2023a).

It is important to note that the EP or NSP is only a very weak function of sample sizes due to the bias correction factor $c_{4,n}$. Therefore, like the ES $|\Delta|$, the EP or NSP cannot be hacked through *N*-chasing.

For nonparametric comparison of two groups, the exceedance probability $\mathrm{EP}_{A \geq B}$ is calculated as (Huang 2023a)

$$\mathrm{EP}_{A \geq B} = \frac{U_{A \geq B}}{n_A n_B}, \tag{14}$$

and the exceedance probability $\mathrm{EP}_{B \geq A}$ is

$$\mathrm{EP}_{B \geq A} = \frac{U_{B \geq A}}{n_A n_B}, \tag{15}$$

where $U_{A \geq B}$ or $U_{B \geq A}$ is the *U* statistic in the Mann–Whitney *U* test.

Accordingly, the NSP of Group A over Group B is

$$\xi = \mathrm{EP}_{A \geq B} - \mathrm{EP}_{B \geq A} = \frac{U_{A \geq B} - U_{B \geq A}}{n_A n_B}. \tag{16}$$

It is worth mentioning that the concept of exceedance probability (EP) and its analysis have been used in many engineering fields. For example, U.S. EPA (Environment protection agency) (1991) established a probabilistic chronic toxics standard: EP=0.0037 to protect aquatic life. Di Toro (1984) performed an exceedance probability analysis of river quality due to runoff. Huang and Fergen (1995) performed an exceedance probability analysis of river BOD (biochemical oxygen demand) and DO (dissolved oxygen) concentration due to point load. Krishnamoorthy et al. (2007) used exceedance probability analysis to assesse the exposure level in work environments. In addition, the term "return period" commonly used in hydraulic engineering and hydrology can be converted into exceedance probability. For example, a 100-year flood is equivalent to EP=1%. Thus, practitioners in engineering fields are more familiar with the term EP than with the term CLES, AUC, or A.

4.4 Example: comparison of old and new flavorings for a beverage

Zaiontz (2020) considered the following problem (example). A marketing research firm conducted experiments on the effectiveness of a new flavoring for a beverage. Eleven people in Group A1 and ten people in Group A2 tasted the beverage with the new flavoring and ten people in Group B tasted the beverage with the old favoring. The people then took a questionnaire to evaluate how enjoyable the beverage was. The scores for the new flavoring (Group A1 and Group A2) and old flavoring (Group B) are shown in Table 1. The sample mean and standard deviation for each group

are shown in Table 2.

Table 1 Scores of the three groups in the beverage flavor taste experiments

New flavoring (Group A1)	New flavoring (Group A2)	Old flavoring (Group B)
13	20	12
17	32	8
19	2	6
10	25	16
20	5	12
15	18	14
18	21	10
9	7	18
12	28	4
15	40	11
16		

Table 2 Sample means and standard deviations of the three groups in the beverage flavor taste experiments

	New flavoring (Group A1)	New flavoring (Group A2)	Old flavoring (Group B)
Sample mean	14.91	19.80	11.10
Sample standard deviation	12.89	12.27	4.33

Zaiontz (2020) performed the two-sample *t*-test (two-tailed) to determine whether there was a significant difference between the two flavorings. He obtained a *p*-value of 0.04 in the two-sample *t*-test for Group A1 versus Group B. He then rejected the null hypothesis at α=0.05, and concluded that there was a significant difference between the two flavorings. That is, the new flavoring was significantly more enjoyable. On the other hand, Zaiontz (2020) obtained a *p*-value of 0.05773 in the two-sample *t*-test for Group A2 versus Group B. He then stated that he could not reject the null hypothesis at α=0.05. That is, there was no significant difference between the two flavorings. It is strange that Zaiontz (2020) did not comment the contradictory results given by the two *t*-tests.

We examined this example using the descriptive statistic analysis method. Table 3 shows the values of six descriptive statistics. Table 4 shows the results of the probabilistic analysis based on the distribution-based comparison and Table 5 shows the results based on the nonparametric comparison.

Table 3 Values of six descriptive statistics for the comparison of beverage flavoring

Statistic	Comparison between Group A1 and Group B	Comparison between Group A2 and Group B
Simple effect size (ES): Eq. (1)	$\lvert\Delta\rvert = 3.81$	$\lvert\Delta\rvert = 8.70$
Relative effect size (RES): Eq. (2)	28.52%	72.12%
Standard uncertainty (SU): Eq. (4)	1.75	4.12
Relative standard uncertainty (RSU): Eq. (5)	45.84%	47.31%

Signal-to-noise ratio (SNR): Eq. (6)	4.76	4.47
Signal content index (SCI): Eq. (7)	0.83	0.82

Table 4 Results of the probabilistic analysis based on the distribution-based comparison

	Comparison between Groups A1 and B	Comparison between Groups A2 and B
Estimated distribution of *Y*: Eq. (9)	$Y \sim N(3.81, 5.78)$	$Y \sim N(8.70, 13.38)$
Exceedance probability (EP) (A ≥ B): Eq. (10)	$\mathrm{EP}_{X_A \geq X_B} = 0.745$	$\mathrm{EP}_{X_A \geq X_B} = 0.742$
Exceedance probability (EP) (B ≥ A): Eq. (12)	$\mathrm{EP}_{X_B \geq X_A} = 0.255$	$\mathrm{EP}_{X_B \geq X_A} = 0.258$
Net superiority probability (NSP): Eq. (13)	$\xi = 0.490$	$\xi = 0.484$

Table 5 Results of the probabilistic analysis based on the nonparametric comparison

	Comparison between Group A1 and Group B	Comparison between Group A2 and Group B
Exceedance probability (EP) (A ≥ B): Eq. (14)	$\mathrm{EP}_{A \geq B} = 0.741$	$\mathrm{EP}_{A \geq B} = 0.725$
Exceedance probability (EP) (B ≥ A): Eq. (15)	$\mathrm{EP}_{B \geq A} = 0.259$	$\mathrm{EP}_{B \geq A} = 0.275$
Net superiority probability (NSP): Eq. (16)	$\xi = 0.482$	$\xi = 0.450$

As can be seen from Table 3, the ES is 3.81 and the RES is 28.52% for the comparison of Group A1 versus Group B, while the ES is 8.70 and the RES is 72.12% for the comparison of Group A2 versus Group B. Our domain knowledge (common sense in this case) tells us the difference between the two flavorings is practically significant. Note that due to the small sample sizes, the RSUs are large: 45.84% and 47.31%. However, the SNRs are large: 4.76 and 4.47, and the SCIs are also large: 0.83 and 0.82, indicating that the effect size estimates are reliable. In other words, the experimental data are credible.

It can be seen from Tables 4 that, the estimated distributions of *Y* for the two comparisons: Group A1 versus Group B and Group A2 versus Group B are significantly different: $Y \sim N(3.81, 5.78)$ versus $Y \sim N(8.70, 13.38)$. However, the difference in the values of the RSU, SNR, SCI, $\mathrm{EP}_{X_A \geq X_B}$, $\mathrm{EP}_{A \geq B}$, and NSP between the two comparisons are not significant. Thus, the two comparisons: Group A1 versus Group B and Group A2 versus Group B should give the same conclusion: the new flavoring is better than the old flavoring.

Note that the values of EPs and NSPs from the distribution-based comparison are consistent with the values from the nonparametric comparison. $\mathrm{EP}_{X_A \geq X_B}$ = 0.745, 0.742, and NSP=0.490, 0.484 based on the distribution-based comparison, while $\mathrm{EP}_{A \geq B}$ =0.741, 0.725, and NSP=0.482, 0.450 based on the nonparametric comparison. These results indicate that the new flavoring is significantly superior to the old flavoring.

Therefore, the comprehensive descriptive statistics given by the descriptive statistic analysis suggest that we should be in favor of the new flavoring over the old flavoring.

5. Why should the *t*-interval method for measurement uncertainty calculation be abandoned?

In measurement science, the half-width of the *t*-interval is defined as the Type A expanded uncertainty for a measurement with a small number of observations (e.g. JCGM 2008). It is called the *t*-based uncertainty. In this section, we discuss two main issues with the *t*-interval and *t*-based uncertainty: rationale and methodology, which explains why the *t*-interval method for measurement uncertainty calculation should be abandoned. We also discuss the issues with the *t*-distribution, which is the basis of the *t*-interval and *t*-based uncertainty.

5.1 Rationale issue: "coverage" is a misleading concept

The rationale behind using the *t*-interval method for measurement uncertainty calculation is "coverage". Indeed, "coverage", as expressed as confidence level or coverage probability, is the central concept in Neyman confidence interval theory (Neyman 1935, 1937). However, it should be noted that confidence level is not the probability in the mathematical sense; it is the so-called "long-term success rate" (e.g. Willink 2010) or "capture rate" (Huang 2018b). In Monte Carlo simulation of the *t*-interval, the success rate or capture rate approaches the nominal confidence level $(1-\alpha)$ asymptotically. That is,

$$\text{success rate or captrue rate} = \lim_{m\to\infty}\frac{k}{m} = 1-\alpha\,, \tag{17}$$

where m is the total number of the simulated intervals and k is the number of the simulated intervals that have captured the true value μ.

Therefore, strictly speaking, confidence level is not the *mathematical probability* that must satisfy Kohnogorov's axioms of probability calculus; it is a relative frequency. However, according to Bunge (1981), "… frequencies alone do not warrant inferences to probabilities …". This is because "… whereas a probability statement concerns usually a single (though possible complex) fact, the corresponding frequency statement is about a set of facts and moreover as chosen in agreement with certain sampling procedures." Bunge (1981) argued, "… the frequency interpretation [of probability] is mathematically incorrect because the axioms that define the probability measure do not contain the (semiempirical) notion of frequency."

It is important to note that "coverage" (the *frequency* of "success" or "capture") is a property of a confidence interval procedure (e.g. the *t*-interval procedure). The "coverage" can be achieved only in the long run of repeated sampling or simulation; it is meaningless for a confidence interval calculated from a sample.

We must distinguish between the *result* of a procedure (statistical method) and the *coverage* of the procedure. In measurement uncertainty analysis, we are interested in the estimated uncertainty (measurement precision), which is the *result* from a procedure. Kempthorne (1976) stated, "…a statistical method should be judged by the result which it gives in practice." However, "coverage" is not a result given by a statistical method. Therefore, "coverage" cannot be used to judge the method's performance in practice. In fact, it would be paradoxical to judge an uncertainty estimation method by "coverage" (Huang 2018c)".

It should be emphasized that, a confidence interval procedure is merely to generate a collection of confidence intervals (called "sticks") with a stated capture rate for the unknown true value (Huang 2018b). Therefore, the *t*-interval provides an "exact" answer to the following question: "What is the *interval procedure* with which the population mean μ would be captured by 1-α of all intervals generated in the long-run of repeated sampling." However, this is a wrong

question for measurmenet uncertianty analysis. The purpose of measurement uncertainty analysis is to determine (or estimate) the measurement precision. The right question is: "How we estimate measurement precision with a given sample?" (Huang 2018d). The *t*-interval procedure is *not* a statistical method for inferring measurement precision. Morey et al. (2016a) stated, "Claims that confidence intervals yield an index of precision, that the values within them are plausible, and that the confidence coefficient can be read as a measure of certainty that the interval contains the true value, are all fallacies and unjustified by confidence interval theory." Therefore, the *t*-interval method is actually misused in measurement uncertainty analysis because it is an "exact" answer to the wrong question (Huang 2018d).

5.2 Methodological issue: the t-interval or t-based uncertainty is a distorted mirror of physical reality

The half-width of the *t*-interval is written as $U_t = t_{\alpha/2}\frac{s}{\sqrt{n}}$ (called the *t*-based uncertainty), where *n* is the number of observations, *s* is the sample standard deviation, and $t_{\alpha/2}$ the *t*-score. The true expanded uncertainty of the sample mean of *n* observations is written as $U_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and is called the *z*-based uncertainty, where $z_{\alpha/2}$ is the *z*-score and σ is the population standard deviation. The *t*-based uncertainty artificially dilates uncertainty. The artificial dilation can be measured by the 'dilation factor' that is defined as the ratio between the expectation of the *t*-based uncertainty and the true expanded uncertainty. That is (Huang 2018b),

$$\text{Dilation factor} = \frac{\mathrm{E}(U_t)}{U_Z} = \frac{c_{4,n}t_p}{z_p}. \tag{18}$$

The dilation factor is extremely high when the sample size is small. For example, at *n*=2, the dilation factor is 5.17 for the nominal coverage probability $1-\alpha = 0.95$ and 19.72 for $1-\alpha = 0.99$. The dilation factor decreases with increasing the sample size. At *n*=30, it is 1.03 for $1-\alpha = 0.95$ and 1.06 for $1-\alpha = 0.99$.

It is important to note that the *z*-based uncertainty $U_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ expresses a physical law, known as the -1/2 power law. The -1/2 power law describes the relationship between the random uncertainty of the sample mean and the number of observations. That is, the random noise of the sample mean decreases as the sample size increases, following $1/\sqrt{n}$. The expectation of the *t*-based uncertainty U_t is $\mathrm{E}(U_t) = c_{4,n}t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, which significantly deviates from the -1/2 power law when the sample size is small as shown in figure 1. Therefore, the *t*-based uncertainty or the *t*-interval is a distorted mirror of the physical reality.

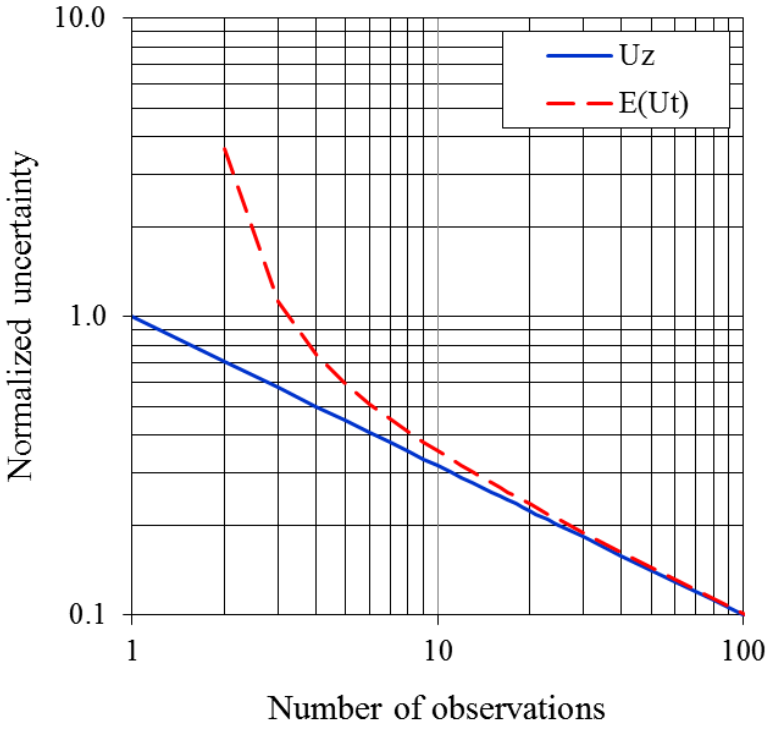

Figure 1. U_z and $\mathrm{E}(U_t)$ (normalized by $z_{\alpha/2}\sigma$ at 1-α=0.95) on the log-log scales (Huang 2018b)

It might be worth mentioning that, prior to Student (William Sealy Gosset), the expanded uncertainty (called "probable error" in Student's 1908 paper) was calculated based on the maximum-likelihood estimate of the population variance. This method significantly underestimates the uncertainty when the sample size is small, leading to the relative biases -43.6%, -20.2%, -7.7% at n=2, 4, and 10, respectively. To solve this underestimation problem, Student (1908) invented the t-distribution. However, the t-based uncertainty U_t, derived from the t-distribution, results in overestimation of the uncertainty, as indicated by the dilation factor, Eq. (18). Also, it is interesting to note that, according to Ziliak and McCloskey (2004), "Student used his t-tables a teensy bit…" They said, "We have learned recently, by the way, that "Student" himself—William Sealy Gosset—did not rely on Student's t in his own work."

5.3 Issues with the t-distribution

The t-interval and t-based uncertainty are constructed based on the t-distribution. Therefore, the methodological issue with the t-interval and t-based uncertainty must be traced back to the t-distribution or the scaled and shifted t-distribution (called the location-scale t-distribution in Wikipedia).

First, the t-distribution is subject to the so-called "t-transformation distortion" (Huang 2018a). The statistic t is a transformed quantity (i.e. the ratio between the sample error and the standard error of the sample mean). The original sample space $\Omega(\varepsilon, s)$ is tranformed into the distorted sample space $\Omega(t)$. The t-transformation itself is mathematically valid, so is the t-distribution. However, the inferences (such as the t-interval) based on the t-distribution may not

be valid because the inferences are actually performed in the distorted sample space $\Omega(t)$ (Huang 2018e). To understand this, consider that plums are dried to make prunes. The drying process is an analogy to the "*t*-transformation"; it distorts the shape of plums, which is analogical to the "*t*-transformation distortion". Therefore, we cannot correctly *infer* the shape of plums based on the shape of prunes.

Second, the scaled and shifted *t*-distribution is not an appropriate sampling distribution for the sample mean of n observations. According to the Central Limit Theorem, the sampling distribution for the sample mean of n observations approximates the normal distribution (also called the scaled and shifted *z*-distribution), regardless of the original distribution. The Central Limit Theorem does not support the scaled and shifted *t*-distribution. Moreover, according to the entropy metric, the scaled and shifted *t*-distribution is not the best distribution among the three candidate distributions considered (Huang 2023b). Two other candidate distributions are the scaled and shifted *z*-distribution and the Laplace distribution. The minimum entropy criterion states that the distribution with the minimum entropy should be chosen because it has the least information loss among a set of candidate distributions. For a given dataset obtained from n observations, the minimum entropy distribution is the scaled and shifted *z*-distribution (Huang 2023b). This is consistent with the Central Limit Theorem. Furthermore, the informity metric confirms the result given by the entropy metric (Huang 2023c). The informity metric is the counterpart of the entropy metric; it can be used as an alternative to the entropy metric (Huang 2023c). In summary, according to the Central Limit Theorem, the entropy metric, and the informity metric, the scaled and shifted *z*-distribution should be used instead of the scaled and shifted *t*-distribution. There is no mathematical or physical principle to support the *t*-distribution or the scaled and shifted *t*-distribution.

It is worth mentioning that, the statistics textbook written by Matloff (2014a) does not cover the *t*-distribution and *t*-intervals. Matloff (2014b) stated, "I advocate skipping the *t*-distribution, and going directly to inference based on the Central Limit Theorem."

6. Alternative to the *t*-interval method for measurement uncertainty calculation: unbiased estimation method

6.1 Unbiased estimation method

Again, for a measurement with a small number of observations, when σ is known, the *z*-based uncertainty U_z is the true expanded uncertainty. In practice, σ may be known from manufacturer's *precision* specification for a measuring instrument. Thus, U_z is can be regarded as the true precision. When σ is unknown, the true precision (i.e. U_z) cannot be known. However, we want to know approximate precision. Therefore, the purpose of uncertainty analysis is to estimate the true precision based on a sample at hand. Note that U_z depends on the population parameter σ. According to the theory of point estimation, σ can be replaced by a sample-based estimator $\hat{\sigma}$ when σ is unknown. Accordingly, U_z can be replaced by a sample-based uncertainty estimator, denoted by $\widehat{U}$. We want $\widehat{U}$ to be the same as U_z on average, i.e. $\widehat{U}$ to be a mean-unbiased estimator of U_z. Note that $s/c_{4,n}$ is a unbiased estimator of σ. Thus, $\widehat{U} = z_{\alpha/2}\frac{s}{c_{4,n}\sqrt{n}}$ is an unbiased estimator of U_z. Note that $\widehat{U} = z_{\alpha/2}\frac{s}{c_{4,n}\sqrt{n}}$ conform to the -1/2 power law.

Hirschauer (2022) stated,

> "What we can extract – at best – from a random sample is an unbiased point estimate (signal) of an unknown population effect size and an unbiased estimation of the uncertainty (noise), caused by random error, of that point estimation, i.e., the standard error, which is but another label for the standard deviation of the sampling distribution."

Indeed, the sample mean $\bar{y}$ (effect size) and the unbiased standard error $\frac{s}{c_{4,n}\sqrt{n}}$ are "what we can extract – at best – from a random sample…"

The unbiased estimation method can provide realistic uncertainty estimates. The "uncertianty paradox" caused by the *t*-interval method disappers when using the unbiased estimation method. For carpenter's laugh example given by D'Agostini (1998) (mentioned in the introduction), the *t*-score at (1-α)=0.999 is 636.62, due to severe *t*-transformation distortion at n=2. On the other hand, the *z*-score at (1-α)=0.999 is 3.29 and the bias correction factor $c_{4,n}$ at n=2 is 0.7979. The unbiased estimator gives $\widehat{U}$= 0.62 mm, which is much more realistic than the ridiculous result U_t=95 mm given by the *t*-interval method. Moreover, unlike the *t*-based uncertainty $U_t = t_{\alpha/2}\frac{s}{\sqrt{n}}$, which cannot be used for measurement quality control due to its high false rejection rate, the unbiased estimator $\widehat{U} = z_{\alpha/2}\frac{s}{c_{4,n}\sqrt{n}}$ can be used for measurement quality control. Importantly, the unbiased estimator $\widehat{U}$ is adopted in the ISO standard for streamflow measurements with acoustic Doppler current profiler (ISO:24578:2021(E)).

It should be emphasized that the unbiased estimation method is based on the theory of point estimation and the unbiasedness criterion. It is not an interval procedure like the *t*-interval method that is based on the confidence interval theory and the "coverage" criterion. These two methods are mutually incompatible and incommensurable. Therefore, the "coverage" criterion should not be applied to the unbiased estimation method. In other words, the performance of the unbiased estimation method should not be judged with the long-term success rate (or capture rate) that is commonly used to evaluate the performance of a confidence interval procedure (Huang 2018d, 2020).

Statistics textbooks usually claim that interval estimation is more informative than point estimation. However, this claim is wrong and misleading. Suppose we employ a statistical distribution model (e.g. normal distribution) in our analysis. As long as the model parameters are given by a valid statistical method (e.g. the method of maximum likelihood) applied to a given dataset, we can obtain an estimated distribution. We can then use this estimated distribution to construct any probability interval we want. For example, in the case of n observations, the sample mean $\bar{y}$ and the unbiased standard error $\frac{s}{c_{4,n}\sqrt{n}}$ are the estimated location and scale parameters. Then, the estimated sampling distribution of the sample mean $\bar{Y}$ is $N(\bar{y}, \frac{s^2}{c_{4,n}^2 n})$ (Huang 2019b). Certainly, this estimated distribution is more informative than any confidence interval.

The unbiased estimation method has been extended to the case where multiple uncertainty components are involved in the measurement uncertainty analysis, which is called the WS-*z* approach (Huang 2016). The WS-*z* approach resolves the Ballico paradox caused by the WS-*t* approach (a *t*-interval method) mentioned in the introduction.

6.2 Example: a comparison of the WS-z and WS-t approaches

Consider two random variables: X and Y. We assume that X is normally distributed with unknown mean and variance, and Y is normally distributed with mean zero and variance σ_Y. We randomly take a sample from X:$\{x_1, x_2, ..x_n\}$ and a sample from Y:$\{y\}$. Then, $Z = \bar{X} + Y$ is an unbiased estimator of the true value of the quantity $Z=X+Y$; the associated variance can be estimated as

$$\mathrm{Var}(Z) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(Y) = \frac{s_X^2}{n} + \sigma_Y^2, \tag{19}$$

where s_X is the sample standard deviation of the n observations $\{x_1, x_2, ..x_n\}$.

Our job is to estimate the expanded uncertainty of the estimate $z = \bar{x} + y$. According to the unbiased estimation method, the expanded uncertainty is calculated as

$$\widehat{U}_{\mathrm{WS}-z} = \frac{z_{\alpha/2}}{c_{4,\upsilon}} \sqrt{\frac{s_X^2}{n} + \sigma_Y^2}\,, \tag{20}$$

where $c_{4,\upsilon}$ is the bias correction factor and υ is the effective degrees of freedom (DOF). The effective DOF can be calculated using the Welch-Satterthwaite (WS) formula. For the two-sample problem considered the Welch-Satterthwaite formula can be written as (Huang 2016)

$$\upsilon = (n-1)\left[1 + n\frac{\sigma_Y^2}{s_X^2}\right]^2. \tag{21}$$

On the other hand, the t-interval method, i.e. the WS-t approach, for this problem is written as (Huang 2016)

$$\widehat{U}_{\mathrm{WS}-t} = t_{\alpha/2,\upsilon} \sqrt{\frac{s_X^2}{n} + \sigma_Y^2}\,. \tag{22}$$

To obtain some numerical results for comparison, we assume $s_X = 3$, σ_Y varies in the range 0 to 3, n=4, and α=0.05. Figure 2 shows the expanded uncertainty estimated by the WS-z and WS-t approaches.

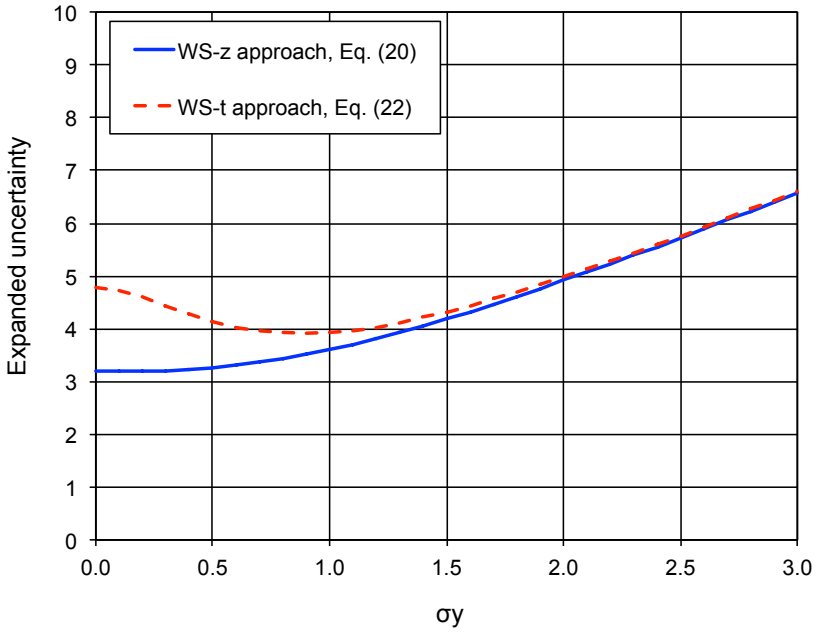

Figure 2. Expanded uncertainty estimated by the WS-z and WS-t approaches ($s_X = 3$, n=4, and α=0.05)

It can be seen from Figure 2 that, the WS-z approach gives realistic estimates of the expanded uncertainty. Importantly, the expanded uncertainty increases continuously with increasing σ_Y, which conforms to our domain knowledge (and common sense) about measurement uncertainty. In contrast, the WS-t approach gives unrealistic estimates of expanded uncertainty, not only because it overestimates uncertainty when σ_Y is small (dilates the uncertainty), but also exhibits a paradoxical behavior: the uncertainty decreases with increasing σ_Y in the range that σ_Y=0 to 0.9. Note that $\widehat{U}_{\mathrm{WS}-t}$ converges to $\widehat{U}_{\mathrm{WS}-z}$ only when σ_Y becomes large. This is expected because when σ_Y is large, σ_Y is more dominant than $\sqrt{s_X^2/n}$. This example shows that the WS-t approach or the t-interval method for measurement uncertainty calculation is inherent flawed.

7. Conclusion and recommendation

According to Jaynes (2003, p758), a paradox is "something which is absurd or logically contradictory, but which appears at first glance to be the result of sound reasoning." Also, "A paradox is simply an error out of control: i.e. one that has trapped so many unwary minds that it has gone public, become institutionalized in our literature, and taught as truth." In this regard, the two-sample t-test and the t-interval are such paradoxes. Statistics textbooks, journals, and computer software packages contribute greatly to spreading these paradoxes. As Hurlbert et al. (2019) stated that, "Many controversies in statistics are due primarily or solely to poor quality control in journals, bad statistical textbooks, bad teaching, unclear writing, and lack of knowledge of the historical literature."

Therefore, in order to implement statistics reform, statistics textbooks and computer software packages should be updated to reflect the paradigm shift from significance testing to

estimation statistics. The author agrees with Hurlbert et al. (2019), "… the term "statistically significant" and all its cognates and symbolic adjuncts be disallowed in the scientific literature except where focus is on the history of statistics and its philosophies and methodologies." Specifically, the two-sample *t*-test and the *t*-interval method for measurement uncertainty calculation (both are bad statistical methods) should be removed from statistics textbooks and computer software packages. On the other hand, good statistical methods such as least squares method and maximum likelihood estimation should withstand statistics reform.

The descriptive statistic analysis should be used instead of the two-sample *t*-test to compare two groups. The eight descriptive statistics: effect size (ES), relative effect size (RES), standard uncertainty (SU), relative standard uncertainty (RSU), signal-to-noise ratio (SNR), signal content index (SCI), exceedance probability (EP), and net superiority probability (NSP) extract the evidence embedded in the data from different aspects. We do not recommend setting a threshold on any of these statistics for inferences. Whether the estimated effect size is of practical importance should be judged on the basis of our domain knowledge with the consideration of these descriptive statistics.

The unbiased estimation method should be used instead of the *t*-interval method for measurement uncertainty calculation or effect size precision estimation. The "uncertainty paradox" and Ballico paradox caused by the *t*-interval method disappear when using the unbiased estimation method.

The author believes that the success of statistics reform depends on collaboration between statisticians and practitioners. The author hopes that this paper will stimulate the discussion among statisticians and practitioners about fundamental issues in the two statistical methods derived from the *t*-distribution: the two-sample *t*-test and the *t*-interval method for measurement uncertainty calculation.

References

Amrhein V, Greenland S, and McShane B 2019 Retire statistical significance *Nature* **567** 305-307

Benjamini Y, De V R, Efron B, Evans S, Glickman M, Graubard B I, He X, Meng X-L, Reid N, Stigler S M, Vardeman S B, Wikle C K, Wright T, Young L J and Kafadar K 2021 ASA President's Task Force Statement on Statistical Significance and Replicability *Harvard Data Science Review* **3**(3) https://doi.org/10.1162/99608f92.f0ad0287

Ballico M 2000 Limitations of the Welch-Satterthwaite approximation for measurement uncertainty calculations *Metrologia* **37** 61-64

Baguley T. Standardized or simple effect size: what should be reported? Br J Psychol. 2009 Aug;**100**(Pt 3) 603-17 doi: 10.1348/000712608X377117 Epub 2008 Nov 17 PMID: 19017432

Berner D and Amrhein V 2022 Why and how we should join the shift from significance testing to estimation *J Evol Biol.* **35**(6) 777-787 doi: 10.1111/jeb.14009. Epub 2022 May 18. PMID: 35582935; PMCID: PMC9322409. https://onlinelibrary.wiley.com/doi/10.1111/jeb.14009

Bonovas S and Piovani D 2023 On *p*-Values and Statistical Significance *Journal of Clinical Medicine* **12**(3) 900 https://doi.org/10.3390/jcm12030900

Bunge M 1981 Four concepts of probability *Applied Mathematical Modelling* 5(5) 306-312

Claridge-Chang A and Assam P 2016 Estimation statistics should replace significance testing *Nat Methods* **13** 108–109 https://doi.org/10.1038/nmeth.3729

Colling L J and Szűcs D 2021 Statistical Inference and the Replication Crisis, *Review of Philosophy and Psychology* **12** 121–147 https://doi.org/10.1007/s13164-018-0421-4
Cumming G 2014 The New Statistics *Psychological Science* **25**(1) DOI: 10.1177/0956797613504966
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, May 30th-June 4th)
Delaney H D and Vargha A 2002 Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples *Psychol Methods* **7**(4) 485-503 doi: 10.1037/1082-989x.7.4.485. PMID: 12530705
Di Toro D M 1984 Probability model of stream quality due to runoff *Journal of Environmental Engineering* ASCE **110**(3) 607-628
Elkins MR, Pinto RZ, Verhagen A, Grygorowicz M, Söderlund A, Guemann M, Gómez-Conesa A, Blanton S, Brismée JM, Ardern C, Agarwal S, Jette A, Karstens S, Harms M, Verheyden G, Sheikh U. 2022 Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors *European Journal of Physiotherapy* (2022) **24**(3) 129-133 DOI: 10.1080/21679169.2022.2073991
Environment protection agency (EPA) 1991 *Technical support document for water quality-based toxics control,* Office of Water, Washington, DC, EPA/505/2-90-001
Etz A 2015 Confidence intervals? More like confusion intervals *The Featured Content Blog of the Psychonomic Society Digital Content Project* https://featuredcontent.psychonomic.org/confidence-intervals-more-like-confusion-intervals/
Fisher R A 1956 *Statistical Methods and Scientific Inference* Edinbrgh: Oliver and Boyd
Grissom R J and Kim J J 2001 Review of assumptions and problems in the appropriate conceptualization of effect size *Psychol Methods* **6**(2) 135-46 doi: 10.1037/1082-989x.6.2.135. PMID: 11411438
Haig B D 2016 Tests of statistical significance made sound *Educational and Psychological Measurement* **77** 489–506. https://doi.org/10.1177/0013164416667981
Halsey L G 2019 The reign of the *p*-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters* **15**(5) 20190174 https://doi.org/10.1098/rsbl.2019.0174
Halsey L, Curran-Everett D, Vowler S *et al.* 2015 The fickle *P* value generates irreproducible results *Nat Methods* **12** 179–185 https://doi.org/10.1038/nmeth.3288
Hand D J 2022 Trustworthiness of Statistical Inference *Journal of the Royal Statistical Society Series A*: *Statistics in Society* 1**85** (1) 329–347 https://doi.org/10.1111/rssa.12752
Hirschauer N 2022 Some thoughts about statistical inference in the 21st century SocArXiv. December 20 doi:10.31235/osf.io/exdfg
Huang H 2010 A paradox in measurement uncertainty analysis '*Global Measurement: Economy & Technology' 1970 - 2010 Proceedings* (DVD) (Measurement Science Conference)
Huang H 2014 Uncertainty-based measurement quality control *Accred Qual Assur* **19** 65-73
Huang H 2016 On the Welch-Satterthwaite formula for uncertainty estimation: a paradox and its resolution *Cal Lab the International Journal of Metrology* **23** 20-28
Huang H 2018a 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** https://doi.org/10.1088/1361-6501/aa96c7

Huang H 2018b More on the *t*-interval method and mean-unbiased estimator for measurement uncertainty estimation *Cal Lab the International Journal of Metrology* **25** 24-33
Huang H 2018c A unified theory of measurement errors and uncertainties *Measurement Science and Technology* **29** 125003 https://doi.org/10.1088/1361-6501/aae50f
Huang H 2018d Uncertainty estimation with a small number of measurements, Part II: a redefinition of uncertainty and an estimator method *Measurement Science and Technology* **29** https://doi.org/10.1088/1361-6501/aa96d8
Huang H 2019a Signal content index (SCI): a measure of the effectiveness of measurements and an alternative to *p*-value for comparing two means *Measurement Science and Technology* **31** 045008 https://doi.org/10.1088/1361-6501/ab46fd
Huang H 2019b Why the scaled and shifted *t*-distribution should not be used in the Monte Carlo method for estimating measurement uncertainty? *Measurement* **136** 282-288 https://doi.org/10.1016/j.measurement.2018.12.089
Huang 2020 Comparison of three approaches for computing measurement uncertainties *Measurement* **163** https://doi.org/10.1016/j.measurement.2020.107923
Huang H 2022 Exceedance probability analysis: a practical and effective alternative to t-tests *Journal of Probability and Statistical Science* **20**(1) 80-97
Huang H 2023a Probability of net superiority for comparing two groups or group means *Lobachevskii Journal of Mathematics* **44**(11) 42-54
Huang H 2023b A minimum entropy criterion for distribution selection for measurement uncertainty analysis *Measurement Science and Technology* **35**(3) 035014 DOI: 10.1088/1361-6501/ad1476
Huang 2023c A theory of informity preprint *ResearchGate* DOI: 10.13140/RG.2.2.28832.97287
Huang H and Fergen R E 1995 Probability-domain simulation - A new probabilistic method for water quality modeling *WEF Specialty Conference "Toxic Substances in Water Environments: Assessment and Control"* (Cincinnati, Ohio, May 14-17, 1995)
Hurlbert SH, Levine RA, and Utts J 2019 Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires *The American Statistician* **73**(sup1):352-357 DOI: 10.1080/00031305.2018.1543616
International Organization of Standards (ISO) 2021 ISO:24578:2021(E), Hydrometry — Acoustic Doppler profiler — Method and application for measurement of flow in open channels from a moving boat, first edition, 2021-3, Geneva Switzerland
Jaynes E T 2003 *Probability Theory*: *The Logic of Science* Cambridge University Press
Jenkins J D 2007 The Student's t-distribution uncovered *Measurement Science Conference Proceedings* Long Beach
Joint Committee for Guides in Metrology (JCGM) 2008 *Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement* (GUM 1995 with minor corrections) Sevres, France
Karlen D 2002 Credibility of confidence intervals *Proceedings of the Conference on Advanced Techniques in Particle Physics* (Durham 18–22 March 2002) Eds. M Whalley and L Lyons
Kempthorne O 1976 Comments on paper by Dr. E. T. Jaynes 'Confidence intervals vs Bayesian intervals' *Foundations of Probability Theory, Statistical Inference, and Statistical Theories and Science* **Vol. II** 175-257 Eds. Harper and Hooker (Dordrecht-Holland: D. Reidel Publishing Company)

Krishnamoorthy K, Mathew T, and Ramachandran G 2007 Upper limits for exceedance probabilities under the one-way random effects model *The Annals of Occupational Hygiene* **51**(4) 397-406 doi:10.1093/annhyg/mem013

Lazzeroni L C, Lu Y, and Belitskaya-Lévy I 2016 Solutions for quantifying *P*-value uncertainty and replication power *Nature Methods* 13(2) 107-108

Lohse K 2022 In Defense of Hypothesis Testing: A Response to the Joint Editorial From the International Society of Physiotherapy Journal Editors on Statistical Inference Through Estimation *Physical Therapy*, **102**(11) 118 https://doi.org/10.1093/ptj/pzac118

Matloff N 2014a Open Textbook: *From Algorithms to Z-Scores: Probabilistic and Statistical Modeling in Computer Science* (University of California, Davis)

Matloff N 2014b Why are we still teaching t-tests? On the blog: Mad (Data) Scientist—data science, R, statistic https://matloff.wordpress.com/2014/09/15/why-are-we-still-teaching-about-t-tests/

McGraw K O and Wong S P 1992 A common language effect size statistic *Psychological Bulletin* **111**(2) 361–365 https://doi.org/10.1037/0033-2909.111.2.361

McShane B B, Gal D, Gelman A, Robert C P, and Tackett J L 2018 Abandon statistical significance *The American Statistician* **73** DOI: 10.1080/00031305.2018.1527253

Morey R D, Hoekstra R, Rouder J N, Lee M D and Wagenmakers E-J. 2016a The fallacy of placing confidence in confidence intervals *Psychon Bull Rev* **23** 103-123 https://rd.springer.com/article/10.3758%2Fs13423-015-0947-8

Morey R D, Hoekstra R, Rouder J N and Wagenmakers E-J 2016b Continued misinterpretation of confidence intervals: response to Miller and Ulrich. *Psychonomic Bulletin & Review* **23** 131-140 https://link.springer.com/article/10.3758%2Fs13423-015-0955-8

Neyman J 1935 On the problem of confidence intervals *Ann Math Stat* **6** 111-116

Neyman J 1937 Outline of a theory of statistical estimation based on the classical theory of probability *Philos Trans R Soc Lond* **A236** 333-380

Normile C J, Bloesch E K, Davoli C C and Scherr K C 2019 Introducing the new statistics in the classroom *Scholarship of Teaching and Learning in Psychology* **5**(2) 162–168 https://doi.org/10.1037/stl0000141

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

Ruscio J and Mullen T 2012 Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve *Multivariate Behavioral Research* **47**(2) 201–223 https://doi.org/10.1080/00273171.2012.658329

Schäfer T 2023 On the use and misuse of standardized effect sizes in psychological research OSF Preprints June 7 doi:10.31219/osf.io/x8n3h

Siegfried T 2010 Odds Are, It's wrong: science fails to face the shortcomings of statistics *Science News* **177** 26 https:// www.sciencenews.org/article/odds-are-its-wrong

Siegfried T 2014 To make science better, watch out for statistical flaws *Science News* Context Blog, February 7, https://www.sciencenews.org/blog/context/make-science-betterwatch-out-statistical-flaws

Stansbury D 2020 p-Hacking 101: N Chasing *The Clever Machine* https://dustinstansbury.github.io/theclevermachine/p-hacking-n-chasing

Student (William Sealy Gosset) 1908 The probable error of a mean *Biometrika* **VI** 1-25

Trafimow D 2018 Confidence intervals, precision and confounding *New Ideas in Psychology* **50** 48-53 https://doi.org/10.1016/j.newideapsych.2018.04.005

Trafimow D 2023 The Story of My Journey Away from Significance Testing A World Scientific Encyclopedia of Business Storytelling, pp. 95-127 DOI: 10.1142/9789811280948_0006
Trafimow D and Marks M 2015 Editorial *Basic and Applied Social Psychology* **37** 1-2
Vargha A and Delaney H D 2000 A critique and improvement of the CL common language effect size statistic of McGraw and Wong *Journal of Educational and Behavioral Statistics* **25** 101–132 doi: 10.3102/10769986025002101
Wadsworth Jr H M 1989 Summarization and interpretation of data *Handbook of Statistical Methods for Engineers and Scientists* 2.1-2.21, Ed. Harrison M Wadsworth Jr. McGRAW-HILL In
Wagenmakers E-J, R. Wetzels D B and Maas H L J van der 2011 Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem *Journal of Personality and Social Psychology* **100** 426–432 https://doi.org/10.1037/a0022790
Wasserstein R L and Lazar N A 2016 The ASA's statement on *p*-values: context, process, and purpose, *The American Statistician* **70** 129-133 DOI:10.1080/00031305.2016.1154108
Wasserstein R L, Schirm A L, and Lazar N A 2019 Moving to a world beyond "p < 0.05" *The American Statistician* **73**:sup1 1-19 DOI: 10.1080/00031305.2019.1583913
Willink R 2010 Probability, belief and success rate: comments on 'On the meaning of coverage probabilities' *Metrologia* **47** 343–346
Zaiontz C 2020 Two sample *t* test: unequal variances *Real Statistics Using Excel* https://real-statistics.com/students-t-distribution/two-sample-t-test-uequal-variances/ accessed on August 22, 2023
Ziliak S T and McCloskey D N 2004 Significance redux *The Journal of Socio-Economics* **33**(5) 665–675 https://doi.org/10.1016/j.socec.2004.09.038
Ziliak S T and McCloskey D N 2008 *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* University of Michigan Press https://doi.org/10.3998/mpub.186351