Research Article

Statistics Reform: Practitioner's Perspective

Hening Huang[1]

1. Teledyne Technologies (United States), Thousand Oaks, United States

It is widely believed 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 to call for statistics reform. As a practitioner in hydraulics and measurement science, the author extensively used statistical methods in environmental engineering and hydrological survey projects. The author strongly concurs with 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, while others are flawed and should be abandoned and removed from textbooks and software packages. This paper focuses on two methods derived from the *t*-distribution: the two-sample *t*-test and the *t*-interval method for calculating measurement uncertainty. We demonstrate why both methods should be abandoned. We recommend using advanced estimation statistics in place of the two-sample *t*-test and an unbiased estimation method in place of the *t*-interval method. Two examples are presented to illustrate the recommended approaches.

1. Introduction

In recent years, the scientific community has become increasingly concerned about the replication crisis. Many scientists believe that one of main causes of the replication crisis is some of the most commonly used statistical methods. Specifically, null hypothesis significance testing (NHST) and its produced *p*-values, and claims of statistical significance, have come in most to blame[1]. Siegfried[2] remarked, "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." Siegfried[3] further claimed, "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies."

 doi.org/10.32388/JVYEJZ.2

In response to these concerns, many scientists have called for the retirement or abandonment of statistical significance and *p*-values (e.g.[4][5][6][7][8]). For instance, since 2015, *Basic and Applied Social Psychology* has banned NHST procedures and *p*-values[9]. Furthermore, many scientists have advocated for statistics reform (e.g.[10][11][12]). Cumming[13] and Cumming and Calin-Jageman[14] proposed the "New Statistics", which primarily involves (1) abandoning NHST procedures and (2) using effect sizes and confidence intervals. Normile et al.[15] introduced the "New Statistics" in classroom settings. Claridge-Chang and Assam[16] suggested replacing significance testing with estimation statistics. A co-published editorial of 14 physiotherapy journals[17] "... 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." More recently, Trafimow et al.[18][19] proposed using a two-step process comprising the APP (*a prior* procedure) and gain-probability analyses to replace the traditional two-step process comprising the power analysis and NHST. Although some authors continue to defend NHST and *p*-values (e.g.[20][21][22][23]) and the debate persists (e.g.[24][25]), Berner and Amrhein[26] noted that "A paradigm shift away from null hypothesis significance testing seems in progress".

As a practitioner in hydraulics and measurement science, the author extensively used statistical methods in environmental engineering and hydrological survey projects (citation omitted). In particular, the author processed thousands of small samples collected during streamflow measurements using acoustic Doppler current profilers (ADCPs). Typically, an ADCP streamflow measurement involves only a few observations (usually around 4). According to the *Guide to the Expression of Uncertainty in Measurement* (GUM; JCGM[27]), the uncertainty of the sample mean from a small sample should be calculated using the *t*-interval method. However, the author found that the *t*-based uncertainty (i.e. the half-width of the *t*-interval) was unrealistic and misleading, leading to the so-called "uncertainty paradox"[28][29] and a high false rejection rate in the quality control of ADCP streamflow measurements[30].

The author is not alone in questioning the validity of the *t*-interval method for calculating measurement uncertainty. Jenkins[31] also found that the *t*-based uncertainty can exhibit significant bias and precision errors. D'Agostini[32] provided a striking 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." Furthermore, Ballico[33] reported a notable counterinstance during a routine calibration at the CSIRO National Measurement Laboratory (NML), Australia. In this instance, a thermometer was calibrated for a 1 mK range (higher precision) and a 10 mK range (lower precision); the uncertainty was calculated using the WS-*t* approach (which combines the Welch–Satterthwaite formula with the *t*-interval). Intuitively, one would expect the thermometer in the higher precision range to have a lower uncertainty than in the lower precision range. However, the WS-*t* approach gave a counterintuitive result: the uncertainty for the 1 mK range was 37.39, compared to 35.07 for the 10 mK range. This counterintuitive result became known as the Ballico paradox[34].

Practitioners in science and industry rely on statistical methods in their work, and the use of flawed methods can have significant negative impacts. Ziliak and McCloskey[35] demonstrated in their extensive 322-page volume that "*Statistical significance is an exceptionally damaging one."* However, over the years, the author has observed that practitioners are frequently accused of misunderstanding and misusing certain statistical methods, particularly NHST procedures and *p*-values, even though the root issue may lie with the methods themselves. In the author's view, if a statistical method or concept is so prone to misunderstanding and misusing that even educational institutions struggle to teach it effectively, then there is likely something inherently wrong with that method or concept. Trafimow[36] argued that, "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 NHST procedures and *p*-values persists, one fact remains clear: after roughly 100 years, NHST procedures and *p*-values have not withstood the test of time.

We understand that no statistical method is perfect, nor can any method be applied without limitations or conditions. However, in the experience of the author and other practitioners, some statistical methods, such as the least squares method and point estimation, have proven to be good and useful. In contrast, other methods, such as the *t*-interval method for calculating measurement uncertainty, have demonstrated serious flaws. The coexistence of sound and problematic methods can be confusing to practitioners, many of whom may not realize that some methods are flawed or controversial and continue to use them inadvertently.

Therefore, the author strongly concurs with the need for statistics reform. This paper offers a practitioner's perspective on statistics reform. We argue that good methods should be preserved, while flawed methods should be abandoned and removed from textbooks and software packages. This paper focuses on two methods derived from the t-distribution: the two-sample t-test and the t-interval method for calculating measurement uncertainty. We will present arguments for why these two methods should be abandoned and propose alternatives.

The rest of the paper is organized as follows. Section 2 briefly reviews examples of good statistical methods that should withstand statistics reform. Section 3 discusses why the two-sample t-test should be abandoned, while Section 4 describes an alternative to this test. Section 5 discusses why the t-interval method for calculating measurement uncertainty should be abandoned, while Section 6 describes an alternative to the t-interval method. 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 possess the following characteristics: (a) it should have clear mathematical meaning and be easily understood, even by those without advanced training in statistics; (b) it should yield realistic results in real-world applications; and (c) it should be relatively uncontroversial in the scientific community. Furthermore, ideally, a good statistical method would be related to a physical principle, thereby giving it with physical meaning. Many good statistical methods meet these criteria. Four examples are listed below.

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

Perhaps, the least squares method is one of the widely used statistical methods in practice. It is hardly controversial in the scientific community, and importantly, it conforms to the principle of minimum energy, a fundamental concept in physics. In this context, the sum of squared errors can be interpreted as representing the internal noise energy of the system under consideration, which naturally tends to a minimum value at equilibrium.

The method of maximum likelihood is another widely used statistical methods. It is also hardly controversial in the scientific community. The method of maximum likelihood is intuitive in nature; as Fisher[37] stated, "The likelihood supplies a natural order of preference among the possibilities under consideration." In other words, the mode of a likelihood function corresponds to the most preferred parameter value given the data[38]. This idea is straightforward and does not require advanced statistical knowledge to understand. In addition, the method of maximum likelihood is essentially consistent with the least squares method.

The Central Limit Theorem states that, given a sufficiently large sample size, the sampling distribution of the sample mean will approximate a normal distribution, regardless of the original distribution. Since measurement error is defined as the difference between the true value and the measured value (e.g. the sample mean), the Central Limit Theorem aligns with the law of error, which is one of the foundational principles in statistics and measurement science.

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

Of course, good statistical methods like 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 widely used procedure among NHST procedures. Therefore, if we are to abandon NHST procedures, the two-sample *t*-test should be abandoned first. However, the literature rarely provides an explicit discussion of the reasons for abandoning the two-sample *t*-test, and usually offer only general debates about the problems with NHST procedures and *p*-values. It is important to note that *p*-values are outputs of statistical methods such as the two-sample *t*-test. Thus, *p*-value problems are not solely with *p*-values but with the statistical methods that produce them. In this section, we address two main issues with the two-sample *t*-test: logic and performance. We argue that these shortcomings provide compelling reasons for its abandonment.

3.1. Logic issue: the two-sample t-test is philosophically flawed and misleading

The two-sample *t*-test is philosophically flawed and misleading. Consider two datasets (groups): Group A from treatment A and Group B from treatment B. We are interested in determining whether treatment A is superior to treatment B (or vice versa). In the standard NHST framework, we begin with

a null hypothesis, a "strawman", that the unknown population means of the two groups are the same, and an alternative hypothesis that they differ. Then, we use the two-sample *t*-test to generate a *p*-value. If $p<0.05$, we conclude that the deference between the two means is "statistically significant" and the null hypothesis is rejected, i.e. the "strawman" is disproven. However, this approach does not answer the question of superiority between treatments A and B. Instead, it misleads us to focus on whether the groups differ in a statistically significant manner, based on an arbitrary *p*-value threshed. In reality, simply examining the data or comparing the group means often suffices to show that treatment A is different from treatment B. We should directly assess the practical significance of the observed difference using our domain knowledge. There is no intrinsic need to construct a "strawman" (the null hypothesis) and then try to disprove it.

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

Even if we accept its logic and use it for comparing the means of two groups, the two-sample *t*-test does not provide reliable results. This can be understood by examining the behaviors of the *p*-value produced by the test. First, as with any sample statistics, the *p*-value itself is subject to uncertainty. Halsey et al.[23] discussed the uncertainty associated with the *p*-value of two-sample *t*-tests through simulations. They demonstrated that "a major cause of the lack of repeatability is the wide sample-to-sample variability in the *P* value." They stated that, "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." Moreover, Lazzeroni et al.[39] 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 emphasized that, "*P* values are variable, but this variability reflects the real uncertainty inherent in statistical results."

Second, the two-sample *t*-test may produce inconsistent results for essentially the same evidence. Bonovas and Daniele[40] illustrated this issue with two trials of a new drug. In a single-center, randomized, double-blind, placebo-controlled trial, the two-sample *t*-test produced a *p*-value of 0.11, suggesting "no difference" between the active drug and placebo. In contrast, a multi-center trial yielded a *p*-value of 0.001, indicating a "significant difference." Despite these conflicting *p*-values, the risk ratio was the same in both trials: 0.70, indicating that the efficacy of the experimental drug was the same. This discrepancy highlights a critical shortcoming of the two-sample *t*-test: its reliance

on *p*-values can lead to inconsistent and potentially misleading conclusions, even when the effect size is consistent.

Third, the *p*-value produced by a two-sample *t*-test is highly dependent on the sample size; it decreases as the sample size increases. Therefore, *p*-values can be easily "hacked" through "*N*-chasing" (a term coined by Stansbury[41]), which guarantees "statistical significance" at any pre-specified threshold even if the effect size (e.g. the difference between the means of two groups) is trivial and lacks practical significance. "*N*-chasing" is one of the most effective ways of *p*-hacking. In the author's opinion, the only viable solution to combat "*N*-chasing" or *p*-hacking is to abandon the two-sample *t*-test.

4. Alternative to the two-sample *t*-test: advanced estimation statistics

We recommend using advanced estimation statistics as in place of the two-sample *t*-test. This framework emphasizes a comprehensive presentation of a set of statistics, including the observed 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.

In this advanced estimation statistics framework, the superiority of treatment A over treatment B (or vice versa) is measured by the observed ES (or RES) along with the EP (or NSP). The reliability of the observed ES (or RES) is then assessed using the SU, RSU, SNR, and SCI. Importantly, we do not specify a fixed threshold for any of these statistics; instead, we make scientific inferences, rather than purely statistical inferences, based on domain knowledge while considering these statistics.

Moreover, this advanced estimation statistics framework avoids the terminology and language associated with the NHST paradigm. Terms such as null hypothesis, alternative hypothesis, *p*-values, statistical significance, and statistical power are eliminated.

4.1. Observed effect size (ES) and relative effect size (RES)

The observed effect size (ES), denoted by Δ, is defined as the absolute deference between the two group means ($\overline{x}_A$ and $\overline{x}_B$, respectively)

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

It is important to note that the observed ES represents the "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. This is in contrast to standardized effect sizes, such as Cohen's *d*, which is dimensionless. Because the simple effect size retains the original physical unit, it is nearly always more meaningful than standardized effect size[42]. Schäfer[43] 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, practitioners' domain knowledge is inherently tied to the physical units of the quantity of interest. Therefore, it is more intuitive for practitioners to assess the practical significance using simple effect sizes. As Baguley[42] noted, "For most purposes simple (unstandardized) effect size is more robust and versatile than standardized effect size." Therefore, we do not recommend using standardized effect sizes such as Cohen's *d* in the advanced estimation statistics framework.

Note also that the observed ES represents the absolute magnitude of the treatment effect. In practice, we are often interested in the relative magnitude of the treatment effect, i.e. the relative effect size (RES). According to Huang[38], RES is defined as the ratio of the observed ES to a baseline measure, such as the average of the two group means. That is,

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

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

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

where $Var\left(\overline{x}_A\right) = s_A^2/n_A$ and $Var\left(\overline{x}_B\right) = s_B^2/n_B$; 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. The RES is usually expressed as a percentage.

The observed ES or RES is independent of sample size. As such, it only emphasizes the treatment effect. Unlike the two-sample *t*-test, which confounds the treatment effect with sample size, increasing the sample size does not alter the observed ES or RES but rather improves its reliability. Therefore, in contrast to *p*-values from *t*-tests, which are vulnerable to "*N*-chasing", the observed ES 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 observed ES is a point estimate of the unknown true effect size. Its reliability must be quantified and assessed. To this end, statistics such as the standard uncertainty (SU), relative standard uncertainty (RSU), signal-to-noise ratio (SNR), and signal content index (SCI) are employed. These statistics collectively provide a comprehensive assessment of the reliability of the observed ES.

Let $u(\Delta)$ denote the SU of the observed ES Δ. $u(\Delta)$ is defined as the standard deviation of $\Delta = |\bar{x}_A - \bar{x}_B|$

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

In measurement science, $u(\Delta)$ is often used as a measure of the precision of a measurement. If we treat the observed ES Δ as a measurement result, then $u(\Delta)$ measures its precision. Note that $u(\Delta)$ has the same physical unit as Δ.

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

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

The signal-to-noise ratio (SNR) is defined as the ratio of signal energy to noise energy. Although it is commonly used in electrical engineering, the concept applies to any signal[44]. For comparing the means of two groups, the observed ES Δ represents the signal, while the SU $u(\Delta)$ represents the noise. Therefore, the SNR is given by

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

Moreover, the signal content index (SCI) is defined as[44]

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

The SCI has a clear physical meaning; it is the relative amount of signal energy contained in the measurement result[44].

Either the SNR or the SCI can be used to measure the reliability of the observed ES. However, because the SCI is bounded between 0 and 1, its interpretation is more intuitive. A high SCI value (e.g. close to 1) indicates that the observed ES is reliable, while a low SCI value (e.g. close to 0) indicates that the observed ES is unreliable due to noise.

It should be noted that, unlike the observed ES or RES, which is independent of sample size, the reliability measures such as the SU, RSU, SNR, and SCI are functions of sample size. As sample size increases, the SU and RSU decrease, while the SNR and SCI increase. This establishes a clear distinction between the observed ES and its reliability measures.

It should also be noted that we do not use confidence interval to quantify the uncertainty (or precision) of the observed ES. This is because the concept of confidence intervals has long been controversial and subject to debate in the scientific community (e.g.[45][46][47][48][49]). In particular, the *t*-interval, which is a confidence interval traditionally used for small samples, is problematic and, as discussed in Section 5, should be abandoned.

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

The observed ES measures the difference, on average, between the two treatments A and B. In other words, when we assume that $\overline{x}_A - \overline{x}_B > 0$, the observed ES quantifies the average superiority of treatment A over treatment B. However, in practice, we are also interested in assessing superiority at the individual level. This means comparing the individual scores in the two groups to determine how often individuals in Group A outperform those in Group B.

The probability that Group A is superior to Group B at the individual level is known as exceedance probability (EP) and is defined as[38]

$$EP_{X_A \geq X_B} = Pr\left(X_A \geq X_B\right) = \int_0^\infty p(y)dy, \tag{8}$$

where X_A and X_B represent the scores of individuals in Groups A and B, respectively, and $p(y)$ is the probability density function for the quantity $Y = X_A - X_B$.

The meaning of the exceedance probability $EP_{X_A \geq X_B}$ is essentially the same as that of several other statistics, including the common language effect size (CLES)[50], the probability of superiority (PS)[51][52], and the area under the receiver operating characteristic curve (AUC) or its nonparametric version (A)[53][54]. However, it is important to note that calculating the CLES requires assumptions of population normality and equal variances, whereas $EP_{X_A \geq X_B}$ does not require these assumptions. In this sense, the CLES is an approximation of $EP_{X_A \geq X_B}$. Additionally, the term "CLES" can be misleading, as it might imply that it is an effect size, when in fact it represents a probability.

Assume that both X_A and X_B are normally distributed with unknown means and unknown variances. The estimated distributions of X_A and X_B are $N(\bar{x}_A, \frac{s_A}{c_{4,nA}})$ and $N(\bar{x}_B, \frac{s_B}{c_{4,nB}})$, respectively, where $c_{4,n}$ 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 Γ(.) stands for Gamma function[55]. In addition, the estimated distribution of $Y = X_A - X_B$ is also a normal distribution

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

Then, the exceedance probability of $X_A \geq X_B$ is given by[56]

$$EP_{X_A \geq X_B} = Pr\left(Z \geq -e^{'}\right) = 1 - \Phi\left(-e^{'}\right) = \Phi\left(e^{'}\right), \tag{10}$$

where *e'* is given by

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

The exceedance probability of $X_B \geq X_A$ is given by

$$EP_{X_B \geq X_A} = Pr\left(Z \geq e^{'}\right) = 1 - \Phi\left(e^{'}\right) = \Phi\left(-e^{'}\right). \tag{12}$$

Furthermore, the net superiority probability (NSP), denoted by ξ, is related to the exceedance probabilities as[56]

$$\xi = EP_{X_A \geq X_B} - EP_{X_B \geq X_A}. \tag{13}$$

Although Eq. (13) is based on the normality assumption, it is considered as a general definition of the NSP for any types of distributions of X_A and X_B[56].

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, similar to the observed ES or RES, the EP or NSP is resistant to manipulation via "*N*-chasing."

The above probabilistic analyses rely on probability distributions. However, these analyses can also be performed without assuming specific distributions, in what is known as nonparametric comparison of two groups. In the distribution-free analysis, the exceedance probability of $A \geq B$ ($EP_{A \geq B}$) is given by[56]

$$EP_{A \geq B} = \frac{U_{A \geq B}}{n_A n_B}, \tag{14}$$

And the exceedance probability of $B \geq A$ ($EP_{B \geq A}$) is given by

$$EP_{B\geq A} = \frac{U_{B\geq A}}{n_A n_B}\ , \tag{15}$$

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 given by

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

It should be noted that the concept of exceedance probability (EP) is essentially equivalent to the gain-probability (G-P) proposed by Trafimow et al.[18][19]. Moreover, EP and its analysis have been used across various engineering fields. For instance, the U.S. EPA[57] established a probabilistic chronic toxics standard of EP = 0.0037 to protect aquatic life. Di Toro[58] conducted EP analysis of river quality affected by runoff. Huang and Fergen[59] applied EP analysis to assess river BOD (biochemical oxygen demand) and DO (dissolved oxygen) concentrations in response to point load. Krishnamoorthy et al. [60] also utilized EP analysis to assess exposure levels in work environments. Furthermore, the concept of exceedance probability is closely related to the term "return period" commonly used in hydraulic engineering and hydrology. For example, a 100-year flood corresponds to an exceedance probability of 1%. Therefore, practitioners in engineering fields are typically more familiar with the concept of exceedance probability than with terms like the CLES, AUC, or A.

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

Zaiontz[61] considered the following problem: a marketing research firm conducted experiments to evaluate the effectiveness of a new flavoring for a beverage. In the study, eleven people in Group A1 and ten people in Group A2 tasted the beverage with the new flavoring, while ten people in Group B tasted the beverage with the old favoring. After tasting, all participants took a questionnaire to evaluate how enjoyable the beverage was. The scores obtained for the new flavoring (Group A1 and Group A2) and old flavoring (Group B) are shown in Table 1, and the corresponding sample means and standard deviations are presented in Table 2.

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 1. Scores 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	3.59	12.27	4.33

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

Zaiontz[61] applied the two-sample t-test (two-tailed) to compare the effectiveness of a new flavoring versus the old flavoring. For the comparison between Group A1 (new flavoring) and Group B (old flavoring), he obtained a p-value of 0.04, which led him to reject the null hypothesis at the $\alpha = 0.05$

doi.org/10.32388/JVYEJZ.2

level and conclude that the new flavoring was significantly more enjoyable. However, for the comparison between Group A2 (new flavoring) and Group B (old flavoring), the p-value was 0.05773, and he could not reject the null hypothesis. It is peculiar that Zaiontz[61] did not address or comment on these contradictory results from the two t-tests.

We examined this example using the advanced estimation statistics. Table 3 shows the estimated effect sizes and their reliability measures. Table 4 shows the results of the probabilistic analysis based on the distribution-based comparison, while Table 5 shows the results based on the nonparametric comparison.

Statistic	Comparison between Group A1 and Group B	Comparison between Group A2 and Group B
Observed effect size (ES): Eq. (1)	$\Delta = 3.81$	$\Delta = 8.70$
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 3. Estimated effect sizes and their reliability measures for the comparison of beverage flavoring

	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) $(A \geq B)$: Eq. (10)	$EP_{X_A \geq X_B} = 0.745$	$EP_{X_A \geq X_B} = 0.742$
Exceedance probability (EP) $(B \geq A)$: Eq. (12)	$EP_{X_B \geq X_A} = 0.255$	$EP_{X_B \geq X_A} = 0.258$
Net superiority probability (NSP): Eq. (13)	$\xi = 0.490$	$\xi = 0.484$

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

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

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

As can be seen from Table 3, the observed ES is 3.81 and the RES is 28.52% for the comparison of Group A1 versus Group B, while the observed ES is 8.70 and the RES is 72.12% for the comparison of Group A2 versus Group B. Our domain knowledge and common sense in this case suggests that the difference

between the two flavorings is practically significant. Although the RSUs are large (45.84% and 47.31%) due to the small sample sizes, the SNRs are high (4.76 and 4.47), and the SCIs are also high (0.83 and 0.82). These values indicate that the observed ES are reliable and that 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, with $Y \sim N(3.81,\ 5.78)$ for the former and $Y \sim N(8.70,\ 13.38)$ for the latter. However, the difference in the values of the RSU, SNR, SCI, EP, and NSP between these two comparisons are not significant. Thus, the two comparisons should lead to the same conclusion: the new flavoring is superior to the old flavoring.

Note that the values of the EP and NSP from the distribution-based comparison (Table 4) are consistent with those obtained from the nonparametric comparison (Table 5). $EP_{X_A \geq X_B}$ = 0.745, 0.742, and NSP=0.490, 0.484 based on the distribution-based comparison, while $EP_{A \geq B}$ =0.741, 0.725, and NSP=0.482, 0.450 based on the nonparametric comparison.

5. Why should the *t*-interval method for calculating measurement uncertainty 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[27]. It is referred to as the *t*-based uncertainty. In this section, we discuss two main issues with the *t*-interval and *t*-based uncertainty: rationale and methodology, which together explain why the *t*-interval method for calculating measurement uncertainty should be abandoned. We also examine problems associated with the *t*-distribution, which is the basis for 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 calculating measurement uncertainty is based on the concept of "coverage". Coverage, expressed as the confidence level or coverage probability, is the central concept in Neyman confidence interval theory[62][63]. However, it is important to note that the confidence level is not a probability in the strict mathematical sense; rather, it represents the "long-term success rate"[64] or "capture rate"[65]. In Monte Carlo simulation of the *t*-interval, the success or capture rate asymptotically approaches the nominal confidence level $(1 - \alpha)$. That is,

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

where *m* is the total number of simulated *t*-intervals and *k* is the number of the intervals that capture the true value μ.

Therefore, strictly speaking, the confidence level is not a *mathematical probability* that satisfies Kohnogorov's axioms of probability calculus; rather, it is a relative frequency. However, as Bunge[66] noted, "... frequencies alone do not warrant inferences to probabilities ..." 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[66] further argued that, "... 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 the "coverage" (the *frequency* of "success" or "capture") is a property of the confidence interval procedure (e.g. the *t*-interval procedure). This coverage can only be realized in the long run through repeated sampling or simulation; it is meaningless for a confidence interval computed from a single sample.

We must distinguish between the *result* of a procedure and the *coverage* of the procedure. In measurement uncertainty analysis, our focus is on the estimated uncertainty given by the procedure. As Kempthorne[67] stated, "...a statistical method should be judged by the result which it gives in practice." However, the concept of coverage does not represent a result produced by the method. Therefore, it is inappropriate and even paradoxical to judge an uncertainty estimation method by its coverage[68].

It should be emphasized that, a confidence interval procedure is merely a mechanism to generate a collection of intervals (or "sticks") with a stated capture rate for the unknown true value[65]. In other words, the *t*-interval method provides an "exact" answer to the 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 the wrong question for measurement uncertainty analysis. The purpose of measurement uncertainty analysis is to determine (or estimate) the measurement precision with a given sample. The correct question is: "How do we estimate measurement precision with a given sample?"[69]. In this context, the *t*-interval procedure is *not* an appropriate method for inferring measurement precision. Morey et al.[47] argued, "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 gives an "exact" answer to the wrong question[69].

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 given by $U_t = t_{\alpha/2}\frac{s}{\sqrt{n}}$ (the *t*-based uncertainty), where *n* is the number of observations, *s* is the sample standard deviation, and $t_{\alpha/2}$ is the *t*-score. In contrast, the true expanded uncertainty of the sample mean, assuming that the population standard deviation σ is known, is given by $U_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, which is called the *z*-based uncertainty, where $z_{\alpha/2}$ is the *z*-score. The *t*-based uncertainty artificially dilates the uncertainty. The artificial dilation can be quantified by a 'dilation factor', defined as the ratio between the expectation of the *t*-based uncertainty and the true expanded uncertainty. That is[65],

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

The dilation factor is extremely high when the sample size is small. For example, when *n*=2, the dilation factor is 5.17 for the nominal coverage probability $1-\alpha = 0.95$ and 19.72 for $1-\alpha = 0.99$. As the sample size increases, the dilation factor decreases significantly. At *n*=30, the dilation factor is only 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, which describes how the random uncertainty of the sample mean decreases as the sample size increases, i.e. in proportion to $1/\sqrt{n}$. In contrast, the expectation of the *t*-based uncertainty is given by $E(U_t) = c_{4,n}t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. For small sample sizes, the expected *t*-based uncertainty significantly deviates from the -1/2 power law as illustrated in Figure 1. Therefore, the *t*-based uncertainty or the *t*-interval acts as a distorted mirror of the physical reality.

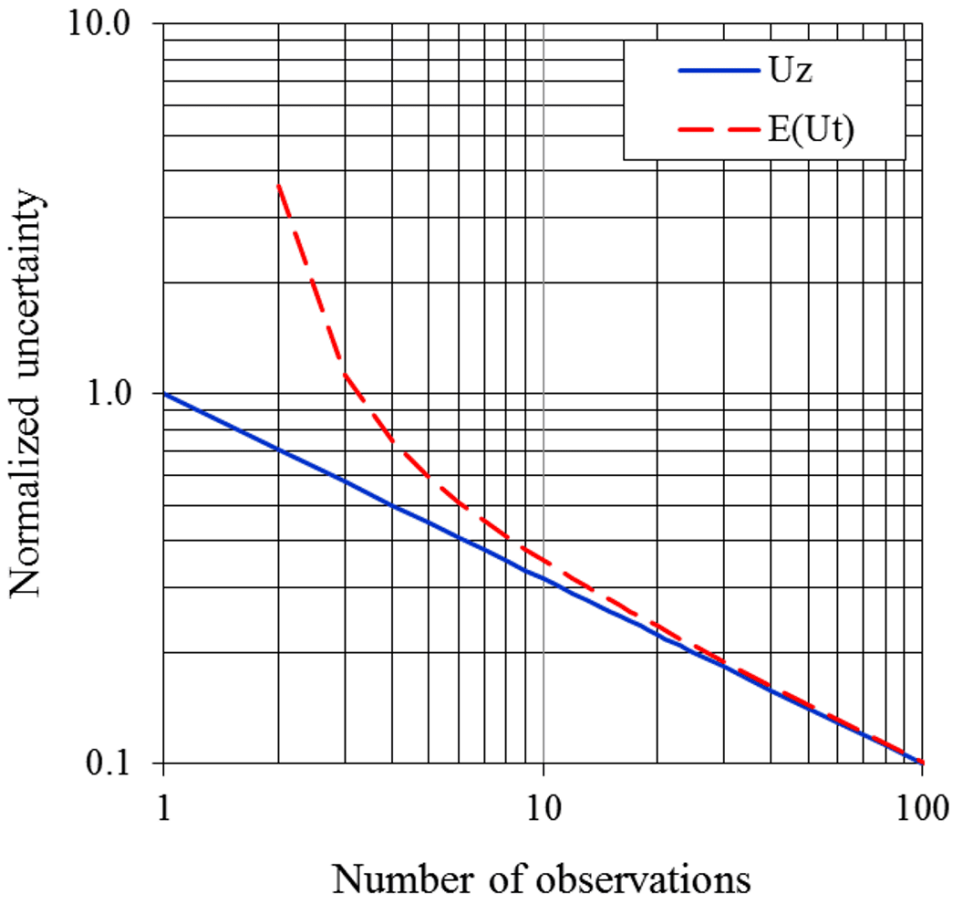

Figure 1. U_z and $E(U_t)$ (normalized by $z_{\alpha/2}\sigma$ at 1-α=0.95) on the log-log scales[65]

It is worth noting that prior to Student (William Sealy Gosset), the expanded uncertainty (referred to as the "probable error" in Student's 1908 paper) was calculated using the maximum-likelihood estimate of the population variance. This approach significantly underestimates the uncertainty when the sample size is small, with relative biases of -43.6%, -20.2%, -7.7% at n=2, 4, and 10, respectively. To correct for this underestimation, Student[70] invented the t-distribution. However, the t-based uncertainty U_t, derived from the t-distribution, leads to an overestimation of the uncertainty, as evidenced by the dilation factor. Interestingly, Ziliak and McCloskey[71] remarked that "Student used his t-tables a teensy bit..." and noted that "We have learned recently, by the way, that "Student" himself—William Sealy Gosset—did not rely on Student's t in his own work."

doi.org/10.32388/JVYEJZ.2

5.3. Issues with the t-distribution

The *t*-interval and *t*-based uncertainty are constructed using the *t*-distribution. Therefore, the methodological issues associated with the *t*-interval and *t*-based uncertainty must ultimately be traced back to the *t*-distribution, or its non-standard version: the scaled and shifted *t*-distribution (referred to as the location-scale *t*-distribution in Wikipedia).

First, the *t*-distribution is subject to what Huang[29] termed the "*t*-transformation distortion." The *t*-statistic is computed as the ratio between the sample error and the standard error of the sample mean, which transforms the original ε-s space $\Omega(\varepsilon, s)$ into a distorted *t*-space $\Omega(t)$. The *t*-transformation itself is mathematically valid, and thus the *t*-distribution is also mathematically sound. However, the inferences made using the *t*-distribution (such as constructing *t*-intervals) may not be valid because the inferences are actually performed in the distorted *t*-space $\Omega(t)$[69]. To illustrate, consider that plums are dried to make prunes. The drying process alters the shape and characteristics of the plums, which is analogous to the "*t*-transformation distortion". Just as one cannot accurately infer the original shape of plums from the dried prunes, we cannot reliably infer the true properties of the original data from inferences made in the distorted *t*-space $\Omega(t)$.

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 approximates the normal distribution (or the scaled and shifted *z*-distribution), regardless of the original distribution. The Central Limit Theorem does not support using the scaled and shifted *t*-distribution. Moreover, among three candidate distributions: the scaled and shifted *t*-distribution, the scaled and shifted *z*-distribution, and the Laplace distribution, Huang[72] demonstrated that the scaled and shifted *z*-distribution is the best choice according to the minimum entropy criterion, while[73] demonstrated that it is also the best according to the maximum informity criterion. The informity metric is the counterpart of the entropy metric; it can be used as an alternative measure to assess distributions[73]. In summary, the Central Limit Theorem, the entropy metric, and the informity metric, all support the use of the scaled and shifted *z*-distribution instead of the scaled and shifted *t*-distribution. There is no mathematical or physical principle that justifies the use of the *t*-distribution or its scaled and shifted version for this purpose.

It is worth mentioning that the statistics textbook by Matloff[74] does not even cover the *t*-distribution and *t*-intervals. In fact, Matloff[75] stated, "I advocate skipping the *t*-distribution, and

going directly to inference based on the Central Limit Theorem." This perspective further emphasizes the argument that inference should be based on the more robust and intuitive foundation provided by the Central Limit Theorem, rather than relying on the *t*-distribution.

6. Alternative to the *t*-interval method for calculating measurement uncertainty: 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 can be regarded as the true precision. When σ is unknown, however, the true precision cannot be determined. In such cases, the purpose of uncertainty analysis is to estimate the true precision based on the available sample data. According to the theory of point estimation, when σ is unknown, it can be replaced by a sample-based estimator $\widehat{\sigma}$. Accordingly, U_z can be replaced by a sample-based estimator $\widehat{U}$. We want $\widehat{U}$ to equal U_z on average, meaning that $\widehat{U}$ should be an unbiased estimator of U_z. Note that $s/c_{4,n}$ is an unbiased estimator of σ. Thus, $\widehat{U} = z_{\alpha/2}\frac{s}{c_{4,n}\sqrt{n}}$ is an unbiased estimator of U_z. This unbiased estimator $\widehat{U}$ conforms to the -1/2 power law.

Hirschauer[76] 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 "uncertainty paradox" caused by the *t*-interval method disappears when using the unbiased estimation method. For the "carpenter's laugh" scenario described by D'Agostini[32] (mentioned in the introduction), the

t-score at the nominal coverage probability (1-α)=0.999 is 636.62 due to severe t-transformation distortion at n=2, leading to the absurd uncertainty estimate U_t = 95 mm . In contrast, 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 estimation method gives $\widehat{U}$= 0.62 mm, which is far more realistic. Moreover, unlike the t-based uncertainty U_t, which is unsuitable for measurement quality control due to its high false rejection rate, the unbiased estimator $\widehat{U}$ can be reliably used for measurement quality control. Importantly, this unbiased estimation method has been adopted in the ISO standard for streamflow measurements with acoustic Doppler current profiler[77].

It should be emphasized that the unbiased estimation method is based on the theory of point estimation and the unbiasedness criterion. Unlike the t-interval method, which is an interval procedure based on confidence interval theory and the "coverage" criterion, the unbiased estimation method is not designed to generate intervals that capture the true value at a specified long-term success rate. These two approaches 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 by the long-term success or capture rate that is commonly used to evaluate confidence interval procedures[65][78].

Statistics textbooks often claim that interval estimation is more informative than point estimation. However, this claim can be misleading. Suppose we employ a statistical distribution model (e.g. normal distribution) in our analysis. If the model parameters are obtained by a valid method (such as maximum likelihood) applied to a given dataset, we can derive an estimated distribution. This estimated distribution, in turn, enables us to construct any probability interval we desire. For example, with n observations, the sample mean $\bar{y}$ and the unbiased standard error $\frac{s}{c_{4,n}\sqrt{n}}$ serve as estimates of the location and scale parameters, respectively. Then, the estimated sampling distribution of the sample mean $\overline{Y}$ is $N(\bar{y}, \frac{s^2}{c_{4,n}^2 n})$[79]. This complete estimated distribution is inherently more informative than any single confidence interval constructed from it, as it provides a full probabilistic description of the uncertainty surrounding the parameter estimate rather than a mere interval estimate.

The unbiased estimation method has been extended to cases involving multiple uncertainty components in measurement uncertainty analysis. This extension is referred to as the WS-z

approach[34]. The WS-z approach resolves the Ballico paradox that arises from the WS-t approach (mentioned in the introduction), by providing more realistic and consistent uncertainty estimates.

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 unknown variance, and *Y* is normally distributed with mean 0 and variance σ_Y. We have *n* observations from *X*:{x_1, x_2, ..x_n} and one observation from *Y*:{*y*}. Then, $Z = \overline{X} + Y$ is the estimator for *Z*=*X*+*Y* and the variance of *Z* is given by

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

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 $z = \overline{x} + y$. According to the unbiased estimation method (i.e. the WS-z approach), the expanded uncertainty is given by

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

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 problem considered, the Welch-Satterthwaite formula can be written as[34]

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

The expanded uncertainty given by the WS-t approach is[34]

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

To obtain numerical results for comparison, we assume that $s_X = 3$, *n*=4, and α=0.05, while σ_Y varies from 0 to 3. Under these assumptions, Figure 2 shows the expanded uncertainty estimates produced by the WS-z and WS-t approaches.

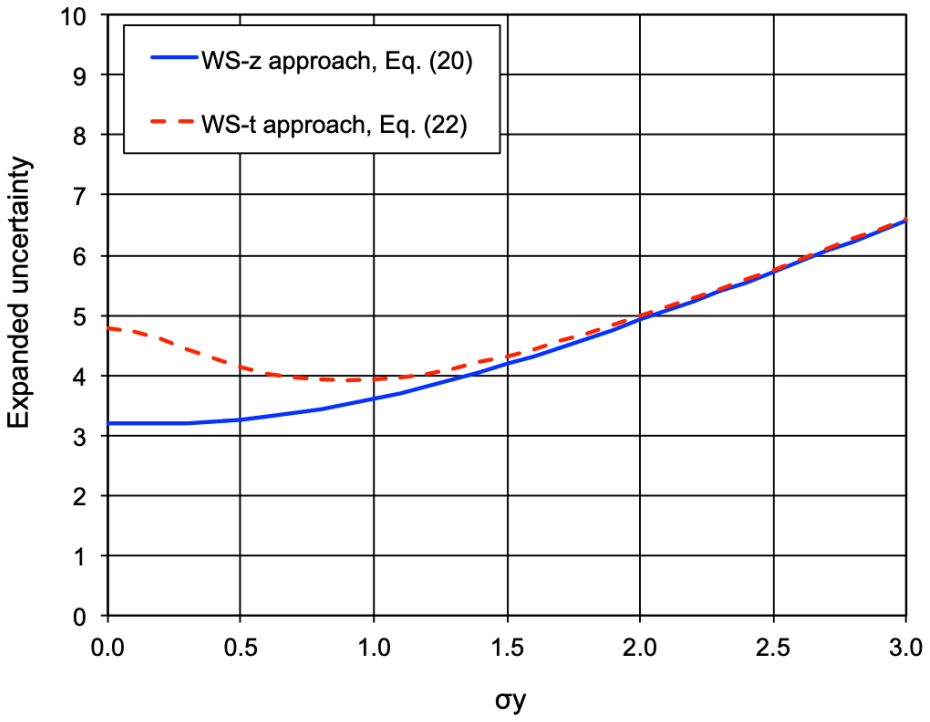

Figure 2. Expanded uncertainty estimated by the WS-*z* and WS-*t* approaches ($s_X = 3$, *n*=4, and α=0.05)

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 as σ_Y increases, which conforms with our domain knowledge and common-sense regarding measurement uncertainty. In contrast, the WS-*t* approach provides unrealistic estimates: it not only overestimates uncertainty when σ_Y is small (dilates the uncertainty), but it also exhibits paradoxical behavior: the uncertainty decreases as σ_Y increases in the range from 0 to 0.9. Notably, the WS-t uncertainty estimate $\widehat{U}_{WS-t}$ converges to the WS-z estimate $\widehat{U}_{WS-z}$ only when σ_Y becomes large. This is expected because when σ_Y is large, the contribution from σ_Y dominates over $\sqrt{s_X^2/n}$. This example clearly demonstrates that the WS-*t* approach, or the *t*-interval method for calculating measurement uncertainty, is inherently flawed.

doi.org/10.32388/JVYEJZ.2

7. Conclusion and recommendation

According to Jaynes[80], a paradox is "something which is absurd or logically contradictory, but which appears at first glance to be the result of sound reasoning." He further explained that "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 represent such paradoxes. Statistics textbooks, journals, and software packages have played a significant role in disseminating these paradoxical methods. As Hurlbert et al. [81] pointed out, "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, to implement statistics reform, statistics textbooks and software packages should be updated to reflect the paradigm shift from significance testing to estimation statistics. The author agrees with Hurlbert et al.[81], who argued that "... 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 calculating measurement uncertainty (both of which are problematic statistical methods) should be removed from textbooks and software packages. In contrast, good statistical methods such as the least squares method and maximum likelihood estimation should withstand statistics reform.

The advanced estimation statistics should be used in place of the two-sample *t*-test for comparing two groups. This approach involves considering multiple statistics, including the observed 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), which collectively extract and reveal the evidence embedded in the data from various perspectives. Importantly, we do not advocate for setting fixed thresholds on any of these statistics to make inferences. Instead, scientific inferences should be made based on domain knowledge while considering the comprehensive information provided by these statistics.

The unbiased estimation method should be used in place of the *t*-interval method for calculating measurement uncertainty. When employing the unbiased estimation method, both the "uncertainty

paradox" and the Ballico paradox (which are inherent to the t-interval method) disappear, leading to more realistic and reliable uncertainty estimates.

The author believes that the success of statistics reform depends on collaboration between statisticians and practitioners. It is hoped that this paper will stimulate discussion on statistics reform and encourage joint efforts to improve statistical practices.

References

1. ^ *Nuzzo R (2014). "Scientific method: Statistical errors". Nature. 506: 150–152. doi:10.1038/506150a.*
2. ^ *Siegfried T 2010 Odds Are, It's wrong: science fails to face the shortcomings of statistics Science News 1 77 26 https:// www.sciencenews.org/article/odds-are-its-wrong*
3. ^ *Siegfried T 2014 To make science better, watch out for statistical flaws Science News Context Blog, Feb ruary 7, https://www.sciencenews.org/blog/context/make-science-betterwatch-out-statistical-flaws*
4. ^ *Amrhein V, Greenland S, McShane B (2019). "Retire statistical significance". Nature. 567: 305-307.*
5. ^ *McShane BB, Gal D, Gelman A, Robert CP, Tackett JL (2018). Abandon statistical significance. The Ame rican Statistician. 73. doi:10.1080/00031305.2018.1527253.*
6. ^ *Halsey L G (2019). "The reign of the p-value is over: what alternative analyses could we employ to fill t he power vacuum?". Biology Letters. 15(5): 20190174. doi:10.1098/rsbl.2019.0174.*
7. ^ *Wasserstein R L, 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.*
8. ^ *Wasserstein R L, Schirm A L, Lazar N A 2019 Moving to a world beyond "p < 0.05" The American Statis tician 73:sup1 1-19 DOI: 10.1080/00031305.2019.1583913*
9. ^ *Trafimow D, Marks M 2015 Editorial Basic and Applied Social Psychology 37 1-2*
10. ^ *Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas H L J (2011). "Why psychologists must chang e the way they analyze their data: The case of psi: Comment on Bem". Journal of Personality and Social Psychology. 100: 426–432. doi:10.1037/a0022790.*
11. ^ *Haig B D (2016). "Tests of statistical significance made sound". Educational and Psychological Measur ement. 77: 489–506. doi:10.1177/0013164416667981.*
12. ^ *Colling L J, Szűcs D (2021). "Statistical Inference and the Replication Crisis". Review of Philosophy and Psychology. 12: 121–147. doi:10.1007/s13164-018-0421-4.*

13. ^Cumming G (2014). "The new statistics: why and how". Psychological Science. 25(1): 7-29. doi:10.1177/0956797613504966.
14. ^Cumming G, Calin-Jageman R (2024). Introduction to the New Statistics Estimation, Open Science, and Beyond. 2nd edition. ISBN 9780367531508. Routledge.
15. ^Normile CJ, Bloesch EK, Davoli CC, Scherr KC (2019). Introducing the new statistics in the classroom. Scholarship of Teaching and Learning in Psychology. 5(2): 162–168. doi:10.1037/stl0000141.
16. ^Claridge-Chang A, Assam P (2016). "Estimation statistics should replace significance testing". Nat Methods. 13: 108–109. doi:10.1038/nmeth.3729.
17. ^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. 24(3): 129-133. doi:10.1080/21679169.2022.2073991.
18. a, bTrafimow D, Hyman M R, Kostyk A, Wang Z, Tong T, Wang T, Wang C (2022). "Gain-probability diagrams in consumer research". International Journal of Market Research. 64(4): 470–483. doi:10.1177/14707853221085509.
19. a, bTrafimow D, Tong T, Wang T, Choy S T B, Hu L, Chen X, Wang C, Wang Z 2024 Improving Inferential Analyses Pre-Data and Post-Data Psychological Methods (to be published)
20. ^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, Kafadar K (2021). "ASA President's Task Force Statement on Statistical Significance and Replicability". Harvard Data Science Review. 3(3). doi:10.1162/99608f92.foado287.
21. ^Hand D J (2022). "Trustworthiness of Statistical Inference". Journal of the Royal Statistical Society Series A: Statistics in Society. 185(1): 329–347. doi:10.1111/rssa.12752.
22. ^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. doi:10.1093/ptj/pzac118.
23. a, bHalsey L, Curran-Everett D, Vowler S et al. (2015). "The fickle P value generates irreproducible results". Nat Methods. 12: 179–185. doi:10.1038/nmeth.3288.
24. ^Heckelei T, Hüttel S, Odening M, Rommel J (2023). "The p-value debate and statistical (Mal) practice–implications for the agricultural and food economics community". German Journal of Agricultural Econ

omics. 72(1): 47–67. doi:10.30430/gjae.2023.0231.

25. ^*Aurbacher J, Bahrs E, Banse M, Hess S, Hirsch S, Hüttel S, Latacz-Lohmann U, Mußhoff O, Odening M, Teuber R (2024). "Comments on the p-value debate and good statistical practice". German Journal of A gricultural Economics. 73(1): 1–3.*
26. ^*Berner D, Amrhein V (2022). "Why and how we should join the shift from significance testing to estim ation". J Evol Biol. 35(6): 777–787. doi: 10.1111/jeb.14009. PMID: 35582935; PMCID: PMC9322409. http s://onlinelibrary.wiley.com/doi/10.1111/jeb.14009.*
27. a, b*Joint Committee for Guides in Metrology (JCGM) (2008). Evaluation of Measurement Data – Guide t o the Expression of Uncertainty in Measurement (GUM 1995 with minor corrections). Sevres, France.*
28. ^*Huang H (2010). "A paradox in measurement uncertainty analysis 'Global Measurement: Economy & Technology' 1970 – 2010 Proceedings (DVD) (Measurement Science Conference)".*
29. a, b*Huang H (2018a). "Uncertainty estimation with a small number of measurements, Part I: new insigh ts on the t-interval method and its limitations". Measurement Science and Technology. 29. doi:10.1088/ 1361-6501/aa96c7.*
30. ^*Huang H (2014). "Uncertainty-based measurement quality control". Accred Qual Assur. 19: 65–73.*
31. ^*Jenkins JD (2007). The Student's t-distribution uncovered. Measurement Science Conference Proceedi ngs. Long Beach.*
32. a, b*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 (Alcosse bre, Spain, May 30th–June 4th).*
33. ^*Ballico M (2000). "Limitations of the Welch–Satterthwaite approximation for measurement uncertain ty calculations". Metrologia. 37: 61–64.*
34. a, b, c, d*Huang H (2016). "On the Welch–Satterthwaite formula for uncertainty estimation: a paradox an d its resolution". Cal Lab the International Journal of Metrology. 23: 20–28.*
35. ^*Ziliak S T, McCloskey D N (2008). The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press. doi:10.3998/mpub.186351.*
36. ^*Trafimow D (2023). "The Story of My Journey Away from Significance Testing". A World Scientific Enc yclopedia of Business Storytelling, pp. 95–127. doi:10.1142/9789811280948_0006.*
37. ^*Fisher R A (1956). Statistical Methods and Scientific Inference. Edinbrgh: Oliver and Boyd.*
38. a, b, c*Huang H (2022). Exceedance probability analysis: a practical and effective alternative to t-tests. Jo urnal of Probability and Statistical Science. 20(1): 80–97.*

39. ^Lazzeroni LC, Lu Y, Belitskaya-Lévy I (2016). Solutions for quantifying P-value uncertainty and replication power. Nature Methods. 13(2): 107-108.

40. ^Bonovas S, Piovani D (2023). "On p-Values and Statistical Significance". Journal of Clinical Medicine. 12(3): 900. doi:10.3390/jcm12030900.

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

42. a, bBaguley T (2009). "Standardized or simple effect size: what should be reported?". Br J Psychol. 100(Pt 3): 603-17. doi: 10.1348/000712608X377117. PMID: 19017432.

43. ^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.

44. a, b, cHuang H (2019). 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. doi:10.1088/1361-6501/ab46fd.

45. ^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.

46. ^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/.

47. a, bMorey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J (2016). 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.

48. ^Morey RD, Hoekstra R, Rouder JN, Wagenmakers E-J (2016). 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.

49. ^Trafimow D (2018). "Confidence intervals, precision and confounding". New Ideas in Psychology. 50: 48-53. doi:10.1016/j.newideapsych.2018.04.005.

50. ^McGraw KO, Wong SP (1992). A common language effect size statistic. Psychological Bulletin. 111(2): 361–365. doi:10.1037/0033-2909.111.2.361.

51. ^Vargha A, 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.

52. ^ Grissom R J, Kim J J (2001). "Review of assumptions and problems in the appropriate conceptualizatio n of effect size". Psychol Methods. 6(2): 135-46. doi: 10.1037/1082-989x.6.2.135. PMID: 11411438.

53. ^ Delaney H D, Vargha A (2002). "Comparing several robust tests of stochastic equality with ordinally sc aled variables and small to moderate sized samples". Psychol Methods. 7(4): 485-503. doi: 10.1037/108 2-989x.7.4.485. PMID: 12530705.

54. ^ Ruscio J, Mullen T (2012). "Confidence intervals for the probability of superiority effect size measure a nd the area under a receiver operating characteristic curve". Multivariate Behavioral Research. 47(2): 2 01–223. doi:10.1080/00273171.2012.658329.

55. ^ Wadsworth Jr H M 1989 Summarization and interpretation of data Handbook of Statistical Methods fo r Engineers and Scientists 2.1-2.21, Ed. Harrison M Wadsworth Jr. McGRAW-HILL In

56. a, b, c, d Huang H (2023). Probability of net superiority for comparing two groups or group means. Lobac hevskii Journal of Mathematics. 44(11): 42-54.

57. ^ Environment protection agency (EPA) (1991). Technical support document for water quality-based tox ics control, Office of Water, Washington, DC, EPA/505/2-90-001.

58. ^ Di Toro D M (1984). "Probability model of stream quality due to runoff". Journal of Environmental Eng ineering. ASCE. 110(3): 607-628.

59. ^ Huang H, Fergen RE (1995). Probability-domain simulation - A new probabilistic method for water q uality modeling. WEF Specialty Conference "Toxic Substances in Water Environments: Assessment and C ontrol" (Cincinnati, Ohio, May 14-17, 1995).

60. ^ Krishnamoorthy K, Mathew T, Ramachandran G (2007). Upper limits for exceedance probabilities und er the one-way random effects model. The Annals of Occupational Hygiene. 51(4): 397-406. doi:10.109 3/annhyg/mem013.

61. a, b, c Zaiontz C 2020 Two sample t test: unequal variances Real Statistics Using Excel https://real-statist ics.com/students-t-distribution/two-sample-t-test-uequal-variances/ accessed on August 22, 2023

62. ^ Neyman J (1935). On the problem of confidence intervals. Ann Math Stat. 6: 111-116.

63. ^ Neyman J (1937). Outline of a theory of statistical estimation based on the classical theory of probabilit y. Philos Trans R Soc Lond. A236: 333-380.

64. ^ Willink R 2010 Probability, belief and success rate: comments on 'On the meaning of coverage probabil ities' Metrologia 47 343–346

65. a, b, c, d, e Huang H (2018). More on the t-interval method and mean-unbiased estimator for measurem ent uncertainty estimation. Cal Lab the International Journal of Metrology. 25: 24-33.

66. ^a, ^b*Bunge M (1981). "Four concepts of probability". Applied Mathematical Modelling. 5(5): 306–312.*

67. ^*Kempthorne O (1976). Comments on paper by Dr. E. T. Jaynes 'Confidence intervals vs Bayesian interva ls'. 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).*

68. ^*Huang H (2018). A unified theory of measurement errors and uncertainties. Meas Sci Technol. 29:1250 03. https://doi.org/10.1088/1361-6501/aae50f*

69. ^a, ^b, ^c*Huang H (2018). Uncertainty estimation with a small number of measurements, Part II: a redefinit ion of uncertainty and an estimator method. Measurement Science and Technology. 29: 015005. doi:10. 1088/1361-6501/aa96d8.*

70. ^*Student (William Sealy Gosset) 1908 The probable error of a mean Biometrika VI 1–25*

71. ^*Ziliak S T, McCloskey D N (2004). "Significance redux". The Journal of Socio-Economics. 33(5): 665–6 75. doi:10.1016/j.socec.2004.09.038.*

72. ^*Huang H (2023). A minimum entropy criterion for distribution selection for measurement uncertainty analysis. Measurement Science and Technology. 35(3): 035014. doi:10.1088/1361-6501/ad1476.*

73. ^a, ^b*Huang H (2023). The theory of informity: a new probabilistic information framework (preprint). Res earchGate. doi:10.13140/RG.2.2.28832.97287.*

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

75. ^*Matloff N (2014). 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/.*

76. ^*Hirschauer N (2022). "Some thoughts about statistical inference in the 21st century". SocArXiv. Decem ber 20. doi:10.31235/osf.io/exdfg.*

77. ^*International Organization of Standards (ISO) (2021). ISO:24578:2021(E), Hydrometry — Acoustic Do ppler profiler — Method and application for measurement of flow in open channels from a moving boa t, first edition, 2021-3, Geneva Switzerland.*

78. ^*Huang H (2020). Comparison of three approaches for computing measurement uncertainties. Measure ment. 163: 107923. doi:10.1016/j.measurement.2020.107923.*

79. ^*Huang H (2019). Why the scaled and shifted t-distribution should not be used in the Monte Carlo meth od for estimating measurement uncertainty? Measurement. 136:282–8. https://doi.org/10.1016/j.measu rement.2018.12.089*

80. ^*Jaynes ET (2003). Probability Theory: The Logic of Science. Cambridge University Press.*

81. a, b *Hurlbert SH, Levine RA, 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.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.