

This paper may be cited as: Huang H (2021) Exceedance probability analysis: a practical and effective alternative to *t*-tests, preprint, DOI: [10.13140/RG.2.2.27268.55687](https://doi.org/10.13140/RG.2.2.27268.55687), available on ResearchGate: [https://www.researchgate.net/publication/348692325\\_Exceedance\\_probability\\_analysis\\_a\\_practical\\_and\\_effective\\_alternative\\_to\\_t-tests](https://www.researchgate.net/publication/348692325_Exceedance_probability_analysis_a_practical_and_effective_alternative_to_t-tests)

## **Exceedance probability analysis: a practical and effective alternative to *t*-tests**

Hening Huang  
Teledyne RD Instruments (retired)  
14020 Stowe Drive, Poway, CA 92064, USA  
e-mail: [heninghuang1@gmail.com](mailto:heninghuang1@gmail.com)

### **Abstract**

This paper presents a practical and effective alternative to the traditional *t*-tests for (1) comparing a sample or a sample mean against a known mean (i.e. one-sample test) and (2) comparing two samples or two sample means (i.e. two-sample test). The proposed method is referred to as exceedance probability (EP) analysis. In one-sample tests, EP is defined as the probability with which a sample or a sample mean is greater than a known mean. In two-sample tests, EP is defined as the probability with which the difference between two samples or between two sample means is greater than a specified value (which is referred to as probabilistic effect size (PES)). The meaning of EP is clear without any confusion; a person even not trained in statistics can understand it. This paper also defines a new statistic, which is referred to as the relative mean effect size (RMES). RMES provides a true measure of the scientific significance (not the statistical significance) of the difference between two means. A case study of preference between two manufacturers is presented to demonstrate the effectiveness of the proposed EP analysis, compared with four existing methods: *t*-tests, common language effect size (CL) analysis, signal content index (SCI) analysis, and Bayesian analysis. Unlike these existing methods that require the assumption of normality, the proposed EP analysis can be performed with any type of distributions. The case study example is examined with a normal distribution model and a raised cosine distribution model. The former is solved with an analytical solution and the latter is solved with a numerical method known as probability domain simulation (PDS).

**Keywords:** alternative to *t*-tests, comparison of samples, effect size, exceedance probability

### **1. Introduction**

There has been a long-standing debate about the validity of *t*-tests (or significance testing in general) and the associated *p*-values in the statistics community for many years. Unlike many statistics textbooks in which *t*-tests and *t*-distribution are part of standard materials, Matloff (2014a) purposely excludes *t*-tests and *t*-distribution in his textbook. An international journal: *Basic and Applied Social Psychology* (BASP) has officially banned significance testing from

BASP since 2015 (Trafimow and Marks 2015). The American Statistician Association (ASA) made an official statement about statistical significance and  $p$ -values (Wasserstein and Lazar 2016). Some authors suggested retiring or abandoning statistical significance and  $p$ -values (e.g. Amrhein et al. 2019, McShane et al. 2018, Halsey 2019, Wasserstein et al. 2019).

It seems that “the reign of the  $p$ -value is over.” Then, an important question is, “what alternative analyses could we employ to fill the power vacuum?” (Halsey 2019). Trafimow and Marks (2015), the editors of BASP that has officially banned significance testing, stated, “... BASP will require strong descriptive statistics, including effect sizes.” The author (Huang 2020a) recently presented a new statistic that is referred to as signal content index (SCI). SCI is defined as the ratio between the signal energy and the total energy of signal and noise contained in observed values, based on the law of conservation of energy. The analysis of the SCI for the difference between two means provides an alternative to the traditional  $t$ -tests. However, the SCI for the difference between two means is a function of the  $t$  statistic. Consequently, a SCI value can be converted to a  $p$ -value resulted from a  $t$ -test and vice versa. In addition, like a  $p$ -value, the SCI is not an absolute measure of the difference between two means. It is an analogy to the heterogeneity index  $I^2$  that is not an absolute measure of the heterogeneity between studies in meta-analysis.

The mathematical basis of  $t$ -tests is the  $t$ -distribution. The  $t$  statistic is a transformed quantity, i.e. the ratio between the sample error and the sample standard deviation. The  $t$ -transformation itself is mathematically valid, and so is the  $t$ -distribution (Huang 2020b). However, the use of  $t$ -distribution for statistical inference may be invalid because of the  $t$ -transformation distortion (Huang 2018a). The  $t$ -transformation distortion is the root cause of extremely high  $t$ -scores when the sample size is very small (Huang 2018b). The  $t$ -based uncertainty is actually misused in measurement uncertainty analysis (Huang 2018c). D'Agostini (1998) also casted doubt on the use of the  $t$ -distribution as the “standard way” for handling small samples. Matloff (2014b) stated, in his blog post titled “Why are we still teaching  $t$ -tests?”, “A precise mathematical understanding of the concepts is crucial to good applications. But stat curricula are not realistic.” He used  $t$ -tests to illustrate and stated, “The  $t$ -test is an exemplar for the curricular ills in three separate senses ... I advocate skipping the  $t$ -distribution, and going directly to inference based on the Central Limit Theorem.” The author (Huang 2019) initiated a discussion with an analogous title “Why are we still teaching  $t$ -distribution?” on ResearchGate, suggesting to revisit all  $t$ -based inferences that may be questionable due to the  $t$ -transformation distortion.

This paper proposes a practical and effective alternative to the traditional  $t$ -tests, referred to as exceedance probability (EP) analysis. The concept of EP is not new in statistics. However, EP seems less known in statistical or scientific inference, although the concept of EP may be implicit in statistical significant testing such as  $t$ -tests. We focus on two tests that are often encountered in practice, (1) one-sample test: comparing a sample or a sample mean against a known mean, and (2) two-sample test: comparing two samples (groups) or two sample means.

In the following sections, section 2 briefly reviews the concept of exceedance probability. Section 3 deals with one-sample tests. Section 4 deals with two-sample tests. Section 5 presents discussion. Section 6 presents a case study: preference between two manufacturers, comparing the proposed EP analysis with four existing methods:  $t$ -tests, common language effect size (CL) analysis, signal content index (SCI) analysis, and Bayesian analysis. Section 7 discusses the case study example with the assumption of non-normal distributions, which is solved with a numerical method. Section 8 presents conclusion and recommendation.

## 2. The concept of exceedance probability (EP)

Consider a random variable  $X$  that has a continuous probability density function (PDF)  $p(x|\theta)$  with the support  $(-\infty, +\infty)$  (or other support), where  $\theta$  is a vector of parameters. For a normal distribution,  $\theta = (\mu, \sigma)$ , where  $\mu$  is the mean and  $\sigma$  is the standard deviation. Exceedance probability (EP) is referred to as the probability that a specified value (or limit), denoted by  $x_{EP}$ , will be exceeded. That is

$$EP(x_{EP}) = \Pr(X > x_{EP}) = \int_{x_{EP}}^{\infty} p(x|\theta)dx = 1 - \Pr(X \leq x_{EP}) \quad (1)$$

For example,  $x_{90}$  is equal to the value, if randomly drawing a huge number of samples thorough Monte Carlo simulation from the probability distribution of  $X$ , where 10% of the samples will be below  $x_{90}$  and 90% will be above  $x_{90}$ . For symmetrical distributions such as a normal distribution,  $x_{50}$  is equivalent to the mean value. In addition, if we specify  $x_{EP} = 0$ ,  $EP(0)$  is the probability that the  $x$  values are greater than 0.

When the model parameters  $\theta$  are known,  $EP(x_{EP})$  will be exact according to Eq. (1). When the model parameters  $\theta$  are unknown,  $\theta$  are replaced with their estimator  $\hat{\theta}$ . Then, Eq. (1) becomes

$$\widehat{EP}(x_{EP}) = \int_{x_{EP}}^{\infty} p(x|\hat{\theta})dx \quad (2)$$

where  $p(x|\hat{\theta})$  is the estimated PDF of  $X$  and  $\widehat{EP}(x_{EP})$  is the estimated EP. For a normal distribution,  $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ , where  $\hat{\mu}$  is an estimator of  $\mu$  and  $\hat{\sigma}$  is an estimator of  $\sigma$ .

The concept of EP has been used in different fields of engineering. For example, the EP analysis is a standard practice in the assessment of receiving water quality. This is because water quality criteria are often set in terms of a concentration level with the associated exceedance probability (or return period) (Huang and Fergen 1995). EPA (Environment protection agency) (1991) sets EP=0.0037 for chronic toxics to protect aquatic life. Di Toro (1984) performed the EP analysis for stream quality due to runoff. Huang and Fergen (1995) performed the EP analysis for BOD and DO concentration along a river due to a point load. The EP analysis is also used for assessing the exposure level in a work environment (Krishnamoorthy et al. 2007).

As mentioned in the introduction section, EP seems less known in statistical or scientific inference. However, the concept of EP may be implicit in statistical significant testing such as  $t$ -tests. In the author's opinion, the EP analysis is more straightforward, more informative, and much easier to understand than  $t$ -tests and the associated  $p$ -values.

## 3. One-sample tests

Suppose that a sample (dataset)  $X=\{x_1, x_2, \dots, x_n\}$  is randomly drawn from a normal distribution  $X : N(\mu, \sigma)$ . Neither  $\mu$  nor  $\sigma$  is known. The dataset gives the sample mean  $\bar{x}_D$  and the sample standard deviation  $s_D$  ("D" means that the sample statistic  $\bar{x}$  or  $s$  is conditioned on the dataset). We are interested in two problems: (a) assessing a sample  $X$  against a known mean  $\mu_0$ , and (b) assessing the sample mean  $\bar{X}$  against  $\mu_0$ .

In our analysis, the location parameter  $\mu$  is estimated with  $\bar{x}_D$  and the scale parameter  $\sigma$  is estimated with  $s_D/c_{4,n}$ , where  $c_{4,n}$  is the bias correction factor that depends on the number of

observations (i.e. the sample size  $n$ ). Accordingly,  $X: N(\bar{x}_D, \frac{s_D}{c_{4,n}})$  is the estimated probability distribution function (PDF)  $X: N(\mu, \sigma)$  given the dataset.

For problem (a), the difference between a sample  $X$  and the known mean  $\mu_0$  is defined as the effect size:  $\Delta X = X - \mu_0$ .  $\Delta X$  is a random variable because  $X$  is a random variable. We are interested in the probability that  $\Delta X > 0$ , or  $X > \mu_0$ , i.e. the exceedance probability (EP) of  $X$  against  $\mu_0$ , denoted by  $EP_a(\mu_0)$ . It is written as

$$EP_a(\mu_0) = \Pr(\Delta X > 0) = \Pr(X > \mu_0) = 1 - \Pr(X \leq \mu_0) \quad (3)$$

As a numerical example, assume that  $\bar{x}_D = 20$ ,  $\frac{s_D}{c_{4,n}} = 5$ , and  $n=25$ . Then,  $EP_a(\mu_0)$  is 25% at  $\mu_0 = 23.372$ , 50% at  $\mu_0 = \bar{x}_D = 20$ , and 75% at  $\mu_0 = 16.628$ . Thus,  $EP_a(\mu_0)$  provides a probabilistic measure for assessing  $X$  against  $\mu_0$ , or a probabilistic assessment of the effect size  $\Delta X = X - \mu_0$ .

Note that the scale parameter in  $X: N(\bar{x}_D, \frac{s_D}{c_{4,n}})$  is only a weak function of the sample size  $n$  because  $c_{4,n}$  approaches unity quickly with increasing  $n$ , say  $n>10$ . Therefore,  $EP_a(\mu_0)$  is nearly independent of  $n$ .

For problem (b), the difference between the sample mean  $\bar{X}$  and the known mean  $\mu_0$  is defined as the effect size:  $\Delta \bar{X} = \bar{X} - \mu_0$ . Similar to problem (a), the EP of  $\bar{X}$  against  $\mu_0$ , denoted by  $EP_b(\mu_0)$ , is written as

$$EP_b(\mu_0) = \Pr(\Delta \bar{X} > 0) = \Pr(\bar{X} > \mu_0) = 1 - \Pr(\bar{X} \leq \mu_0) \quad (4)$$

The estimated PDF of the sample mean is  $\bar{X}: N(\bar{x}_D, \frac{s_D}{c_{4,n}\sqrt{n}})$ , given the dataset.

For the same assumed values:  $\bar{x}_D = 20$ ,  $\frac{s_D}{c_{4,n}} = 5$ , and  $n=25$ ,  $EP_b(\mu_0)$  is 25% at  $\mu_0 = 20.674$ , 50% at  $\mu_0 = \bar{x}_D = 20$ , and 75% at  $\mu_0 = 19.326$ . Thus,  $EP_b(\mu_0)$  provides a probabilistic measure for assessing  $\bar{X}$  against  $\mu_0$  or a probabilistic assessment of the effect size  $\Delta \bar{X} = \bar{X} - \mu_0$ .

Note that the scale parameter in  $\bar{X}: N(\bar{x}_D, \frac{s_D}{c_{4,n}\sqrt{n}})$  is a function of the sample size  $n$ , so  $EP_b(\mu_0)$  depends on  $n$ . Also note that, in general,  $EP_a(\mu_0) \neq EP_b(\mu_0)$ . However, they take the same value, 50%, at  $\mu_0 = \bar{x}_D$ . That is  $EP_a(\bar{x}_D) = EP_b(\bar{x}_D) = 50\%$ .

In addition, we define the ratio between  $\bar{x}_D - \mu_0$  and  $\mu_0$  as the relative mean effect size (RMES) of a one-sample test

$$RMES_{\text{one-sample}} = \frac{\bar{x}_D - \mu_0}{\mu_0} \quad (5)$$

$RMES_{\text{one-sample}}$  provides a true measure of the scientific significance (not the statistical significance) of the difference between the two means. Importantly, RMES does not depend on the sample size  $n$ .

#### 4. Two-sample tests

Suppose that two samples (dataests)  $X_1 = \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$  and  $X_2 = \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$  are randomly drawn from two independent normal distributions  $X_1 : N(\mu_1, \sigma_1)$  and  $X_2 : N(\mu_2, \sigma_2)$ , respectively. Both  $\mu_1$  and  $\mu_2$  are unknown, neither  $\sigma_1$  nor  $\sigma_2$ , and in general  $\sigma_1 \neq \sigma_2$ . The two datasets give the sample means  $\bar{x}_{1,D}$  and  $\bar{x}_{2,D}$ , and the sample standard deviations  $s_{1,D}$  and  $s_{2,D}$ , respectively. The sample sizes are  $n_1$  and  $n_2$  respectively. Again, we are interested in two problems: (a) assessing the difference between the two samples  $X_1$  and  $X_2$ , and (b) assessing the difference between the two sample means  $\bar{X}_1$  and  $\bar{X}_2$ .

Similar to the analysis in the one-sample test, the estimated distributions of  $X_1$  and  $X_2$  are  $X_1 : N(\mu_1, \sigma_1) \approx X_1 : N(\bar{x}_{1,D}, \frac{s_{1,D}}{c_{4,n_1}})$  and  $X_2 : N(\mu_2, \sigma_2) \approx X_2 : N(\bar{x}_{2,D}, \frac{s_{2,D}}{c_{4,n_2}})$ , respectively.

For problem (a), the difference between  $X_1$  and  $X_2$  is defined as the effect size:  $\Delta X = X_1 - X_2$ .  $\Delta X$  is a random variable because  $X_1$  and  $X_2$  are random variables. The estimated PDF of  $\Delta X$  is also normal and is written as

$$p(\Delta X) = N\left(\left(\bar{x}_{1,D} - \bar{x}_{2,D}\right), \sqrt{\left(\frac{s_{1,D}}{c_{4,n_1}}\right)^2 + \left(\frac{s_{2,D}}{c_{4,n_2}}\right)^2}\right) \quad (6)$$

We are interested in the probability that  $\Delta X > 0$ , or  $X_1 > X_2$ , i.e. the EP of  $\Delta X$  against 0, denoted by  $EP_a(0)$ . It is written as

$$EP_a(0) = \Pr(\Delta X > 0) = \Pr(X_1 > X_2) \quad (7)$$

$EP_a(0)$  is the estimated probability that a sample  $X_1$  from one distribution (or group) is greater than a sample  $X_2$  from another distribution (or group). A person even not trained in statistics can understand this statement.

Furthermore, we define the EP of  $\Delta X$  against a specified value  $\Delta x_{EP}$  as

$$EP_a(\Delta x_{EP}) = \Pr(\Delta X > \Delta x_{EP}) = 1 - \Pr(\Delta X \leq \Delta x_{EP}) \quad (8)$$

Note that  $EP_a(0)$  is a special case where  $\Delta x_{EP} = 0$ .

Because the distribution of  $\Delta X$  is symmetric,  $EP_a(\Delta x_{50} = \bar{x}_{1,D} - \bar{x}_{2,D})$  means that 50% of the  $\Delta x$  samples will be greater than  $\Delta x_{50}$ . In other words,  $X_1$  is greater (or smaller) than  $X_2$  by  $\Delta x_{50}$  at the odds of 1:1. On the other hand,  $EP_a(\Delta x_{75})$  means that 75% of the  $\Delta x$  samples will be greater than  $\Delta x_{75}$ , or  $X_1$  is greater than  $X_2$  by  $\Delta x_{75}$  at the odds of 3:1. Thus,  $\Delta x_{EP}$  provides a probabilistic measure of the effect size  $\Delta X = X_1 - X_2$ , the difference between two samples (groups). It is therefore referred to as the probabilistic effect size (PES).

Note that the scale parameter in  $p(\Delta X)$  is only a weak function of the sample sizes  $n_1$  and  $n_2$  because  $c_{4,n}$  approaches unity quickly with increasing  $n$ , say  $n > 10$ . Therefore,  $EP_a(0)$  or  $EP_a(\Delta x_{EP})$  is nearly independent of the sample sizes  $n_1$  and  $n_2$ .

For problem (b), the difference between the two sample means  $\bar{X}_1$  and  $\bar{X}_2$  is defined as the effect size:  $\Delta \bar{X} = \bar{X}_1 - \bar{X}_2$ . The estimated PDF of  $\Delta \bar{X}$  is also normal and is written as

$$p(\Delta\bar{X}) = N \left( (\bar{x}_{1,D} - \bar{x}_{2,D}), \sqrt{\left(\frac{s_{1,D}}{c_{4,n_1\sqrt{n_1}}}\right)^2 + \left(\frac{s_{2,D}}{c_{4,n_2\sqrt{n_2}}}\right)^2} \right) \quad (9)$$

Similar to problem (a), the EP of  $\Delta\bar{X}$  against 0 (i.e.  $\bar{X}_1 > \bar{X}_2$ ), denoted by  $EP_b(0)$ , is written as

$$EP_b(0) = \Pr(\Delta\bar{X} > 0) = \Pr(\bar{X}_1 > \bar{X}_2) \quad (10)$$

Also similarly, the EP of  $\Delta\bar{X}$  against a specified value  $\Delta\bar{x}_{EP}$  is written as

$$EP_b(\Delta\bar{x}_{EP}) = \Pr(\Delta\bar{X} > \Delta\bar{x}_{EP}) = 1 - \Pr(\Delta\bar{X} \leq \Delta\bar{x}_{EP}) \quad (11)$$

Analogously,  $\Delta\bar{x}_{EP}$  provides a probabilistic measure of the effect size  $\Delta\bar{X} = \bar{X}_1 - \bar{X}_2$ , the difference between two sample means (group means). Accordingly,  $\Delta\bar{x}_{EP}$  is also a probabilistic effect size (PES). We have  $\Delta x_{50} = \Delta\bar{x}_{50} = \bar{x}_{1,D} - \bar{x}_{2,D}$ .

Note that the scale parameter in  $p(\Delta\bar{X})$  is a function of the sample sizes  $n_1$  and  $n_2$ . Therefore,  $EP_b(0)$  and  $EP_b(\Delta\bar{x}_{EP})$  depend on the sample sizes  $n_1$  and  $n_2$ . Also note that, in general,  $EP_a(\Delta x_{EP}) \neq EP_b(\Delta\bar{x}_{EP})$ , except that both are 50% at  $\Delta x_{50} = \Delta\bar{x}_{50} = \bar{x}_{1,D} - \bar{x}_{2,D}$ .

In addition, we define the ratio between  $\Delta x_{50}$  (or  $\Delta\bar{x}_{50}$ ) and a weighted-average of the two sample means as the relative mean effect size (RMES) of a two-sample test

$$RMES_{two-sample} = \frac{\bar{x}_{1,D} - \bar{x}_{2,D}}{\bar{x}_w} \quad (12)$$

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

$$\bar{x}_w = \frac{\frac{n_1\bar{x}_{1,D}}{s_{1,D}^2} + \frac{n_2\bar{x}_{2,D}}{s_{2,D}^2}}{\frac{n_1}{s_{1,D}^2} + \frac{n_2}{s_{2,D}^2}} \quad (13)$$

$RMES_{two-sample}$  provides a true measure of the scientific significance (not the statistical significance) of the difference between the two sample means. Note that if  $s_{2,D}^2/n_2$  goes to zero,  $RMES_{two-sample}$  reduces to  $RMES_{one-sample}$  with  $\bar{x}_{D,1}$  replaced by  $\bar{x}_D$  and  $\bar{x}_{2,D}$  replaced by  $\mu_0$ .

It should be pointed out that the EP analysis is not limited to normal distributions. In fact, the EP analysis does not require the assumption of normality. The PDF of  $X$  in a one-sample test, or the PDFs of  $X_1$  and  $X_2$  in a two-sample test can be any type of distributions. For example,  $X_1$  can be normally distributed and  $X_2$  can be uniformly distributed; Eqs. (7) and (8) still apply. However, in this situation, a numerical procedure may be needed to generate a solution. A numerical method known as probability domain simulation (PDS) is described in Appendix. The use of PDS for a case study example is presented later in this paper.

## 5. Discussion

### 5.1. Comparison with the analysis of common language effect size (CL)

In the problem (a) of two-sample tests, the EP of  $\Delta X$  against zero, i.e.  $EP_a(0) = \Pr(X_1 > X_2)$ , is the probability that a sample  $X_1$  from one distribution (or group) is greater than a sample  $X_2$  from another distribution (or group). Thus, the meaning of  $EP_a(0)$  is the same as the meaning of an effect size statistic known as common language effect size (CL) proposed by McGraw and Wong (1992). CL may be under other names such as the probability of superiority (PS), area under the receiver operating characteristic (AUC), or A for its nonparametric version (Ruscio and Mullen 2012). However, the formulation of  $EP_a(0)$  is different from that of CL.  $EP_a(0)$  is calculated based on the estimated PDF of  $\Delta X$ , Eq. (6); it does not require the assumption of normality. In principle,  $EP_a(0)$  can be calculated from the difference between two distributions of any types. In contrast, the calculation of CL requires the assumption of normality. CL is calculated based on the standardized (mean) effect size that is equivalent to a 'z-score' of a standard normal distribution (Coe 2002). For two independent samples, the z-score for CL is written as (Coe 2002)

$$z_{\text{CL}} = \frac{\bar{x}_{1,D} - \bar{x}_{2,D}}{\sqrt{s_{1,D}^2 + s_{2,D}^2}} \quad (14)$$

Therefore, CL is the upper tail probability associated with  $z_{\text{CL}}$  on a table of the normal cumulative distribution. Note that  $z_{\text{CL}}$  is different from the usual z-score for a z-test, in which the standard error, e.g.  $s_1/\sqrt{n_1}$ , is used, instead of the sample standard deviation  $s_1$ .

If both  $X_1$  and  $X_2$  are normally distributed, and the variances  $\sigma_1^2$  and  $\sigma_2^2$  are known or the sample sizes  $n_1$  and  $n_2$  are very large, the numerical value of  $EP_a(0)$  will be the same or approximately the same as CL. If  $s_{1,D}^2$  and  $s_{2,D}^2$  are estimated with small samples,  $EP_a(0)$  will be slightly different from CL. This can be seen in the case study example presented later in this paper.

### 5.2. Comparison with the z-test and t-test

It is important to note that the z-test or t-test and the associated  $p$ -value are applicable to problem (b) only. They are not applicable to problem (a).

There is a relationship between  $EP_b(\mu_0)$  and the one-tailed  $p$ -value resulted from a one-sample z-test or t-test; there is also a relationship between  $EP_b(0)$  and the one-tailed  $p$ -value resulted from a two-sample z-test or t-test.

For a one-sample z-test, the one-tailed  $p$ -value for the null: the effect is greater than zero, can be calculated as

$$p_{\text{one-tailed}} = \Pr\left(z = \frac{\bar{x} - \bar{x}_D}{\frac{\sigma}{\sqrt{n}}} < -z_p = -\frac{\bar{x}_D - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = \Pr(\bar{X} < \mu_0) = 1 - EP_b(\mu_0) \quad (15)$$

For a two-sample equal-variance z-test, the one-tailed  $p$ -value for the null: the effect is greater than zero, can be calculated as

$$\begin{aligned}
p_{\text{one-tailed}} &= \Pr \left( z = \frac{(\bar{X}_1 - \bar{X}_2) - (\bar{x}_{1,D} - \bar{x}_{2,D})}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_p = -\frac{\bar{x}_{1,D} - \bar{x}_{2,D}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) \\
&= \Pr((\bar{X}_1 - \bar{X}_2) < 0) = \Pr(\bar{X}_1 < \bar{X}_2) = 1 - EP_b(0)
\end{aligned} \tag{16}$$

It can be readily demonstrated that these results are the same for the one-tailed  $p$ -values for the null: the effect is less than zero.

Since the  $t$ -test approaches the  $z$ -test when the sample size is large (say,  $n>30$ ), the  $p$ -value resulted from a one-tailed  $t$ -test will be approximately equal to  $1 - EP_b(\mu_0)$  for a one-sample test and will be approximately equal to  $1 - EP_b(0)$  for a two-sample test.

The EP analysis for problem (b) is superior to the one-tailed  $t$ -test. First,  $EP_b(\mu_0)$  or  $EP_b(0)$  has a clear mathematical meaning without any confusion. For example,  $EP_b(0) = \Pr(\bar{X}_1 > \bar{X}_2)$  is the estimated probability that a sample mean  $\bar{X}_1$  from one distribution (or group) is greater than a sample mean  $\bar{X}_2$  from another distribution (or group). A person even not trained in statistics can understand this statement.  $EP_b(\mu_0)$  or  $EP_b(0)$  provides a probabilistic assessment of the effect size  $\bar{X}_1 - \bar{X}_2$ . However, like a  $p$ -value,  $EP_b(\mu_0)$  or  $EP_b(0)$  does not provide an absolute measure of the effect size. We therefore do not suggest any cut-off value of  $EP_b(\mu_0)$  or  $EP_b(0)$  for assessing statistical significance. We do not use the term “statistical significance” in the EP analysis at all. Instead, we suggest using  $EP_b(\mu_0)$  or  $EP_b(0)$  in conjunction with RMES for scientific inference. In contrast, the meaning of a  $p$ -value is often misinterpreted or misunderstood. Particularly, statisticians tend to interpret evidence dichotomously based on whether or not a  $p$ -value crosses the conventional 0.05 threshold for statistical significance (McShane and Gal 2017). For example,  $p>0.05$  is often misinterpreted as “the probability that the null hypothesis is true” or “ $p>0.05$  means that no effect was observed.” However, statistical significance is not equivalent to scientific significance. Moreover, since the  $p$ -value decreases with increasing the sample size and approaches zero as sample size goes to infinity, statistical significance will be always achieved with sufficiently large samples in  $t$ -tests, even if the absolute difference between two means (i.e. the mean effect size) or RMES is very small and meaningless.

Second, the calculation of  $EP_b(\mu_0)$  or  $EP_b(0)$  does not require the normality assumption. It does not require the assumption of equal variance or using a pooled variance either. In principle, the EP analysis can be performed for any type of distributions. In contrast, the calculation of  $p$ -values in a  $t$ -test requires the normality assumption. It also requires the assumption of equal variance or using a pooled variance. Therefore, the EP analysis for problem (a) has less restriction and less limitation than  $t$ -tests;  $EP_b(\mu_0)$  or  $EP_b(0)$  is more accurate (or has less uncertainty) than  $p$ -values.

### 5.3. The Bayesian view

It should be pointed out that the proposed EP analysis is based on the frequentist view of model parameters. That is, the unknown parameters  $\mu$  and  $\sigma$  are considered and treated as constants that are estimated with the point estimation method with the mean-unbiased criterion. In the Bayesian view, however, the unknown parameters  $\mu$  and  $\sigma$  are treated as random variables. The author (Huang 2020d) recently presented a new Bayesian method based on a rule of

transformation between the frequentist view and the Bayesian view and demonstrated that the frequentist sampling distribution, estimated with a given dataset, is virtually the same as the Bayesian probability distribution of an unknown parameter (e.g.  $\mu$ ) in the case that prior information is not involved. The author (Huang 2020d) also demonstrated that, in the light of the law of aggregation of information (LAI) (Huang 2020c) and the frequentist-Bayesian transformation rule, the frequentist and Bayesian inference are virtually equivalent so they can be unified, at least in measurement uncertainty analysis. Thus, the EP analysis may also be performed based on the Bayesian view. Take the problem (b) of two-sample tests as an example. According to the frequentist-Bayesian transformation rule,  $\bar{X}_1 \rightarrow \mu_1$ , and  $\bar{X}_2 \rightarrow \mu_2$ . Consequently,  $\Delta\bar{X} \rightarrow \Delta\mu$ , the PDF of  $\Delta\mu$  is the same as Eq. (9). Furthermore, Eq. (10) and Eq. (11) also apply to  $\Delta\mu$ . However, the Bayesian view may not be applicable to the problem (a) of two-sample tests that may have to be dealt with the frequentist sampling theory only.

The interested reader is referred to Huang (2020c) for the law of aggregation of information (LAI) and Huang (2020d) for detailed discussion on the new Bayesian method and the potential unification of frequentist and Bayesian inference.

## 6. Case study: preference between two manufacturers

We consider a well-posted example that is originally given in a textbook of Roberts (1964). Two manufacturers, denoted by A and B, are suppliers for a component. We are concerned with the lifetime of the component and want to choose the manufacturer that affords the longer lifetime. Manufacturer A supplies 9 units for lifetime test. Manufacturer B supplies 4 units. The test data give the sample means 42 and 50 hours, and the sample standard deviations 7.48 and 6.87 hours, for the units of manufacturer A and B respectively.

Roberts (1964) discussed this example with a two-tailed  $t$ -test and concluded that, at the 90% level, the samples afford no significant evidence in favor of either manufacturer over the other. Jaynes (1976) discussed this example with a Bayesian analysis. He argued that our common sense tell us immediately, without any calculation, the test data constitutes fairly substantial (but not overwhelming) evidence in favor of manufacturer B. The author (Huang 2020a) recently discussed this example with a SCI analysis.

### 6.1 The $t$ -tests

In this study, we conducted the  $t$ -tests with the pooled variance and Welch's  $t$ -test. We considered both the one-tailed and two-tailed tests. Table 1 shows the results.

**Table 1.** Results of the  $t$ -tests

	$t$ -test with pooled variance	Welch's $t$ -test
Degrees of freedom	11	6.72
$t$ statistic	1.8436	1.9568
$p$ -value (one-tailed)	0.0462	0.0465
$p$ -value (two-tailed)	0.0923	0.0930

According to the dichotomous interpretation of evidence based on whether or not a  $p$ -value crosses the conventional 0.05 threshold for statistical significance, the estimated  $p$ -values from both of the two-tailed  $t$ -tests suggested that, at the 95% level, the samples afford no significant

evidence in favor of either manufacturer over the other. On the other hand, the one-tailed  $t$ -tests at the 95% level barely suggest significance. Apparently, the  $t$ -tests fail to extract evidence that is already clear to our unaided common sense that we should prefer manufacturer B.

The estimated  $p$ -value from a  $t$ -test depends on the sample sizes, or degrees of freedom. For this example, the  $p$ -values would be smaller if the sample sizes were greater than 9 and 4 for manufacturer A and B respectively. However, the mean effect size  $\bar{x}_{A,D} - \bar{x}_{B,D}$ , which is a true measure of scientific significance, does not depend on the sample sizes. While we certainly prefer large samples when making scientific inference, our decision should not be made based on the  $p$ -value that varies with the sample size.

## 6.2 The Bayesian analyses

In the Bayesian analysis of Jaynes (1976), the location parameters, i.e. the unknown mean lifetime of manufacturer A's and manufacturer B's components, are treated as random variables and are denoted by  $a$  and  $b$  respectively. Jaynes (1976) calculated the probability that  $b > a$ , conditioned on all available data. That is

$$\Pr(b > a) = \int_{-\infty}^{\infty} da \int_a^{\infty} p(a)p(b) db \quad (17)$$

where  $p(a)$  is the posterior distribution of  $a$ , based on the sample of  $n=9$  items supplied by manufacturer A, and  $p(b)$  is the posterior distribution of  $b$ , based on the sample of  $n=4$  items supplied by manufacturer B. By using the Jeffreys' prior, Jaynes (1976) found that both  $p(a)$  and  $p(b)$  are the scaled and shifted  $t$ -distributions. The integration of Eq. (17) gives a probability of 92.0% or odds of 11.5:1 that manufacturer B's components have a greater mean lifetime, which conforms to the indication of common sense (Jaynes 1976).

However, the estimated probability with Eq. (17) may be questionable because of two issues in the Bayesian formulation for this example. The first issue is the use of the Jeffreys' prior. In fact, there has been a debate on the validity of Jeffreys priors among Bayesians. For example, D'Agostini (1998), a leading proponent of Bayesian methods in particle physics, argued that "...it is rarely the case that in physical situations the status of prior knowledge is equivalent to that expressed by the Jeffreys priors, ..." D'Agostini further stated, "The default use of Jeffreys priors is clearly unjustified, especially in inferring the parameters of the normal distribution, ...." Moreover, the author (Huang 2018b) revealed that the scaled and shifted  $t$ -distribution is a distorted sampling distribution due to the Bayesian 'transformation'.

The second issue is that the conventional Bayes Theorem in its conventional form: posterior  $\propto$  prior  $\times$  likelihood, is actually flawed. In fact, the conventional Bayes Theorem is a reformulated form of the original Bayes Theorem that is correct. The author (Huang 2020c) recently discovered that it is a mistake to use likelihood function in the conventional (or reformulated) Bayes Theorem. This mistake is the root cause of the inherent bias of the conventional Bayesian method.

These two issues can be solved with a new Bayesian method that is derived based on the law of aggregation of information (LAI) and the rule of transformation between the frequentist view and the Bayesian view (Huang 2020d). According to the new Bayesian method, the posterior distribution of  $a$  is written as

$$p(a) = N(a | \bar{x}_{A,D}, \frac{s_{A,D}}{c_{4,n_A} \sqrt{n_A}}) \quad (18)$$

where  $c_{4,n_A} = 0.9693$  at  $n_A = 9$ ,  $\bar{x}_{A,D} = 42$  hours, and  $s_{A,D} = 7.48$  hours.

The posterior distribution of  $b$  is written as

$$p(b) = N(b | \bar{x}_{B,D}, \frac{s_{B,D}}{c_{4,n_B} \sqrt{n_B}}) \quad (19)$$

where  $c_{4,n_B} = 0.9213$  at  $n_B = 4$ ,  $\bar{x}_{B,D} = 50$  hours, and  $s_{B,D} = 6.87$  hours.

Substituting Eqs. (18) and (19) into Eq. (17) yields a probability of 96.7% or odds of 29.1:1 that manufacturer B's components have a greater mean lifetime.

It is important to note that, the estimated probability or odds from both of the above Bayesian analyses is only for the *mean lifetime of the samples* with  $n=9$  and 4 for manufacturer A and B respectively; it is not for the *lifetime of individual units*. Apparently, the Bayesian analysis depends on the sample sizes. The estimated probability or odds would be large if the sample sizes were large.

It should be remarked that the two-sample one-tailed  $t$ -test is essentially equivalent to Jaynes' Bayesian analysis. According to the discussion in subsection 4.2,  $(1 - p_{\text{one-tailed}}) \approx \Pr(b > a)$ . The one-tailed  $t$ -test using the pooled variance gives a  $p$ -value of 0.0462, leading to  $(1 - p_{\text{one-tailed}}) \times 100 = 95.38\%$ . This result is comparable to Jaynes' Bayesian analysis result  $\Pr(b > a) = 92\%$ .

### 6.3 The SCI analysis

The signal content index (SCI) for the difference between two sample means is defined as (Huang 2020a)

$$\text{SCI} = 1 - \frac{1}{(\bar{x}_1 - \bar{x}_2)^2} \left( \frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2} \right) \quad (20)$$

For this example, 1 → A and 2 → B. Substituting the values that  $n_A = 9$ ,  $\bar{x}_{A,D} = 42$  hours,  $s_{A,D} = 7.48$  hours,  $n_B = 4$ ,  $\bar{x}_{B,D} = 50$  hours, and  $s_{B,D} = 6.87$  hours into Eq. (18) gives  $\text{SCI} = 0.74$ . This SCI value suggests that there is substantial difference between the mean lifetimes of two manufacturers' components.

In addition, the author (Huang 2020a) defines the sample signal energy (SSE) as

$$\text{SSE} = \bar{x}^2 - \frac{s^2}{n} \quad (21)$$

For this example, the mean lifetime is the signal. The SSE values are 1758 and 2490 (hour)<sup>2</sup> for manufacturer A's and B's sample means respectively. That is, manufacturer B's sample mean contains significantly greater signal energy than manufacturer A's sample mean does. Based on the SCI value and the SSE values, we should have a preference of manufacturer B.

However, like the  $t$ -tests and the Bayesian analyses, the SCI analysis also depends on the sample sizes. The SCI approaches to unity when the sample sizes are very large even if the mean effect size  $\bar{x}_1 - \bar{x}_2$  is small and insignificant. That is, the SCI is not an absolute measure of the mean effect size. Therefore, a scientific inference must consider both the SCI and SSE

values.

#### 6.4 The analysis of common language effect size (CL)

According to Eq. (14), the *z*-score  $z_{\text{CL}}$  is calculated as

$$z_{\text{CL}} = \frac{\bar{x}_{B,D} - \bar{x}_{A,D}}{\sqrt{s_{B,D}^2 + s_{A,D}^2}} = \frac{50 - 42}{\sqrt{6.87^2 + 7.84^2}} = 0.808 \quad (22)$$

The resulting CL is 0.791 (or 79.1%). That is, the lifetime of manufacturer B's components (individual units) is greater than the lifetime of manufacturer A's components (individual units) with a probability of 79.1% or at an odds of 3.8:1. According to this CL analysis, we should have a preference of manufacturer B.

#### 6.5 The exceedance probability (EP) analysis

It is important to note again that all of the above analyses, except for the CL analysis, are for the difference between the *two sample means*: the two mean lifetimes of the tested units. That is, these analyses consider the problem (b) of two-sample tests. Indeed, *t*-tests, Bayesian analyses, and SCI analysis, are for the problem (b) of two-sample tests only; they are not for the problem (a) of two-sample tests. However, this example should be considered as a problem (a) of two-sample tests because we are more concerned about the *lifetime of all individual units* in a group than the *mean lifetime of the group*. Therefore, we conducted the EP analysis for the effect size  $\Delta X$  in this study.

Under the normality assumption, the estimated PDF of manufacturer A's lifetime  $X_A$  (individual units) is written as

$$p(X_A) = N(\bar{x}_{A,D}, \frac{s_{A,D}}{c_{4,n_A}}) \quad (23)$$

Also under the normality assumption, the estimated PDF of manufacturer B's lifetime  $X_B$  (individual units) is written as

$$p(X_B) = N(\bar{x}_{B,D}, \frac{s_{B,D}}{c_{4,n_B}}) \quad (24)$$

Let  $\Delta X = X_B - X_A$ . The PDF of  $\Delta X$  is estimated as

$$p(\Delta X) = N\left(\left(\bar{x}_{B,D} - \bar{x}_{A,D}\right), \sqrt{\left(\frac{s_{A,D}}{c_{4,n_A}}\right)^2 + \left(\frac{s_{B,D}}{c_{4,n_B}}\right)^2}\right) = N(8, 10.44) \quad (25)$$

The probability that manufacturer B's components (individual units) has a greater lifetime, i.e.  $X_B > X_A$ , is calculated as

$$EP_a(0) = \Pr(X_B > X_A) = 77.8\% \quad (26)$$

In other words, the lifetime of manufacturer B's components (individual units) is greater than the lifetime of manufacturer A's components (individual units) at an odds of 3.5:1.

Note that  $EP_a(0) = \Pr(X_B > X_A) = 77.8\%$  and  $CL=79.1\%$ . The CL value is slightly greater than the  $EP_a(0) = \Pr(X_B > X_A)$  value. This is because the CL calculation does not account for the negative bias of the sample standard deviation when the sample size is small.

Moreover,  $\Delta x_{50} = 8$  hours and  $\Delta x_{75} = 0.958$  hours. That is, the lifetime of manufacturer B's components (individual units) is greater than the lifetime of manufacturer A's components (individual units) by 8 hours at the odds of 1:1, and by 0.958 hours at the odds of 3:1. Thus, we should have a preference of manufacturer B.

We also calculated the relative mean effect size (RMES) according to Eq. (12). The estimated RMES is 17.79%, which indicates that the mean (i.e. on average) lifetime of manufacturer B's components is greater than the mean lifetime of manufacturer A's component by 17.79%. This RMES value is considered to be "scientifically" or "practically" significant.

## 7. Calculating $EP_a(0)$ for the case study example with non-normal distributions

We have pointed out in section 4 that the EP analysis does not require the assumption of normality. The PDFs of  $X_1$  and  $X_2$  in a two-sample test can be any type of distributions. If  $X_1$  and/or  $X_2$  are not normally distributed, however, an analytical solution of  $p(\Delta X)$  may not be available so that a numerical procedure may be needed. Two numerical methods may be used. One is the well-known Monte Carlo simulation (MCS) and the other is known as probability domain simulation (PDS). Detailed discussion on PDS and a comparison with its counterpart: sampling-domain simulation, i.e. MCS, can be found in Huang and Fergen (1995). Appendix briefly describes the PDS for a two-dimensional problem.

In this section, we consider the case study example again, but assume that the lifetime of a manufacturer's component follows a raised cosine distribution. We present the PDS results for  $EP_a(0)$  to demonstrate the effectiveness of PDS for the EP analysis with non-normal distributions.

According to Castrup (2004), the PDF of a raised cosine distribution centered at zero is written as

$$p(x) = \begin{cases} \frac{1}{2\alpha} \left[ 1 + \cos \frac{\pi x}{\alpha} \right] & -\alpha \leq x < +\alpha \\ 0 & \text{elsewhere} \end{cases} \quad (27)$$

where  $\alpha$  is the bounding limit.

The standard deviation, denoted by  $\sigma$ , of the raised cosine distribution is written as (Castrup 2004)

$$\sigma = \frac{\alpha}{\sqrt{3}} \sqrt{1 - \frac{6}{\pi^2}} = 0.3615\alpha \quad (28)$$

Castrup (2004) discussed four candidate distributions with finite bounding limits for a measurand (or for a calibration error): uniform, triangular, quadratic, and raised cosine. Castrup (2004)

commented that the uniform distribution is not appropriate because it is not a physically credible distribution. The triangular distribution is not appropriate either because it displays abrupt transitions at the bounding limits and at the center, which are physically unrealistic. The quadratic distribution does not have a discontinuity at the center, but it rises abruptly at the bounding limits, which diminishes its physical validity. The raised cosine distribution overcomes all shortcomings of the other three distributions; it exhibits a central tendency and can be determined from minimum containment limits (Castrup 2004). Moreover, the shape of a raised cosine distribution looks similar to the shape of a normal distribution except that it has finite bounding limits. In fact, a raised cosine distribution is more reasonable than a normal distribution for describing a measurand (or calibration error) with finite bounding limits. A normal distribution has infinitely long tails that are physically unrealistic for the component lifetime of a manufacturer because the component lifetime must have a lower and an upper limit due to the quality control program at manufacturers. However, a raised cosine distribution is not as mathematically convenient as a normal distribution; it may not lead to an analytical solution in the EP analysis.

The lifetime of manufacturer A's component is centered at  $x=\bar{x}_{A,D}$  and is within the interval between the lower limit  $\bar{x}_{A,D} - \alpha_A$  and the upper limit  $\bar{x}_{A,D} + \alpha_A$ . According to Eq. (27), its PDF can be written as

$$p(x_A) = \begin{cases} \frac{1}{2\alpha_A} \left[ 1 + \cos \frac{\pi(x_A - \bar{x}_{A,D})}{\alpha_A} \right] & (\bar{x}_{A,D} - \alpha_A) \leq x_A < (\bar{x}_{A,D} + \alpha_A) \\ 0 & \text{elsewhere} \end{cases} \quad (29)$$

In order to be compatible with the normal distribution, we assume that  $\sigma_A = s_{A,D}/c_4$ . Thus,  $\alpha_A = s_{A,D}/(0.3615c_4) = 21.35$  hours.

Similarly, the PDF for the lifetime of manufacturer B's component is centered at  $x=\bar{x}_{B,D}$  with the lower and upper limits  $\bar{x}_{B,D} - \alpha_B$  and  $\bar{x}_{B,D} + \alpha_B$  respectively. It is written as

$$p(x_B) = \begin{cases} \frac{1}{2\alpha_B} \left[ 1 + \cos \frac{\pi(x_B - \bar{x}_{B,D})}{\alpha_B} \right] & (\bar{x}_{B,D} - \alpha_B) \leq x_B < (\bar{x}_{B,D} + \alpha_B) \\ 0 & \text{elsewhere} \end{cases} \quad (30)$$

where  $\alpha_B = s_{B,D}/(0.3615c_4) = 19.46$  hours.

Table 2 summarizes the parameter values for the raised cosine distributions of the lifetime of two manufacturers' components.

Table 2. Parameter values for the raised cosine distributions (units: hours)

Manufacturer	Center	Bounding limit	Lower limit	Upper limit
A	42	21.35	20.65	63.35
B	50	19.46	30.54	69.46

We implemented the two-dimensional PDS using an Excel spreadsheet. The range of  $x$  is divided into  $m=100$  intervals. Thus,  $\Delta x_A = 0.4270$  and  $\Delta x_B = 0.3892$ .

We are interested in  $EP_a(0) = \Pr(\Delta X > 0) = \Pr(X_B > X_A)$  only. The resulting  $EP_a(0)$  from the PDS with the raised cosine distributions is 77.1%. This value is compatible with  $EP_a(0) = 77.8\%$  resulted from the analytical solution with the normal distributions.

## 8. Conclusion and recommendation

The proposed exceedance probability (EP) analysis provides a probabilistic assessment of the effect size (e.g. the difference between two samples or two means). The meaning of EP is clear without any confusion; a person even not trained in statistics can understand it. The EP analysis, in conjunction with the proposed relative mean effect size (RMES), provides the basis for scientific inference.

The EP analysis for the problem (a) of two-sample tests under the assumption of normality is essentially the same as the CL (common language effect size) analysis. The EP analysis for the problem (b) of two-sample tests is essentially the same as a one-tailed  $z$ -test (comparing two sample means from two independent normal distributions with known standard deviations); it is approximately equal to the one-tailed  $z$ -test. However, the EP analysis is more straightforward, more informative, and much easier to understand than the  $t$ -test and the associated  $p$ -value. Moreover, unlike the CL analysis,  $z$ -tests, or  $t$ -tests, the EP analysis does not require the assumption of normality. In principle, the EP analysis can be performed for any type of distributions. In addition, the calculation of EP does not require the assumption of equal variance or using a pooled variance. Therefore, the EP analysis has less restriction or less limitation; it is more accurate (or has less uncertainty) than either the CL analysis or  $t$ -tests.

In the author's opinion, problem (a) makes much more sense than problem (b) in many real-world applications. It is important to note that the  $t$ -test and the associated  $p$ -value are applicable to problem (b) only. The  $p$ -value, which depends on the sample sizes, is misleading in statistical and scientific inference, because "significance", in terms of the conventional 0.05 threshold, can be always achieved with sufficiently large samples, even if the absolute difference between two means (i.e. the mean effect size) or RMES is very small and meaningless. Therefore, we suggest considering problem (a) whenever it applies and always reporting RMES. We do not suggest any threshold in the EP analysis. We do not use the term "statistical significance" in the EP analysis at all.

When dealing with non-normal distributions, two numerical methods may be used for the EP analysis: Monte Carlo simulation (MCS) and probability domain simulation (PDS). The case study of preference between two manufacturers has demonstrated the effectiveness of PDS with non-normal distributions.

### Appendix: Probability domain simulation (PDS)

Consider a random quantity  $Z$  that relates to a random vector  $\mathbf{X}$  in a general form

$$Z = f(\mathbf{X}) \quad (31)$$

where  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  and  $n$  is the number of input quantities.  $n$  is also referred to as the dimension of the problem.

PDS is based on the following proposition: for any value  $z$  of  $Z$ , the probability of  $P(Z=z)$  is equal to the sum of all probabilities that all  $[x]$  of  $[\mathbf{X}]$  satisfy  $z = f(\mathbf{x})$  (Huang and Fergen 1995). This proposition is an extension of the proposition for two special cases:  $Z = X + Y$  and  $Z = X/Y$  discussed in Berman (1969). The proposition can be demonstrated with the law of total probability and a rule for combining the number of ways in which events can occur (e.g. Huntsberger 1970).

We illustrate the algorithm of PDS, based on the proposition, for a two-dimensional problem  $\mathbf{X} = (X_1, X_2)$  and  $Z = f(X_1, X_2)$ . We assume that  $X_1$  and  $X_2$  are independent random variables having PDFs  $p_1(x_1)$  and  $p_2(x_2)$  respectively. The ranges of  $x_1$  and  $x_2$  are divided into  $m$  intervals of  $\Delta x_1$  and  $\Delta x_2$ , respectively. Let  $x'_{1,i}$  denote the realization of  $X_1$  within the interval  $(x_{1,i} - \frac{\Delta x_1}{2}, x_{1,i} + \frac{\Delta x_1}{2})$ , and analogously for  $x'_{2,j}$ . The output of  $Z = f(X_1, X_2)$  at  $X_1 = x'_{1,i}$  and  $X_2 = x'_{2,j}$ , denoted by  $z_{i,j}$ , is written as

$$z_{i,j} = f(x'_{1,i}, x'_{2,j}) \quad (32)$$

The occurrence probability of  $z_{i,j}$  is the probability that  $x'_{1,i}$  and  $x'_{2,j}$  occur simultaneously. It is written as

$$P(z_{i,j}) = P_1(x'_{1,i})P_2(x'_{2,j}) = p_1(x'_{1,i})p_2(x'_{2,j})\Delta x_1\Delta x_2 \quad (33)$$

Both Eqs. (32) and (33) need to be implemented for all combinations of  $x'_{1,i}$  and  $x'_{2,j}$ , ( $i, j = 1, 2, 3, \dots, m$ ). This will yields an  $m \times m$  matrix of  $z$  and an  $m \times m$  matrix of the associated probability  $P$ . The results need to be manipulated to obtain the PDF of  $Z$  or the exceedance probability (EP) of  $Z$  against a specified value. To obtain the PDF of  $Z$ , the range of  $z$  is divided into a number of intervals of  $\Delta z$ . Then, according to the proposition, the PDF of  $Z$  is estimated as

$$p(z_i) = \frac{1}{\Delta z} \sum_k P\left[(z_k - \frac{\Delta z}{2}) \leq z_i < (z_k + \frac{\Delta z}{2})\right] \quad (34)$$

where  $z_k = z_{i,j}$ .

The EP of  $Z$  against a specified value  $z_s$  is estimated as

$$\text{EP}(z_s) = \sum_k P(z_k > z_s) = \sum_{i=1}^m \sum_{j=1}^m P(z_{i,j} > z_s) \quad (35)$$

A two-dimensional PDS can be easily implemented using an Excel spreadsheet. A high-dimensional ( $n > 2$ ) PDS may need a computer program.

For the case study example considered in section 6 with the PDS,  $\mathbf{X} = (X_A, X_B)$ , and  $Z = f(X_A, X_B) = \Delta X = X_B - X_A$ . It is a two-dimensional problem.

In addition, it should be mentioned that the PDS algorithm naturally complies with the discretized formula of the Bayesian method (Huang 2020d). The Markov Chain Monte Carlo (MCMC) sampling is often employed to generate Bayesian posterior distributions. However, the MCMC method in general associates with computational difficulty and lack of transparency. PDS might be an effective alternative to the MCMC method.

## References

- Amrhein V, Greenland S, and McShane B 2019 Retire statistical significance *Nature* **567** 305–307  
 Berman S M 1969 *The elements of probability* Addison-Wesley Publishing Company

- Castrup H. (2004) Selecting and applying error distributions in uncertainty analysis, presented at the Measurement Science Conference, Anaheim, 2004.
- Coe R 2002 It's the Effect Size, Stupid What effect size is and why it is important *The Annual Conference of the British Educational Research Association*, University of Exeter, England, 12-14 September 2002 <https://dradamvolungis.files.wordpress.com/2012/01/its-the-effect-size-stupid-what-effect-size-is-why-it-is-important-coe-2002.pdf>
- D'Agostini G 1998 Jeffreys priors versus experienced physicist priors: arguments against objective Bayesian theory *Proceedings of the 6<sup>th</sup> Valencia International Meeting on Bayesian Statistics* (Alcossebre, Spain, May 30<sup>th</sup>-June 4<sup>th</sup>)
- Di Toro D M 1984 Probability model of stream quality due to runoff *Journal of Environmental Engineering ASCE* 110(3) 607-628.
- Environment protection agency (EPA) 1991 *Technical support document for water quality-based toxics control*, Office of Water, Washington, DC, EPA/505/2-90-001
- 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: 20190174 <https://doi.org/10.1098/rsbl.2019.0174>
- Huang H 2018a 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 2018b Uncertainty estimation with a small number of measurements, Part I: new insights on the *t*-interval method and its limitations *Meas. Sci. Technol.* 29(1) <https://doi.org/10.1088/1361-6501/aa96c7>
- Huang H 2018c Uncertainty estimation with a small number of measurements, Part II: a redefinition of uncertainty and an estimator method *Meas. Sci. Technol.* 29(1) <https://doi.org/10.1088/1361-6501/aa96d8>
- Huang 2019 Why are we still teach *t*-distribution? ResearchGate [https://www.researchgate.net/post/Why\\_are\\_we\\_still\\_teaching\\_t-distribution](https://www.researchgate.net/post/Why_are_we_still_teaching_t-distribution)
- Huang H 2020a 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(4) 045008, <https://doi.org/10.1088/1361-6501/ab46fd>
- Huang 2020b Comparison of three approaches for computing measurement uncertainties *Measurement* 163 <https://doi.org/10.1016/j.measurement.2020.107923>
- Huang H 2020c Two simple and practical methods for combining prior information with current measurement in uncertainty analysis *Cal Lab: the International Journal of Metrology* 27(3) 22-32 available on ResearchGate: [https://www.researchgate.net/publication/344502279\\_Two\\_simple\\_and\\_practical\\_methods\\_for\\_combining\\_prior\\_information\\_with\\_current\\_measurement\\_in\\_uncertainty\\_analysis](https://www.researchgate.net/publication/344502279_Two_simple_and_practical_methods_for_combining_prior_information_with_current_measurement_in_uncertainty_analysis)
- Huang H 2020d A new Bayesian method for measurement uncertainty analysis and the unification of frequentist and Bayesian inference, preprint, DOI: 10.13140/RG.2.2.35338.08646, available on ResearchGate: [https://www.researchgate.net/publication/344552280\\_A\\_new\\_Bayesian\\_method\\_for\\_measurement\\_uncertainty\\_analysis\\_and\\_the\\_unification\\_of\\_frequentist\\_and\\_Bayesian\\_inference?channel=doi&linkId=5f7fd8a5458515b7cf71d5ec&showFulltext=true](https://www.researchgate.net/publication/344552280_A_new_Bayesian_method_for_measurement_uncertainty_analysis_and_the_unification_of_frequentist_and_Bayesian_inference?channel=doi&linkId=5f7fd8a5458515b7cf71d5ec&showFulltext=true)
- 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), available on ResearchGate: [https://www.researchgate.net/publication/285589247\\_Probability-domain\\_simulation\\_-\\_A\\_new\\_probabilistic\\_method\\_for\\_water\\_quality\\_modeling](https://www.researchgate.net/publication/285589247_Probability-domain_simulation_-_A_new_probabilistic_method_for_water_quality_modeling)

- Huntsberger D V 1970 *Elements of statistical inference* second edition Allyn and Bacon Inc.
- Jaynes E T 1976 Confidence intervals vs Bayesian intervals in *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, eds. Harper and Hooker, Vol. II, 175-257, D. Reidel Publishing Company Dordrecht-Holland
- Krishnamoorthy K, Mathew T, 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
- 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 and Gal D 2017 Statistical Significance and the Dichotomization of Evidence *Journal of the American Statistical Association*, 112: 885-895  
DOI: [10.1080/01621459.2017.128984](https://doi.org/10.1080/01621459.2017.128984)
- 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
- Roberts N A 1964 *Mathematical Methods in Reliability Engineering* McGraw-Hill Book Co. Inc. New York
- 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>
- Trafimow D and Marks M 2015 Editorial *Basic and Applied Social Psychology* 37 1-2
- 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