Selecting the best product alternative in a sea of uncertainty

Reinout Heijungs

Received: 19 May 2020 / Accepted: 1 December 2020 / Published online: 2 February 2021
© The Author(s) 2021, corrected publication 2022

Abstract

Introduction Most LCA studies are comparative and to an increasing extent the effects uncertainty are included in LCA results. This raises the question how the best option from a set of product alternatives can be selected when the product scores are uncertain. The starting point of this article is a set of Monte Carlo results for a number of alternative products.

Indicators for single product alternatives First we discuss different ways of expressing results for product alternatives separately. This includes a discussion of centrality (mean, median, geometric mean, etc.) and dispersion (standard deviation, standard error, confidence interval, etc.).

Indicators of difference for two product alternatives A critical review of approaches to single out the superior option on case of a comparison of two is given. This includes familiar approaches such as t tests, but also lesser known ones such the Bhattacharyya coefficient and Cohen’s d. All approaches are defined, discussed, and illustrated with one consistent, downloadable, example.

More than two product alternatives The findings for two products are generalized for the multi-product situation. In particular, the issue of inflation of type I errors in multiple comparisons is discussed.

Discussion Two main questions are identified: (1) What is the probability that a randomly selected specimen of product A performs better than a randomly selected specimen of product B? (2) How much will a randomly selected specimen of product A perform better than a randomly selected specimen of product B? These two options can both be relevant, but existing approaches for distinguishing product alternatives address one of these two only, or they even turn out to answer a different, less relevant, question. A proposal for a new indicator that addresses both questions simultaneously is offered and its use is illustrated.

Keywords Comparative LCA · Uncertainty · Null hypothesis significance testing · Comparison index · Interpretation

1 Introduction

A substantial part of the LCA studies is comparative, comparing a set of alternative products in order to select the best one, or comparing a product with a proposed redesigned product, in order to check if the redesign is an improvement in terms of the environmental performance.

At the same time, it is common knowledge that LCA results are subject to uncertainty. A statement on the environmental preference of product A to product B must therefore be made with due acknowledgment of possible mistakes due to such uncertainty. After all, knowing the probability of making the wrong decision may affect the decision you make. The situation is comparable to those of weather forecasts: if the probability of rain is stated to be 30%, you probably will not have any rain, but still may decide to bring your umbrella. Uncertainty is an indispensable ingredient of decision-making (Lipschitz and Strauss 1997).

The last decade an increasing number of LCA studies report results (such as impact category results) with uncertainty information, for instance, standard deviations or ranges (Igos et al. 2019). In most cases, these uncertainty indications originate from Monte Carlo simulations (Lloyd...
and Ries 2007; Igos et al. 2019), which in turn are based on estimates of the uncertainties of the input data (Weidema et al. 2013). Several issues have been pointed out with that, such as the underestimation of the input uncertainties (Kuczynski 2019), the ignorance of correlated uncertainties (Groen and Heijungs 2017), and the problem of increased computation time (Peters 2007). In this article, all these points are assumed to be solved or otherwise unimportant, and it is just assumed that LCA results have been obtained with correct and complete uncertainty indication. Another issue is that most LCA studies address multiple impact categories, and therefore, a decision-maker is facing the problem of weighting or otherwise trading-off conflicting outcomes. That problem will also be assumed to be solved here, for instance, because a weighting scheme has been established, or because the LCA study focuses on one impact category only, such as the carbon footprint.

As a concrete example, let us suppose there are two alternative products, A and B, and that for both a carbon footprint has been calculated, say 100 with a standard deviation of 10 for product A, and 110 with a standard deviation of 10 for product B (to facilitate the discussion, units will be left out here). An understudied question is then: which product is better? In their comprehensive review of the uncertainty practice in LCA, Igos et al. (2019) do address comparative LCA, but they mainly see it in as a communication problem (“a proper LCA study should communicate all the different components of uncertainty treatment: the identification, qualitative characterization, and sensitivity analysis for non-comparative evaluation; as well as the quantitative characterization and uncertainty analysis for comparative study. The question of how to present them is still open to the practitioner.”) A similar remark applies to the recent framework by Heijungs et al. (2019). Also, in a recent critique on the use of statistical analyses in LCA (Von Brömssen and Röös 2020), the emphasis is on visualization and description, because those authors reject the use of more formal methods.

Most LCA studies these days look at the central value, 100 for product A and 110 for product B, and if the standard deviations are sufficiently narrow conclude that product A is better, because its impact is somehow “significantly” lower. But precise and convincing criteria for preferring one option to another option have hardly been discussed in the context of LCA. It is the main objective of this article to change this.

A few papers have addressed this question before. For instance, Gregory et al. (2016) explicitly focus on “robust comparative life cycle assessments incorporating uncertainty.” Likewise, Mendoza Beltrán et al. (2018) devote a paper to the question how to draw conclusions. But these papers have limitations. Gregory et al. (2016) basically rely on one approach, the comparison indicator (see below), while there are many more options. And Mendoza Beltrán et al. (2018) mix the uncertainty problem with the multi-criteria problem, so that they can devote only a part of the discussion to the uncertainty problem proper. This article does not discuss approaches for multi-criteria decision making, that also in some cases capture the problem of deciding on the basis of uncertain information, but always in a context of multiple criteria (Prado-Lopez et al. 2014). Here the emphasis is purely on single-criterion decision making, under uncertainty. Another restriction we make here is that the uncertainty calculations are made using the Monte Carlo approach, so not using one-at-a-time scenario variations, Taylor series, fuzzy numbers, or other approaches. Because Monte Carlo is by far the most-used approach (Lloyd and Ries 2007; Igos et al. 2019), this restriction is not problematic. Moreover, several of the insights obtained are also valid for other approaches.

In conclusion, the purpose of this article is to study methods to decide between probabilistic single-score product alternatives.

In the next two sections, comparisons of two product alternatives are discussed, first for separate products, and then involving one integrated metric. A brief generalization to three or more products is presented in Sect. 4. Section 5 discusses the findings and proposes a novel approach. The supplementary information provides details of many of the intermediate steps. For general statistical formulas and concepts, only a general reference is given here: Agresti and Franklin (2013), Rice (2007) and Zwilling and Kokoska (2000).

2 Indicators for single product alternatives

2.1 General introduction

Unfortunately, while comparative LCA is probably the most widely-used type of LCA, there are few texts that discuss the principles and terminology and that propose a mathematical notation. For instance, the ISO 14044 standard (ISO 2006) mentions the terms “compar*” (with * coding for “e,” “ative,” “ison,” etc.) not less than 44 times, but it gives no indication how to actually do comparisons. More extensive and detailed guides, such as the ILCD Handbook (ILCD 2010), provide not much more on this. Perhaps one of the few exceptions is Heijungs and Suh (2002), who devote a section (8.2.6) to “comparative analysis.” In the slipstream of this, Heijungs and Kleijn (2001) and Heijungs et al. (2005) discuss it as one of the key methods in life cycle interpretation. However, these texts primarily focus on comparative LCA without uncertainty.

Despite the lack of good guidance, many case studies actually are comparative, and they do, at least in some cases, include uncertainty information. So our discussion below
is based on a small number of theoretical treatments and a larger number of case studies. Because there is no standard for terminology and notation, we will rephrase the published work in a common language.

Before the different approaches are presented, a small hypothetical case will be introduced. The data and model equations have been entered in Excel and are available as supplementary information. The data is for two stochastic systems, A and B, featuring correlation and non-normality. From these populations, a Monte Carlo sample of size \( n = 1000 \) has been drawn. These \( 2 \times 1000 \) numbers are used to illustrate the various approaches.

We also need to introduce a notation. The Monte Carlo sampling yields two series of values, for product A (say, \( \mathbf{a} = (a_1, a_2, \ldots, a_n) \)) and for product B (say, \( \mathbf{b} = (b_1, b_2, \ldots, b_n) \)). Notice that we assume that the same number of Monte Carlo runs has been performed for both products. Although this may not be strictly necessary, it is at least the usual practice. Moreover, it simplifies the formulas. For many approaches, the supplementary information contains the general formula with unequal \( n_A \) and \( n_B \), and the main text simplifies to \( n_A = n_B = n \).

### 2.2 Informal conclusions

In literature, quite a few examples of presentations of comparative LCA results can be found. Sometimes, this assumes a tabular form (for instance, Romero-Gámez et al. (2017)’s Table 7) and/or a graphical form (for instance, Cespi et al. (2014)’s Fig. 2). Table 1 and Fig. 1 below show typical examples, for an unspecified impact category in an unspecified unit (it might be global warming in kg CO\(_2\)-equivalent, for instance). In most cases, no further analysis is done, but a verbal discussion concludes the presentation. In the example case, for instance, an analyst might conclude that product A performs slightly, but not always, better than product B.

It is not always clear what the tabulated or plotted numbers indicate. What do the “value” and the “uncertainty” represent? For the “value” component, the following options are imagined:

- the outcome of the “deterministic” LCA (i.e., the LCA result without uncertainties);
- the mean of the outcomes of the Monte Carlo series;
- the median of the outcomes of the Monte Carlo series;
- the geometric mean of the outcomes of the Monte Carlo series.

A similar confusion may surround the “uncertainty” number:

- ranges (min–max) of the outcomes of the Monte Carlo series;
- standard deviations of the outcomes of the Monte Carlo series;
- \( 2 \) (or 1.96) times the standard deviations of the outcomes of the Monte Carlo series;
- geometric standard deviations of the outcomes of the Monte Carlo series;
- squared geometric standard deviations of the outcomes of the Monte Carlo series;

| Table 1 | Typical format to communicate the results from a comparative LCA with uncertainty in tabular form |
|---------|--------------------------------------------------------------------------------------------------|
| Product A | Value ± uncertainty | 50.27 ± 10.41 |
| Product B | 60.22 ± 11.85 |

| Fig. 1 | Typical format to communicate the results from a comparative LCA with uncertainty in graphical form |
percentile values (e.g., \( P_{2.5} \) and \( P_{97.5} \)) of the outcomes of the Monte Carlo series;

- the standard error of the mean of the outcomes of the Monte Carlo series;

- 2 (or 1.96) times the standard error of the mean of the outcomes of the Monte Carlo series.

As long as this is not clearly stated, any further decision making becomes shaky. So, a first recommendation is that in forms of communication like Table 1 and Fig. 1, more complete and more standardized terms are employed, for instance, not “value” but “mean value” and not “uncertainty” but “standard deviation” (that is at least how Table 1 and Fig. 1 were made). Sometimes more information is added, but in a way which is not really helpful. For instance, Cespi et al. (2014) give a table (their Table 5) which contains “uncertainty scores … in terms of the squared standard deviation at 95% confidence interval (SD²).” Frankly speaking, we have no clue what this means. A squared standard deviation is a variance, but a confidence interval has little to do with that. And variances are, due to their quadratic unit (such as square kilogram when the impact score would be kilogram) difficult to interpret. In other cases (e.g. Messagie et al. 2014), the meaning of the numbers is clearly specified (“minimum, arithmetic mean, and maximum values”), but the extreme value minimum and maximum are perhaps not robust enough.

Figure 2 shows two variations on the barplot with error bars: the boxplot and the histogram. These forms can be seen as a more unambiguous presentation than Fig. 1, because the information in boxplots and histograms is more or less standardized. For instance, boxplots display the three quartiles (Q₁, median, and Q₃) in the central area. But the precise rules for the position of the whiskers and dots or stars outside these whiskers differ per software, and they are not always reported in case studies. And for histograms, the choice of the bin width affects the shape details of the histogram.

Yet, also these graphical forms do not immediately provide an undisputable and objective comparison of product alternatives. For instance, both figures suggest that A is better than B, but not with full confidence. The next few sections discuss a few attempts from literature. Some of these have been used in LCA, but others have, as far as we know, not.

2.3 Means, standard deviations, standard errors, and confidence intervals

The Monte Carlo results provide two data vectors (\( \mathbf{a} \) and \( \mathbf{b} \)), and we can compute the mean value from it. For product A, the mean will be indicated by \( \bar{a} \) and it is found through

\[
\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i
\]

(A similar expression holds for \( \bar{b} \), but only the formulas for A will be worked out below). Because the Monte Carlo simulation generates a random sample of finite size, the means \( \bar{a} \) and \( \bar{b} \) will not be exact representations of the population values (\( \mu_A \) and \( \mu_B \)) but will deviate from it. The standard error of the mean (often abbreviated plainly as the standard error) is an estimate of the standard deviation of the estimated value of \( \bar{a} \) and \( \bar{b} \). When \( n \) is large (say, larger than 30, which typically is the case in Monte Carlo experiments), this standard error of the mean (\( s_{\bar{a}} \)) is given by

\[
s_{\bar{a}} = \frac{s_A}{\sqrt{n}}
\]

In this expression, \( s_A \) is the standard deviation of the Monte Carlo results of product A

\[
s_A = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\bar{a} - a_i)^2}
\]
The standard error is in LCA not often reported, but Ross and Cheah (2017) do so. Ribal et al. (2017) also provide standard errors, but not on the basis of the \( \frac{s}{\sqrt{n}} \) formula, but on the basis of more numerically intensive process, namely, bootstrapping (Efron 1979). In our example, we have concentrated on the classical expression for the standard error of the mean.

The standard error can be used to construct a confidence interval for the mean of \( A \). This is an interval (defined by two numbers, the lower value, and the upper value), within which the true mean (\( \mu_A \)) is located with large confidence. A usual choice for this confidence level is 95\%, but other choices (e.g., 99\%) are also defensible. The confidence interval at confidence level \( 1 - \alpha \) is then given by

\[
\text{CI}_{\mu_A,1-\alpha} = \left[ \bar{a} - t(n - 1, \alpha/2)s_{\bar{a}}^{-1} + t(n - 1, \alpha/2)s_{\bar{a}}^{-1} \right]
\]

where \( t(\nu, p) \) is the value of the \( t \)-distribution that belongs to a cumulative right-tailed probability \( p \) at \( \nu \) degrees of freedom (for a 95\% confidence interval, use \( p = 0.025 \), which gives a \( t \) of approximately 1.96 when \( n \) is large). In that case, the width of the confidence interval for \( \mu_A \) is

\[
2 \times 1.96 \times s_{\bar{a}}^{-1} \approx 4s_{\bar{a}}^{-1}
\]

Table 2 shows the results for the example systems. Figure 3 shows the means and the 95\% confidence intervals of the means.

Because there is no overlap at all in Fig. 3, one might conclude that product A performs better than product B. This conclusion, however, is not necessarily justified, because it depends on what one means by “performs better.” At any case, one can conclude that the mean score of product A is “significantly” (see below) lower than the mean score of product B. But a lower mean score is not the same as a lower score.

| Statistic | Product A | Product B |
|-----------|-----------|-----------|
| Mean (\( \bar{a} \) and \( \bar{b} \)) | 50.27 | 60.22 |
| Standard deviation (\( s_{\bar{a}} \) and \( s_{\bar{b}} \)) | 10.41 | 11.85 |
| Standard error of the mean (\( s_{\bar{a}}^{-1} \) and \( s_{\bar{b}}^{-1} \)) | 0.33 | 0.37 |
| 95\% confidence interval for the mean (\( \text{CI}_{\mu_{\bar{a}},0.95} \) and \( \text{CI}_{\mu_{\bar{b}},0.95} \)) | [49.63, 50.92] | [59.49, 60.96] |

Fig. 3 A graphical representation of the means and 95\% confidence intervals of the means. Mind that the vertical scale differs from the one in Fig. 1 and that the vertical axis does not start at 0.

Examples of confidence intervals in LCA can be found with Aktas and Bilec (2012), Kim et al. (2013) and Niero et al. (2014). Ribal et al. (2017) also give confidence intervals, on the basis of bootstrap estimates.

An important caveat is that the standard error or the mean, and therefore the width of the confidence interval for the mean, depends on the sample size:

\[
s_{\bar{a}}^{-1} \propto \frac{1}{\sqrt{n}}
\]

The consequence is that both standard error of the mean and width of the confidence interval for the mean can shrink arbitrarily small if the number of Monte Carlo trials is increased. This can be regarded as a variant of “p-hacking” (Ioannidis 2019). We will return to this issue below.

A final comment is that standard errors and confidence intervals cannot only be calculated for the mean but also for other parameters, e.g., for the standard deviation and the median. But they always refer to a parameter, not to the distribution in itself. So, we can speak of a confidence interval of the mean or median CO\(_2\)-emission, but not of
2.4 Other centrality statistics

The previous section took the mean of A and B (a and b) as the basis of the “value” of Table 1. Instead of the mean, we may decide to emphasize other statistics that represent a centrality parameter. Two examples will be briefly elaborated here: the median and the geometric mean.

The median (M) is the value for which half of the data is lower and the other half is higher. Finding the median from a vector of numbers is easy: sort and take the middle value, or the mean of the two middle values in case of an even sample size. The issue of the standard errors and confidence intervals is more involved (for instance, using the bootstrap). The same caveat as for the mean shows up: the standard errors and confidence intervals shrink with increasing sample size, and sample size is basically the very large number of Monte Carlo runs. The median has occasionally been used, for instance by Muller et al. (2017). Geisler et al. (2005) also mention the median, but they define it in a different way (“the square root of the product of the minimum and maximum value,” so more as a geometric analogue to the midrange).

The geometric mean is another centrality measure, and it has become increasingly popular in LCA (Sonnemann et al. 2003; Hong et al. 2010). The geometric mean (here denoted as GM) of a data vector is calculated as

\[ GM_A = \exp \left( \frac{1}{n} \sum_{i=1}^{n} \ln(a_i) \right) \]

It is difficult to draw clear and justified conclusions on the basis of the median and/or geometric mean with or without standard errors and/or confidence intervals. For completeness, Table 3 shows the values for the example case.

In the case of approximately symmetric distributions as here, mean, median, and geometric mean are very close by. But for highly skewed distributions, large differences may occur. For instance, the small data vector (2, 3, 100) has mean 35, median 3, and geometric mean 8.43.

2.5 Other uncertainty statistics

Section 2.3 was based on the idea that an uncertainty indication was a standard deviation or standard error, calculated from a vector of Monte Carlo results (s_A and s_B) above. A few other options are discussed here.

Kim et al. (2013) present results with a coefficient of variation (CV). In general, it is defined as

\[ CV_A = \frac{s_A}{a} \]

The CV is dimensionless, so it cannot be drawn together with the mean or median in one plot.

The geometric standard deviation (GSD, or its square, GSD^2) is also sometimes seen in LCA (Hong et al. 2010; Muller et al. 2017). It is defined as

\[ GSD_A = \exp \left( \sqrt{\frac{\sum_{i=1}^{n} \ln\left( \frac{a_i}{GM_A} \right)^2}{n - 1}} \right) \]

In contrast to the usual standard deviation, the geometric standard deviation is dimensionless. This has the advantage of a more straightforward interpretation as “pure numbers.” The downside is that expressions such as “central value ± geometric standard deviation” cannot be made.

The interquartile range (IQR) is the distance between the first and the third quartile, so the size of the “box” of a boxplot:

\[ IQR_A = Q_{3.A} - Q_{1.A} \]

The quartiles are 25 and 75-percentiles, so the IQR spans the central 50% of the data. Heijungs and Lenzen (2014) employ it within LCA.

We also see sometimes a form that spans 95% instead, running from P_{2.5} to P_{97.5} (where P indicates percentiles). As stated above, some texts incorrectly refer to this as a 95% confidence interval. Incorrectly, because confidence intervals apply to the estimate of a parameter, such as the mean and the standard deviation, not to the entire confidence interval.
distribution. But of course we can calculate this distance, which might be called an interpercentile range (IPR):

\[ \text{IPR}_{95,A} = P_{97.5,A} - P_{2.5,A} \]

Some texts do not determine it from the Monte Carlo-based distributions, but they use a formula based on the normal distribution

\[
\left( \bar{a} + 1.96 s_A \right) - \left( \bar{a} - 1.96 s_A \right) = 3.92 s_A \approx 4 s_A
\]

There are also authors who base this on lognormal distributions:

\[
\frac{\text{GM}_A \times \text{GSD}_A^{1.96}}{\text{GM}_A / \text{GSD}_A^{1.96}} = \text{GSD}_A^{3.92} \approx \text{GSD}_A^4
\]

Further, Geisler et al. (2005) define an “uncertainty range” (UR) as a 90% percentile interval:

\[ \text{UR}_{90,A} = \frac{P_{95,A}}{P_{5,A}} \]

Table 4 lists the values of these other statistics for the example case.

As a final comment, for several statistics there are two fundamentally different ways to calculate them. Take, for instance, the 95-percentile value \( P_{95} \). We can calculate from the data the mean and standard deviation as above, and then apply

\[ P_{95,A} = F^{-1}_{N\left(\bar{a},s_A\right)}(0.95) \approx 67.39 \]

where \( F^{-1}_{N\left(\bar{a},s_A\right)}(0.95) \) is the inverse cumulative distribution of the normal distribution. Or if we assume another distribution (say, lognormal), apply an equivalent formula. Alternatively, we may just order the data and look for the value at position 950 (at least, when \( n = 1000 \)). In our example this was

\[ P_{95,A} \approx 68.57 \]

In this case, the difference is small. But the difference can be large, and there is a fundamental difference due to the fact that the second approach does not assume a particular distribution, but purely works with the data. It is often unclear which of these two approaches is used in a specific article.

### 3 Indicators of difference for two product alternatives

#### 3.1 General introduction

The previous section discussed how to express a result including uncertainty for each product separately. This section will take up the challenge of focusing on the comparison of two products in one integrated metric. Before doing so, a small intermezzo is needed to discuss the issue of dependent and independent sampling.

Typically, product systems are modeled with different foreground processes but with identical background processes. For instance, in comparing two types of light bulb, we assume electricity for using the light bulb comes from the same power plant or market mix. It therefore makes sense to use the same electricity sample realization in a particular Monte Carlo run for both types of light bulb. This issue has been discussed in various places (Heijungs and Kleijn 2001; Henriksson et al. 2015; Suh and Qin 2017; Qin and Suh 2018; Lesage et al. 2019; Heijungs et al. 2019), and this is not the place to further elaborate on this. But the distinction is important to keep in mind. In the next sections, we will add information on the sampling method (dependent or independent) whenever appropriate, and provide results for both approaches whenever applicable. Briefly stated, we do all sampling in a dependent way, but we consider the samples as independent if we want to test the effect of independent sampling. For instance, we apply the independent samples \( t \)-test to simulate independent sampling.

#### 3.2 Null hypothesis significance testing

Some authors (e.g., Henriksson et al. 2015; Ross and Cheah 2017) have used null hypothesis significance testing (NHST), as a formal criterion to decide if one product beats the other product. The raw material of NHST is the samples, which are considered to be obtained from populations.
with hypothesized properties. Then it is calculated how likely or unlikely the sample properties are in the light of the population properties. If the outcome is unlikely, the hypothesis is rejected, and the opposite alternative hypothesis is considered to hold. The \( p \)-value measure the degree of likelihood.

Hypothesis tests exist for different parameters (mean, median, standard deviation, etc.) and for different set-ups (one sample, two samples, multiple samples). Below, we will focus on two-sample hypothesis tests for the mean. Near the end of this section we will discuss the case of the median, and the next section will discuss the case of more than two samples.

A relevant incarnation of such a test is one in which the null hypothesis is that the mean scores of the two products are the same. In symbols

\[
H_0 : \mu_A = \mu_B
\]

The independent samples \( t \) test (see supplementary information) proposes a test statistic

\[
t = \frac{\bar{a} - \bar{b}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}
\]

The value of the test statistic \( t \) obtained from the samples is then compared with the critical value, for a pre-defined significance level \( \alpha \). The value \( \alpha = 0.05 \) is a conventional choice, but just as with the confidence level, other choices can be made, for instance \( \alpha = 0.01 \). The critical value in this case comes from a \( t \)-distribution with \( 2n - 2 \) degrees of freedom. But because \( n \) is typically large or very large (e.g., 1000 or more), the critical value effectively reduces to the standard normal value. For the example data, the results are as in Table 5.

The value of \( t \) is clearly in the rejection region, so the conclusion is that the null hypothesis of equal means is to be rejected, in favor of the alternative hypothesis of unequal means, so \( \mu_A \neq \mu_B \).

Table 5 also contains the results for the dependent (or paired) samples \( t \)-test. See the supplementary information for more detail.

| Statistic               | Independent comparison | Dependent comparison |
|-------------------------|------------------------|----------------------|
| Value of \( t \)-statistic | \(-19.95\)            | \(-64.40\)           |
| Degrees of freedom      | 1998                   | 999                  |
| Critical values of \( t \) | \(-1.96\) and \(1.96\) | \(-1.96\) and \(1.96\) |
| \( p \)-value           | \(< 10^{-6}\)          | \(< 10^{-6}\)        |

If the null hypothesis \( \mu_A = \mu_B \) is rejected as above, the NHST procedure “proves” that the mean scores \( \mu_A \) and \( \mu_B \) differ “significantly.” A number of remarks are due on the terms “prove” and “significant”:

- The NHST procedure can never exclude the possibility that the null hypothesis of equal means has been incorrectly rejected. In fact, the chance of drawing this wrong conclusion is set by \( \alpha \) (5% in this example). Such an event is referred to a type I error.
- The procedure may also fail to detect a difference. Depending on the test details, the probability of such a type II error (\( \beta \)) can be quite high.
- The term “significant” may suggest to a quick reader that the difference is large or important. In fact, it is only jargon for the fact that NHST has established that there is (probably) a difference, but the difference can be small or large. Indeed, Vercauteren et al. (2010) write that “the reusable PC cup has a significant more favorable environmental score than the one-way cups,” without a hypothesis test, demonstrating the danger inherent in the term “significant.”
- After a rejection of the null hypothesis of equality of means, one can conclude that the means are not equal. But a decision-maker wants of course to know which product is then better. Therefore, a so-called post-hoc analysis is needed to find out more.

In addition, it is important to observe that the \( t \)-statistic contains a term with \( \sqrt{n} \). Effectively, this means that a difference between \( \mu_A \) and \( \mu_B \), however tiny, will always be detected with a sufficiently large \( n \). And \( n \) is here the number of Monte Carlo runs, so basically a user-defined aspect. When the population mean for A is 100.12345 and for B it is 100.12346, this difference will inevitably come to the surface with a huge sample size \( n \), and NHST will then declare the negligible difference to be “significant.”

A variation to the test \( \mu_A = \mu_B \) with the \( t \) test is the Wilcoxon-Mann-Whitney test for comparing medians. This test can also be used to compare means in case the populations do not follow a normal distribution. However, there is hardly any need for this, because even for very skew distributions, the \( t \) test performs extremely well when \( n \) is large.

Results for the example case are as in Table 6. The table also shows, for completeness, the dependent (paired) samples version, using the Wilcoxon signed-rank test. See the supplementary information for details.

The conclusion is here the same as for the comparison of means. The evidence of a “significant difference” is slightly smaller than before (due to the replacement of values by ranks, some information is lost and the power of the test is therefore smaller), but it is still overwhelming. However, it
Table 6 Results of a null hypothesis significance testing \((H_0 : \mu_A = \mu_B)\) of the medians of the example system

| Statistic                  | Independent comparison | Dependent comparison |
|----------------------------|------------------------|---------------------|
| Sum of ranks \((S_A \text{ or } S_B)\) | 759626                 | 22686               |
| Value of z-statistic       | −18.65                 | −24.91              |
| Critical values of z       | −1.96 and 1.96         | −1.96 and 1.96      |
| p value                    | < 10⁻⁶                 | < 10⁻⁶              |

should be kept in mind that this only implies that the median product A has a lower impact than the median product B.

3.3 The standardized mean difference

Cohen (1988) introduced the standardized mean difference as the difference of the means of the two groups, divided by a common standard deviation. As a formula, in the case of equal sample size \(n_A = n_B = n\):

\[
d = \frac{\bar{a} - \bar{b}}{\sqrt{\frac{s_a^2 + s_b^2}{2}}}
\]

The standardized mean difference measures how many standard deviations the two means are separated. It has the convenient property of being dimensionless, which enables the setting of absolute standards (e.g., a policy might prescribe that products with a \(d\) of 0.4 or more should be regulated.) Cohen considered a value around 0.2 as “small,” around 0.5 as “medium” and around 0.8 as “large.” There is no limit to \(d\), it can exceed 1 very well. It can also be negative, in which case the absolute value is used for the labels “small,” etc. As far as we know, the standardized mean difference has never been proposed within LCA (but see below).

Cohen also related the \(d\) to another type of effect size, namely the Pearson correlation coefficient, here between the impact scores and the dichotomous variable that indicates which product it is from, A or B:

\[
r = \frac{d}{\sqrt{d^2 + 4}}
\]

Its square, \(r^2\), can be interpreted as the fraction of variance in the score that is “explained” by the subdivision in two groups. The correlation coefficient has been mentioned in LCA, but never in the context of ranking products.

The value of the standardized mean difference is not dependent on the sample size \(n\). Of course a larger sample size will return a more accurate value for the standardized mean difference, but the result will not be systematically smaller, as is the case with standard errors, confidence intervals and \(p\)-values, or systematically larger, as is the case with \(t\)-statistics. As such, it appears to be more natural in the context of arbitrarily large sample sizes, as with Monte Carlo. That is also a reason why Cumming (2012) mentions it in relation to a critique on NHST.

Standardized mean differences themselves can also be the topic of a null hypothesis significance test, and it is likewise possible to calculate a confidence interval for it (Hedges and Olkin 1985; Cumming 2012). Easier, however, is the familiar test for \(r\). Again, there is a \(\sqrt{n}\) dependence, so with large \(n\), the population value (\(\delta\)) will always be found to differ from 0, even for extremely small values of \(\delta\).

Results for the example system are in Table 7.

The standardized mean difference has been mentioned in relation to LCA by Heijungs et al. (2016) and Aguilera Fernández (2016), but as far as we know, it has not been actually applied to LCA case studies.

3.4 “Modified” null hypothesis significance testing

Recognizing the drawbacks of normal NHST, Heijungs et al. (2016) introduced a modification to the NHST scheme (referred to as “modified NHST” by Mendoza Beltrán et al. (2018)), in which the event of a negligible but significant difference is removed through requiring a minimum threshold difference. But because differences can pertain to variables that are measured in any unit or scale, the minimum difference is expressed as a relative threshold, similar to the standardized mean difference.

Together with the significance level \(\alpha\), a second number, \(\delta_0\), is set, for instance \(\delta_0 = 0.2\). The null hypothesis then takes the one-sided form

\[
H_0 : \left| \frac{\mu_A - \mu_B}{\sigma} \right| \leq \delta_0
\]

It can be tested with a slightly different test statistic:

Table 7 Result of the standardized mean difference of the example system

| Statistic               | Independent comparison | Dependent comparison |
|-------------------------|------------------------|---------------------|
| Cohen’s \(d\)           | −0.89                  | −2.04               |
| Pearson’s \(r\)         | −0.41                  | −0.90               |
| Fraction explained \(r^2\) | 0.21                   | 0.81                |
| Value of \(t\)-statistic | −14.10                 | −64.40              |
| Critical values of \(t\) | −1.96 and 1.96         | −1.96 and 1.96      |
| Degrees of freedom      | 998                    | 999                 |
| \(p\) value             | < 10⁻⁶                 | < 10⁻⁶              |
\[ t = \frac{\bar{a} - \bar{b}}{\sqrt{\frac{s_a^2 + s_b^2}{n}}} - \delta_0 \]

and \( t(2(n-1)) \). The example system yields the results of Table 8.

### 3.5 Nonoverlap statistics

Some authors (e.g., Ross and Cheah 2017; Prado and Heijungs 2018) present histograms (or variations on it, such as violin plots) of the Monte Carlo distributions of comparative LCA. Typically, such histograms have a degree of overlap.

For the case that the histograms represent two normal distributions with the same standard deviation, Cohen (1988) defined three statistics, \( U_1 \), \( U_2 \), and \( U_3 \), that measure the degree of nonoverlap of two probability distributions. The supplementary information defines the three; here, we will only provide their values for the example case (Table 9). No equivalent for the dependent case is available.

Two variations should be mentioned. Grice and Barrett (2014) introduce another form for \( U_1 \), here referred to as \( U'_1 \). And McGraw and Wong (1992) introduce another variation: the common language effect size (CLES), which is in between \( U_2 \) and \( U_3 \), and which is supposed to reflect that probability that a randomly select specimen of group B has a higher value than a randomly select specimen of group A. In the example case, it is 0.74, suggesting that there if we buy 100 times a specimen of product A and B, in 74% of the cases product A will perform better.

### 3.6 Other nonoverlap statistics

While the nonoverlap statistics above only apply to normal distributions, the Bhattacharyya coefficient is a measure of the similarity of two arbitrary probability distributions (Everitt and Skrondal 2010). It is defined for theoretical probability distribution functions, in which case the value is 0 for perfectly separated distributions, and 1 for perfectly coinciding distributions. It can also be found for empirical data by partitioning the values into a number of bins, like we do for a histogram. The details of the binning influence the value of the Bhattacharyya coefficient, so we tried a few bin sizes in the example case, concluding that the Bhattacharyya coefficient is around 0.85 (see Fig. 4).

The Bhattacharyya coefficient can also be transformed in a distance, the Bhattacharyya distance, which has an exactly opposite interpretation. Other distances (Kullback-Leibler distance, mutual entropy, etc.) are possible as well. In the context of LCA, we further mention a paper by Qin and Suh (2018), where an overlapping coefficient (OVL) was defined and calculated (see Supplementary information). In general, the results are difficult to interpret. For instance, in the case study the Bhattacharyya distance was around 0.15, but a decision-maker will not find that very helpful. But they do have an advantage over the indicators by Cohen in case of clearly non-normal distributions.

### 3.7 Comparison indicator and discernibility analysis

Several authors (Huijbregts 1998; Huijbregts et al. 2003; Geisler et al. 2005) use the ratio of two scores

\[ CI_i = \frac{b_i}{a_i} \]

for every Monte Carlo run \( i = 1, \ldots, n \). This yields a distribution of values. Whenever a certain percentage (Huijbregts et al. 2003) use 95%, Geisler et al. (2005) use 90%) of the CI values is below or above 1, the product with the lower impact is declared to be significantly better.

A related analysis is the discernibility analysis, introduced by Heijungs and Kleijn (2001). The procedure counts how often A beats B, and expresses this result as a fraction, which can be interpreted as a probability (see below). This can easily be implemented by using the Heaviside step function \( \Theta(x) \) (which returns 1 if \( x > 0 \) and 0 otherwise):

### Table 8 Result of the modified NHST procedure for the example system, with \( \delta_0 = 0.2 \)

| Statistic            | Independent comparison | Independent comparison |
|----------------------|------------------------|------------------------|
| Value of \( t \)-statistic | 19.75                  | 64.40                  |
| Critical value of \( t \) | 1.65                   | 1.65                   |
| Degrees of freedom  | 1998                   | 999                    |
| \( p \) value        | < 10^{-6}              | < 10^{-6}              |

### Table 9 Result of the nonoverlap statistics \( U_i \), \( U_2 \), and \( U_3 \) for the example system. Also included are two later modifications

| Statistic           | Independent comparison | Dependent comparison |
|---------------------|------------------------|----------------------|
| Cohen’s \( U_1 \)   | 0.51                   | –                    |
| Cohen’s \( U_2 \)   | 0.67                   | –                    |
| Cohen’s \( U_3 \)   | 0.81                   | –                    |
| Grice and Barrett \( U'_1 \) | 0.34                 | –                    |
| McGraw and Wong (CLES) | 0.74            | –                    |
If $K$ is around 0.5, a random specimen of product A is equally often better and worse than a random specimen of product B. Here, it has been proposed that a product is significantly better when $K > 0.95$ or $K < 0.05$.

Both approaches are by definition a “per run” analysis, working with $b_i - a_i$ or equivalently $\frac{b_i}{n}$, implying a dependence assumption. Yet, the Monte Carlo runs themselves may be created not only per run, but also within one run per alternative. Both dependently and independently sampled results are shown in Table 10.

The approaches are easy to implement, but of course a judgment which fraction of CI or which value of $K$ is convincing remains unclear. Probably it will depend on the interests at stake and the goal of the study, so it might be set in the goal and scope definition. Another issue is that a large discernibility can be associated with a small difference. In that sense, it has the same problems as NHST.

Results from these two types of analysis have been visualized. Figure 5 gives the range of the CI in terms of its 5- and 95-percentile and the median, following Huijbregts et al. (2003). The dashed line indicates the line of indifference. The figure also illustrates the probability of dominance of one product versus another product, following Guo and Murphy (2012). Notice that we speak of “probability of dominance,” employing the frequentist interpretation of probability as the number of favorable events divided by the total number of events.

This approach has actually been implemented quite a few times, most often in relation to the CI but also occasionally in relation to discernibility. A few examples include De Koning et al. (2010), Röös et al. (2010, 2011), Gregory et al. (2013), Cespi et al. (2014), Gregory et al. (2016) and AzariJafari et al. (2018), and Von Brömssen and Röös (2020).

### 3.8 Other measures of superiority

The idea of the common language effect size is intriguing, because it tries to address the question “What is the probability that a randomly selected specimen of product A performs better than a randomly selected specimen of product B?” And that is not the same as: “Is the mean of product A’s score lower than the mean of product B’s score?” But still the common language effect size has been defined for normal distributions only. This is also the case for the other nonoverlap statistics. In that sense, the comparison index, the discernibility analysis, and the Bhattacharyya coefficient have a wider applicability, because they rely on the empirical distribution of A and B, not on a fitted theoretical distribution. With a powerful computer and 1000 (or even more) data points, we can do a much better analysis than when we force a not entirely fitting normal or other distribution with the data. Below a few options are worked out. Because they are variations to the discernibility analysis, we will label their results as $K_2$, $K_3$, etc.
A first option is to find the fraction of cases of B that exceeds the value of one specific run for A \( (a_i) \). As a formula:

\[
S_i = \frac{1}{n} \sum_{j=1}^{n} \Theta(b_i - a_j)
\]

Next we average this over all cases of A:

\[
K_{2,A} = \frac{1}{n} \sum_{i=1}^{n} S_i = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Theta(b_i - a_j)
\]

This number can be interpreted as the probability of superiority of A. We can also swap the positions of A and B to find the probability of superiority of B. The two numbers add up to 1. Wei et al. (2016) use an indicator of this style, referring to it as the decision confidence probability.

Another option is to argue that half of the time, we draw a value from A that is less than the median, and half of the time the value is larger. It therefore suffices to count how many values of B exceed the median of A. In formula:

\[
K_{3,B} = \frac{1}{n} \sum_{i=1}^{n} \Theta(b_i - M_A)
\]

This number can be interpreted as the probability of superiority of A with respect to the median B. Also here, we can swap the positions of A and B to find the probability of superiority of B with respect to the median A. These two numbers do not necessarily add up to 1. See Table 11 for the results of the example. This type of comparison is intrinsically based on independent analysis.

All these superiority measures are expressed as fractions of count and apply to the entire distribution. They are not susceptible to the problem of "p-hacking" that was identified in relation to significance tests. On the other hand, they share with significance the neglect of the size of the difference.

### 4 More than two product alternatives

So far, the discussion has been restricted to the case of two products, A and B. It frequently happens that the comparison involves three or even more products. Some of the approaches mentioned above can just be applied for each pair of products, but for some other approaches, care is needed. Unfortunately, most texts do not explicitly discuss the multi-comparison case. A few exceptions are Heijungs and Kleijn (2001), Copius Peereboom et al. (1999), Hung et al. (2009), and Mendoza Beltrán et al. (2018).

First, the issue of computational demand. If there are \( k \) alternatives, there are \( \frac{1}{2} k(k + 1) \) different pairs to consider. With \( k = 10 \) alternatives, for instance, the number of pairwise comparisons is 55, and with \( k = 15 \) this has increased to 120. The number of comparisons therefore increases rapidly. Most types of analysis described above, however, do not take much time, but it might become an issue, especially if the number of Monte Carlo runs is very large. For instance, \( K_2 \) requires a double summation. With \( n = 100,000 \) this implies 10 billion steps, for each pair of alternatives. It may also become an issue when we try to use uncertainty considerations to prioritize data collection efforts for a very large number of products, for instance, when we would combine the methods of Reinhard et al. (2019) and Patouillard et al. (2019).

Another important thing to consider is the increase of the probability of a type I error in null hypothesis significance tests. If there are \( k = 10 \) alternatives and 55 pairwise comparisons must be made, each with a type I error probability of 5%, the overall error rate becomes...
so 94%, while we intended to keep this at 5%. The Bonferroni correction then would require to test the individual hypotheses at $\alpha \approx 0.001$. A better idea is then to use an omnibus test that test all hypotheses in one go. The analysis of variance (ANOVA) is such an omnibus test for testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

It provides an $F$-statistic instead of a $t$-statistic, and it is tested only with an upper critical value, but the basic idea is very much the same. The conclusion is, however, not. When the null hypothesis is rejected, the conclusion is that at least one of the product alternatives has a significantly deviating mean score, but it is not known which it is (or are), and neither if it is (or they are) higher or lower than the rest. To find this out, a post-hoc test is needed. There are many such post-hoc tests, an often-used one is Tukey’s honestly significant difference (HSD) test. For testing equality of $k$ medians, an omnibus version of the Wilcoxon-Mann-Whitney test is available as the Kruskal-Wallis test. Here the issue of post hoc tests is more complicated.

The approaches that do not do a significance test, such as the comparison indicator, the discernibility analysis, the nonoverlap statistics, and the Bhattacharyya coefficient, all rely on pairwise comparison, but they do not suffer from the issue of inflating the probability of a type I error, because they do not perform a significance test. A convenient scheme for presenting the results is then in the form of Table 12, which is based on Mendoza Beltrán et al. (2018). In this table, an indication like “$X \leftrightarrow Y$” means that the cell contains the preference information of product X versus Y. For instance, it contains Cohen’s $d$ or the Bhattacharyya coefficient of X and Y. It can also contain two types of information, for instance a value and a significance level, a value and a standard deviation, or a confidence interval.

The table above contains all information for a specific pair (say, X and Y) two times, for “$X \leftrightarrow Y$” and for “$Y \leftrightarrow X$.” In that respect there are some observations to be made:

- For some of the statistics, there is a situation of symmetry, so the cell with “$X \leftrightarrow Y$” contains the same result as “$Y \leftrightarrow X$.” This is, for instance, the case for $p$-values from NHST.
- For other statistics, the two cells contain exactly opposite (antisymmetric) information. For instance, for Cohen’s $d$ we have that “$X \leftrightarrow Y$” has the same numerical result as “$Y \leftrightarrow X$,” except for a minus sign.
- Still other statistics have complementing information. For instance, for the probability of superiority, the values in “$X \leftrightarrow Y$” and “$Y \leftrightarrow X$” add to exactly 1.
- The most interesting situation occurs for those statistics for which “$X \leftrightarrow Y$” and “$Y \leftrightarrow X$” are not uniquely related, so for which the second cell contains information that we could not guess from the first cell alone. This is, for instance, the case for the modified NHST (see also Mendoza Beltrán et al. 2018).

For some of the statistics, there is a situation of symmetry, so the cell with “$X \leftrightarrow Y$” contains the same result as “$Y \leftrightarrow X$.” This is, for instance, the case for $p$-values from NHST.

Table 12 can be made more readable not only by filling out numbers in the different cells but also by adding a color scheme, for instance, using “cold” colors (blue) for small differences and “hot” colors (red) for large differences.

### Table 11

| Statistic                              | Independent comparison | Dependent comparison |
|----------------------------------------|------------------------|----------------------|
| Probability of superiority of A w.r.t. B (K2, A) | 0.74                   | –                    |
| Probability of superiority of B w.r.t. A (K2, B) | 0.26                   | –                    |
| Probability of superiority of A w.r.t. the median B (K1, A) | 0.82                   | –                    |
| Probability of superiority of B w.r.t. the median A (K1, B) | 0.18                   | –                    |

### Table 12

| product A | product B | product C | product D |
|-----------|-----------|-----------|-----------|
| product A | –         | A ↔ B     | A ↔ C     | A ↔ D     |
| product B | B ↔ A     | –         | B ↔ C     | B ↔ D     |
| product C | C ↔ A     | C ↔ B     | –         | C ↔ D     |
| product D | D ↔ A     | D ↔ B     | D ↔ C     | –         |

5 Discussion

We have by now met quite a few approaches for selecting the best product alternative out of two options. Altogether, it appears that two questions need separate attention.

Question 1: What is the probability that a randomly selected specimen of product A performs better than a randomly selected specimen of product B?

Table 12 Proposed format for communicating comparative results in case of more than two products. A cell like “$X \leftrightarrow Y$” contains information about the difference or preference of X with respect to Y.
• If the answer is around 50%, the two products are indifferent, and tossing a coin is equally reliable as using the LCA result.
• If the answer is much less than 50%, we should choose product B.
• If the answer is much more than 50%, we should choose product A.

Question 2: How much will a randomly selected specimen of product A perform better than a randomly selected specimen of product B?
• If the answer is just a bit, it is questionable if we should invest the money and time to switch.
• If the answer is that it matters a lot, a choice will pay off.

But because we should address both questions simultaneously, a decision matrix as in Table 13 appears. Then a final question is which of the various statistics introduced in this article help to answer which problem. Table 14 presents an answer to that nagging question.

Except for the last one (to be discussed below), none of these approaches offers a simultaneous Yes to both questions. This is, after all, understandable. The mean size of an effect and the probability of an effect remain two different aspects, and reconciling them into one remains an issue. NHST (including its modification) covers neither of the two, and the other approaches cover only one.

First, we discuss the striking fact that NHST performs bad here. After all, hypothesis testing is often considered to be the summmum bonum of scientific research, and it assumes an important place in many curricula. But note, we are not arguing that it is useless, we only conclude that it is not useful in answering the two questions that we just formulated. In fact, it answers a third question: What is the probability that the observed effect is the effect of chance due to a limited sample size? In a context of Monte Carlo runs, that question is not relevant (Heijungs, 2020).

Next, we discuss the fact none of the approaches answers both questions. With a little bit of recombination work, something must be possible. We propose a “modified comparison index” as follows. First we define two versions of the comparison index:

\[
CI_{A,i} = \frac{b_i}{a_i} \quad \text{and} \quad CI_{B,i} = \frac{a_i}{b_i}
\]

for \( i = 1, \ldots, n \). Next, we define a minimum threshold value, for instance \( \gamma_0 = 1.2 \). This is used for assessing the superiority of A:

\[
K_{4,A} = \frac{1}{n} \sum_{i=1}^{n} \Theta(CI_{A,i} - \gamma_0)
\]

and the superiority of B:

\[
K_{4,B} = \frac{1}{n} \sum_{i=1}^{n} \Theta(CI_{B,i} - \gamma_0)
\]

See Table 15 for the results. In the case of independent sampling, we see that product A beats product B by a factor of at least 1.2 (so 20%) with 51% probability, and it is beaten by product B by a factor of at least 1.2 with 10% probability. In the end, three arbitrary, goal and scope related, numbers must be defined in order to arrive at a decision:

• the threshold value of the comparison index (\( \gamma_0 \));
• the minimum probability of beating an inferior product alternative (is 51% enough?);
• the maximum probability of being beaten by an inferior product alternative (is 10% acceptable?).

Clearly, more experience is needed before rules of thumb can be designed for this. Perhaps lessons can be learned from IPCC (Mastrandrea et al. 2010), who developed a likelihood scale, running from “exceptionally unlikely” (less than 1%) to “virtually certain” (more than 99%). Another possible source of inspiration can be found in the usual false positive (type-I) probability (\( \alpha \)) and false negative (type-II) probability (\( \beta \)) (Agresti and Franklin 2013).

There are multiple advantages of this procedure:

### Table 13: Framework for deciding between two products A and B

| Question                                                                 | Low     | ≈ 50% | High    |
|--------------------------------------------------------------------------|---------|-------|---------|
| How much will a randomly selected specimen of product A perform better than a randomly selected specimen of product B? | a bit   | questionable | never mind | questionable |
|                                                                          | a lot   | choose B    | questionable | choose A    |

© Springer
it works on the empirical distribution of results, so no assumptions about normality, lognormality, or other distributions are needed;
• it does not require to make a choice for a specific centrality or dispersion statistic, such as the mean and the standard deviation, but it employs the full distribution;
• it can be easily generalized to comparisons of three or more product alternatives;
• it is easy to implement in software;
• the procedure is not sensitive to p-hacking;
• the results are easy to grasp for less statistically trained people.

With respect to this final advantage: the procedure can be easily explained in terms of counting. One simply counts how often the CI favoring product A is more than 1.2, and how often the CI favoring product B is more than 1.2.

Another layer of complexity is added when multiple impact categories are assessed, including multi-criteria methods that allow for trade-off. Prado-Lopez et al. (2014), Wei et al. (2016), and Mendoza Beltrán et al. (2018) discuss a few issues, but obviously a renewed treatment makes sense in the light of our discussion for the one-category case.

### Table 14 Suitability of the various comparison statistics for answering the two relevant questions

| Statistic                                      | Answers the question | How much will a randomly selected specimen of product A perform better than a randomly selected specimen of product B? |
|-----------------------------------------------|----------------------|----------------------------------------------------------------------------------------------------------------|
| difference in mean or median                  | no                   | yes                                                                                                           |
| NHST with t-test or Wilcoxon–Mann–Whitney test| no                   | no                                                                                                            |
| modified NHST                                 | no                   | no                                                                                                            |
| Cohen’s d, Pearson’s r                        | no                   | yes                                                                                                           |
| nonoverlap statistics (U₁, U₂, U₃, CLES)      | yes                  | no                                                                                                            |
| Bhattacharyya coefficient and overlapping coefficient | yes       | no                                                                                                            |
| comparison index and discernibility (K₂, K₃) | yes                  | no                                                                                                            |
| modified comparison index (K₄; see below)     | yes                  | yes                                                                                                           |

### Table 15 Result of the proposed superiority statistics K₄ for the example system, using γ₀ = 1.2

|                               | Independent comparison | Dependent comparison |
|-------------------------------|-------------------------|----------------------|
| Probability of threshold superiority of A (K₄,a) | 0.51                    | 0.49                 |
| Probability of threshold superiority of B (K₄,b) | 0.10                    | 0.00                 |

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11367-020-01851-4.

**Acknowledgments** Two reviewers gave very helpful suggestions which helped to improve the text.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

AzariJafari H, Yahia A, Amor B (2018) Assessing the individual and combined effects of uncertainty and variability sources in comparative LCA of pavements. Int J Life Cycle Assess 23:1888–1902

Aguilera Fernández EM (2016) The influence of management practices on the greenhouse gas balance of Mediterranean cropping systems. Identifying the climate change mitigation potential through quantitative review and life cycle assessment. Doctoral thesis Universidad Pablo De Olavide, Seville

Agresti A, Franklin C (2013) Statistics. The art and science of learning from data. Third edition, Pearson

Aktas CB, Bilec MM (2012) Impact of lifetime on U.S. residential building LCA results. Int J Life Cycle Assess 17:337–349

Cespi D, Passarini F, Ciacci L, Vassura I, Castellani V, Collina E, Piazzalunga A, Morselli L (2014) Heating systems LCA: comparison of biomass-based appliances. Int J Life Cycle Assess 19:89–99

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd ed. Academic Press

Cumming G (2012) Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis. Routledge

Efron B (1979) Bootstrap methods. Another look at the jackknife. Ann Stat 7:1–26

Everitt BS, Skrondal A (2010) The Cambridge dictionary of statistics. Fourth Edition. Cambridge University Press
Röös E, Sundberg C, Hansson P-A (2011) Uncertainties in the carbon footprint of refined wheat products: a case study on Swedish pasta. Int J Life Cycle Assess 16:338–350
Qin Y, Suh S (2018) Does the use of pre-calculated uncertainty values change the conclusions of comparative life cycle assessments?—an empirical analysis. PLoS ONE 13:e0209474
Reinhard J, Wernet G, Zah R, Heijungs R, Hilty LM (2019) Contribution-based prioritization of LCI database improvements: the most important unit processes in ecoinvent. Int J Life Cycle Assess 24:1778–1792
Ribal J, Ramírez-Sanz C, Estruch V, Clemente G, Sanjuán N (2017) Organic versus conventional citrus. Impact assessment and variability analysis in the Comunitat Valenciana (Spain). Int J Life Cycle Assess 22:571–586
Rice JA (2007) Mathematical statistics and data analysis. Third edition, Thomson
Romero-Gámez M, Antón A, Leyva R, Suárez-Rey EM (2017) Inclusion of uncertainty in the LCA comparison of different cherry tomato production scenarios. Int J Life Cycle Assess 22:798–811
Ross SA, Cheah L (2017) Uncertainty quantification in life cycle assessments. Interindividual variability and sensitivity analysis in LCA of air-conditioning systems. J Ind Ecol 23:335–346
Sonnemann GW, Schuhmacher M, Castells F (2003) Uncertainty assessment by a Monte Carlo simulation in a life cycle inventory of electricity produced by a waste incinerator. J Cleaner Prod 11:279–292
Suh S, Qin Y (2017) Pre-calculated LCIs with uncertainties revisited. Int J Life Cycle Assess 22:827–831
Vercalsteren A, Spriensma C, Geerken T (2010) Life cycle assessment and eco-efficiency analysis of drinking cups used at public events. Int J Life Cycle Assess 15:221–230
Von Brömssen C, Röös E (2020) Why statistical testing and confidence intervals should not be used in comparative life cycle assessments based on Monte Carlo simulations. Int J Life Cycle Assess 25:2101–2105
Wei W, Larrey-Lassalle P, Faure T, Dumoulin N, Roux P, Mathias J-D (2016) Using the reliability theory for assessing the decision confidence probability for comparative life cycle assessments. Environ Sci Technol 50:2272–2280
Weidema BP, Bauer C, Hirschier R, Mutel C, Nemecek T, Reinhard J, Vadenbo CO, Wernet G (2013) Overview and methodology. Data quality guideline for the ecoinvent database version 3. Ecoinvent Report 1 (v3). The ecoinvent Centre
Zwillinger D, Kokoska S (2000) CRC Standard probability and statistics tables and formulae. CRC Press

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.