The International Journal of Life Cycle Assessment (2020) 25:2101–2105
https://doi.org/10.1007/s11367-020-01827-4

COMMENTARY AND DISCUSSION ARTICLE

Why statistical testing and confidence intervals should not be used in comparative life cycle assessments based on Monte Carlo simulations

Claudia von Brömssen[1] · Elin Röös[2]

Received: 24 September 2020 / Accepted: 28 September 2020 / Published online: 6 October 2020

Abstract

In the last years, it has been suggested to use statistical inferential methods, such as hypothesis testing or confidence intervals, to compare different products, services, or systems within comparative life cycle assessments based on Monte Carlo simulation results. However, the use of statistical inferential methods in such settings is fundamentally incorrect and should not be continued. In this article, we explain why and look closer at some related topics.

Keywords Statistical hypothesis testing · Inferential statistics · Monte Carlo simulations · Uncertainty · Paired simulations

1 Introduction

Descriptive statistics, e.g., mean values, medians, variances, or percentiles, can be computed to summarize basically any dataset available. Some caution needs to be taken when choosing the most informative metric for a dataset, depending on dataset characteristics and data collection processes, e.g., while mean values are a good measure for symmetric distributions, they are easily affected by single outliers or when the distribution is skewed. In such cases, a median could be a better representation of the data.

Commonly, due to data collection being expensive and time-consuming, data for the complete population cannot be collected. Imagine, for example, that we want to know the average pesticide use for different crops for a specific year and region and that this information only sits with each individual farmer. It would be too expensive to collect the data from every farmer; instead we would randomly select a sample (a number of observations) of farmers and collect the data from these farmers only. We can then calculate, e.g., the sample mean of pesticide use for a specific crop and then use this information together with the uncertainty of the mean (the standard error) to draw conclusions on the true mean for the whole population (of all farmers growing this crop). This is called inferential statistics. Another typical example would be to investigate if the difference between mean pesticide use for different crops is significant, i.e., if it is reasonable to assume that the two population means of pesticide use are, in fact, different.

Inferential statistics is, thus, the area of statistics where descriptive measures are combined with probability theory to make conclusions from a data sample to an underlying population. In contrast to descriptive statistics, several important assumptions need to be fulfilled before inferential statistics, such as hypothesis tests and confidence intervals, can be used. Many of the basic statistical tools require that observations are randomly collected, i.e., independent of each other. It is also often required that data follow a certain probability distribution, e.g., normal or Poisson, or that the dispersion within different groups of observations is equal.

In the last years, it has been suggested to use statistical inferential methods to compare different products, services, or systems within comparative life cycle assessments, when results are simulated by Monte Carlo runs (see, e.g., Henriksson et al. (2015) and Mendoza Beltran et al. (2018)). However, the use of statistical inferential methods in such settings is fundamentally incorrect as basic assumptions are not fulfilled. In this article, we will point out the problems that

Communicated by Matthias Finkbeiner

✉ Claudia von Brömssen
Claudia.von.Bromssen@slu.se

[1] Division of Applied Statistics and Mathematics, Department of Energy and Technology, Swedish University of Agricultural Sciences, P.O. Box 7032, 750 07 Uppsala, Sweden

[2] Department of Energy and Technology, Division of Agricultural Engineering, Swedish University of Agricultural Sciences, P.O. Box 7032, 750 07 Uppsala, Sweden

arise when inferential statistical methods are falsely applied and discuss on which prerequisites Monte Carlo simulations in life cycle assessment are useful.

2 The idea of independent observations as basis of statistical inference

Statistical inference is based on the information given by one or several samples with data that were randomly selected from one or several populations. The sample data is used to make generalizations about the population as a whole. Data is usually collected using a random sampling process within an experimental design or an observational study. The most important basic component in this collection of data is independence of the selected observations, certifying that the information added by an additional observation is new. While there are sampling designs that relax this assumption, they are always matched with specialized estimation theory or specific statistical modelling to account for this relaxation. Typical examples of sampling that is not completely independent are repeated measures designs, where one experimental unit is observed at several occasions over a longer time span, or paired samples, where two observations are matched according to one of several background variables before these observations are assigned to two different treatment groups.

Monte Carlo simulation for any mathematical model, e.g., an LCA model, can never produce data that meets the requirement of independence, as the output from the simulation is given by the known simulation inputs propagated through the model, and no new knowledge is gained. Simulations can be useful for visualization of uncertainty or further studies of the underlying system, but are not appropriate as input to statistical inferential methods.

3 Are there additional issues with statistical approaches for simulated data?

Even if there would be some way to justify the use of simulated data in statistical inference, there are other fundamental problems with their use in this setting. White et al. (2014) describe how the use of p values as measure of statistical significance is meaningless when data is simulated, as the p value would inflate with increased simulation runs. This means, it is easy to get significant results by doing enough simulation runs, even though no new information is added and no generalizations can be made from the results. For comparative LCA, a similar observation is made by Heijungs (2020), who however incorrectly advises to handle this by putting an upper limit to the number of simulations. As the simulations are a theoretical construct, there would be no way to determine the right amount of simulations, and any results from inferential statistics methods would still be meaningless.

An additional point raised by White et al. (2014) that is equally true for LCA is that the null hypothesis of no difference is known to be false a priori. Unless the same input parameters are given for several products or services, the simulation results cannot be the same in the mean.

4 Is there any possibility to use inferential statistical with LCA?

In case a LCA input parameter is described by actual observed data for a number of independent sites or situations, these individual observations could be propagated through the model providing equally many observations for the output parameters. In that case, the sample of output values can be viewed as transformed input data, and the independence of the observations is maintained. Then these output values could be used to relate the mean difference between products to the variability of the collected data, and a statistical test can determine if this difference is statistically significant. However, such an approach is only possible if there is a complete setup of observations for each of the input parameters within the same data sample. If data for different input parameters is collected in different studies, which is usually the case in LCA, there is no obvious way to combine these samples in a LCA and still allow the use of inferential statistics.

5 What is the use of Monte Carlo simulation in LCA?

Monte Carlo simulations are still very applicable in LCA. They are, for example, useful for producing an overall estimate of dispersion in the output parameter, which typically cannot be computed analytically due to the complex relationships and assumptions within the LCA (Röös et al. 2010; Röös et al. 2011). We can also use sensitivity analysis to identify the parameters that are most influential or uncertainty analysis to apportion the total variation to different inputs. For an overview, see Wei et al. (2015). If the LCA is very complex and a model that runs faster is needed, proxy models or emulators can be derived from simulations (Masnadi et al. 2020). What is important to remember is that with these approaches, we study the constructed LCA model, not the real world.

If the goal is comparative LCA, Monte Carlo simulations can be used to produce visualizations or simulation intervals for non-inferential comparisons between products. Examples of this are to count the frequency of pairwise preferences (Heijungs et al. 2005; Heijungs and Kleijn 2001), i.e., to compute the number of simulations where product A is better than product B. As this is a descriptive frequency measure, it is not affected by the number of simulations.

Simple simulated 95% intervals could be also be computed by selecting the 2.5% and 97.5% quantile from the simulations of the output parameter (Röös et al. 2010; Röös et al. 2011). Such an interval would include 95% of the outcomes produced by simulations and could be used to quantify simulation results instead of a plot of simulation outcomes. Such simulation intervals are not confidence interval in the meaning that they say anything about the location of the true population mean but quantify the uncertainty in the simulated system around the simulated mean.

6 Paired simulation to improve comparability

While statistical significance testing is inappropriate for simulated data, one concept described in connection to this by Henriksson et al. (2015) can still be helpful. They suggest dependent sampling to reduce variability in comparative LCA. The name given is misleading, as no data is sampled. Instead the concept could be called paired or blocked simulations. The principle behind this is that some choices or parameters are the same for several products or services and should be simulated as such. Imagine, for example, that we look at environmental impacts from agricultural products A and B and that in the production of these, the same fuels and fertilizers are used. This means that although the uncertainty in the emissions and resource use associated with these input contribute to the total uncertainty of each product, this uncertainty does not contribute to the same extend to the uncertainty in a comparison.

As an example, a very simple model of emission E is run twice: once with a common input parameter X which is varied independently for product A and B and once with the same common input parameter held constant within one simulation run (paired simulation). The simulation results are given as density plot in Fig. 1. We see hardly any difference as all input factors contribute to the overall dispersion of the simulated emissions. The advantage of paired simulation is, however, obvious when we look at the differences in emission E between paired simulations compared with the independent ones. In Fig. 2, the density plot shows the simulated differences in emission for the two products (emission of product A – emission of product B), once in yellow for all inputs simulated independently and once in blue for one of the input parameters having the same value within a simulation for both product A and product B. Paired simulations lead to substantially less variation in this comparison as much of the variation is in common for both products and, thus, is removed when differences are computed.

In addition to quantifying the simulation results using mean difference and variance of the difference, we can also use the number of simulation that results in product B having higher emissions (Röös et al. 2010; Röös et al. 2011). In independent simulations, this happens in 1675 out of 2000 simulations, while in paired simulations, product B is simulated with higher emissions in 1975 out of 2000 simulations.

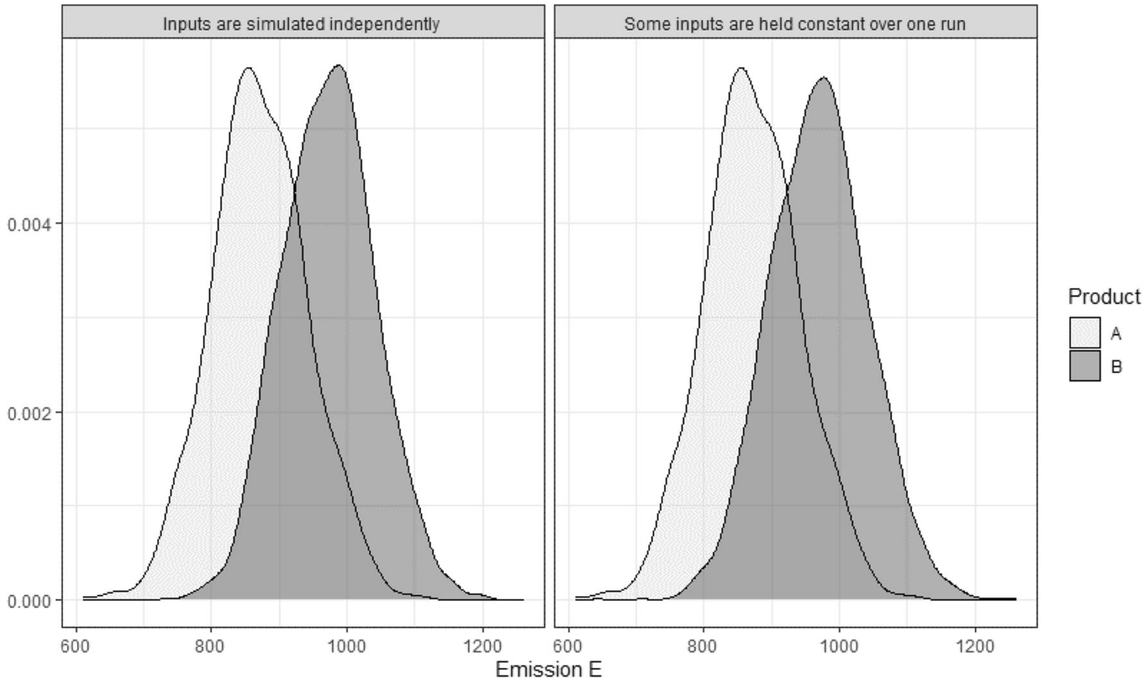

Fig. 1 Density plots for simulations of emissions from two different products A and B. In the plot to the left, input parameters were simulated independently from each other; in the plot to the right, one input parameter was held constant over a simulation run for both product A and product B. Two thousand simulations were made for each product

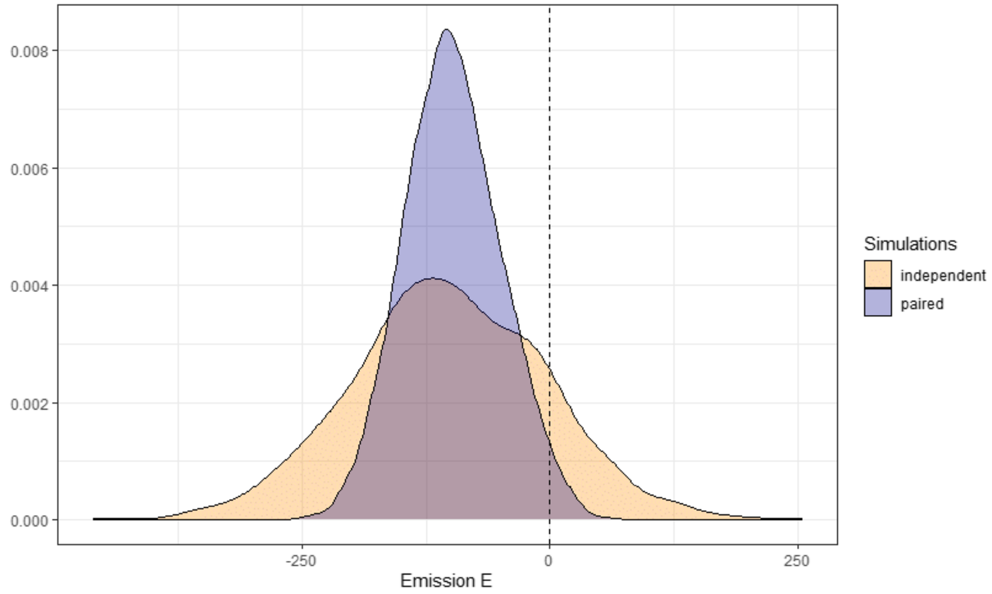

Fig. 2 Density plot for simulations of differences in emissions from two different products (emission from product A – emission from product B). Simulating all input parameters independently gives more variability in the difference in emissions (yellow). Keeping one of the inputs constant for both products removes the corresponding variation in the computation of differences (blue). The dashed line indicates the value 0, i.e., no difference in emission E between the two products

Here, it is also important to issue a caution. It is important to carefully consider which inputs are really shared between products, services, and systems and which are not. Simply pairing simulation of many inputs to get rid of variability is not a feasible way, as it would lead to a severe underestimation of the simulated variation, and no conclusions could be drawn.

7 Imperfect inputs to LCA

Heijungs (2020) discusses the use of imperfect or imprecise inputs in Monte Carlo based LCA. They describe how the use of a large number of simulations (10,000) yield a very precise result (low confidence interval) although the accuracy in the result is not improved with the number of simulation (which it would be in the case of additional samples from the real world). While Heijungs (2020) identifies the imperfect input distributions as the problem and recommends to not use Monte Carlo simulations at all when, e.g., pedigree approaches (Weidema and Wesnæs 1996) are used, the actual problem is that they use inferential statistics when this is not a valid option.

However, imperfect input data and distributions are a concern in all empirical sciences. Measurements can deviate systematically in the mean due to mistakes or bad planning of the studies, e.g., by the use of faulty instruments, unclear measurement instructions, or selection biases in the units chosen for the experiment or observational study. Such problems need, obviously, to be addressed at their source as the validity of any analysis or simulation result depends on reasonably correct inputs or data. Additionally, also the uncertainty can be imperfectly inputted into an LCA, either by under- or overestimating the variability of individual input parameters or by missing out on relevant input factors that contribute to the overall uncertainty.

Both of these imprecision do, however, not influence the functionality of Monte Carlo simulations as long as the results are interpreted considering the quality of the input data. In this context, it is also important to remember that the LCA itself is imperfect. It is a simplified model of reality often constructed for a very specific purpose and based on assumptions that are generally false and known to be false (Beven 2002; Morton 1993) and thereby never meant to be a perfect description of the real world.

8 Conclusions

Statistical inferential methods cannot and should not be used for LCA based on Monte Carlo simulations. As soon as LCA input parameters and their variability, as well as the LCA model structure, are defined, all information about this system is available, and no new knowledge can be gained by simulating. The objective of Monte Carlo simulations is instead to visualize and describe the LCA results in order to learn more about the system that has been set up. We can, for example, learn which input parameters contribute most to the variability of an output parameter, which might not be obvious if the LCA construct is complex. We can also estimate the total variation in the output parameter or compare simulation outputs for different products or services and use that information for decision support. Any kind of quantification and analysis is conditional on the assumptions made when constructing the LCA. We reproduce the world we have created and can learn from simulations about properties of this world only.

Acknowledgements Open access funding provided by Swedish University of Agricultural Sciences.

References

Beven KJ (2002) Chapter 12 uncertainty and the detection of structural change in models of environmental systems, in: Developments in Environmental Modelling. Elsevier, pp. 227–250. https://doi.org/10.1016/S0167-8892(02)80013-6

Heijungs R (2020) On the number of Monte Carlo runs in comparative probabilistic LCA. Int J Life Cycle Assess 25:394–402. https://doi.org/10.1007/s11367-019-01698-4

Heijungs R, Kleijn R (2001) Numerical approaches towards life cycle interpretation five examples. Int J Life Cycle Assess 6:141. https://doi.org/10.1007/BF02978732

Heijungs R, Suh S, Kleijn R (2005) Numerical approaches to life cycle interpretation - the case of the Ecoinvent'96 database (10 pp). Int J Life Cycle Assess 10:103–112. https://doi.org/10.1065/lca2004.06.161

Henriksson PJG, Heijungs R, Dao HM, Phan LT, de Snoo GR, Guinée JB (2015) Product carbon footprints and their uncertainties in comparative decision contexts. PLoS One 10:e0121221. https://doi.org/10.1371/journal.pone.0121221

Masnadi MS, Perrier PR, Wang J, Rutherford J, Brandt AR (2020) Statistical proxy modeling for life cycle assessment and energetic analysis. Energy 194:116882. https://doi.org/10.1016/j.energy.2019.116882

Mendoza Beltran A, Prado V, Font Vivanco D, Henriksson PJG, Guinée JB, Heijungs R (2018) Quantified uncertainties in comparative life cycle assessment: what can be concluded? Environ Sci Technol 52:2152–2161. https://doi.org/10.1021/acs.est.7b06365

Morton A (1993) Mathematical models: questions of trustworthiness. Br J Philos Sci 44:659–674

Röös E, Sundberg C, Hansson P-A (2010) Uncertainties in the carbon footprint of food products: a case study on table potatoes. Int J Life Cycle Assess 15:478–488. https://doi.org/10.1007/s11367-010-0171-8

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. https://doi.org/10.1007/s11367-011-0270-1

Wei W, Larrey-Lassalle P, Faure T, Dumoulin N, Roux P, Mathias J-D (2015) How to conduct a proper sensitivity analysis in life cycle assessment: taking into account correlations within LCI data and interactions within the LCA calculation model. Environ Sci Technol 49:377–385. https://doi.org/10.1021/es502128k

Weidema BP, Wesnæs MS (1996) Data quality management for life cycle inventories—an example of using data quality indicators. J Clean Prod 4:167–174. https://doi.org/10.1016/S0959-6526(96)00043-1

White JW, Rassweiler A, Samhouri JF, Stier AC, White C (2014) Ecologists should not use statistical significance tests to interpret simulation model results. Oikos 123:385–388. https://doi.org/10.1111/j.1600-0706.2013.01073.x

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