Simulation Studies

Angelo Moretti

Department of Computing and Mathematics

Manchester Metropolitan University, UK

A.Moretti@mmu.ac.uk

What are simulation studies and why are they important?

Simulation studies are computer experiments where data is created via pseudo-random sampling. In this way, the "truth", under some assumptions, is known and further evaluations of statistical methods are possible (Morris et al, 2017). In other words, the real world is emulated.

Simulations studies are crucial in the developing and evaluations of any statistical method. In fact, as pointed in Münnich (2014), practitioners, by changing settings of the simulation, can learn many aspects of the behaviour of a statistical method "which hardly can be found with mathematical proofs". This is an important point; in fact, it might impossible, or very difficult, to obtain always analytical results (Morris et al, 2017).

Furthermore, a vast number of scenarios can be carefully investigated and guidelines for users can be drawn from simulation studies. In the real world, only one sample is available, therefore, prior evaluations are needed.

There are many reasons why practitioners and researches should use simulation studies in the quantitative social sciences, and some of these are listed below:

- Evaluate the bias and variance of an estimator in finite samples and its consistency,
- Comparing estimators,
- Investigate how the sample size impact to the method' performances,
- Choosing the sample size when designing a study (Morris et al, 2017),
- Check whether any mistake has been made in the algebra or software (Morris et al, 2017).

In this work, the focus has been paid to simulation studies in Statistics to evaluate statistical methods. The word simulations appear in many areas of quantitative methods, such as microsimulation and agent-based models. These topics are not discussed here due to limited space.

The Components of a Simulation Study

Any simulation study should be based on some core elements: i) universe, Ω with dimension N, which is the "true" and known population, ii) an estimand/target, θ, for example a population mean or proportion or a model parameters e.g. regression coefficient, iii) sample size n, iv) number of repetitions of the sampling experiment (by construction), B. The repetitions are indexed as follows $b = 1, \dots, B$. These components need to be carefully evaluated and chosen in the designing of the simulation study.

The Steps of a Simulation Study

Morris et al. (2017) suggest some keys steps of simulation studies. In their paper, they also discuss each one in details. Therefore, the reader may want to refer to that paper for details, especially if their interest is in medical Statistics.

The first step of a simulation study is the planning. Here it is extremely important to identify the aims and the data-generating mechanisms of the simulation study. The next section on classification will discuss some important points regarding the data-generating mechanisms. It is also worthwhile to define correct and meaningful quantities we aim to estimate, if the aim is estimation (estimands). Since with the simulation study the aim is to provide guidelines to practitioners, test and evaluate methods or check whether any mistakes are present in the statistical methods, in this phase the outputs that need to be reported should be defined. In fact, these outputs will need to be included in the programme written used for the simulation study.

The next step involves the writing of the programme used to run the simulation. Morris et al. (2017) stress some useful points that researchers should consider. A crucial piece of advice is "Start small and build up code, including plenty of checks". In fact, since the simulation study will take some time to be completed, it is important to check that the programme is working correctly. If errors are made in this step, some precious time may be wasted. Another important, piece of advice is to store the random number at the start of each repetition. This is useful for replicability and in case some results are missing.

After this important step, the analysis of the results can begin. Exploratory analysis using descriptive statistics and quality measures are conducted here.

After the analysis of the results, the most important and relevant outputs, depending on the audience, are summarised and presented. Graphs, diagrams and tables (or a combination of these) may be of help to the readers.

A classification

In order to introduce the classification of simulation studies, we follow the interesting work by Münnich and Burgard (2014). The authors classify Monte Carlo simulation studies into two main categories: design-based and model-based. Other three groups of simulations are possible i.e. quasi design-based, quasi model-based and design-based under model data. However, these are not discussed in this work and we refer to Burgard, et al (2015) for more details on these.

Although, Burgard, et al. (2015) deals with simulation studies in a specific area of survey statistics i.e. small area estimation. Nevertheless, their proposed crucial principles can be extended to all the simulation studies in statistics and quantitative methods. Therefore, readers interested in simulation study maw want to refer to Burgard, et al (2015).

In design-based simulation studies, random samples, following a sampling design, are drawn from a *fixed* and *finite population*. By finite population we mean a population where the dimension is finite. The true parameters are obtained from this population. For example, if the target parameter is the mean of a variable, then the mean is calculated on the population. Thus, the impact of role of the sampling design to the statistical method is studied. In this case, the population needs to be constructed by considering real characteristics of the real population. For example, variables from the Census can be taken into account and reproduced in the experiment. These simulations studies are always important when the role of the sampling design is crucial, for example, in Official Statistics. This latter point is discussed in the following sections with examples.

The other group of simulation studies relates to model-based simulations. In this type of simulations, random samples are drawn from a *superpopulation* model. This consists on a population, generated according to a model, and this is generated every time a random sample is drawn from it. In other words, a population is generated following a model, a random sample is drawn from it and this is then repeated many times. Therefore, the true parameters are derived from the superpopulation. The concept of superpopulation is discussed in Skinner at al. (1989). As highlighted in Münnich and Burgard (2014), this setting is useful in scenarios where model

assumptions need to be evaluated under certain conditions, and to check whether the method is programmed correctly.

How does a Simulation Study Work and What are the Performance Measures?

Simulation studies are largely used to study the properties of estimators. We remind to the reader, that an estimator $\hat{\theta}$ is a statistic used to infer the value of an unknown population parameter θ. Since in the real world the population parameter will always be unknown and only one sample is observed, simulation studies may help here to evaluate the estimator properties. In order to discuss this point, we follow a simple but informative example.

Let us assume we want to estimate the population mean θ of a variable Y using two estimators i.e. the sample same $\hat{\theta}^{(1)}$ and the sample median $\hat{\theta}^{(2)}$. We also assume that Y is Normally distributed with mean θ and variance σ^2 i.e. $N(\theta, \sigma^2)$ and this is a *fixed* finite population with N dimension.

The steps of the simulation study designed in order to evaluate the problem are the following:

1. *Population Generation*: Generate independent draws $Y_1, Y_2, \dots, Y_N$ from $N(\theta, \sigma^2)$.
2. *Sampling Experiment*: Select S (many) simple random sample without replacement with indexes $s = 1, \dots, S$ from the population.
3. *Compute the Estimators*: Calculate the estimators in each sample s i.e. $\hat{\theta}_s^{(1)}$ and $\hat{\theta}_s^{(2)}$ for every sample drawn from the population $s = 1, \dots, S$.
4. *Evaluation of the Precision, Accuracy* and *Efficiency*:

   Bias. For example the bias of $\hat{\theta}^{(1)}$ is given by $Bias\left(\hat{\theta}^{(1)}\right) = S^{-1} \sum_{s=1}^{S} \hat{\theta}_s^{(1)} - \theta$.

   Mean Squared Error of $\hat{\theta}^{(1)}$. $MSE\left(\hat{\theta}^{(1)}\right) = S^{-1} \sum_{s=1}^{S} \left(\hat{\theta}_s^{(1)} - \theta \right)^2$. The same applies for estimator 2.

   Relative Efficiency of $\hat{\theta}^{(2)}$ over $\hat{\theta}^{(1)}$. $RE = \frac{MSE\left(\hat{\theta}^{(1)}\right)}{MSE\left(\hat{\theta}^{(2)}\right)}$.

The simulation steps between 2 and 4 are repeated for a large number of samples S.

Performance Measures

As pointed in Morris et al (2017), measures of performance should be analysed jointly.

The three performance measures described at step 4 are common indicators in Statistics and particularly in Official Statistics to evaluate the quality of estimators (see for example, Correa-Onel et al., 2016).

The bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated in the population. It is a measure of precision. An estimator with zero bias is called unbiased. The bias is usually an important focus of simulation studies since this quantifies whether an estimator, or more general, a statistical method "targets θ on average" (Morris et al., 2017). The mean of the estimates across all the repetitions e.g. if estimator (1) is used, $\hat{\theta}^{(1)} = \sum_s \hat{\theta}_s^{(1)} / S$, can be also presented and this can be compared with the "truth" θ. Of course, zero bias is the ideal outcome of an estimation process, but often, a "small" bias is also acceptable by considering other good properties (see e.g. Moretti et al. 2018).

The mean squared error of an estimator measures the average squared difference between the estimated values and the true value. It can be also seen as a risk function, corresponding to the expected value of the squared error loss (see Lehmann et al., 1998). This indicator measures the accuracy of an estimator.

The relative efficiency is a measure used to compare two estimators. In particular, considering the above example, when $RE < 1$ $\hat{\theta}^{(1)}$ is preferred, Thus, we say that estimator $\hat{\theta}^{(1)}$ is more efficient than estimator $\hat{\theta}^{(2)}$. This measure is widely used in Survey Statistics when variances of estimators computed under different sampling designs are compared. The reader may want to refer to Skinner at al. (1989).

Of course, true values depend very much on the scale of the variables. For example, the variable 'Income' is very different from the variable 'Height'. When estimators based on different variables are compared, relative measures need to be computed. Therefore, relative bias or relative root mean squared error should be used.

An example of a simulation study in survey statistics can be found in Moretti et al. (2018). In this paper they propose an estimator of the Mean Squared Error of a small area mean based on a mixed model. They evaluate their estimator via simulation and the performance measures described above are used to support their algorithm. The aim of the paper was to compare univariate to multivariate small area estimation and the simulation study found that multivariate small area estimation produces more efficient estimates than the univariate case. This is an

example on how simulation studies can help in the formulation and evaluation of a statistical method.

Applications in Official Statistics

Simulation studies are widely used by Statistical Agencies. In fact, in Official Statistics, there is the need to investigate whether new statistical methods can be used to produce good quality statistics before these are released. Here, some applications of simulation studies in Official Statistics are presented. Particularly, the focus has been paid to some examples used by researchers in the UK Office for National Statistics (ONS) and Statistics Netherlands.

Bates, et al. (2019) used simulation studies in synthetic data generation in order to give guidelines to users working in this area. One way to overcome privacy and confidentiality issues is the release of synthetic data in place of observed values. This is a method of statistical disclosure control for the public dissemination of microdata. Indeed, it is important to evaluate the statistical similarity between the synthetic and the true population of interest. Thus, simulation studies may help here.

De Waal (2015) compares different statistical matching methods considering a population from the Dutch Population Census 2001. National statistical institutes are interested in constructing datasets by combining available data from multiple data sources. Statistical matching can be used when different data sources contain different units with a set of common variables. The performance of the different methods in statistical matching are evaluated in his work via simulation. Since the true is assumed to be known and being the Dutch Population Census 2001, the properties of the matched data sets can be compared to it and investigate.

Another example on how simulation studies can be useful in Official Statistics is Mayhew (2016). The report deals with the issue of missingness in web scraped prices. This generates problems price indices are made from web scraped data. The aim of the work is to find the imputation method that minimises the relative imputation biases. The main finding of the simulation study it is that carrying forward the previous price is the best approach if the aim is to minimise the imputation bias.

How and What to Report

Presenting the results of a simulation study is one of the most important steps. Tables and graphs are crucial here and these need to be self-explanatory. Presenting useless and redundant outputs will mislead users.

As stressed in Morris et al. (2017), there are four dimensions that must be accounted in the summary of the results:

1. Data Generation Mechanisms
2. Methods
3. Estimands / Target of Analysis
4. Performance measures.

These dimensions need to be collocated in the right place inside a table or graphs, in order to present the results *clearly* and *unambiguously*.

Here, transparency should always be the priority. Thus, the way the data (e.g. the true population) is generated should be clearly stated at the beginning of the simulation description. For example, considering the Moretti et al. (2018), where a model-based simulation study is conducted, the model and its parameters used in order to generate the data are clearly described. Therefore, users could replicate the experiment. This is extremely important. In fact, if one of the parameters was omitted, it would be impossible to generate the population, and thus, replicate the simulation study. Moreover, if the initial parameters are obtain using some real data, this should be also mentioned in the article. When different data generation mechanisms are used in the simulation study, for example in Buil-Gil et al. (2020), these should be clearly highlighted in the Tables and Figures. Buil-Gil et al. (2020) evaluate a spatial predictor in small area estimation with applications to confidence in police work and the role of spatial autocorrelation and number of geographical areas was investigated. The authors analyse many different combinations of these and therefore, as the reader can see in their paper, the tables were created in such a way that comparisons are straightforward. For example, in the rows there are the different levels of spatial autocorrelation, and the number of areas are located in the columns. Thus, in each cell there is the performance measure of a combination of the two.

It is likely that at least two methods are compared in simulation studies, for example, comparing the mean to the median. If the outputs e.g. Tables are not too large, it is very much useful to the readers to see the performance measures close to each other, thus the different methods can

be compared visually. Here, graphs are crucial. Let us consider the following example. The goal is to compare two estimators computed for 100 geographical areas. Each area has got a different sample size, and this might affect the performance of the estimator, e.g. the design variance decreases when the sample size increases. A useful output here is a line graph where in the horizontal axis all the areas are places ordered, for example, by growing sample size, and in the vertical axis there is a performance measure, e.g. the mean squared error, averaged across all the simulation. In this way, it is possible to evaluate the behaviour of the estimator for each area varying the sample size.

Estimands or more general the targets of analysis should also be described extensively in the simulation description section. For example, if the aim is to estimate the mean of a variable in the population using an estimator developed in the paper, it is useful to relate the results to the formula described in the method section of the paper e.g. by mentioning the equation number. This is also very helpful to avoid repetition of formulas that are already given in other sections of the paper. Morris et al (2017) discuss different possible targets of analysis: estimation, testing, model selection, prediction and design of a study. We refer to their paper for a discussion on this and more examples.

The importance of the performance measures in simulation studies is already discussed in the previous section with examples. In the results section of a simulation study, it is crucial to present measures that be compared across the different methods and scenarios. Therefore, relative measures are appropriate in case of comparisons. It is important to highlight how these may be affected by all the parameters focus of the simulation; for example, sample size, intra-class correlation in case of clustered data and correlations.

Final Remarks

In this entry we discussed how simulation studies are helpful in statistics and quantitative methods. As we widely discussed in the previous sections, there are many different reasons why researchers can design and conduct simulation studies. Here, we mainly focus on

simulation studies designed in order to evaluate estimators. However, the key principles can be extended to other types of simulation studies.

An important recommendation also discussed in the introductory sections, is that before designing which type of simulation study needs to be designed, the aim should be clear. A wrong simulation study setting may give misleading information and poor evaluations.

It is also extremely important to write efficient programmes. Since simulation experiments may take a long time to run, efficient programmes will facilitate the computations involved in the experiment. Moreover, all the outputs needed for the analysis should be automatized (i.e. included in the programmes); hence, if the initial parameters are changed, new outputs are returned automatically.

Another important advantage of simulation studies is that these can be used to teach students statistical methods. By changing parameters and assumptions, students can learn how these affect the results.

A last important piece of advice is the following: always check that the generated population was generated correctly before starting the simulation. For example, if the population is generated following a linear model with some parameters, it is good practice to estimate the model on that population to see whether the true parameters values can be obtained.

References

Bates, A.G. Špakulová, I. Dove, I, Mealor, A. (2019). Synthetic data pilot. ONS methodology working paper series number 16.

Buil-Gil, D., Moretti, A., Shlomo, N. and Medina, J. (2020). Applying the Spatial EBLUP to Place-Based Policing. Simulation Study and Application to Confidence in Police Work. *Applied Spatial Analysis and Policy*.

Burgard, J. P. Münnich, R. Seger, J. and Zimmermann, T. (2015). Simulation studies for small area estimation. New Challenges for Statistical Software - The Use of R in Official Statistics in Bucharest, Romania.

Correa-Onel, S., Whitworth, A., and Piller, K. (2017). Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England. GSS Methodology Series No 42.

De Waal, T. (2015) Statistical matching: Experimental results and future research questions

Lehmann, E. L.; Casella, George (1998). Theory of Point Estimation (2nd ed.). New York: Springer. ISBN 978-0-387-98502-2. MR 1639875.

Mayhew, M. (2016). Imputing Web Scraped Prices. Report from Office for National Statistics. Available at https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/imputingwebscrapedprices.

Moretti, A., Shlomo, N. and Sakshaug, J. (2018). Parametric bootstrap mean squared error of a small area multivariate EBLUP, *Communications in Statistics - Simulation and Computation.*

Münnich, R. and Burgard, J. P. (2014). SAE Teaching using Simulations. Statistics in Transition new series and Survey Methodology, 16(4).

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). Analysis of complex surveys. New York: Wiley.