Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

INFERENTIAL STATISTICS AND SIMULATION GENERATED SAMPLES: A CRITICAL REFLECTION

Marko A. Hofmann
Silja Meyer-Nieberg
Tobias Uhlig

Department of Computer Science
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, Neubiberg 85577, GERMANY

ABSTRACT

A review of recently published papers demonstrates: simulation practitioners apply the standard methods of inferential and descriptive statistics for their reasoning with simulation generated samples without much critical reflection. Yet, simulation-generated samples differ in important aspects from empirical samples, for which the standard statistical methods have been developed. Simulation models do have inherent epistemic and computational limits for replication that do not exist with empirical data sets. Consequently, neither is simulation-based data generation the same as the collection of empirical data nor is the analysis of synthetic data equally beneficial as of empirical data. These differences are much more fundamental for computer simulation than the problems of specific techniques of inferential statistics which have been criticized recently. If simulation generated data is used for testing research hypotheses the core issue is not the method of statistical reasoning but the assurance of what might be called evidential content.

1 INTRODUCTION AND MOTIVATION

Standard statistical methods are commonly used in computer simulation without critical reflection, as our survey of recently published papers in the field can prove (see Section 3). If statistical methods are critically reflected from simulation practitioners at all, the focus lies on technical issues of specific methods. For example, White et al. (2014) and Troitzsch (2014) have (independently from each other) criticized the use of null hypothesis significance testing (NHST) in scientific applications of computer simulation. Both articles recommend to focus on *effect sizes* instead of p-values for the interpretation of simulation results. With respect to the inadequacy of NHST they reason as follows:

1. Statistical significance expressed as p-values depends on effect size AND sample size (significance $\approx$ effect size $\times$ sample size (Rosenthal and Rosnow 1991) (Sedlmeier 1996, 44)).
2. Increasing sample size with simulation systems is trivial, hence:
3. Even a minuscule effect can be demonstrated as significant using a model that allows thousands or even millions of simulation runs.

A hidden assumption of this reasoning, important for the third proposition, is made explicit by Law (2014) who used the same argument in his textbook in order to explain why he considers NHST inadequate for simulation model validation: “Since the model is only an approximation to the actual system, a null hypothesis that the system and model are the ’same’ is clearly false (p. 269)”. Subsequently, he recommends confidence intervals (p. 273), which, by the way do not solve the problem with sample size, since confidence intervals become ridiculously tiny with huge samples.

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

The formulation of the arguments (1)-(3), which can be considered as the most prominent criticism of standard statistical procedures in computer simulation, is mathematically impeccable, and in some simulation applications indeed among the arguments against statistical methods, especially if the statistical information is exclusively condensed into single p-values (Hofmann 2015b; Hofmann and Meyer-Nieberg 2018) Yet, the reasoning misses the essential two weaknesses of all standard statistical methods in computer simulation.

As an illustration of the standard reasoning consider the following experiment: We observe a process (Reference) that produces samples from the standard normal distribution $N(0,1)$. Using our Reference process we generate two sample sets containing 50 measurements. Applying a NHST (Wilcoxon-Mann-Whitney-Test) to both samples returns the expected result, we reject the hypothesis of different underlying distributions. In a second step, we fit a simple model based on a normal distribution to each sample (see Figure 1).

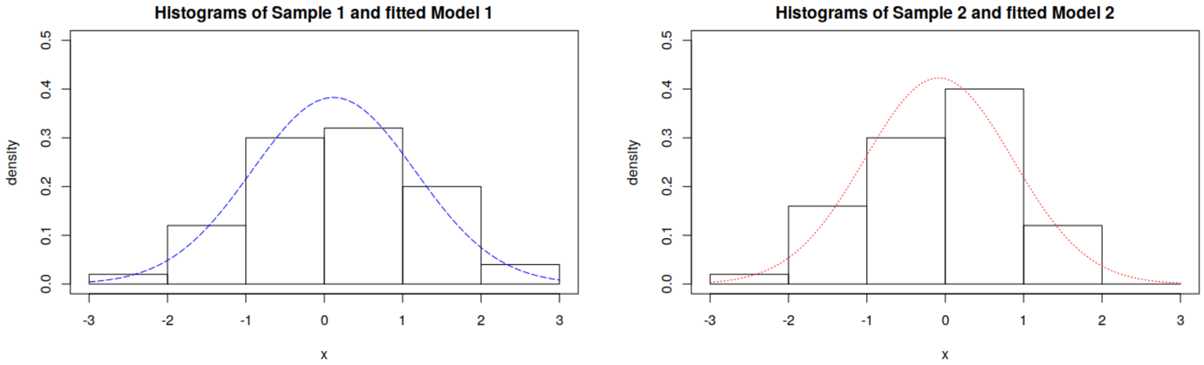

Figure 1: Histograms of two samples drawn from the same distribution with the models fitted to the samples (sample size = 50).

The resulting models differ slightly from the Reference, however, this deviation is well within the expected margins for small sample sizes. Figure 2 illustrates the probability density function (PDF) of the two models in comparison to the actual Reference process.

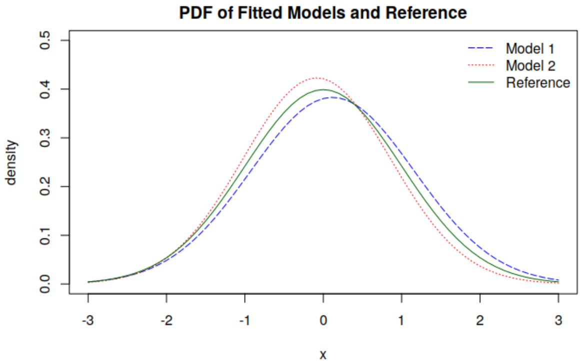

Figure 2: Comparison of the PDFs of the fitted models and the Reference.

In the next stage of the experiment, we repeat our NHST but instead of using the original samples we generate new samples with our fitted models. When we generate 50 samples with each model the NHST once again rejects the null hypothesis of different underlying distributions. This is certainly a result we would expect since both models are actually fitted to samples stemming from the same original distribution.

However, having a simulation model enables us to generate arbitrarily many additional samples. Looking at our models we see that they actually have slightly differently parametrized distributions. Accordingly, when we generate 500 samples for each model a subsequent NHST now confirms the null hypothesis. This implies that our models are meaningful different, although we know that given our experiment setup this is far from true. Arbitrarily increasing the number of samples boosts our confidence that the small differences in the models are meaningful, even though they clearly are not. This overconfidence is not an inherent problem with NHST, but is caused by putting too much trust in our simulation model. Depending on the setup and the data used for fitting our simulation models do have inherent epistemic and computational limits for the number of meaningful replications.

However, NHST can make perfect sense for simulation generated output if some preconditions are met (see Section 3.3). Futhermore, it is the equalization of generated simulation data and empirical data which is the main problem not a specific method. This paper argues that computational and the epistemological arguments are more important than statistical methods whenever simulation-generated output is interpreted.

2 CURRENT USE OF STATISTICS IN COMPUTER SIMULATION PAPERS

So, the question arises what is the current practice in simulations studies? To provide an answer, we focused on simulation studies from 2012 to 2018. We considered three dedicated general simulation journals, i.e., Simulation: Transactions of the Society for Modeling and Simulation International, the Journal of Simulation, and Simulation Modeling Practice and Theory, and identified studies using simulation models with random factors. The simulation studies stem from diverse fields ranging from health care over transportation to social sciences. In general, the research aims to compare and to assess competing strategies or to identify key factors contributing to the simulation outcome. While the topics vary considerably, discrete event simulation dominates the techniques applied. The papers covered in the literature review are summarized and categorized in Table 1.

Several studies focus on networks and communication. Imputato and Avallone (2018) for example investigate the impact of network device buffers on packet schedulers. Additionally, dynamic queue limits (DQL) recently introduced in the Linux kernel (Imputato and Avallone 2018) are analyzed. To this end, they set up a model using the discrete-event network simulator ns-3. The simulation is first compared to the results of experiments with a Linux system before a scenario-based analysis is carried out. The results are reported with the help of box-plots.

Cognitive radio ad-hoc networks (CRAHNs) which consider primary and secondary users are studied by Dung et al. (2016). They focus on the number of hops distribution between source and destination. Especially, they consider the question on how the usage of the network by the primary users affects the performance for the secondary users. The networks are modeled as geometric random graphs and investigated for varying control parameters. The analysis derives average values and the empirical hop count distribution which is provided as figures.

Romano and ElAarag (2012) presented a study in the area of Web 2.0. focusing on Web cache replacement strategies. Three quality measures were introduced and are discussed in detail in the experimental section. However, aside from reporting the frequencies and derived measures no statistical analysis appears to have been carried out.

We identified several papers addressing problems in health care. For example, Vile et al. (2017) conducted analyses of demand management strategies for emergency departments. They modeled the patient flow with a discrete event model (DEM) based on a hospital in Wales. After a model validation, the authors conducted what if analyses which were set into comparison with the baseline of the current operational procedures. The results were compared based on extremal and average values.

Bedoya-Valencia and Kirac (2016) conducted a simulation study of an emergency department. The goal was to identify means to improve the efficiency of the operations. To this end they designed a discrete event model that was implemented with Simio. After verification and validation, the model was used to simulate a base scenario and five alternatives. Statistical tests appear to have been conducted since the

Table 1: Summary and categorization of the papers covered in the literature review. The abbreviations read as follows: discrete event system (DES), agent-based system (ABS), box-plot (bp), analysis of means (ANOM), analysis of variance (ANOVA).

Author	Year	Simulation Type	Area	Analysis Type	Statistical Analysis Techniques
Imputato and Avallone	2018	DES	networks	comparison	visualization (bp)
Rahimikelarijani et al.	2018	DES	transportation	comparison	moments, ANOVA, significance tests, *p*-value
Conrads et al.	2017	DES	engineering	comparison	extremal values, average, quantiles, histograms
Vile et al.	2017	DES	health care	comparison	extremal values, average
Calle et al.	2016	DES	economics	factors	ANOM, ANOVA
Bae et al.	2016	hierach.	defense	comparisons	significance tests, *p*-value, ANOVA
Bova et al.	2016	ABS	defense	comparisons	mean, extremal values, confidence intervals
Bedoya-Valencia Kirac	2016	DES	health care	comparison	significance tests
Dung et al.	2016	random graphs	networks	factors	visualization (distribution) average values
Kim et al.	2016	real-life	health care	comparison	visualization (bp), significance tests, *p*-value
Henchey et al.	2014	DES	transportation	comparison	mean, extremal values, confidence intervals
Yates et al.	2014	ABS	social sciences	factors	ANOVA *p*-value
Lee at al.	2013	DES	health care	comparison	moments, significance tests, confidence intervals
Romano and ElArag	2012	DES	Web 2.0	comparison	frequencies, accumulated values

authors report whether comparisons with the base case yielded statistical significant results or not. However, no information is provided concerning the tests or the significance level.

Kim et al. (2016) provide a simulation analysis of the impact of wearing personal protective equipment on the performance of medical personnel during live-saving interventions in the case of warm temperatures. To this end, they considered chemical, biological, radiological, and nuclear (CBRN) disasters and conducted a randomized real-life simulation study with 20 participants. Each participant was asked to perform the tasks with and without the life-saving equipment. To evaluate the performance quality measures were defined and sampled. The comparison between the groups was carried out with the help of statistical tests – Fisher, paired t-test, and log-rank test depending on the measure. The authors additionally derive the 95% confidence intervals and provide the p-value.

Lee et al. (2013) consider patient appointment scheduling in outpatient clinics and focus on two scheduling strategies. Since their aim is to provide strategic guidelines as to how to decide when to use which strategy, they design a discrete event simulation model and provide a comparison and an analysis for several scenarios. Their analyses report statistical measures as mean and standard deviation and mentions statistical tests ($\alpha = 0.05$) without specifying the employed used. Furthermore, confidence intervals are taken into account.

Production and industrial problems are also investigated. Maintenance strategies for cutting tool replacements were considered in (Conrads et al. 2017). Conrads et al. focused on tunnel boring machines for mechanized tunneling and constructed a discrete event process model in AnyLogic. They compared several strategies and scenarios using Monte Carlo simulations with 1000 runs. Their analysis is carried out with the help of histograms, minimal, average and maximal values, and the 95% quantile.

Calle et al. (2016) present a simulation study considering inventory and production systems. They focus on hybrid fulfillment strategies and analyze the impact of different information sources concerning product availability on the performance. To this end, they first implement stochastic models of stochastic systems in ARENA before carrying out the analysis using a full factorial design based. They apply the analysis of means (ANOM) and the analysis of variance (ANOVA) to identify the main effects and the interactions of factors. Statistical significant results are reported in the case of the mean analysis whereas the p-value is provided in the case of ANOVA.

Bae et al. (2016) are among the few identified researchers that consider defense research studies. To address the problem of time-expensive simulation they suggest to re-use simulation results applying a hierarchical model tabulation technique. They present a case study for naval air defense. The focus of their study lies on demonstrating the accuracy and efficiency of their approach. The evaluation is based on error measures, visualizations, and on statistical significance tests, namely t-test, F-tests, and Kruskal-Wallis as well as the analysis of variance. Generally, they report key figures as mean and standard deviation as well as common test statistics and the p-value.

Another study considering military topics is given by Bova et al. (2016). They use an agent-based system to model secondary threats stemming from a ballistic impact events and report the design and the validation of their model. They focus on the cascading effects of impacting fragments with an aircraft. Their model validation is based on two scenarios and provides key figures as mean, minimal and maximal values as well as the half-width of the 95% confidence interval.

Transportation questions are the focus of several studies. A discrete event model was used by Rahimikelarijani et al. (2018) to assess the quality of closure strategies in the case of the Houston ship channel. After building the model, they validate and verify it based on the t-test with a significance level of 0.1. Here, the p-value is reported. The assessment of the different closure strategies utilizes the mean and standard deviation and is based on the analysis of variance. Several key figures as e.g. F-value and p-value are given. The final decision uses Fisher least significant difference (LSD) test at the significance level of 0.1.

Chen et al. (2015) address capacity planning problems for airports. To this end, they develop a discrete event model for the air side processes of the airport. The model is intended for a use in strategic planning focusing on trade-off analyses of different options. After validation, the authors provide scenario-based analyses in order to show the capabilities of their model. Here, they report average values augmented with graphical representations of the distributions.

Henchey et al. (2014) do not carry out an analysis of different strategies. Instead, they present the first stages of their research into emergency response and provide details concerning model development and validation. The focus of their study lies on a specific transportation network. Here, they built a model using ARENA and simulated vehicular crashes. The outcome of the simulation experiments is validated against real-life data. Here, the authors use the average, minimal and maximal values as well as the 95% confidence interval.

The driving forces of civil violence are investigated by Yates et al. (2014). To this end, they extend Epstein's agent-based model allowing it to operate with a GIS based representation of the space. The implementation makes use of MASON and GeoMason and is analyzed using the example case of Iran. The study aims to investigate the main factors concerning predefined key measures as e.g. the number of jailed persons. They use the analysis of variance providing the p-value.

The most impressive finding of this review is that the methods of descriptive and inferential statistics developed for *empirical data sets* are used without any reflection for *simulation-generated data*. It seems as if the interpretation of both data types can be done in exactly the same manner.

3 EMPIRICAL AND SIMULATION-GENERATED DATA SETS

3.1 A Short Overview of the Issue

The *ideal* empirical sample is composed of individual units drawn randomly from a population. Each new element of sampling has the same evidence ("evidentiary value" or "evidential content" to suggest two more specific expressions) with respect to a research hypothesis. The only basic limit for the size of empirical samples is the population size itself. A simulation-based sample, in contrast, is not drawn from a population, it is calculated on the basis of an assumed probability density function (PDF) and pseudorandom numbers. This difference creates two important additional limits for simulation-based samples. The statistical evidence that can be generated by stochastic simulation is, first, technically limited by the variability of the stochastic model (the "*variational limit*" for short). The question is simply: How many meaningful simulation runs are produced by the simulation? Second, the evidence is fundamentally or *epistemically limited* by the data used to calibrate and validate the model.

The fallacy of equating empiric and simulation-generated samples can be illustrated with a trivial example: A barber wants to enlarge his shop. He recruits a job applicant on a trial basis and estimates the possible throughput of the new shop with two barbers on the basis of a sample from a test day. The barber develops a simple stochastic queuing simulation model based on a service time distribution that best fits the data from this day. He performs 1000 simulation runs with the model and compares the results with his empirical data ($n = 1000$) from the last years. Using a significance test he is now able to demonstrate that the (non-nil) hypothesis "the average throughput cannot be augmented above the factor 1.8 (the barber's limit of profitability)" can be rejected with a p-value of 0.001. The well-educated barber concludes that this is a "3σ"-result (Grafarend 2006, 553), and employs his new colleague.

Obviously, the fault in this reasoning is not the significance test but equating the simulation-generated sample with the empiric sample. The sample from a single day, to which the simulation model is calibrated, might be biased (causing a severe epistemic limit for the number of instructive replications). This problem is independent from the method of statistical reasoning. An effect size, for example a raw effect size like the increase in throughput from that first day, is as affected by any bias from that specific day as a test. The same is true for confidence intervals (for effect sizes), which could be rendered ridiculously narrow with the same argumentation.

In addition, a second problem could be that the stochastic model of the simulation is too simple (causing a variational limit) – it might only produce, for example, 500 different trajectories. It is variationally too limited for 1000 replications. Potential reasons for such a technical restriction are bad pseudorandom number generators, simplistic PDFs, or a deficient stochastic model logic.

3.2 Going into Details

The premise (2) of the reasoning from Section 1 against NHST ("increasing sample size with simulations is trivial") confuses empirical and simulation-generated data sets with respect to the evidence they create for or against a research hypothesis in just the same way as the barber does. From 20 years of experience in the military simulation domain by the first author of this paper, we are inclined to conclude that the confusion is common: The unreflecting transfer of statistical techniques invented for empirical data into the realm of computer simulation seems to be more the rule than the exception. Any (correct) empirical data set, however, consists of a collection of facts. Increasing the (empirical) sample size implies increasing the scientifically relevant evidential content of this sample. A data set generated using a stochastic simulation is, by its own, only a list of numbers, generated via stochastic algorithms based on pseudorandom number generators and assumed distributions. All the scientific content of a simulation-generated data set depends on the empirical data set used to calibrate and validate the simulation model. Although the exact evidential content of a simulation model is, a priori, unknown, increasing the sample size via more simulation runs cannot indefinitely increase the scientific relevance of a sample. Above model specific thresholds larger samples contribute no further evidence: Either they do not generate new trajectories or they overextend the model into regions not supported by theory or fact. The fundamental difference between the two types of data sets becomes evident looking again at the "barber example": If you observe the real queue (in the service line), each new observation is a contribution to the facts known about the real distribution of service times. If, in contrast, you generate a new service time from your simulation model you only "unfold" your implemented service distribution by drawing a new pseudorandom number. The distribution in the simulation already contains all the information. In other words, if a simulation-generated sample is large enough to deduce the underlying distribution nothing further can be learned from additional sampling.

The issue is also illustrated by exhaustive sampling: Exhaustive sampling of a simulation model (the simulation reproduces only already known trajectories) is by no means equal to exhaustive sampling of a real world population. In the latter case you have assembled all the available information about a referent system, in the former case you have only unfolded model-specific assumptions based on an empirical snapshot (the underlying empirical data set) of the system. Consequently, "increasing sample size with simulations is trivial" only if one completely ignores the evidence within such a sample. Essentially. "Simulations can only deliver us results that fall within the deductive closure of our prior knowledge" (Arnold and Kästner 2013), meaning, simulation cannot provide new empirical data but only data that is already implied in its setup. The special limits of computer simulation (variational and epistemic) can be further elucidated using extreme examples.

1. A *deterministic model* has been calibrated and validated using a single empirical data set. (A deterministic model is technically equal to a stochastic model that allocates all probability mass into a single value.) Obviously, the limit for a sensible sample size per parameter setting is one. This is the technical limit addressing the number of substantially differing simulation runs produced by the model. In other words, the variational limit of this "simulation-generated" sample – indicating the technically restricted variability of the sample – is exactly one.
   On the other hand, if you change the input parameter settings of a deterministic model to new values (creating substantially new replications), the credibility into results decreases with the distance between the original (validated) and the new parameter setting. One can expect to eventually "overextend" the model's validity. This is the limit with regard to the secured epistemic content of the model (the epistemic limit of the sample). It is a fundamental but generally unknown limit.

2. Now, reconsider a *simple stochastic model* based on a single discrete PDF with only a few values for the random variable. We assume that this model is based on a bad pseudorandom number generator (PRNG) with a period of 500. The model's variational limit is 500 simulation runs.
   The empirical basis of this model is supposed to be a single historical event, a combat. A normal distribution has been assumed in order to explore how this historical scenario would have run under less or more favorable circumstances with respect to the weapons' hit probabilities. The transition from valid to plausible, implausible, and invalid values in this example is inscrutable (Hofmann 2015a). The epistemic limit for the sample size depends on the range of variation of the normal distribution, that is, σ. A very small σ might assure validity but also render the exploration uninteresting. A large σ will surely overextend the model into invalid regions. In any case, there is definitely a limit for the variation beyond which the model becomes pure speculation. This epistemic limit is often at the heart of controversial discussions about social simulation models, in general.

3.2.1 Epistemic Limits

From these simplified examples a small step leads to practice, at least with respect to epistemic limits. Most modern combat simulation systems and many social agent-based simulations are epistemically much more limited than their current usage suggests. These models regularly overextend their range of validity – often for good reasons: In essence, these models are explorative, not evaluative. An explorative model, however, does not provide statistical samples that can be treated like empirical samples. In particular, almost all the methods of inferential statistics are ill-suited for explorative simulation, which speculates on the basis of plausibility, sometimes far away from validity. It is self-deceiving to calculate narrow confidence intervals or tiny *p*-values based on such samples. Exploratory data analysis (EDA) (Tukey 1977; Hoaglin et al. 2000) is much better for exploration – as the name also suggest. One will find the same skepticism towards inferential statistics and praise of EDA in the first publications on data farming (Horne and Meyer 2005). The authors emphasize that the method has the ability "to discover trends and outlier in results [p. 1082]" and "to process large parameter spaces, makes possible the discovery of surprises (both positive and negative) and potential options [p. 1082]". Unfortunately, current usage often seems to neglect the restrictions of validity in data farming (Hofmann 2013). The reports from military data farming studies (unfortunately classified), for example, are loaded with ostentatious inferential statistics.

3.2.2 Variational Limits

The reasoning for the variational limit seems more difficult to transfer to practice, since modern PRNGs have periods (Mersenne Twister's period, for example: $2^{19937} - 1$) much higher than the computer's range of number representations (something close to 2^{64}). The actual variability of the model's output function, however, does not only depend on the PRNG. The cumulative effect of several interacting randomized algorithms on the output is, in general, not assessable in advance, but the concatenation of stochastic processes invariably creates a fundamental limit for the sensible size of all the samples generated with the model. In principle, it would always be possible to "exhaust" the model "informationally" with larger and larger samples: Above a certain sample size nothing new is ever generated with respect to the output distribution. What you see is "pseudorandom noise". The logic of the simulation model might also trivialize the output distribution despite complicated PDFs and good PRNGs in the model. Every simulation practitioner knows that a complex model can crumble into a trivial output function. The sensible sample size limit in such cases is very small. It is relatively easy, for example, to design a queuing simulation with complex service time distributions, service rules, forks and service paths that, nevertheless, produces a trivial output function. Unfortunately, such a trivialization is not always detected. If one simulates thousands of input configurations with methods like data farming, one will presumably not scrutinize each output function. Only aggregated parameters like means and variances (hiding the triviality of the output function) will be of any interest. The unreflected application of explorative simulation methods like data farming with simple models producing millions and billions of simulation runs might therefore often be seriously flawed: The model only repeats already produced trajectories. The same skepticism is recommended for agent-based simulation. These models are generally applied without any computational analysis of the output variability. One focuses on the search for interesting macro-phenomena generated by simple rules. As anecdotal evidence such a demonstration might be interesting, but it cannot be simply extended to statistical reasoning without checking for variability. It is true, that even some simple cellular automata can produce impressive complexity (some of them are even Turing-complete (Cook 2004)), others, however, are periodic or trivial (Wolfram 2002). In any case, a change of input parameters or a change of random numbers does not, in itself, guarantee sufficient output variability of computer simulations. The assessment of the computational variability of stochastic algorithms is a task in its own.

3.3 A Perfect Counterexample against the Standard Criticism against NHST: PRNGs

On the other hand, the number of substantially differing, meaningful simulation runs can be tremendous. If the epistemic content of these simulation runs can be assured (by theory and fact) or if epistemic content is dispensable than it should make perfect sense to use millions of simulation runs for inferential statistical reasoning. Fortunately, pseudorandom number generators (PRNGs) are ideal examples for the second case (irrelevance of epistemic content). Basically, PRNGs are approximate simulations of true random processes (TRP; atmospheric noise or radioactivity, for example). Huge samples (up to 10^8!) are considered to be most adequate for testing PRNGs (Soto 1999; Haramoto 2009; Rukhin et al. 2010; L'Ecuyer 2015) – as long as the samples are not larger than the total number of different numbers generated by the PRNG (its period). Only extremely large samples have enough power to find the tiny effects that can reveal a bias in modern PRNGs. With respect to such specific attributes like independence, goodness-of-fit, or lengths of runs, a PRNG is only acceptable if it is not possible to demonstrate a significant deviation from a true random process even with huge samples. Obviously, PRNGs are practical counterexamples against the standard reasoning against NHST in simulation: Although PRNGs are not equal to TRP for theoretical reasons (the nil hypothesis of randomness is false) they are tested via significance tests based on huge samples generated by algorithms that mimic (simulate) randomness. The trick is simply to consider p-values down to 10^{-4} as inconclusive. A PRNG is finally rejected only if the p-values are as low as 10^{-8} (Haramoto 2009).

With respect to the issue of this article sampling from PRNGs is special for several reasons. First, the variability of the "simulation output", (the number of different pseudorandom numbers) is central to the statistical investigation. The assessment of the computational variability of stochastic algorithms is the pivotal task of investigation. Most simulation practitioners are aware of the problem of bad PRNGs (Hellekalek 1998), and that a very bad PRNG has a small period. Hence, problems of insufficient technical variability are more visible than in applied computer simulations. Second, PRNG do not simulate anything beyond randomness, they are not intended to imitate a reference system. They are tested for specific mathematical attributes. Thus, the output of PRNGs is not further processed using probability density functions that represent the uncertainty with respect to real world processes. This post-processing renders the assessment of "normal" simulation output much more complicated, since it links PRNGs with model logic and PDFs. One can trivialize even the Mersenne Twister with a binary logic or a simple discrete PDF. In addition, the lack of further "reference to originals" releases PRNGs from the burden of any epistemic content. There is no need to validate with respect to a sample from a reference system. The only reference is the abstract mathematical true random process and its theoretical attributes. Further "epistemic content" is dispensable.

4 SUMMARY AND CONCLUSIONS

A review of recent papers shows that the difference between empirical and simulation-generated data is not considered to be a problem by most simulation practitioners. In addition, the methods of descriptive and inferential statistics are taken to be equally suitable for the interpretation of simulation results. If statistical issues are taken into consideration, it is the use of NHST and its p-values that is criticized, since they depend on sample size, which is trivial to increase with computer simulations. Yet, the methods of inferential statistics have been developed for empiric data sets. Ideally, such sets consist of individual facts, and the only limit for sample size is the population itself. Simulation-generated samples, in contrast, are based on stochastic algorithms that might have (undetected) variational limits, and surely have epistemic limits. One of these limits is generally the reason why increasing sample size with simulations does not make much sense above a certain context-dependent threshold - not the dependency of NHST on sample size.

PRNG are perfect examples for simulations that can produce meaningful huge samples that can be adequately handled by NHST. Good PRNGs are not variationally limited (periods exceed 2^{64}), and their epistemic limits are confined to specific aspects of randomness – they do not depend on further empiric data from a reference system.

The very basis of *inferential statistics* is that the whole sample consists of individual facts (units) with the same evidential content. In empirical research, this precondition is met by a correctly drawn sample from a reference population. Computer-simulation-based samples are not drawn from populations, they are generated by algorithms and may include numerous artifacts including exact repetitions and invalid trajectories. Artifacts, however, render inferential statistics useless, because it has to take every single data element seriously. The judgment-based methods of exploratory data analysis (*descriptive statistics*!) are much more adequate in such a situation.

In order to judge the epistemological limits of a simulation model providing the reader with the distribution of the empirical data used for validation is mandatory. To further assess the computational limits of a model one would like to see, how the output distributions or core parameters change for, let us say, 100, 1000 and 10000 runs.

REFERENCES

Arnold, E., and J. Kästner. 2013. "When can a Computer Simulation act as Substitute for an Experiment?". *Preprint series, Stuttgart Research*.

Bae, J. W., J. H. Kim, I.-C. Moon, and T. G. Kim. 2016. "Accelerated Simulation of Hierarchical Military Operations with Tabulation Technique". *Journal of Simulation* 10(1):36–49.

Bedoya-Valencia, L., and E. Kirac. 2016. "Evaluating Alternative Resource Allocation in an Emergency Department using Discrete Event Simulation". *Simulation* 92(12):1041–1051.

Bova, M. J., F. W. Ciarallo, and R. R. Hill. 2016. "Development of an Agent-Based Model for the Secondary Threat Resulting from a Ballistic Impact Event". *Journal of Simulation* 10(1):24–35.

Calle, M., P. L. Gonzlez-R, J. M. Leon, H. Pierreval, and D. Canca. 2016. "Integrated Management of Inventory and Production Systems Based on Floating Decoupling Point and Real-Time Information: A Simulation Based Analysis". *International Journal of Production Economics* 181:48–57.

Chen, X., J.-h. Li, and Q. Gao. 2015. "A Simple Process Simulation Model for Strategic Planning on the Airside of an Airport: a Case Study". *Journal of Simulation* 9(1):64–72.

Conrads, A., M. Scheffer, H. Mattern, M. König, and M. Thewes. 2017. "Assessing Maintenance Strategies for Cutting Tool Replacements in Mechanized Tunneling using Process Simulation". *Journal of Simulation* 11(1):51–61.

Cook, M. 2004. "Universality in Elementary Cellular Automata". *Complex Systems* 15:140.

Dung, L. T., T. D. Hieu, and S.-G. Choi. 2016. "Simulation Modeling and Analysis of the Hop Count Distribution in Cognitive Radio Ad-Hoc Networks with Shadow Fading". *Simulation Modelling Practice and Theory* 69:43–54.

Grafarend, E. W. 2006. *Linear and Nonlinear Models: Fixed Effects, Random Effects, and Mixed Models*. Berlin: Walter de Gruyter.

Haramoto, H. 2009. "Automation of Statistical Tests on Randomness to Obtain Clearer Conclusion". In *Monte Carlo and Quasi-Monte Carlo Methods 2008*, edited by P. L' Ecuyer and A. B. Owen, 411–421. Springer Berlin Heidelberg.

Hellekalek, P. 1998. "Good Random Number Generators are (not so) Easy to Find". *Mathematics and Computers in Simulation* 46:485–505.

Henchey, M. J., R. Batta, A. Blatt, M. Flanigan, and K. Majka. 2014, May. "A Simulation Approach to Study Emergency Response". *Journal of Simulation* 8(2):115–128.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 2000. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hofmann, M. 2013. "Simulation-Based Exploratory Data Generation and Analysis (Data Farming): a Critical Reflection on its Validity and Methodology". *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 10(4):381–393.

Hofmann, M. 2015a. "Reasoning beyond Predictive Validity: The Role of Plausibility in Decision-Supporting Social Simulation". In *Proceedings of the 2015 Winter Simulation Conference*, edited by M. D. Rossetti et al. Piscataway, New Jersey: IEEE.

Hofmann, M. 2015b. "Searching for Effects in Big Data: Why p-Values are not advised and what to use instead". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al. Piscataway, New Jersey: IEEE.

Hofmann, M., and S. Meyer-Nieberg. 2018. "Time to Dispense with the p-Value in OR?". *Central European Journal of Operations Research* 26(1):193–214.

Horne, G. E., and T. E. Meyer. 2005. "Data Farming: Discovering Surprise". In *Proceedings of the 2005 Winter Simulation Conference*, edited by J. A. Joines et al. Piscataway, New Jersey: IEEE.

Imputato, P., and S. Avallone. 2018. "An Analysis of the Impact of Network Device Buffers on Packet Schedulers through Experiments and Simulations". *Simulation Modelling Practice and Theory* 80:1–18.

Kim, T. H., C. H. Kim, S. D. Shin, and S. Haam. 2016. "Influence of Personal Protective Equipment on the Performance of Life-Saving Interventions by Emergency Medical Service Personnel". *Simulation* 92(10):893–898.

Law, A. M. 2014. *Simulation Modeling and Analysis (5th ed.)*. New York: Mcgraw-Hill, Inc.

L'Ecuyer, P. 2015. "Random Number Generators with Multiple Streams for Sequential and Parallel Computing". In *Proceedings of the 2015 Winter Simulation Conference*, edited by M. D. Rossetti et al. Piscataway, New Jersey: IEEE.

Lee, S., D. Min, J.-H. Ryu, and Y. Yih. 2013. "A Simulation Study of Appointment Scheduling in Outpatient Clinics: Open Access and Overbooking". *Simulation* 89(12):1459–1473.

Rahimikelarijani, B., A. Abedi, M. Hamidi, and J. Cho. 2018. "Simulation Modeling of Houston Ship Channel Vessel Traffic for Optimal Closure Scheduling". *Simulation Modelling Practice and Theory* 80:89–103.

Romano, S., and H. ElAarag. 2012. "A Quantitative Study of Web Cache Replacement Strategies using Simulation". *Simulation* 88(5):507–541.

Rosenthal, R., and R. L. Rosnow. 1991. *Essentials of Behavioral Research: Methods and Data Analysis (2nd ed.)*. New York: McGraw-Hill, Inc.

Rukhin, A., J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo. 2010. *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. Number 800-22 in NIST Special Publication. National Institute of Standards and Technology.

Sedlmeier, P. 1996. “Jenseits des Signifikanztest-Rituals: Ergaenzungen und Alternativen”. *Methods of Psychological Research Online* 1(4):41–63.

Soto, J. 1999. “Statistical Testing of Random Number Generators”. In *Proc. of the 22nd National Information Systems Security Conference*, 1–12. NIST.

Troitzsch, K. 2014. “Analysing Simulation Results Statistically: Does Significance Matter?”. In *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*, edited by H. Coelho et al., 88–105. PA, USA: Hershey.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Mass.

Vile, J. L., E. Allkins, J. Frankish, S. Garland, P. Mizen, and J. E. Williams. 2017. “Modelling Patient Flow in an Emergency Department to better understand Demand Management Strategies”. *Journal of Simulation* 11(2):115–127.

White, J., A. Rassweiler, J. Samhouri, A. Stier, and C. White. 2014. “Ecologists should not use Statistical Significance Tests to interpret Simulation Model Results”. *Oikos* 123:385–388.

Wolfram, S. 2002. *A New Kind of Science*. Champaign, Ilinois, US, United States: Wolfram Media.

Yates, J., A. Ford, and J. Kuglics. 2014. “Identifying Key Parameters and Trends in Civil Violence: a Sub-Regional, Agent-Based Simulation Approach using GIS”. *Journal of Simulation* 8(3):179–194.

AUTHOR BIOGRAPHY

MARKO HOFMANN is Chief Scientist at ITIS GmbH in Neubiberg, Germany since 2000, and adjunct Professor at the University of the Federal Armed Forces in Munich, Germany since 2010. He holds a M.S., a Ph.D. and the venia legendi in Computer Science. His email address is marko.hofmann@unibw.de.

TOBIAS UHLIG is a postdoctoral researcher at the Universität der Bundeswehr München, Germany. He holds a M.Sc. degree in Computer Science from Dresden University of Technology and a Ph.D. degree in Computer Science from the Universität der Bundeswehr München. His research interests include operational modeling, natural computing and heuristic optimization. He is a member of the ASIM and the IEEE RAS Technical Committee on Semiconductor Manufacturing Automation. He is one of the founding members of the ASIM SPL work group BeESPL. His email address is tobias.uhlig@unibw.de.

SILJA MEIER-NIEBERG is a postdoctoral researcher at the ITIS GmbH. She holds a Ph.D. degree in Computer Science from the Technical University of Dortmund. Her research interests include modeling, simulation-based optimization, and metaheuristics. Her email address is silja.meyer-nieberg@unibw.de.