World Transactions on Engineering and Technology Education
Vol.6, No.2, 2007

© 2007 UICEE

Identifying an educational void and new teaching approach for input modelling

Monica Minadeo-Cook & Mary Court

University of Oklahoma
Norman, United States of America

ABSTRACT: One of the most critical elements to a simulation study is input modelling where data are collected, prepared and analysed so as to accurately capture the activity of entities and processes for the system of interest. The ability to represent the data through probability distributions aids the computational efficiency of the simulated model. However, before utilising well-known distributions, the analyst must first ensure that the data are independent. Very few textbooks provide information on this crucial step; if they do, they either lack examples or only provide graphical means for examining patterns in the data – patterns that may suggest independence. These graphical examinations rely on the *judgement skills* of the analyst for their interpretation – a skill not usually found in students taking their first simulation course. In this article, the authors examine a set of non-parametric runs tests that relieve students from having to use their own judgement when determining independence. This study fills the educational void on testing for independence, and examines the advantages and disadvantages of employing the runs tests based on characteristics of the underlying data. The goal here is to provide guidance on teaching input modelling for an introductory simulation course.

INTRODUCTION

In general, today's simulation modelling languages handle the execution of discrete-event logic with ease, allowing a first course in simulation to focus on having the students learn discrete-event logic and simulation modelling through the use of a particular simulation language. Here, students are usually given nicely-worded problems that contain a complete description of the system in terms of its inputs and functions (processes) where the goal is to develop a simulation model of the system so as to answer a specific question (or sets of questions) on system performance. In practice, however, *nicely-worded* system descriptions are rarely, if ever, available and the practicing engineer/simulationist must conduct a simulation study from *crib-to-death* (a completely open-ended problem, ie there are no textbook descriptions in industry!).

A simulation study involves the execution of approximately eight high-level reiterative steps (as shown in Figure 1) while the conventional course objective for an introductory undergraduate course in simulation is to have students be able to take a complete description of the system of study and encode the description into the simulation language of choice (almost always chosen by the instructor). Introductory course textbooks for teaching simulation languages provide examples and problem sets where the arrival processes and service mechanisms are entirely described; the student is left with only the abstraction tasks (Step 4) and the process of verifying that his/her code accurately reflects the behaviour of the described system (Steps 5 and 6). Additionally, the student is often asked to perform some type of system analysis (Step 7), such as obtaining a confidence interval on a parameter of interest or performing a *what-if* analysis on various system levels (eg the number of resources available or their scheduling schemas). Hence, the student tends to become well versed in Steps 4 through 7 of a simulation study (see Figure 1), while the first three steps are often overlooked or perhaps provided to the student by the instructor or course textbook. Step 8 is usually not encountered until the student is able to utilise simulation in practice or is allowed to implement the results of his/her simulation study through an internship, project or capstone course. However, Step 3 of Figure 1, input modelling, can be covered in an introductory simulation course – and it should be as it is one of the most critical elements to a simulation study, since as with all computer programs, garbage-in-garbage-out!

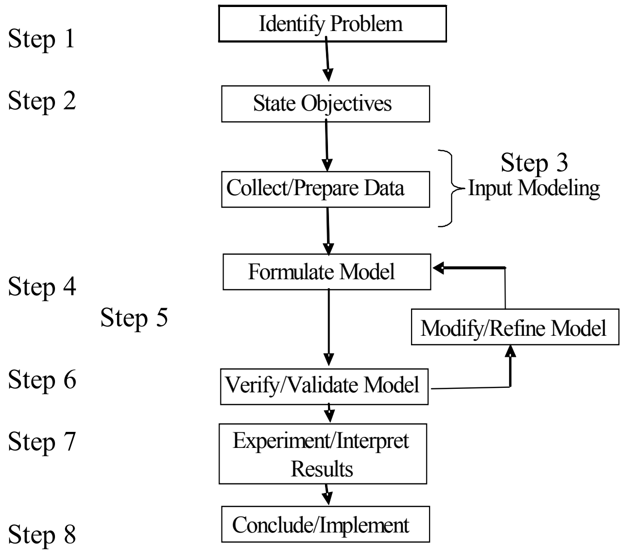

Figure 1: Eight reiterative steps of a typical simulation study.

Imperative to the ability of implementing a well-known distribution as one of the input processes is that the analyst/student ensures the data collected on that known activity are independent. Until recently, very few textbooks provided information on this crucial step, testing the data for independence [1]. Additionally, textbooks used to introduce simulation modelling languages do not contain chapters or even sections on input modelling. If they do, then they either omit the issue of independence or only provide graphical means for visualising the data [1][2]. These graphical means rely on the *judgement skills* of the analyst/student for his/her interpretation (qualitative tests). So students taking their first introductory simulation course may not have the necessary experience to make sound decisions, or they may not trust their own judgement skills. There are very few non-parametric statistical (quantitative) tools available to assist the student/ simulationist in determining whether the process shows statistical evidence that the data are independent.

A set of non-parametric independence tests are presented in this article that are used to calculate a test statistic for determining whether a set of data pass a test of independence. The calculable test statistics relieve the student/analyst from having to use his/her own judgement. The work presented fills the educational void found in textbooks that may be used to teach an introductory course in simulation modelling – particularly when the course is model or language focused and not data or analysis focused.

BACKGROUND

There are essentially four *activities* (Activity 0 - Activity III) to follow when fitting data to a well-known probability distribution. The initialisation activity assumes that the data are independent, but it has very little documentation available to test that critical assumption. Court first documented and named an *Activity 0* in 1994 for the IE5573 course, *Statistical Analysis of Simulation*, at the University of Oklahoma [3]. Activity 0 initially only involved using graphical means to see if trends or patterns existed in data via scatter plots or perhaps correlation plots. But Court also borrowed the methodologies that were being used on testing the independence of pseudo-random number generators and applied them to Activity 0 [3]. She presented those examples on her IE5573 course Web site (as taken from Banks and Carson [4]).

In 1994, the remaining activities (Activity I-III) were named and well documented in Law and Kelton [5]. However, Activity 0 is where the authors' research contributions lie – a thorough search of the literature on non-parametric tests, identifying those that may be suitable for Activity 0 and determining the best for an introductory simulation course. So only Activities I-III are briefly described in order to provide a complete definition of the activities involved in input modelling and note that Activities I-III can only be performed on data that are independent.

Activity I entails hypothesising the family using shape, where the independent data are first categorised as discrete or continuous. Activity II involves the estimation of parameters for the hypothesised distributions identified from Activity I. Typically, Maximum Likelihood Estimators (MLEs) are used to obtain estimators for the unknown distribution parameters. A list of MLEs and their calculations or associated algorithms for many well-known probability distributions are available in ref. [1]. Activity III is the goodness-of-fit tests. Here, various tests are performed with the hypothesised distribution and with the data fitted *against* the hypothesised distribution to determine a measure of the fit. The statistical tests may include the Komolgorov-Smirnov (KS) test, the Anderson-Darling (AD) test and the Chi-square test. All tests provide a p-value and hence give the analyst information on the power of the fit.

So, Activity 0 is in need of a non-parametric test that produces a test statistic that could assist the analyst/student when determining whether the data set is independent, thus providing the student with the justification to continue on to Activities I-III. Unfortunately, while there are several tests for data independence in literature, most are parametric or require knowledge about the hypothesised distribution, so they are more suitable for Activity III of input modelling and not Activity 0.

After a thorough review of literature, five potential candidates were found for supporting Activity 0: poker and gap tests [6], Pearson's Chi-square test of independence [7], Spearman's rank order correlation test [8], Sobolev test [9] and runs tests [6]. Upon further evaluation, several issues were detected with all but one of the tests: run tests. The Person's Chi-square test of independence is one of the most well-known independence tests; however, it is a parametric test – in particular, this test assumes that the data are normally distributed. The Spearman's rank order correlation test, while it is a non-parametric test, measures the correlation between observed *data pairs*. Thus, the application of this test will require the student to manipulate the data into *pseudo-pairs*. The Sobolev test, also non-parametric, determines the *geographica*l independence of data points. Again, manipulation (the implementation of a *pseudo x-y plane*) of the data is required for supporting Activity 0. While the poker and gap tests are non-parametric and leave the data in its original form, they are typically used for testing pseudo-random number generators and are usually constrained to those that generate three-digit data points. Here, the sequences of digits *within* a data point are examined for randomness, ie they do not test for independence *between* data points. The runs tests are a set of statistical tests also used to test the independence of data generated by pseudo-random number algorithms, but they allow the data to stay in its original form and they do test for independence *between* data points.

One of the authors' main observations about students in introductory simulation courses is that they tend to be confused as to the suitability and application of statistical tests when the tests require the data to be manipulated (eg the batch means method – an output analysis technique). So the goal is to eliminate tests that require data manipulation for Activity 0. Additionally, if the tests require a strong mathematical or statistical background on the part of the user and/or the test statistics is not readily calculable (eg cannot be calculated in an electronic spreadsheet), again, students tend to lack the skills set to implement the test. Thus, the runs tests are the only potential candidates most suitable for Activity 0.

Banks et al define four runs tests that are used to test the independence and uniformity of data generated by pseudo-random number algorithms: runs up and down (R_UD), runs above and below the mean (R_ABM), runs length up and down (RL_UD), and runs length above and below the mean (RL_ABM) [6].

R_UD looks at trends within the data – whether the data exhibit positive trends/behaviour and/or negative trends/ behaviour. The analyst compares the N collected data points and assigns + signs and - signs to the data indicating a positive or negative relationship between data pairs, respectively. A series of like signs constitutes a *run*, with a *run up* being made up of all + and a *run down* being made up of all -. The R_UD hypothesis test is defined as follows:

Let H_0: X_i's ~ independently, H_1: X_i's not ~ independently with α = P(reject H_0 | H_0 true):

Test statistic: $Z_0 = a-[(2N-1)/3]/[(16N-29)/90]^{1/2}$ (1)

Failure to Reject H_0: $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$

where a is the observed total number of runs both up and down.

R_ABM examines the relationship of the data points to their mean. Here, the analyst uses the sample mean to compare each data point, and all of the + and - signs are reassigned according to the distance from the sample mean. The number of *runs up* and *runs down* is also adjusted accordingly. While the acceptance region for H_0 is the same as R_UD, the R_ABM test statistic for $n_1$or n_2>20 is defined as:

Test statistic: $$Z_0 = \frac{b-(2n_1n_2/N)-½}{[2n_1n_2(2n_1n_2-N)/N^2(N-1)]^{1/2}} \quad (2)$$

where b is the observed total number of runs, n_1 is the number of individual observations above the mean (+) and n_2 is the number of individual observations below the mean (-).

RL_UD examines the number of data points with increasing (positive) and decreasing (negative) trends, and looks at the length of these run types. So here, the number of + and - signs in a *series-of-like-signs* is kept track of (the length within a run) for this test. Rather than an approximately normal test statistic, this test statistic follows the Chi-square distribution. So, RL_UD for N>20 data points has the following test statistic and acceptance criteria:

Test statistic: $$\chi^2_0 = \sum_{i=1}^{L} \frac{[O_i - E(Y_i)]^2}{E(Y_i)} \quad (3)$$

Failure to Reject H_0: $\chi^2_0 \leq \chi^2_{\alpha,L-1}$ (4)

where O_i are the observed runs of length I; *E(Y$_i$)* are the expected runs of length i; L=N-1; and:

$$E(Y_i) = [2/(i+3)!][N(i^2+3i+1)-(i^3+3i^2-i-4)]$$
$$\text{for } i \leq N-2 \quad (5)$$

and $$E(Y_i) = 2/N!$$
$$\text{for } i = N-1 \quad (6)$$

Note that the mean total number of runs, μ_a, for all run lengths i is denoted below as:

$$\mu_a=(2N-1)/3 \text{ for all i} \quad (7)$$

The same acceptance region and test statistic is used for the fourth test, RL_ABM, except that L=N for this test. Here, the sample mean is calculated and the + and - signs are determined according to the data's relationship with the mean. Next, the lengths of the runs are observed, hence the *length of runs above and below the mean*. But new calculations are required for the expected values and, as before, only hold when N>20:

$$E(Y_i) = Nw_i/E(I) \quad (8)$$

$w_{i,}$ the approximate probability that a run has length i, is:

$$w_i = (n_1/N)^i(n_2/N) + (n_1/N)(n_2/N)^i, \quad (9)$$

and E(I), the approximate expected length of a run, is:

$$E(I) = n_1/n_2 + n_2/n_1 \quad \text{for } N›20 \quad (10)$$

and E(A) is the approximate expected total number of runs of all lengths:

$$E(A) = N/E(I) \quad (11)$$

Of the four runs tests, the last two tests require the most number of calculations and rely on the student to have an understanding of the Chi-square test. So, if the student has had an experimental design course or a statistical analysis course prior to the introductory simulation course, the student should have enough prerequisite knowledge to understand the mechanics of all four tests – as is the case with the introductory simulation course mentioned above. Thus, all four runs tests are carried into the methodology section below.

METHODOLOGY AND RESULTS

The methodology described here has been built to test the robustness of the runs tests for their ability of testing data sets for independence. The following sets of distributions are proposed as the test bed:

- *Data sets known to be dependent*: dependent data sets are chosen to see if the runs tests can identify data sets that are dependent, ie reject $H_0|H_0$ is false. Here, the waiting time in queue (W_q) of an M/M/1 queue with 90% utilisation (an arrival rate of one customer per minute and a service completion rate of one customer every 0.9 minutes) was chosen. Obviously, waiting times in queues are dependent and the higher the utilisation, the higher the dependency;
- *Data sets known to be independent and symmetric*: since runs tests serve as a means for testing the robustness of pseudo-random number generators, they are most likely robust for the uniform distribution of U[0,1]. However, another well-known symmetric distribution was selected: the normal distribution with parameter set N(5,2.5);
- *Data sets known to be independent but skewed*: for the fourth data set, a distribution was chosen that is quite frequently encountered in simulation studies, ie the exponential distribution – a highly skewed distribution. This data set was chosen in order to determine the ability of the runs tests to avoid Type I errors (ie reject $H_0|H_0$ is true). The intention was to see if the RL_UD and RL_ABM tests have a tendency to reject data that come from highly skewed distributions. The reader should recall that these two tests are actually Chi-square tests; it is believed that these tests will have a tendency to reject skewed distributions merely due to the data's underlying shape. An Exp(5) is used for this.

A *rule-of-thumb* in simulation analysis is to have at least 200 data points before one fits data to a distribution, so all of the data sets for each run test contain the minimum number of points, ie 200. If the runs tests are robust with the minimum number of points, then it should be robust for larger data sets.

Each of the four runs tests are applied to 40 sets of the 200 data points generated from each test bed, ie where α = P(reject $H_0|H_0$ is true) or the p-value is set at 0.05 (5%). Thus, for α to be at its stated level of significance, no more than two of the 40 tests for a particular run test will reject H_0 when testing a set of data known to be independent (ie H_0 is true).

To generate the random variates for the data sets, the inverse transformation method was used and executed in Microsoft® *Excel* for most of the data generation. For example, -1/λ*ln(rand()) is used to generate the exponential data sets with λ =5. The Arena Input Analyzer was used to generate the normal data and then it was imported into in *Excel* for post processing [2]. The W_q data were obtained by performing the discrete-event logic in an *Excel* spreadsheet, which is a conventional homework assignment in an operations research course (typically held for second semester undergraduate industrial engineering students prior to their first simulation course). All runs tests were performed in *Excel*. So the data analysis tools chosen represent a common set of tools that an undergraduate student would possess when taking an introductory course in simulation.

The results of the calculated Type I error (α') for each of the independent data sets are shown in Table 1. The values represent the fraction of tests (out of 40) that rejected H_0. The authors suspect that the Chi-square tests, particularly RL_ABM, would be more likely to reject skewed data sets, eg Exp(5), since runs tests were designed to favour symmetric distributions. Surprisingly, the symmetric data sets (normal and uniform) also failed the RL_UD and RL_ABM tests.

Table 1: Resulting Type I errors for the 40 sets of 200 data points on each independent test bed.

α' = [number of tests rejecting H_0 | H_0 true)]/40				
Test Bed Distribution	R_UD	R_ABM	RL_UD	RL_ABM
U[0,1]	0.05	0.00	0.30	0.50
Exp(5)	0.05	0.05	0.35	0.85
N(5,2.5)	0.00	0.00	0.30	0.45

The RL_UD and RL_ABM tests also had difficulty in rejecting the dependent data sets (W_q), ie H_0 was accepted 7% and 10%, respectively, when H_0 was false, while the R_UD and R_ABM tests rejected H_0 100% of the time for the dependent data.

A PROPOSED TEACHING APPROACH

Only one set of non-parametric tests – runs tests – was found in the literature to suit the objective to have the data remain in its original form for the analysis to be calculated. Of the four runs tests, those involving the run length (RL_UD and RL_ABM) failed to uphold the stated level of significance. Additionally, those tests had a tendency to accept H_0 when H_0 was false. So the evidence suggests that both RL_ABM and RL_UD should not be used in input modelling when testing data for independence. However, R_UD and R_ABM did meet or exceed their stated level of significance for all test beds. Additionally, these tests were easier to calculate and less computationally tedious than the RL_UD and RL_ABM tests.

R_ABM and R_UD are the two recommended non-parametric runs tests to teach in an introductory simulation course for the subject, *testing data for independence when fitting data to well-known probability distributions – Activity 0*. It is recommended that students be taught these tests in the following manner:

- Students are given independent sets of data points that are known to be independent and from well-known probability distributions (eg exponential or normal). Each student has his/her own independent data set (ie while the data set is from a particular distribution, random number generators are used to provide different sets of the random variates to each student);
- Each student is *walked through the analysis* for the two runs tests, R_UD and R_ABM, eg the assignment of + and - to the data; they are then asked to calculate the corresponding test statistic;
- The results of the tests are openly discussed in the classroom. A good approach is to have the class tally the number of data sets with *failure to reject H_0* and the number of data sets with *reject H_0*. Hopefully, the percentage of *reject H_0* will match the stated level of significance. Then a discussion of the p-value will bring to light the need for students to understand how important the concept is to statistical analysis;
- Students are then given independent sets of dependent data, and steps 2 and 3 are repeated. However, a discussion of Type I and Type II errors can now take place, eg false positives and false negatives in hypothesis testing;
- To solidify the topic, students should be required to collect data from a *real-life* system (eg the time of car arrivals at an intersection) and apply the two runs tests to the collected data. This last task will give students some experience in utilising the runs tests in practice.

CONCLUSIONS AND FUTURE RESEARCH

This work represents the first study conducted on the quantitative testing of data for independence in simulation input modelling and provides the first documented approach for teaching this topic. Future research could be aimed at providing a larger test bed of the data sets – more replications of the distributions to calculate a confidence ban on the power of the tests. Additionally, the test bed could be expanded to include other distributions, eg beta and gamma distributions. The size of the data sets could also be expanded past 200 data points to see the impact of N on the power of the runs tests.

ACKNOWLEDGEMENT

The authors would like to acknowledge that the preliminary results of this research were published in the *2007 Industrial Engineering Research Conference Proceedings* CD-ROM in the paper entitled *Filling the educational void in input modelling: non-parametric tests for independence*.

REFERENCES

1. Law, A., *Simulation Modeling and Analysis* (4th edn). Boston: McGraw Hill (2000).
2. Kelton, W.D., Sadowski, R.P. and Sturrock, D.T., *Simulation with Arena* (4th edn). New York: McGraw Hill (2007).
3. Court, M.C., IE5573, Statistical Analysis of Simulation (1994), www.ecn.marycc.IE5573sp94.input.html
4. Banks, J. and Carson, J., *Discrete-Event System Simulation*. Upper-Saddle River: Prentice Hall (1984).
5. Law, A. and Kelton, W.D., *Simulation Modeling and Analysis* (2nd edn). New York: McGraw Hill (1990).
6. Banks, J., Carson, J. and Nelson, B., *Discrete-Event System Simulation* (2nd edn). Upper-Saddle River: Prentice Hall (1999).
7. Moore, D.S., The effect of dependence on Chi-squared tests of fit. *Annals of Stat., J. of Inst. of Mathematical Statistics*, 10, **4**, 1163 (1982).
8. Sheskin, D.J., *Handbook of Parametric and Nonparametric Statistical Procedures* (2nd edn). London: Chapman & Hall, 1353 (2000).
9. Jupp, P.E. and Spurr, B.D., Sobolev tests for independence of directions. *Annals of Stat., J. of Inst. of Mathematical Statistics*, 13, **3**, 1140-1141 (1985).