Calculating and Using Confidence Intervals for Model Validation

Mikel D. Petty
University of Alabama in Huntsville
301 Sparkman Drive, Shelby Center 144, Huntsville, AL 35899 USA
pettym@uah.edu 256-824-4368

Keywords
Confidence interval, Interval estimate, Validation

Abstract: *A confidence interval is an interval (i.e., a range of values) estimate of a parameter of a population (e.g., a mean) calculated from a sample drawn from the population. A confidence interval has an associated confidence level, which is frequency with which a calculated confidence interval is expected to contain the population parameter. Confidence intervals are of interest in modeling and simulation because they are often used in model validation. Typically, a set of executions of the model to be validated, which is a sample from the population of all possible executions of the model, are run and from their results a confidence interval is calculated as an estimate of the population parameter (e.g., mean model output value) that would result if all possible model executions had been run. Then, if the corresponding known or observed value for the simuland is within the confidence interval calculated from the model executions, or within some acceptable tolerance of the confidence interval's endpoints, the model is considered to be valid for the parameter in question. This paper is an introductory tutorial and survey on confidence intervals in model validation. Confidence intervals are introduced in a statistical context, their interpretation and use in model validation is explained, and examples of the application of confidence intervals in validation are presented.*

1. Introduction

Conceptually, a confidence interval is a range of values which is expected, with some quantifiable degree of confidence, to contain the value of an unknown value of interest. For example, suppose a random sample of 100 boxes of cereal is selected from among all of the boxes filled by an automatic filling machine during a work shift. The mean weight of the 100 boxes in the sample is found to be 12.05 ounces and the standard deviation to be 0.1 ounces. Using the procedures to be described in the next section, we can calculate an interval [12.0304, 12.0696] for the mean weight of all boxes filled at the station and associate a *confidence level* of 0.95 (95%) with that interval.[1] We call the calculated interval [12.0304, 12.0696], together with its associated confidence level, a *confidence interval*.

In modeling and simulation, confidence intervals are frequently used as a quantitative method of validation [1] [2]. Essentially, a confidence interval is calculated for one of the model's response variables[2], and if that confidence interval contains the known or observed value for the simuland[3] for the same response variable, the model is considered valid for that response variable.

This paper is a tutorial and survey on model validation using this method. In this paper, the statistical mathematics and their application methods are presented at a pragmatic level suitable for simulation practitioners, following the similar expository approach of [3], [4], [5], [6] and especially [7]. For readers with different interests, the same material is covered in a more conceptual and intuitive manner in [8] and [9], and with considerably more mathematical formality in [10], [11], and [12].

Following this introductory section, sections 2 and 3 of this paper constitute the tutorial material. Section 2 provides essential statistical background on point and interval estimates, the concept of a confidence interval, and procedures for calculating them, all from a statistical (i.e., non-validation) perspective. Section 3 explains the use of confidence intervals in model validation; it provides a procedure for calculating and using confidence intervals for validation, explains the conventional interpretation of confidence interval in the context of validation, and identifies when the confidence interval method is appropriate. Section 4 is the survey portion of the paper; it briefly surveys a sample of practical applications of confidence intervals in validation drawn from the simulation literature. Finally, section 5 discusses some issues associated with model validation using confidence intervals.

[1] This example is from [6].

[2] A *response variable*, also known as a dependent or output variable, is a value of interest produced by running a simulation. Examples include mean queue length in a bank lobby simulation or Red losses in a combat simulation.

[3] A *simuland* is the subject of a model; it is the object, process, or phenomenon to be simulated [2].

2. Statistical background

This section provides brief but essential statistical background.[4] Topics covered include the concepts of point and interval estimates and confidence intervals and standard procedures for calculating confidence intervals. This section is entirely statistical in content; the validation interpretation and uses of confidence intervals are covered in later sections.[5]

2.1 Confidence interval concept

Often we are interested in estimating the value of some parameter of a population, e.g., the mean income of all households in a metropolitan area. Although calculating the mean of a set of values is a simple matter, actually collecting the data for every member of a population is often impractical or infeasible. Instead, data values from a subset or sample of the population are collected, the mean of the sample values is calculated, and the sample mean statistic is interpreted as an estimate of the population mean parameter.[6] Similarly, the variability in a population parameter can be estimated by calculating the variance or standard deviation of a sample from that population.

A single valued estimate such as a sample mean is referred to as a *point estimate*. However, it is generally unlikely that the value of a point estimate will be precisely equal to the population parameter it estimates. In contrast, an *interval estimate* is a range, or interval, of values which is expected to include the population parameter value. Because the value of the population parameter is unknown, it can not be said with certainty whether a given interval includes the parameter value. A confidence interval is an interval estimate of an unknown population parameter, calculated from a sample drawn from that population, and for which there is a known and statistically justified level of confidence that the unknown population parameter falls within that interval. It is important that the confidence level be statistically justifiable; this justification will arise from the method used to calculate the confidence interval.

2.2 Calculating confidence intervals

Here we will confine our attention to calculating confidence intervals for the population mean μ, although the same principles apply to other population parameters.[7] Given a population X and a sample $x_1, x_2, \ldots, x_n$ drawn from X, the sample mean $\overline{x}$ is easily calculated as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the sample standard deviation s as

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}.$$

Intervals will be written $[L, U]$, where L is the lower bound and U is the upper bound of the interval. The general conceptual form of a confidence interval is *point estimate* ± *margin of error*, or in interval form $[L, U]$ = [*point estimate* – *margin of error*, *point estimate* + *margin of error*]. For a confidence interval for the population mean μ, the point estimate is the sample mean $\overline{x}$. The procedure for calculating the margin of error depends on the characteristics of the population from which the sample is drawn and of the sample.

To begin, we make the simple but unrealistic assumptions that the population from which the sample was drawn is known to be normally distributed and that the standard deviation σ of the population is known. Then the confidence interval for the population mean μ is

$$\left[\overline{x} - z_c \frac{\sigma}{\sqrt{n}}, \overline{x} + z_c \frac{\sigma}{\sqrt{n}} \right],$$

where z_c is the critical value for the normal distribution for confidence level c. The values for z_c can be found in statistical tables or generated by software or statistical calculators. Table 1 shows some of most commonly used critical values.[8]

[4] Because of length limits, this section can not provide complete details regarding confidence intervals from a statistical perspective; this section is meant only as a brief introduction (for readers unfamiliar with the topic) or a refresher (for readers familiar with the topic). For complete details, see [5] or [7].

[5] Readers with a strong statistical background may safely skip this section.

[6] A numerical value or measure, such as a mean, is termed a *parameter* when it applies to a population, and a *statistic* when it applies to a sample [3].

[7] Confidence intervals for population means are most often used in validation.

[8] The Student *t* distribution and the meaning of the degrees of freedom (d.f.) entries will be explained later.

Confidence level c	Normal z	Student t			
		d.f. = 5	d.f. = 10	d.f. = 20	d.f. = 30
0.80	1.282	1.476	1.372	1.325	1.310
0.90	1.645	2.015	1.812	1.725	1.697
0.95	1.960	2.571	2.228	2.086	2.042
0.99	2.576	4.032	3.169	2.845	2.750

Table 1. Common critical values for confidence intervals.

Example 1.[9] A jogger runs the same 2 mile route every day. Suppose a random sample of 90 of her times (in minutes) is taken. The population of all of her times is known or assumed to be normally distributed with a known standard deviation $\sigma = 1.80$. The sample mean $\bar{x}$ of the 90 times in the sample is found to be 15.60. Then a 95% confidence interval for the population mean μ is

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}}\right] =$$
$$\left[15.60 - 1.960\frac{1.80}{\sqrt{90}}, 15.60 + 1.960\frac{1.80}{\sqrt{90}}\right] \approx$$
$$[15.23, 15.97]$$

A more realistic assumption is that the population standard deviation σ is not known. In this situation, the population standard deviation σ is estimated using the sample standard deviation s and the Student t distribution is used in place of the normal z distribution when calculating the confidence interval. The confidence interval for the population mean μ is

$$\left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}}\right],$$

where t_c is the critical value for the Student t distribution for confidence level c.[10] The values for t_c, which can be found in statistical tables or generated by software or statistical calculators, depend not only on the confidence level c as with the z distribution, but also on the quantity $n - 1$, also known as the *degrees of freedom* (commonly abbreviated d.f.). Table 1 shows some of the most commonly used critical values for the Student t distribution. Note that for a given confidence level c, the critical values for the Student t distribution are larger than the values for the normal z distribution (although the difference decreases as d.f. increases). Consequently, for a given c using the t distribution will produce a larger interval than using the z distribution.

Example 2.[11] Using seismograph readings, a scientist estimates the yield (in kilotons) of 6 underground tests of a covert nuclear weapon by a hostile nation. The 6 sample values (45.3, 47.1, 44.2, 46.8, 46.5, 45.5) are taken from a population known or assumed to be normally distributed. The population standard deviation σ is unknown. The sample mean $\bar{x} = 45.9$ and the sample standard deviation $s \approx 1.10$. Because the sample size $n = 6$, the degrees of freedom $n - 1 = 5$, and the critical value for the t distribution for confidence level $c = 0.99$ and d.f. = 5 is $t_c = 4.032$. Then a 99% confidence interval for the population mean μ is

$$\left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}}\right] =$$
$$\left[45.9 - 4.032\frac{1.10}{\sqrt{6}}, 4.59 + 4.032\frac{1.10}{\sqrt{6}}\right] \approx$$
$$[44.09, 47.71]$$

In both of the examples, it has been assumed that the population is either known or assumed to be normally distributed, and the decision to use the z distribution or the t distribution to calculate the confidence interval was based solely on whether the population standard deviation σ was known or unknown. In fact, the question of which distribution to use is somewhat more complicated.

[9] The example is from [7].

[10] Confusingly, the t_c notation for the critical value of the Student t distribution, and the earlier z_c notation for the critical value of the normal z distribution, are only one of several notations used for critical values in the literature. For example, for the t distribution notations used to denote the critical value include t [3], t_c [7] [8], $t_{\alpha/2}$ [5] [10] [11], $t_{n-1,1-\alpha/2}$ [23], and $t_{(1-\gamma)/2}(n-1)$ [12]. (Notations for the z distribution are similar.) The different notations all refer to the same value. The confidence level $c = (1 - \alpha)$, where α is the level of significance. The subscript $\alpha/2$ denotes the area under the distribution's probability density curve in one of the two "tails" when $c = (1 - \alpha)$ area is in the center.

[11] The example is from [7], with modifications.

When choosing the distribution, three considerations are involved:

1. Population distribution: normal, approximately normal,[12] unknown.
2. Population standard deviation σ: known, unknown.
3. Sample size n: ≥ 30, < 30.

With these considerations in mind, these guidelines are used to select the distribution: [13]

If ((the population distribution is normal or approximately normal) or (the population distribution is unknown and the sample size $n \geq 30$)) and (the population standard deviation σ is known),
then calculate the confidence interval using z and σ, as shown in Example 1.

If ((the population distribution is normal or approximately normal) or (the population distribution is unknown and the sample size $n \geq 30$)) and (the population standard deviation σ is unknown),
then calculate the confidence interval using t and s, as shown in Example 2.

If (the population distribution is unknown and the sample size is < 30),
then a confidence interval can not be calculated.

2.3 Statistical interpretation of a confidence interval

It is tempting to assume that a given confidence interval $[L, U]$ with confidence level c has a probability c of containing the population mean μ. While this is intuitive, it is imprecise. In fact, for any given confidence interval $[L, U]$, the population mean μ and the confidence interval's lower and upper bounds L and U are all constants, so the confidence interval either does, or does not, contain μ; in other words, the probability that the confidence interval contains the population mean is either 1 or 0, not c [7]. The correct interpretation of confidence level c is that if many samples were taken from the population, and a confidence interval calculated from each of them at confidence level c, then $(100 \cdot c)$% of those confidence intervals would contain the true population mean μ. Thus, once a particular confidence interval has been calculated, we may be $(100 \cdot c)$% confident that it is one of the intervals that does contain μ.

3. Confidence intervals in validation

This section explains the use of confidence intervals in model validation. It presents a simple procedure for the confidence interval validation method, discusses the conventional interpretation of confidence interval in the context of validation, and identifies when the method is appropriate.

3.1 Validation method procedure

Because it involves executing the model, the confidence interval validation method is considered a dynamic method in the categorization scheme given in [1]. As with all verification and validation methods, the method involves a comparison [2]; here a given or observed value for the behavior or performance of the simuland is compared to a confidence interval for that value calculated from data obtained by executing the model.

In its simplest form, the confidence interval validation method is as follows:

1. Based on model outputs and available simuland data, select a model response variable x to use for validation.
2. Based on model execution time and statistical considerations, select a number of model executions, i.e., the sample size n.
3. Execute the model n times, recording the response variable x_i from each execution i, to produce the sample $x_1, x_2, \ldots, x_n$.
4. Calculate the sample mean $\overline{x}$ and sample standard deviation s for the model response variable from the sample $x_1, x_2, \ldots, x_n$.
5. Based on the available knowledge of the distribution of the model response variable, the availability of the population standard deviation, and the sample size, select a distribution to use (z or t) to calculate the confidence interval.
6. Select the desired confidence level c.
7. Using the selected distribution, confidence level c, and the sample statistics $\overline{x}$ and s, calculate a confidence interval $[L, U]$ for the model's mean response variable value.
8. Determine if the known simuland value y for the response variable is within the confidence interval $[L, U]$, i.e., if $L \leq y \leq U$; if it is, declare the model valid (or not invalid) for the response variable x.

---

[12] In [7], "approximately normal" is defined as "reasonably symmetrical and mound-shaped". One way to check this is by plotting and visually examining a histogram of the sample. Larger samples make this method more reliable. See [18] for a discussion of the subtleties of setting the proper bin size when plotting histograms for data from an unknown distribution.

[13] Every statistics textbook provides guidelines for selecting either z or t for constructing the confidence interval. Dismayingly, they often disagree with each other. In fact, a few even disagree with themselves, giving contradictory guidelines at different places in the same textbook. Here we present the guidelines as given in [7], although the statement of them in the form of an if-then statement is new and is not taken from [7].

Several comments regarding the simple procedure are needed. In step 2, if model execution time does not preclude it, it is recommended that at least 30 model executions be run (sample size $n \geq 30$), as this improves the statistical reliability of the calculations. In step 5, be cautious about assuming that the distribution of the model response variable is normal; even if the simuland's values for that response variable are thought to be normally distributed, assuming the same is true for the model is effectively assuming an unproven degree of validity in the model. The population distribution should be examined before making such an assumption. In step 6, there are no firm guidelines for selecting the confidence level. A confidence level of $c = 0.95$ is most frequently used in practice, and using it will likely raise no methodological objections. On the other hand, some simulation experts recommend a confidence level of $c = 0.80$ for validation [13] [14] [15], and that value is also used in practice, e.g., [16] [17]. In step 8, the simple inclusion test as described is common in practice, but it is not the only way to use the confidence interval; a more sophisticated approach is described in [18].

Example 3.[14] A discrete event simulation model of a bank drive up window is used to study customer delays, defined as the time (in minutes) a customer spends waiting in line before service begins. The model is executed 6 times, producing a sample of mean customer waiting times (2.79, 1.12, 2.24, 3.45, 3.13, 2.38). For this sample, the sample mean $\bar{x} = 2.51$ and the sample standard deviation $s = 0.82$. The population distribution is known or assumed to be normal, but the population standard deviation σ is unknown, so the t distribution will be used. Because the sample size $n = 6$, the degrees of freedom $n - 1 = 5$, and the critical value for the t distribution for confidence level $c = 0.95$ and d.f. = 5 is $t_c = 2.571$. Then a 95% confidence interval for the population mean μ is

$$\left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}}\right] =$$
$$\left[2.51 - 2.571\frac{0.82}{\sqrt{6}}, 2.51 + 2.571\frac{0.82}{\sqrt{6}}\right] \approx$$
$$[1.65, 3.37]$$

However, a value of 4.3 was previously obtained for the actual simuland mean delay time by observing customers waiting in line at the drive up window. Because 4.3 is outside the calculated confidence interval [1.65, 3.37], the model is not considered valid for customer delay.

Note that the confidence interval validation procedure as written is for a single response variable; this was for explanatory simplicity only. In practice it is more common to record multiple response variables from each model execution and calculate separate confidence intervals for each response variable. Then the model's overall validity is assessed based on whether or not a predetermined number of the confidence intervals include the corresponding simuland value, e.g., "three out of four" simuland values must be within the model confidence intervals. When using multiple response variables, practitioners often treat multiple response variables and confidence intervals as if they were independent of each other and consider each one separately. However, the independence of a model's response variables is not a given; for example, in a combat model, low Blue losses may be expected to correlate to high Red losses, and vice versa. More sophisticated approaches that consider multiple confidence intervals simultaneously have been developed [1] [13]; a simple method is shown in [16] and [17].

3.2 Validation method interpretation

For any non-trivial model executed on a digital computer, there are an extremely large but finite number of possible simulations, i.e., executions of the model.[15] The set of all possible executions of a given model may be regarded as a population, and any subset of those possible executions as a sample. As discussed earlier, a sample of values for one of the model response variables drawn from the population of all possible executions of the model can be obtained by executing the model and recording the values, and the sample mean $\bar{x}$ and standard deviation s can be calculated from the sample values. From the sample statistics and a confirmed assumption about the distribution of the population's response variable values, it is possible to calculate a confidence interval for the mean value of the response variable for the population, i.e., for all possible executions of the model. But note carefully that such a confidence interval is for the *model* (the population of all possible model executions), not for the *simuland*.

Nevertheless, the conventional validation interpretation of the confidence interval is that if the observed simuland value for the response variable is included in the model confidence interval, then the model is considered valid (or not invalid) for that response variable. There is no

[14] The example is from [18].

[15] The number of possible simulations is finite, not infinite, because any real computer has a finite memory capacity and thus there are a finite number of possible inputs (including the random number seed) to the model. Once the input is fixed, the output is also fixed, because programs executing on digital computers are deterministic. The apparent run-to-run variability in most simulations is usually due to run-to-run changes in the random number seed, of which there are a large but finite number of possible values.

statistical justification or refutation for this interpretation. Statistically, the calculated confidence interval relates to the population of possible model executions and has nothing to do with the simuland. Relating the model confidence interval to the simuland is a non-statistical application and interpretation. However, it is an interpretation that has been widely accepted and frequently used by knowledgeable modeling and simulation researchers and practitioners.

3.3 Using this validation method

As suggested in the procedure and the example, this method is most appropriate when a single value is available for the simuland for the response variable (or each of the response variables) of interest. That single value is then tested for inclusion in the confidence interval calculated from the sample of model executions. This situation may arise in validating a combat model by retrodiction, i.e., comparing the model's output to the outcome of an actual historical battle [19]. In that application, typically only one historical outcome is available. However, comparing a single historical value to a model's confidence interval is assuming that the observed historical value is the mean of the underlying distribution of historical outcomes, an assumption that can be "quite tenuous" [17]. The historical outcome could have been anomalous, and if the same battle had been fought multiple times[16] the mean outcome may be significantly different from the actual historical outcome. If instead there are multiple values (i.e., a sample) available for the simuland response variable of sufficient number to determine a distribution and its parameters, additional statistical validation methods are available, such as a hypothesis test comparing the simuland means and the model mean.[17]

4. Confidence interval validation in practice

This section is a survey of examples of actual uses of confidence intervals in validation. Selected representative and illustrative examples drawn from the literature are presented.

Example 4. In [20], workflow management in a medical clinic was studied using a discrete event simulation model of the clinic. The medical clinic offers five different types of appointments: first visit (FV), magnetic resonance imaging (MRI), computed tomography (CT), surgery pre-assessment (PRE), and surgery under anesthetic (SU). Patients at the clinic often progress through a series of appointments of different types, e.g., FV, MRI, SU. The response variables of interest were the mean time (in minutes) patients spent waiting for each type of appointment.[18] The model was complex and had a long run time,[19] so only 10 model executions were performed. Confidence intervals were calculated for each of the appointment types' waiting times as generated by the model and compared to the mean waiting times for those appointment types as observed at the actual clinic.

Table 2 shows the results. For each of the five appointment types, the table lists the observed simuland mean wait time, the sample mean $\bar{x}$ and sample standard deviation s for the wait time in the model executions, and the lower and upper bounds for the calculated confidence interval.[20] The confidence intervals were calculated using the t distribution, the sample standard deviation s as an estimate for the population standard deviation σ, confidence level $c = 0.95$, and degrees of freedom d.f. $10 - 1 = 9$. The critical value t_c for $c = 0.95$ and d.f. $= 9$ is 2.262.

Inspection of Table 2 reveals that the observed simuland mean wait time is within the confidence interval for the model mean wait time for only one of the five appointment types (MRI). In spite of this, the model developers considered the discrepancies to be small and considered the model to be "valid" [20].

Example 5. In [21], alternative improvements to a seaport's infrastructure were studied using a discrete event simulation model of the port. The port has four quays (areas where ships may be berthed, unloaded, and loaded). Ships using the port can be grouped into three categories: G1, < 60 meters in length; G2, 60-120 meters in length; and G3, > 120 meters in length. The different quays are able to process different types of ships at different speeds. The response variables of interest were the number of ships of each type berthed, unloaded, and reloaded over the course of one year. A large sample of 45 model executions was run. Confidence intervals were calculated for each of the ship type counts as generated by the model and compared to the ship type counts as observed at the actual port over a one year period.

[16] This is not an experiment anyone would like to conduct.
[17] See [18] for an example of a hypothesis test comparing two means applied to the bank drive up window example in Example 3.

[18] Although waiting times were measured in minutes, successive appointments for a patient were usually days apart, so the time values are large.
[19] Each execution required 15 hours [20].
[20] Three of the confidence interval bounds in the table are slightly different from those in [20]. This paper's author recalculated the confidence intervals using the sample statistics given in [20] and the recalculated values are reported here.

Appointment Type	Simuland mean	Model		Confidence interval	
		Mean $\bar{x}$	Std dev *s*	Lower bound *L*	Upper bound *U*
FV	11,333	11,070	182.5	10,939	11,201
MRI	7,489	7,534	451.5	7,211	7,857
CT	8,853	9,064	173.2	8,940	9,188
PRE	4,030	3,761	90.7	3,696	3,826
SU	13,733	13,069	169.3	12,948	13,190

Table 2. Confidence interval values for a medical clinic model [20].

Ship Type	Simuland count	Model		Confidence interval	
		Mean $\bar{x}$	Std dev *s*	Lower bound *L*	Upper bound *U*
G1	109	111.14	14.45	106.8	115.5
G2	169	174.42	16.07	169.6	179.2
G3	19	17.28	5.26	15.7	18.8
Total	297	303.68	35.89	292.9	314.5

Table 3. Confidence interval values for a port model [21].

Vehicle Type	Simuland count	Model		Confidence interval	
		Mean $\bar{x}$	Std dev *s*	Lower bound *L*	Upper bound *U*
Firefly	4	1.6	0.502	1.365	1.835
Cromwell	10	5.3	1.695	4.510	6.093
Halftrack	10	9.2	2.745	7.915	10.485

Table 4. Confidence interval values for a combat model [22].

Table 3 shows the results. For each of the three ship types and for the total of all types, the table lists the observed simuland ship count, the sample mean $\bar{x}$ and sample standard deviation *s* for the ship count in the model executions, and the lower and upper bounds for the calculated confidence interval.[21] The confidence intervals were calculated using the *t* distribution, the sample standard deviation *s* as an estimate for the population standard deviation σ, confidence level $c = 0.95$, and degrees of freedom d.f. 45 – 1 = 44. The critical value t_c for $c = 0.95$ and d.f. = 44 is 2.015.

Inspection of Table 3 reveals that the observed simuland ship count is within the confidence interval for the model ship count for one of the three ship types (G1) as well as the total of all three types. For ship types G2 and G3, the simuland count was outside the confidence interval, but by a very small amount. The model was considered to be valid [21].

Example 6. In [22], an entity-level constructive model of modern combat (OneSAF) was modified to model World War II combat. The modified model was validated using retrodiction, i.e., comparing the results of the model for a specific historical event with the actual historical outcome. The historical Battle of Villers-Bocage, which took place in the Normandy region of France on June 13 1944, was a small tank battle between British and German forces. Three different types of British vehicles were lost during the battle: Firefly, a U. S. Sherman tank modified to carry a British 17-pounder gun; Cromwell, a British tank armed with a 75 mm gun; and Halftrack, a lightly armored vehicle for transporting troops, weapons, and supplies. The response variables of interest were the number of British vehicles of each type destroyed during the battle. A sample of 20 model executions was run. Confidence intervals were calculated for the destroyed counts for each of the British vehicle types as generated by the model and compared to the historical destroyed counts.

Table 4 shows the results. For each of the three vehicle types, the table lists the historical destroyed count, the sample mean $\bar{x}$ and sample standard deviation *s* for the destroyed count in the model executions, and the lower and upper bounds for the calculated confidence interval.[22]

[21] One of the confidence interval bounds in the table is slightly different from those in [21]. This paper's author recalculated the confidence intervals using the sample statistics given in [21] and the recalculated values are reported here.

[22] [22] does not report the sample standard deviation *s*. However, using the other information given in [22] (sample mean, sample size, confidence level, and

The confidence intervals were calculated using the *t* distribution, the sample standard deviation *s* as an estimate for the population standard deviation σ, confidence level $c = 0.95$, and degrees of freedom d.f. 20 – 1 = 19. The critical value t_c for $c = 0.95$ and d.f. = 19 is 2.093.

Inspection of Table 4 reveals that the historical destroyed count is within the confidence interval for the model destroyed count for one of the three vehicle types (Halftrack). For vehicle types Firefly and Cromwell, the historical count was outside the confidence interval. Nevertheless, the model was considered by the developers to be "historically reasonable" [22].

5. Discussion

Although the confidence interval validation method is reassuringly quantitative, and as a validation method it is certainly more reliable than the too-frequently used "that looks about right" method, there nevertheless remain aspects of the method that call for subjective judgment by the analyst, model developer, model user, or accrediting authority.[23] Four are mentioned here.

First, in all three of the examples of the method in practice (Examples 4, 5, and 6), the simuland response variable values were outside the model confidence intervals for at least half of the response variables. In spite of those apparent problems, the models were considered "valid" or "reasonable" in every case. A decision of this sort is a judgment call that is outside the technical scope of the validation method.[24] The method provides objective quantitative input (the simuland values are inside, or outside, or the model confidence intervals) to a subjective qualitative decision (the model is valid enough, or not valid enough, for the intended application).

Second, all else being equal, a larger value for the confidence level *c* produces a larger interval, because the critical values of *z* and *t* are larger for larger values of *c* (see Table 1 to confirm), and the critical value is a multiplier in the confidence interval calculations. For example, a confidence level of $c = 0.95$ will produce a larger interval than a confidence level of $c = 0.80$. Consequently, larger confidence levels make it more likely that the simuland response variable will be inside the model confidence interval for that variable, and thus result in a less rigorous validation test. When choosing a confidence level *c*, be aware of the consequences of a Type II error (using an invalid model).[25]

Similarly, for a given confidence level *c* using the Student *t* distribution rather than the normal *z* distribution produces a larger confidence interval, as noted earlier. This again makes it more likely that the simuland response variable will be inside the model confidence interval for that variable. However, the analyst may have no choice but to use the *t* distribution if the population standard deviation is unknown.

Finally, all else being equal, a larger sample size *n* produces a smaller interval, because *n* is in the denominator of the margin of error term in the confidence interval calculations. Consequently, larger samples (i.e., more model executions) make it less likely that the simuland response variable will be inside the model confidence interval for that variable, and thus result in a more rigorous validation test. When choosing a sample size *n*, be aware of the consequences of a Type I error (not using a valid model).[26]

6. References

[1] O. Balci, "Verification, Validation, and Testing", in J. Banks (Ed.), *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, John Wiley & Sons, New York NY, 1998, pp. 335-393.

[2] M. D. Petty, "Verification, Validation, and Accreditation", in J. A. Sokolowski and C. M. Banks (Editors), *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*, John Wiley & Sons, Hoboken NJ, 2010, pp. 325-372.

[3] J. A. Parsons, Practical Mathematical and Statistical Techniques for Production Managers, Prentice-Hall, Englewood Cliffs NJ, 1973.

[4] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition*, Chapman & Hall, Boca Raton FL, 2004.

[5] D. R. Anderson, D. J. Sweeney, and T. A. Williams, *Modern Business Statistics with Microsoft Excel, Second Edition*, Thomson South-Western, Mason OH, 2006.

[6] W. Navidi, *Statistics for Engineers and Scientists*, McGraw-Hill, New York NY, 2006.

[7] C. H. Brase and C. P. Brase, *Understandable Statistics: Concepts and Methods, Ninth Edition*, Houghton Mifflin, Boston MA, 2009.

confidence interval lower and upper bounds), it was possible to solve the confidence interval equation to find the sample standard deviation.

[23] The contrast between the empirical and the social processes involved in validation is pointed out in [24].

[24] A formalization of the notion of "close enough" is described in [18].

[25] See [1] or [2] for a discussion of Type I and Type II errors in verification and validation.

[26] See [1] or [2] for a discussion of Type I and Type II errors in verification and validation.

[8] S. K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction, Second Edition*, Radius Press, New York NY, 1991.
[9] D. J. Hand, *Statistics: A Very Short Introduction*, Oxford University Press, Oxford UK, 2008.
[10] G. K. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*, John Wiley & Sons, New York NY, 1977.
[11] G. E. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery, Second Edition*, John Wiley & Sons, Hoboken NJ, 2005.
[12] M. J. Evans and J. S. Rosenthal, *Probability and Statistics: The Science of Uncertainty, Second Edition*, W. H. Freeman and Company, New York NY, 2010.
[13] O. Balci and R. Sargent, "Validation of simulation models via simultaneous confidence intervals", *American Journal of Mathematical and Management Science*, Vol. 4, No. 3-4, 1984, pp. 375-406.
[14] O. Balci, "Validation, Verification, and Testing Techniques throughout the Life Cycle of a Simulation Study", *Annals of Operations Research*, Vol. 53, No. 1, 1994, pp. 121-173.
[15] J. P. C. Kleijnen, "Statistical Validation of Simulation Models", *European Journal of Operational Research*, Vol. 87, Iss. 1, 1995, pp. 21-34.
[16] L. E. Champagne, *Development Approaches Coupled With Verification and Validation Methodologies for Agent-based Mission-level Analytical Combat Simulations*, Ph.D. Dissertation, Air Force Institute of Technology, AFIT/DS/ENS/03-02, 2003.
[17] L. E. Champagne and R. R. Hill, "Agent-Model Validation Based on Historical Data", *Proceedings of the 2007 Winter Simulation Conference*, Washington DC, December 9-12 2007, pp. 1223-1231
[18] J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation, Fifth Edition*, Prentice Hall, Upper Saddle River NJ, 2010.
[19] S. E. Barbosa and M. D. Petty, "A Survey and Comparison of Past Instances of Combat Model Validation by Retrodiction", *Proceedings of the Spring 2010 Simulation Interoperability Workshop*, Orlando FL, April 12-16 2010.
[20] R. S. Mans, N. C. Russell, W. van der Aalst, P. J. M. Bakker, and A. J. Moleman, "Simulation to Analyze the Impact of a Schedule-aware Workflow Management System", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 86, Iss. 8-9, August-September 2010, pp. 510-541.
[21] E. Demirci, "Simulation Modelling and Analysis of a Port Investment", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 79, Iss. 2, February 2003, pp. 94-105.
[22] K. M. Kelly, C. Finch, D. Tartaro, and S. Jaganathan, "Creating a World War II Combat Simulator Using OneSAF Objective System", *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference*, Orlando FL, December 4-7 2006, pp. 510-520.
[23] R. R. Mielke, "Statistical Concepts for Discrete Event Simulation", in J. A. Sokolowski and C. M. Banks (Editors), Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains, John Wiley & Sons, Hoboken NJ, 2010, pp. 25-56.
[24] S. Denize, S. Purchase, and D. Olaru, "Using Case Data to Ensure 'Real World' Input Validation within Fuzzy Set Theory Models", in A. Meier and L. Donzé, (Editors), *Fuzzy Methods for Customer Relationship Management and Marketing: Applications and Classifications*, IGI Global, Hershey PA, 2012.

7. Author's biography

Mikel D. Petty is Director of the University of Alabama in Huntsville's Center for Modeling, Simulation, and Analysis. He is also an Associate Professor of Computer Science and a Research Professor of Industrial and Systems Engineering and Engineering Management. Prior to joining UAH, he was Chief Scientist at Old Dominion University's Virginia Modeling, Analysis, and Simulation Center. He received a Ph.D. in Computer Science from the University of Central Florida in 1997. Dr. Petty has worked in modeling and simulation research and development since 1990 in areas that include human behavior modeling, simulation interoperability and composability, multi-resolution simulation, and applications of theory to simulation. He has published over 170 research papers and has been awarded over $14 million in research funding. He served on a National Research Council committee on modeling and simulation, is a Certified Modeling and Simulation Professional, and is an editor of the journals SIMULATION and Journal of Defense Modeling and Simulation. He was the dissertation advisor to the first two students to receive Ph.D.s in Modeling and Simulation at Old Dominion University.