Advanced Topics in 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, Validation, Bonferroni correction, Difference of means

Abstract

A confidence interval is an interval estimate of a parameter of a population, such as a mean, calculated from a sample drawn from the population. In addition to its endpoints, a confidence interval has an associated confidence level, which is a statistically justified degree of confidence that the interval actually contains the population parameter. Confidence intervals are often used in model validation. The model to be validated is executed multiple times; those executions compose a sample from the population of all possible executions of the model. A confidence interval is calculated from the results of the model executions as an estimate of the model's response variable that would be found if all possible model executions had been run. If the known or observed value for the simuland corresponding to the response variable is within the confidence interval, or within some acceptable tolerance of its endpoints, the model is considered to be valid for the variable in question.

This paper is a continuation of a Fall 2012 Simulation Interoperability Workshop paper; that earlier paper was an introductory tutorial and survey on the calculation and use of confidence intervals for model validation. This paper covers three advanced topics in the same area. The first is a useful quantification of the notion of "close enough" with respect to confidence interval inclusion. The second is a confidence interval adjustment applicable when multiple potentially non-independent model response variables are being validated. The third is the calculation of confidence intervals for the difference of two means. For all three of these topics, the explanations are motivated and illustrated with examples from the literature of their practical application in model validation.

1. Introduction

A confidence interval is an interval estimate of a parameter of a population, such as a mean, calculated from a sample drawn from the population [1]. In addition to its endpoints, a confidence interval has an associated confidence level, which is a statistically justified level of confidence that the interval contains the population parameter. Confidence intervals are frequently used as a quantitative method of validation [2] [3]. Essentially, a confidence interval is calculated for one of the model's response variables[1], and if the confidence interval contains the known or observed value for the simuland[2] for the same response variable, the model is considered to be valid for the response variable.

This paper is a continuation of an earlier Simulation Interoperability Workshop paper on the use of confidence intervals in validation [4]. The earlier paper provided essential statistical background on confidence intervals and how to calculate them, specified a procedure for using confidence intervals for model validation, explained the conventional validation interpretation of confidence intervals, stated when the use of confidence intervals in validation is appropriate, surveyed several practical applications of confidence intervals in validation, and discussed some issues associated with model validation using confidence intervals. In the earlier paper, the statistical mathematics and their means of application were presented at a pragmatic level suitable for simulation practitioners.

This paper is similarly pragmatic in its presentation; as before, the simulation practitioner is the target reader. However, this paper covers three more advanced topics in using confidence intervals in validation. Throughout this paper, familiarity with the earlier paper and its content will be assumed and no attempt will be made herein to re-explain the introductory material.

[1] A *response variable*, also known as an output variable, a dependent variable, a measure of interest, or a performance measure, is a variable produced by running a simulation that is of interest to the model user. Examples include mean queue length in a bank lobby simulation or Red losses in a combat simulation.

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

Following this introductory section, each of the three advanced confidence interval topics is covered in a separate section. Section 2 covers the quantification of the notion of "close enough", section 3 covers the Bonferroni correction for non-independent response variables, and section 4 covers the calculation of confidence intervals for the difference of two means. Each of the sections includes practical examples of model validation.

2. Quantifying "close enough" in confidence interval inclusion

Three examples of actual uses of confidence intervals for validation were surveyed in the earlier confidence interval paper [4]: validation of a discrete event simulation model of workflow in a medical clinic [5], validation of a discrete event simulation model of ship loading and unloading in a seaport [6], and validation of a real-time constructive model of entity-level combat [7]. In each of the applications confidence intervals were calculated for several model response variables, and in each case, at least one of the simuland values was outside the corresponding model confidence interval. Nevertheless, in each case the model was judged to be "close enough" to be useful. Clearly, the simple rule stated in [4] that the simuland value should be within the model confidence interval for the model to be considered valid is applied flexibly in practice.

As far as could be determined from the sources, these practical assessments of "close enough" were subjective and qualitative. However, an objective and quantitative means of assessing whether a simuland value is "close enough" to a calculated confidence interval is available [8]. Not only does the method enable increased consistency in assessing whether a simuland value is "close enough" to a confidence interval, it also provides objective and quantitative guidance as to whether a confidence interval is narrow enough to be suitable for validation; if the interval is found to be too wide, performing additional model executions (i.e., larger n) will narrow it [4].

To explain the method, we begin by introducing essential notation:[3]

X	Population of all possible model executions
x_i	Model response variable value for execution i
n	Number of model executions, i.e., sample size
$\bar{x}$	Model response variable mean for sample, (n executions)
s	Model response variable standard deviation for sample (n executions)
$[L, U]$	Confidence interval; L is lower bound, U is upper bound
μ	Model response variable mean for all model executions; unknown
μ_0	Simuland value for response variable; known
ε	Error tolerance for "close enough"

Of particular interest in this method is the error tolerance value ε. This value is selected by the person performing the validation based on the model's intended application. For some applications, e.g., a model of the area damaged by a nuclear explosion, a larger error tolerance might be acceptable; for other applications, e.g., a model of the spatial position of a scalpel wielded by a robotic surgery device, a smaller error tolerance might be preferred. In any case, the value of ε should be set based on the application before the validation calculations are performed so as to preserve objectivity.

The quantity of interest in validation is the difference $|\mu_0 - \mu|$ between the simuland value μ_0 and the model response variable mean μ. We would like to know if the difference is less than the error tolerance, i.e., if $|\mu_0 - \mu| \le \varepsilon$. Unfortunately, the model response variable mean μ for all model executions is unknown; consequently, we estimate its value with a confidence interval $[L, U]$ calculated from a sample of n executions of the model. Therefore the calculated confidence interval $[L, U]$ for μ will be used to determine if $|\mu_0 - \mu| \le \varepsilon$.

To apply the method, first the confidence interval $[L, U]$ for μ is calculated in the usual way from sample mean $\bar{x}$ and sample standard deviation s found by executing the model n times. Two additional quantities, "best case error" b and "worst case error" w, are calculated as $b = \min(|\mu_0 - L|, |\mu_0 - U|)$ and $w = \max(|\mu_0 - L|, |\mu_0 - U|)$. Then the following rules and sub-rules are applied:

(1) if $\mu_0 < L$ or $\mu_0 > U$ then
  (1a) if $b > \varepsilon$ then model not valid
  (1b) if $w < \varepsilon$ then model valid
  (1c) if $b \le \varepsilon$ and $w > \varepsilon$ then more executions needed

(2) if $\mu_0 \ge L$ and $\mu_0 \le U$ then
  (2a) if $w \le \varepsilon$ then model valid
  (2b) if $w > \varepsilon$ then more executions needed

These rules can be explained through reference to Figure 1. The two cases labeled (1) in the figure are those where μ_0 is outside the confidence interval $[L, U]$; the difference between the cases in the figure is whether μ_0 is $> U$ or $< L$, but rule (1) applies to both. In either case, without an error tolerance, this situation might normally be interpreted as the model being not valid for the response variable.

[3] All of this notation is consistent with the earlier paper [4], except the following: model response variable mean μ was y in that paper and simuland response variable value μ_0 and error tolerance ε are new to this paper.

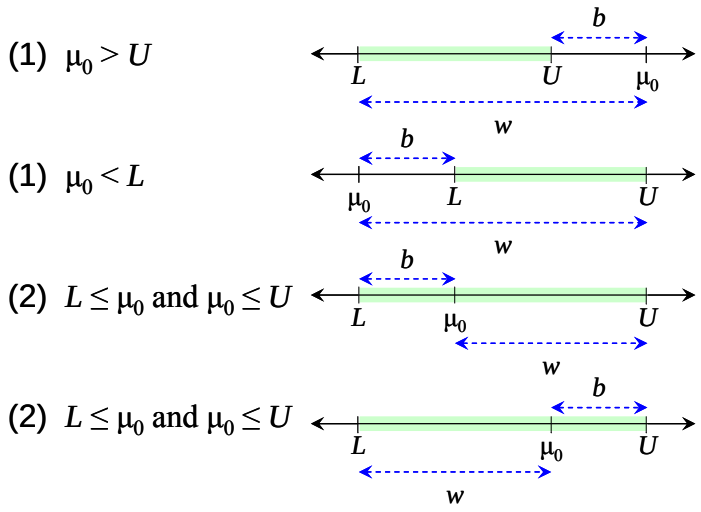

Figure 1. Cases for the "close enough" rules.

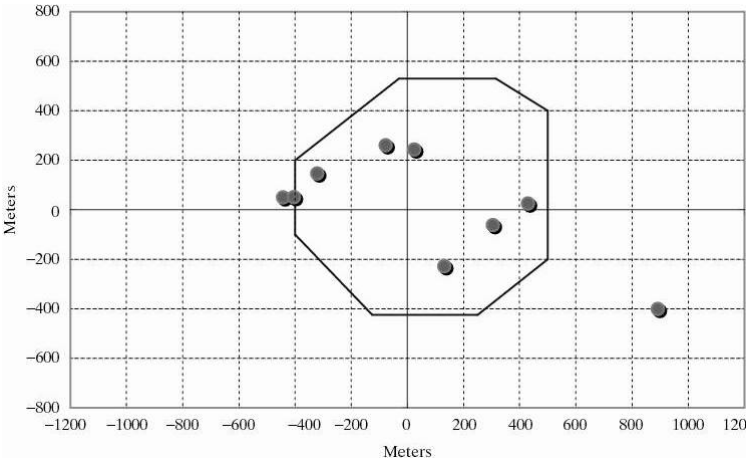

Figure 2. Example Monte Carlo bomb impact points [8].

With the error tolerance, the interpretation is as follows. In sub-rule (1a), μ_0 is outside the interval and outside the error tolerance ε, i.e., not "close enough", and the model is not valid. In sub-rule (1b) μ_0 is outside the interval but inside the error tolerance ε, i.e., "close enough", and the model is valid. In sub-rule (1c) the confidence interval is too wide with respect to the error tolerance ε and additional executions are needed to narrow the interval before an assessment can be made.

The two cases labeled (2) in the figure are those where μ_0 is inside the confidence interval [*L*, *U*]; the difference between the cases in the figure is whether μ_0 is closer to *U* or to *L*, but rule (2) applies to both. In either case, without an error tolerance, this situation might normally be interpreted as the model being valid for the response variable. With the error tolerance, the interpretation is as follows. In sub-rule (2a) μ_0 is inside the interval and the model is valid. In sub-rule (2b), μ_0 is inside the interval but the confidence interval is too wide with respect to the error tolerance ε and additional executions are needed to narrow the interval before an assessment can be made.

Example 1. In [8], a Monte Carlo model of bombing accuracy is presented. The model calculates the number of conventional (unguided) bombs that hit within a prescribed target perimeter out of a planeload of 10 bombs. For each bomb, the model stochastically generates random variates for the *x* and *y* errors based on probability distributions modeling the bomb aiming system's accuracy and then uses those *x* and *y* errors to deterministically determine whether the bomb impacts within the target's perimeter. Figure 2 illustrates the impact points generated by the model for a single planeload of 10 bombs and a typical target perimeter.[4]

[4] Only nine bomb impact points are visible in the figure; the tenth is outside the area represented by the figure.

i	x_i	i	x_i
1	6.48	11	5.38
2	5.27	12	5.73
3	7.24	13	6.67
4	8.73	14	7.48
5	7.70	15	6.04
6	8.59	16	9.62
7	9.37	17	7.77
8	7.08	18	8.75
9	7.65	19	9.10
10	5.23	20	7.09

Table 1. Response variable results from version 1 of the bombing accuracy model.

i	x_i	i	x_i	i	x_i
1	7.15	11	6.58	21	5.28
2	7.67	12	7.72	22	7.37
3	4.32	13	6.85	23	6.27
4	5.93	14	4.22	24	6.81
5	4.00	15	5.66	25	6.99
6	5.03	16	3.90	26	5.85
7	6.82	17	5.51	27	6.05
8	6.68	18	5.16	28	5.54
9	4.59	19	5.86	29	7.49
10	4.55	20	6.50	30	6.25

Table 2. Response variable results from version 2 of the bombing accuracy model.

n	d.f.	$\bar{x}$	s	t_c	$[L, U]$	$U - L$	b	w	Rule	Outcome
5	4	7.08	1.30	2.776	[5.47, 8.70]	3.23	0.53	2.70	(2b)	More runs
10	9	7.33	1.39	2.262	[6.34, 8.33]	1.99	0.34	2.33	(1c)	More runs
20	19	7.35	1.39	2.093	[6.70, 8.00]	1.30	0.70	2.00	(1a)	Not valid

Table 3. "Close enough" results for bombing accuracy model version 1.

n	d.f.	$\bar{x}$	s	t_c	$[L, U]$	$U - L$	b	w	Rule	Outcome
5	4	5.81	1.64	2.776	[3.78, 7.85]	4.07	1.85	2.22	(2b)	More runs
10	9	5.67	1.34	2.262	[4.72, 6.63]	1.91	0.63	1.28	(2b)	More runs
20	19	5.74	1.23	2.093	[5.16, 6.31]	1.15	0.31	0.84	(2b)	More runs
30	29	5.95	1.12	2.045	[5.53, 6.37]	0.84	0.37	0.47	(2a)	Valid

Table 4. "Close enough" results for bombing accuracy model version 2.

For the sake of this example, we assume that through some means, such as live testing or data collection from operational experience, the simuland value μ_0 for the expected number of bomb hits is known to be 6 out of 10. Two versions of the bombing accuracy model will be validated using the confidence interval "close enough" method. Version 1 of the model is moderately biased with respect to the simuland; whereas the simuland value $\mu_0 = 6$, for the population of all possible model executions the model response variable mean $\mu = 6.8$ and the standard deviation $\sigma = 1.50$. It is important to recall that these values for μ and σ for the population of all possible model executions are not available to the validation analyst; he or she estimates them based on a sample of n model executions. The unknown values are given in this example for expositional clarity.

Based on the intended uses for the bombing accuracy model, the validation analyst selects an error tolerance $\varepsilon = 0.5$. He or she then executes the model. Each execution of the model performs 400 Monte Carlo trials of 10-bomb planeloads. Table 1 lists the values returned for 20 executions of version 1 of the model; the values in the table are the mean number of hits x_i over the 400 trials in execution i for each of the 20 executions.[5] Table 2 shows data from 30 executions of version 2 of the bombing accuracy model, which will be discussed later.

Table 3 shows the results of applying the "close enough" method to the results of model version 1 in Table 1. In the table, n is the number of model executions (i.e., the sample size), d.f. is the degrees of freedom to use when calculating the confidence interval (d.f. = $n - 1$), $\bar{x}$ is the mean number of hits from executions 1 to n, s is the standard deviation in the number of hits for the n executions, t_c is the critical value for the Student t distribution for confidence level $c = 0.95$ and degrees of freedom d. f., $[L, U]$ is the confidence interval calculated

[5] The x_i values in the tables were generated by an actual implementation of the bombing accuracy model as described briefly in this example and in more detail in [8]. The values in the table are rounded to two digits after the decimal point.

for μ from the n executions, $U - L$ is the width of the interval, b and w the best and worst case errors calculated as described earlier, Rule identifies the rule that fits the values in the row, and Outcome is the assessment of model validity corresponding to the Rule. The first row in the table, with sample size $n = 5$, shows the outcome after the validation analyst performs 5 model executions, generating results $x_1, x_2, \ldots, x_5$ in Table 1; for $n = 5$, $\mu_0 = 6$ is inside the confidence interval [5.47, 8.70] but the best case error b is larger than the error tolerance ε, leading to an assessment via sub-rule (2b) that more executions are required to narrow the interval.

Based on that assessment, the validation analyst performs 5 additional executions of the model; combined with the five already performed, the model has been executed 10 times, generating results $x_1, x_2, \ldots, x_{10}$ in Table 1. The second row in the table, with sample size $n = 10$, shows the outcome. The additional executions reduced the width of the confidence interval from 3.23 to 1.99. Now μ_0 is outside the interval, but quite close to it; in fact, the best case error b is less than the error tolerance ε. However, the confidence interval is still too wide with respect to ε, again leading to an assessment that more executions are required, this time via sub-rule (1c). Finally, the analyst performs 10 additional executions of the model, giving a total of 20 executions, and analyzes results $x_1, x_2, \ldots, x_{20}$. This time μ_0 is again outside the confidence interval and the best case error b is larger than the error tolerance ε, leading to an assessment via sub-rule (1a) that the model is not valid. Recall that simuland value $\mu_0 = 6$, the unknown model response variable mean $\mu = 6.8$, and the error tolerance was set at $\varepsilon = 0.5$, so this conclusion regarding the model is correct.

Motivated by the assessment that the model is not valid, the model is improved. Version 2 of the model is slightly biased with respect to the simuland; for the population of all possible model executions the model response variable mean $\mu = 6.3$ and the standard deviation $\sigma = 1.30$. Again, these values for μ and σ are not available to the validation analyst. Table 2, on the left, lists the values returned for 30 executions of version 2 of the model.

As with version 1 of the model, the validation analyst performs an iteratively sequence of model executions and analyses using the "close enough" rules. The results are shown in Table 4. For $n = 5$, $n = 10$, and $n = 20$, sub-rule (2b) is selected and more executions are required. Finally, for $n = 30$, μ_0 is inside the confidence interval and the worst case error w is less than the error tolerance ε, leading to an assessment via sub-rule (2a) that the model is valid. Recall that simuland value $\mu_0 = 6$, the unknown model response variable mean $\mu = 6.3$, and the error tolerance was set at $\varepsilon = 0.5$, so this conclusion regarding the model is correct.

This method has both advantages and disadvantages. Most importantly, it allows the notion of "close enough" to be applied in a quantitative and objective way, which is clearly an improvement over the flexible and informal way it is often applied in practice. It also enables, in situations where an execution of the model may be expensive or time consuming, an incremental approach to determining the number of model executions required with respect to the error tolerance ε. On the other hand, the method depends on the validation analyst making a reasonable selection of the value for ε.

Finally, note that if the error tolerance $\varepsilon = 0$, these rules do not simplify to the simple interval inclusion condition for model validity given in [4]. If $\varepsilon = 0$ and μ_0 is outside the confidence interval, sub-rule (1a) will assess the model as not valid as expected, but if μ_0 is inside the confidence interval, sub-rule (2b) will assess any confidence interval with non-zero width, i.e., with $L \neq U$, as too wide. Therefore, care should be exercised in applying this method with very small values for error tolerance ε.

3. Applying the Bonferroni correction for multiple confidence intervals

It is often the case that multiple response variables are analyzed from a single model execution. For example, from an execution of a bank lobby model, both mean queue length and server utilization may be studied, or from an execution of a combat model, both Red and Blue losses may be studied. In these situations, it is potentially a mistake to apply analysis methods intended for single response variables (i.e., *univariate* methods), such as the calculation of confidence intervals, to each of the response variables individually. In doing so, the validation analyst is implicitly assuming that the response variables are independent. It is easily seen at an intuitive level that the response variables may in fact not be independent; for example, in a bank lobby model, mean queue length is likely to be positively correlated with server utilization (if the teller is busy, more customers will have to wait), and in a combat model, Red losses are likely to be negatively correlated with Blue losses (if Blue inflicts many losses on Red, Red will be less able to inflict losses on Blue).[6] The likelihood of non-independence has been recognized in the literature; e.g., "when two or more confidence intervals are computed from data generated on the same simulation run, they are rarely independent" [9] and "multiple parameter estimates from the same system are likely to be dependent" [8]. In

[6] In [10], correlation between response variables is termed *cross-correlation*, so as to distinguish it from correlation of a response variable with itself, which is *autocorrelation*. See [8] for a useful discussion of autocorrelation.

[10], the use of univariate methods to analyze multiple response variables without being aware of the limits of doing so is characterized as "naïve".

Nevertheless, the application of univariate procedures for confidence interval calculation to multiple response variables is often performed in practice; indeed, this was done in all three examples studied in the earlier confidence interval paper [4]. In [5], confidence intervals were calculated using univariate methods for waiting times for five different types of medical appointments, but the waiting times may not be independent if the same resources (e.g., physicians) are required for different types of appointments. In [6], confidence intervals were calculated using univariate methods for the number of ships processed by the port for three different ship types, but the counts may not be independent if the same resources (e.g., quays or cranes) could service different types. Finally, in [7], confidence intervals are calculated using univariate methods for vehicles lost in combat for three different British vehicle types, but the counts may not be independent if the German forces had to make target selection decisions between the different British vehicle types.

However, analysis methods intended for multiple response variables (i.e., *multivariate* methods) are available. For confidence interval calculation for multiple model response variables that can not be shown or assumed to be independent, a simple multivariate method is available. The method, which is variously referred to as Bonferroni intervals [10], the Bonferroni adjustment [11], or the Bonferroni correction [12], is mathematically justified by a mathematical result known as the Bonferroni Inequality, which characterizes the probability of multiple confidence intervals simultaneously containing their true population parameters.[7] Here we are not concerned with the theory of the inequality, but with how to apply the correction in practice. It requires only a small adjustment to the univariate method for calculating confidence intervals. In this situation, the multiple intervals are referred to as *joint* or *simultaneous* intervals.

Recall from the earlier confidence interval paper [4] that a (univariate) confidence interval for a population mean may be calculated using the Student *t* distribution as

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

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and t_c is the critical value for the Student t distribution for confidence level c. The critical value notation t_c used in [4], which follows [1] and [13], is simple but not the only notation used in the literature for the critical value.[8] To explain the calculation of Bonferroni intervals, we will switch to the alternative but equivalent notation $t_{\alpha/2,n-1}$ for the critical value, where α is the level of significance and the subscript $\alpha/2$ denotes the area under the distribution's probability density curve in one of the two "tails" when confidence level $c = (1 - \alpha)$ area is in the center. Thus, if $c = 0.95$ (95% confidence), $\alpha = 0.05$ and $\alpha/2 = 0.025$, and the critical value is chosen so that the area in each "tail" of the distribution is 0.025.[9] For example, suppose $c = 0.80$, $\alpha = 0.20$, and $n = 30$; then $t_{\alpha/2,n-1} = t_{0.01,29} = 1.311$ [1].[10]

To apply the Bonferroni correction, simply use critical value $t_{\alpha/2k,n-1}$, rather than $t_{\alpha/2,n-1}$, where k is the number of joint confidence intervals, to calculate each of the intervals.[11,12] This adjustment reduces the area in each of the tails, thus increasing the critical value and widening the confidence interval. For example, suppose $c = 0.80$, $\alpha = 0.20$, and $n = 30$ as before, but $k = 2$; then $t_{\alpha/2k,n-1} = t_{0.005,29} = 1.699$ [1], compared with $t_{\alpha/2,n-1} = t_{0.01,29} = 1.311$. The effect of the Bonferroni correction is that we may be at least $(c \cdot 100)$% confident that each of the confidence intervals simultaneously contains the population mean for its response variable.

---

[7] Recall that here the population parameter is the mean value for the model response variable over all possible execution of the model. For more mathematically detailed discussions of the Bonferroni Inequality and its implications for multivariate analysis of simulation output, see [10] or [18].

[8] Alternatives include t [22], t_c [1] [13], $t_{\alpha/2}$ [11] [23] [24], $t_{n-1,1-\alpha/2}$ [25], $t_{\alpha/2,n-1}$ [15], and $t_{(1-\gamma)/2}(n-1)$ [26]. The different notations all refer to the same value.

[9] In many conventional statistical tables of values for the t distribution, $\alpha/2$ is cross-referenced with the degrees of freedom $n - 1$ to look up the critical value.

[10] A confidence level of $c = 0.80$ for model validation is recommended by some simulation experts [21] [27] [28]. However, while it is sometimes used in practice, e.g., [14] [15], the conventional value of c = 0.95 is more common, e.g., [5] [6] [7].

[11] The Bonferroni correction as stated here is found in [8], [9], [10], [11], [12], [14], [15], [17], and [18]. Unfortunately, not all sources agree, e.g., for k joint confidence intervals, the critical value is given as $t_{\alpha/2k,n-k}$ in [29]. Using the latter form will produce wider intervals.

[12] It is not strictly required that all of the intervals be adjusted by the same amount, as shown here. For details, see [8] or [18].

i	x_i	i	x_i
1	108	11	102
2	129	12	159
3	129	13	107
4	150	14	116
5	128	15	131
6	143	16	120
7	147	17	120
8	130	18	149
9	184	19	156
10	168	20	130

Table 5. Simulated U-boat sightings in the Bay of Biscay, October 1942–March 1943 [15].

i	x_i	i	x_i
1	2	11	4
2	5	12	3
3	3	13	2
4	3	14	2
5	4	15	4
6	2	16	4
7	5	17	3
8	3	18	5
9	5	19	5
10	6	20	4

Table 6. Simulated U-boat sinkings in the Bay of Biscay, October 1942–March 1943 [15].

Variable	n	d.f.	$\bar{x}$	s	$t_{0.01,19}$	$[L, U]$	Historical	Outcome
Sightings	20	19	135.3	21.44	1.729	[127.01, 143.59]	135	Valid
Sinkings	20	19	3.7	1.22	1.729	[3.23, 4.17]	3	Not valid

Table 7. Confidence intervals for U-boat sightings and sinkings calculated using the Bonferroni correction.

Example 2. In [14] and [15], an agent-based model is used simulate Allied aircraft operations against German U-boats in the Bay of Biscay. Two historical periods were simulated (October 1942–March 1943 and April 1943–September 1943); historically the technologies and procedures used by the Allies differed during these two periods [15]. For this example we will examine only the first period. The model calculates U-boat sightings and U-boat sinkings for each month during a model execution; the total sightings and sinkings for the execution are the response variables. Clearly, the two response variables should not be assumed to be independent, because a U-boat must be sighted before it can be sunk. Table 5 shows the U-boat sightings generated by the model, and Table 6 shows the U-boat sinkings generated by the model, for of 20 model executions. The historical (simuland) values for October 1942–March 1943 were 135 sightings and 3 sinkings [16].

Confidence intervals were calculated using the Bonferroni correction for the mean values of each of the two model response variables from the results shown in the tables. As described in [15], $c = 0.80$ (80% confidence), $\alpha = 0.20$, and $k = 2$ were used, thus the critical value was $t_{\alpha/2k,n-1} = t_{0.05,19} = 1.729$ [1]. Table 7 summarizes the calculations and outcomes.[13] By way of comparison, if the Bonferroni correction had not been used, then the critical value would have been $t_{\alpha/2,n-1} = t_{0.01,19} = 1.328$ [1] and the confidence intervals would have been [128.93, 141.67] (sightings) and [3.34, 4.06] (sinkings).

The Bonferroni correction allows the calculation of joint confidence intervals for potentially non-independent response variables, is quite easy to use, and applies in "very general circumstances" [17]. Note also that for a single interval, i.e., $k = 1$, then the correction formula gives the univariate critical value, as expected. However, applying the Bonferroni correction widens the confidence intervals, thus weakening the stringency of the resulting validation test. Care should be exercised in applying this method for large numbers of intervals. As a rule of thumb, maximum value of $k = 10$ has been recommended [18].

4. Calculating confidence intervals for the difference of two means

As discussed in the earlier paper [4], confidence intervals are often used in validation when a single value or observation for the simuland is available for comparison with the results of multiple executions of the model. This is often the case when validating combat models; typically we have only a single occurrence of a historical battle for comparison with multiple simulations of the event [4] [15]. However, in some validation applications sufficient simuland values or observations are available to allow calculating means for the response variables to be used for validation for both the simuland and the model. A range of statistical hypothesis tests for comparing two

[13] The confidence interval bounds in the table are slightly different from those in [14] and [15], the sources of the example. This paper's author recalculated the confidence intervals using the sample statistics given in [14] and [15] and the recalculated values are reported here.

means are available for use in these situations; for example, the Student *t* test for comparing two means is used in [8] to validate a discrete event model of a bank drive-up window. A discussion of hypothesis tests for validation is outside the scope of this paper. However, confidence intervals, and in particular confidence intervals for the difference between two means, may also be used for validation in this situation. A confidence interval for the difference between two means can in general provide an estimate of how different the means may be, and in particular may indicate, if the interval includes 0, that the two means could be the same. The model may be interpreted as valid if the confidence interval shows that the difference between the means is 0 or acceptably small.

Our presentation of the method follows [1]. For clarity, we will explain it in terms of a single response variable. There are two populations of possible values for the response variable, the simuland and the model, and a sample has been taken from each, presumably by observations of the simuland and executions of the model. We begin by introducing essential notation:

$\bar{x}_1$	Simuland response variable mean for sample
$\bar{x}_2$	Model response variable mean for sample
s_1	Simuland response variable standard deviation
s_2	Model response variable standard deviation
n_1	Number of simuland observations, i.e., sample size
n_2	Number of model executions, i.e., sample size
μ_1	Simuland response variable mean for all simuland observations; unknown
μ_2	Model response variable mean for all model executions; unknown
$[L, U]$	Confidence interval for $\mu_1 - \mu_2$; L is lower bound, U is upper bound

The two population means μ_1 and μ_2 are unknown; however, from the two samples a $(c \cdot 100)$% confidence interval for the difference $\mu_1 - \mu_2$ between them may be calculated using the Student *t* distribution as

$$\left[(\bar{x}_1 - \bar{x}_2) - t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right],$$

where t_c is the critical value for the Student *t* distribution for confidence level *c*.

Finding the critical value requires the degrees of freedom. In the case of a single sample of size *n*, we have seen that the degrees of freedom d.f. = $n - 1$. Here there are two samples, thus d.f. = $\min(n_1 - 1, n_2 - 1)$.[14] For example, suppose $c = 0.90$, $\alpha = 0.10$, and $n_1 = 20$ and $n_2 = 14$; then $t_c = 1.771$ [1]. Using the *t* distribution to calculate a confidence interval in this way requires that either (1) both populations are normal or approximately normal ("mound-shaped and symmetric" [1]), or (2) both $n_1 \geq 30$ and $n_2 \geq 30$.

Example 3. In [19], three constructive combat models are compared: OneSAF, VR-Forces, and Alt Agg. All three are constructive combat models. OneSAF and VR-Forces are entity-level models, whereas Alt Agg is a unit-level model based on entity-level data and procedures.[15] Comparing the results of multiple models, a method referred to as comparison testing in the validation literature [2], is a validation method when at least one of the models is assumed *a priori* to be valid; even if no model is assumed to be valid, the differences between the models' results can signal validity issues [19]. For the comparison of the three models, the 1991 Battle of 73 Easting was simulated in each model and the Blue and Red losses were compared.[16] In [19], these response variables were compared using a statistical hypothesis test. In this example the results are reanalyzed using confidence intervals for the difference between two means.

Table 8 summarizes the results of the model executions.[17] In the table, $\bar{x}$ is the mean and *s* is the standard deviation for the losses, Blue or Red, for *n* executions of the model. The loss counts are in vehicles destroyed.

[14] A slightly more complicated and less conservative means of calculating d.f. in the two sample case, known as Satterthwaite's approximation, is available and is implemented in most statistical software; see [1] for more details.

[15] OneSAF is the U. S. Army's standard entity-level constructive model. VR-Forces is a commercial product of VT MAK. Alt Agg was developed by a collaboration of Gnosys Systems, Science Applications International Corporation, and the University of Alabama in Huntsville under the sponsorship of the Defense Advanced Research Projects Agency. For more information on the models, see [19].

[16] In addition to [19], see [30] for a detailed discussion of the difficulties of recreating the Battle of 73 Easting in a constructive combat model.

[17] In [19] multiple versions of OneSAF and Alt Agg were compared. For this example, we report and compare only the results of the "standard" versions of the models.

Model	Side	$\bar{x}$	s	n
OneSAF	Blue	27.04	6.20	25
	Red	47.28	8.35	25
VR-Forces	Blue	32.08	4.07	25
	Red	48.00	3.67	25
Alt Agg	Blue	23.40	7.93	25
	Red	50.68	5.32	25

Table 8. Red and Blue losses for the Battle of 73 Easting as generated by three constructive models [19].

Model 1	Model 2	Side	$\bar{x}_1 - \bar{x}_2$	t_c	E	$[L, U]$	$L \le 0 \le U$?
OneSAF	VR-Forces	Blue	-5.04	2.064	1.483	[-8.10, -1.98]	No
		Red	-0.72	2.064	1.824	[-4.49, 3.05]	Yes
VR-Forces	Alt Agg	Blue	8.68	2.064	1.783	[5.00, 12.36]	No
		Red	-2.68	2.064	1.293	[-5.35, -0.01]	No
Alt Agg	OneSAF	Blue	-3.64	2.064	2.013	[-7.80, 0.52]	Yes
		Red	3.40	2.064	1.980	[-0.69, 7.49]	Yes

Table 9. Confidence intervals for the differences between mean Blue and Red losses.

Table 9 summarizes the calculation of the confidence intervals for the differences between the means. The table shows six confidence intervals calculated for the difference of means: there are three possible pairs of models, and for each pair of models an interval is calculated for difference of mean Blue and mean Red losses. In the table, $\bar{x}_1 - \bar{x}_2$ is the difference between the sample means, t_c is the critical value for $c = 0.95$ and d.f. = 24, E is the error term in the confidence interval formula (see below), and $[L, U]$ is the calculated confidence interval.

$$E = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note that three of the six confidence intervals contain 0. Because in this example models are being compared, an interval containing 0 is interpreted as supporting the conclusion that the models are statistically equivalent for the response variable in question. If one of the models was assumed *a priori* to be valid, or if the comparison was between a model and the simuland rather than two models, the conventional validation interpretation is that if the interval for the difference of means contains 0, the model is valid for that response variable. In either case, the (model compared to model or model compared to simuland), the mathematics of the comparison process are the same.

5. Additional topics

Even together, the introductory paper [4] and this paper do not exhaust the topics related to the use of confidence intervals in model validation. Additional relevant topics include:

1. Analyzing multiple potentially non-independent output variables from a single model execution [10]
2. Scheffé confidence intervals as an alternative to the Bonferroni correction [10]
3. Steady-state analysis of non-terminating discrete event models using confidence intervals [10]
4. Discussion of the limitations applying confidence intervals to historical events [15]
5. Validation methods using joint or simultaneous confidence intervals or regions [2] [20] [21]

6. References

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

[2] 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.

[3] 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.

[4] M. D. Petty, "Calculating and Using Confidence Intervals for Model Validation", *Proceedings of the Fall 2012 Simulation Interoperability Workshop*, Orlando FL, September 10-14 2012, pp. 37-45.

[5] 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, No. 8-9, August-September 2010, pp. 510-541.

[6] E. Demirci, "Simulation Modelling and Analysis of a Port Investment", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 79, No. 2, February 2003, pp. 94-105.

[7] 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.

[8] 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.

[9] G. S. Fishman, *Discrete-Event Simulation: Modeling, Programming, and Analysis*, Springer-Verlag, New York NY, 2001.

[10] J. M. Charnes, "Analyzing Multivariate Output", *Proceedings of the 1994 Winter Simulation Conference*, Arlington VA, December 3-6 1995, pp. 201-208.

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

[12] Wikipedia contributors, "Bonferonni correction", Wikipedia, The Free Encyclopedia, Online at http://en.wikipedia.org/w/index.php?title=Bonferroni_correction&oldid=529961001, Accessed January 5 2013.

[13] S. K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction, Second Edition*, Radius Press, New York NY, 1991.

[14] 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.

[15] 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

[16] B. McCue, *U-boats in the Bay of Biscay: An Essay in Operations Analysis*, National Defense University Press, Washington DC, 1990.

[17] C. Alexopoulos and A. F. Seila, "Output Data Analysis", in J. Banks (Ed.), *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, John Wiley & Sons, New York NY, 1998, pp. 225-272.

[18] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis, Third Edition*, McGraw-Hill, New York NY, 2000.

[19] M. D. Petty, J. Panagos, J. P. Joseph, and R. W. Franceschini, "Validation Using Comparison Testing of Three Constructive Combat Models", *Proceedings of the Fall 2011 Simulation Interoperability Workshop*, Orlando FL, September 19-23 2011, pp. 201-212.

[20] O. Balci and R. G. Sargent, "Validation of multivariate response models using Hotelling's two-sample T^2 test", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 39, No. 6, December 1982, pp. 185-192.

[21] 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.

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

[23] G. K. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*, John Wiley & Sons, New York NY, 1977.

[24] 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.

[25] 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.

[26] 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.
[27] 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.
[28] J. P. C. Kleijnen, "Statistical Validation of Simulation Models", *European Journal of Operational Research*, Vol. 87, Iss. 1, 1995, pp. 21-34.
[29] X. Yan and X. Su, *Linear Regression Analysis: Theory and Computing*, World Scientific, Singapore, 2009.
[30] W. E. Daniels and M. D. Petty, "Recreating the Battle of 73 Easting in a Constructive Combat Model", *Proceedings of the 2012 AlaSim International Modeling and Simulation Conference*, Huntsville AL, May 1-3 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 175 research papers and has been awarded over 100 research projects totaling over $15 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 journal *SIMULATION*. He was the dissertation advisor to the first two students to receive Ph.D.s in Modeling and Simulation at Old Dominion University and is currently coordinator of UAHuntsville's Modeling and Simulation degree program.