# How Many Times Should a Stochastic Model Be Run? An Approach Based on Confidence Intervals

**Michael D. Byrne (byrne@rice.edu)**
Departments of Psychology and Computer Science, 6100 Main Street
Houston, TX 77005 USA

## Abstract

A persistent problem in computational cognitive modeling is that many models are stochastic. If a model is stochastic, what is the prediction made by the model? In general, this problem is solved via Monte Carlo simulation. This raises the question of how many runs of the model are adequate to produce a meaningful prediction, a question that has received surprisingly little attention from the community. This paper proposes a systematic approach to the selection of the number of model runs based on confidence intervals and provides tables and computational examples.

**Keywords:** stochastic models; Monte Carlo simulation; number of model runs.

## Introduction

One of the primary advantages of computational modeling over purely informal theorizing is that computational models are often capable of producing quantitative predictions about human performance. That is, beyond saying "condition A will generate more errors than condition B" models can often produce numerical predictions for the error rates in A and B, which has considerable advantages, both theoretically and practically.

However, it turns out that "the prediction" of a model is not always exactly known. If the model in question is deterministic or can be represented by a closed-form equation, then this is not an issue. However, if the model is stochastic, which many computational cognitive models are, then this is a potential problem—one that has received almost no attention in the literature. This is most likely because a solution is apparent: run the model multiple times and take the average across those runs to be the model's prediction. This is termed Monte Carlo (MC) simulation.

However, this raises another question that has generally not been systematically addressed: how many times should the model be run? Most people have the intuitive sense that more runs are better. We trust an average across 100 model runs more than an average of 5 model runs. But are there cases when 5 is enough? Are there cases when 100 is not enough? How many model runs are "enough" for the point(s) reported as "the model's predictions(s)" to be an accurate representation of what the model really predicts?

Running the model repeatedly is sampling from an infinitely large population: all possible runs of the model. In this sense, the prediction of a model is a single point—the mean of all possible runs. However, since "all possible runs" is an infinite population, we must sample, and therefore we must tolerate some uncertainty about the true prediction made by the model. This is the domain of inferential statistics, one that should be familiar to anyone who has taken a statistics course, yet there is a peculiar lack of discussion of this question in modeling papers and almost no evidence of statistical methods being applied to this problem. "How large a sample should be taken?" is not a new question in statistics, but it has received little attention from cognitive modelers.

Understanding this issue is important for both model builders and people who evaluate stochastic models. For model builders, it is critical to understand what the predictions of the models really are. Consider a model of an experiment where subjects make errors. There are two conditions in the experiment. In condition A, the subjects made errors on 3.6% of the trials. In condition B, the subjects made errors on 7.0% of the trials. Inferential statistics performed on the human data indicate that this difference is statistically reliable. How many runs of the model are necessary to be sure that the model legitimately makes different predictions in the two conditions?

Similarly, a surprisingly high percentage of papers that present such models contain no information whatsoever about how many times the model was run in order to generate the point predictions, or use questionable methods for determining the number of model runs. How credible are the model's "predictions" under such conditions? If a model of the just-described experiment were presented where the model was run 50 times in each condition, is the modelers' claim that the model captures the difference between the conditions credible?

## Previous Approaches

One of the only systematic discussions of this issue appears in Ritter, Schoelles, Quigley, and Klein (2011). One of the most salient parts of that paper is their Table 1, where they present a list of papers in the 2004 proceedings of the International Conference of Cognitive Modeling and note both how many human subjects were run and how many times the model was run to generate the reported results. Nearly half (14 of 33 papers) report no information at all about how many model runs were used. It is thus difficult to know what approach was used.

A common approach (and one I admit I have frequently relied on in the past) is to simply choose an arbitrarily large number (e.g., 100) and assume that number is "large enough" that the issue is addressed. This is at least better than reporting nothing, and running the model many times is almost certainly better than running it few times, but this approach is not well-motivated mathematically and there may be cases where 100, while large, still is not "large enough."

Another approach that appears common (at least anecdotally; this does not appear often in the cited data) is to run the model the same number of times as there were human subjects in the experiment being simulated. There is a certain intuitive appeal to this approach—if *n* was enough in the human data, then *n* should be enough for the model. This logic is fundamentally flawed, though. If collecting human data could be done in almost no time and at almost no cost, every experiment would have a great many subjects and inferential statistics would be unnecessary. However, there are often substantial costs to collecting human data, and the same cost function often does not apply to running a model. Collecting 1,000 subjects worth of data is practically impossible for many human experiments but is eminently tractable for many simulation models run on modern computing hardware. If modelers want to claim that their models are truly making point predictions, the number of model runs should be determined by the mathematics of sampling and not the practical costs associated with running subjects.

However, it should be noted that for some models equalizing the number of subjects and the number of model runs may indeed be an appropriate approach. For example, Daily, Lovett, and Reder (2001) ran the same number of ACT-R models as human subjects because each ACT-R model run was parameterized to match a specific subject in the data set, and model predictions were compared with individual subject data rather than group means. This is certainly not the norm, however, as most model data takes averages of model runs and compares those to mean human data.

Furthermore, some models do have higher costs associated with each run. For example, the ACT-R model described in Zemla, et al. (2011) was coupled somewhat unreliably to a real-time flight simulator and each run of the model could take more than 10 minutes, and some runs failed because the coupling was lost. Thus, even 40 runs of the model could take many hours to perform.

In the more general case, however, most cognitive models compare the mean prediction of the model with a mean based on many subjects, and runs of the model are relatively cheap. What then?

One approach is to run the model repeatedly until the mean converges. That is, until additional model runs do not change the mean model prediction by more than some small threshold value. For an example of this approach in practice, see Teo, John, and Blackmon (2012). This approach is often only practical if model runs are not just cheap, but extremely cheap. For example, the cited example used more than 20,000 model runs.

Another possible, though not recommended, strategy is to use an inferential statistical test (such as a *t*-test) in an attempt to show "no significant difference" between the model and the human data. The logic behind this approach is faulty for two reasons. First, failure to reject with an inferential test does *not* provide evidence that two samples are equal; this is a gross misunderstanding of statistical tests.[1] Second, such tests are most likely to fail to find a difference when small sample sizes are used, so this method rewards higher levels of uncertainty about the model's true prediction.

Ritter, et al. (2011) present an approach based on power analysis to provide general guidance on the number of model runs to use. This is an interesting strategy, but it has some limitations. In particular, it assumes that the model and the data each have only two points with a well-defined statistical effect size between them, and does not apply to situations where proportions are used. The current method takes a different, though somewhat related, approach that is intended to be more general.

## Confidence Intervals

A confidence interval is a statistical construct that provides information about the location of a population parameter based on a sample statistic. Modelers want to know the value of a population parameter (the mean of all possible runs of the model), but cannot know it because the population is infinite. Thus a sample statistic (the mean of the MC runs) is used to estimate the population parameter.

A confidence interval requires the specification of a confidence level, which is generally a large percentage; 95% is the modal choice in many domains. A confidence interval is a range of values such that the confidence level proportion of those intervals will contain the true population parameter. That is, 95% of the 95% confidence intervals for the mean will contain the true sample mean.

In general, confidence intervals are computed by taking the sample mean and constructing the interval around that mean, given a desired confidence level and the sample size. The method proposed here essentially does this in reverse: it computes the appropriate number of MC runs (denoted *n*) that should be conducted in order to achieve a confidence interval of a particular width given a desired confidence level. The confidence level will be assumed to be 95% but computations could be carried out for other confidence levels if desired. For example, if a modeler wanted to be 95% certain that the true prediction of their model was within 1% of the mean of the sample, this method can then determine the appropriate *n*. This will be made more concrete with examples in the following sections.

Note that constructing confidence intervals requires that certain assumptions be met. The two most critical assumptions are:

- *Random sampling*. It is critical that the MC methods used do not introduce systematic bias such that some random numbers are more likely than others. This is unlikely unless poor random number generation schemes are used.
- *Independent and identically distributed*. Each run of the model must be statistically independent from the other runs, that is, what happens in one run of the model cannot affect the behavior of other runs, and all runs must come from the same distribution. The independence assumption can be

[1] Note that there are more recent Bayesian methods that do allow testing for equality. However, these methods are not in widespread use and are more mathematically complex. Full development of methods based on such techniques is beyond the scope of the current paper.

potentially problematic for some models, particularly those models that learn over time. The identical distribution assumption may be violated by some models as well. For instance, if sometimes the model follows one strategy and sometimes it follows others, and those strategies affect the outcomes, then not every sample comes from an identical distribution.

## Models of Proportions

Some models produce proportions as their output. The most common form of this is probably the proportion of trials correct. The standard equation for a confidence interval for a proportion is this:

$$CI = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad [1]$$

where $CI$ is the confidence interval, $p$ is the observed proportion, $n$ is the sample size, and $z_{\alpha/2}$ is the value of the normal distribution that represents the upper tail of the distribution set by the confidence level. (For a 95% interval, this value is 1.96.) The width of the confidence interval, denoted $w$, is thus:

$$w = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad [2]$$

This equation can then be solved for $n$ to produce an equation to specify the minimum number of model runs necessary to produce a confidence interval with the desired width:

$$n = p(1-p)\left(\frac{z_{\alpha/2}}{w}\right)^2 \qquad [3]$$

Note that an integer number of model runs is required, so the ceiling of this value should be taken.

In the case of proportions, the minimum sample size is a function of the proportion. In general, the model should be producing the same proportion as the human data, so $p$ should usually be known; generally the modeler should only have to specify the desired width $w$. Table 1 presents the minimum $n$ for a range of values of $w$ and $p$ assuming 95% confidence. $w$ appears on the vertical and $p$ along the horizontal. For values of $p$ greater than 0.5, the table is simply mirrored; a $p$ of 0.7 is equivalent to a $p$ of 0.3.

Note that Equation 3 has some potential issues with statistical assumptions. This is an approximation, and this particular approximation starts to break down as $p$ approaches 0 or 1, particularly as $n$ gets smaller. While the appropriate correction for this is the subject of some discussion in statistical circles, for the purpose of estimating required sample size, the simple Yates correction (Blyth & Still, 1983) is almost certainly adequate, and has been applied whenever the uncorrected $np$ < 10. (See the Appendix for details, including complete equations.)

An example will help illustrate. In the scenario described in the introduction, there were two conditions, A and B, that had error proportions of 3.6% and 7%. The modeler needs two values for $n$, one for each group. For group A, $p$ is 0.036. For group B, $p$ is 0.07. The difference between the two values is 3.4%, so to be sure the model predicts a real difference between conditions, $w$ should be half that value, or 0.017. Assuming 95% confidence, meaning a $z$ of 1.96, then then by Equation 3 the number of model runs needed in condition A is 462 and in condition B it is 866.

These are large numbers of model runs. It is highly unlikely that the experiment being modeled involved over 1300 subjects, so matching the number of model runs to the number of subjects is not likely to produce large enough numbers. In fact, even a moderately large number of subjects (e.g., 100 or 200) would produce a number of model runs that would be grossly inadequate if the modeler simply matched to the number of subjects. 1,000 model runs in each condition would have worked, but a brief glance at Table 1 shows that there may be circumstances when even 1,000 runs will not be sufficient to guarantee confidence in the true value of the model's prediction, particularly as $p$ approaches 0.5 and the desired interval width gets smaller. Discriminating 50% from 45%, for example, would require several thousand model runs for each group.

Note that the width $w$ is not necessarily a function of the difference between two groups; this value is at the discretion of the researcher, and of course larger values of $w$ require fewer model runs. Researchers should be aware, however, that small numbers of model runs can produce substantial uncertainty about what exactly the model's predictions really are. Reviewers of modeling research should be aware of this as well.

Table 1. Minimum number of model runs ($n$) to achieve desired confidence interval width ($w$) for model fitting to proportion ($p$), assuming 95% confidence level. Shading indicates continuity correction was applied (see the Appendix for full details).

| | | Proportion ($p$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** | **0.3** | **0.35** | **0.4** | **0.45** | **0.5** |
| Width ($w$) | **0.005** | 7300 | 13830 | 19593 | 24587 | 28812 | 32270 | 34959 | 36880 | 38032 | 38416 |
| | **0.01** | 1825 | 3458 | 4899 | 6147 | 7203 | 8068 | 8740 | 9220 | 9508 | 9604 |
| | **0.02** | 457 | 865 | 1225 | 1537 | 1801 | 2017 | 2185 | 2305 | 2377 | 2401 |
| | **0.05** | 109 | 139 | 196 | 246 | 289 | 323 | 350 | 369 | 381 | 385 |
| | **0.1** | 36 | 53 | 68 | 62 | 73 | 81 | 88 | 93 | 96 | 97 |
| | **0.15** | 19 | 27 | 34 | 40 | 45 | 36 | 39 | 41 | 43 | 43 |

## Models of Ratio or Interval Data

Proportion data are actually simpler to work with, because the variance in a sample of proportion data is known exactly when *p* and *n* are known, and there are no issues with units. This is not the case for interval or ratio data where the variance in a sample is not necessarily dependent on the mean. Furthermore, the meaning of a given interval width is dependent on the units. ±1% has a clear interpretation in the case of a proportion, but the value of saying ±100 units depends heavily on the units, and possibly also the scale of data. For example, saying the prediction of a model is known with ±100 ms when the task being modeled takes 500 ms is not particularly precise, but ±100 ms for a task that takes 6 minutes is probably much greater precision than the actual data. So, unlike with proportion data, magnitudes and units matter.

This makes determining the appropriate number of model runs more challenging. However, there are certain simplifications that can be made to make the problem more tractable. These simplifications can be applied to any type of interval or ratio data and the equations still apply, but the underlying assumptions are likely more valid for certain types of measurements than for others. Because response time is still one of the most common dependent measures in experimental psychology, there are many models being fit to response time data, and so these simplifications are intended to be most appropriate when applied to response time data.

The first simplification is that a unitless measure of *w*, the target confidence interval width, is desirable. This allows intervals to be compared across measurements and simplifies the equations. In order to do this, *w* is defined here as "proportion of the mean." That, a *w* of 0.1 represents 10% of the mean, whatever that mean is. So, if the mean of the human data is 1200 ms and the desired precision is ±100 ms, then *w* would be 0.083. This is not unusual in engineering applications, where the goal is often to be accurate within a certain percentage of the mean.

The second simplification involves variance. The standard equation for a confidence interval for interval or ratio data is the following:

$$CI = M \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad [4]$$

where is *M* is the sample mean, *z* is the usual value of standard normal (1.96 for 95% confidence), σ is the standard deviation, and *n* is the sample size. Thus, the width is a function of the standard deviation (square root of the variance) of the sample. However, in modeling human performance, it is often the case that the standard deviation scales with the mean. That is, models of tasks that take 5 minutes have larger variances than models of tasks that take 800 ms; this is generally true of human data but also many models. This allows the use of the *coefficient of variation* (*CV*) as a measure of variability. The CV is defined as:

$$CV = \frac{\sigma}{\mu} \qquad [5]$$

So, if the standard deviation and the mean scale proportionately, the CV will be constant. This is not always true in either human or model data, but it provides a starting point. If *w* is taken as a multiplier on the mean, then the width of a confidence interval can be computed by solving the following equation:

$$w\mu = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad [6]$$

for *n*. The intermediate form of the solution looks like this:

$$\sqrt{n} = z_{\alpha/2} \frac{\sigma}{w\mu} \qquad [7]$$

Notice, however, that the standard deviation divided by the mean is the coefficient of variation. Thus, the final form of the equation for determining the minimum number of model runs is the following:

$$n = \left( \frac{z_{\alpha/2}}{w} CV \right)^2 \qquad [8]$$

Again, non-integer values of *n* should be rounded up to the nearest integer.

Note that these equations (starting with Equation 4) assume that the population standard deviation (that is, σ) is known. In practice this assumption is fine as long as the sample size is moderately large, over about 100. For smaller values of *n*, the *t*-distribution should be used to determine the critical value rather than the normal (*z*) distribution.

Table 2 shows the minimum number of model runs required to produce the desired interval with *w* for a given

Table 2. Minimum number of model runs (*n*) to achieve desired confidence interval width (*w*) for model with coefficient of variation (*CV*), assuming 95% confidence level. Shading indicates correction for small *n* (see the Appendix for full details).

| | | Coefficient of variation (*CV*) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** | **0.3** | **0.35** | **0.4** | **0.45** | **0.5** |
| Width (*w*) | **0.005** | 385 | 1537 | 3458 | 6147 | 9604 | 13830 | 18824 | 24587 | 31117 | 38416 |
| | **0.01** | 99 | 385 | 865 | 1537 | 2401 | 3458 | 4706 | 6147 | 7780 | 9604 |
| | **0.02** | 27 | 99 | 217 | 385 | 601 | 865 | 1177 | 1537 | 1945 | 2401 |
| | **0.05** | 7 | 18 | 37 | 64 | 99 | 139 | 189 | 246 | 312 | 385 |
| | **0.1** | 5 | 7 | 11 | 18 | 27 | 37 | 50 | 64 | 81 | 99 |
| | **0.15** | 5 | 5 | 7 | 10 | 13 | 18 | 24 | 30 | 37 | 46 |

coefficient of variation (*CV*), assuming 95% confidence. Table values with shading have been adjusted for smaller *n* using the *t*-distribution to compute the required critical value. Note that the rightmost column in Table 2 is nearly the same as in Table 1. This is not accidental, as both are ultimately based on the normal distribution.

Unsurprisingly, the more variable the model is, the more model runs will be required. Similarly, the narrower the desired width is, the more model runs will be required. For example, a model with a standard deviation of 180 ms being fit to a mean of 900 ms produces a coefficient of variation of 0.2, so to produce an interval where the sample is 95% certain to be within 5% of the mean (that is, *w* is 0.05) will require only 64 runs of the model. Of course, 5% of the mean is 45 ms, so if there are many conditions with small differences, a narrower width may be required and thus more model runs may be necessary to differentiate the model's predictions across conditions.

Unfortunately, this approach leads to something of a chicken-and-egg problem: how can one know the CV of a model prior to running it? Unfortunately, for many models there is no *a priori* way to know what the CV for the model will be. Sometimes this can be estimated based on similar models, or if a model must be run in multiple conditions the CV used in one condition is often a useful guide for what it will be in other, similar conditions. However, this still does not address a brand new model in a new domain. So where to begin?

A conservative way to start is to use the CV of the human data. Many stochastic modeling systems produce data that are less noisy than the human data upon which they are based, in which case the CV from the human data is a worst-case scenario. This may be somewhat more expensive in terms of number of model runs, but it should guarantee that the target interval width is achieved.

Unfortunately, the CV from the human data is not always available. Models of experiments where the complete original human data are not available are not uncommon, and variability is not always reported by experimenters. What then? Unfortunately, there is no obvious answer. A practical suggestion is to begin by running the model 20 times, estimate the CV from those 20 runs, and then consult Equation 8 (or the table) to see if additional model runs are necessary. This will sometimes produce too many model runs but at least provides a starting point. (Note that, strictly speaking, this violates the assumptions of computing the confidence interval if more runs are required. However, this is not likely to be a problem as long as this procedure is not repeated often for the same interval.)

Note that *w* could be specified as an absolute tolerance in the original units—e.g., milliseconds—and essentially the same equations could be applied, but then scale invariance would be sacrificed. This may be desirable for certain modeling contexts and would certainly be a reasonable alternative.

## Discussion

Stochastic models, by definition, imply uncertainty. This uncertainty is manageable, but only if modelers collect samples of adequate size. Samples that are too small make it difficult to be confident that the reported prediction of the model is close to the true prediction of the model. Samples that are too large have no negative consequences other than simulation expense, though that can be a legitimate issue for some models. Therefore, it behooves the field to take a principled approach to determining the number of runs to use when performing MC simulations.

Most critically, common and simple heuristics like "perform the same number of runs as the number of human subjects" and "pick an arbitrarily large number" fail to respect the mathematics of the situation. Fundamentally, the field needs to honor known relationships between sample size and uncertainty in estimation. One way of accomplishing this is through the use of confidence intervals. This is not a panacea, but forms a principled basis for computing sample size and will hopefully serve as a starting point for future modeling work.

Current reporting practices make this difficult. As noted by Ritter, et al. (2011), many papers do not even report the number of model runs used. This should be a basic requirement for all stochastic models. Furthermore, papers should report not only how many runs were used, but the basis for selecting that number. Matching to the number of subjects and informal "guesstimation" are not principled ways of choosing this number, and requiring modelers to report their rationale should encourage more principled approaches.

Given the size of the numbers in most of Table 1, it seems highly likely that many extant models of proportional data are using insufficient sample sizes. Shortcomings like these need to be addressed. Fortunately, Table 1 can be used straightforwardly as a guide to help those evaluating such models and should be able to help identify cases where insufficient model runs have been performed.

However, in the case of interval and ratio data, knowledge of the sample size and the justification for choosing it is still not sufficient information, because the number of model runs required is also a function of the variability in those runs. Here that variability is expressed as the coefficient of variation, but reporting of the variance or standard deviation of model runs would allow computation of the CV, and therefore computation of the appropriate sample size by reviewers or other researchers. It does not seem unreasonable to require reporting of this information as well. Such a requirement would encourage modelers to consider whether or not they had a sufficient number of model runs, and would allow more informed evaluation of models.

Note that the methods described here are not intended to be completely comprehensive; there are certainly situations modelers will encounter that are not covered by these equations. For instance, if the subjects being modeled produce multinomial proportions, (that is, split responses between three categories rather than two), then neither of the approaches here will apply. There are many kinds of data modeled and it would be nearly impossible to cover ever possibility. However, there are ways to compute confidence intervals for many other kinds of data, and thus the general approach outlined here could serve as a guide.

The ultimate goal of principled methods of selecting the number of model runs is increasing the accuracy and transparency of the modeling process. Note that the proposed procedure does not remove all judgment from the decision. The appropriate value for *w* is still at the discretion of the modeler, and reasonable disagreement about what the appropriate value of *w* should be is possible. Simply initiating such discussion would be an important step forward.

Finally, resources for computing sample sizes, including code in both R and Python, as well as web-based calculators, can be found at:

http://chil.rice.edu/research/nomr/

## Acknowledgments

I would like to thank Frank Ritter for getting me started on this topic, and Bonnie John for encouraging me to work out the math and lay it out in a paper.

This work was supported in part by National Institute of Standards and Technology (NIST) grant #70NANB12H215. The views and conclusions expressed are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of NIST, the U.S. government, or any other organization.

## References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119–126.

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association, 78*, 108–116.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science, 25*, 315–353.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of simulation runs:Treating simulations as theories by not sampling their behavior. In S. Narayanan & L. Rothrock (Eds.), *Human-in-the-loop simulations: Methods and Practice* (pp. 97–116). London: Spring-Verlag.

Teo, L., John, B. E., & Blackmon, M. H. (2012). CogTool-Explorer: A model of goal-directed user exploration that considers information layout. In *Human Factors in Computing Systems: Proceedings of CHI 2012* (pp. 2479–2488). New York: ACM.

Zemla, J. C., Ustun, V., Byrne, M. D., Kirlik, A., Riddle, K., & Alexander, A. L. (2011). An ACT-R model of commercial jetliner taxiing. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting*, (pp. 831–835). Santa Monica, CA: Human Factors and Ergonomics Society.

## Appendix

As noted in the main text, Equations 1 and 4 are large-sample approximations that require adjustment under certain conditions. Equation 1 is particularly difficult because the "true" confidence interval for a proportion is not, in fact, symmetric about the sample mean. A great deal of debate has gone on in the statistical literature on how to best estimate the confidence interval for a proportion (see, e.g., Agresti & Coull, 1998). However, while there are many more sophisticated methods for estimating the confidence interval for a sample, these methods have not been applied in reverse, that is, in order to compute the sample size necessary to produce a desired interval. In general, the recommendation is that the standard (or "Wald" form) of Equation 1 is adequate for sample size estimation. However, this equation is know to be a particularly bad approximation when the product *np* is less than 10. The simplest correction to solve for *n* is the Yates correction (Blyth & Still, 1983), where *w* is computed this way:

$$w = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} + \frac{.5}{n} \qquad \text{[A.1]}$$

This can be solved for *n* as follows:

$$n = \frac{\frac{z_{\alpha/2}}{w} + p(1-p) + \sqrt{(\frac{z_{\alpha/2}}{w} + p(1-p))^2 - \left(\frac{z_{\alpha/2}}{w}\right)^2}}{2\frac{z_{\alpha/2}}{w}} \qquad \text{[A.2]}$$

As usual, this should be rounded up to the nearest integer. Values in Table 1 were computed first using Equation 3, but then if that produced *np* < 10, equation A.2 was used in its place.

For Equation 4, the approximation starts to break down because the population standard deviation is unknown and must be estimated from the sample. For large samples, this is not a problem, but when *n* drops below 100, a correction should be applied to compensate. The correction for this is simple and uncontroversial: the critical value for *t* is substituted for the critical value for *z*. This yields the following equation for *n* as a function of *w* and the *CV*:

$$n = \left(\frac{t_{\alpha/2}(n-1)}{w}CV\right)^2 \qquad \text{[A.3]}$$

The alert reader will note that the degrees of freedom for *t* depends itself on *n*, meaning the equation is indeterminate. The equation for the *t* distribution is solvable, but this is unnecessarily complicated; the equation can be computed starting with a large *n* and then *n* adjusted downward until the equation converges. This is the method used to generate the entries in Table 2 whenever Equation 8 generated an *n* less than 100. In practice, the adjustment is not large; this simply added 2 or 3 to the value of *n* required.

Note that the *t*-based correction assumes the population being sampled from is normally distributed. The *t*-based approximation should be fine for any population that is roughly continuous, unimodal, and symmetric. Since many models use Gaussian or logistic (which is nearly normal) noise, this will usually not be a problem. For more non-normal populations, other corrections would have to be adopted on a case-by-case basis depending on the shape of the population distribution.