doi: 10.1255/tosf.48

Building confidence intervals around the obtained value of a sample

Dominique M. Francois-Bongarcon, PhD
Agoratek International Consultants Inc., North Vancouver, Canada. E-mail: dfbgn2@gmail.com

Common practice in sampling for the TOS erudite consists of using the sampling variance obtained from Gy's numerical theory to build confidence intervals around the true sample value. This is usually done to characterise the 'precision' of the sample, and, by centring that interval on the sampled value, one states for instance that "the true value has 95% chances of being between values x and y", those two values usually being centred on the sampled value". The somewhat naïve rationale behind this practice is reviewed in some details and criticised. It is suggested the confidence interval of real interest to the user of the sampled value, is more difficult to define and more delicate and indirect to build. Some methods for doing so are examined and a methodology is recommended.

Introduction

Gy's Theory of Sampling[1] (TOS) has a powerful numerical section that gives us a wealth of information about the behaviour of a sample, provided we know enough about some physical characteristics of the matter being sampled and basic parameters about the sample such as its mass. Armed with it, we can in particular predict the variance we are likely to encounter should the sample be taken many times, i.e. characteristics about the distribution of the possible sample values. That predicted variance, which measures the dispersion of that sample distribution, allows for a characterisation what is often termed the 'precision' of the sample, or in other words, its goodness.

It is not uncommon then to use that variance to build some kind of a confidence interval around the obtained sample value to state where the true value of the variable to be measured may lie. Indeed, what is the use of the sampled value if we have no notion of what it really means regarding the unknown, true value we are trying to best guess? Building this confidence interval also clearly requires, implicitly or not, not only the variance, but also an idea of the distribution type or shape.

This practice, however, can often be applied quite naively, as we are going to see, starting with the fundamental question: "What distribution exactly are we speaking about?"

Better definition of the problem

So, here we are, with a sample value in hand, and the ability to predict the dispersion variance attached to it. Now, experience and knowledge also give us an idea of the shape of the sample distribution:

- Normal-like if the variance is relatively small (and the sample 'precision relatively good); this is a consequence of the symmetrisation of such distributions when their variances diminish, itself deeply and implicitly rooted in the general mechanism underlying the famed Central Limit Theorem.
- Lognormal-like or binomial-like in the opposite case.

We can therefore predict a 'histogram of sorts' of the possible sample values. And this, in practice, may not be hugely rigorous, but in reality, experience shows it works well enough: when that histogram is built experimentally as the result of repeated sampling, this method is usually reasonably validated. But there lies an often unseen difficulty: we then need to define very clearly the nature and full range of what it is, exactly, we are trying to guess.

When a sample is taken, hopefully in a representative fashion, we are obviously hoping to be able to use its value in lieu of the unknown, true value of the variable measured/estimated by that sampling operation, and we would like to know how imperfect doing so can eventually be. That is where a confidence interval may come into play: a very explanatory view to it consists of trying to attach probabilities to the unknown value underlying the sampling, saying for instance that there are 95% chances that it is in a specific, known interval around the obtained sample value.

To quickly understand/illustrate why using the sample distribution shape to do so is a rather naive idea, and for the sake of the exercise, let us assume a true value T and that the sample distribution around it is skewed towards high values ('to the right') like in the 3-bar histogram of Figure 1 where the true value is the centre value of the 3 possible sample outcomes. When we take a sample, we do not know the true value, and in this simplistic case, all we know is that the sample is one of the possible outcomes, in this case one of three, but we do not know which one.

Going in turn to every possible sample value in the distribution, and looking where the true value lies in each case with respect to that sample value, i.e. on which side of the sample value and how often this will happen, the histogram of the possible true values that could generate this sample can be drawn (Figure 2). Clearly, this is not the distribution of the sample values, it is, at best its mirror image. Skewed distributions calling for asymmetric confidence intervals, it becomes clear using the sample distribution directly would be very wrong in this case.

It does not mean, however, the solution lies in symmetry. This example was simple but also itself quite naive. The models of TOS tell us that the variance of the sample distribution is heteroscedastic, meaning it changes with the true value being sampled, i.e. it is concentration-dependent. In this example, we had ignored this important fact.

It is nevertheless possible to reach the following conclusions:

- The distribution of sample values around a given true value is not the same (in dispersion and shape) as the underlying distribution of potential true values around a known sample.
- The first one is usually simple and fairly well known (to a good enough degree in practical terms), the latter, conversely, is not readily known, and its determination would be very complex.
- For confidence intervals characterising the unknown true value, unfortunately it is that second, problematic one that really counts.

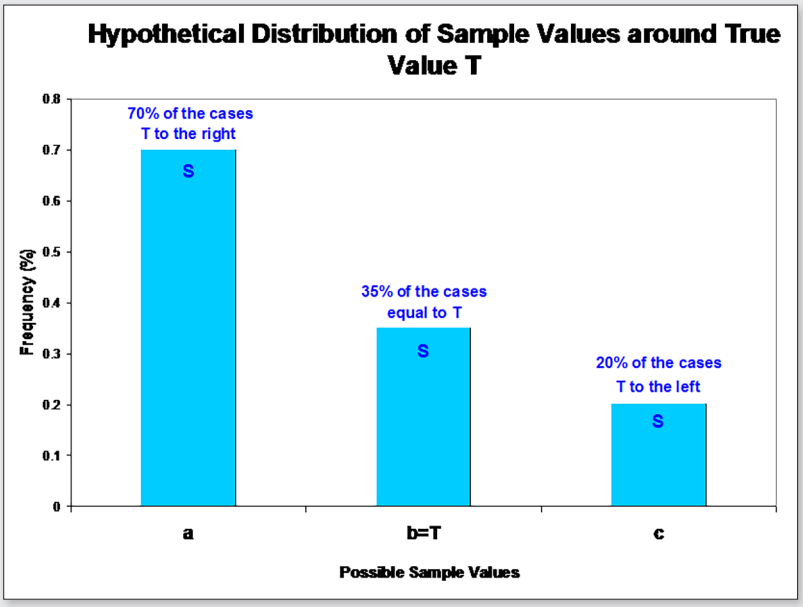

Figure 1. Hypothetical distribution of sample values around true value T

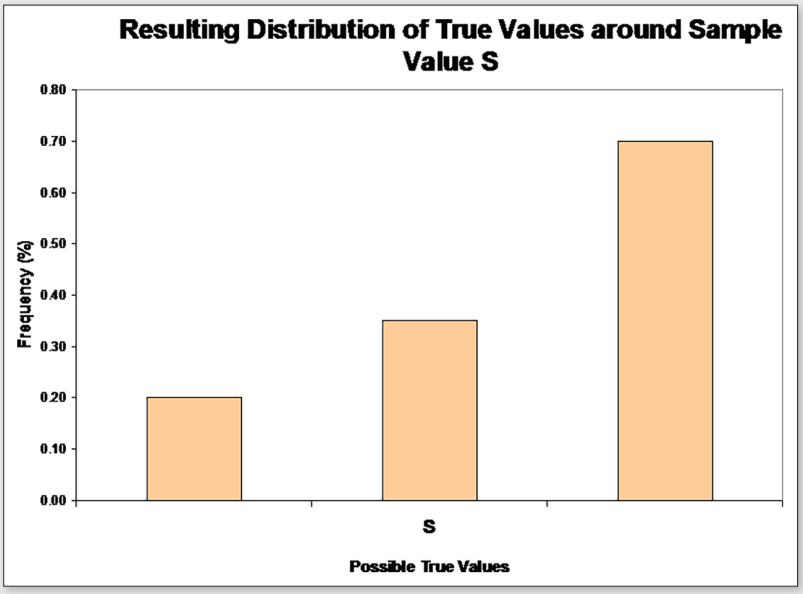

Figure 2. Resulting distribution of true values around sample value S

The lookup method

The ignorance in which we are of the underlying distribution of the possible true values that are able to give raise to a known sample value (obtained experimentally), indeed makes the problem of building a proper confidence interval around the sample value, a rather complex one. A method sometimes used is the lookup method, based on the likelihood concept. In this method, each possible true value (concentration) is considered in turn, and a confidence interval (e.g. 95% confidence as an example) is built around it for the sample values, based on what is known of the sample distribution, including its concentration-dependent variance. By definition each interval contains the 95% most likely sample values for a given true value. These intervals are plotted on a diagram. The upper and lower limits of these intervals define a region in the [True Value, Sample Value] space, containing all the sample values belonging to their respective 95% confidence intervals around their true values (in the example of Figure 3, the sample distributions were assumed to be binomial). We will call it the '95% Domain'.

Then, when considering a specific sample value, it defines a horizontal line on the diagram. The intersection of the line with the 95% Domain is then used as a confidence interval (the red segment on Figure 3). It is assumed (intuitive, but not demonstrated) that this interval contains approximately the 95% most likely true values able to generate that specific sample value.

Testing

The proportion selected by the lookup method was therefore put to a test by spreadsheet simulation of sample binomial distributions,

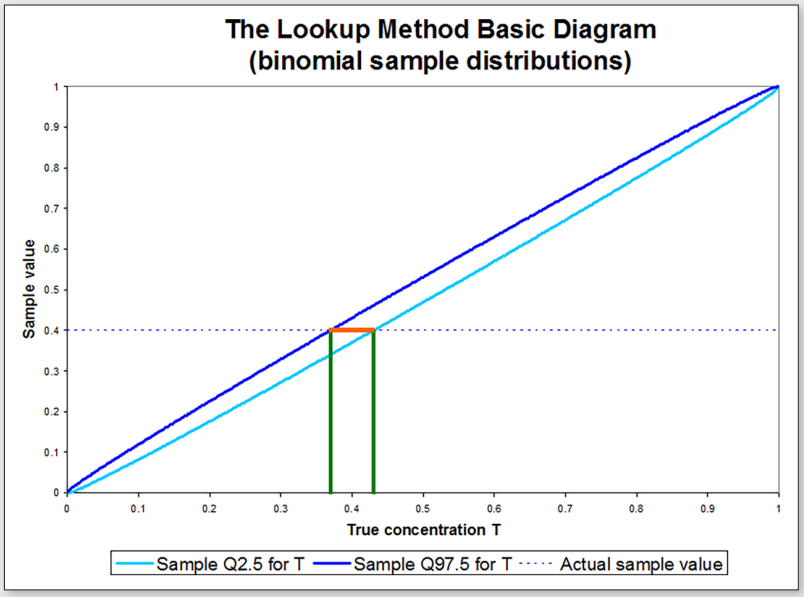

Figure 3. The lookup method basic diagram

for a full range of true values (concentrations) varying between the minimum of 0 and the maximum of 1, using 1,000 binomial trials for each. For numbers of success draws lesser than 5, i.e. concentrations lesser than 0.005, numerical stability problems altered the results to some degree. As kindly pointed out by a reviewer, the discrete nature of the binomial distribution is a significant factor in this observation. In any case, for these low numbers, the proportion of values within the lookup interval averaged to 95%, but with large variations, between 90.4% and 98.8%, with no pattern. Above these, the variations around 95% tend to become increasingly smaller, still without pattern, and still averaging to 95%. Given the numerical limitations imposed by the spreadsheet precision, the method was therefore reasonably validated, in conformity with our initial intuition.

Method comparisons

The simulation, however, is too heavy a process for routine applications, and as mentioned, of imperfect numerical stability. Using it as a benchmark, simpler - initially considered naive - methods were compared to it, namely:

- Gaussian confidence interval around the experimental sample value using the estimated binomial sampling variance.
- Lognormal and binomial confidence interval variants of the latter.
- Mirror images of the above two variants (skewed the other way).

As a comparison score, the maximum, relative, unsigned difference obtained for the two limits of the interval was used, along with an eyeball examination.

The following was observed:

- Surprisingly, the mirror images did not perform well, as the lookup interval was always slightly skewed to the right, likely a consequence of the variance heteroscedasticity we had previously ignored.
- The binomial intervals fared very erratically, possibly due solely to numerical problems. Where they seemed to behave properly, their results were however rather poor.

Table 1. Comparison of 95% Confidence Intervals on Simulated Sample Distributions

Sample Concentration	Lookup		Lognormal		Normal	
	LL	UL	LL	UL	LL	UL
0.005	0.002	0.010	0.002	0.011	0.001	0.009
0.010	0.005	0.017	0.005	0.017	0.004	0.016
0.050	0.038	0.064	0.038	0.065	0.036	0.064
0.100	0.083	0.119	0.083	0.120	0.081	0.119
0.250	0.224	0.277	0.224	0.278	0.223	0.277
0.270	0.243	0.297	0.244	0.299	0.242	0.298
0.350	0.321	0.379	0.321	0.380	0.320	0.380
0.500	0.469	0.530	0.470	0.532	0.469	0.531
0.900	0.880	0.916	0.882	0.919	0.881	0.919

- The normal and lognormal intervals both performed well at concentrations of 0.005 and above. Below these, numerical problems made the comparison unreliable.
- An approximate concentration threshold of 0.26 on the concentration was found to exist, that differentiated their performances: below 0.26 the lognormal intervals worked best, with the normal intervals performing better above 0.26.

A selection of these results is offered in Table 1.

Conclusion

In the case of binomial-like sample distributions, the lognormal and normal confidence intervals can be used, lognormal below concentrations of 0.26, normal ones above. When normal distribution are simulated instead, the normal confidence intervals are winning over lognormal at all concentrations, which is not surprising, but violates the expected distribution shapes at low concentrations. The simple rule described above and its concentration threshold of 0.26, should heuristically give good results in all practical cases.

Acknowledgments

An anonymous reviewer is hereby acknowledged for a very valuable suggestion.

Reference

1. Gy, P. M. (1998) *Sampling for Analytical Purposes*, John Wiley & Sons Ltd, Chichester, England.