Defence Research and Development Canada | Recherche et développement pour la défense Canada

DRDC | RDDC
technologysciencetechnologie

2016-04-28

DRDC-RDDC-2016-L130

Produced for: Capt(N) Garnier, Commanding Officer CFMWC

Scientific Letter

Challenges in interpreting multi-modal data from trials and simulations

Background

The Operational Research Team (ORT) is integrated at the Canadian Forces Maritime Warfare Centre (CFMWC) to provide an in situ analytical function, as part of the Centre's overall vision "to deliver valid and effective maritime warfare tactics and maritime force development advice" [1]. One of the tasks assigned to the ORT by the CFMWC Commanding Officer (CO) is to better integrate the analysis of results produced by the Maritime Modelling and Simulation Cell (MMSC) and those produced by the Operational Analysis sections for each battlespace. Doing so requires understanding some fundamental limitations of analyzing limited amounts of data. This letter addresses concerns related to the particular challenges of dealing with *multi-modal* data, i.e., with data that has more than one peak in its distribution, generated at in-water trials and/or simulations. This can be of a particular concern if the data set is small—as it normally must be for in-water trials due to the cost—and thus more complex methods of analyzing multi-modal data sets would not work, and the modes may not even be well defined. In order to illustrate the difficulties of interpreting this type of data while enabling wide distribution, and thus hopefully wide discussion and understanding of the issue, notional data was generated for the purposes of this letter. However, it has been designed to resemble a case seen in real in-water and computer-based models; the details of these observations will be reported on via appropriate means.

Constructing the data

The data used to illustrate the challenges posed by having multiple modes are based on a modified version of the sinc function on the interval [0,3], evaluated at increments of 0.1. For values less than or equal to 0.6, the function is set to the linear function (x / 4.2).[1] For all other values it is set to |sinc(x*4)|. These parameters were selected in such a way that the shape of the generated distribution resembles data distributions obtained from selected field trials. Finally, the data set was normalized (i.e., to have a total integrated value over the entire interval equal to one) so that it would correspond to a probability density function (PDF). This generates a function with four clear modes (local maxima), as seen in Figure 1. Note that for the purposes of this letter *mode* will be used in a somewhat informal sense to refer to the values clustered around each of the four local maxima at 0.6, 1.1, 1.9, and 2.7, with those clusters separated by

[1] These values were tuned somewhat in order to create a case that clearly illustrates the point that the mean may not lie on a mode, and to dampen the very high first peak of the sinc function.

the local minima. The single most common value in the distribution (at 1.1) will be referred to as the 'primary mode.' As the underlying generating function creates these modes, no amount of additional sampling will make the distribution normal (i.e., the central limit theorem does not apply in this case).

This distribution is representative of a situation where some underlying factor can take on one of four values, combined with some other factor or factors producing normal or near-normal variation on top of that. Common examples of multi-modal distributions in nature include average body weight and height, with the distinct peaks due to the underlying factor of gender (most mammals) or caste (e.g., ants). Another example could be the time taken to go between two cities, where multiple routes are available with distinct mean travel times (the underlying factor), but with there always being some variation in individual trip time on top of that. In military trials, examples can include some change in manoeuvre by a weapon or target that causes a jump (positive or negative) in the performance of the weapon, rather than a smooth variation. In order to simplify the dataset, no outliers have been generated. Outliers could be usually attributed to some sort of unique cause (possibly even to non-linear feedback loops in the system). The discussion of the non-linear behaviour of a system is beyond the scope of this letter.

In order to generate random numbers from this empirical distribution, the cumulative density function (CDF) was calculated and used as a look-up table for random numbers generated uniformly on the interval [0,1]. For the purpose of this paper, 40 random numbers were generated; these were divided into sets of five and ten. The distribution of their associated modes (defined here as the local maximum that would be found from that point by a hill climbing algorithm) are in Tables 1 and 2, with the overall mean for each set given as well.

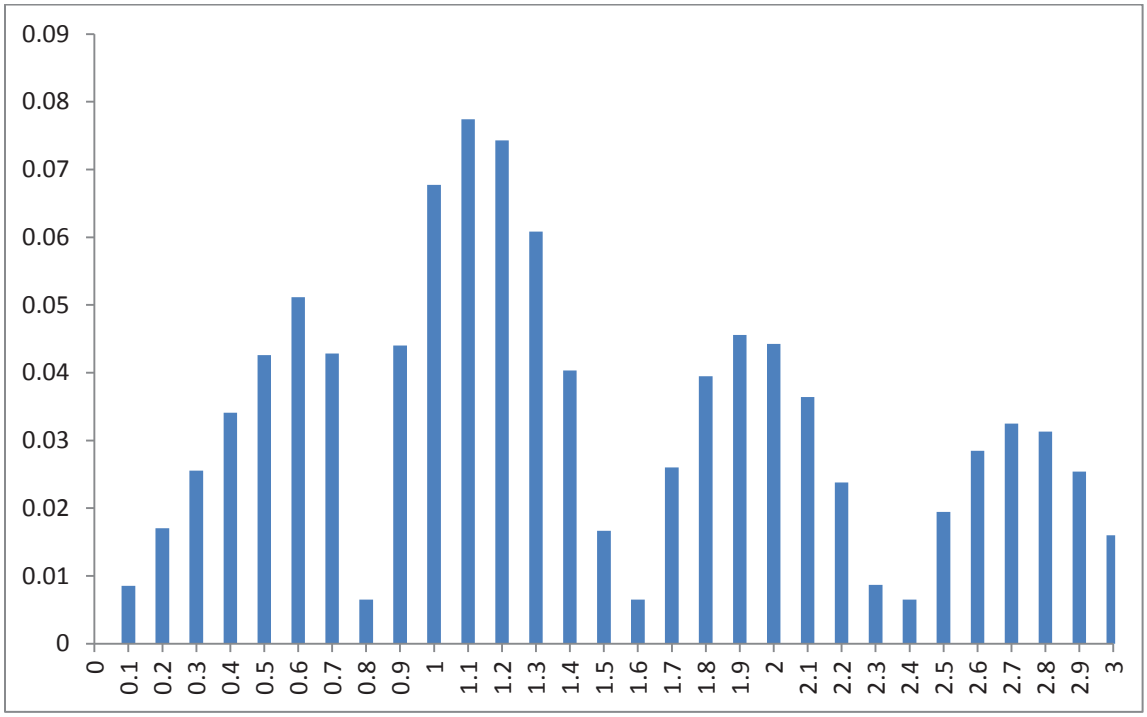

***Figure 1**: Notional multi-modal data distribution.*

Table 1: *Associated mode of random values in eight sets of five values, with 95% confidence intervals (C.I.) based on standard error.*

	Associated Mode						
Random Set	0.6	1.1	1.9	2.7	Mean		95% C.I.
1	1	1	2	1	1.64	±	0.75
2	0	3	1	1	1.62	±	0.61
3	1	4	0	0	0.98	±	0.23
4	1	2	2	0	1.36	±	0.55
5	1	3	0	1	1.28	±	0.70
6	0	3	1	1	1.56	±	0.52
7	2	2	1	0	1.10	±	0.52
8	0	1	2	2	2.08	±	0.48

Table 2: *Associated mode of random values in four sets of ten values, with 95% confidence intervals (C.I.) based on standard error.*

	Associated Mode						
Random Set	0.6	1.1	1.9	2.7	Mean		95% C.I.
1	1	4	3	2	1.63	±	0.48
2	1	7	1	1	1.17	±	0.32
3	1	6	1	2	1.42	±	0.44
4	2	3	3	2	1.59	±	0.47

Discussion of the data

A key point to consider when looking at Figure 1 is that this distribution represents the behaviour of a system under a single set of conditions. Generally, what one is hoping to understand during experimentation is how the behaviour of the system changes under different conditions (e.g., when adopting different tactics, or when evaluating different threats). Consider an in-water trial where one hopes to compare two variables with two possible outcomes each. At the very most perhaps twenty individual runs can be done at a typical trial, so one can only hope to collect five data points for each set of conditions. The challenge is then how one can compare several distributions of these types to each other, where e.g., the whole distribution may shift slightly left or right, or the relative size of the modes may change.

First one can consider an ideal situation where it is possible to build up a relatively accurate estimate of the underlying distribution (as in Figure 1). The mean of this distribution is approximately 1.4, and the median is 1.3. Neither of them is a particularly good estimate of the primary mode (the global maximum) at 1.1, nor do they convey anything about the local maxima whose peak values are each about two thirds of the primary maximum (at 0.6 and 1.9), let alone the shortest maximum at 2.7 whose peak is nonetheless over 40% the height of the global maximum. Given this lack of representativeness and descriptive value, comparisons made between multiple distributions of this type using these overall mean, median, or global mode values are potentially not very meaningful. Using just the primary mode (i.e., the global maximum) is perhaps somewhat more representative, but does not allow one to compare potential shifts between the relative peak sizes (i.e., if the number of peaks is the same, but the relative magnitude of the peaks is different).

The situation is even less clear when one only has a small number of samples to work with (e.g., the cases in Tables 1 and 2). When working with only five samples per set (a typical number for in-water trials), mean values between 0.98 and 2.08 are obtained (Table 1). The range is a bit narrower when looking at sets of ten, but sample means still vary between 1.17 and 1.63 (Table 2). Looking at the 95% confidence intervals on the samples, it would be very difficult to detect changes smaller than about 0.8 or 1.0, without even considering comparing changes in the relative frequency of the modes. Notably sets 3 and 8 in Table 1 do not have overlapping confidence intervals, despite having been sampled from the same underlying distribution.

Looking at how the sample points fall into the modes for samples of size five (Table 1), only in half of the sets is the primary mode of the distribution the sole primary mode of the sample. In two of eight it has the same value as another mode, and in another two it is a secondary mode. In half of the sets one of the modes is completely missed, and in one case two of them were missed. When trying to compare such samples across conditions, one would have no information about the missing modes, which are nonetheless important aspects of the underlying distribution. We could also easily incorrectly select the primary mode of some sets, and perhaps conclude that different conditions had different primary modes when they in fact did not.

Things are slightly better for the samples grouped in sets of ten, with every mode represented in each sample, and the underlying primary mode being the sample's primary mode in three of the sets and tied for the primary in the fourth case. However, the pattern of the secondary modes is somewhat erratic. If only some of the modes correspond to 'success' in an encounter, not having an accurate picture of these secondary modes may lead to incorrect conclusions about which conditions have better overall success.

This division into modes also assumes one knows how many modes there are beforehand, and roughly where they are. This would likely be not at all obvious. For reference, histograms of the underlying data corresponding to Table 2 are reproduced in Annex A. Set 2 looks unimodal with a strong peak at 1.2, set 4 looks bimodal, and sets 1 and 3 look to be positively skewed, with set 1 having one mode and set 3 perhaps having two.

Discussion and recommendations

Given the above, it is apparent that comparing relatively small samples of data generated from systems that are inherently multi-modal requires care, and one cannot rely on simple measures of central tendency such as median and mean. To build up an accurate estimate of relative peak sizes and overall shifting of peaks between several conditions, one will likely need many tens if not hundreds of points for each condition. The only feasible way to generate that many points for warship manoeuvres is to use computer-based simulations. However, there will remain a challenge in validating how well the distributions will fit real-world data, as the in-water data for each of the conditions may only be representative of some of the modes. It is also worth noting that the challenges noted above do not even consider the effect of outliers, which are often observed in physical trials due to the intervention of unanticipated external factors (e.g., equipment malfunction, shifting environmental conditions).

However, there are means to resolve this conundrum. Rather than focusing only on the overall multi-modal performance factor, one can attempt to measure the underlying variables driving the distribution. What is perhaps more realistically achievable from a validation perspective is to understand separately a) the conditions that cause the disturbances that move the final result

between the different modes, and b) how the overall distribution shifts when the effect of the disturbances is accounted for. This could potentially be dealt with by running a non-linear multi-variate regression with the number of disturbances observed as one of the predictors. This has the advantage of allowing many trials with different factors to be combined into one model treating the data together in a controlled manner, facilitating a forecasted outcome as well as identifying significant factors contributing to that outcome. However the needed number of samples will be large, perhaps on the order of 10–40 [2]. This will increase as the number of predictive variables increases, perhaps more if the underlying distributions of those variables are non-linear and non-normal. An advantage of focusing on the underlying drivers is that they may occur more than once per run, allowing many times the measurements to be taken. As those underlying distributions are better characterized, simulations can be run to build up a picture of the resulting higher level distribution.

With all that said, the overall recommendation is that care must be taken when presenting and comparing data from in-water trials where the distribution of the underlying measurement is known (or can be expected) to be multi-modal. While there are some techniques to cope with this, they still may require more trials than is realistic and affordable to run. When presenting results, one must also be on guard for the tendency of individuals to be insensitive to sample size, and misjudge how representative a small sample is of the overall population [3]. In particular, straight comparison of measures such as average or mean should be avoided. Also, alternate measures of effect or performance that are known to be unimodal can be presented alongside the multi-modal measures.

This also indicates that there will be difficulties in doing detailed validation of computer-based models based only on overall measures of performance. What will perhaps be more productive is ensuring the lower level behaviours of the models are representative of reality, to provide some confidence that the complex performance distributions generated are realistic. As the average number of disturbances per run may be higher than one, data on these may accumulate faster than the number of test runs.

Finally, the interested reader working in this area can contact the author to obtain practical examples of this phenomenon.

Prepared by: Matthew R. MacLeod, CFMWC ORT (DRDC – Centre for Operational Research and Analysis).

References

[1] National Defence (2016), *Canadian Forces Maritime Warfare Centre*, http://halifax.mil.ca/CFMWC/ [intranet] (accessed 16 April 2016).

[2] M. Wulder (1998), A Practical Guide to the Use of Selected Multivariate Statistics, Pacific Forestry Centre, Natural Resources Canada.

[3] A. Tversky and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185 (4157), 1124–1131.

Annex A Histograms of generated data

This annex contains the histograms of the data sets in Table 2, in order to illustrate the difficulty in inferring the shape of the original distribution in Figure 1, even with 10 samples. The sets in Table 1 are not reproduced as they are even less clear, containing only 5 samples.

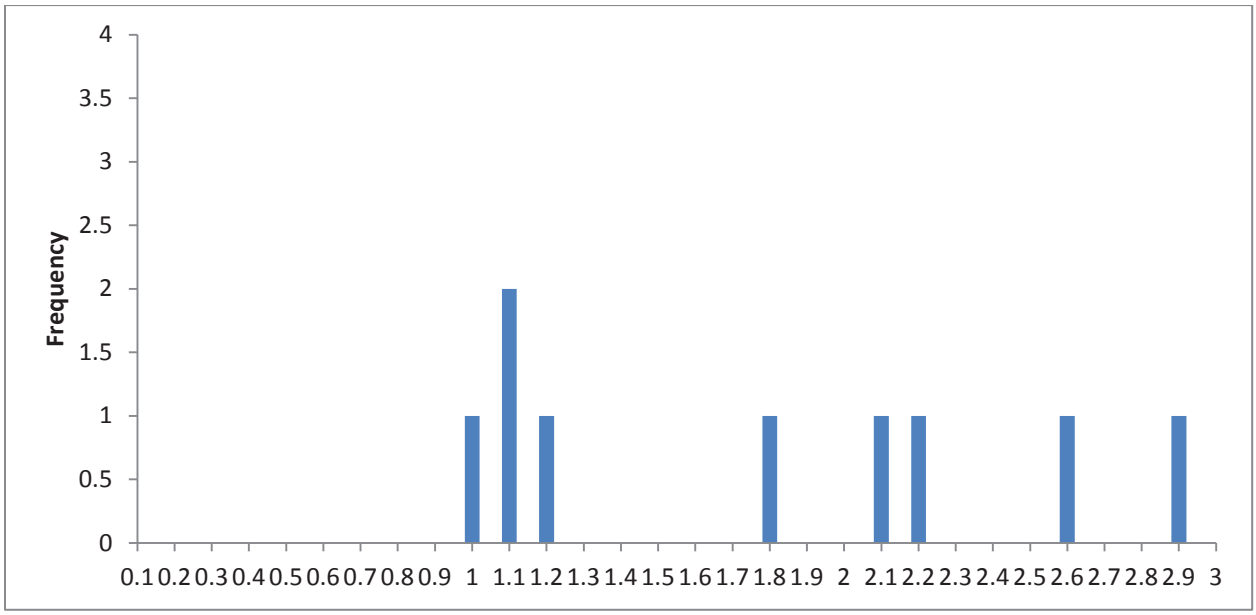

Figure A.1: *Histogram corresponding to set 1 from Table 2.*

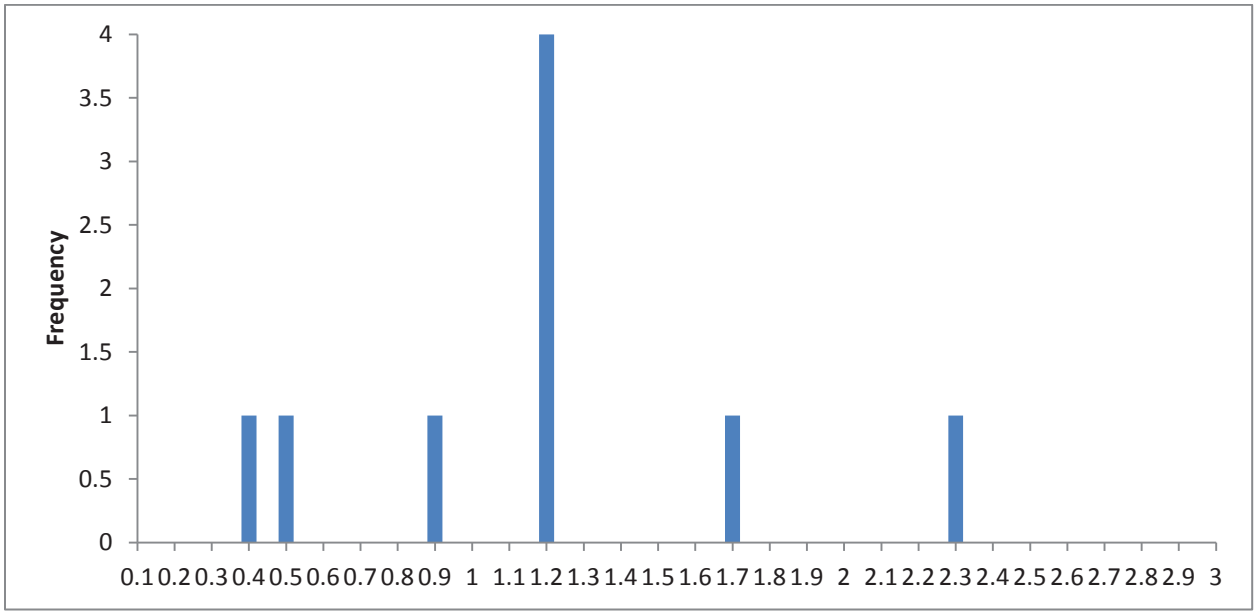

Figure A.2: *Histogram corresponding to set 2 from Table 2.*

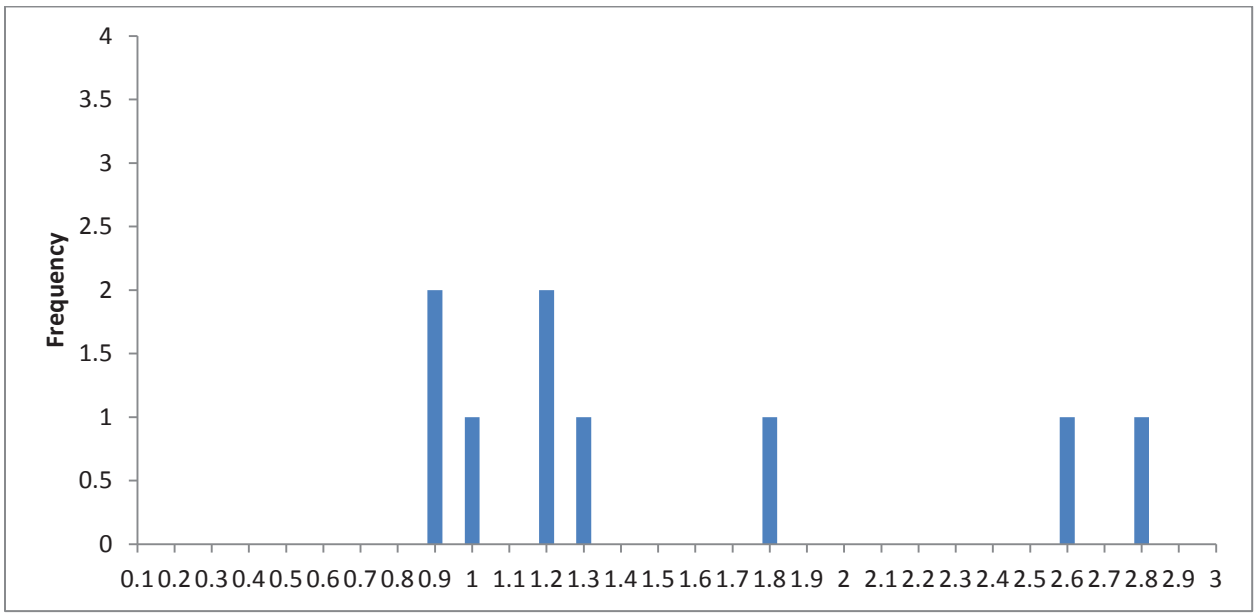

Figure A.3: *Histogram corresponding to set 3 from Table 2.*

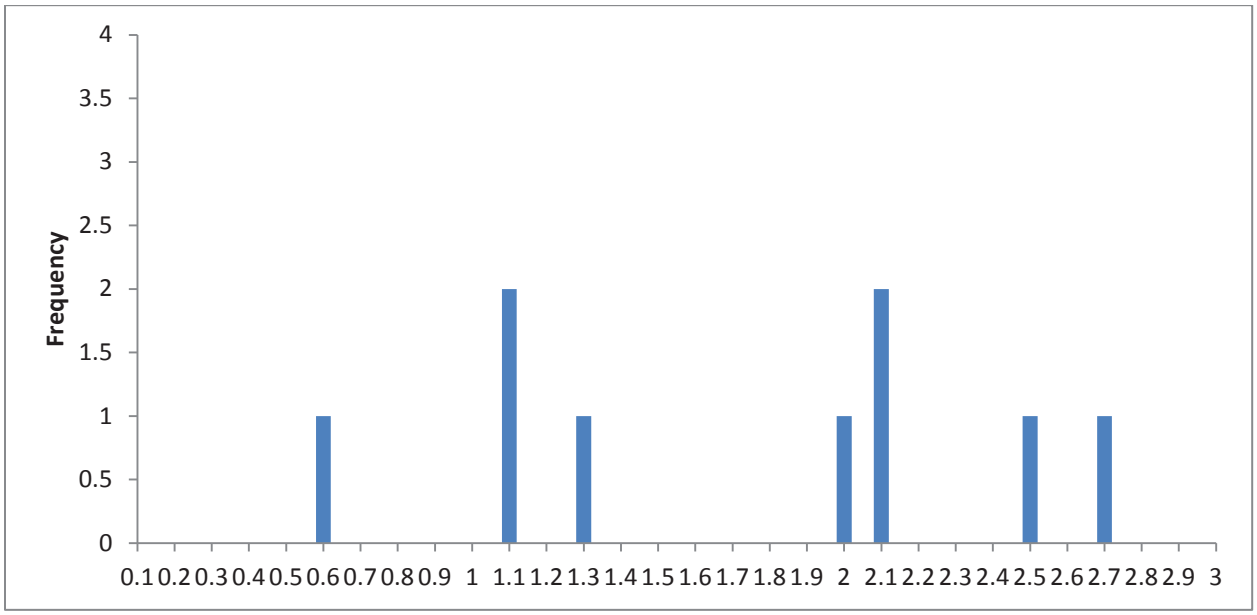

Figure A.4: *Histogram corresponding to set 4 from Table 2.*