Benchmarking of Direct Counting Approaches

Vladimir Shigunov, *DNV GL SE Maritime, Hamburg, Germany*, vladimir.shigunov@dnv.com
Cleve Arnauld Wandji, *Bureau Veritas, Paris, France*, cleve.wandji@bureauveritas.com
Vadim Belenky, *David Taylor Model Basin, W. Bethesda, Maryland, USA*, vadim.belenky@navy.mil

ABSTRACT

Benchmarking and comparative testing of three approaches for direct counting of stability failures are described. These approaches are based on estimation of failure rate from sample data using exponential distribution, statistical frequency of failures and binomial distribution. All three approaches were included in the draft Explanatory Notes for the Second-Generation IMO Intact Stability Criteria. The benchmarking is carried out using synthesized data following Poisson distribution. Brief description of the step-by-step approaches is included in the paper for the sake of reader's convenience.

Keywords: *IMO, Second-generation intact stability criteria, Direct stability assessment, Direct counting, Failure rate.*

1. INTRODUCTION

The second-generation intact stability criteria, published by IMO for a trial use as MSC.1/Circ.1627, contain a provision for application of state-of-the-art numerical simulations, referred as *direct stability assessment*. Requirements for the direct stability assessment are detailed in the Explanatory Notes; the draft was recently approved at the 8th session of the Ship Design and Construction Subcommittee (SDC 8/WP.4) of IMO. These Explanatory Notes are expected to be approved and published by the IMO Maritime Safety Committee (MSC) in 2022. Both probabilistic and deterministic criteria are applicable with direct stability assessment (MSC.1/Circ.1627 paragraph 3.5.3.1.4).

A probabilistic criterion is formulated in terms of the rate of stability failures, i.e. a number of failures per time unit. Procedure of estimation of the failure rate from time series with observed failures is referred to as *direct counting*. Three direct-counting procedures are described in sections 3.3, 3.4 and 3.5 of the Appendix 4 of the draft Explanatory Notes. All these procedures use Poisson process model to relate probability of failure with time of exposure. Brief overview of a Poisson process is given in the next section of this paper.

Since application of Poisson process model requires adoption of certain assumptions, it makes sense first to test the direct counting procedures for data that actually follow Poisson distribution, where the event rate is known. For this purpose, the data are not obtained from numerical simulation of ship motions in waves but generated in such a way that they comply with Poisson process assumptions. The objective of this testing is to check if these direct counting procedures are capable to recover this given event rate. The second objective is to evaluate uncertainty quantification techniques included with these three direct counting procedures.

2. POISSON PROCESS

Poisson distribution (e.g. Hayter, 2012, or Ryan, 2007) of a discrete random variable is used to describe a number of random events that occur within certain specified boundaries. A Poisson process is a model for a series of these discrete events. Application of Poisson distribution for stability failures in the context of the Second-Generation Intact Stability Criteria is described by Shigunov (2019); a summary of useful properties of a Poisson process is provided here.

For a Poisson process with a constant rate $r > 0$, the number of events N in a time interval of length t satisfies the Poisson distribution

$$f(k) = p\{N(t) = k\} = (rt)^k \cdot e^{-rt}/k! \qquad (1)$$

which expresses the probability of occurrence of $k = 0, 1, \ldots$ events during a time interval t. A special case of eq. (1) is $k = 0$, which corresponds to the probability that no failures occur during time t:

$$p \equiv p\{N(t) = 0\} = e^{-rt} \qquad (2)$$

From eq. (2), it follows that the probability that at least one failure happens during time *t*, i.e. that $k>0$, (loosely formulated: "probability of stability failure during time *t*") is

$$p^* \equiv p\{N(t) > 0\} = 1 - p\{N(t) = 0\} \\ = 1 - p = 1 - e^{-rt} \qquad (3)$$

The mean of a Poisson process, i.e. the mean number of events per interval *t*, is

$$\mu_N = \sum_{k\geq 0} kf(k) = rt \qquad (4)$$

i.e. the rate *r* is equal to the expected number of events per time unit:

$$r = \mu_N/t \qquad (5)$$

A useful property of a Poisson process is that time intervals between events are independent random variables, exponentially distributed with rate *r* (and vice versa: if the time intervals between events are not exponentially distributed, the process will not be a Poisson process). If T denotes the time to the next event, eq. (2) leads to

$$p\{T > t\} = e^{-rt} \qquad (6)$$

for $t > 0$ and 0 otherwise.

The independence of stability failures can be violated in practice by several effects, one of which is the *clustering* of big roll amplitudes: big roll amplitudes tend to appear in groups. The direct counting techniques have to include a way to "de-cluster" big roll amplitudes. The three methods, described in the Explanatory Notes (section 3 of the Appendix 4 of SDC 8/WP.4), ensure such de-clustering in different ways.

The first method (further referred to as M1 for brevity) is based on estimation of failure rate from sample data using exponential distribution, as described in section 3.3 of Appendix 4 of SDC 8/WP.4. In this method each simulation is conducted for arbitrary simulation time, but not longer than the occurrence of the first stability failure (note that simulation time is limited in any case due to self-repetition effects). The total simulation time t_t and the total number of encountered stability failures *N* are used to define the *maximum likelihood estimate of the stability failure rate* from eq. (4) as

$$\hat{r} = N/t_{\mathrm{t}} \qquad (7)$$

where total simulation time $t_{\mathrm{t}} = \sum_{i=1}^{N} T_i$; T_i are time intervals to each failure. Since the *sample mean time to failure* is $\bar{T} = t_{\mathrm{t}}/N$, eq. (7) can be also written as $\hat{r} = 1/\bar{T}$.

The second method (further referred as M2 for brevity) is based on estimation of failure rate from statistical frequency of failures, as described in section 3.4 of Appendix 4 of SDC 8/WP.4. This method employs eq. (3), i.e. makes use of the probability that at least one failure occurs during time t – hence any stability failures encountered in a simulation after the first one do not affect the estimate. Numerical simulations are carried out for a constant time Δ*t*, and the probability of at least one failure in a single simulation is estimated as $\hat{p} = N/M$, where *N* is the number of simulations in which at least one stability failure was encountered, and *M* is the total number of simulations.

The third method (further referred as M3 for brevity) is based on estimation of the rate from sample data using binomial distribution (Leadbetter et al. 2019). The method is described in section 3.5 of Appendix 4 of SDC 8/WP.4. In this method, numerical simulations are carried out for arbitrary time. All stability failures are recorded. To achieve independence of events, only one failure is counted during decorrelation time of roll motion. The latter is introduced in the section 3.8 of Appendix 4 of SDC 8/WP.4 and defined as a time for the envelope of autocorrelation function of roll motion to decrease to a specified threshold level, set to 0.05.

3. STEP-BY-STEP PROCEDURES

The step-by-step procedures are convenient for practical application and ensure that the described methods are applied in a uniform way. These procedures are provided in the Explanatory Notes (sections 3.3, 3.4 and 3.5 of Appendix 4 of SDC 8/WP.4). For the sake of reader's convenience these procedures are briefly summarized below.

In the M1 method, each simulation is conducted for arbitrary simulation time, but not longer than the occurrence of the first stability failure (note that simulation time is limited in any case due to self-repetition effects). After each such simulation,

1. record number of simulation *M*, number of encountered stability failures ΔN_M (1 or 0) and duration of simulation Δt_M;
2. calculate N^* as *N* before the simulation plus 1;

3. update the total number of failures as $N+\Delta N_M$; and the total simulation time t_t as $t_\mathrm{t}+\Delta t_M$;
4. update the maximum likelihood estimate (MLE) of failure rate as $\hat{r} = N/t_\mathrm{t}$;
5. update the conservative estimate of MLE of failure rate as $\hat{r}^* = N^*/t_\mathrm{t}$;
6. update the upper boundary of the 95%-confidence interval of failure rate using equation $r_\mathrm{U} = 0.5\chi^2_{1-0.05/2,2N^*}\hat{r}^*/N^*$; and
7. update the lower boundary of 95%-confidence interval of failure rate, $r_\mathrm{L} = 0.5\chi^2_{0.05/2,2N}\hat{r}/N$.

In the M2 method, numerical simulations are carried out for a constant simulation time Δt (which is limited by self-repetition effects). After each simulation,

1. record the number of realization M and whether this realization led to at least one failure ($\Delta N_M = 1$) or not ($\Delta N_M = 0$);
2. update the total number of failures as $N + \Delta N_M$;
3. calculate the probability of at least one failure in single simulation as $p = N/M$ and estimate of failure rate as $r = -\ln(1-p)/\Delta t$;
4. update the upper boundary of 95%-confidence interval of probability of at least one failure in a single simulation as $p_\mathrm{U} = 1$ for $N = M$ or $p_\mathrm{U} = v_1 F_{v_1,v_2,1-0.05/2}/(v_2 + v_1 F_{v_1,v_2,1-0.05/2})$ otherwise, with $v_1 = 2(N+1)$ and $v_2 = 2(M-N)$;
5. update the lower boundary of 95%-confidence interval of probability of at least one failure in single simulation as $p_\mathrm{L} = 0$ for $N = 0$ or $p_\mathrm{L} = v_1 F_{v_1,v_2,0.05/2}/(v_2 + v_1 F_{v_1,v_2,0.05/2})$ otherwise, with $v_1 = 2N$ and $v_2 = 2(M-N+1)$;
6. estimate the upper boundary of 95%-confidence interval of failure rate, $r_\mathrm{U} = -\ln(1-p_\mathrm{U})/\Delta t$;
7. estimate the lower boundary of 95%-confidence interval of failure rate, $r_\mathrm{L} = -\ln(1-p_\mathrm{L})/\Delta t$.

In the M3 method, numerical simulations are carried out for arbitrary simulation time (limited by self-repetition effects); all stability failures are recorded, but not all are counted. Binomial distribution is applied to describe the probability that there are $N_{a\mathrm{U}}$ independent stability failures (i.e. up-crossing events of a level a or down-crossing events of a level $-a$) out of total $N_a = \sum_{k=1}^{N_\mathrm{r}} N_k$ instances of observation of roll motion or lateral acceleration amplitude. Here, N_r is the total number of records comprising the data set of observation and N_k, $k = 1, \dots, N_\mathrm{r}$, denotes the number of observations in each record (the records may contain different numbers of observations and be of different durations).

The first stability failure after initial transition time is an independent event; the next independent stability failure is counted only after decorrelation time T_dc has passed. The total number of independent stability failures is $N_{aU} = \sum_{k=1}^{N_\mathrm{r}} N_{Uk}$, where N_{Uk} is the number of independent stability failures observed during the k-th record.

The failure rate is estimated as $\hat{p} = N_{aU}\Delta t/T_a$, where Δt is time increment used in simulation and $T_a = \sum_{k=1}^{Nr}(N_k\Delta t - T_\mathrm{ramp})$ is the total time of all records, with the constant ramp time T_ramp excluded to account for initial transients.

The number of independent stability failures N_{aU} is a random binomial-distributed variable. The binomial distribution has only one parameter, the probability that the event will occur at any particular instant of time. This probability can be estimated as $\hat{p} = N_{aU}\Delta t/T_a$. The variance of N_aU can be estimated as $\hat{V}_{NU} = T_a\hat{p}(1-\hat{p})/\Delta t$, and the boundaries of the confidence interval of the failure rate estimate were computed, using normal approximation for binomial distribution, as $r_\mathrm{U,L} = (N_{aU} \pm Q_N(0.5(1+P_\beta))\hat{V}_{NU}^{1/2})/T_a$, where Q_N is the quantile of the standard normal distribution and P_b is the accepted confidence probability. For $P_b = 0.95$, $Q_N(0.5(1+P_\beta)) = 1.96$.

4. INPUT DATA AND CALCULATIONS

Three authors independently executed the step-by-step procedures described in the previous section. The objectives of testing were to find out, whether

- the procedures are uniformly understood,
- there could be any improvements in the text,
- misinterpretation is possible,
- all the authors obtain the same results using the same procedure, and
- all procedures are able to recover correct result.

The overall objective of this study was to test the direct counting methods in "ideal" conditions, where the data are generated in such a way that they comply with Poisson process assumptions and the "true" rate of events is known. Based on this experience, understanding and uniform interpretation of the detailed "step-by-step" descriptions could be ensured. Two tests were undertaken, using

- a single data set to focus on comparison of numerical results against a known answer and verify the interpretation of the detailed "step-by-step" procedures, and
- multiple data sets to focus on verification of the calculation of confidence interval: the number of successful estimates should be close to the confidence probability.

In the single data set test, the rate of events r was set to $7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$ to generate a sample of $N = 25$ exponentially distributed times T_i between failures, $i = 1,2, \dots, N$. Note that a variable T, exponentially distributed with the rate r, can be generated as $T = -\ln x / r$, where x is a random variable drawn from a uniform distribution on the unit interval $(0,1)$ (in MS Excel, `rand()` function can be used).

The generated time intervals between events are shown in Table 1 (the total time is 28093.6081 s). The maximum likelihood estimate of the rate from the full generated sample is $\hat{r} = 8.899 \cdot 10^{-4}\ \mathrm{s}^{-1}$, and the estimates of the mean and standard deviation of time between events are $\hat{T} = 1123.74$ s and $\hat{\sigma}_T = 1005.55$ s, respectively.

Table 1. Generated time intervals between events used in test concerning single data set

2733.980	2679.500	445.665	258.192	1073.380
2792.510	280.590	1820.620	942.395	237.282
524.140	2362.350	546.241	1218.310	1121.510
1217.190	288.416	465.511	74.271	24.568
48.523	2658.140	2993.350	855.247	431.727

This test verified the estimates of the failure rate and upper and lower boundaries of its 95%-confidence interval provided by the three methods.

What are the expected results of the test? In an ideal case, the direct counting methods should be able to capture the full data set in Table 1. Thus, the expected result is the maximum likelihood estimate, i.e. $\hat{r} = 8.899 \cdot 10^{-4}\ \mathrm{s}^{-1}$, further referred to as the *benchmark estimate*. However, the compared direct counting methods are intended for practical post-processing of ship motion simulation data and include provisions to insure independence of events. As a result, the outcome of the test may not necessarily recover the benchmark estimate exactly, hence one of the checks is to compare the rate estimates by the three methods with the benchmark estimate.

On the other hand, the ultimate aim of direct counting is the true rate value, i.e. $7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$. As the dataset is finite, the rate estimate is a random number, comparison of which with the true rate is meaningless. However, the confidence interval, if it is correctly constructed about this estimate, should contain the true value with the specified confidence probability, i.e. 95%-confidence interval is expected to contain the true rate $r = 7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$ with a 95%-chance. Since each considered method applies own technique to construct a confidence interval, the first logical step would be to see whrther the three confidence intervals do contain the true value. However, a more conclusive test would be to check whether the true rate is within the confidence interval with 95%-confidence probability. Such a test requires multiple data sets.

In the test concerning multiple data sets, the same rate of events $r = 7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$ was applied to generate $M = 10^4$ data sets, each consisting of $N = 25$ exponentially distributed time intervals between events $T_{i,j}$, where $i = 1, \dots, N$ and $j = 1, \dots, M$. The confidence intervals were verified, for $N = 1,2, \dots, 25$, by counting the number of cases, out of $M = 10^4$, where the true rate value $r = 7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$ is within the confidence interval, above its upper boundary or below its lower boundary: if the confidence intervals are correct, such cases should comprise 95%, 2.5% and 2.5%, respectively, of all cases (if the confidence probability is set to 0.95).

5. RESULTS: SINGLE DATA SET

In this test, estimates of the failure rate and its 95%-confidence interval with the three methods were compared.

First, present the interval data from Table 1 in a format, typical for outcome of numerical simulation of ship motions. For the M1 method, duration of a simulation is arbitrary, and the result does not depend on the duration of individual simulations. The maximum length of a simulation is defined by self-repetition effects. For comparison purposes, the maximum length of a simulation was set to 1800.0 s. The transformation of data in Table 1 into an input for the M1 method is shown in Fig. 1, where observed events are depicted as dots.

Reformatting the data from Table 1 to Fig. 1 is straight forward. The time until the first stability failure is 2733.980 s. It is larger than the simulation length of 1800 s, so the fisrt record does not have any observed events. The fisrt event is observed during the second record at the time instant 2733.98 s −

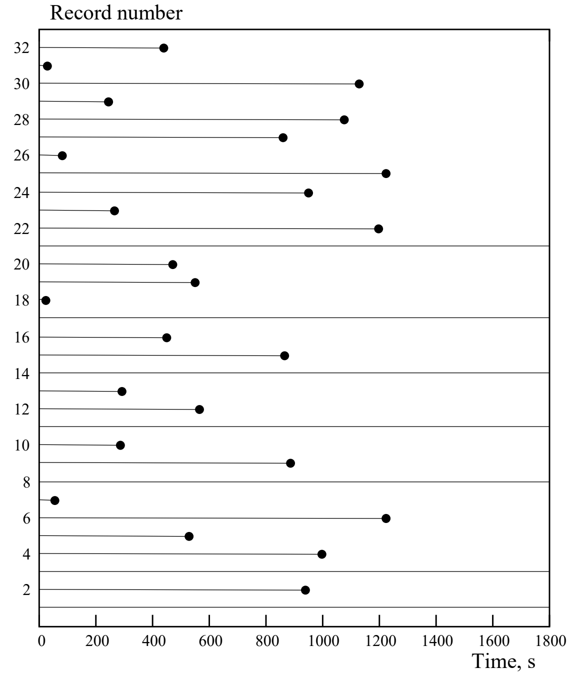

Fig. 1. Representation of events observed per record based on Table 1, formatted as input for M1

$1800.0\ \text{s} = 933.98\ \text{s}$ etc. Since some time intervals between failures in Table 1 exceed the assumed simulation duration, Fig. 1 contains 32 records, while there are only 25 events: seven record do not contain any observed events.

Table 2 shows results including the index of a record M, observed number of failures ΔN_M, time before failure Δt_M (if no failure was observed, $\Delta t_M = 1800\ \text{s}$), the total time (cumulative for all records) t_t, maximum likelihood estimate of failure rate $\hat{r}$ and upper r_U and lower r_L boundaries of the 95%-confidence interval of the failure rate. Since the time until the first stability failure was 2733.980 s, no stability failure occurred ($\Delta N_1 = 0$) in the first simulation ($M = 1$) of duration $\Delta t_1 = 1800$ s. The first stability failure occurred in the second simulation ($\Delta N_2 = 1$) at the time instant 2733.980 s – 1800.0 s = 933.98 s after its start, at which this simulation stopped ($\Delta t_2 = 933.98\ \text{s}$). In the third simulation of duration $\Delta t_3 = 1800\ \text{s}$, again no stability failure occurred ($\Delta N_3 = 0$) until its end: the time to the second failure was 2792.510 s, i.e. 2792.510 s−1800.0 s=992.51 s after the start of the fourth simulation ($\Delta t_4 = 992.51\ \text{s}$) etc. Since the time until the last stability failure was 431.727 s, one stability failure ($\Delta N_{32} = 1$) occurred in the last, 32-nd simulation, at the time instant $\Delta t_{32} = 431.727$ s after ist start (at this instant, the simulation stopped). In total, 32 simulations of the total duration 28093.6081 s were conducted, in which 25 stability failures were encountered. For the complete dataset (32 records), $\hat{r} = 8.899 \cdot 10^{-4}\ \text{s}^{-1}$ (which agrees with the benchmark estimate), $r_\text{U} = 1.271\ 10^{-3}\ \text{s}^{-1}$ and $r_\text{L} = 5.759 \cdot 10^{-4}\ \text{s}^{-1}$.

Table 2. Application example of M1-procedure

M	ΔN_M	Δt_M, s	N	N^*	t_t, s	$\hat{r}$, s^{-1}	$\hat{r}^*$, s^{-1}	r_U, s^{-1}	r_L, s^{-1}
1	0	1800.0	0	1	1800.0	0.0	5.556e-4	2.049e-3	-
2	1	933.98	1	1	2733.98	3.658e-4	3.658e-4	1.349e-3	9.260e-6
3	0	1800.0	1	2	4533.98	2.206e-4	4.411e-4	1.229e-3	5.584e-6
4	1	992.51	2	2	5526.5	3.619e-4	3.619e-4	1.008e-3	4.383e-5
5	1	524.14	3	3	6050.64	4.958e-4	4.958e-4	1.194e-3	1.022e-4
6	1	1217.19	4	4	7267.83	5.504e-4	5.504e-4	1.206e-3	1.500e-4
7	1	48.52	5	5	7316.35	6.834e-4	6.834e-4	1.400e-3	2.219e-4
8	0	1800.0	5	6	9116.35	5.485e-4	6.582e-4	1.280e-3	1.781e-4
9	1	879.5	6	6	9995.85	6.002e-4	6.002e-4	1.167e-3	2.203e-4
10	1	280.55	7	7	10276.4	6.812e-4	6.812e-4	1.271e-3	2.739e-4
11	0	1800.0	7	8	12076.4	5.796e-4	6.624e-4	1.194e-3	2.330e-4
12	1	562.4	8	8	12638.8	6.330e-4	6.330e-4	1.141e-3	2.733e-4
13	1	288.4	9	9	12927.2	6.962e-4	6.962e-4	1.219e-3	3.183e-4
14	0	1800.0	9	10	14727.2	6.111e-4	6.790e-4	1.160e-3	2.794e-4
15	1	858.2	10	10	15585.4	6.416e-4	6.416e-4	1.096e-3	3.077e-4
16	1	445.6	11	11	16031.0	6.862e-4	6.862e-4	1.147e-3	3.425e-4
17	0	1800.0	11	12	17831.0	6.169e-4	6.730e-4	1.104e-3	3.080e-4
18	1	20.6	12	12	17851.6	6.722e-4	6.722e-4	1.103e-3	3.473e-4
19	1	546.3	13	13	18397.9	7.066e-4	7.066e-4	1.139e-3	3.762e-4
20	1	465.5	14	14	18863.4	7.422e-4	7.422e-4	1.178e-3	4.058e-4
21	0	1800.0	14	15	20663.4	6.775e-4	7.259e-4	1.137e-3	3.704e-4
22	1	1193.3	15	15	21856.7	6.863e-4	6.863e-4	1.075e-3	3.841e-4
23	1	258.2	16	16	22114.9	7.235e-4	7.235e-4	1.119e-3	4.135e-4
24	1	942.4	17	17	23057.3	7.373e-4	7.373e-4	1.127e-3	4.295e-4
25	1	1218.3	18	18	24275.6	7.415e-4	7.415e-4	1.121e-3	4.395e-4
26	1	74.3	19	19	24349.9	7.803e-4	7.803e-4	1.168e-3	4.698e-4
27	1	855.3	20	20	25205.2	7.935e-4	7.935e-4	1.177e-3	4.847e-4
28	1	1073.3	21	21	26278.5	7.991e-4	7.991e-4	1.175e-3	4.947e-4
29	1	237.3	22	22	26515.8	8.297e-4	8.297e-4	1.211e-3	5.200e-4
30	1	1121.5	23	23	27637.3	8.322e-4	8.322e-4	1.205e-3	5.275e-4
31	1	24.6	24	24	27661.9	8.676e-4	8.676e-4	1.248e-3	5.559e-4
32	1	431.7	25	25	28093.6	8.899e-4	8.899e-4	1.271e-3	5.759e-4

For the methods M2 and M3, the simulations were assumed to be of the same length 1800.0 s for comparison. Fig. 2 shows the data from Table 1 as events (depicted as dots) observed per record.

Reformatting the data from Table 1 to Fig. 2 is also straight forward. The only difference compared to the method M1 is that a record may have mutiple events (if the time between them is small enough to fit into a single simulation).

Note that records 5, 14 and 16 contain events, which are very close to each other. If the data were obtained from numerical simulations of ship motion, these events may be expected to be dependent. However, as the data in Table 1 follow Poisson process per the definition, such cases when events are too close do not represent a concern in this study.

The result of the method M2 depends on the exposure time Δt (or number M of simulations for the same total simulation time), therefore, several values of Δt were used for testing and comparison. Using the duration of each simulation $\Delta t = 1800$ s leads

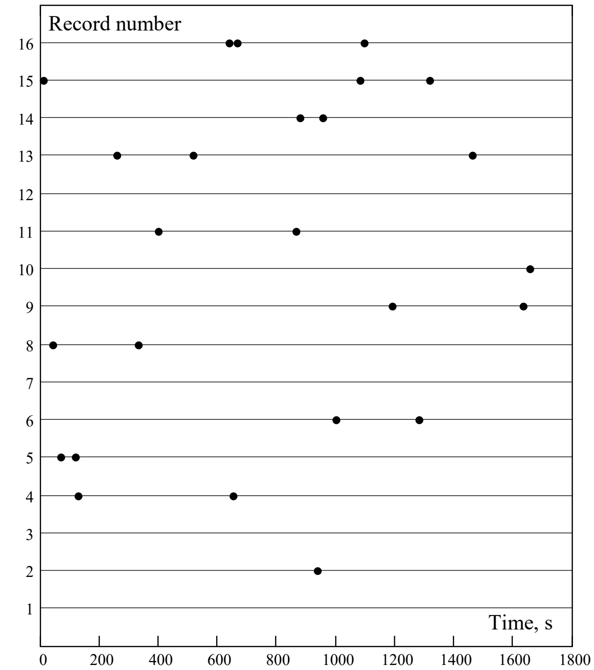

Fig. 2. Representation of events observed per record based on Table 1, formatted as input for M2 and M3

to 16 records in total and the following estimates: $r = 8.443 \cdot 10^{-4}\ \mathrm{s}^{-1}$, $r_\mathrm{U} = 1.321 \cdot 10^{-3}\ \mathrm{s}^{-1}$ and $r_\mathrm{L} = 5.094 \cdot 10^{-4}\ \mathrm{s}^{-1}$, which deviate from the benchmark estimate and results of the other two methods.

Reducing the exposure time Δt (i.e. increasing the number M of simulations) improves accuracy: for example, for a duration of each simulation of 1 s (which means $M = 28094$ simulations), results are very close to the benchmark estimate and results of the other two methods: $r = 8.902 \cdot 10^{-4}\ \mathrm{s}^{-1}$ (vs. the benchmark estimate $8.899 \cdot 10^{-4}\,\mathrm{s}^{-1}$), $r_\mathrm{U} = 1.314 \cdot 10^{-3}\ \mathrm{s}^{-1}$, $r_\mathrm{L} = 5.761 \cdot 10^{-4}\ \mathrm{s}^{-1}$.

Table 3 shows examples of results for M from 1 to 2000. For $M = 1$, only one ($M = 1$) simulation of the total duration 28093.6081 s was conducted, in which, the first stability failure occurred at the time instant 2733.980 s after the start ($N = 1$), after which, everything that happened in the simulation was ignored. As $M = 1$ and $N = 1$, $p = N/M = 1$. For $M = 2$, two ($M = 2$) simulations, each of the duration $0.5 \cdot 28093.6$ s $= 14046.8$ s were conducted. In the first of these simulations, the first stability failure occurred at the time instant 2733.980 s after its start, and the remaining part of this simulation was ignored. In the second simulation, the first stability failure (which is the tenth stability failure in Table 1) occurred at the time instant 15585.4 s – 14046.8 s=1538.6 s after its start, and the remaining part of this simulation was ignored. Since $M = 2$ and $N = 2$, $p = N/M = 1$ etc. For $M = 2000$, the duration of each simulation was $28093.6081\ \mathrm{s}/2000\ \approx 14.05$ s, thus all 25 stability failures were counted, and $p = 25/2000 = 0.0125$.

Table 3. Application examples of M2-procedure

M	Δt, s	N	p	$\hat{r}$, s^{-1}	r_U, s^{-1}	r_L, s^{-1}
1	28093.6	1	1	-	-	9.012e-7
2	14046.8	2	1	-	-	1.225e-5
3	9364.53	3	1	-	-	3.693e-5
4	7023.4	4	1	-	-	7.217e-5
5	5618.72	5	1	-	-	1.158e-4
6	4682.27	6	1	-	-	1.662e-4
7	4013.37	7	1	-	-	2.224e-4
8	3511.7	8	1	-	-	2.836e-4
9	3121.51	9	1	-	-	3.491e-4
10	2809.36	10	1	-	-	4.186e-4
11	2553.96	10	9.091e-1	9.389e-4	2.379e-3	3.465e-4
12	2341.13	11	9.167e-1	1.061e-3	2.632e-3	4.079e-4
13	2161.05	10	7.692e-1	6.785e-4	1.383e-3	2.867e-4
14	2006.69	13	9.286e-1	1.315e-3	3.148e-3	5.395e-4
15	1872.91	12	8.000e-1	8.593e-4	1.676e-3	3.909e-4
16	1755.85	12	7.500e-1	7.895e-4	1.493e-3	3.683e-4
17	1652.56	12	7.059e-1	7.405e-4	1.375e-3	3.513e-4
18	1560.76	13	7.222e-1	8.207e-4	1.495e-3	4.010e-4
19	1478.61	12	6.316e-1	6.753e-4	1.227e-3	3.272e-4
20	1404.68	15	7.5e-1	9.869e-4	1.742e-3	5.063e-4
50	561.87	22	4.4e-1	1.032e-3	1.576e-3	6.346e-4
100	280.94	23	2.3e-1	9.303e-4	1.398e-3	5.857e-4
200	140.47	24	1.2e-1	9.101e-4	1.355e-3	5.814e-4
500	56.19	25	5.0e-2	9.129e-4	1.348e-3	5.902e-4
1000	28.09	25	2.5e-2	9.012e-4	1.330e-3	5.829e-4
2000	14.05	25	1.25e-2	8.955e-4	1.322e-3	5.794e-4

Using simulations of constant duration 1800 s in the method M3 led to 16 equal records of total duration 28800 s and $r = 6.944 \cdot 10^{-4}\ \mathrm{s}^{-1}$ (vs. the benchmark estimate $8.899 \cdot 10^{-4}\,\mathrm{s}^{-1}$), $r_\mathrm{U} = 9.666 \cdot 10^{-4}\ \mathrm{s}^{-1}$ and $r_\mathrm{L} = 4.223 \cdot 10^{-4}\ \mathrm{s}^{-1}$. Cutting the duration of the last record up to the time instant of event (i.e. setting the total duration to 28093.6 s) led to $\hat{r} = 8.899 \cdot 10^{-4}\,\mathrm{s}^{-1}$ (which is equal to the benchmark estimate and the rate provided with M1 method), $r_\mathrm{U} = 1.239 \cdot 10^{-3}\ \mathrm{s}^{-1}$ and $r_\mathrm{L} = 5.441 \cdot 10^{-4}\,\mathrm{s}^{-1}$. The mathematical reason for the observed behavior of the rate estimate is not clear.

Comparison

Fig. 3 compares the estimates of the failure rate and the upper and lower boundaries of its 95%-confidence interval between the three methods (the exposure time in the method M2 was set to 1800 s) vs. the number of events, and Fig. 4 compares the failure rate and the upper and lower boundaries of its 95%-confidence interval for $N = 25$.

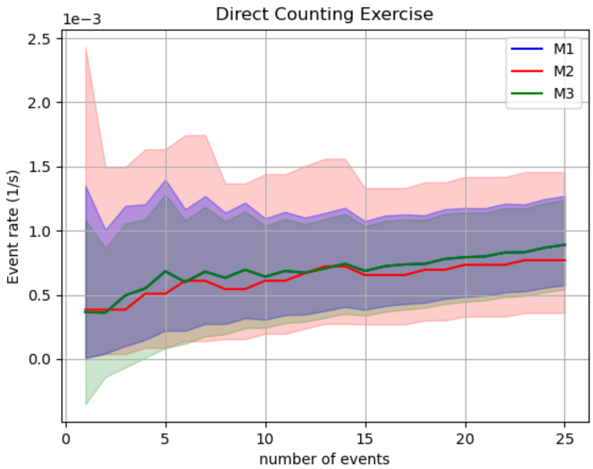

Fig. 3. Estimates of failure rate and upper and lower boundaries of its 95%-confidence interval (exposure time in M2-procedure is 1800 s) vs. number of events

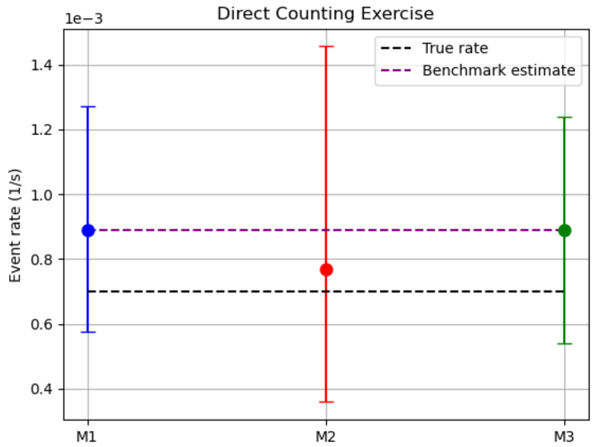

Fig. 4. Estimate of failure rate and upper and lower boundaries of its 95%-confidence interval for $N = 25$ (using exposure time 1800 s in M2-procedure)

All three confidence intervals do contain the true value of the rate $7.0 \cdot 10^{-4}\,\mathrm{s}^{-1}$, moreover, their boundaries are very close. The benchmark estimate $\hat{r} = 8.899 \cdot 10^{-4}\,\mathrm{s}^{-1}$ is reproduced by M1 and M3 procedures, while the M2 estimate is slightly different, but happen to be closer to the true rate for the considered sample.

To analyse the reason why the procedure M1 reproduces the benchmark estimate, note that the maximum likelihood estimate (step 4 for the M1 procedure) is essentially the same as the one applied to the data in Table 1 if the last simulation ends with a stability failure. However, if the last simulation does not end with a stability failure, this procedure provides a conservative estimate by assuming a stability failure just at the instant of stopping.

Similarly, the M3 procedure uses the same maximum likelihood estimate of the rate as was applied to the data in Table 1. As in this test the data points are assumed independent, the decorrelation time is zero and therefore, no data points were excluded. However, to reproduce the benchmark estimate exactly, the duration of the last records needs to be corrected by excluding the time after the last event.

To understand why the M2 procedure provides a different estimate, note that the formulation “at least one event”, used in the M2 procedure, means that the number of events per simulation can be one or two, or three etc. Thus if a simulation contains more than one event (which is a case for 10 records in Fig. 2), the events beyond the first one do not change the estimate. To see a limit behavior of the M2 procedure, the exposure time was set to 1.0 s; Fig. 5 and Fig. 6 show respective results (the other two procedures are unchanged). When the exposure time is significantly reduced, the rate estimate obtained with the M2 procedure becomes almost identical to those obtained with the M1 and M3 procedures.

This can be expected for two reasons: first, the number of failures N correctly captured if the simulation time Δt is sufficiently small, so that each simulation contains not more than one failure (in the example in Table 3, this for $\Delta t \leq 56.19$ s, which corresponds to $M \geq 500$. Second, the rate in the M2 method is estimated as $r = -\ln(1-p)\,/\Delta t$, where $p = N/M$. Therefore, $r = -\ln(1 - N/M)\,/\Delta t$, which converges to $r = N/(M\Delta t) = N/t_t$ for $\Delta t \to 0$ while $M\Delta t = \text{const} = t_t$, i.e. the rate estimate in the M2 method converges to the maximum likelihood estimate in the zero-limit exposure time.

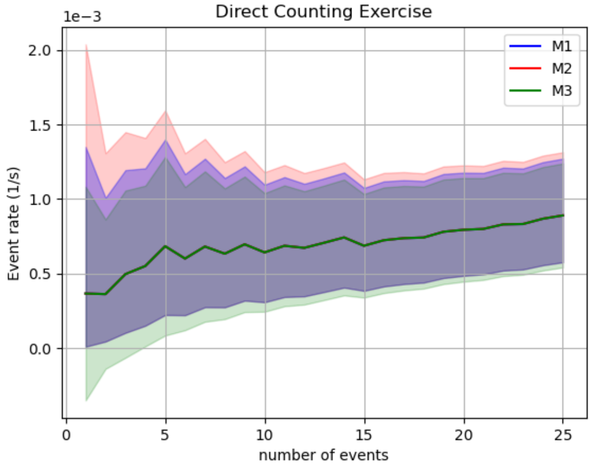

Fig. 5. Estimates of failure rate and upper and lower boundaries of its 95%-confidence interval (exposure time in M2-method 1 s) vs. number of events

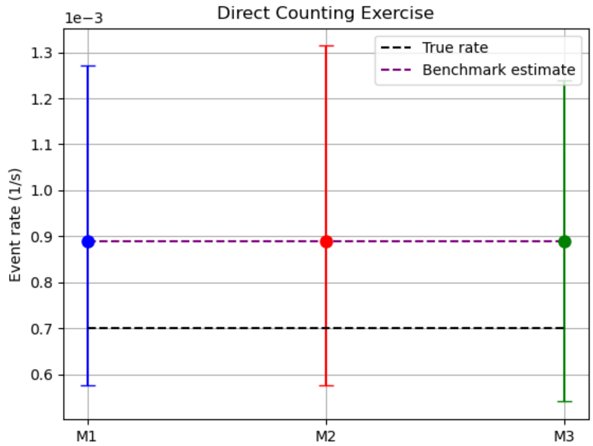

Fig. 6. Estimate of failure rate and upper and lower boundaries of its 95%-confidence interval for $N = 25$ (using exposure time 1 s in M2-method)

6. RESULTS: MULTIPLE DATA SETS

In this test, multiple ($M = 10^4$) data sets were generated, each consisting of $N = 25$ events. The number of cases, out of $M = 10^4$, was counted, when the true rate value $r = 7.0 \cdot 10^{-4}\ \mathrm{s}^{-1}$ is within the estimated confidence interval, above its upper boundary or below its lower boundary. If the procedures are as accurate as expected, the true rate value should be within the estimated confidence interval, above its upper boundary and below its lower boundary in about 95%, 2.5% and 2.5% of all cases, respectively (although random deviations from these numbers are expected).

Table 4 shows results for the sample sizes $N = 1,2, \dots, 25$ (in all cases, the number of data sets is the same $M = 10^4$). For M1-method, the number of "misses" in both directions, i.e. $r > r_\mathrm{U}$ and $r < r_\mathrm{L}$, is close to 2.5% for all sample sizes N, i.e. M1 method accurately estimates both the upper and lower boundaries of the 95%-confidence interval of failure rate for all sample sizes.

For the M2-method, the number of misses $r > r_\mathrm{U}$ is lower than 2.5% (significantly lower for small sample sizes N), i.e. the upper boundary of the confidence interval is slightly high.

Table 4. Sample size *N*, number of estimates above estimated upper boundary $r > r_\mathrm{U}$ and below estimated lower boundary $r < r_\mathrm{L}$ of 95%-confidence interval of failure rate, as well as number of estimates r_inside, which are within estimated 95%-confidence interval, depending on sample size for 10^4 data sets

N	$r > r_\mathrm{U}$, %			$r < r_\mathrm{L}$, %			r_inside, %		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	2.50	0.40	5.76	2.46	2.46	0.01	95.04	97.14	94.23
2	2.36	0.60	4.93	2.54	2.54	0.00	95.10	96.86	95.07
3	2.50	0.64	4.51	2.58	2.58	0.00	94.92	96.78	95.49
4	2.34	0.76	4.46	2.44	2.44	0.00	95.22	96.80	95.54
5	2.54	0.91	4.22	2.63	2.63	0.07	94.83	96.46	95.71
6	2.49	1.06	4.32	2.57	2.57	0.15	94.94	96.37	95.53
7	2.60	1.05	4.20	2.48	2.48	0.13	94.92	96.47	95.67
8	2.53	1.01	4.22	2.44	2.44	0.24	95.03	96.55	95.54
9	2.53	1.11	4.10	2.44	2.44	0.33	95.03	96.45	95.57
10	2.48	1.31	4.05	2.42	2.42	0.47	95.10	96.27	95.48
11	2.62	1.25	3.99	2.32	2.32	0.55	95.06	96.43	95.46
12	2.68	1.25	4.07	2.30	2.30	0.62	95.02	96.45	95.31
13	2.38	1.37	4.04	2.28	2.28	0.78	95.34	96.35	95.18
14	2.53	1.46	3.77	2.31	2.30	0.77	95.16	96.24	95.46
15	2.48	1.48	3.73	2.43	2.42	0.8	95.09	96.10	95.47
16	2.57	1.53	3.79	2.49	2.48	0.83	94.94	95.99	95.38
17	2.80	1.56	3.81	2.27	2.26	0.88	94.93	96.18	95.31
18	2.83	1.61	3.85	2.28	2.27	0.99	94.89	96.12	95.16
19	2.59	1.68	3.66	2.42	2.42	0.94	94.99	95.90	95.40
20	2.62	1.64	3.65	2.39	2.39	1.03	94.99	95.97	95.32
21	2.63	1.62	3.50	2.50	2.50	1.00	94.87	95.88	95.50
22	2.49	1.69	3.40	2.51	2.51	1.05	95.00	95.80	95.55
23	2.61	1.55	3.56	2.43	2.43	1.12	94.96	96.02	95.32
24	2.48	1.56	3.40	2.48	2.48	1.04	95.04	95.96	95.56
25	2.41	1.52	3.48	2.52	2.52	1.13	95.07	95.96	95.39

The number of under-estimates $r < r_{\mathrm{L}}$ is close to 2.5% for all sample sizes N, which means that this method accurately estimates the lower boundary.

For the M3-method, the number of misses $r > r_{\mathrm{U}}$ is slightly but systematically more than 2.5%, and the number of misses $r < r_{\mathrm{L}}$ is slightly but systematically less than 2.5%, especially at small N. The method shows some asymmetry in the evaluation of confidence interval. The total number of successful estimates, however, remains very close to the given confidence probability of 0.95.

7. CONCLUSIONS

The objective of the described effort was to compare three approaches to direct counting, included in the draft Explanatory Notes for the IMO Second-generation intact stability criteria. These three approaches are based on the estimation of failure rate from sample data using exponential distribution, statistical frequency of failures and binomial distribution.

A comparison of these methods was carried out using synthesized data set following Poisson distribution. The ability of these approaches to "de-cluster" large roll response remains outside of the scope of this paper. The advantage of using synthesized data is that the events are known to be independent, which is assumed in the derivation of the three tested procedures.

All three approaches were able to correctly estimate the failure rate – the true value of the failure rate was within the confidence interval. It was noted that the accuracy of the procedure using statistical frequency of failures improves with decreasing exposure time.

Constructing confidence intervals was benchmarked by repeating the estimation procedure 10^4 times and counting the number of successes (when the confidence interval contains the benchmark value). The estimate of the confidence interval was considered to be correct when percentage of successes was close to the accepted confidence probability. All three approaches demonstrated correct techniques for construction of confidence intervals.

The described effort used synthesized data following Poisson distribution. Further study should use data derived from simulation of ship motion, so that de-clustering capabilities of the three approaches could also be addressed. Another characteristic to compare is the practicability of the three methods in actual assessment.

8. ACKNOWLEDGEMENTS

The participation of the third author in the described effort was partially supported by the US Office of Naval Research under the supervision of Dr. Woei-Min Lin.

The authors are also grateful to Mr. William Peters of the US Coast Guard for fruitful discussions and encouragement.

9. REFERENCES

Hayter, A. 2012, Probability and Statistics for Scientists and Engineers, 4th edition, Brook/Cole, ISBN 978-1111827045, 826 p.

IMO MSC.1/Circ.1627, "Interim guidelines on the second-generation intact stability criteria", London, December 2020.

IMO SDC 8/WP.4, "Development of Explanatory Notes to The Interim Guidelines on Second Generation Intact Stability Criteria", Report of the Drafting Group on item 6 of addenda, London, January 2022.

Leadbetter, M. R., Rychlik, I. and K. Stambaugh. 2019, Estimating Dynamic Stability Event Probabilities from Simulation and Wave Modeling Methods. Chapter 22 of Contemporary Ideas on Ship Stability. Risk of Capsizing, Belenky, V., Spyrou, K., van Walree F., Neves, M.A.S., and N. Umeda, eds., Springer, ISBN 978-3-030-00514-6, pp. 381-391.

Ross, S. M. 2009, Introduction to probability and statistics for engineers and scientists. 4th ed., Academic Press, ISBN 978-0123704832

Ryan, T. R., 2007 *Modern Engineering Statistics*, Wiley-Interscience, ISBN 978-0470081877, 586 p.

Shigunov, V., 2019, "Direct counting method and its validation", Proc. 17-th Int. Ship Stability Workshop, 10-12 June, Helsinki, Finland, pp. 119-128