Safety in Extreme Environments (2020) 2:79–91
https://doi.org/10.1007/s42797-019-00010-9

ORIGINAL RESEARCH ARTICLE

Comparative statistical analysis of simulated ice management effectiveness

Amy Price[1] · Maria Yulmetova[1] · Sarah Khalil[1]

Received: 28 June 2019 / Revised: 23 October 2019 / Accepted: 28 October 2019 / Published online: 13 December 2019

Abstract

Ice management is critical for safe and efficient operations in ice-covered waters; thus, it is important to understand the impact of the operator's experience in effective ice management performance. This study evaluated the confidence intervals of the mean and probability distributions of two different sample groups, novice cadets and experienced seafarers, to evaluate if there was a difference in effective ice management depending on the operator's level of experience. The ice management effectiveness, in this study, is represented by the "clearing-to-distance ratio" that is the ratio between the area of cleared ice (km^2) and the distance travelled by an ice management vessel (km) to maintain that cleared area. The data analysed in this study was obtained from a recent study conducted by Memorial University's "Safety at Sea" research group. With the distribution fitting analysis providing inconclusive results regarding the normality of the data, the confidence intervals of the dataset means were obtained using both parametric approaches, such as *t*-test, Cox's method, and Johnson *t*-approach, and non-parametric methods, namely Jackknife and Bootstrap methods, to examine if the assumption of normality was valid. The comparison of the obtained confidence interval results demonstrates that the mean efficiency of the cadets is more consistent, while it is more varied among seafarers. The noticeable difference in ice management performance between the cadet and seafarer sample groups is revealed, thus, proving that crew experience positively influences ice management effectiveness.

Keywords Risk management of offshore operations · Confidence interval · Johnson *t*-approach · Cox's method · Bootstrap · Jackknife

Introduction

An ice management system helps forecast and mitigate ice within an operational region to keep personnel safe and reduce operational downtime (Hotzel and Miller 1985). There are many components within an ice management system, such as ice detection, tracking, forecasting, and physical ice management (ISO 2010).

Ice management operations have been going on since the mid-1970s, but there is very little data available to help model the complex system that is ice management, which is required to conduct comprehensive risk assessments (Haimelin et al. 2017). One of the hardest factors to collect data on is the 'human element,' with their varying backgrounds, experiences, educations and training (Haimelin et al. 2017). The analysis conducted in this paper helps quantify some of these hard to define variables, using the data obtained from a marine simulator (Veitch et al. 2018).

This analysis studied the effect of vessel operator experience within the physical ice management component of the complex ice management system. Physical ice management has two main types of operations: towing or changing the course of icebergs and clearing or breaking up of ice (Haimelin et al. 2017). In this study, the data analysed was collected during the clearing of pack ice in an ice management simulation. One of the main reasons for the removal of

✉ Amy Price
ahprice@mun.ca

Maria Yulmetova
mayulmetova@mun.ca

Sarah Khalil
skhalil@mun.ca

[1] Memorial University of Newfoundland, St. John's, Canada

pack ice is to keep the evacuation zone ready for an emergency, allowing space for the lifeboats to be launched (ISO 2010). Clearing of pack ice is also conducted to reduce the global and local loads, and to help with station-keeping most often with floating production storage and offloading platforms (FPSOs) (ISO 2010).

The ice management data used in this analysis was collected during a recent study conducted by Memorial University's "Safety at Sea" research group. Their study investigated the effect of crew experience on ice management operations with the use of a marine ice management simulator (Veitch et al. 2018). The study examined two groups, novice cadets and experienced seafarers, to see if there was a difference in the effectiveness of the ice management operation. After analysing the data from the experiments, 18 usable sample points were collected for the cadets, and 15 sample points were usable for the seafarers (Veitch 2018).

With such small sample-sets, it can be challenging to properly capture the population that the sample-sets are trying to represent. In this investigation, both distributional fitting methods and confidence intervals (CIs) were used to assess if there is a difference in the ice management performance between the novice cadets and experienced seafarers, as well as to examine if the assumption of normality used in the original study was creditable.

When analysing the effect of crew experience on ice management performance, Veitch et al. (2018) assumed that the collected experimental data was normal by only interpreting a normal probability plot. However, this assumption may be inaccurate due to the small sample sizes, since normality tests for small datasets have little power to reject the null hypothesis (Ghasemi and Zahediasl 2012). In most cases, small sample sizes pass normality tests, thereby hinder the detection of non-normality (Helsel and Hirsch 2002). Thus, this study investigated if the assumption of normality is accurate for the ice management performance of the cadets and seafarers.

With small sample-sets, the *t*-test is typically used to find the CIs of the parameters (Johnson 1978), but this approach requires an assumption of normality. When the data seems skewed parametric methods like Cox's (Land 1972; Zhou and Goa 1997) and modified *t*-approaches (Johnson 1978; Banik and Kibria 2010) can be used to find CIs as they take into account some degree of skewness. For many applications and comparisons, the various bootstrap methods seem to be favoured when distributional assumptions do not want to be made regarding a sample-set (Zhou and Goa 1997; Henderson 2005; Perera 2008; Banik and Kibria 2010; Lee et al. 2010; Mamun et al. 2017).

Therefore, both parametric and non-parametric methods were implemented to investigate the CIs of the means and the difference between them, to see if the parametric *t*-test method that requires an assumption of normally distributed data produces similar conclusions to the non-parametric results. For the non-parametric methods, three different types of Bootstrap, standard normal, percentile and Bootstrap *t*-method, were investigated along with Jackknife.

CIs quantify an expected range for a parameter value, such as a mean or median, to fall within (Liu 2009). CIs are also very useful when comparing the difference between two populations. Traditionally, especially in the medical and psychology fields, null hypothesis tests have been used to see if there is a difference in two populations but cannot be used to tell if this difference is significant (Corty and Corty 2011). Therefore, CIs are a suitable approach to compare the two different experience levels of ice management operators. For this analysis, a 95% confidence level was set for the CI range.

One of the focuses of this paper is to investigate if there is a difference between the probability distributions for the ice management effectiveness of the seafarers and cadets sample groups. Experience and expertise are valuable to employers; for example, 81% of U.S. employers pay for and encourage their employees to learn and enhance their skill sets (Germain 2011), but how are experts distinguished from novice employees. Simulators have been used extensively in the medical field for training, as well as for studies investigating the difference in performance between different experience levels (Bick et al. 2013; Conway et al. 2014; Mandava et al. 2015). None of these studies found here have distinguished any type of distributions for either the novice or expert sample groups (Ueda et al. 2010; Bick et al. 2013; Conway et al. 2014; Mandava et al. 2015; Mazomenos et al. 2016; Cahill et al. 2018). One paper that studied expert and novice performs in a specific task that neither had extensive experience in, found that there were a few cases where the students had a higher performance value than the experts, but overall the experts had a lower variance. This paper also noted that a lower variance in outcomes is more valuable in clinical situations where patients are expecting a certain outcome every time (Ueda et al. 2010). This expectation or standard of expecting similar outcomes every time could also be translated to operation in harsh or dangerous environments, like in ice management in the offshore field.

There is some criticism of expert vs novice style experiments since the expert, in theory, has a lot more practice and knows what to expect from the activity (Giskegjerde 2011), in particular like the medical procedure simulations. In this case, the simulation of ice management operations can be considered a dynamic environment, and is considered a better environment to conduct the novice vs expert type of experiments since the exact situation is much harder to predict (Giskegjerde 2011). Dynamic environments change despite the operator's actions and allows examiners to analyse experts as they process and recognize patterns in the emerging environment (Cellier et al. 1997). In dynamic environments researchers have observed that experts observe larger patterns than novices, which allows them to focus on the more relevant aspects of the operation. With practice, experts catalogue a large collection of patterns which allows them to narrow down more critical states, and prevent events opposed to correcting missteps (Cellier et al. 1997).

The analysis in this paper can be broken up into four steps. First, an investigation of the samples' distributions is conducted to assess the distributional assumptions required with parametric CIs methods. Second, the CIs of the sample means are developed with both parametric and non-parametric methods; this is followed by finding the CIs of the difference in the sample means. Finally, the results are analysed by comparing the different methods used to generate the CIs with interval length, assessing the assumption of a normal distribution with small sample-sets, and determining if bridge experience influences ice management effectiveness.

Description of data

The data used in this analysis was collected by Memorial University's "Safety at Sea" research group. The study conducted by Veitch et al. (2018), investigated the effect of crew experience and ice concentration on ice management operations with the use of a marine ice management simulator. The study examined two groups, novice cadets and experienced seafarers, to see if there was a difference in the effectiveness of the ice management operation. The study was conducted with the use of a marine simulator (Fig. 1). The simulator has a simplified bridge, to help reduce the impact of inexperience with operating an ice management type vessel and also helps control the number of variables during a scenario.

Various parameters were collected in the initial study to try and investigate effective ice management, such as average and peak ice concentration drop, total ice clearing area, clearing-to-distance ratio, and cumulative ice-free life-boat launch time (Veitch 2018). For this analysis effective ice management was quantified with the clearing-to-distance ratio, where the ratio is between the area cleared (km^2) and the distance the vessel travelled (km) during the operation. Ice management effectiveness was used to help quantify good and efficient ice management operations. An operator's work was considered effective if they managed to clear a larger area with less distance travelled. Therefore, the larger clearing-to-distance ratios indicate better operator performance (Veitch 2018).

The original study tested operators in two different ice management scenarios, an "emergency" scenario and a "precautionary" scenario, but talking to the operators once they completed the "precautionary" ice management scenarios, it was found that that scenario was not feasible with only one vessel (Veitch et al. 2018). Therefore, the data used in this analysis was collected during the "emergency" ice management simulation, where the operators were asked to keep a lifeboat launch area (seen as the box in Fig. 2) cleared of ice for 30 min, to allow for lifeboat evacuation.

In the original experiment, two factors were investigated: experience level and ice concentration (Veitch et al. 2018). In this analysis, only experience level was analysed. Thus, the data collected for the two different ice concentrations was mixed for this analysis.

Since the focus of the original experiment was to see if there was a difference between cadets and seafarers operating an ice management simulator in two different levels of ice concentration, the original experiment was set up and conducted as a factorial design of experiments to see if any significant effects were detectable. This was also advantageous due to the limited number of local cadets and seafarers, as well as the amount of time each session required. The size of both

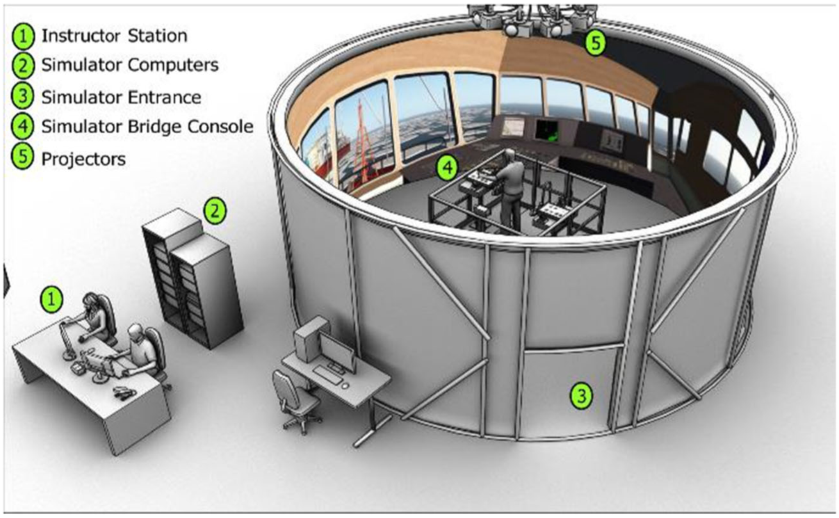

Fig. 1 Marine simulator schematic (modified from Veitch et al. 2018)

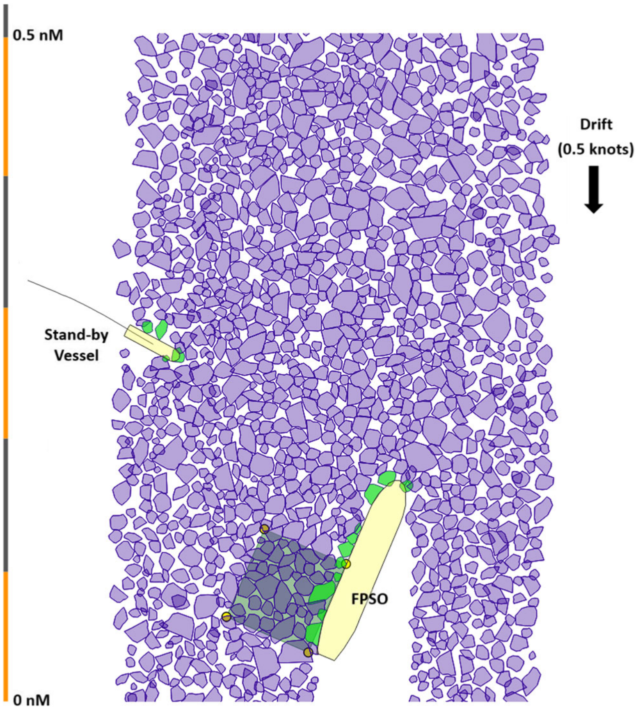

Fig. 2 "Emergency" ice management scenario (modified from Veitch et al. 2018)

sample groups was based on the minimum required number of runs that would allow for significant effects of a 2-factor factorial design to be detectable. This indicated that each sample group should have at least 18 participants (Veitch et al. 2018), which in statistics is a relatively small sample size. Once that data was analysed, the clearing-to-distance ratio had 18 usable sample points for the cadets and 15 usable sample points for the seafarers (Veitch 2018).

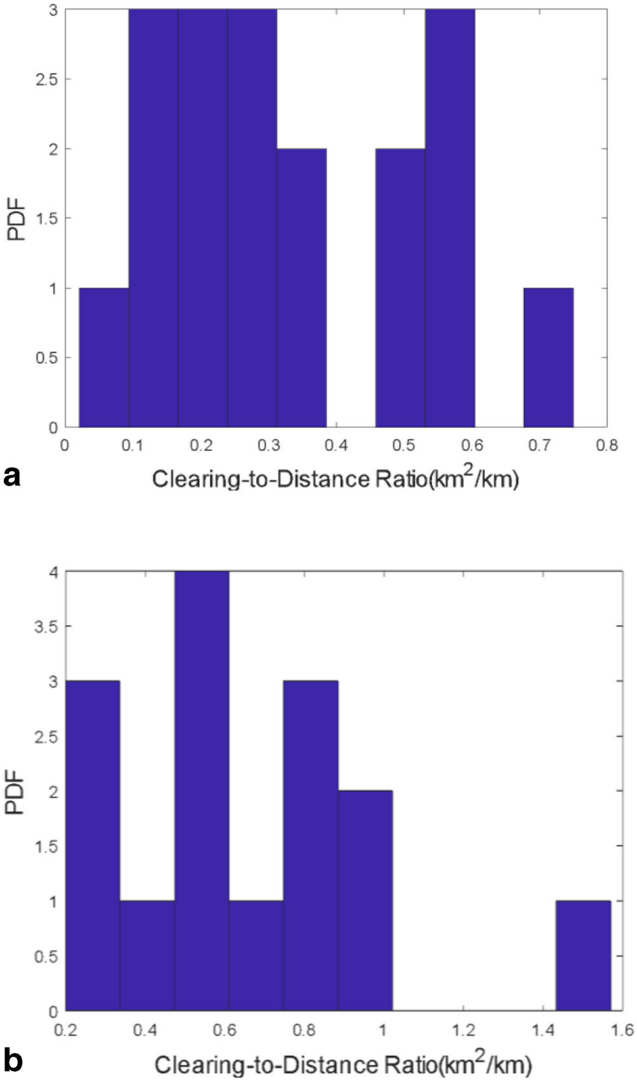

Fig. 3 a) Cadet and b) seafarer ice management effectiveness probability density

Methods

Determination of probability distribution model

Since these data sets are unique and nothing similar appears to have been collected and statistically analysed before, there is no known expected probability distribution. Thus, exploratory data analysis was completed with probability distribution fitting methods to see if the assumption of a normal distribution is acceptable.

Due to the small sample-set sizes, the probability density plots (Fig. 3) did not help give any general idea of what distributions may be present. A normal distribution was investigated since that was the assumption of the original analysis (Veitch et al. 2018) and is now being analysed and compared in this study. Additionally, it is common to assume normality with small sample-sets to allow the use of the *t*-test. A lognormal distribution was also chosen due to the possible unevenness of the probability density plots and the common occurrence of lognormal distribution in nature (Casella and Berger 2002; Banik and Kibria 2010). Therefore, normal and lognormal distributions were selected to be investigated.

First, Box-Cox transformations were generated to see if a sample-set may require a transformation to approach a normal distribution (Montgomery 2013). In other words, the Box-Cox transformations were applied to see if the initial assumptions of either a normal or lognormal distribution were sufficient, or if the sample-sets required a different transformation that should also be analysed.

From there, the chosen distributions were examined with visual (cumulative probability distribution and probability plots) and formal (Kolmogorov-Smirnov goodness of fit test) methods.

One visual method compares the observed and theoretical cumulative density function (CDF) of the sample data to see what theoretical distribution follows the observed CDF plotted data the best (Ang and Tang 2007). Theoretical distribution, in this case, is the distribution created using a cumulative density function and parameters from the sample-sets, such as the mean and standard deviation. The theoretical distribution is compared to the observed cumulative density where the collected data is plotted.

The other visual method was probability plots, which see if a set of random or observed variables follows a specific CDF. First, the observed CDF is plotted, then the linearized theoretical CDF of a particular distribution is overlaid. If the observed data points follow the linear predicted probability function, this suggests that the sample-set follows the distribution under investigation (Haldar and Mahadevan 2000). Some generated probability plot also included CIs to help gauge how close the observed data follows the linearized probability function.

The final distribution fitting method was the analytical approach of the goodness of fit tests. The goodness of fit tests are simple null hypothesis tests and only indicate if a chosen distribution under investigation is not rejected. The Kolmogorov-Smirnov goodness of fit test (K-S test) was selected over the Chi-Square test due to the small sample size preferring a CDF comparison over a probability density function comparison (Ang and Tang 2007). Since the overall distribution was of more concern, then the tail ends of the distribution, the K-S test was also chosen over the Anderson-Darling test (A-D test) (Ang and Tang 2007). The K-S test uses the CDF to compare an observed sample-set to a theoretical distribution using the sample's parameters. The maximum difference between the observed and theoretical distribution, D_n, is calculated, then compared to the critical difference, D_n^{α}, found for a given significance level, α, and the number of sample points, n. If $D_n \leq D_n^{\alpha}$, then the null hypothesis is not rejected (Haldar and Mahadevan 2000).

Parametric methods

If the distributional goodness of fit tests do not reject the assumption of normality, the parametric *t*-distribution test, also known as the *t*-test, is implemented to find the CIs for the sample means as well as the CI for their difference. The *t*-test is used when the sample size is small ($n < 30$), such as this case (Navidi 2006).

In the case where the data appears to have some degree of skewness, the Johnson *t*-approach is applied. This approach still uses the *t*-distribution, but the *t* variable is modified using properties from the sample to make it more robust and remove bias (Johnson 1978). The Cornish-Fisher expansion is used to derive the correction factor as it allows the correction of bias and skewness effects, by correcting the difference between the mean and median that arrives with asymmetrical distributions (Johnson 1978); the CI relationship that is derived can be seen as Eq. 1.

$$\left[x + \left(\frac{\hat{\mu}_3}{6s^2N}\right)\right] \pm t_{\alpha/2,v}\frac{s}{\sqrt{N}} \tag{1}$$

The Johnson *t*-approach uses the sample mean $\overline{x}$, the third moment of the population $\hat{\mu}_3$, the variance s^2, the sample size N, and the *t*-value $t_{\alpha/2,\ v}$. This approach is for sample sizes as small as 13, and appropriate for a sample with distributions ranging from a normal *t*-distribution to samples with asymmetrical Chi-square distribution degree of skewness (Johnson 1978).

In cases where samples have a higher degree of skewness that approach a lognormal distribution, Cox's method can be used (Land 1972; Zhou and Gao 1997). This method is an effective approach to approximate the CI of a lognormal mean. The Cox's method, even though suggested for larger sample sizes, seems to reasonably meet the nominal confidence level for small samples as long as the variance is low (Land 1972). The Cox's method finds the CI in its lognormal form using Eq. 2.

$$\theta = \overline{Y} + \frac{s^2}{2} \pm Z_{1-\alpha/2}\sqrt{\frac{s^2}{n} + \frac{s^4}{2(n-1)}} \tag{2}$$

Where $\overline{Y}$ is the sample mean of the log transformed sample, in this case, s^2 is also the variance of the log transformed sample, and $Z_{1-\alpha/2}$ is the z-value for the chosen confidence level (Land 1972).

Non-parametric methods

If the sample sizes are too small to make a confident assumption of the data's distributions, the non-parametric methods are beneficial for computing the uncertainties of the data means. Both the Bootstrap and the Jackknife methods are computer-intensive resampling methods that allow the estimation of the statistical parameters and their uncertainties when the sample data is small, without making distributional assumptions about the data. The main difference of these techniques is the sampling type, which is conducted with replacement in the Bootstrap and without replacement in the Jackknife (Chernick 1999; Henderson 2005).

The Bootstrap procedure starts by creating many samples artificially from an original dataset by applying multiple resamplings with replacement. From there, statistical conclusions are drawn from the obtained resampled data.

The general Bootstrap algorithm for computing confidence interval for the mean is followed. First, m Bootstrap samples are created of size n with replacement from the original data sample $X_{1,}$ $X_{2,}$ ..., X_n. Next, the estimated parameters for each Bootstrap sample are computed, such as the means $\overline{Q}_1^*, \ldots, \overline{Q}_m^*$ or standard errors $se_1^*, \ldots, se_m^*$. The Bootstrap mean $\overline{Q}$, the mean of the means for all Bootstrap samples, is also calculated. Finally, the CIs of the Bootstrap means are determined (Efron 1979).

There are various methods for computing CI for the mean. In this study, three Bootstrap methods, standard normal

(NORM), percentile (PER) and Bootstrap *t*-method (STUD), are analysed with a 95% CI probability (Betta et al. 2006; Stepien 2016; Liguori et al. 2017).

The NORM method computes a CI for the mean as:

$$\overline{Q} \pm z_{\alpha/2} * SE \tag{3}$$

where $z_{\alpha/2}$ is the point where the standard unit normal distribution is exceeded with an α/2 probability, and *SE* is the standard error (Betta et al. 2006).

The PER method uses the percentiles of the Bootstrap means, so that the lower and upper limits of the CI are $\overline{Q}^*_{\alpha/2}$ and $\overline{Q}^*_{1-\alpha/2}$, the quantiles of the Bootstrap means distribution. For example, the endpoints of the 95% CI are the 2.5, and 97.5 percentiles of the Bootstrap sample means.

The STUD method constructs the CI using a Student's *t*-statistic. First, the *t*-statistic $t_1^*, \ldots, t_m^*$ are calculated for each Bootstrap sample as:

$$\dot{t}_1^* = \left(\overline{Q}_i^* - \overline{Q}\right) / se_i^* \tag{4}$$

where $\overline{Q}_i^*$ and se_i^* are the mean and standard error of i^{th} Bootstrap sample. Then, the CI is defined by:

$$\overline{Q} \pm \dot{t}^*_{\alpha/2} * SE \tag{5}$$

The Bootstrap method is also used to compute the CI of the difference between two sample means, to define how much the means of two data samples differ. The bootstrap algorithm for computing the CI of two means is similar to that described above. First, *m* bootstrap samples of size *n* were created for both data samples. Then, *m* bootstrap samples of their difference are computed, and the means for each of them are obtained. Finally, the CIs are extracted using the percentile method.

Similar to the Bootstrap method, the Jackknife approach is applied to estimate the statistical parameters of the random variables. The technique is mainly conducted to estimate the uncertainties associated with the statistical parameter of interest. The Jackknife method may be perceived as less effective than the Bootstrap method (Efron and Gong 1983); however, both Jackknife and Bootstrap are known to perform adequately when the distribution is undecided or when the normality assumption is invalid. Jackknife is considered a special case of the Bootstrap method but has the advantage of requiring fewer calculations (Efron and Gong 1983).

The Jackknife technique starts by resampling without replacement, that is, by leaving out a fixed number of observations from the original sample each time to obtain a new sample-set. In this case, one different data point is removed every time to obtain a new sample. Thus, 18 and 15 new samples with an *n-1* sample size were obtained for the cadet and seafarer groups, respectively, where *n* is the original sample size. Next, the means of the new Jackknife samples $\overline{x_{(j)}}$ are calculated, followed by calculating the overall mean of these samples' means, $\overline{x_{(*)}}$, Eq. 6.

$$x_{(*)} = \frac{\sum\left(x_{(j)}\right)}{n} \tag{6}$$

Next, the pseudo-values, which are defined as the differences between the original sample mean and the new sample means, are calculated using the Eq. 7.

$$x^*_{(j)} = nx - (n-1)x_{(j)} \tag{7}$$

The pseudo mean $\overline{x}^*$ is then obtained for the datasets; this value is called the Jackknife estimate, Eq. 8.

$$x^* = \sum \frac{\left(x^*_{(j)}\right)}{n} \tag{8}$$

This calculation is then followed by Eq. 9, the Jackknife variance, V-jack, which is the variance of the mean.

$$Vjack = \frac{1}{n(n-1)} \sum \left(x^*_{(j)} - x^*\right)^2 \tag{9}$$

Finally, a 95% CI for the sample-set mean is calculated with Eq. 10.

$$x^* \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{Vjack}{n}} \tag{10}$$

Results

Probability distribution model

First, Box-Cox transformations were created to check the original assumption of the sample-sets likely having either a normal or lognormal distribution. From the Box-Cox transformation, the best-rounded lambdas were found to be 0.5 and 0.0 for the cadets and seafarers, respectively. These lambda values suggested that a square root and log transformation were the most appropriate transformation for these sample-sets. The results of the Box-Cox transformation were similar to the initial assumptions of the sample-sets having either a normal or lognormal distribution. Therefore, the normal and lognormal distributions were investigated along with a chi-square distribution.

Next, the CDF plots were examined, looking for similarities in the observed and theoretical curves of the distribution functions. Analysing Fig. 4, it can be observed that both the cadet and seafarer sample-sets do not favour either the normal, lognormal, or chi-squared distributions.

Then, the normal probability plots for both the original, log and square root transformed data were created (Fig. 5). It may be observed in most cases both original and transformed data follow the straight line reasonably well, only having a few outliers at the tail ends of the plots. These outliers were checked with a Grubb's test, but can also be observed on the probability plots as the data points outside of the 95% CIs, that are shown as (the curved red lines). Examining the normal probability plots (Fig. 5a), only one obvious outlier was found on the seafarer plot. The probability plots of log-transformed data (Fig. 5b) only have one outlier as well, but, in this case, with the cadet sample-set. Finally, the square root transformed data probability plots (Fig. 5c) have no outliers. A single outlier is not enough to reject a distribution, but in cases like these where all the distributions under investigation are producing similar results it is a means to distinguish the better distribution; in this case the Chi-Squared distribution stands out.

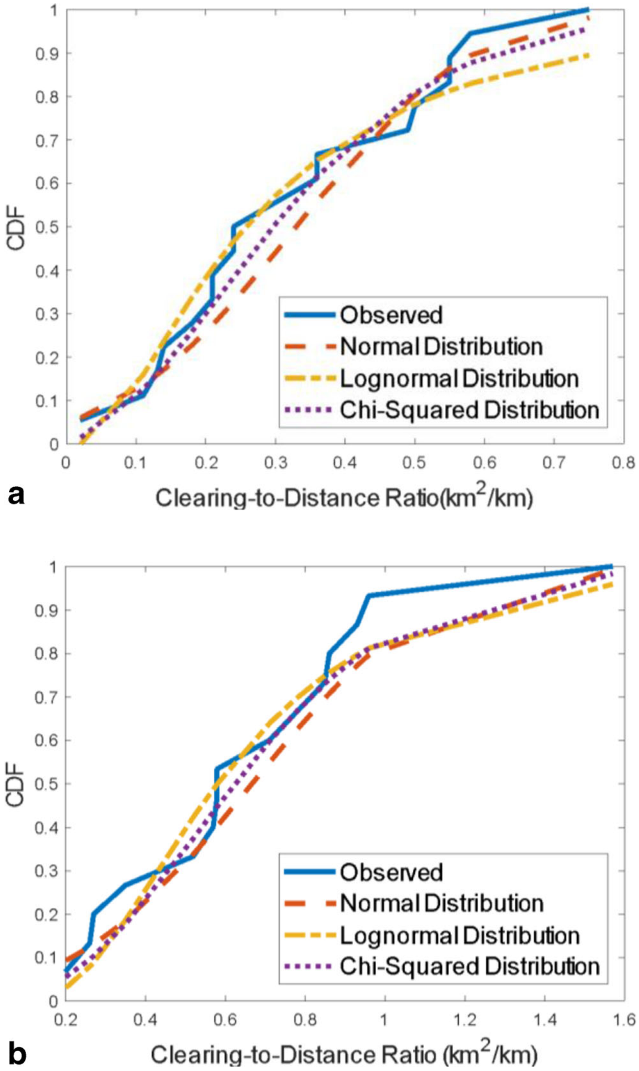

Fig. 4 a) Cadet and b) seafarer ice management effectiveness, observed and theoretical CDF

Thus, considering only the probability plots, the normal, lognormal, and chi-squared distributions are probable for both sample-sets. The *p* values found with the probability plots from the Anderson-Darling (AD) values in Table 1, for the normal and lognormal distributions are observed to have similar significances. However, for the chi-squared distribution, much larger p values are evident for both sample-sets, which agrees with the probability plots containing no outliers.

Finally, the K-S test compared the maximum difference of the three distributions, against the critical values, Table 2. For all cases, $D_n \leq D_n^{\alpha}$, indicating that none of the null hypotheses were rejected for critical values (α) ranging from 0.01 to 0.20. Thus, all three distributions are possible for the two sample-sets.

Even though the normal distribution was not rejected, neither were the lognormal or chi-squared distributions. Therefore, the assumption of normality may not be appropriate due to the effect of the small sample sizes. Additionally, the Box-Cox transformation suggested that a transformation is required for both sample-sets. Thus, from the exploratory data analysis, the most appropriate distributions for the two sample-sets only has a slight preference towards a chi-squared distribution. Hence, one should be careful with parametric tests when used for computation of CIs, as the results of the analysis depend on the chosen distributions. Therefore, using non-parametric methods such as Bootstrap and Jackknife is advantageous in cases like these, where the population distribution is unknown, and the samples are too small to distinguish an obvious distribution from distribution fitting analyses.

Parametric and non-parametric confidence interval estimation of the means

The assumption of normality was not rejected with the distribution fitting methods, but nor were the lognormal and the chi-squared distribution; thus, a confident assumption of normality cannot be granted. Nonetheless, the *t*-test was still conducted to allow for a comparison to the other parametric methods along with the Bootstrap and Jackknife resampling methods. The *t*-test was also still evaluated to allow for an assessment of the results when a small sample-set may falsely indicate normality of the data, against methods were distribution assumptions are not required. The CIs calculated with the *t*-test are shown in Tables 3, 4, and 5, and discussed in the following sections.

Similarly, since the Chi-Square and lognormal distributions were also not rejected, the CIs using the Cox's method and Johnson *t*-approach were determined and are presented in Tables 3 and 4, as well.

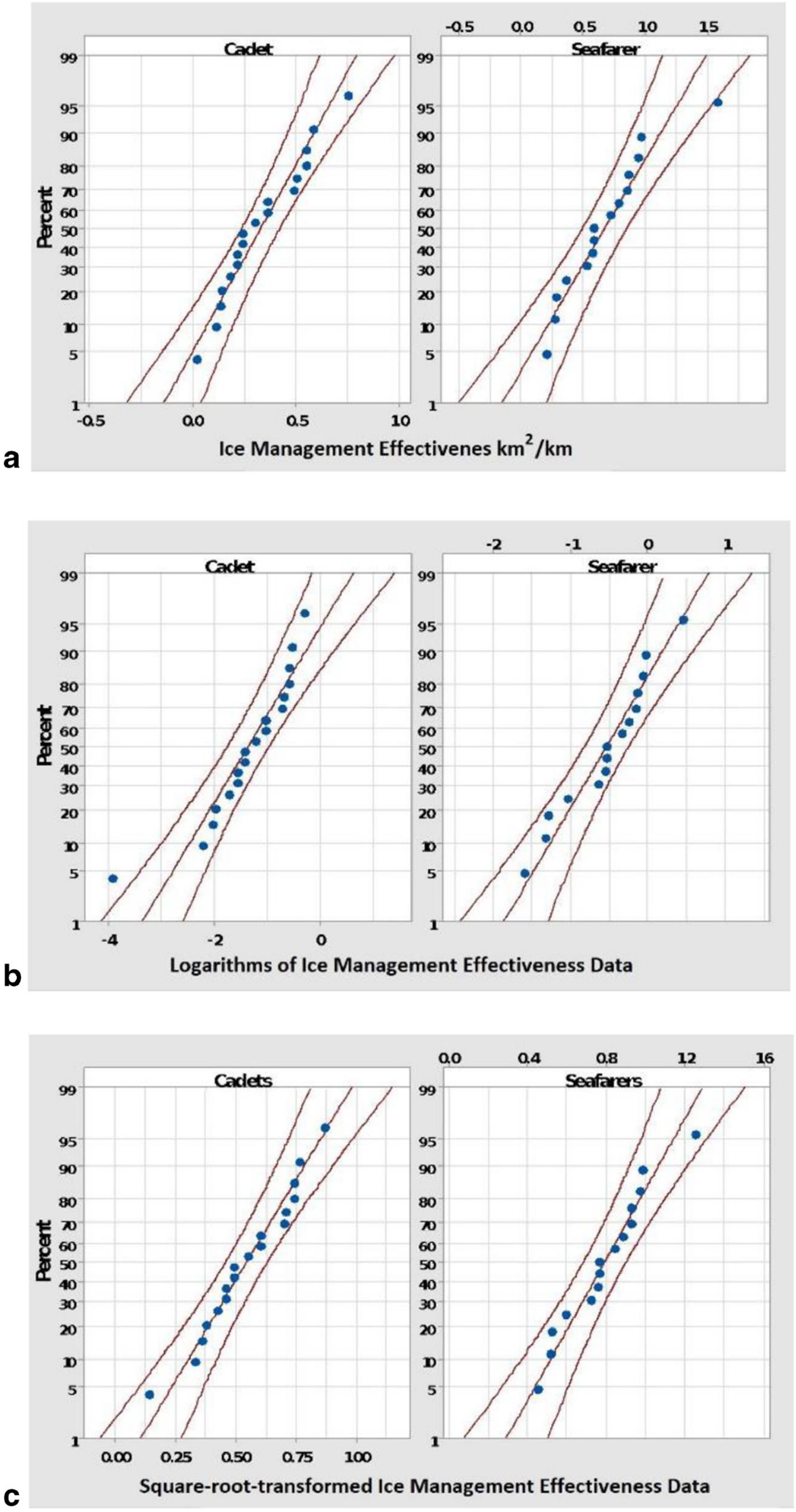

Fig. 5 Probability plots of a) original, b) log-transformed and c) square-root-transformed data of cadet and seafarer ice management effectiveness

Before computing the CIs using the Bootstrap methods, the minimum number of Bootstrap replications, *m,* was defined, typically the required number of samples range from 1000 to 2000 samples (Banik and Kibria 2010). The specific number of Bootstrap samples depends on the characteristics of the data, such as the original sample size, variance, estimated

Table 1 Probability plot distributions AD‘s *P*-values

Distribution	P-Value		AD Value	
	Cadet	Seafarer	Cadet	Seafarer
Normal	0.274	0.360	0.431	0.379
Lognormal	0.301	0.316	0.601	0.401
Chi-Squared	0.689	0.587	0.254	0.282

parameters, etc. Therefore, to find the optimal *m,* the line graph of the Bootstrap samples was inspected. As shown in Fig. 6, the variation of the Bootstrap means appears fairly regular, starting from approximately 150 samples. However, the CI values converge after roughly 1000 sample replications. Therefore, 1000 sample replications were used in this analysis. It can further be observed that the Bootstrap *t*-method required the highest number of replications, thus governed the minimum number of sample replications required.

All the obtained results of the CIs for the cadet and seafarer dataset means are shown in Table 3, while the CI widths are displayed in Table 4.

The CIs obtained using the different parametric, and non-parametric methods are plotted in Fig. 7. Examining the cadet CIs, the Cox’s method produced the widest interval, while the NORM Bootstrap technique obtained the narrowest CI. Then, when analysing the seafarer CIs the NORM Bootstrap method was observed to have the widest interval, while the rest of the CIs are relatively similar, with the PER Bootstrap method producing the narrowest interval. With that being said, all seven methods implemented to calculate the CIs demonstrate that the seafarer sample-set has wider CIs than the cadet sample-set.

The widest cadet dataset CI for the mean has a width of 0.252 km^2/km, and the narrowest seafarer dataset CI for the mean has a width of 0.342 km^2/km. As well, the CIs of the means for the seafarer sample-sets can be observed to be shifted towards higher ratios than the cadets. Both the difference in the CIs widths and the noticeable difference in the values of the CIs of the mean indicate the apparent difference between the ice management performance of the two sample groups.

A CI for the difference in the sample set means was calculated to define how much they differ. The CI for the difference in means extends from 0.150 to 0.539 km^2/km and from 0.124 to 0.551 km^2/km based on the results obtained from the PER Bootstrap and *t*-test methods, respectively (Table 5). Put another way, the difference in the means of the effectiveness between seafarer and cadet sample-sets using the boostrap method is 0.344 km^2/km and estimated within the accuracy of ±0.195 km^2/km. It should be noted though, that for a proper analysis for the CI of two sample means, both sample-sets should be the same size, which is not the case with these sample-sets. Therefore, the results in Table 5 can only be taken as an estimation of the CI of the difference in the sample means.

Table 2 Kolmogorov-Smirnov goodness of fit test results

	Cadet *n* = 18	Seafarer *n* = 15
D_n^α Critical Value (for $0.01 \leq \alpha \leq 0.2$)	0.373 to 0.245	0.404 to 0.266
D_n Normal Distribution	0.1715	0.1357
D_n Lognormal Distribution	0.1137	0.1206
D_n Chi-Squared Distribution	0.1144	0.1207

Discussion

With only 18 sample points for the cadet group and 15 sample points for the seafarer group, it is difficult to notice any distinct distribution. Even though the normal distribution was not rejected for the two sample-sets, but neither were the lognormal distribution or the chi-squared distribution. Therefore, stating that normal distribution is incorrect is not necessarily true. This statement, which is likely made from the other distributions also not being rejected, might be wrong due to sample sizes causing the difficulties in rejecting the null hypotheses. Thus, the data may be normal or may not be, but one cannot be sure that it is normal based only on a single normal probability plot test. That is why one must be careful when assuming the normality of small datasets such as this case. Thus, due to the indecision of the probability distribution fitting methods, this restricts the practicality of parametric methods. Further experiments need to be conducted to decide on one distribution, or to confirm if the sample-sets are a combination of several distributions, to allow for proper use of parametric methods.

With the use of the Box-Cox transformation, it was found that the ice management effectiveness of seafarers may be described by a lognormal distribution, while the cadets were better represented by a chi-squared distribution. Therefore, a chi-squared distribution was added to the list of distributions being analysed. The analysis found that chi-squared distribution is favoured for both sample-sets when comparing the probability plots AD’s *p*-values, but when comparing K-S test values the lognormal values are slightly lower for both groups, suggesting a preference towards the lognormal distribution. Therefore, the Box-Cox transformation recommendations of a lognormal distribution and chi-squared distribution were reasonably close to what was found preferred in the exploratory data analysis.

Since the parametric approaches require an idea of an expected distribution or an assumption; therefore, non-parametric methods appear to be more appropriate in this case. However, since the distributional assessments failed to reject

Table 3 CIs of the mean

Method		Cadets Dataset (km^2/km)	Seafarers Dataset (km^2/km)
Bootstrap	PER	0.2414–0.4300	0.5163–0.8580
	NORM	0.2806–0.4150	0.4838–0.9688
	STUD	0.2296–0.4405	0.4891–0.8880
Jackknife		0.2293–0.4582	0.4706–0.8614
Student's *t*-Distribution		0.2294–0.4284	0.4706–0.8614
Cox's Method		0.1997–0.4516	0.4639–0.8326
Johnson *t*-approach		0.2302–0.4292	0.4741–0.8649

the null hypotheses, all three of the parametric methods used to find the CIs were compared to the results obtained from the Bootstrap and Jackknife methods.

Therefore, with the absence of an explicit distribution assumption, Bootstrap and Jackknife were used to evaluate the uncertainties of the means of the data associated with the ice management effectiveness. Bootstrap and Jackknife analysis show that the CI width (Table 4) of the cadet sample-set is less than the seafarer sample-set. Thus, it may be concluded that the mean efficiency of the cadets is more consistent, while it is more varied among seafarers. This counters one article where it saw the experienced personal tended to have smaller CI widths since there was less variance in their performances (Ueda et al. 2010). This wider CI width for the seafarers' ice management effectiveness could be explained by the cadets all having minimal exposure to ice management, while the seafarers had various levels of experience in sea ice, fluctuating from less than three seasons to more than ten seasons (Veitch 2018).

When just the CI widths are compared, it was found that with both sample-sets the PER Bootstrap method produced narrower intervals, compared to the other extreme where the STUD Bootstrap method tended to produce wider intervals. This could be interpreted that the STUD Bootstrap method is more conservative than the PER Bootstrap method. The *t*-test and the Johnson *t*-approach also have relatively consistent interval widths compared to each other, falling in the middle of the interval widths both times.

All the methods applied agree that CIs of the seafarer sample-set are broader than those of the cadet sample-set (Table 4). The CI for the cadet dataset mean can vary from 0.200 to 0.458 km^2/km depending on the method used, while the seafarer dataset CI of the mean can vary anywheres between 0.464 and 0.969 km^2/km. It shows the significant difference between the ice management performance of these two groups. The CI for the difference between two means results confirms this fact as well (Table 5); therefore, proving the statement that crew experience influences ice management effectiveness.

When the assumed normal distribution parametric *t*-test results are compared to those of the skewed parametric methods and the non-parametric resampling methods, similar values can be observed. An interesting observation between the *t*-test and the Johnson *t*-approach can be made. Both methods produced very similar CIs for both sample sets. As mentioned previously, the Johnson *t*-approach approaches a *t* distribution as the sample approaches a symmetrical distribution; these similarities in both the *t*-test and the Johnson *t*-approach could also suggest that the samples are indeed normally distributed. This similarity of the results from the different tests is not unexpected since the normal distribution was not rejected in the exploratory data analysis, but from the results of the distribution fitting methods, a confident assumption of a normal distribution could not be made. Thus, for the investigation of small sample-sets with unknown distributions, the uncertainty of deciding on distribution for parametric analysis should be replaced with non-parametric analysis methods such as Bootstrap and Jackknife.

Future work may include:

- The collection of more data would help decide on the most appropriate distribution.
- The ice concentration factor may be considered by splitting the results into four categories to be further compared

Table 4 Width of CIs of the mean

Method		Cadets Dataset (km^2/km)	Seafarers Dataset (km^2/km)
Bootstrap	PER	0.1886	0.3417
	NORM	0.1344	0.4850
	STUD	0.2109	0.3989
Jackknife		0.2289	0.3908
Student's *t*-Distribution		0.1990	0.3908
Cox's Method		0.2519	0.3687
Johnson *t*-approach		0.1990	0.3908

Table 5 CI of the difference in the means

Method	Difference in the Data Set Means (km^2/km)
Bootstrap Percentile Method	0.1496–0.5389
Student's *t*-Distribution	0.1237–0.5505

and examined for normality or appropriate distributions, but this may be more difficult to investigate due to the smaller sample sizes.

- It is observed that both data sets distributions are skewed to some degree and influenced by a few outliers in the sample-sets. Since the sample means are affected by data skewness and highly sensitive to outliers, an analysis of the CIs of the medians, as a more resilient measure of location (Helsel and Hirsch 2002), may be more appropriate for these datasets.

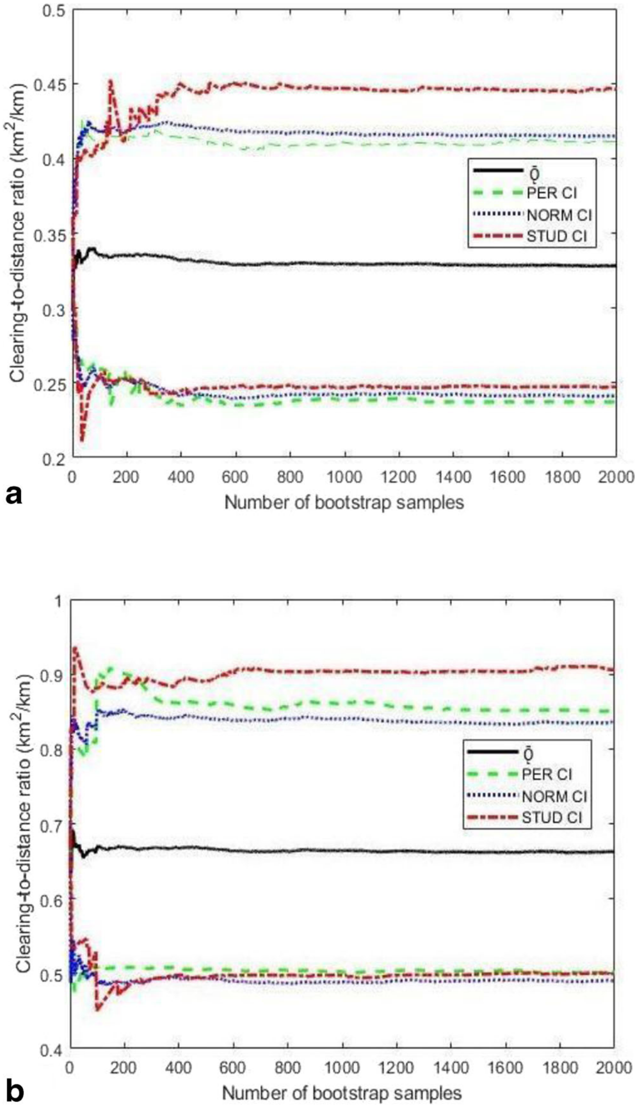

Fig. 6 The a) cadet and b) seafarers datasets Bootstrap means $\overline{Q}$ and their CIs, obtained by PER, NORM and STUD methods, depending on the number of Bootstrap replications

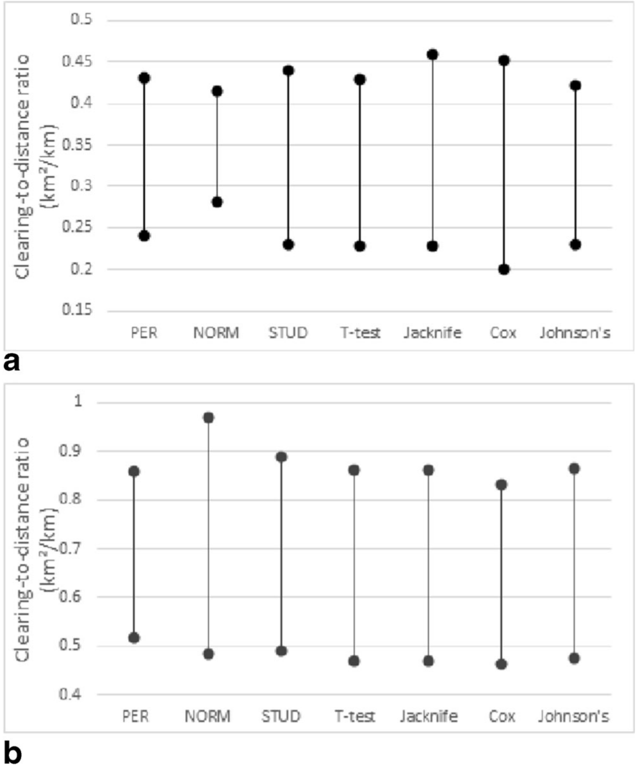

Fig. 7 CIs for the a) cadets and b) seafarers means obtained by Bootstrap (PER, NORM, STUD), *t*-test, Jackknife, Cox's method and Johnson *t*-approach

Conclusion

In this study, the ice management effectiveness data obtained from a marine simulator experiment was evaluated. The assumption of normality used by Veitch et al. (2018) was investigated by conducting exploratory data analysis, along with a comparative statistical analysis of the ice management effectiveness data. The statistical analysis was performed through an assessment of the CIs of the means, with the average length, obtained from the *t*-test, Cox's method, Johnson *t*-approach, Bootstrap methods and Jackknife analyses.

There is not much data available on the CIs of novice and experienced personnel in dynamic environments, which makes it impossible to compare the results to other relevant studies. Nonetheless, based on the study's results, the following conclusions were made. First, since the null hypotheses, for all three distribution fitting methods, for the distributions under investigation were not rejected, the assumption of a normal distribution of the datasets is inconclusive. Nevertheless, both parametric and non-parametric tests were applied to find CIs of the sample means. Both parametric and non-parametric tests produced similar results suggesting that the initial assumption of normality may be valid. Moreover, the CIs indicate that there is less variation of cadet efficiency in comparison with the seafarer group. This reduction in disparity may be explained by the limited exposure of cadets' to ice management, while the seafarers' experience in sea ice varied significantly.

Finally, based on the analysis of the CI of the difference between the two sample-set means, the conclusion of crew experience positively influencing the ice management performance is confirmed.

Acknowledgements We would like to thank and acknowledge Erik Veitch for allowing us to use and further extend the analysis completed by him.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ang A, Tang W (2007) Determination of Probability Distribution Models. In: Probability Concepts in Engineering, John Wiley & Sons, Inc, Hoboken, pp 278-301

Banik S, Kibria G (2010) Comparison of some parametric and nonparametric type one sample confidence intervals for estimating the mean of a positively skewed distribution. Communication in Statistics-Simulation and Computation 39:361–389

Betta G, Capriglione D, Pietrosanto A, Sommella P (2006) A Reliable and Robust Methodology for Testing Measurement Software. IEEE Instrumentation and Measurement Technology Conference Proceedings 2101–2106. doi:https://doi.org/10.1109/IMTC.2006.328465

Bick J, DeMaria S, Kennedy J et al (2013) Comparison of expert and novice performance of a simulated transesophageal echocardiography examination. Simul Healthc 8:329–334

Cahill P, Samdani A, Brusalis C et al (2018) Youth and experience: the effect of surgeon experience on outcomes in cerebral palsy scoliosis surgery. Spine Deformity 6:54–59

Casella G, Berger RL (2002) Common families of distributions. In: Statistical inference, Second Edition, Duxbury, Pacific Groove, pp 85–138

Cellier J-M, Eyrolle H, Marine C (1997) Expertise in dynamic environments. Ergonomics 40:28–50

Chernick MR (1999) Bootstrap methods: a practitioner's guide. John Wiley & Sons, Inc., California

Conway N, Romanelli J, Bush R, Seymour N (2014) Ramifications of single-port laparoscopic surgery: measuring differences in task performance using simulation. Surg Innov 21:106–111

Corty E, Corty R (2011) Setting sample size to ensure narrow confidence intervals for precise estimation of population values. Nurs Res 60: 148–153

Efron B (1979) Bootstrap methods: another look at the Jackknife. Ann Stat 7:1–26

Efron B, Gong G (1983) A leisurely look at the bootstrap, Jackknife and cross validation. Am Stat 37:36–48

Germain M (2011) A chronological synopsis of the dimensions of expertise: toward the expert of the future. Perform Improv 50:38–46

Ghasemi A, Zahediasl S (2012) Normality tests for statistical analysis: a guide for non-statisticians. International Journal of Endocrinology and Metabolism 10:486–489

Giskegjerde G (2011) Expertise or Safety Climate? Approaching Human Factors in Demanding Martime Operations. Master of Philosophy in Psychology Thesis, University of Oslo

Haimelin R, Goerlandt F, Kujala P, Veitch B (2017) Implication of novel risk perspective for ice management operations. Cold Region Science and Technology 133:82–93

Haldar A, Mahadevan S (2000) Determination of distributions and parameters from observed data. In: Probability, Reliability, and Statistical Methods in Engineering Design, John Wiley & Sons Inc., New York, pp 112–117

Mandava SH, Liu J, Maddox M et al (2015) Stratification of expert vs novice Laparoscopists using the basic laparoscopic urological surgery (BLUS) curriculum at a single institution. Journal of Surgical Education 72:964-968

Helsel DR, Hirsch RM (2002) Statistical methods in water resources. U.S. Geological Survey. https://pubs.usgs.gov/twri/twri4a3/pdf/twri4a3-new.pdf. Accessed 12 March 2019

Henderson AR (2005) The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. Clin Chim Acta 359:1–26

Hotzel IS, Miller JD (1985) Ice Management on the Grand Banks. OCEANS '85 - Ocean Engineering and the Environment 416–419. doi:https://doi.org/10.1109/OCEANS.1985.1160119

International Organization of Standards (ISO) (2010) Petroleum and natural gas industries —Arctic offshore structures. International Organization of Standards (ISO), Geneva

Johnson N (1978) Modified t tests and confidence intervals for asymmetrical populations. J Am Stat Assoc 73:536–544

Land C (1972) An evaluation of approximated confidence interval estimation methods for lognormal means. American Society for Quality 14:145–158

Lee S, Bolic M, Groza V, Dajani H, Rajan S (2010) Confidence intervals estimation for blood pressure measurements with nonparametric bootstrap approach. 2010 IEEE International Workshop on Medical Measurements and Applications 130-133. https://doi.org/10.1109/MEMEA.2010.5480216

Liguori C, Ruggiero A, Sommella P, Russo D (2017) Choosing bootstrap methods for the estimation of the uncertainty of traffic noise measurements. IEEE Transaction on Instrumentation and Measurement 66:869–878

Liu XS (2009) Sample size and the width of the confidence interval for mean difference. The British Journal Of Mathematical And Statistical Psychology 62:201–215

Mamun A, Hussin A, Zubairi Y, Imon R, Rana S (2017) Small-sample confidence interval for the slope of linear structural relationship model. Electronic Journal of Applied Statistical Analysis 10:374–383

Mazomenos E, Chang P, Rippel R et al (2016) Catheter manipulation analysis for objective performance and technical skills assessment in transcatheter aortic valve implantation. Int J CARS 11:1121–1131

Montgomery DC (2013) 15.1 Nonnormal Responses and Transformations. In: Design and analysis of experiments, 8th edn, John Wiley & Sons, Inc, Hoboken, pp 643–652

Navidi W (2006) Statistics for engineering and scientists. McGraw-Hill, New York

Perera S (2008) Normal theory and bootstrap confidence interval estimation in assessing diagnostic performance gain when combining two diagnostic tests. Communication in Statistics-Simulation and Computation 37:2076–2088

Stepien B (2016) Bootstrap confidence intervals for noise indicators. Acustica United with Acustica 102:389–397

Ueda M, Mine A, Munck J, Hakogi T, Vann Meerbeek B (2010) The effect of clinical experience on dentine bonding effectiveness: student versus trained dentists 37:653–657

Veitch E (2018) Influence of bridge officer experience on ice management effectiveness. Ocean Engineering Research Centre, St. John's

Veitch E, Molyneux D, Smith J, Veitch B (2018) Investigating the influence of bridge officer experience on ice management effectiveness using a Marine simulator experiment. Journal of Offshore Mechanical and Arctic Engineering 141:1-12. doi:https://doi.org/10.1115/1.4041761

Zhou XH, Goa S (1997) Confidence intervals for the log-Normal mean. Stat Med 16:783–790