Universitetet
i Stavanger

DET TEKNISK-NATURVITENSKAPELIGE FAKULTET

Institutt for industriell økonomi, risikostyring og planlegging

MASTEROPPGAVE

Studieprogram/spesialisering: Risikostyring - Master i teknologi/siv.ing. offshore sikkerhet.	Vår semesteret, 2011. Åpen
Forfatter: Trine E. Haaland	…………………………………………… (signatur forfatter)

Fagansvarlig: Jan Erik Vinnem

Veileder(e): Bjørnar Heide

Title of thesis: Develop a method to compare the barrier performance for the operator companies on Norwegian Continental Shelf and use this method to investigate if there is a statistically significant difference between the companys' barrier performances.

Tittel på masteroppgaven: Lage en metode for å sammenligne barrieytelse for operatørselskapene på norsk kontinentalsokkel og bruke mehoden for å undersøke om det er en statistisk signifikant forskjell mellom operatørselskapenes barriereytelse.

Studiepoeng: 30 SP

Emneord: Method development Barrier performance Failure rates Classical Statistics Predictive Bayesian approach RNNP Regression analysis ANOVA	Sidetall: 78 + vedlegg/annet: 19 Stavanger, 29/06/2011

Contents

Preface

After 3 years of working and studying part time, this master degree finally comes to an end with this thesis. This thesis option was presented to me at a meeting with Safetec Nordic in the middle of the autumn semester. I chose this option because it was something I could relate to with my experience working offshore where our safety is very much depending on sound and functional safety barriers, and also because it is interesting to be able to compare the operator companies with each other and from this, be able to identify best practices that may benefit all.

I appreciate the opportunity to write this thesis for Safetec Nordic, where all the employees have been very encouraging and helpful, which has actually made it fun to write the thesis. Many thanks go to my supervisor Bjørnar Heide from Safetec Nordic, giving me honest and helpful feedback about my thesis, as well as Jan Erik Vinnem from the University of Stavanger, who has been able to answer my questions where others could not.

I would also like to thank Husebø Torleif and Inger Danielsen from PSA for their cooperation and providing the data for my thesis.

Lastly I would like to thank my parents and my husband for their help and support and encouraging words throughout this hectic part of my life.

Abstract

In this thesis a method has been developed to investigate if there is a statistically significant difference between the barrier performances of different operator companies.

The method is developed using the predictive Bayesian approach, as described in Aven (2003). It includes two formulas; one formula for the prediction interval, which will be used to compare the operators' failure rates, and the second formula for making a criterion to define the number of tests required in order to obtain an acceptable level for the failure rates. The method is comparable to the classical methods of "tests of two proportions" and "choice of sample size".

The Classical approach has also been compared to the predictive Bayesian approach. The predictive Bayesian approach uses the previous data as background knowledge to find the predicted value of p. It could be argued that the predictive Bayesian approach is more reliable than the Classical approach because the result from the predictive Bayesian approach tries to describe how the states of the observable quantities are at present or in the future. The Classical approach on the other hand only calculates the probability as it was, implicitly excluding the evolution of the events.

The method, in its essence, compares two of the operators' failure rates at a time, for seven barriers. It determines the operator which has a statistically significant lower failure rate. The method has its main motivation in assuming that operator companies will show trends that reveal different maintenance, inspection and testing schemes and expertise for the barriers.

The results from the developed method show that three comparisons of the operators' barrier performance were found to have a statistically significant lower failure rate, while the results from the Classical approach found twelve such comparisons that showed a statistically significant lower failure rate. The Classical approach found more comparisons with statistically significant lower failure rates, because the developed method performed fewer comparisons due to the criterion for number of tests for next period was not fulfilled for each comparison. However, the developed method results in the same conclusions as the Classical approach, but the Classical approach gives more exaggerated indications compared to the developed method.

1. Introduction

In 1999-2000 a pilot project was launched by the Norwegian Petroleum Directorate (NPD) to develop a tool to measure the risk levels on the Norwegian Continental Shelf (NCS). The reason was that there was mistrust and disputes between the parties in the Norwegian petroleum sector in the latter part of the 1990s. There were extreme concerns from the representatives of the unions and the authority that the risk level was increasing, while the company managers claimed that safety had never been better. In order to create an unbiased and objective channel of information, the NPD started the pilot project “The risk level project” (Vinnem et al, 2006). The result of this project is the tool that has become the RNNP (“Risikonivå I Norsk Petroleumvirksomhet”), which translates: “Trends in risk level in the petroleum activity”. The purpose of “Risk Level project” (commonly referred to as RNNP), as stated in ”Trends in risk level in the petroleum activity” summary report from Petroleum Safety Authority (PSA) (2009), is to monitor the risk level development using various quantitative and qualitative indicators. It is an important part of the universal understanding of the risk level development and aims to improve health, safety and environmental conditions in the Norwegian petroleum sector.

The RNNP has been developed in collaboration with the partners from the industry, who have in consensus agreed that the RNNP is a sensible and rational tool that establishes a common understanding of the level of risk in the petroleum industry.

Before the project started, there was a limited amount of indicators being recorded. Among them; “frequency of work related accidents” and “loss of work hours due to work related accident” were the most common ones measured. It was agreed that these indicators did not give a representative picture of the risks and so additional indicators were necessary. DFU’s (Definert Fare og Ulykkesituasjoner which translates to: “defined situations of hazard and accident”) were developed, and cover all known scenarios that could lead to loss of life. The occurrences of DFU were chosen as the indicators for the frequency of potential major accidents, and the performance of safety and preparedness barriers were chosen as indicators for the barriers quality (PSA, 2000).

As of today, eleven annual RNNP reports have been published and many more papers that discuss methods and study the data from the risk indicators. These risk indicators have been collected by the PSA from the operator companies. The RNNP is based on a triangulation of statistics, engineering and social science to provide a broad and commonly accepted picture of the risk. RNNP emphasizes that there is no single set of indicators that can represent all the relevant aspects of health, environment and safety. It is therefore necessary to show/present the risk indicators in a variety of ways so that more information about the risk levels can be provided. Simply put; the RNNP takes into account the information from observations, indicators and structured processes and uses this information to analyze and evaluate the risks in order to provide input on what risk reducing measures are needed.

1.1 Background

This thesis is a continuation of the core idea that more information can be extracted from the data that is collected if viewed and modeled in a different way than the previous studies. In 2009 Safetec performed the study called “Regression analysis of HC leaks against other indicators in RNNP”. The purpose was to search for correlations between hydrocarbon leaks and other barriers (such as Christmas trees, gas detection, deluge, BOP etc.) for the installations on the NCS. This study concluded that there was a correlation between the HC leaks and the safety culture on board an installation. Other studies have studied the correlation between two specific DFU's, which prompted new ideas. One of these ideas was to compare the barrier performance of the operators, which is the basis for this thesis.

1.2 Thesis problem

By using the barrier data from RNNP, a method will be developed to compare the operator's failure rate, using the predictive Bayesian paradigm. The RNNP also uses the predictive Bayesian approach to develop their risk indicators. The method will then be used to investigate if there is a statistically significant difference between the company's barrier performances. The developed method will also be compared to the equivalent Classical approach of comparing the company's barrier performance to investigate the differences in the two methods.

Why is it important to do this?

The purpose of comparing operator companies is to search for differences in barrier performances between the companies. By dividing the data in groups of operators; it is possible to see if one operator company is better at keeping a lower risk level in one certain area, than another operator. By singling out which barriers the operator companies have the lowest failure rate, a best practice for barriers can be developed and help lift the safety level for the petroleum activity in Norway.

1.3 Limitations

As there are many barriers that can be considered and analyzed, time doesn't allow for all to be considered in this thesis, so only seven barriers will be used as the basis to "check" the developed method. Most of the uncertainty in the data is related to the activity levels and test culture, which will be argued later to be important contributions in the understanding of the barrier performances.

2. Theory

In this chapter different methods will be presented, where analyses of barriers and incidents have been performed, to introduce what this thesis is based on and what has already been done by others.

2.1 Definitions

To be able to achieve the goals of this thesis, some terms need to be clearly defined. The first term that needs to be defined is probability. In school, students are thought that probability is the chance of something happening. This is the layman term used for describing probability in classical statistics.

In classical statistics, probability as defined in Aven (2009) is: “the relative fraction of times the event occurs if the situation analyzed where to be hypothetically “repeated” an infinite number of times”. The belief is that there exists a true probability and the focus is to estimate this underlying true probability through repetitive trails.

In predictive Bayesian approach however, probability is defined as” the means to express the uncertainty about possible occurrences, as seen through the eyes of the assessor, based on the background knowledge”. This probability is regarded as knowledge based expressions of uncertainty, which cannot be described independently of the analyst (Heide 2009). Therefore, contrary to the classical statistics definition of probability, the predictive Bayesian approach focuses on describing the ”state of the world” through observable quantities.

The next term that needs to be defined is risk. There are many definitions of risk, but not all can be used in the same context. Risk as described by Kaplan (1997) is the complete set of scenarios (S_i), the likelihood (L_i) and the consequences (C_i) of each scenario that is the set of all triplets $\{S_i, L_i, C_i\}$.

Another definition of risk is “the combination of the probability or an event and its consequences” from ISO 2002.

Yet another definition comes from Aven (2009 & 2010): risk is understood as the two dimensional combination of: “(i) event A and the consequences of this event C, and (ii) the associated uncertainties U (whether A will occur and what value C will take)”. Ergo: “the uncertainty about and severity of the consequences of an activity”. This is what is called the (A, C, U) perspective.

This last definition is in line with the definition of probability used in the predictive Bayesian approach and will be used to define risk in this thesis.

Next to define is an indicator. An example of an indicator used in everyday life is a thermometer, which is an indicator for temperature. A general definition for indicator is an instrument or variable that is used to measure a condition. According to Øien (2001a) an indicator is defined as: “a measurable or operational variable that can be used to describe the condition of a broader phenomenon or aspect of reality”. A risk indicator is an indicator that measures risk. According to Vinnem et al, (2003b) a risk indicator is defined as: “a measurable quantity which provides information about risk”.

As the thesis will be comparing operators' barriers, a barrier must also be defined. A barrier can be defined as "a measure that reduces the likelihood of triggering a potential risk of damage or reduces the potential for damage or harm" (ISO 17776).

2.2 Classical vs. Bayesian approach

Since barrier indicators provide information about the performance of barriers and systems which can be used to evaluate trends, it is important to choose a suitable statistical approach that will indicate the most realistic trend of the barriers or safety system. As this thesis will be comparing two different types of approaches, the statistical inference for the two approaches will be explained.

From Walpole et al. (8^{th} ed.) it is stated that the theory of statistical inference consists of methods that make generalizations about a population. The classical method is based on inference strictly from information obtained from a random sample selected from the population. The Classical approach is relative frequency based, which means that there is usually an event, x_i, that is divided by the number of times the event has occurred, n_i, in an experiment or study, when the experiment is hypothetically repeated an infinite amount of times. To measure the performance of a barrier, a barrier is repeatedly tested with a result of either success or failure.

$$p = \frac{x_i}{n_i}$$

P stands for the probability of success (or failure in this case as the interest is in measuring the failure rate).
x_i stands for the number of trails that has resulted in a success, and
n_i stands for the number of repeated tests.

The probability distribution for these tests is called the binomial distribution and is denoted by

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0,1,2,3, \dots, n$$

where, each independent trial can result in a success or failure, with probability p or q = 1-p respectively.

As it is not possible to repeat the test or study infinite amount of times, so the Classical approach uses a sample size to represent the population. The average number of failures that is calculated from this sample is given the sign $\bar{X}$. If the expected value of the average number of failures from the sample is the mean for the population; $E(\bar{X}) = \mu$, it is possible to consider that the probability found from the sample is representative for the population.

The RNNP method uses the predictive Bayesian approach, which will also be used in this thesis. According to Aven (2009) this perspective is based on" interpretation, and makes a sharp distinction between historical data and experiences, future quantities of interest such as loss of lives, injuries etc. (referred to as observables) and predictions and uncertainty assessment of these."

According to Aven & Røed (2009), Heide et al. (2007) and Aven (2003), the predictive Bayesian (epistemic) approach focuses on the real population and the observable quantities that can be observed in the future as well. The knowledge that is available to the analyst is used to predict the future values/outcomes of the observations. When using the predictive Bayesian approach, it is assumed that the reference population n_0 is relatively large and that x_0 denotes the number of results that are in the failure mode. The probability of an outcome, for example $P(X_1=1)$ (where X_i is the outcome), is not the same probability or "chance", that is found in relative frequency based classical statistics, but it is the assessors assigned probability that is based on the available knowledge. (If p were known to the analyst, it would be the analyst's probability of X=1 or any other outcome that would be applied). In this context however, the available knowledge is not in the form of expertise knowledge but the knowledge from the data from the previous years. Probability in this context means the measure of uncertainty of observable quantities/barriers in the future.

This means is that from the knowledge or historic data that is available, the assessor assigns a subjective probability, which is the measure of uncertainty of future observable quantities, to say something or express the state of the barrier.

This knowledge can be expressed in this manner:

$$P(X = 1) = \frac{x_0}{n_0}$$

where x_0 is the number of failures from the previous years and
n_0 is the number of tests from the previous years.

Thus the prediction of the state of a barrier can be expressed in this manner:

$$(p|K) = \frac{x_0}{n_0} = p$$

p is the measure of uncertainty that is assigned about the future observable quantity,
K is the knowledge which is represented by the previous years.

This formula expresses that the assessor assigned the measure of uncertainty to be the number of failures divided by the number of tests from the previous years. To make sure that the background knowledge is strong, the reference population should be at least twice the size of the new populations, so that any new knowledge from the first new test will provide little new information to affect the background knowledge that is already established.

The overall expression can be written like this;

$$X{\sim}bin(n, \frac{x_0}{n_0})$$

where X is the observable quantity in the future and has a binomial distribution with parameters
n – which is the number of tests that will be done in the future
x0 – which is the number of failures from previous years and
n0 – which is the number of tests from previous years.

Even though the trials are not independent, it is still valid as long as the previous number of tests n0 is at least twice as large as n (Røed & Aven,2009).

The reason why predicting the probability or proportion of failure from the predictive Bayesian statistics is more valid than Classical statistics, is because the predictive Bayesian approach uses the same data as in the Classical approach, but as a source of knowledge to predict the real probability instead of saying something about how the probability was before, which isn't necessarily representative to what it is now or in the future. When predicting the future it is important to keep in mind that the uncertainties related to the future observable quantities are epistemic, which means, they result from lack of knowledge.

2.2.1 Confidence interval and Prediction interval

In classical data testing, the data are treated as samples from a larger population, and from the samples it is possible to calculate an estimated expected failure rate. However, as point estimates are not good estimates for population parameters, it is preferable to determine an interval where the population parameters are expected to be found. A confidence interval (Classical approach) shows graphically where the true population parameter is likely to be, in the form of (1-α) % degree of confidence.

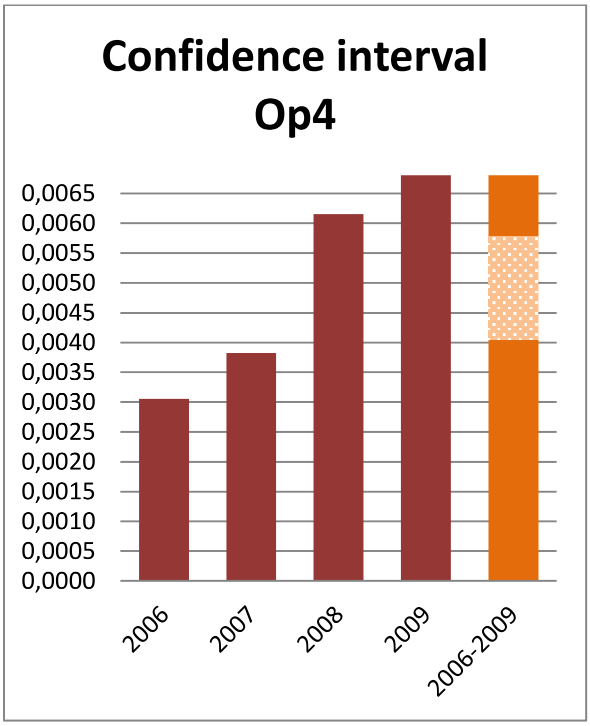
As the data that is being used has a binomial distribution (success or failure) and has many tests (n>>30) it is possible to use approximate normalized distribution.

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{(Eq. 2.2.1)}$$

Using formula (Eq. 2.2.1) where;
p – is probability of failure,
α – level of significance = 10,
z – is 1.65 (for α/2 = 5) and
n is the sample size.

As an example a confidence interval will be made. The data can be found in Appendix B-1.

A 90 % **confidence interval** is made for a 4 year period to be between (α=10);
[0,21 % and 0.25 %] for Operator 1 and [0.40 % and 0.58 %] for Operator 4.



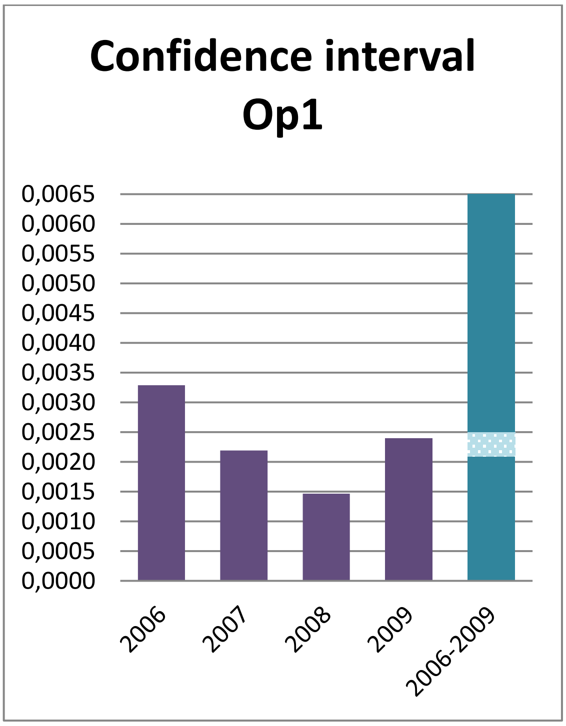


Figure 2.2.1.1 shows graphically the confidence interval for the period 2006-2009. The shaded area represents the 90 % confidence interval.

As can be seen from Figure 2.2.1.1, Operator 1 has a narrower confidence interval than Operator 4 which gives Operator 1 a "more reliable" confidence interval. In other words, the width of the interval shows how much uncertainty there exists for the degree of confidence, therefore the narrowest interval has less uncertainty. This reflects the point stated above that intervals are more reliable than point estimators. The confidence interval shows that Operator 4 has a weaker ground for their "good" failure rate.

In the predictive Bayesian approach the assumption is that the data present represents the population and the goal is to figure out how the population will look in the future. A **prediction interval** can then be constructed to show where the failure rate is likely to lay.

$$P(X = x_0|K) = \binom{n}{x_0} p^{x_0}(1-p)^{n-x_0} \quad \text{Eq.2.2.2}$$

The data observed and collected is considered to be the previous knowledge and is used to determine the "future" values of the probability, by using the number of tests that are planned for the next period.

The parameters;
x_0 – number of failures from "the previous knowledge",
n – number of tests next period (the period that the prediction interval is for),
p – probability of a failure from "the previous knowledge",
α – 10, i.e. 90 % prediction interval.

If the number of tests in the next period for Operator 1 is 37 222 and 4 372 for Operator 4, the prediction interval (from formula Eq.2.2.2) for the failure rate is [0.19 % to 0.27 %] and [0,32 % to 0,66 %] respectively.

In both the prediction interval and confidence interval, the same data is used to calculate the probability of a failure "p". But in the prediction interval the probability includes the number of tests that will be performed next period. This means that the confidence interval and the prediction interval don't describe the same thing; actually confidence interval describes the past while the prediction interval describes the future.

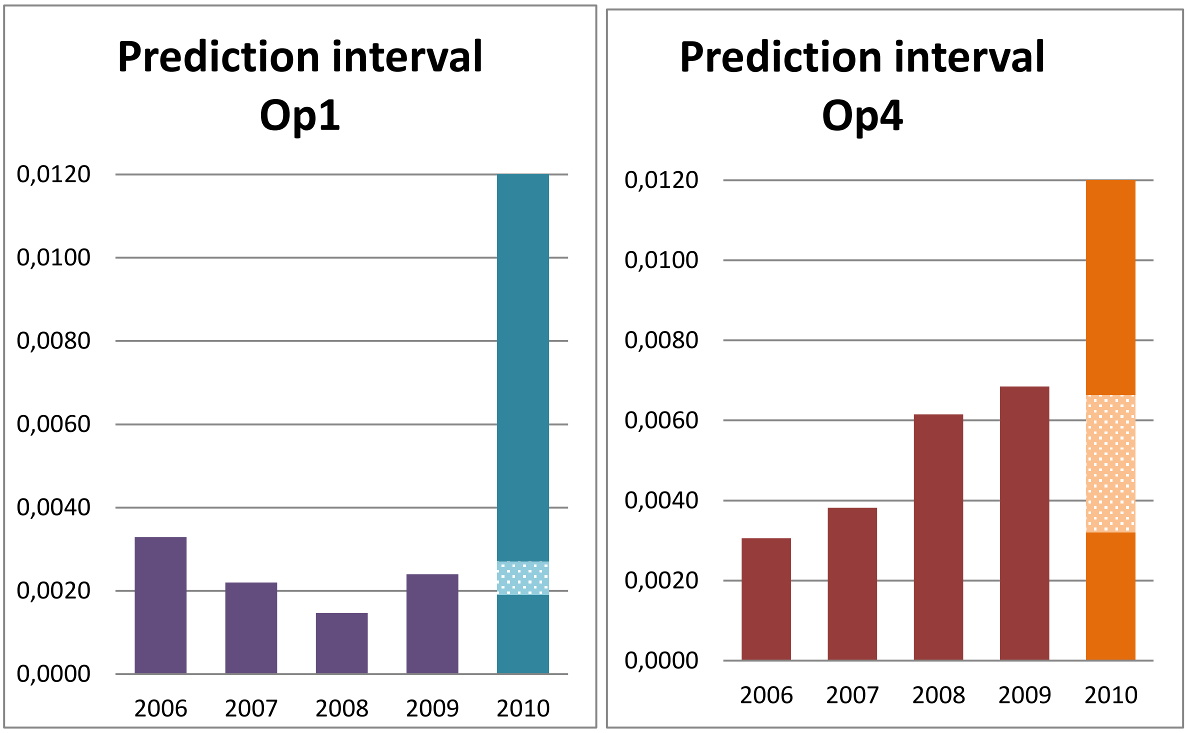


Figure 2.2.1.2 shows the prediction interval for Operator 1 and Operator 4. The shaded area represents the 90 % prediction interval.

Figure 2.2.1.2 shows that Operator 4' prediction interval is less narrow than Operator 1's prediction interval.

In this chapter it has now been presented by means of an example, how the failure rates can be calculated in the Classical approach and the predictive Bayesian approach.

2.2.2 Summary: classical vs. pred. Bayesian approach.

List of the differences for classical and pred. Bayesian approach:

- The Classical approach estimated the probability, p, of a population from the sample by relative frequency; $p = \frac{x_i}{n_i}$, where x_i is the number of failures and n_i is the number of tests. The predictive Bayesian approach assigns a probability from the reference group (x_0 is the number of failures made in the past, and n_0 is the number of tests performed in the past); $(p|K) = \frac{x_i}{n_i} = p = \frac{x_0}{n_0}$, which is called a conditional perspective.

- In the pred. Bayesian approach the parameters are viewed as unknown and random variables while in the Classical approach the parameters are unknown but fixed quantities.
- Pred. Bayesian approach uses data (or previous knowledge) that is ***at least*** twice the amount of the new population to keep a solid background knowledge, while the Classical approach only has a recommendation that n >30.

By comparing the Classical approach and the predictive Bayesian approach, the reader is shown how the predictive Bayesian approach uses a learning process which allows the method to supplement in new information or "trends" to be accounted for in the "future" data testing.

2.3 Indicators

From Vinnem et al. (2006) it is stated that when considering the risk of major hazard during the stay aboard an offshore installation there are two types of indicators that are developed:

1) Event based and,

2) Performance based.

Barrier indicators are performance based and, therefore, it is easier to acquire more of this data than the event based indicators, like near-miss. It is in the company's best interest to record this data as often as reasonably possible to ensure the quality of the barriers that is supposed to protect the people on board from major hazards.

Occurrence based indicators are indicators that measure the occurrence of occupational injuries, and exposure of employees to selected hazards with occupational illness potential.

It is important to note that indicators also fall under two other categories; leading and lagging.

Leading and lagging indicators may have several definitions but not very different.
According to Heide (2009), a Leading/lagging indicator can defined by how quickly the indicator reacts to change. So with this definition, indicators aren't fix to be either leading or lagging, but can be sorted on a continuous scale with leading in one end and lagging on the other, ergo, degree of leading or lagging.

For instance, the number of hydrocarbon leaks is more lagging than gas detection because changes in the gas detection failure rate are found in a shorter time than for hydrocarbon leaks.

According to Vinnem et al (2003a), leading and lagging indicators have another definition:

A proactive (leading) risk indicator is "A measurable quantity which provides information about risk, explicitly addressing an aspect of future performance (for example, anticipated number of hot work hours next year).
A reactive (lagging) risk indicator is "A measurable quantity based on outcomes of accidents and incidents.

Both definitions are quite similar, and barrier indicators will fall under leading indicators in both cases.

2.4 The RNNP method

The RNNP uses the predictive Bayesian approach where the indicators are normalized over a 3 year-period to predict the trends for next year. The trends give a picture of whether the increase or decrease in the amount of, for example, failures of a certain barrier is reasonable, for the period in question. The narrower the prediction interval is, the better it is at noticing trends.

In method report, two different methods are presented for proportion of failures for barrier indicators. The first method is the "total proportion of failure" where the sum of failures on installation j (x_j) is divided by the sum of the number of tests (X_j).

$$\text{Total proportion of failure} = \frac{\sum_{j=1}^{N} x_j}{\sum_{j=1}^{N} X_j} \qquad \text{Eq. 2.4.1}$$

The total proportion of failure reflects the quality of the barrier for installations which perform sufficient number of tests, but it doesn't necessarily reflect the quality of the barrier for the entire continental shelf.

The second is "facilitated proportion of failure" where the sum of the ratio of failures on installation j (x_j) divided by the number of tests (X_j), is divided by the number of installations that has done tests for the barrier element.

$$\text{Facilitated proportion of failure} = \frac{1}{N}\sum_{j=1}^{n} \frac{x_j}{X_j} \qquad \text{Eq. 2.4.2}$$

The facilitated proportion of failure avoids the problem where installations with many test dominate the result, but the statistical data for the installation with fewer test will be poorer. It is therefore seen as necessary to use both methods to gain a greater picture for the barrier performance.

The normalization of data is common in most approaches when analyzing trends. As frequency data usually don't take into account for relevant information about the activity level or exposure level, the trends that are produced contain false signals which need to be eliminated. Normalization is therefore an important aspect when comparing the frequency by a parameter like man-hour, installation years or wells drilled etc. If normalization is not done, the data will most likely give a misguided picture of the risk level.

For onshore plants a method has been devised for compensating for too little data when analyzing status and trends. It is a weighted combination indicator which is produced when multiplying the weighted number of observed incidents (event based data) with observed fraction of test failure (barrier performance based data). It describes the amount of incidents that isn't detected automatically by a certain detector in the area. When there is missing data and "zero occurrences" the assigned number of observed incidents and observed fraction of test failure are replaced with the average values for the continental shelf (Heide et al., 2007). In this way an installation with missing data will not be totally overlooked. (Method report, 2010)

2.4.1 Disadvantages

What are the advantages for using method RNNP? What are the disadvantages by using the RNNP method?

- The disadvantages with the 3 year rolling average is that when there are years that have exceptionally large amounts of failures for a test, the 3 year rolling averages doesn't show this trend and the impression that there is a "new" downwards trend can be misinterpreted. If the data isn't treated with a 3 year rolling and only the yearly failure rate are shown the failure rates may show different trends, where the dominant trend is the more realistic one and the indicator can be analyzed properly. An improvement could be to either omit the years that are not representative for the trend, or use them in a different way when analyzing the data. Maybe the 2 last year's average vs. the last 5 years would represent the trend better etc. The point being made is that the data should be not be analyzed in the same way as trends may be lost.

2.5 Regression analysis

As a part of a research project to investigate if there are possible indicators that can be used to high-light exposures to major accidents and unveil possible important cause factors (either indirectly or directly), Safetec performed the study "Regression analysis of HC leaks against other indicators in RNNP" (2009). Correlations coefficients with number of (non-ignited) HC leaks were calculated for the barrier data, falling objects, serious personnel injuries and noise data in the period 2003 to 2008.

The methods that have been used to perform the regression analysis are explained and presented where they appear to be relevant for the understanding of the concept of developing a method.

Not all the information that is available can be used in the form they are collected. That is why some of this information is normalized, scaled or weighted before being used in a correlation or regression analysis. Below some of these methods are presented.

As the information about the leak points for all the process modules is not available, and not all the people on board can contribute to leak frequencies, the weight of the process module, and personnel related to leak frequency is used instead. When considering the number of personnel that has access to equipment that can cause a leak, the POB is far too high number. Therefore the number used is the number of people that work with process and maintenance. In the analysis, the weight of the process module had a correlation of 65% (from 18 installations), which makes the relationship between the weight of the process module and the number of leak points satisfactory.

Traditional linear regression and Poisson regression are then performed to find if there is a correlation between these variables and the other barriers and occurrences mentioned previously.

Substantial weight was given to the questionnaires that were performed, which are used to determine the risk level of the safety culture. As the method as to which the questionnaires are analyzed and processed will not be used in the thesis, it is not seen as necessary to say more about them that they resulted in a good correlation to the Hydrocarbon (HC) leaks.

To be able to analyze the data for number of serious personnel injuries for the specific time period, some scaling needs to be done.

Scaling serious personnel injuries:

$$\frac{number\ of\ serious\ personnel\ injuries\ in\ the\ period\ 2003-2007+0.5}{number\ of\ people\ per\ shift\ on\ installation\ in\ period} \quad \text{Eq. 2.5.1}$$

In this equation an additional value of 0.5 is added to the injuries in the period so that the installation without registered personnel injuries with high POB received a lower personnel injury value then an installation with a lower POB. This means that installations with a higher POB with no injuries are seen as safer than one with lower POB.

Weighting leak categories: Due to the fact that some HC leaks have a higher potential of harm than others, the HC leaks are be weighted accordingly. From the RNNP report of 2008 the weightings of hydrocarbon leak rates are shown below.

Table 2.5.1: weighting of leakage category based on RNNP weighting

Leak rate	**Weighting of leakage**
<0.1 kg/s	0.0054
0.1 – 1 kg/s	0.0108
1 – 10 kg/s	0.0296
>10 kg/s	0.412

The study makes use of the total proportion of failure (Eq. 2.4.1) for the barrier data to calculate the average total proportion of failure per year per installation. The study also takes into account that there are installations that perform significantly more tests than others by outlining a criterion to filter what data is significant to use. The criterion is; the number of tests for an installation for a given year has to be at least 10% of the average number of test per installation to be included in the study. This applies for the barriers fire- and gas- detection. The remaining barriers must have 20 % to be included.

To be able to compare the probability level for a HC leak for the installations, it helps to assign all installations with a small probability for a HC leak. All installations are given “half a leakage” with rate <0.1 kg/s, which doesn’t reflect a real leak, but rather a size which is used to differentiate the installations with 0 leaks when the number of leaks are scaled.

The study uses Spearman correlation calculations and simple linear –, multiple linear- and Poisson – regression models. The reason for choosing these models and advantages and disadvantages to these methods are explained below.

Spearman correlation is used because it is easily adaptable to a generalized linear model. Advantages to this correlation are; that it can measure non-linear relationships and it is robust against extreme values. The disadvantage is that some information is lost when observations have the same value, as it is not possible to differentiate between observations that are the same.

Simple linear regression is performed with the cause variables listed on p. 8 of the report. Then a ***multiple linear regression*** is adapted to investigate if it is possible to explain parts of the variation in the scaled number of leaks from the cause variables.

From Walpole, (8^{th} Ed. Chpt. 11.1) ***linear regression*** is used to find the best relationship between the dependent variable (Y) and the independent variable (regressor) (x). Most cases are not deterministic (a given x does not always give the same value for Y), and the problems are probabilistic in nature. When a linear regression uses only one regressor, it is called a simple linear regression. This means that the interest lays in the what influence one cause variable has on the relationship that is being analyzed. When there is more than one regressor, it is called multiple linear regression and is used when the interest is to have several cause variables to help explain the relationship. The **reason** why linear regression is so often used, is because more often than not, the relationships between the parameters α, β_1, β_2 etc. are linear in nature. The variance in a linear regression is often of interest because it explains how well the regression line fits to the data. Therefore to analyze the variance, an ANOVA (Analysis of Variance) table can be produced to test the linearity of the regression.

An ANOVA table looks like this:

Table 2.5.2 ANOVA table

Source of variance	**SS**	**df (degrees of freedom)**	**MS**	**F**	**p-value**
Between groups	SSA	k-1	$\frac{SSA}{k-1}$	$\frac{MSA}{MSE}$	$P(F > f_{obs})$
Within groups	SSE	N-k	$\frac{SSE}{k(n-1)}$		
Total	SST	N-1			

SST is the total sum of squares, SSA is the treatment sum of squares and SSE is the error sum of squares. SSE can be due to “noise”, and SSA can be due to different expected values in the groups. The relationship between them is:

$$SST = SSA + SSE$$

k is the number of groups or “treatments” and N is the total number of tests done.

The p – value which is produced from the ANOVA table can be explained as the probability to observe something just as extreme, as what has just been observed, given that the null hypothesis is correct (Kvaløy, 2005).

Also from the ANOVA calculations, the coefficient of multiple determination; R^2, can be calculated. R^2 can be used to find the variables that are useful predictors to find the “best regression”. R^2 measures the proportion of variability in the response y, explained by the fitted regression model. Simply put, if the value of R^2 = 1.0, the regression fits perfectly. The pitfall of having too many variables is that R^2, increases artificially and the model becomes over fitted. $R^2_{adjusted}$ is a variation of R^2 that provides an adjustment for degrees of freedom. Therefore $R^2_{adjusted}$ is calculated as well to show if R^2 has too many variables included, and if a reduced model fits better. This is why it is possible, in multiple linear regression, to analyze the multicollinearity (linear dependency) of the variables to see which are significant and which are not.

When considering using simple linear regression it is important to check that there are no violations of the assumptions, so residual plots (studentized and normal probability) are essential. In multiple linear regressions it is important to keep in mind that linear statistical models are empirical approximations and the true linear model cannot be found. Also, too much attention to R^2 when choosing the so-called best model is unfavorable since the value of R^2 can be as large as one wish.

The advantage and reason for using ***Poisson regression*** (GLM) is that it is not necessary to model for observations with the value 0. This means that it is not necessary to perform subjective scaling to take into account the size of the installation. The only drawback is that it is difficult to take into account that different types of leaks, as they have different types of severity and therefore should be weighted differently. This is however achieved in linear regression so the severity of the leak types is taken into account.

The study concludes that there is a correlation between the safety culture and the frequency of hydrocarbon leakages on an installation. By using different cause variables and different regression methods, it is possible to see if some cause variables have a stronger or weaker correlation to HC leaks. The barrier performance showed that there was a significant correlation with HC leaks as well, with the exception of Christmas tree barrier, however this could also be explained by older installations that have shown to have a negative effect with respect to HC leaks.

2.5.1 Advantages vs. disadvantages

- An advantage in the regression analysis is the criterion that was determined on what data is satisfactory to use, with regard to barrier test data. This criterion is very useful and helps the analysis to give a realistic picture of the barrier performance quality.
- The disadvantages or weaknesses of the regression analyses is that the criterion for what results in a good barrier test are not discussed or taken into account. For example, a pressure test criterion on rig A may be keeping the pressure at 200 bar for 10 minutes, and on rig B it has to hold 150 bar for 30 minutes, and both are considered to be a good test. Another example is that if rig A has some problems with a down hole safety valve (DHSV), and fail to get a good test on four attempts due to complications reaching the valve. In the end they managed to reach the valve and get a good test on the fifth try. On rig B the same test was successful at first attempt. In both these examples it could be considered that both rig A and B have achieved good tests because in their round of testing they managed to achieve good tests at the end. Unless a criterion is established, that for example test need to be successful at first attempt be counted as acceptable, one could in theory have several failed tests every time a round of test is scheduled but if the last test they take is good then it is considered to be a good test. (In a real emergency there may not be a second chance if barrier fail at first attempt to activate)
- Another disadvantage is that the activity level in the production/process and drilling “area” are not considered when for example looking at how many barrier test and HC leaks are registered for process and how many barrier tests are done for drilling. This could give a

good indication for the reasons behind the number of HC leaks and barrier tests. If one assumes that larger installations have more active production wells than smaller installations, it would be logical to think that they have a higher activity level, but that doesn't make it is necessarily true that they have more HC leaks or barrier failures. This is something that could be looked deeper into to reveal if there is any correlation between activity level and HC leaks.

- The regression analyses seem to indicate (but do not conclude) that there are more noise, more falling objects and more HC leaks on installations that are older and larger. Older installations are sometimes assumed to be larger. Therefore, it could be investigated to see if indeed older installations in general are larger, and if the designs of older installations are also less convenient to operate and maintain than the newer ones. This could explain the observations that larger installations seem to have a higher level of risk than smaller ones.

In general the regression analysis doesn't point at any single indicator that has an effect on the HC leaks, but in connection with other indicators (multiple linear regressions) it is shown that there is some relationship to the safety culture.

2.5.2 Improvements & suggestions

- Including the activity level in the regression analysis could give more correlations than what was found.
- Obtaining the barrier testing criterion that each operator use would help understand the background of the testing program, culture and results, which could lead to different conclusions that has previously not been found.

2.6 Requirements for barrier reporting

When analyzing barriers/ safety functions, criteria need to be defined to determine when test is acceptable and when it failed to meet these criteria. If a specific criterion doesn't exist for a barrier, then the decision whether the test was acceptable or failed will be subjective and the analyses will be misleading as mentioned in chapter 2.5.1.

According to ISO 17776, the definition of a barrier is "a measure that reduces the likelihood of triggering a potential risk of damage or reduces the potential for damage/harm". This definition also has to match/fulfill the description of barrier that is stated in the management regulations (styringsforskriften) §1, risk reduction, part 2: "barriers are to be established that a) reduces the likelihood that such errors, hazards and accident situations develop, b) limiting the possible damage and disadvantages".

PSA has made a document where the requirements for the companies reporting for barrier testing are given and is summarized below (PSA, 2010). The barrier data that is available can be divided into 6 categories:

1. Fire and Gas detection:
    a. Proportion of fire detection test failure
    b. Proportion of gas detection test failure
2. Shutdown:
    a. Proportion of riser, ESDV (Emergency shutdown valve) test failure
    b. Proportion of valve closing test failure
    c. Proportion of internal leak rate test failure/breach
    d. Proportion of BVD (Blowdown valve) test failure
    e. Proportion of PSV (pressure safety valve) test failure
3. Wellbore isolation:
    a. Proportion of Christmas tree test failure
    b. Proportion of valve closing (for Christmas tree) test failure
    c. Proportion of leak rate(Christmas tree) test failure/breach
    d. Proportion of DHSV (Down hole safety valve) test failure
4. Wellbore isolation w/BOP:
    a. Proportion of BOP (Blow Out Preventer) test failure
5. Fire safeguard:
    a. Proportion of deluge test failure
    b. Proportion of start test (fire pumps) test failure
6. Emergency preparedness:
    a. Proportion of muster time test failure

The definition for a failure for fire and gas detection test is when the “F&G logic” (Fire and gas logic) doesn’t receive a signal from the detector that it has detected a fire, smoke, heat or gas and activate an alarm in the “F&G panel” (Fire and gas panel). The indicator takes into account each detector, so if one of the detectors fails, the indicator will register one failure, and if two detectors fail; two failures will be registered. For gas detection the gas needs to reach the upper alarm limit to be considered an approved test.

For testing of ESDV, these have to close within the specified time interval, and the leak rate cannot exceed the rates specified for the individual valves. The indicator takes into account each detector like in the fire and gas detection. For BDV the pressure relief valve has to open within a specified time and also takes into account each BDV valve, so the failures will count for each valve. The PSV test is a failure if the PSV doesn’t open at 120% of the set point, or over 50 bar, whichever is the lowest pressure criterion. Also the PSV indicator counts for each valve.

For the wellbore isolation testing, each valve is tested separately. The Christmas tree has 2 separate reporting indicators; first, the Christmas tree valves has specific closing times which need to be satisfied, and second it has to withstand the required pressure so that the internal leak rates are within the boundaries of acceptance. The time requirement doesn’t apply for the DHSV, but the leak rate requirement does. The indicator takes into account each valve when counting the tests and failures.

The BOP is pressure tested, and needs to keep a constant pressure for a certain time period to be approved. The number of failures is counted per sealing element and has a reference to NORSOK D-

010 table A.1 (Appendix A). The amount of tests is defined as the number of pressure tests per sealing element in the BOP.

The fire safeguard includes the deluge valves and the fire pumps. The failure mode for the deluge valve is when it doesn’t open. The indicator counts per deluge control valve, including the signal path from manual and automatic activation in the deluge panel. The fire water supply results in a failure if the firewater pumps don’t manage to start within a specified time and supply the minim volume and pressure. The indicator counts per pump, independently of its capacity, in addition the status of the pumps are continuously monitored.

The emergency preparedness is measured in the form of the time it takes to muster according to emergency instructions. The indicator includes real mustering alarms and practice drills. The attendances of the drills are recorded (although it is not required to report it) and the time for all personnel to be accounted for is registered. The required mustering times are different for the individual installations due to different size and POB. The failure is registered if the required mustering time is exceeded.

All these requirements are understood in the sense that once a barrier has failed, the failure is investigated to find the source of the problem and rectify this before testing the barrier again. So from the example in chapter 2.5.1 if a barrier is tested 5 times and the 4 first tests were failures, then the indicator registers 4 failures. It is also clear now from NORSOK D-010 that the requirement for pressure testing the BOP has to fulfill a minimum requirement as stated in table A.1. The time intervals in which they have to be tested are also stated. It is assumed that the other barriers have similar regulatory requirements, specified for each company, to ensure that a minimum standard has to be achieved.

3. Method

To be able to compare the operator companies, it is first necessary to define the method to compare them. In this section the methods to analyze the data for the different types of barriers will be explained. To continue the comparison of classical vs. pred. Bayesian, the Classical approach will be presented first.

3.1 The Classical Approach – Comparing two proportions.

To find a method that will distinguish which installations/operators have the lowest proportion of failure for their barriers, the simplest way is to compare two installations and find the probability that one of the installations has a lower failure rate than the other.

The normal distribution is a very suitable approximation for errors (distributions) in scientific measurements. It depends on the two parameters variance and mean. If the sample contains more than n ≥ 30 observations or tests, the sample distribution of the estimated mean, $\bar{X}$, can be said to be a good normal approximation. The central limit theorem (Walpole et al. 8th ed. Theorem 8.2), states:

"If $\bar{X}$ is the mean of a random sample of size n taken from a population with a mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$$

As n→∞, is the standard normal distribution n (z;0,1)"

where:
Z – is a standard normal random variable/standardizes normal random variables
$\bar{X}$– is the average of the sample collected from the population
σ^2 – is the true variance of the population
n – is the number of tests from the sample

The central limit theorem allows the analyst to quantify the statistical uncertainty that is associated with a single sample mean. A confidence interval can be constructed around the sample mean. As n → ∞ the distribution becomes a standard normal distribution.

This means according to the central limit theorem, $\bar{X} \approx N(\mu,\sigma^2)$, $Z_1 = \frac{\bar{X}_1-\mu_1}{\sigma_1/\sqrt{n_1}} \approx N(0,1)$ when $n_1 \to \infty$, and $Z_2 = \frac{\bar{X}_2-\mu_2}{\sigma_2/\sqrt{n_2}} \approx N(0,1)$ when $n_2 \to \infty$ etc.

To be able to compare two distributions with each other, another theorem is necessary to introduce; theorem 7.11 (Walpole et al. 8th ed.) about linear combination of random variables. It states that:

"If $X_1, X_2, \dots, X_n$ are independent random variables having normal distribution with means $\mu_1, \mu_2, \dots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, then the random variable

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Has a normal distribution with mean

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

And variance

$$\sigma_Y{}^2 = a_1{}^2\sigma_1{}^2 + a_2{}^2\sigma_2{}^2 + \cdots + a_n{}^2\sigma_n{}^2 \text{ .”}$$

This means that $Z_1 - Z_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}\right)$.

By using the central limit theorem (theorem 8.2) and theorem 7.11 it is possible to compare two populations with each other from the formula below;

Eq. 3.1.1 (Theorem 8.3)

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}}$$

Z – is a standard normal random variable/standardizes normal random variables
$\overline{X_1}$ – is the average of the sample collected from population 1
$\overline{X_2}$ – is the average of the sample collected from population 2
μ_1 – is the true average of population 1
μ_2 – is the true average of population 2
σ_1^2 – is the true variance of population 1
σ_2^2 – is the true variance of population 2
n_1 – is the number of tests from sample population 1
n_2 – is the number of tests from sample population 2

The two distributions are independent and normally distributed; $N\ (\mu_1, \sigma_1^2)$ and $N\ (\mu_2, \sigma_2^2)$, because n_1 and n_2 for the two distributions are very large (n>>30) and $np \geq 5 \ or\ n(1-p) \geq 5$. As n→∞ the distributions become standard normal distributions, and therefore the differences between the sample means are: $\overline{X_1} - \overline{X_2}$ and the sampling distribution: $\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}\right)$.

As mentioned in Walpole et al (8th ed. chapter 9.11), it is stated that when $X_1 \sim bin(n_1, p_1)$, and $X_2 \sim bin(n_2, p_2)$, and where the proportions (of failure) are $\widehat{p_1} = x_1/n_1 \approx N(p_1, p_1(1-p_1))$ and $\widehat{p_2} = x_2/n_2 \approx N(p_2, p_2(1-p_2))$, a confidence interval for $p_1 - p_2$ can be established by using $\widehat{p_1}$ and $\widehat{p_2}$. $\widehat{p_1}$ and $\widehat{p_2}$ can safely be assumed to be approximately normal distributed as long as $n_2\widehat{p_2} \geq 5$ and $n_1\widehat{p_1} \geq 5$. From theorem 7.11, as previously explained, it can be concluded that;

$$\widehat{p_1} - \widehat{p_2} \approx N\left(p_1 - p_2, \frac{p_1(1-p_{1)}}{n_1} + \frac{p_2(1-p_{2)}}{n_2}\right)$$

With this information it is possible to make the assertion that $P\left(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right) = 1-\alpha$, where

$$Z = \frac{(\widehat{p_1}-\widehat{p_2})-(p_1-p_2)}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right)+\left(\frac{p_2(1-p_2)}{n_2}\right)}}$$

Eq. 3.1.2

and α is the level of significance, usually 10%, and $z\alpha/_2$ is the Z-value leaving an area of $1-\alpha/2$ outside the interval. As long as $n_1\widehat{p_1},\ n_2\widehat{p_2},\ n_1(\widehat{1-p_1})\ and\ n_1(\widehat{1-p_1})$ are all greater than or equal to 5, it is possible to replace p_1, p_2 with their estimates $\widehat{p_1}$ and $\widehat{p_2}$. The resulting formula becomes:

$$Z = \frac{(\widehat{p_1}-\widehat{p_2})-(p_1-p_2)}{\sqrt{\left(\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}\right)+\left(\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}\right)}}$$

where;

$\widehat{p_1}$ – is the estimated probability of population 1
$\widehat{p_2}$ – is the estimated probability of population 2
p_1 – is the true probability of population 1
p_2 – is the true probability of population 2
n_1 – is the number of tests from sample population 1
n_2 – is the number of tests from sample population 2

Table 3.1.1 Data table for the PSV barrier for the two installations.

2004-2006			
Installation	Total number of test for PSV (n)	Total number of failures (x)	Proportion of failure for PSV ($\hat{p}$)
AB (1)	1685	176	0,1044
AH (2)	826	65	0,0787

Width: $(\widehat{p_1}-\widehat{p_2} = 0{,}0257)$

As an example, random samples from two installations, for the PSV barrier, are chosen; installation AB (p_1) and installation AH (p_2). See Appendix B-2 for data used to find the values in Table 3.1.2.

The first thing that needs to be done is to test if p_1 actually is different from p_2. Null hypothesis; H_0: $p_1 = p_2$ and alternative hypothesis H_1: $p_1 \neq p_2$. Using Eq. 3.1.2 it is possible to test the hypothesis.

$$Z = \frac{(\widehat{p_1}-\widehat{p_2})-0}{\sqrt{\left(\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}\right)+\left(\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}\right)}} \sim N\left(p_1-p_2\,, \left(\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}\right)+\left(\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}\right)\right) under\ H_0$$

The rejection area is then outside a two tailed 95 % confidence interval. Using the values in Table 3.1.1., the values in the formula give; Z= 2.152 (98.4%), which is larger than the confidence interval Z ± 1.96 (97.50%, 2.50%). Therefore H_0 is rejected in favor of H_1.

As values for p_1 and p_2 are estimated values instead of true, a t-distribution would be more suitable to use instead of the normal distribution where;

$$T = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{z}{\sqrt{V/(n-1)}},$$

where $V = \frac{(n-1)S^2}{\sigma^2}$ which then results in:

$$T = \frac{Z}{\sqrt{V/v}}$$

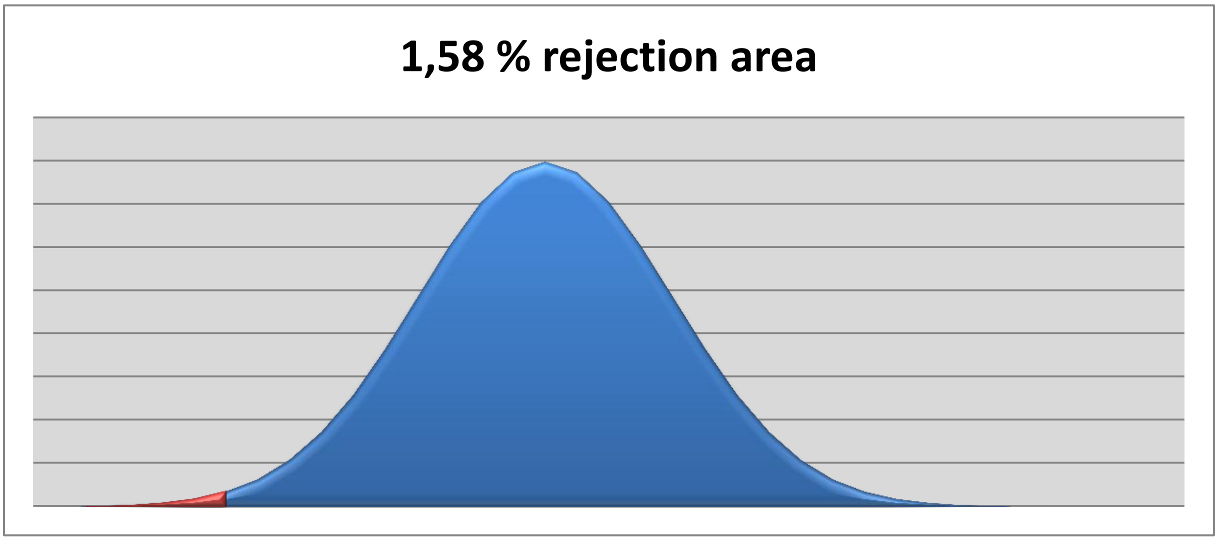
T – is a student-t random variable
Z – is a standard normal random variable
$\bar{X}$ – is the average of the sample collected from the distribution
μ – is the true average of the distribution
σ^2 – is the true variance of the distribution
n – is the number of tests from sample the distribution
s^2 – is the sample variance
V – is a chi-squared random variable
v – is the degrees of freedom

Where Z is normally distributed: $= \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.
The t-distribution is quite similar to the standard normal distribution, in fact, when n→∞ they are identical. But when n ≤ 30 (like for example when determining the height of the boys in a class), the values of S^2 fluctuate considerably from sample to sample and is very different from the standard normal distribution. This is because the T-value depends on both $\bar{X}$ and S^2, while the Z value only depends on $\bar{X}$. Since the variance in the t-distribution depends on the sample size, n, the variance is always greater than 1, whereas the variance for the standard normal distribution is 1. The tail of a t-distribution is usually larger, giving the shape of the bell curve to be lower and wider than the normal distribution. As the number of degrees of freedom increase, the distribution becomes more and more normal.

Another difference between the t-distribution and the normal distribution is that the t-distribution doesn't relate to the central limit theorem. However as the number of tests in this case is so large, S from the sample is treated as a sufficiently good estimator for σ, or as stated in Walpole et al (8th ed.); " S will be very close to the true σ and thus the central limit theorem prevails". This is the concept of a "Large-sample confidence interval" (chapter 9.4, Walpole et al 8th Ed.).

To get a picture of the sampling distribution, a 90 % confidence interval for $p_1 - p_2$ is constructed: 0.0061 < p1 – p2 < 0.0454, which gives an interval width of 0.039. The confidence interval shows that with a 90 % confidence that p_1-p_2 will always be higher than zero. To find the probability that P (p_1 -

$p_2 > 0$), a one tailed confidence interval is constructed, where the lower limit is zero. The area under the graph to the right of the lower bound is the probability that P ($p_1 - p_2 > 0$).



Graph 3.1.1 Graphical representation of the probability; P ($p_1 - p_2 > 0$).

With a low end, one tailed confidence interval of 98,42 %, the value of the limit is 0, 00002255 ≈ 0. It is therefore found that P ($p_1 - p_2 > 0$) = 98,42 %.

The probability that installation AB has a higher proportion of failure than installation AH seems reasonable, as installation AB always has a higher proportion of failure than AH in each of the years.

3.2 Pred. Bayesian Approach – Comparing two proportions.

Today's literature doesn't show how to compare proportions based on the pred. Bayesian approach. In this chapter, one approach on how this could be done will be presented.

As the same criteria are fulfilled in the pred. Bayesian approach as in the Classical approach, the theorems are briefly summarized below.

From chapter 3.1 it was shown that by using the central limit theorem with the theorem about linear combination it is possible to compare two populations with each other giving;

$$Z_1 - Z_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}\right)$$

Or the more useful;

$$\widehat{p_1} - \widehat{p_2} \approx N\left(p_1 - p_2, \frac{p_1(1 - p_{1})}{n_1} + \frac{p_2(1 - p_{2})}{n_2}\right)$$

And the formula becomes;

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}}$$

Or the equivalent:

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}}$$

Eq.3.2.1

Z – is a standard normal random variable/standardizes normal random variables
$\widehat{p_1}$ – is the estimated failure rate from the sample of population 1
$\widehat{p_2}$ – is the estimated failure rate from the sample of population 2
p_1 – is the true failure rate of population 1
p_2 – is the true failure rate of population 2
n_1 – is the number of tests from sample population 1
n_2 – is the number of tests from sample population 2
as $n \to \infty$

The pred. Bayesian approach uses the previous knowledge (or data) to calculate the failure rate in the future by using the number of tests for the next testing period.

The data can be considered approximately normally distributed because the barrier data are binomial random variables with parameters n (number of tests next period) and p (the probability of a failure), with $\mu = np$ and $\sigma^2 = np(1-p)$. The requirements are still valid as μ is the equivalent future mean np, and σ^2 is the future variance np(1-p).

As long as $np \geq 5$ and $n(1-p) \geq 5$, the data can be considered approximately normally distributed even in the pred. Bayesian approach. Below the relationship is shown using the formula;

$$\mu = \bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \approx p_f = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n_f}}$$

where;
p_f – is the future probability
μ – is the true mean
$\hat{p}$ – is the probability calculated from the previous data
n_f – is the number of tests that will be performed for the next period
n – is the number of tests already performed
σ^2 – is the variance
α – level of significance = 10,
z – is 1.65 (for α/2 = 5)

It is therefore possible to conclude that the same criteria are fulfilled here as in the Classical approach and the limiting form of the normal approximation can be used to express the future failure rate p_f;

$$Z\alpha_{/2} = \frac{\hat{p} - p_f}{\sqrt{\hat{p}\hat{q}/n_f}} \rightarrow \hat{p} \pm Z\alpha_{/2}\sqrt{\hat{p}\hat{q}/n_f} = p_f$$

Substituting this in for Eq.3.2.1 the results become;

$$Z = \frac{(p_{1C} - p_{2C}) - (p_{1F} - p_{2F})}{\sqrt{\left(\frac{p_{1C}(1 - p_{1C})}{n_{1F}}\right) + \left(\frac{p_{2C}(1 - p_{2C})}{n_{2F}}\right)}}$$

where;
n_{1F} and n_{2F} are the future number of test,
p_{1F} and p_{2F} are the future proportion of failures,
and n_{1C}, n_{2C}, (and consequently the values x_{1C} and x_{2C}), p_{1C} and p_{2C} are the number of tests (and number of failures) and proportion of observed failures in the experience data that have been determined by the analyst. The prediction interval then becomes;

$$(p_{1C} - p_{2C}) \pm Z\alpha_{/2}\sqrt{\left(\frac{p_{1C}(1 - p_{1C})}{n_{1F}}\right) + \left(\frac{p_{2C}(1 - p_{2C})}{n_{2F}}\right)} = (p_{1F} - p_{2F})$$

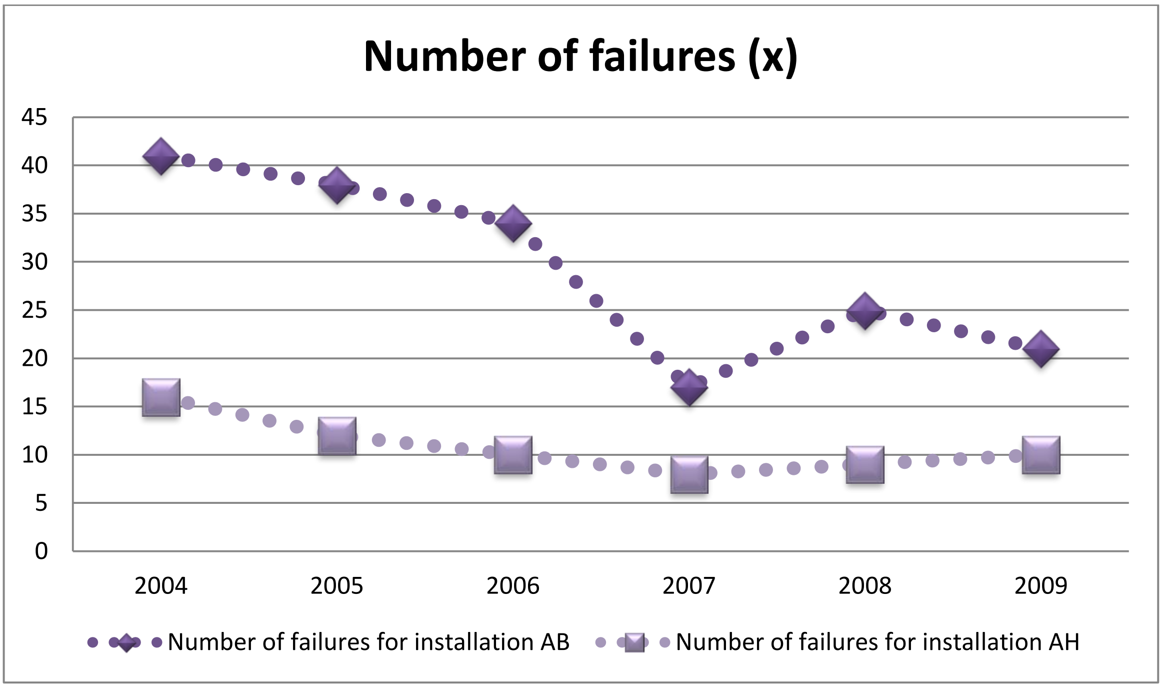
Eq 3.2.2

To find what factors can influence the results using the previous formula, the method for implementing (analyzing) the formula will be divided into steps. First all the normal data used in the Classical approach will be used to see what the formula actually does. This gives an interval of approximately 0.039. However, as it is unrealistic that the number of test in the next period is going to be as high as the past 6 years, the second step is to use realistic values for the number of tests for the next period. Using Eq 3.2.2 with n_{1F} = 232 and n_{2F} = 158 (which are values that are calculated with the least square method below), the results become:

$$-0{,}0225 < (p_{1F} - p_{2F}) < 0{,}0741$$

The interval is now broader, 0.097, because of the lower values of the number of test for the next period, and the interval now includes values below 0. This means that it is not possible to say with 90 % certainty that the proportion of failure for installations AH is higher than installations AB.

A possible third step could be to choose the appropriate value of the proportion of failure. One way that this can be done is to look for trends in the data, and exclude the data that doesn't follow the dominant trend. Practically this means choosing the amount of failures and of tests the installations have performed the last years. From the table below, the number of failures for the last 6 years is shown:

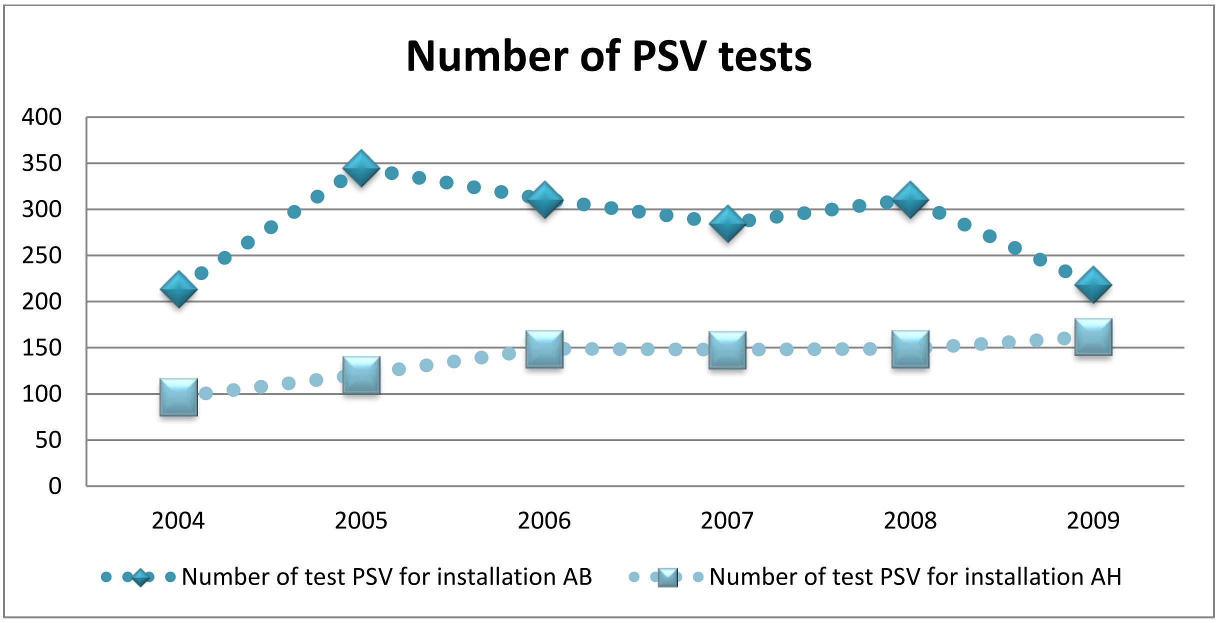


Graph 3.2.1 The number of failures for installation AB and AH for a 6 year period.

From Graph 3.2.1 a decreasing trend is visible for installation AB, where the number of test failures is decreasing every year. By only looking at the graph, it is possible to see that in 2007, there was a stronger decrease than the previous years, and then in 2008 the rate goes back to the trend shown before 2007. If the trend had continued in 2007, the number of failures that should have occurred that year would have been around 30. For installation AH, it seems that the number of test failures have first decrease a little and then started to slightly increase again. Overall it would seem it isn't changing much, but if the focus would be on the last 4 years, it could seem like a new trend might have arisen, where the number of failures doesn't change with more than 1 or 2 failures, as if it is flat lining a little.

To make a proper assumption on which years gives a representative picture of the number of test failures, the years that don't follow the overall trend are not included in determining the value for x_{1C} an x_{2C}. For installation AB, x_{1C} is determined to be 159 (excluding the data from 2007) and x_{2C} for installation AH to be 37 (excluding the data from 2004 and 2005).

This next graph shows the number of tests that have been performed the last 6 years:



Graph 3.2.2 The number of PSV tests for installation AB and AH for a 6 year period

From Graph 3.2.2 there are two peaks that seem pronounced for installation AB. A decrease trend from 2005 is visible with a small increase in 2008 before the trend continues as before. For installation AH an increasing trend is shown, however the two first years, 2004 and 2005 has a steeper increase than the last 4 years, where 2006 – 2008 are basically the same with the only change is seen in 2009.

To determine the value for n_{1C} and n_{2C} the same method is used as when determining x_{1C} and x_{2C}. As installation AB has a main decrease trend the first year, 2004 and 2008 are excluded since these don't represent the trend very well. The value for n_{1C} is then determined to be 1160. For installation AH the newest trend from 2006 to 2009 is chosen to be the representative trend and therefore 2004 and 2005 are excluded. The value for n_{2C} is then 608.

From Graph 3.2.2, the number of test for the next period can be estimated by the help of the least square methods. Using the data points chosen by the analyst and making a regression line, the next point can be estimated, the number of test for the next period.

What the method of least squares does is that it finds sum, where the squares of the residuals (the error in the fit of a linear model $\boldsymbol{y} = a + b\boldsymbol{x}$) becomes the minimum. This means that it finds the line were the residuals are at the minimum, which is known as the regression line.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_1}{n} = \bar{y} - b\bar{x}$$

From the values that are chosen, and listed in Appendix B-2, the values of a and b for installation AB are: a = 304.42 and b = -14.43, the resulting regression line becoming: y = 304.42 – 14.43x. The estimated number of test for the next period is then calculated to be 232.

For installation AH the values for a and b are: a = 150.57 and b = 1.43, and the resulting regression line becomes: y = 150.57 – 1.43x. The estimated number of test for the next period is calculated to be 158.

Table 3.2.2 Prediction interval data for the two installations.

n_{1f}	232	x_{1c}	159
n_{2f}	158	x_{2c}	37
n_{1C}	1160	p_{1C}	0,137
n_{2C}	608	p_{2C}	0,061

The 90% prediction interval for the future proportion of failures can be found using Eq. 3.2.2 and the data in Table 3.2.2:

$$(p_{1F} - p_{2F}) = (p_{1C} - p_{2C}) \pm Z_{\alpha/2} \sqrt{\left(\frac{p_{1C}(1 - p_{1C})}{n_{1F}}\right) + \left(\frac{p_{2C}(1 - p_{2C})}{n_{2F}}\right)}$$

$0.0276 < (p_{1F} - p_{2F}) < 0.1248$, giving an interval width of 0,097.

The interval shows that the probability of p_{1f} and p_{2f} is always larger than zero with a 90 % prediction interval. To find the probability that p_{1F} is larger than p_{2F} the same method as for the Classical approach can be used; by construction a one tailed prediction interval. Assigning the lower limit to the value of 0,00000246 ≈ 0, it is found that P (p_{1f} - p_{2f} ≥ 0) = 99,51%.

This shows that p_{1F} in a 99,51% prediction interval will have a significantly higher failure rate than p_{2F}. As the probability that p_{1F} is larger than p_{2F} for the Classical approach is a little lower (with 98.42%) than the pred. Bayesian approach, it could be argued that with the pred. Bayesian approach the results will show a stronger indication of the real risk. The point of constructing a pred. Bayesian interval is to illustrate a more “real” trend of the barrier performances of different installations.

The main point is to find the general trend and not include misleading data points in the trend. Therefore some data points that do not follow the general trend are excluded.

The prediction interval width for the pred. Bayesian approach is wider than for the Classical approach in chapter 3.1.But it is important to consider that the data that has been used for making the prediction interval is only the data that is relevant. Consider the scenario that a risk analysis is performed on an installation that has fewer data than what is common or satisfactory. The analyst might look at the interval and think that the interval is too broad and therefore gives an uncertain representation of the risk. However if the analyst were to take in more data that isn’t as relevant to the analysis, the interval width could become narrower. But is it really making the results more accurate? In the predictive Bayesian approach this is taken into consideration because when the analyst selects the data, the data is passed through the analyst’s filter of what is realistic or relevant and what is not. When the data then is represented in a prediction interval, the picture that it illustrates is the most realistic combination of the data and the understanding of the barrier performance.

This method might seem very time consuming to use when the data that is available consist of many installations and many years of operation. One way this method can be used effectively is to choose one installation that is of particular interest and compare it to the average of the rest of the installations. In this way the installation that is of interest can be analyzed for major differences compared to the average “normal” installation on the Norwegian continental shelf.

Another method, that will be the subject of this thesis, is to find the overall failure rate for each operator and use the method to find which of the operators are statistically significant better than the others.

3.3 Number of tests criterion.

When considering how to implement the method, some ideas were investigated but were found to be not suitable for the purpose. One of these ideas was to use a one way ANOVA table for each barrier type and test each barrier type has a statistically significant lower failure rate than the others. The hypothesis test would be defined as; the null hypothesis, H_0: p_1=p_2=p_3=p_4 and the alternative hypothesis, H_1: at least one of the failure rates will be different from the others. However the test would only show if there exists a statistically significant difference and would not show which one(s) it was. This poses as a problem as the interest is to find which operator has a statistically significant lower failure rate, and therefore this approach was not used.

As mentioned in the introduction 6-7 barriers will be considered, one from each of the categories presented in chapter 2.6:

a. Proportion of muster time test failure
b. Proportion of start test (fire pumps) test failure
c. Proportion of BOP (Blow Out Preventer) test failure
d. Proportion of fire detection test failure
e. Proportion of Christmas tree test failure or DHSV
f. Proportion of riser, ESDV (Emergency shutdown valve) test failure

From chapter 3.2 the developed method requires the number of tests that will be performed for the next period. One way to estimate this number is using the least square methods, if this information isn’t available. In this case the data for the next year, 2010, is available and it is possible to find how many tests were performed.

To make this number more useful, a method to calculate the number of tests that is required to satisfy an accuracy criterion for the failure rates is developed. In other words; the method will calculate the number of tests that an installation must perform, to be within the accepted error of the failure rate. The installations that have performed at least this number of tests for the next period will be the contributing installations in calculating the average number of tests performed for the next period.

How many tests should a barrier have performed to be within a certain safety criteria? Consider the probability, p, as an estimated failure rate criterion for a barrier. Estimating the failure rate with few tests is not very accurate, so if an error uncertainty, e, also is considered together with a prediction interval, the failure rate will be more accurate and reliable.

For a binomial random variable x, the expected future value μ_f is defined as n_f multiplied by p, where n_f is the number of tests for next period, multiplied by p the failure rate criterion, which in this case is determined by the number of failures observed from previous tests. The operator companies have a set failure rate criterion for each barrier, and should be used instead when available.

$$\mu_f = E(x)_f = \bar{x}_f = n_f p$$

The variance for a binomial distribution is defined by n multiplied by p, multiplied by q, where q = 1-p.

$$\sigma^2 = Var(x) = npq$$

When the number of tests multiplied with the observed failure rate is larger than 5, a standard normal distribution can be assumed giving;

$$Z\alpha/_2 = \frac{X - np}{\sqrt{n\hat{p}\hat{q}}}$$

with α level of significance.

This shows that the expected value for x with its error uncertainty can be expressed as;

$$E(X)_f \pm e$$

which gives;

$$e = Z\alpha/_2\sqrt{n_f\hat{p}\hat{q}}$$

This is how the error uncertainty is expressed for the random variable X, however the expression for a proportion, p, the accepted error will look like this;

$$Z\alpha/_2 = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n_f}}} \rightarrow e = Z\alpha/_2\sqrt{\frac{\hat{p}\hat{q}}{n_f}}$$

Rearranging the formula to give the number of tests needed to satisfy a certain accepted error becomes:

$$n_f = \frac{Z\alpha/_2{}^2\hat{p}\hat{q}}{e^2}$$

$Z\alpha/_2$ is the Z-value leaving an area of $1 - \alpha/_2$ to the right, out of the prediction interval
$\hat{p}$ is the estimated failure rate criterion
$\hat{q} = 1 - \hat{p}$
n_f is the number of future tests that is needed and
e is the "accepted" error that will define the accuracy of the calculated failure rate.

The value for the estimated failure rate criterion would in this case mean the accepted failure rate each operator had for their barriers. As these accepted failure rates are unknown at present time, the values have been estimated from the lowest or "best" failure rates for each operator that are available. Table 3.3.1 shows the accepted failure rate for 7 barriers for each operator, rounded down to 1 decimal.

Table 3.3.1 failure rate criteria for 7 barriers

Barrier	Operator 1	Operator 3	Operator 4	Operator 5
Muster time	14,8%	3,6%	10,5%	21,7%
Start test fire pumps	0,2%	0,3%	0,3%	0,2%
BOP	0,3%	3,7%	2,0%	3,8%
Fire detector	2,4%	0,1%	0,3%	0,3%
Christmas tree	0,3%	0,1%	0,9%	0,1%
DHSV	2,4%	0,1%	4,9%	2,8%
ESDV	1,0%	0,1%	0,5%	3,8%

As some of these failure rate criteria are quite high, the values have been reduced to a more suitable value, shown in table 3.3.1-1

Table 3.3.1-1 failure rate criteria revised.

Barrier	Operator 1	Operator 3	Operator 4	Operator 5
Muster time	8%	1,9%	6,0%	10%
Start test fire pumps	0,2%	0,3%	0,3%	0,2%
BOP	0,3%	2,5%	2,0%	2,7%
Fire detector	2,4%	0,3%	0,3%	0,3%
Christmas tree	0,3%	0,3%	0,9%	0,1%
DHSV	2,4%	0,3%	4,0%	2,8%
ESDV	1,0%	0,3%	0,5%	3,0%

These failure rates might seem very strict for some of the installations for the operators but as these are estimated values in the place of the operator's real failure rate criteria it isn't necessary to put too much emphasis on them. If the failure rate criteria are used in the way they are meant to, a barrier should be properly examined and tested when it is over the failure rate set by the operator.

The next value that needs to be determined is the accepted error, ***e***. This could be done in different ways; there could be an e for each barrier, or for each operator of both. For simplicity and comparability, the accepted error will be chosen to be for each barrier as these are the values that are supposedly most similar to each other.

Table 3.3.2 error, e, in estimating p, for each barrier

Barrier	Accepted error, *e*
Muster time	0,070 – 7%
Start test fire pumps	0,0100 – 1%
BOP	0,0200 – 2%
Fire detector	0,008 – 0,8%
Christmas tree	0,007 – 0,7%
DHSV	0,012 – 1,2%
ESDV	0,012 – 1,2%

In order to be sure that the error in the calculated failure rate is small, the value for ***e*** should be as small as possible, but still be realistic in practice.

Using a 90% critical region, the number of tests each barrier must perform next period to satisfy the accepted error criterion for the failure rates are shown in Table 3.3.3.

Table 3.3.3 number of tests required for each installation for each operator.

Barrier	Operator 1	Operator 3	Operator 4	Operator 5
Muster time	41	10	31	50
Start test fire pumps	54	81	81	54
BOP	20	165	133	178
Fire detector	990	126	126	126
Christmas tree	165	165	493	165
DHSV	440	56	722	511
ESDV	186	56	93	547

These results are quite interesting, for example it shows that Operator 3 only needs to take 10 tests per installation per year instead of 26 which is the normal amount of tests for mustering on offshore installations. This is because their observed failure rates are low and given their criteria, they have established a good barrier for mustering time. It also shows that the other operators don't test the mustering barrier enough and therefore need to perform more tests than the normal 26.

The results presented in Table 3.3.3 are sensitive to the estimated values of the failure rate $\hat{p}$ and ***e***. If the values of these two estimates were to change, the value for n will also change correspondingly.

The values for ***e*** are estimated with respect to how many tests can be performed and still be achievable for the operators. If ***e*** were to be too small, the value for n will be very high and it would be nearly impossible to perform that amount of tests of the barrier. The way this problem has been solved is to choose one e for each barrier to make the operators comparable to each other and so that none of the operators exceed the assumed maximum number of tests. The results will then show that one or several of the operators might be close to the maximum number of tests and the rest will be lower.

Now that the estimated number of tests for each operator is established, it is possible to choose the installation that has at least these amounts of test and compare them using the method described in chapter 3.2.

4. Results

In this thesis a method for comparing barrier performances from different operators has been developed using the arguments and views of the predictive Bayesian approach. The core idea of the method is that by estimating the failure rate for a barrier from previously observed data, and combining it with the number of tests that will be performed for the next period, it is possible to calculate a more accurate prediction interval for the failure rate than making a confidence interval in classical statistics. Furthermore, the value for the number of tests that will be performed for the next period is calculated by choosing an accepted error that defines the quality of the barrier, together with a chosen failure rate criterion for each barrier. This is given by the formula;

$$n_f = \frac{Z{\alpha/_2}^2 \hat{p}\hat{q}}{e^2}$$

The method has then been used to compare two failure rates with each other (the overall operator failure rate), to determine if one has a statistically significant lower failure rate than the other. This is given by the formula;

$$(p_{1C} - p_{2C}) \pm Z\alpha/_2 \sqrt{\left(\frac{p_{1C}(1-p_{1C})}{n_{1F}}\right) + \left(\frac{p_{2C}(1-p_{2C})}{n_{2F}}\right)} = (p_{1F} - p_{2F})$$

The interval width's for the prediction intervals varied, but it is worth mentioning that when the accepted error is integrated in the method, the interval widths are at their narrowest in these comparisons.

The following results were found when comparing the operators overall failure rate for each barrier using this method, and the Classical approach for comparison.

4.1 Fire detection barrier

From Appendix C, the graphs show the number of failures and number of tests two operators (Operator 1 and Operator 4), from 2004 to 2009. From these graphs a trend is determined and with the data that participate in the general trend, a prediction interval is made with the number of tests that are performed for the next year. The prediction interval is tested to find if there is statistically significant difference between the two compared operators for a critical region of 90%, i.e. $1.645 < z_\alpha < -1.645$.

The graphs for the failure rate for all the comparisons can also be found in Appendix C.

Excluding the data for 2004 for both operators for number of failures, and excluding 2006 for Operator 1 and 2009 for Operator 4 for number of tests, gives the following values;

Table 4.1 Data table for Operator 1 vs. Operator 4 Fire detector barrier

n_{1f}	1735	x_{1c}	507
n_{4f}	363	x_{4c}	92
n_{1c}	188189	p_{1c}	0,0027
n_{4c}	21397	p_{4c}	0,0043

This gives the prediction interval of $-0.00762 < (p_{1F} - p_{4F}) < 0.004407$.The interval width is 0,012 which has a considerably narrow prediction interval. A hypothesis tests (with null hypothesis; H_0: $p_{1F}=p_{4F}$, H_1: $p_{1F}<p_{4F}$) reveals that -0,4393 > - 1,645, H_0 is not rejected, there is no significant difference between the two operators.

The same procedure has been performed for the other 5 comparisons, and is listed in the Table 4.1.1 below:

Table 4.1.1 Results for fire detection barrier comparisons from the method.

Lower bound	**Compared operators**	**Upper bound**	**z**	**Interval width**	**%**
-0,0076	$< P_{1f} - P_{4f} <$	0,0044	-0,4393	0,0120	33,02 %
-0,0090	$< P_{3f} - P_{4f} <$	0,0038	-0,6616	0,0128	25,41 %
-0,0048	$< P_{1f} - P_{5f} <$	0,0015	-0,8715	0,0063	19,18 %
-0,0064	$< P_{3f} - P_{5f} <$	0,0013	-1,1072	0,0077	13,41 %
-0,0059	$< P_{4f} - P_{5f} <$	0,0068	0,1108	0,0127	54,41 %
-0,0026	$<P_{1f} - P_{3f} <$	0,0045	0,4212	0,0071	66,32 %

For the fire detection barrier it was found that there was no statistically significant differences between the operators when hypothesis testing with a 90% significance level.

For comparison a 90% confidence interval for the Classical approach is also performed for the barrier, the results are shown in table 4.1.2

Table 4.1.2 Results for fire detection barrier comparisons for the Classical approach

Classical approach					
Lower bound	**Compared operators**	**Upper bound**	**z**	**Interval width**	**%**
-0,0055	$< P_{1f} - P_{4f} <$	0,0010	-1,1414	0,0065	12,69 %
-0,0064	$< P_{3f} - P_{4f} <$	0,0002	-1,5591	0,0066	5,95 %
-0,0027	$< P_{1f} - P_{5f} <$	0,0011	-0,6793	0,0039	24,85 %
-0,0036	$< P_{3f} - P_{5f} <$	0,0003	-1,3734	0,0040	8,48 %
-0,0020	$< P_{4f} - P_{5f} <$	0,0049	0,7012	0,0069	75,84 %
-0,0008	$< P_{1f} - P_{3f} <$	0,0025	0,8549	0,0033	80,37 %

The results show that that there was no statistically significant differences between the operators when hypothesis testing with a 90% confidence interval.

4.2 Start test fire pumps barrier

The results for the start test for fire pumps barrier is shown in table 4.2.1. The z value shows that none of the operators has a statistically significant lower failure rate than any of the others, as all of the z values are within the 90% prediction interval; $-1.645 < z_{\alpha} < 1.645$.

4.2.1 Results for the Start test barrier comparison.

Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0137	$< P_{1f} - P_{4f} <$	0,0082	-0,4108	0,0219	34,06 %
-0,0072	$< P_{1f} - P_{3f} <$	0,0083	0,1140	0,0155	54,54 %
-0,0095	$< P_{1f} - P_{5f} <$	0,0078	-0,1638	0,0173	43,49 %
-0,0099	$< P_{4f} - P_{5f} <$	0,0136	0,2623	0,0235	60,34 %
-0,0108	$< P_{3f} - P_{5f} <$	0,0131	0,1635	0,0239	56,49 %
-0,0144	$< P_{3f} - P_{4f} <$	0,0130	-0,0825	0,0274	46,71 %

For the Classical approach the results are shown in table 4.2.2.

Table 4.2.2 Results for the start test barrier comparison for the Classical approach.

Classical approach					
Lower bound	**Compared operators**	**Upper bound**	**z**	**Interval width**	**%**
-0,0080	$< P_{1f} - P_{4f} <$	0,0025	-0,8576	0,0105	19,56 %
-0,0030	$< P_{1f} - P_{3f} <$	0,0041	0,2462	0,0072	59,72 %
-0,0047	$< P_{1f} - P_{5f} <$	0,0030	-0,3667	0,0077	35,69 %
-0,0038	$< P_{4f} - P_{5f} <$	0,0075	0,5464	0,0113	70,76 %
-0,0045	$< P_{3f} - P_{5f} <$	0,0069	0,3424	0,0114	63,40 %
-0,0074	$< P_{3f} - P_{4f} <$	0,0060	-0,1684	0,0134	43,31 %

The z value for the Classical approach doesn't show that any of the operators has a statistically significant lower failure rate than any of the others

4.3 BOP barrier

The BOP stands for Blow out preventer and is a large specialized valve used to seal, control and monitor the fluids in the well that is being drilled. For the BOP barrier not all of the barriers met the requirements for the number of tests and therefore only two operators were compared. For this comparison it was found that Operator 1 has a statistically significant lower failure rate than operator 4 as $-2,5098 < -1,645$.

Table 4.3.1 Results for the BOP barrier comparison.

Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,06915	$< P_{1f} - P_{4f} <$	-0,0144	**-2,5098**	0,0835	0,60 %

For the Classical approach the results are shown in table 4.3.2.

Table 4.3.2. Results for the BOP comparison for the Classical approach.

Classical approach					
Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0581	< P_{1f}-P_{4f} <	-0,0254	**-4,2012**	0,0835	0,00 %
-0,0746	< P_{1f}-P_{3f} <	0,0076	-1,3406	0,0822	9,00 %
-0,1006	< P_{1f}-P_{5f} <	-0,0237	**-2,6617**	0,1243	0,39 %
-0,0522	< P_{3f}-P_{4f} <	0,0357	-0,3103	0,0166	37,82 %
-0,0951	< P_{3f}-P_{5f} <	0,0201	-1,0709	0,1152	14,21 %
-0,0728	< P_{4f}-P_{5f} <	0,0143	-1,1032	0,0871	13,50 %

For the Classical approach two comparisons were found to have a statistically significant difference; Operator 1 has a statistically significant lower failure rate than Operator 4 and Operator 5.

4.4 Muster barrier

Mustering time is the time it takes for all the people on board an installation to gather at the lifeboats or muster station when then general alarm is sound. The barrier for mustering time did not include all operators as some didn't meet the criteria for number of tests. Therefor only 3 operators were compared. Among these operators it was found that operator 3 had a statistically significant lower failure rate than operator 1 as 2, 1645 > 1,645.

Table 4.4.1 Results for mustering barrier comparison.

Lower bound	Compared Operators	Upper bound	z	Interval Width	%
-0,0842	< P_{1f} - P_{4f} <	0,1671	0,5428	0,2512	70,64 %
0,0298	< P_{1f} - P_{3f} <	0,2182	**2,1645**	0,1884	98,48 %
-0,1842	< P_{3f} - P_{4f} <	0,0192	-1,3351	0,2033	9,09 %

For the Classical approach the results are show in Table 4.4.2.

Table 4.4.2 Results for the muster barrier comparison for the Classical approach.

Classical approach					
Lower bound	Compared Operators	Upper bound	z	Interval Width	%
-0,0269	< P_{1f} - P_{4f} <	0,1098	0,9981	0,1366	84,09 %
0,0689	< P_{1f} - P_{3f} <	0,1791	**3,6999**	0,1102	99,99 %
-0,1313	< P_{3f} - P_{4f} <	-0,0338	**-2,7846**	0,1650	0,27 %
-0,1267	< P_{1f} - P_{5f} <	0,0634	-0,5476	0,1901	29,20 %
-0,2377	< P_{3f} - P_{5f} <	-0,0735	**-3,1168**	0,3112	0,09 %
-0,1646	< P_{4f} - P_{5f} <	0,0184	-1,3139	0,1830	9,44 %

The results show that there are 3 comparisons that gives a statistically significant difference; Operator 3 has a statistically significant lower failure rate compared to the 3 other operator companies.

Also for the comparison of operator 3 and 4 the Classical approach results in a significantly lower failure rate than the developed method. From the other comparisons it is possible to see that the Classical approach has a tendency to result in more exaggerated z-values than the developed method. A reason for this could be that the Classical approach does not restrict data like the developed method and only relies on previous data to calculate the confidence interval. The developed method however, updates its knowledge by including a criterion on how many tests needs to be performed next period.

4.5 Christmas tree barrier

A Christmas tree is an assembly of valves, spools and fittings that directs and controls the flow of formation fluid for completed wells. Operator 1, 3 and 5 were the operators that met the number of tests criteria for the next period and none are shown to have statistical significant lower failure rates than the other as all z values lie within the $-1.645 < z_\alpha < 1.645$ interval.

Table 4.5.1 Results Christmas tree barrier comparison.

Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0045	$< P_{1f} - P_{3f} <$	0,0115	0,7212	0,0160	76,46
-0,0111	$< P_{1f} - P_{5f} <$	0,0058	-0,5163	0,0169	30,28
-0,0142	$< P_{3f} - P_{5f} <$	0,0019	-1,2532	0,0162	10,51

The result for the Classical approach is shown in table 4.5.2.

Table 4.5.2 Results for the Christmas tree barrier comparison for the Classical approach.

Classical approach					
Lower bound	**Compared operators**	**Upper bound**	**z**	**Interval width**	**%**
-0,0014	$< P_{1f} - P_{3f} <$	0,0084	1,1825	0,0097	88,15 %
-0,0095	$< P_{1f} - P_{5f} <$	0,0042	-0,6404	0,0136	26,10 %
-0,0123	$< P_{3f} - P_{5f} <$	-0,0001	**-1,6627**	0,0123	4,82 %
-0,0147	$< P_{1f} - P_{4f} <$	0,0019	-1,2723	0,0166	10,16 %
-0,0176	$< P_{3f} - P_{4f} <$	-0,0022	**-2,1194**	0,0198	1,70 %
-0,0053	$< P_{4f} - P_{5f} <$	0,0128	0,6810	0,0181	75,21 %

The results show that for 2 of the comparisons there is a statistically significant difference in the failure rate; Operator 3 has a statistically significant lower failure rate compared to operator 5 and operator 4. Here, as in the mustering barrier, it is possible to see that for the comparison of Operator 3 and 5, the classical approach has found a significant lower failure rate where the developed method has not. The same explanation can be used; the Classical approach is less restricted and therefore results in more exaggerated z-value.

4.6 DHSV barrier

A DownHole Safety Valve is a valve that acts like a failsafe to prevent unwanted fluids to reach the surface. For the DHSV barrier only two operators met the criteria for number of tests required for the next period; operator 1 and operator 3. It was found that operator 3 had a statistically significant lower failure rate than operator 1 with 1,645 < 1,6796.

Table 4.6.1 Results for DHSV barrier comparison.

Lower bound	Compared operator	Upper bound	z	Interval width	%
0,0003	$<P_{1f}-P_{3f}<$	0,0317	**1,6796**	0,0320	95,35 %

The result for the Classical approach is shown in table 4.5.2.

Table 4.5.2 Results for the DHSV barrier for the Classical approach.

Classical approach					
Lower bound	**Compared operator**	**Upper bound**	**z**	**Interval width**	**%**
0,0072	$< P_{1f} - P_{3f} <$	0,0248	**2,9869**	0,0320	99,86 %
-0,0568	$< P_{1f} - P_{4f} <$	-0,0148	**-2,8062**	0,0717	0,25 %
-0,0223	$< P_{1f} - P_{5f} <$	0,0094	-0,6706	0,0317	25,12 %
-0,0715	$< P_{3f} - P_{4f} <$	-0,0322	**-4,3428**	0,1037	0,00 %
-0,0365	$< P_{3f} - P_{5f} <$	-0,0085	**-2,6413**	0,0450	0,41 %
0,0057	$< P_{4f} - P_{5f} <$	0,0530	**2,0423**	0,0587	97,94 %

The results show that for 5 of the comparisons there is a statistically significant difference in the failure rate; Operator 3 has a statistically significant lower failure rate compared to the three other operators, Operator 1 has a statistically significant lower failure rate than Operator 4, and Operator 4 has a statistically significant lower failure rate than Operator 5.

4.7 ESDV barrier

The Emergency shutdown valve is a valve that stops the flow of hazardous fluid when a dangerous event has occurred on an installation. For the ESDV barrier, only operator 1 and operator 3 met criteria for the number of tests for next period, but no statistical significant differences was found.

Table 4.7.1 Results for ESDV barrier comparison.

Lower bound	Compare operators	Upper bound	z	Interval width	%
-0,0094	$< P_{1f} - P_{3f} <$	0,0228	0,6816	0,0322	75,22 %

The result for the Classical approach is shown in table 4.7.2.

Table 4.7.2 Results for the ESDV barrier comparison for the Classical approach.

Classical approach					
Lower bound	**Compare operators**	**Upper bound**	**z**	**Interval width**	**%**
-0,0039	$< P_{1f} - P_{3f} <$	0,0173	1,0362	0,0212	85,00 %
-0,0592	$< P_{1f} - P_{4f} <$	0,0348	-0,4273	0,0940	33,46 %
-0,0912	$< P_{1f} - P_{5f} <$	0,0235	-0,9706	0,1147	16,59 %
-0,0648	$< P_{3f} - P_{4f} <$	0,0270	-0,6767	0,0918	24,93 %
-0,0969	$< P_{3f} - P_{5f} <$	0,0159	-1,1807	0,1128	11,89 %
-0,0943	$< P_{4f} - P_{5f} <$	0,0510	-0,4894	0,1453	31,23 %

The results show that for the Classical approach none of the operators had a statistically significant lower failure rate than any of the others.

5. Discussion

In this section the results, methods and influencing factors are discussed.

5.1 Discussion of the results

The results from the method are dependent on many uncertain factors such as; design, test culture, activity level and maintenance program. This makes it difficult to say for certain that one operator is better than another because they may have different design, different test culture, and different activity level and different maintenance program. This is why it is interesting to develop a method that can give an indication of the results these differences contribute to in the barrier performance for the operators.

The results in chapter 4 show the comparisons in the developed method (which follow the predictive Bayesian paradigm) and comparisons in the Classical approach. In the developed method not all operators have been compared because some of the operators didn't satisfy the requirement for the number of tests for next period, but in the Classical approach all operators have been compared with each other.

The results for the developed method show that there were three barriers that showed statistically significant lower failure rate for one of the operators.

Below some of the statistically significant differences are shown.

Table 5.1.1: The BOP comparisons shown for the developed method and the Classical approach.

Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,06915	$< P_{1f} - P_{4f} <$	-0,0144	-2,5098	0,0835	0,60 %
Classical approach					
Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0581	$<P_{1f}-P_{4f}<$	-0,0254	-4,2012	0,0835	0,00 %

Table 5.1.2: The Muster time comparisons shown for the developed method and the Classical approach.

Lower bound	Compared Operators	Upper bound	z	Interval Width	%
0,0298	$< P_{1f} - P_{3f} <$	0,2182	2,1645	0,1884	98,48 %
-0,1842	$< P_{3f} - P_{4f} <$	0,0192	-1,3351	0,2033	9,09 %
Classical approach					
Lower bound	Compared Operators	Upper bound	z	Interval Width	%
0,0689	$< P_{1f} - P_{3f} <$	0,1791	3,6999	0,1102	99,99 %
-0,1313	$< P_{3f} - P_{4f} <$	-0,0338	-2,7846	0,1650	0,27 %

Table 5.1.3: The Christmas tree comparisons shown for the developed method and the Classical approach.

Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0142	< $P_{3f} – P_{5f}$ <	0,0019	-1,2532	0,0162	10,51
Classical approach					
Lower bound	Compared operators	Upper bound	z	Interval width	%
-0,0123	< P_{3f} - P_{5f} <	-0,0001	-1,6627	0,0123	4,82 %

Table 5.1.4: The DHSV comparisons shown for the developed method and the Classical approach.

Lower bound	Compared operator	Upper bound	z	Interval width	%
0,0003	< P_{1f} - P_{3f} <	0,0317	1,6796	0,0320	95,35 %
Classical approach					
Lower bound	Compared operator	Upper bound	z	Interval width	%
0,0072	< P_{1f} -P_{3f} <	0,0248	2,9869	0,0320	99,86 %

- For the BOP barrier; Operator 1 was found to have statistically significant lower failure rate than Operator 4.
- For the mustering barrier; Operator 3 was found to have a statistically significant lower failure rate than Operator 1.
- For the DHSV barrier; operator 3 was found to have a statistically significant lower failure rate than operator 1.

In the Classical approach there were 12 comparisons that showed to have statistically significant lower failure for four barriers.

- For the BOP barrier; Operator 1 was found to have a statistically significant lower failure rate for comparison with Operator 4 and Operator 5.
- For the mustering barrier; Operator 3 was found to have a statistically significant lower failure rate compared to the 3 other operator companies.
- For the Christmas tree barrier; Operator 3 was found to have a statistically significant lower failure rate compared to operator 5 and operator 4.
- And lastly for the DHSV barrier; Operator 3 was found to have a statistically significant lower failure rate compared to the three other operators, Operator 1 was found to have a statistically significant lower failure rate than Operator 4, and Operator 4 was found to have a statistically significant lower failure rate than Operator 5.

As mentioned before the developed method didn't compare all operators for each barrier because of the number of test criteria, and therefore the Classical approach shows more comparisons that have resulted in statistically significant differences in failure rates than for the developed method. However, for the three comparisons that show statistically significant different failure rates in the developed method are also shown in the Classical approach, but with lower or higher z-values than in the developed method.

From the table 5.1.1-5.1.4 above there are some differences that need to be commented. The Classical approach shows a stronger difference compared to the developed approach: for example, for the BOP barrier, the z value is -2,5098 for the developed method and for the classical the z value is -4,2012 i.e. the Classical approach shows a lower z value. For the DHSV the z value is 1,6796 for the developed approach and 2,9869 for the Classical approach, showing a higher value for the Classical approach. This is the general trend when comparing the two methods; the Classical approach will show a stronger difference, but in the “same” direction as the developed method. From Appendix D, the relationship between the z-value and the probability percentage for the operators are shown in graphs for each barrier. The general trend shown in these graphs is that the Classical approach will give values that are both lower in value on the negative side, and higher on the positive side, when compared to the developed method. Therefore the graph for the Classical approach occupies a larger interval of the graph than the developed method.

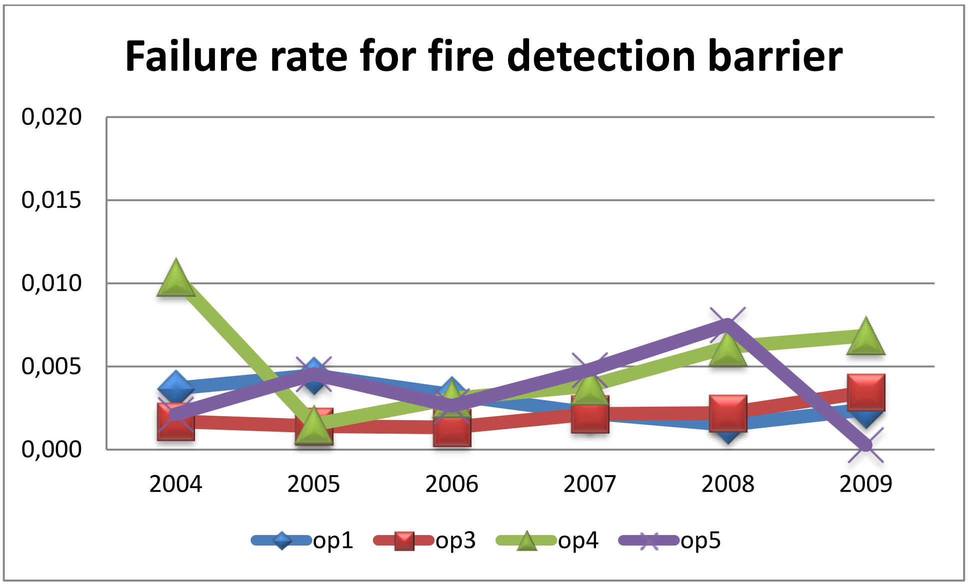
Whether the Classical approach shows a better indication of the quality of the barrier performance due to the “stronger” z-values could be argued for, however it could also mean that the Classical approach overestimates and/or underestimates the barrier performances. As mentioned in the start of this chapter, there are uncertainties that have to be taken into account as well before a definitive conclusion about the methods can be made.

5.1.1 The barriers

The barriers that were used in the method were chosen because they are considered to be some of the most important barriers on an installation, there is at least one barrier represented for each of barrier categories:

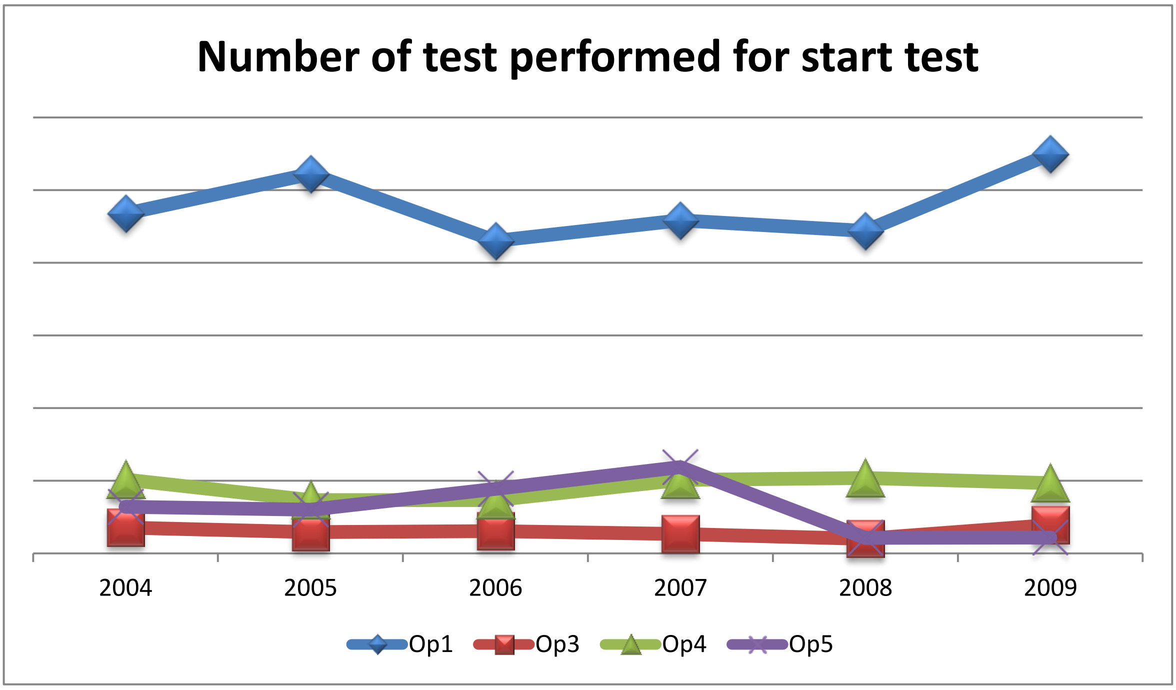
1. Fire and Gas detection,
2. Shutdown,
3. Wellbore isolation,
4. Wellbore isolation w/BOP,
5. Fire safeguard and
6. Emergency preparedness.

The fire detector barrier comes from the first category and is one of the barriers which aren’t expected to show much difference between the operators, as there is usually a standard type of fire detector used for most installations. The fire detectors are usually abundant on installations, where there is at least one in each room. Thus there exists quite a lot of redundancy for this barrier to make sure that the barrier works.



Graph 5.1.1.1 Failure rate for fire detection barrier for the last six years

The **start test for the fire pumps** is a very important barrier because in the case there is a large fire, the fire pumps will help reduce the damage dramatically. The start test for the fire pumps, on most installations, are connected to a computer system, with a manual start button somewhere closer to the fire pumps. In the requirement for barrier testing it is specified that a failure is registered for both procedures. This means that a failure for this barrier could be due to an electrical or computer failure from the computer system or physical damage or deterioration to the motor or the pumps themselves. Unlike most of the other barriers, this one isn't dependent on activity level. As long as the installation is active there should be routine tests performed to make sure that it works. This can been seen also from the graph below;



Graph 5.1.1.1 Number of test performed for start test the last six years.

All the operators have a relatively stable testing procedure for this barrier; independent of the type activity that is being performed.

When calculating the failure rate criteria it was found that the poorest barrier criteria for all of the operators were mustering time, BOP and DHSV. By poorest, it means that their failure rates were the highest. Mustering time and BOP are quite different types of barriers. **Mustering time** depends on all the workers on an installation to act as they should in an emergency situation, who know and remembers what to do. It is strongly dependent on the human factor, such as if a worker takes the mustering drill seriously or not, if they get scared or confused from the scenario that is being simulated, and so on. This barrier is hard to control because it involves all the workers at the same time. Another factor that could affect the mustering time is how strict the time criteria for the operators are. The average mustering time for Operator 1 is: 19 mins, Operator 3: 15 mins, Operator 4: 22,5 mins and Operator 5: 32,5 mins. It is assumed that the times depends on the “slowest” person to reach the life boats, which would mean that the time criteria are estimated by the time it would take for the person that has his/her station the farthest away, to secure their work station and get to the nearest lifeboat. According to the failure rates in Table 3.3.1 the operator with the lowest failure rate is Operator 3, which also coincidently has the fastest mustering time requirement. The operator with the highest failure rate is Operator 5, which also has the slowest mustering time requirement. This could mean that there are fewer people on the installations of Operator 3 than Operator 5, or that the installations are smaller than Operator 5. It could also mean that there is more focus on the mustering time on the installations of Operator 3 than Operator 5, in the form of incentives or cultural norms that have been cultivated by the operator, therefore improving the affect the human factor has.

The BOP however is a mechanical barrier that is in use most of the time when there is drilling activities. Its main purpose is to seal, control and monitor the well to prevent any unwanted fluids or gases to reach the atmosphere. The state of the BOP can be monitored by regular maintenance and BOP tests to ensure the quality of the barrier. The number of BOP tests that is required to be performed is dependent on how many valves are needed to be pressure tested. This can range from 4 to 10 dependent on how complicated the BOP is. In some cases a BOP test can take up to 24 hours and the company man on the installation has to consider the costs of having to take “too many” tests to ensure the quality. Another important point to mention is that when testing BOP’s, it would be misleading to say that one installation that drills few new wells a year (like 3) has a less accurate failure rate than another installation that drills many new wells a year (like 10). It depends on how many sections the well has and how long or deep the well is. The BOP is a barrier that is very dependent on the activity level. When there is little activity, there is less deterioration, and so the BOP will be tested less. What deteriorated the BOP is mainly the sand or solids that become part of the drilling fluid when drilling. It grinds and polished the pipe and valves which could compromise the integrity of the sealing function of the BOP. This is why it is important to have routine BOP tests according to the drilling time. It would also depend on the drilling parameters – more vigorous drilling could lead to faster deterioration.

The same arguments used for the BOP can be used for **Christmas trees.** One installation might have more producing wells than another and therefore the one that has more wells will test more than the other. Also the production rate can be a factor in the deterioration of the Christmas tree. For a formation that has high pressure, water and sand problems, these Christmas tree might be tested more than other Christmas trees which produce from "problem free" formations.

The **DHSV**'s frequency of testing should be determined on the problems of the wells, like scaling or water intrusion which are the most common and serious problems. This barrier, like the Christmas tree, is dependent on the formation it is producing from. Together with the Christmas tree, these are the two barriers that have to be approved in order to produce from the well. This is why it is so important that they have strict failure rate criteria and regular maintenance and testing. For wells that have a lot of problems, the DHSV is tested very often and if there is a test failure the DHSV is replaced with a new one.

ESDV are in place to isolate the hydrocarbon containing sections on an installation. ESDV's are mounted on the pipelines in order to isolate these and prevent hydrocarbons from these to sustain a fire on the platforms. This is one of the barriers that contribute when calculating the FAR values for an installation, and has much focus in HAZOP's. These valves can be very large, and is very specialized, takes a long time to manufacture, and therefore it is very difficult to get a new one in a hurry if something were to go wrong with it. A test failure for this barrier is dependent on leakage rate, when the closing time is too short (giving a hammer shock) or to slow. To fix this problem the closing time is adjusted to the correct closing time interval. When the leak rate is too high, it means that the barrier has too much wear and is deteriorating. In the case where there is too much deterioration for the ESDV to be an effective barrier, a new one needs to be ordered if repairs cannot bring it to an acceptable condition. This doesn't happen very often as the ESDV normally is very robust and is designed for a long life time.

5.1.2 Test culture

The test culture on an installation is a good indicator on how well the barriers are performing. An installation that tests its barriers often would automatically have a good record of their barriers and if a barrier is performing poorly it can be rectified before any incidents occur.

However testing a barrier often is quite time consuming and in the petroleum industry, time is money. The operators have to decide how much time they will spend on a barrier to ensure that it is efficient and at the same time be cost efficient. For barrier tests that are very time consuming, the time-cost-efficiency can be hard to determine. For example, when analyzing the data to calculate realistic failure rate criteria, it was observed that some of the barriers, like BOP, have very different test cultures for the different installations. Some installation had an extraordinary amount of tests being performed and on others very few were performed.

When looking for other explanation for this, it was found according to the RNNP report of 2010 (page 3), that there exists a lot of uncertainty around the data collected for BOP's. There could be many reasons for these uncertainties in data collecting; some installations may only report one test when they have tested the whole BOP, while other installation considers each value as one test. There are some installations that aren't having much drilling activity one year and will therefore not test the BOP as many times as the previous year and could be misunderstood as the amount of testing has gone down for this installation. The same is the case for DHSV. When testing the DHSV, a coil has to be run in hole to get to the DHSV to be tested and sometimes the coil can get stuck and has to be pulled out of hole. Should this count as one test? The DSVH has actually not been tested because of the restriction, but it has been attempted to test it. So after several tries the coil reached the DHSV and the DHSV is tested and it might be counted as one test.

Even so, the number of tests that are registered performed, should be an indication of what is actually being performed. The testing culture and their misunderstandings should be quite constant each year so the barrier data can be considered to be useful in the sense of indicating the state of the barriers.

According to some experts, the PSV tests are difficult barriers to compare on an operator scale because they are tested differently. This is why there should be a common census on how much overpressure is required to get a good test for all the operators. But as there are different designs and different options on what should result in a good test, these differences may not be so easy to change.

One observation that can be seen when looking at the data, is when an installation doesn't have a drilling rig on it and needs assistance to drill new wells or to drill re-entry wells, the BOP tests aren't registered for that operator but might be registered for the jack-up rig that is placed over the installation to drill the re-entry wells. As the responsibility for the activities on the rig lies with the operator who is renting the rig, this data isn't registered in the correct manner. This wouldn't pose a problem if all these jack up rigs data were registered for the operator and available to be analyzed, but in some cases they are not. This doesn't give the correct picture of the operator's procedures and requirements for the BOP barrier.

5.1.3 Activity level

The activity level on an installation is crucial to include when the goal is to find the performance of a barrier. The first thing that has to be considered is what type of activity is being done. In this thesis the activities are divided into 3 categories: drilling, production and maintenance. The barriers consider different type of activity so it is important for the explanation and understanding of the test frequency to know what activity this installation is concerned with this year.

As some of the barriers, that is be used in the method, are in connection with drilling activities, like BOP, the activity level should have been in the form of how many wells are being drilled that year. Christmas tree or DHSV are production related barriers and will be in the form of how many active wells the installation has.

Some barriers are tested more often because of problems related to that particular barrier, while others are seldom tested are working just fine. Examples of critical barriers that needs to be tested regularly are DHSV, Christmas trees, ESDV and BOPs. These barriers ensure the safety of the main operations on the installations and needs to be functioning all the time.

The fact that the activity level for each installation isn't registered in a common database can lead to misunderstandings when considering barriers that are registered.

5.1.4 Maintenance

Maintenance can be categorized into two types; corrective maintenance and preventive maintenance. Corrective maintenance is done on the barrier when it has failed and needs to be restored to the accepted functioning state again. Preventive maintenance is maintenance done in planned periodical intervals and before the system fails to prevent that the system doesn't fail or breakdown. The goal is to preserve and enhance the quality of the system so that the lifespan of the system will at least last as long as it was designed for.

From the RNNP report of 2010, it is also mentioned that several operators have challenges with maintaining an expected level of maintenance management in the light of the regulations. Barriers that don't have sufficient preventive maintenance can be shown to require corrective maintenance more often because the barrier hasn't been taken care of as it was designed for. The number of barrier failures and corresponding corrective maintenance can therefore also be used as an indicator to measure the effectiveness of the preventive maintenance program

5.2 The method

One of the most difficult assertions made in the method is the approximate normal distribution assumptions for the developed method. As the "normal approximation to the binomial distribution" (Walpole et al. 8th Ed.) states; if X (here meaning the barrier data) is a binomial random variable with parameters n and p, the X has approximately a normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$. The approximation is good if np and n(1-p) are greater or equal to five. In the pred. Bayesian paradigm the probability that is of interest is the one that lays in the future, which could be comparable to the Classical approaches' "true" probability. As this thesis is comparing these two types of probability, it is seen as reasonable to also integrate this "comparison" in developing the method. It is therefore considered that the approximation to normal distribution is a valid argument.

Continuing looking at the differences in the Classical approach and the developed method, there were some interesting results that were observed. In table 4.4.1 and 4.4.2, a distinction can be seen between the developed method and the Classical approach. From the comparisons in the Classical approach the results show exaggerated z-values, to the extent that a comparison in the Classical approach resulted in a statistically significant lower failure rate compared to the developed method. Overall the Classical approach has more exaggerated z-values than the developed method. In Appendix D, the relationship between the z-value and the probability percentages that a specific operator has a lower failure rate then another are shown. The graphs show that there is a trend, where the Classical approach always has more extreme values compared to the developed method, but both methods show indications in the same direction. This means that where the developed method has found that Operator 1 has a lower failure rate (like 10 %); the Classical Approach has found that Operator 1 has a failure rate that is lower (like 6 %) than what the developed method found.

A possible reason for this trend could be that the Classical approach does not restrict data like the developed method and only relies on previous data to calculate the confidence interval. The developed method however, updates its knowledge by including a criterion on how many tests needs to be performed next period, and therefore has "less" data included in the prediction interval.

When it comes to the prediction intervals, there are some figures that need elaborated comments. Figure 2.2.1.2 shows the prediction interval for Operator 1 and Operator 4. It is interesting to notice that the interval is much narrower for Operator 1 than for Operator 4. This could be explained by the fact that Operator 1 performed much more tests than Operator 4, and therefore has a smaller error in the failure rate than Operator 4.To be able to compare the Operator companies on the same standard, it was important to develop the formula for the number of required tests to be performed for the next period, as has been done in chapter 3.3 and can be seen in the result tables in chapter 4. An example is the start test for the fire pumps, table 4.2.1-4.2.2, where the interval widths are very similar, which means that their possible errors in failure rates are comparable to each other.

The criteria for the failure rates of the barriers that have been tested are only educated guesses as information of the operators' test philosophy and acceptance criteria for the selected barriers were not available. Some of the failure rate criteria have been reduced more than the estimated criteria showed, because the failure rates found in the data were too high. If the failure rate criteria are too high it may be necessary to perform much more test than an installation may be able to take per year. One example is mustering time, where the failure rate was between 14 and 20% at "best" for some of the operators. A barrier such as mustering time is an important barrier and it is almost certain that the operators have tried and is still trying to develop new routines and encourage the workers on the installation to improve the performance and obtain compliance with the requirement.

Ideally, to make criteria for the failure rates for the barriers and determining how many tests the operators should be performing, careful consideration should be given for each barrier with their respective activity level and consequence of failure for the respective installation. Out of the four uncertainty variables mentioned earlier, the activity level is probably the most unbiased variable because it is a recording of what has happened, the facts. When using the data to test the method, activity level wasn't available but could make the results look very different if normalized according to their activity level. However if it is assumed that all the operators have installations with little and high activity, then they could be considered to be at approximately the same average activity level. It would then be possible to calculate, on operator level, the failure rates and number of tests that should be performed for each of their installations.

5.3 Lessons learned from major accidents

The petroleum industry has experienced many major accidents that have taken the lives of hundreds of people during the last 40 years or so. One of the most famous major accidents is the Piper Alpha accident that happened on June 6th 1988, where 167 people lost their lives. The Piper Alpha accident was one of the worst offshore oil disasters at the time, in terms of lost lives and industry impact, and started a movement that resulted in the rewriting and renewal of the safety procedures for the English petroleum authority (Vinnem, 2007).

In recent years there have been at least 4 incidents that have caught the public's eyes;

- December 12th 2007 – Statfjord A (Statoil). 4 000 m^3 of crude oil was released during an offshore loading of the tanker Navion Britannica.

- May 24th 2008 – Statfjord A (Statoil). An oil leak was discovered in one of the shafts of the platform and most of the people were evacuated.
- August 21st 2009 – West Atlas jackup (Seadrill/PTTEP). 4 500 – 34 000 m^3 Oil and gas leak into the Timor Sea (shallow water).
- April 20th 2010 – Deepwater Horizon (BP). Blow out that caused explosion and one of the largest oil leaks in history (deep water).

There are also other accidents which aren't mentioned here that have helped motivated the PSA to investigate and publish recommendations for the petroleum industry to reduce the risks of loss of life, loss of property and pollution of the environment.

As most accidents that occur in the petroleum industry are so divers in nature, it would seem natural that they all are representative to situations that could occur in any country. From PSA's report on the Deepwater Horizon accident (PSA, 2010), reviews and recommendations are outlined so that the Norwegian petroleum industry can learn from the mistakes done on DwH (Deepwater Horizon). The report highlights that more attention needs to be given to the continued development to improve the barrier performance and barrier control, with respect to the level of ambition for the performance criteria, for testing, maintenance and monitoring of all types of barriers.

In the DwH report the BOP has been given extra attention even though it wasn't the triggering failure that caused the blow out, but it was a big contributor to the disastrous outcome of it. When using the barrier data to compare the operator's barrier performance of the BOP, it was found that there is a lot of inconsistency with how the tests are being performed and how often they are tested, as previously stated. From the RNNP 2010 it is also mentioned that there was a negative trend for the BOP, where less tests were reported and a higher percentage of test failure was observed. The BOP is such an important barrier and PSA has expressed that more attention should be given to make sure that all operators do the best they can to make sure that it is kept in optimal condition at all times.

The DwH report also mentioned that there were challenges related to the safety culture onboard the DwH. This has already been analyzed and presented the regression analysis in chapter 2.5; where the conclusion was that there is a correlation between the safety culture and the frequency of hydrocarbon leakages on an installation. From the RNNP report of 2010 it was also mentioned that over the last three years there is a negative trend in reported hydrocarbon leaks on production facilities offshore. The fact that a regression analysis was performed shows that the industry is actively looking for answers.

Lastly the DwH report mentions several times that the whole industry and all its participants from all stages must come together and contribute in reducing the risks together. This is the motivation that lies behind making a method to compare operator companies' barriers so that it is possible to learn from each other and increase the level of quality of all the barriers.

History has shown that the neglect of physical and organizational barriers is the reasons why major accidents occur. Also there is often more than one barrier failure that leads to a major accident. If the Deepwater Horizon accident should teach the petroleum industry anything, it is that all the barriers that are set in place to ensure the safety of the people on board and the quality of the activity done, has to be given continuous attention.

6 Conclusion

By using the values and views of the predictive Bayesian approach, a method has been developed to conclude whether an operator will have a statistically significant different failure rate than another. The method includes two formulas; one formula for the prediction interval, which will be used to compare the operators' failure rates, and the second formula for making a criterion to define the number of tests required to satisfy an accepted error for the failure rates.

Two operators are compared at a time to investigate if one of these operators will have a statistically significant lower failure rate than the other. In order to have a statistically significant lower failure rate, the z-value for operator's barrier must lie outside the critical region of 90%, i.e. outside the interval $1.645 < z_{\alpha} < -1.645$.

The developed method found three comparisons that showed one operator having statistically significant lower failure rate then another. The results show that:

- For the BOP barrier, Operator 1 has a statistically significant lower failure rate than Operator 4.
- For the mustering time barrier, Operator 3 has a statistically significant lower failure rate than Operator 1.
- And for DHSV barrier – Operator 3 has a statistically significant lower failure rate than Operator 1.

The Classical approach found twelve comparisons where an operator had a statistically significant lower failure rate then another.

- For the BOP barrier; Operator 1 was found to have a statistically significant lower failure rate for comparison with Operator 4 and Operator 5.
- For the mustering barrier; Operator 3 was found to have a statistically significant lower failure rate compared to the 3 other operator companies.
- For the Christmas tree barrier; Operator 3 was found to have a statistically significant lower failure rate compared to operator 5 and operator 4.
- And lastly for the DHSV barrier; Operator 3 was found to have a statistically significant lower failure rate compared to the three other operators, Operator 1 was found to have a statistically significant lower failure rate than Operator 4, and Operator 4 was found to have a statistically significant lower failure rate than Operator 5.

Operator 3 was found to have the most statistically significant lower failure rates out of the comparisons that were made.

The results show that there are operators with statistically significant lower failure rate than others. In this case the interesting question is: what makes one operator better than another? It is most likely because of the different combination of the design, test culture, activity level and maintenance program. The operators that were found to have statistically significant lower failure rates for certain barriers have most likely found the right combination of these variables to manage the barriers in a safer way.

The developed method, which is based on the predictive Bayesian approach, is compared to the Classical approach of comparing two proportions. In the Classical approach, the probability or “chance” of an event is relative frequency based; while in the predictive Bayesian approach the probability is defined as the assessors’ assigned uncertainty about an activity (whether it will occur) and the severity of the consequences of the activity, based on the available knowledge (A,C,U perspective).

Because of this difference, the Classical approach and predictive Bayesian approach are not actually comparable as they don’t belong to the same set of views and values, i.e. they belong to different paradigms. The reason why the Classical approach and the predictive Bayesian approach is compared in this thesis, is to show the actual differences in results that these two methods produce.

If the two approaches are considered as methods to find the most reliable probability, p, it is possible to compare them, but keeping in mind that they don’t express the same type of probability.

In the predictive Bayesian approach the uncertainty in the data is taken into consideration and is given attention when calculating p. This is because the predictive Bayesian approach is an epistemic approach. This means that when predicting the future, the uncertainty related to the future observables is a result from lack of knowledge.

In the Classical approach the method is considered as objective and therefore the uncertainty that lies in all data is hidden and not considered.

The predictive Bayesian approach uses the previous data as background knowledge to find the predicted value of p. It could be argued that the predictive Bayesian approach is more reliable than the Classical approach because the result from the predictive Bayesian approach tries to describe how the states of the observable quantities are at present or in the future. The Classical approach on the other hand only calculates the probability as it was, because present data about the observable quantities is not available, implicitly excluding the evolution of the events. The uncertainty in the data that is available is also not scrutinized because the Classical approach is considered to be objective. It is therefore considered that the predictive Bayesian approach is a more valid method for measuring barrier performance than the Classical approach.

7 Possible further work

A possible continuation of this work could be the integration of the activity level and/or test culture into the method to see if they have an impact on the outcome or not. In this thesis they have been discussed and highlighted because of their believed importance in the understanding of the barrier performances.

Another possibility that could be investigated is to see if there is a correlation between the barrier performance and the different designs of the barriers, or maintenance programs, or test culture, if this information were to become available. It has been stated several times that it is believed that the barrier performances are connected to how well the test culture and maintenance program is managed in accordance to the activity level and the design of the barriers. If some any of these correlations are found to make the barriers more reliable, then a new best practice can be formulated.

While working on this thesis it has become apparent that these aspects could lead to very different results than what has been obtained. The developed method only gives an indication of which operators have found the correct balance between test culture, maintenance program, activity level and design of the barriers. When these four aspects have been properly investigated in accordance to the barrier performance, a much better understand will have been obtained as how to properly manage the barriers.

8 Abbreviations

RNNP – RisikoNivå Norsk Petroleumsindustri / "Risk level project"
NCS – Norwegian Continental Shelf
POB – People on Board
FAR - Fatal Accidental Rate
PLL - Potential Loss of Life
DFU – Defined situations of hazard and accident
ANOVA – Analysis of Variance
R^2 – Coefficient of multiple determination
$R^2_{adjusted}$ – Adjusted coefficient of multiple determination that provides adjustment for degrees of freedom.
ESDV – Emergency shutdown valve
BDV – Blowdown valve
PSV – Pressure safety valve/ pressure relief valve.
DHSV – Down Hole safety valve
HAZOP – Hazard and Operational study

9 Appendix

Appendix A

NORSOK standard D-010 Rev 3, August 2004

Annex A
(Normative)
Leak test pressures and frequency for well control equipment

Table A.1 - Routine leak testing of drilling BOP and well control equipment

	Frequency / Element	Stump	Before drilling out of casing		Before well testing	Periodic		
			Surface	Deeper casing and liners		Weekly	Each 14 days	Each 6 months
BOP	Annulars	MWDP 1)	Function	MSDP 1)	TSTP 1)	Function	MSDP 1)	WP x 0,7
	Pipe rams	MWDP	Function	MSDP	TSTP	Function	MSDP	WP
	Shear rams	MWDP	Function	MSDP	TSTP	Function		WP
	Failsafe valves	MWDP	Function	MSDP 3)	TSTP	Function	MSDP 3)	WP
	Well head connector	MWDP	MSDP		TSTP		MSDP	WP
	Wedge locks	Function						
Choke/kill line and manifold	Choke/kill lines	MWDP	MSDP	MSDP	TSTP		MSDP	WP
	manifold	MWDP	MSDP	MSDP	TSTP		MSDP	WP
	Valves	Function	Function	Function	Function		Function	
	Remote chokes							
Other equipment	Kill pump	WP 2)		MSDP			MSDP	WP
	Inside BOP	MWDP 2)		MSDP	TSTP		MSDP	WP
	Stabbing valves	MWDP 2)		MSDP	TSTP		MSDP	WP
	Upper kelly valve	MWDP 2)		MSDP			MSDP	WP
	Lower kelly valve	MWDP 2)		MSDP			MSDP	WP

Legend

WP	working pressure
MWDP	maximum well design pressure
MSDP	maximum section design pressure
Function	Function testing: testing shall be done from alternating panels/pods.
TSTP	tubing string test pressure
1)	Or maximum 70 % of WP
2)	Or at initial installation
3)	From above if restricted by BOP arrangement

NOTE 1 All tests shall be 1,5 MPa to 2 MPa/5 min and high pressure/10 min.

NOTE 2 If the drilling BOP is disconnected/re-connected or moved between wells without having been disconnected from its control system, the initial leak test of the BOP components can be omitted. The wellhead connector shall be leak tested.

NOTE 3 The BOP with associated valves and other pressure control equipment on the facility shall be subjected to a complete overhaul and shall be recertified every five years. The complete overhaul shall be documented.

Appendix B-1

Operator 1	Observed data				future data
	2006	2007	2008	2009	2010
x	110	88	57	88	...
n	33435	40064	38711	36676	37222
p	0,0033	0,0022	0,0015	0,0024	0

Operator 4	Observed data				future data
	2006	2007	2008	2009	2010
x	13	18	30	25	...
n	4249	4712	4874	3651	4372
p	0,0031	0,0038	0,0062	0,0068	0,0000

	Operator 1	Operator 4
Confidence interval		
p	0,0023	0,0049
n	148886	17486
$Z_{1-\alpha/2}$	1,645	1,645
Lower bound p	0,0021	0,0040
Upper bound p	0,0025	0,0058
Interval width	0,0004	0,0017

	Operator 1	Operator 4
Prediction interval		
p	0,0023	0,0049
x	148886	17486
nf	37222	4372
	0,05	0,05
	0,95	0,95
Lower bound x	71	14
Upper bound x	101	29
Lower bound p	0,0019	0,0032
Upper bound p	0,0027	0,0066
Interval width	0,0008	0,0034

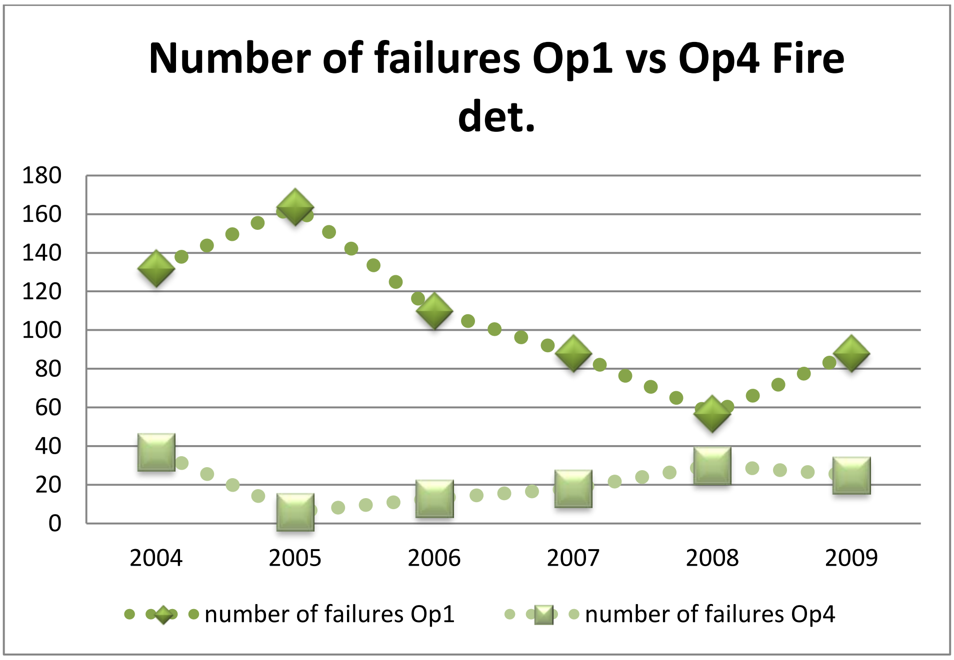
Appendix B-2

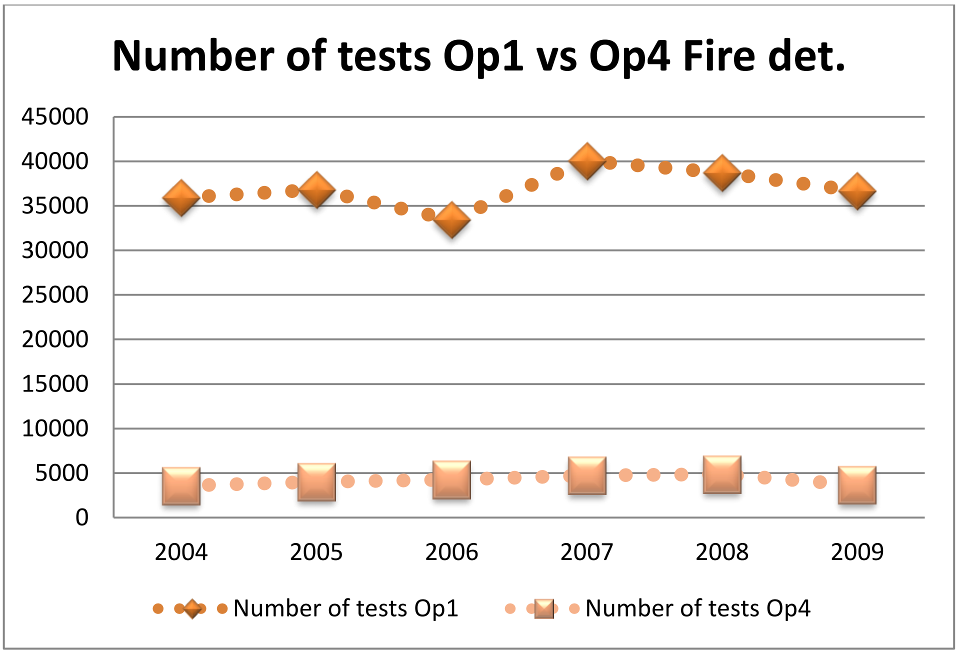
This diagram shows the number of tests, number of failures and the probability for the PSV barrier for the two installations AB and AH.

year	2004		
Installation:	number of test PSV (n)	number of failures (x)	PSV (p)
AB	214	41	0,191588785
AH	97	16	0,164948454
Year	2005		
Installation:	number of test PSV (n)	number of failures	PSV
AB	345	38	0,110144928
AH	121	12	0,099173554
Year	2006		
Installation:	number of test PSV (n)	number of failures	PSV
AB	311	34	0,109324759
AH	149	10	0,067114094
Year	2007		
Installation:	number of test (n)	number of failures	PSV
AB	285	17	0,059649123
AH	148	8	0,054054054
Year	2008		
Installation:	number of test (n)	number of failures, (x)	PSV
AB	311	25	0,080385852
AH	149	9	0,060402685
Year	2009		
Installation:	number of test (n)	number of failures (x)	PSV
AB	219	21	0,095890411
AH	162	10	0,061728395
	sum		
Installation:	number of test for PSV (n)		proportion of failure for PSV
AB	1685	176	0,104451039
AH	826	65	0,078692494

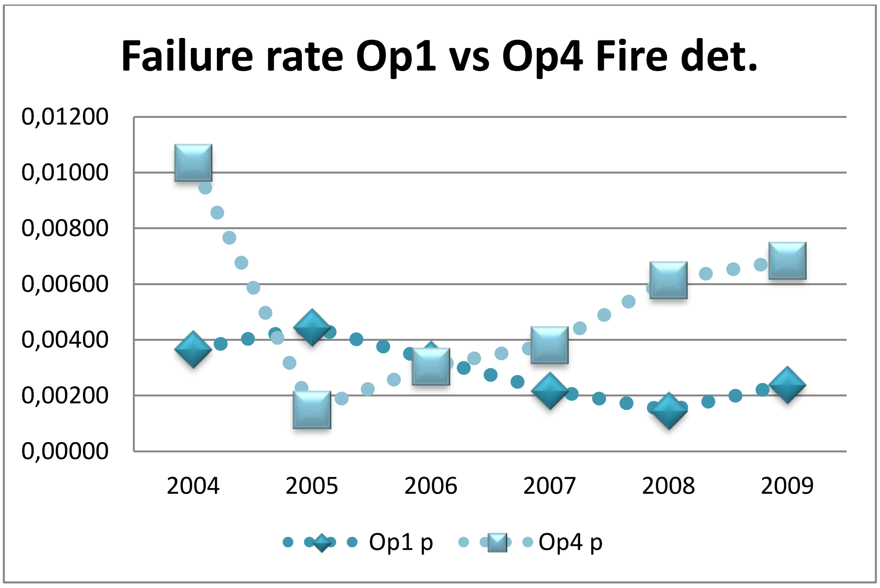
Appendix C

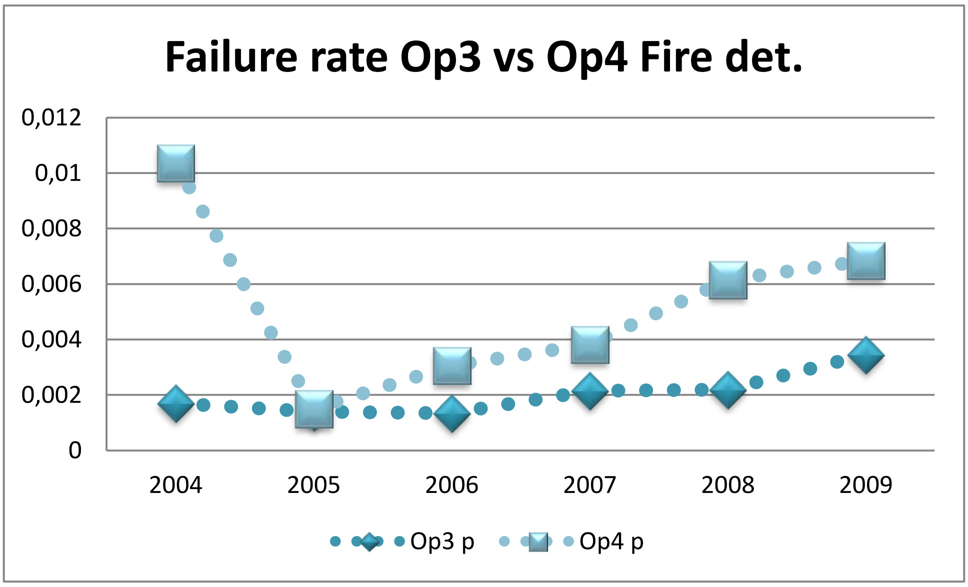
Graphs to determine the general trend as explained in chapter 4.1.

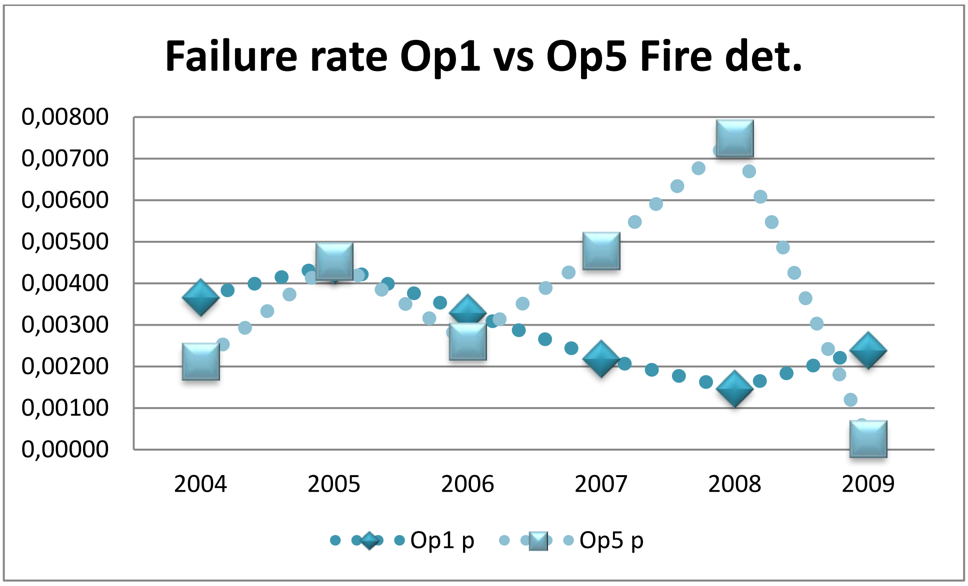


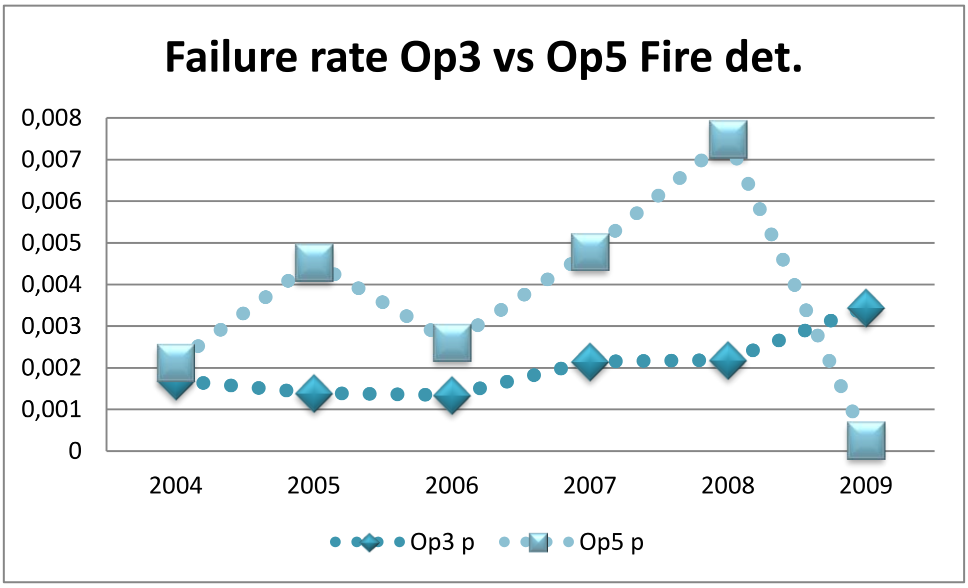


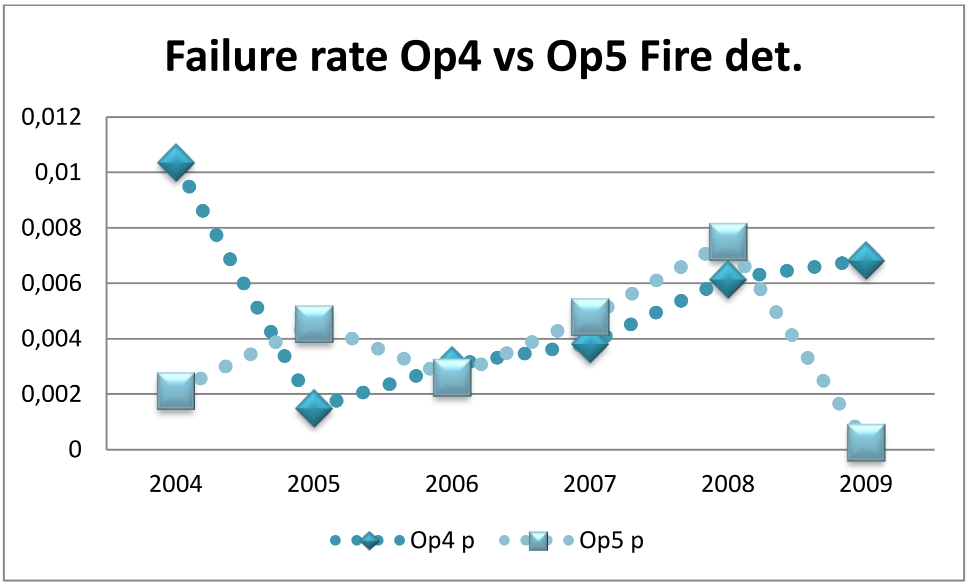
Fire detection barrier:

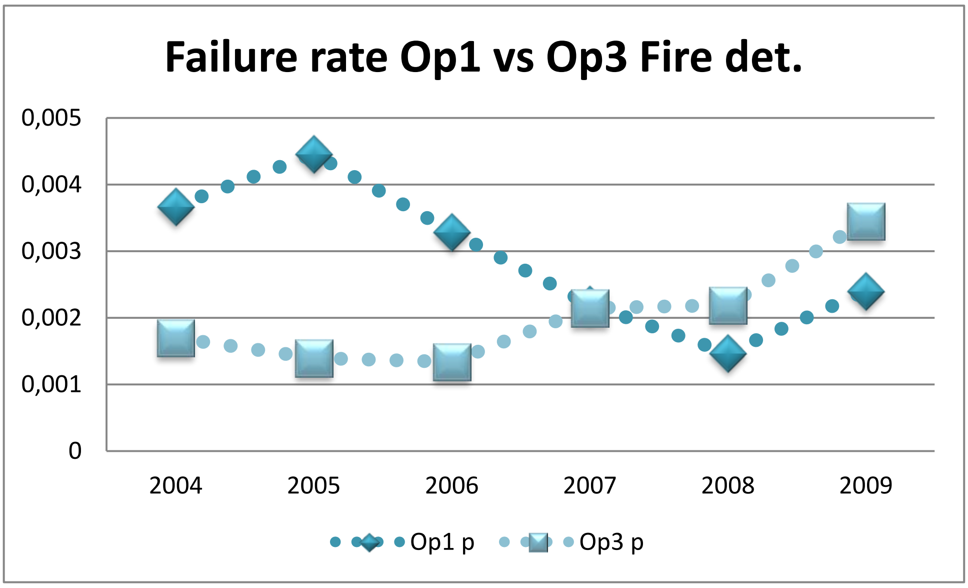




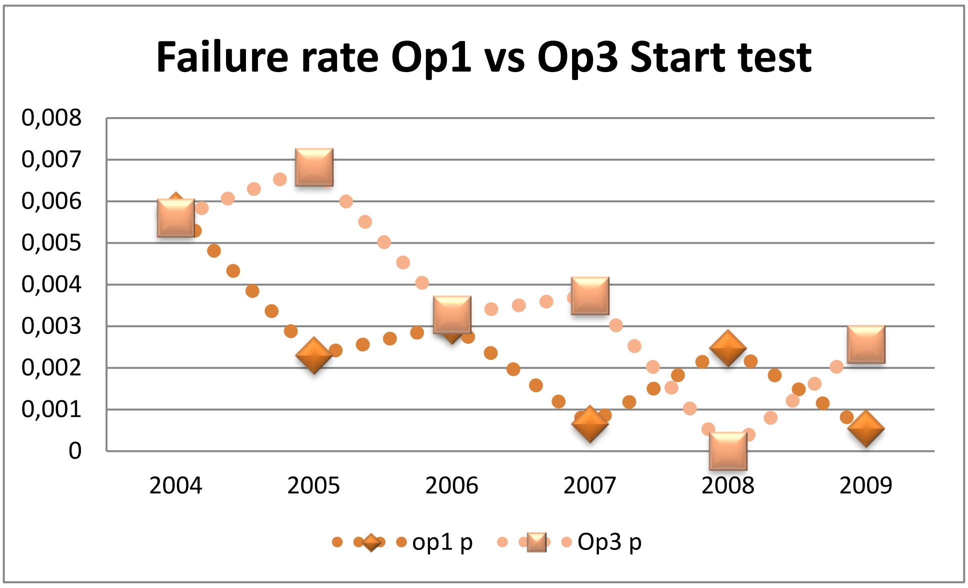

Failure rate Op1 vs Op5 Fire det.
0,00800
0,00700
0,00600
0,00500
0,00400
0,00300
0,00200
0,00100
0,00000
2004
2005
2006
2007
2008
2009
Op1 p
Op5 p

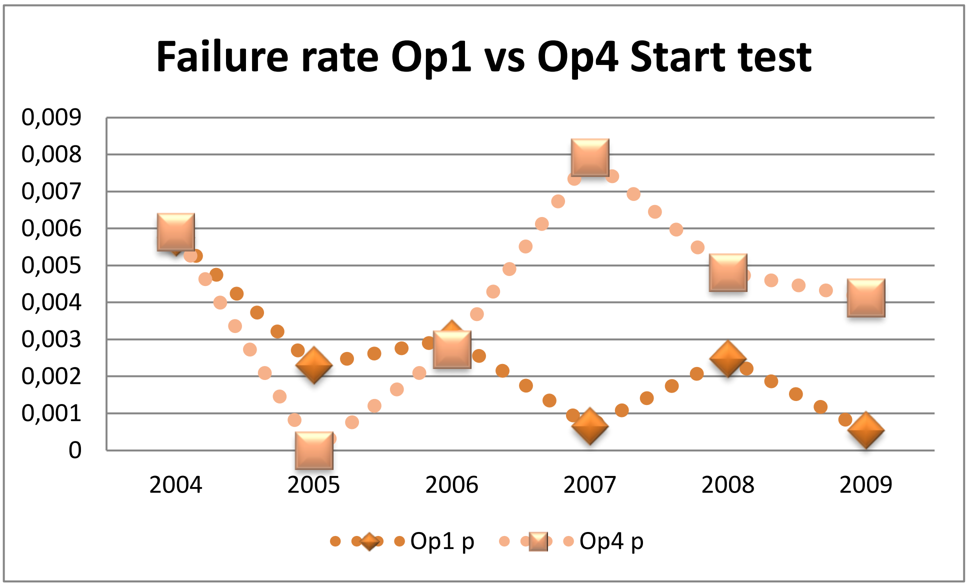

Failure rate Op3 vs Op5 Fire det.
0,008
0,007
0,006
0,005
0,004
0,003
0,002
0,001
0
2004
2005
2006
2007
2008
2009
Op3 p
Op5 p

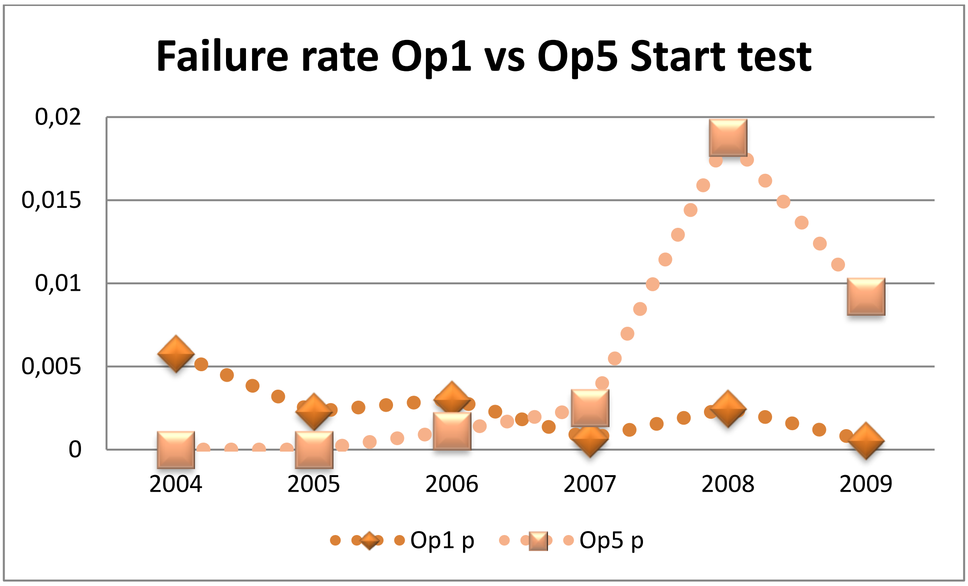

Failure rate Op4 vs Op5 Fire det.
0,012
0,01
0,008
0,006
0,004
0,002
0
2004
2005
2006
2007
2008
2009
Op4 p
Op5 p

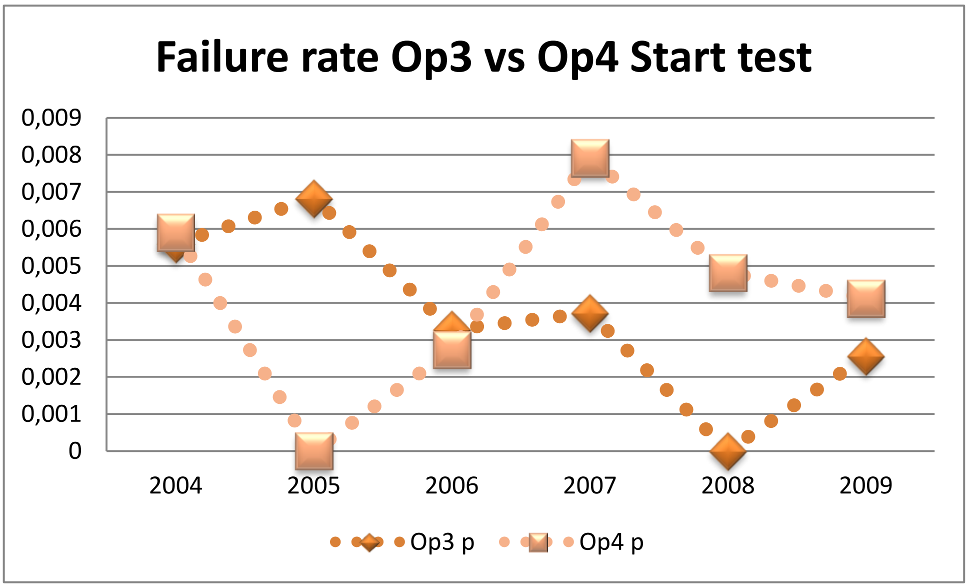

Failure rate Op1 vs Op3 Fire det.
0,005
0,004
0,003
0,002
0,001
0
2004
2005
2006
2007
2008
2009
Op1 p
Op3 p

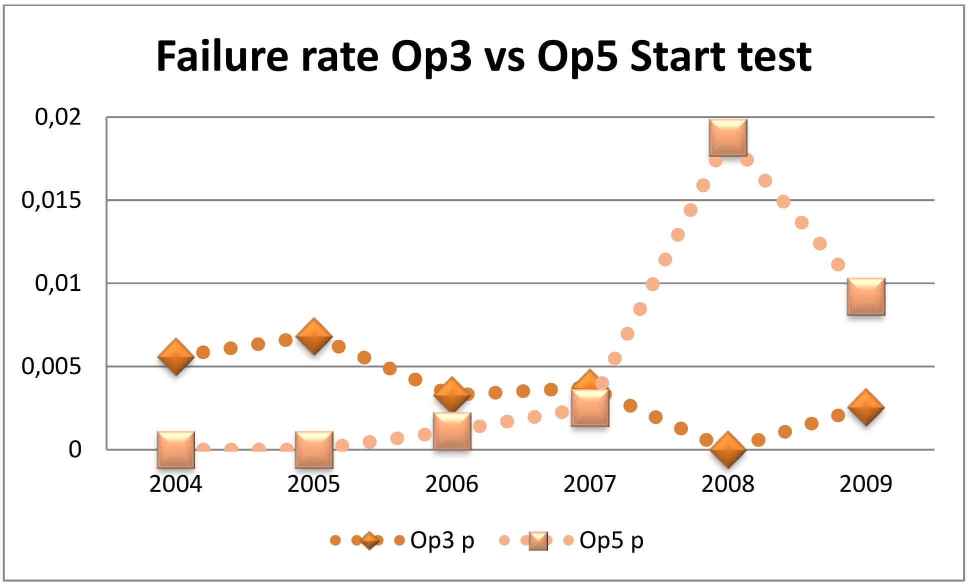
Start test fire pump barrier:

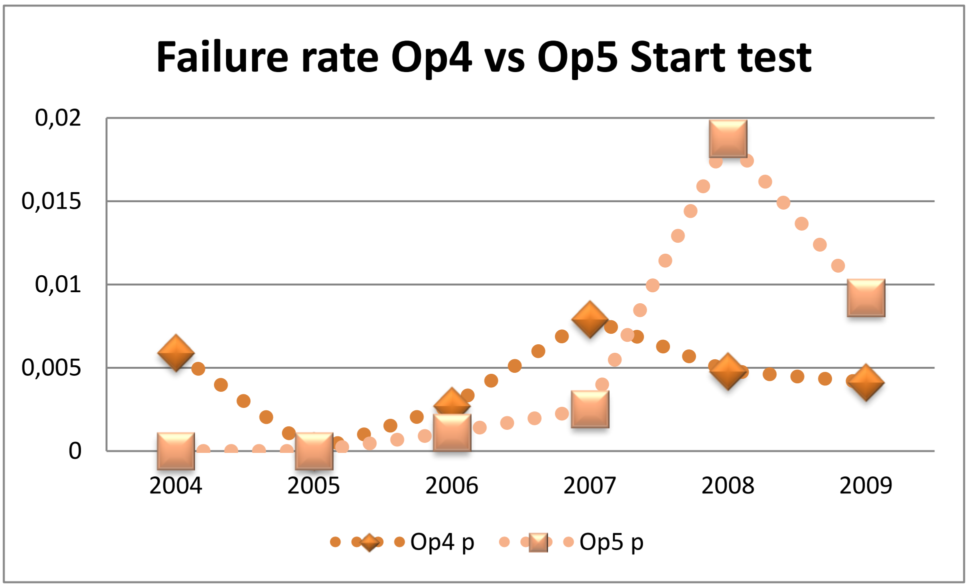




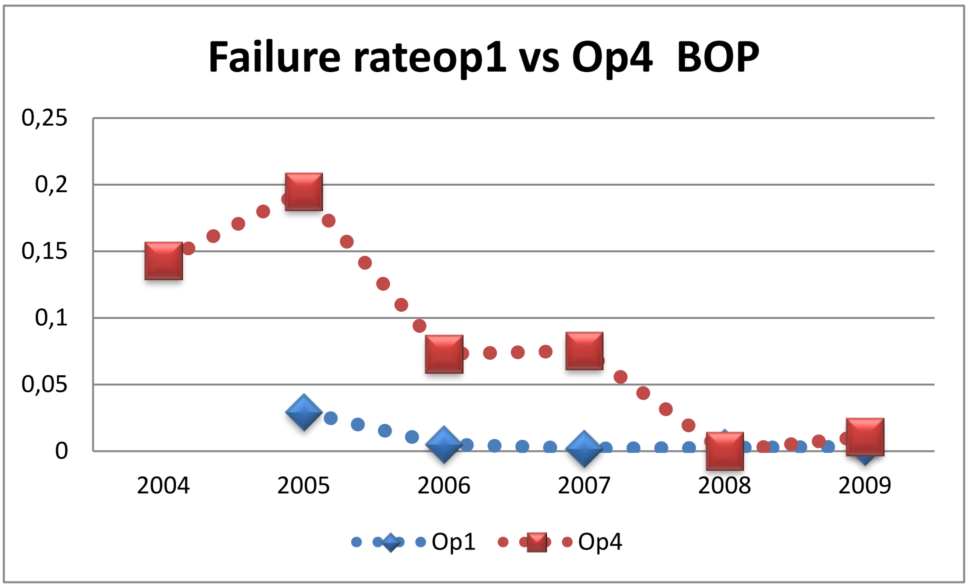

Failure rate Op1 vs Op5 Start test
0,02
0,015
0,01
0,005
0
2004
2005
2006
2007
2008
2009
Op1 p
Op5 p


Failure rate Op3 vs Op4 Start test
0,009
0,008
0,007
0,006
0,005
0,004
0,003
0,002
0,001
0
2004
2005
2006
2007
2008
2009
Op3 p
Op4 p

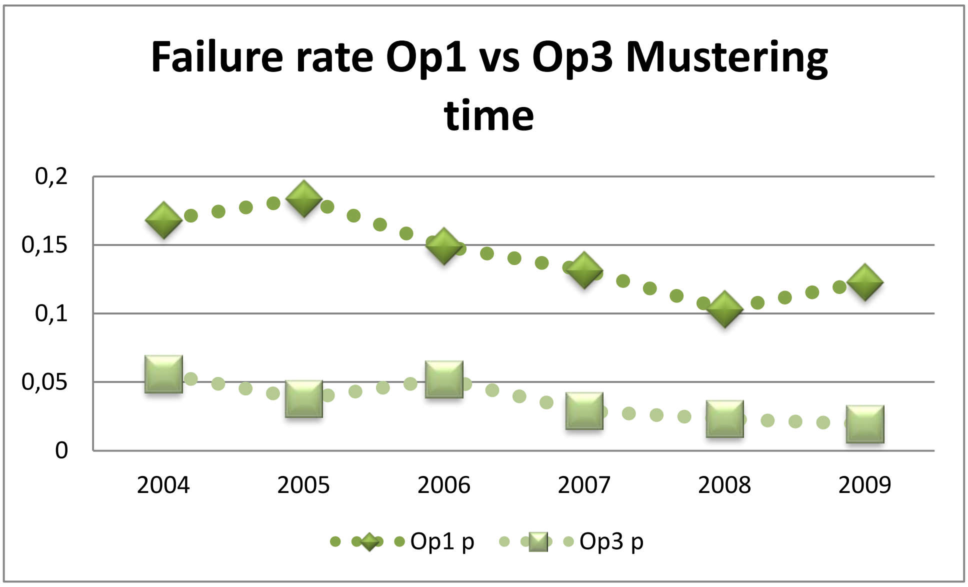

Failure rate Op3 vs Op5 Start test
0,02
0,015
0,01
0,005
0
2004
2005
2006
2007
2008
2009
Op3 p
Op5 p

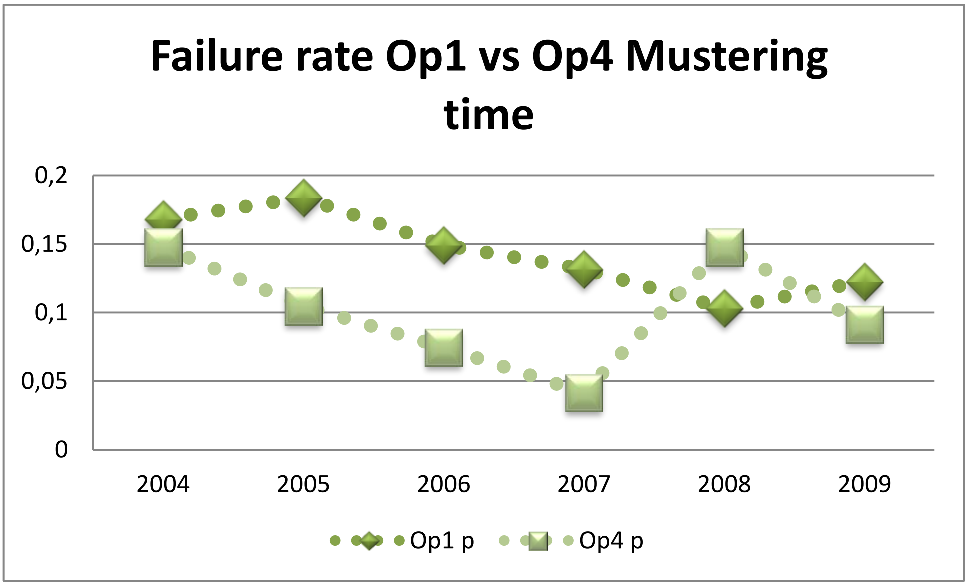

Failure rate Op4 vs Op5 Start test
0,02
0,015
0,01
0,005
0
2004
2005
2006
2007
2008
2009
Op4 p
Op5 p

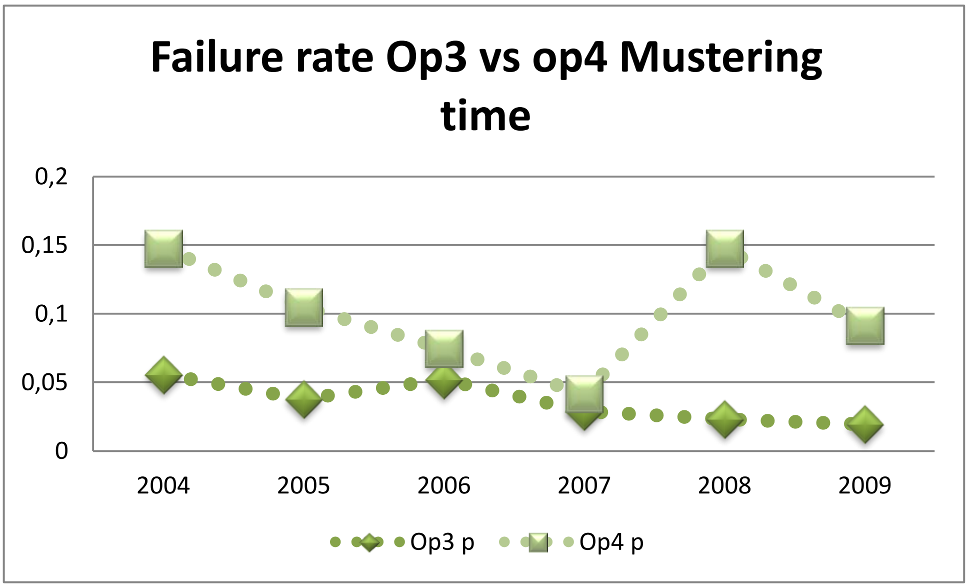
BOP barrier:



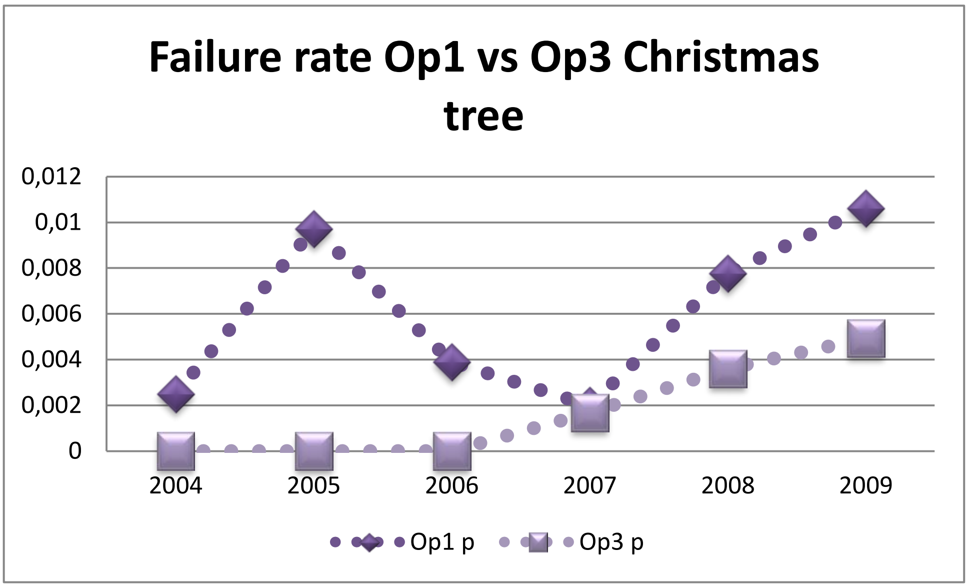
Mustering time barrier:

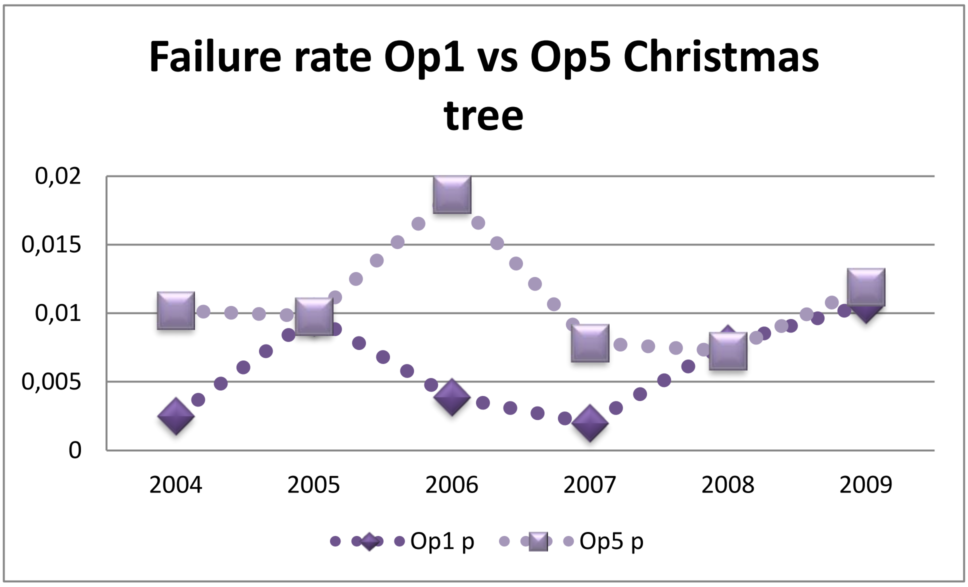


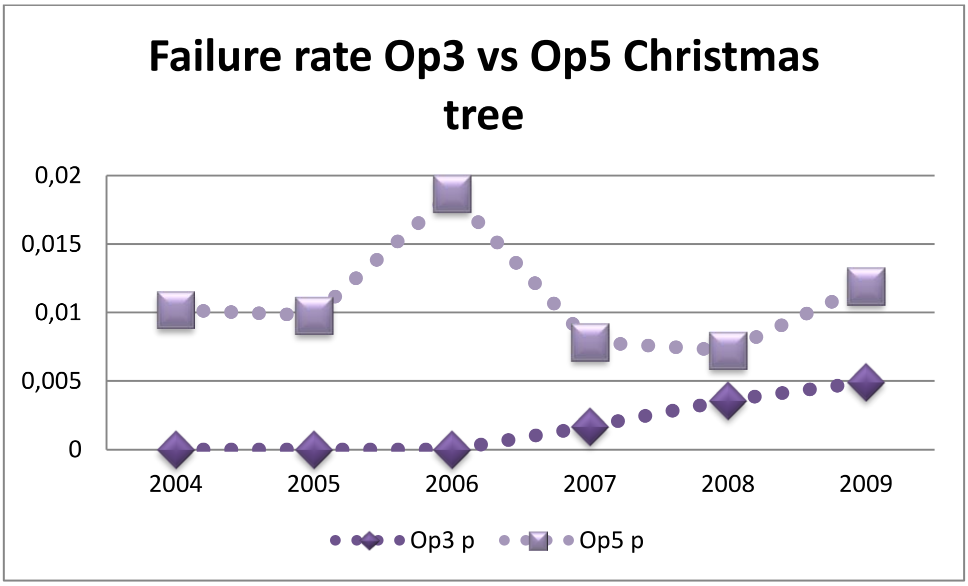

Failure rate Op1 vs Op4 Mustering time
0,2
0,15
0,1
0,05
0
2004
2005
2006
2007
2008
2009
Op1 p
Op4 p


Failure rate Op3 vs op4 Mustering time
0,2
0,15
0,1
0,05
0
2004
2005
2006
2007
2008
2009
Op3 p
Op4 p

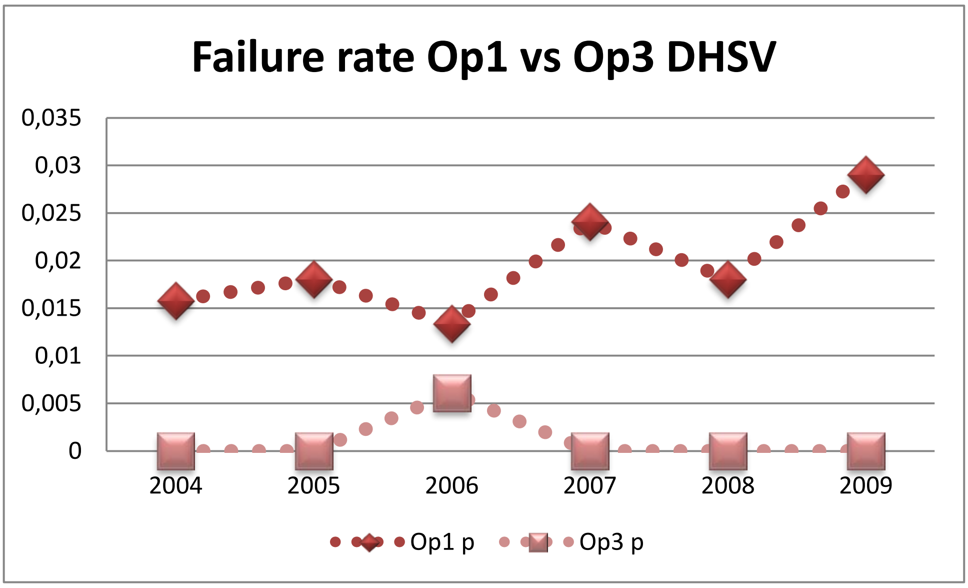
Christmas tree barrier:



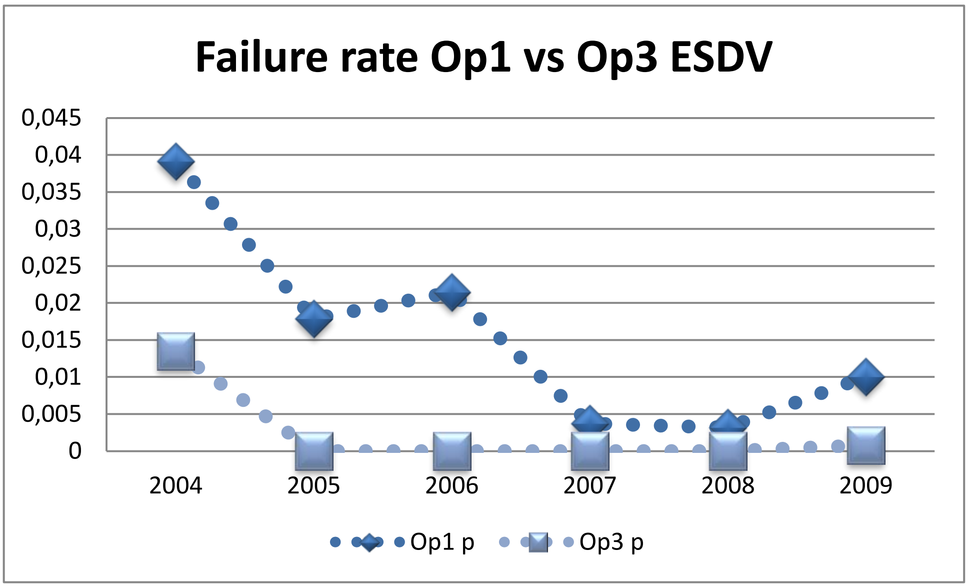




DHSV barrier:

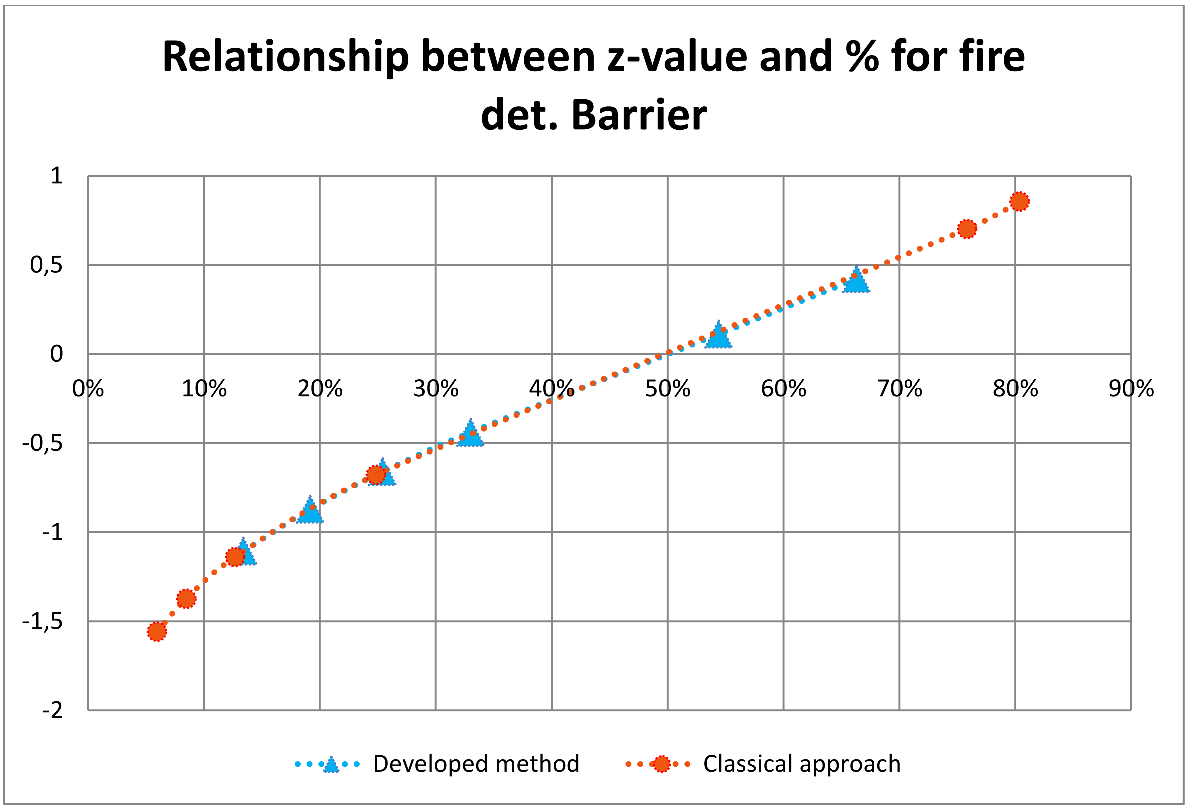


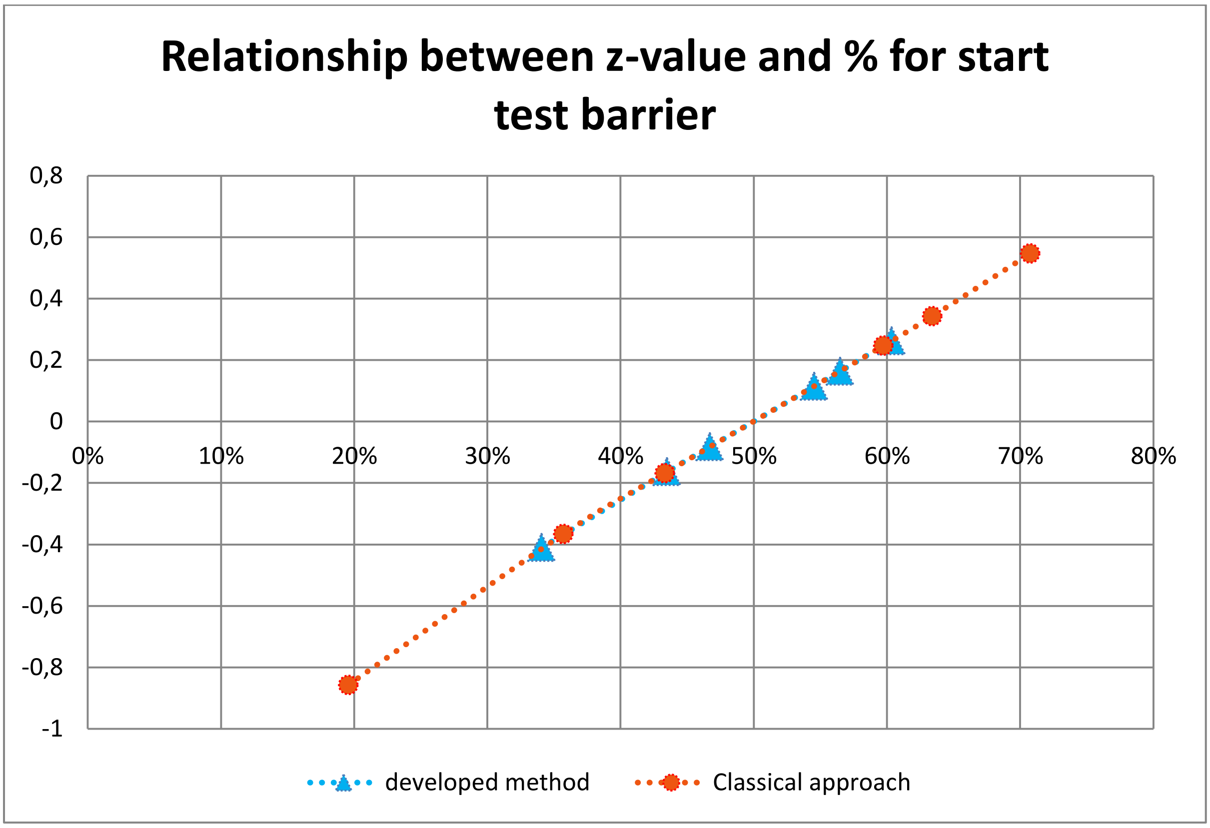
ESDV barrier:

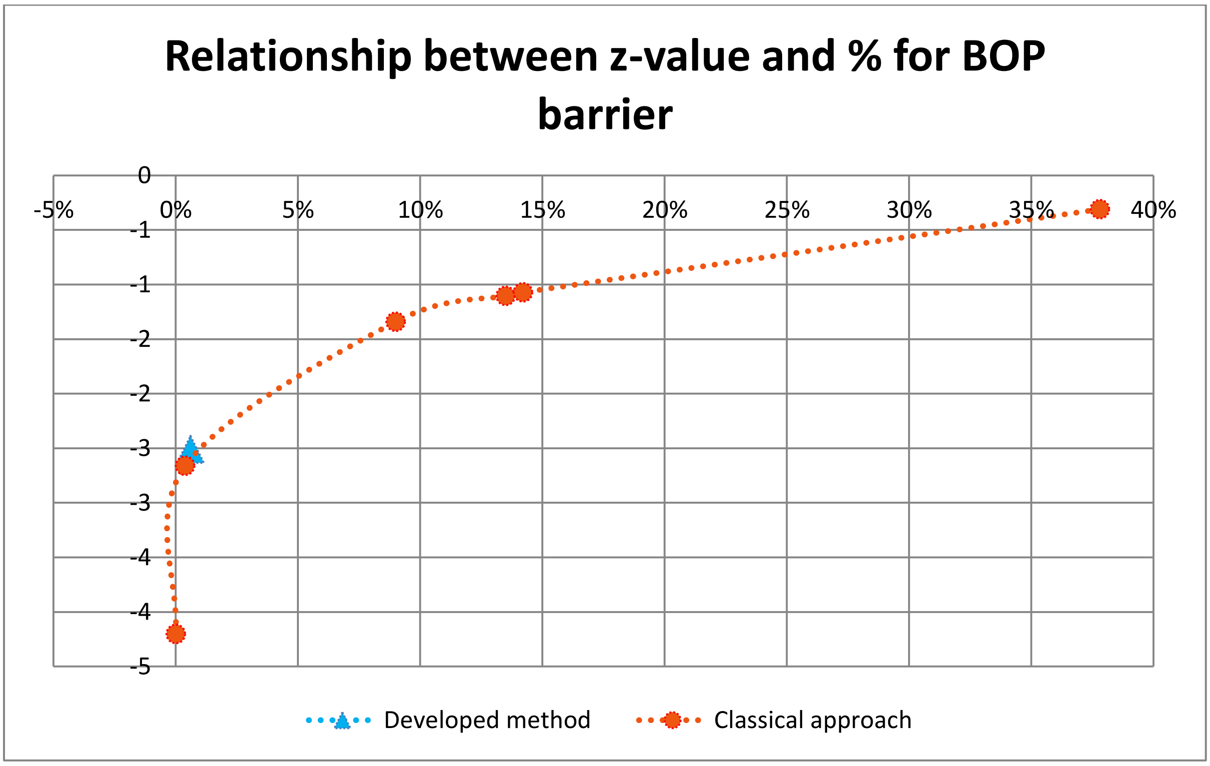


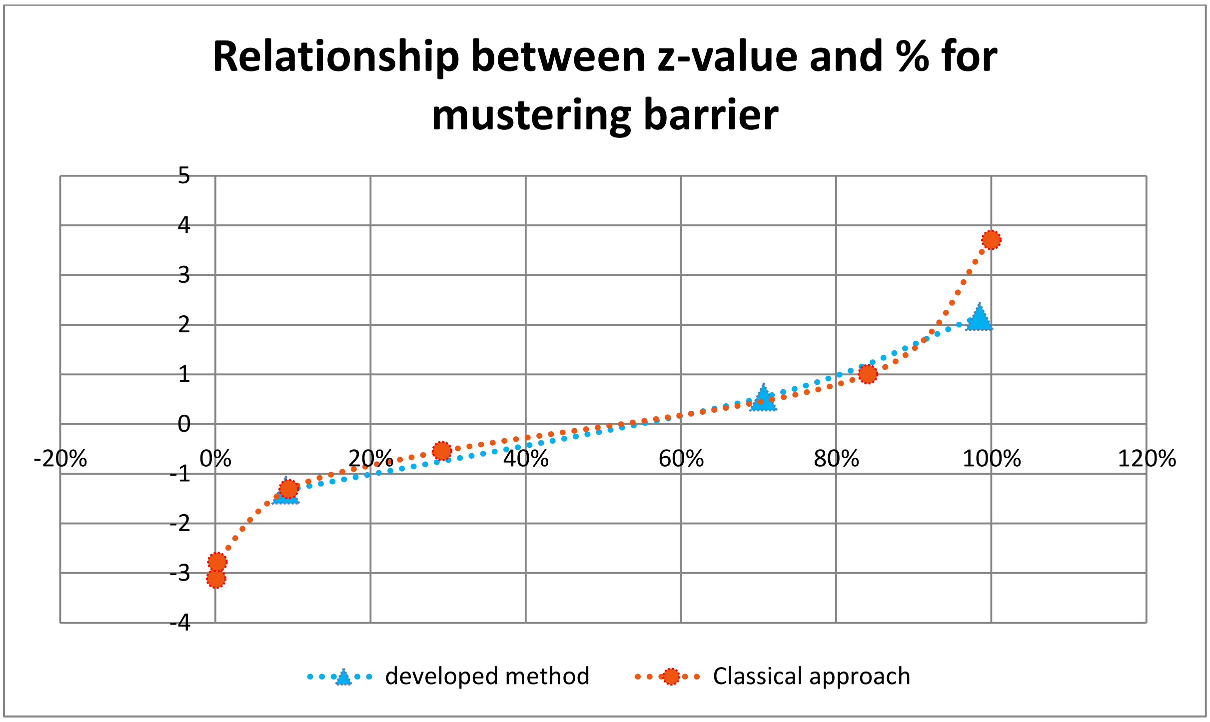
Appendix D

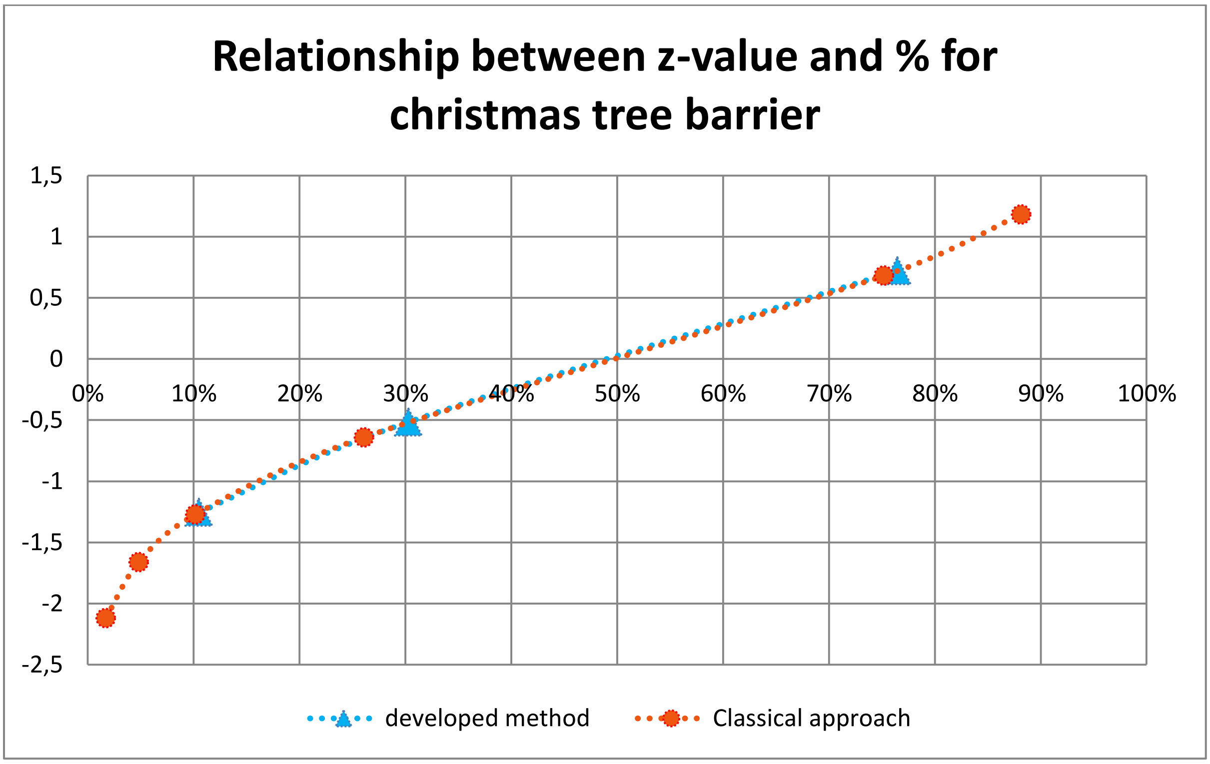
The relationship between the z-value and the probability percentages that a specific operator has a lower failure rate then another are shown

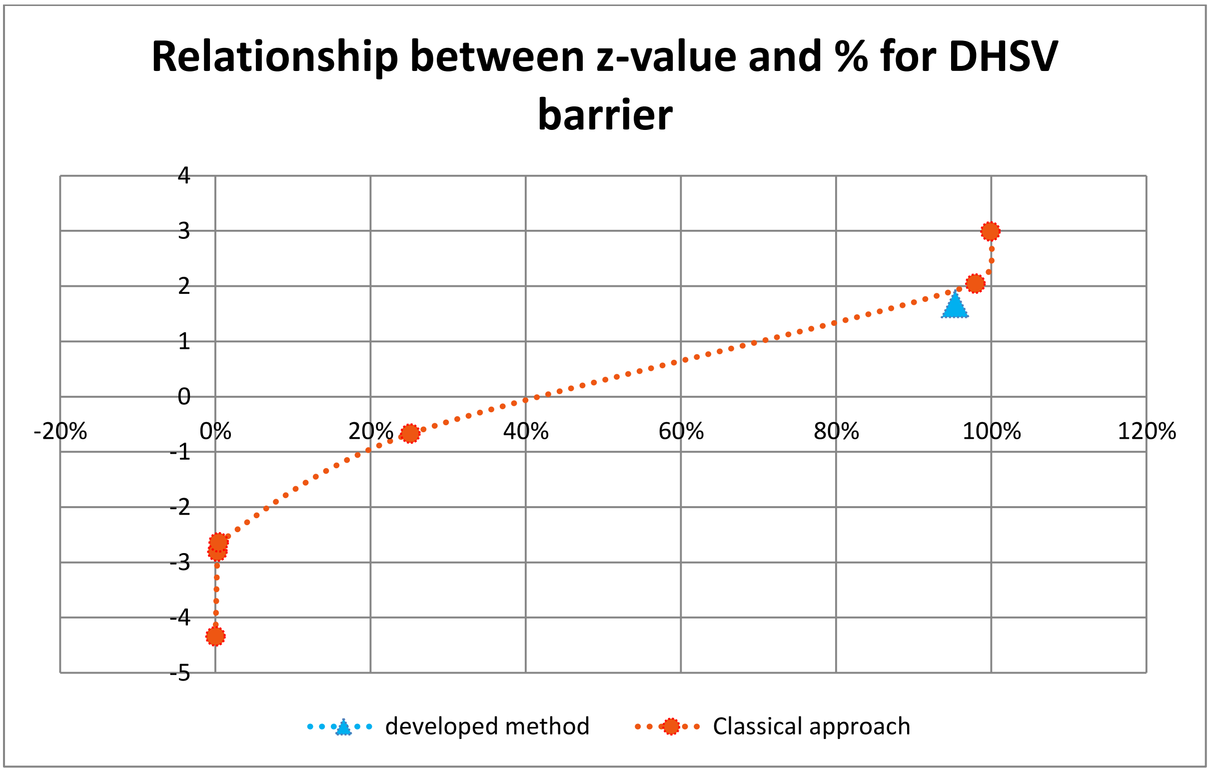


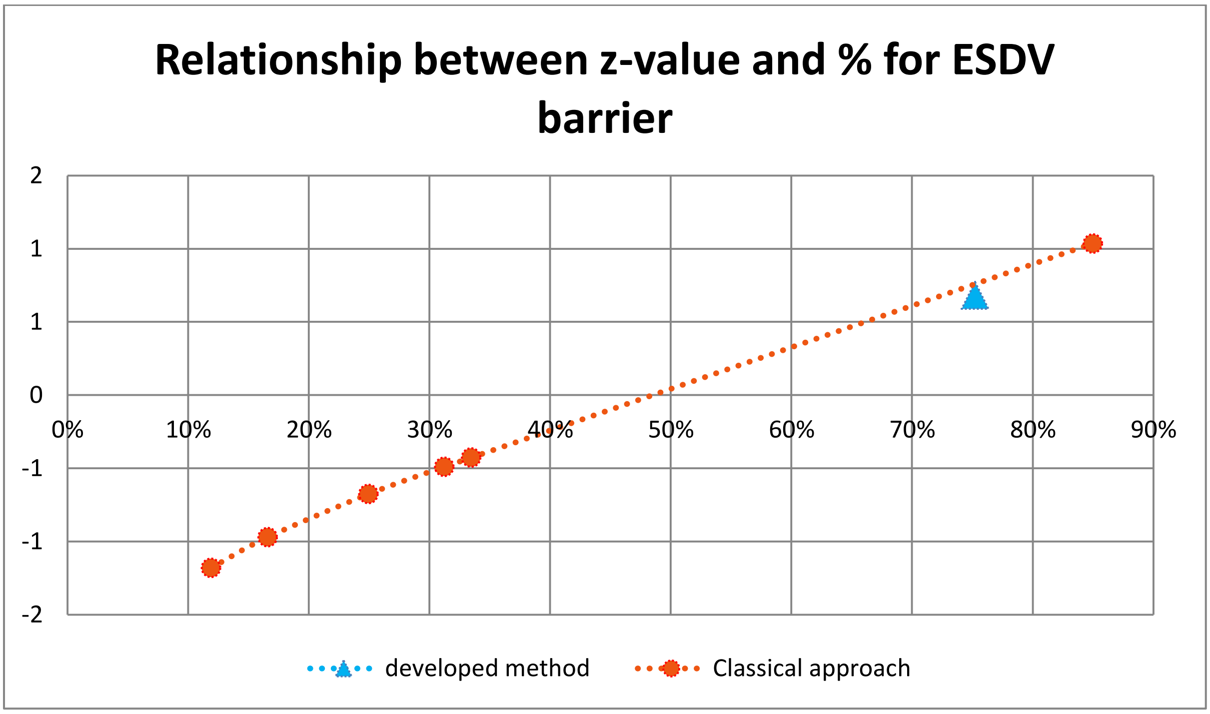



Relationship between z-value and % for BOP barrier
0
-1
-1
-2
-2
-3
-3
-4
-4
-5
-5%
0%
5%
10%
15%
20%
25%
30%
35%
40%
Developed method
Classical approach


Relationship between z-value and % for mustering barrier
5
4
3
2
1
0
-1
-2
-3
-4
-20%
0%
20%
40%
60%
80%
100%
120%
developed method
Classical approach


Relationship between z-value and % for christmas tree barrier
1,5
1
0,5
0
-0,5
-1
-1,5
-2
-2,5
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
developed method
Classical approach


Relationship between z-value and % for DHSV barrier
4
3
2
1
0
-1
-2
-3
-4
-5
-20%
0%
20%
40%
60%
80%
100%
120%
developed method
Classical approach


Relationship between z-value and % for ESDV barrier
2
1
1
0
-1
-1
-2
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
developed method
Classical approach

10 Bibliography

Aven, T., Kvaløy, J.T., (2002) *Implementing the Bayesian paradigm in risk analysis*. Reliability engineering and system safety 78: 195-201.

Aven , T., (2003) "Foundation of risk analysis"

Aven,T., (2009) "perspective on risk in decision-making context – review and discussion" Safety science 47 (2009) 798-806

Aven, T., (2010) "Misconceptions of risk" Wiley, NJ

Heide Knudsen, B., Vinnem, J.E., and Aven, T. , (2007) *Methods to monitor risk for onshore petroleum plants.*

Heide, B., (2009) "Monitoring Major Hazard Risk for Industrial Sectors" PhD thesis UiS no. 66 May 2009

Kaplan S. The words of risk analysis. Risk Anal 1997;17(4):407-17

Kvaløy, J. T., (2005) *Brukt statistikk riktig!*. University of Stavanger

Safetec. *Regresjonsanalyse av hydrokarbonlekkasjer mot andre indikatorer i RNNP* – Norsk sokkel, hovedrapport, September, (2009)

PSA (2000), Risk level on the Norwegian continental shelf, pilot project report; 2000

PSA (2009), Trends in Risk level in the petroleum activity, summary report; 2009.

PSA (2010) – Development of risk level –Norwegian continental shelf, method report (2010*)/ Utvikling I risikonivå – norsk sokkel, Metoderapport for 2010*, rev 2, 20.01.2011.

PSA (2010) *Krav til selskapets rapportinger av ytelse av barrierer 2008*, rev. 12, 22.11.2010. Risikonivå norsk petroleumsvirksomhet.

PSA (2011) Hovedrapport RNNP – Risikonivå i norsk petroleumsvirksomhet, rev.1, 27.04.2011. Risikonivå i petroleumsvirksomhet Norsk sokkel.

PSA (2011) sammmendragsrapport RNNP– Risikonivå i norsk petroleumsvirksomhet, 27.04.2011

PSA (2011) "*Deepwater Horizon – vurderinger og anbefalinger for norsk petroleumsvirksomhet, sammendrag.*"09.06.2011

Røed, W., Aven, T., (2009) *Bayesian approaches for detecting significant deterioration*. Reliability engineering and system safety 94; 604-610.

Vinnem, J. E., Aven, T., Sørum, M., and Øien, K., (2003a) "Risk indicators for major hazards in the offshore petroleum industry" (abbreviated version) ESREL 2003

Vinnem, J. E., Aven, T., Sørum, M., and Øien, K., (2003b)"Structural approach to risk indicators for major hazards". Proceedings of ESREL 2003 conference. Maastrict, Netherlands. 1615-1621.

Vinnem, J.E, Aven, T., Husebø, T., Seljelid J. & Tveit O.J. (2006) *Major hazard risk indicators for monitoring of trends in the Norwegian offshore petroleum sector*. Reliability Engineering and System Safety 91: 778-791.

Vinnem, J.E., (2007) "*Offshore risk assessment – principles, modeling and applications of QRA Studies"*, 2 ed.

Walpole, myers, myers, ye – Probability and statistics for engineers & scientist, 8th edition

Øien, K., (2001a) "Risk control of offshore installations. A framework for the establishment of Risk indicators. Department of production and quality engineering, PhD thesis. Norwegian University of Science and technology (NTNU), Trondheim, Norway

Øien, K., (2001b) "Risk indicators as a tool for risk control. Reliability engineering and system safety 74, 129 – 145.