Lessons Learned

Test and Evaluation of Mechanical Demining Equipment according to the CEN Workshop Agreement (CWA 15044:2004)

Part 4: Statistical methods used to calculate demining machine performance, performance confidence intervals and performance differences

Note that this document is still under review. The final version will be published at the end of April 2010

ITEP Working Group on Test and Evaluation of Mechanical Assistance Clearance Equipment (ITEP WGMAE)

Drafted by F.Borry, ITEP Secretariat
First version, 21/01/2010

Table of Contents

1. Introduction

The design of the performance trial described in the CEN Workshop Agreement on Test and Evaluation of Demining Machines (CWA 15044) as well as the calculation of the resulting mechanical demining machine performances uses common statistical principles and methods.

This document highlights the statistical principles and methods used in the CWA 15044 context and further lists links to references which can give more exhaustive information. The general statistical procedures are described step by step in the main body of the text and then converted to the CWA 15044 application in the corresponding text boxes.

2. Statistical Principles

2.1. Terminology and definitions

Statistics is a field in mathematics concerned with methods and procedures for collecting, presenting and summarising data (descriptive statistics) as well as to draw inferences or make predictions (inferential statistics). Typically, in inferential statistics, sample data are employed to draw inferences about one or more populations from which the samples have been derived. Whereas a **population** consists of the sum total of subjects/objects that share something in common with each other, a **sample** is a set of subjects/objects which have been derived from a population [1]. The basic objectives of statistics are 1) the **estimation of population parameters** (values that characterise a particular population) and 2) the **testing of hypotheses** about these parameters [2].

When the mine neutralisation capability of a mechanical demining machine is assessed in a CWA 15044 performance trial we are determining how a population of Anti-Personnel (AP) mines will likely be handled (neutralised or not) by the machine. The main aim of the test is to estimate the AP mine clearance capability of the machine as the percentage of neutralised AP mines in a test lane (sample). The test set-up prescribed by the CWA 15044 test guidelines further allows for the testing of a series of hypotheses about the machine's AP mine clearance capability. The following hypotheses could for instance be tested:

- is the machine's clearance capability equal in all three soil types?
- is the machine's clearance capability in sandy soil equal for the three mine burial depths?
- is the clearance capability in topsoil of two different machines equal for flush buried mines?
- etc.

For a sample to be useful in drawing inferences about the larger population from which it was drawn, it must be **representative** of the population. The ideal sample to employ is a **random sample**. A random sample must adhere to the following criteria:

- each subject/object in the population has an equal likelihood of being selected as a member of the sample,
- the selection of each subject/object is independent of the selection of all other subjects/objects, and
- for a specified sample size, every possible sample that can be derived from the population has an equal likelihood of occurring [1].

If the CWA 15044 test lane set-up specifications are followed, i.e. test targets[1] are buried in the test lanes at randomly located positions, then each test lane represents a random sample of the mine population under consideration. The latter population could for instance be AP mines buried at 10 cm depth in gravel.
Note that a position scheme devised by an operator burying test targets is not considered random. To allocate test targets randomly, a random number generator can be used such as the RAND () in Microsoft ExCel or a freely available random number generator on the web (see for example [31] and [32])

A **statistic** refers to a characteristic of a sample, i.e. it is a number which may be computed from the data observed in a random sample. A **parameter**, on the other hand, refers to a characteristic of a population, i.e. it is a number describing a population [1] [6]. A critical aspect of statistics is the estimation of **parameters** with **statistics**. Statistics, derived from samples, are used as estimators of the corresponding population parameters [3].

[1] The test targets used are representative of the generic class of AP mines or a specific class of AP mines depending on the objectives of the trial

The statistic determined in the CWA 15044 is the number of neutralised AP mine test targets, expressed as a percentage of the total number of AP mine test targets buried in the test lane. This statistic is then used as an estimate of the machine's capability to neutralise AP mines for the conditions represented by the test lane conditions.

The **sampling distribution** is the distribution[2] of the statistic calculated from a sample. If a person repeatedly took samples of size n from the population and computed a particular statistic for that sample each time, the resulting distribution of all the values obtained for the statistic is called the sampling distribution of that statistic. Every statistic has a sampling distribution [3].

Suppose a demining machine is run over a test lane with 50 test targets buried at 10 cm depth in sand (sample size n=50) and the percentage of neutralized targets is determined. Next, the machine is run again over a test lane with the same characteristics, i.e. 50 test targets buried at 10 cm depth in sand (sample size n=50) and the percentage of neutralized targets is determined again. The second percentage obtained will not necessarily be the same as the first percentage. Hence, when the test is repeated an infinite number of times, an infinite number of neutralization percentages would be obtained. The distribution of this infinite number of neutralization percentages is called the sampling distribution of the mine neutralization percentage.

Keeping the sampling distribution in mind, it should be realized that while the statistic obtained from one's sample is probably near the center of the sampling distribution (because most of the samples would be there) one could have gotten one of the extreme samples just by the luck of the draw. If one took the average of the sampling distribution -- the average of an infinite number of samples -- one would be much closer to the true population average -- the parameter of interest. So the average of the sampling distribution is essentially equivalent to the population parameter [5]. The range of statistics that can be obtained for the estimate of the population parameter through sampling is referred to as the **sampling error**. Sampling error gives some idea of the precision of the statistical estimate. A low sampling error means that there is relatively less variability or range in the sampling distribution and that it is therefore more likely that the obtained estimate is close to the real population value of interest.

[2] Definition and examples of *distributions* can be found at [33] and [34]

In the CWA 15044 case a test is run only once for a particular condition (for instance test targets at 10 cm depth in sand). Hence, we need to be aware that although the neutralization percentage we obtain is most likely near the center of the sampling distribution, i.e. close to the parameter we are looking for (the real neutralization percentage of the machine for the given conditions), we could have obtained an extreme value of the sampling distribution and hence we could be relatively far of the parameter we are looking for. It is therefore important to know the sampling error in order to assess how likely our estimate is close to the real neutralization percentage.

In practice, the sampling error is indicated by a **confidence interval** calculated using the **standard deviation** of the sampling distribution[3]. The standard deviation is the most commonly used measure of distribution spread. The confidence interval provides a range of values which is likely to contain the population parameter of interest.

Confidence intervals are constructed at a **confidence level** selected by the user. The confidence level tells one how sure one can be that the population parameter lies within the range of values given by the confidence interval around the estimate. The confidence level is expressed as a percentage. The 95% confidence level means one can be 95% certain; the 99% confidence level means one can be 99% certain. Most researchers use the 95% confidence level. A 95% confidence level means that if the same population is sampled on numerous occasions and confidence interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95% of the cases. The higher the confidence level one is willing to accept, the more certain you can be that the true population parameter lies within the range given by the confidence interval [8] [9].

In order to calculate the sampling error and confidence interval for an estimate the sampling distribution should be known. For the CWA 15044 application, the sampling distribution follows the **binomial distribution** (see text box). The binomial distribution describes the behaviour of a count variable *X* if the following conditions apply:

- the number of observations *n* is fixed,
- each observation is independent,
- each observation represents one of two outcomes ("success" or "failure"), and
- the "probability of success" *p* is the same for each outcome.

If these conditions are met, then the count variable *X* follows a binomial distribution with parameters *n* and *p*. The binomial distribution indicates the probability of obtaining X successful counts when sampling *n* objects/subjects in a population with a theoretical success rate *p* [7] [10].

[3] More information on the standard deviation of a distribution and formulas to calculate the standard deviation can, amongst other, be found at [35]

The observations in the CWA 15044 performance test are count values, i.e. we count the number of neutralised targets. The number of observations *n* has been fixed to 50 (50 mine targets in a test lane) and each observation can only have two outcomes: mine target neutralised (success) or not neutralised – life (failure). The probability of success *p* is the real neutralisation capacity of the machine for the specific conditions represented by the test lane (soil type, mine burial depth) and hence is a fixed value. The neutralisation of any AP mine test target in the test lane is independent of the neutralisation of the other AP mine test targets. In summary, the number of AP mine test targets neutralised in the test lane follows a binomial distribution.

Figure 1 depicts an example of a binomial distribution for a test lane with 20 test targets (n=20) and a machine with a theoretical AP mine clearance capacity of 80% (p=0.8). The graph shows that if we run the defined machine on a test lane with 20 test targets, we are most likely to obtain a clearance result of 16 out of 20 mines (21.8 % of the tests), i.e. the real clearance capacity of the machine, but there is also a relatively high probability we obtain 14 out of 20 mines (10.9 % of the tests) or 18 out of 20 mines (13.7 % of the tests). We can even find results indicating a machine clearance capacity of 10 out of 20 mines (0.2 % of the tests) or 20 out of 20 mines (1.2% of the tests).

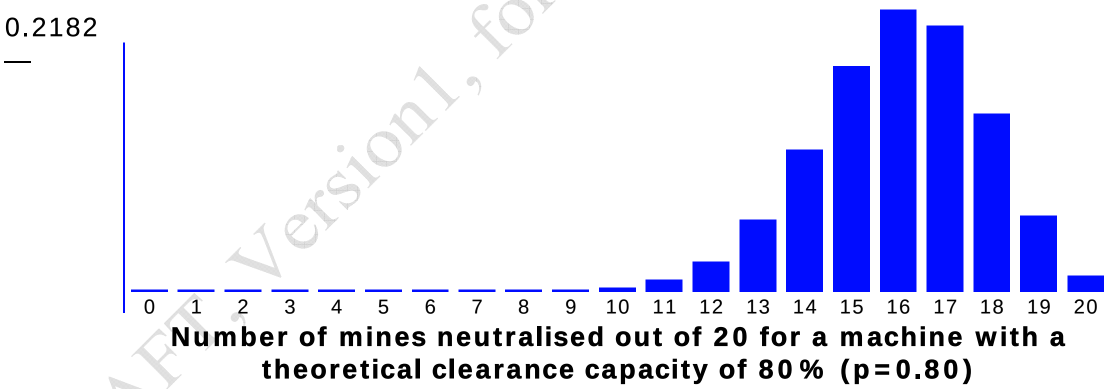


Figure 1: binomial distribution for n= 20 and p= 0.8

Note that according to statistical theory the sampling distribution of a count variable is only well-described by the binomial distribution in cases where the population size is significantly larger than the sample size. As a general rule, the binomial distribution should not be applied to observations from a random sample unless the population size is at least 10 times larger than the sample size [7].

The binomial distribution is a mathematical function with two variables n and p. Probabilities from a binomial distribution can be calculate directly using the appropriate formula (available at [7]). However, these calculations involve a lot of computation and faster methods are available, such as treading the values from a table (available at [36]), using a binomial distribution calculator available on the web (example at [10]) or using a spreadsheet which contains the statistical function (example: Microsoft ExCel – BINOMDIST).

2.2. Calculation of confidence interval

The confidence interval of a binomially distributed statistic can be calculated using different methods (formulas), each resulting in slightly different interval estimates for the same level of confidence. The most commonly used/cited are the Normal Approximation Method, the Wald Method, the Adjusted Wald Method, the Clopper-Pearson or Exact Method, and the Score Interval (Wilson) Method. Corresponding confidence interval formulas as well as the advantages and disadvantages of the cited methods are highlighted in [11] and [12]. Confidence intervals are easily calculated using open source calculators such as the ones at [13] and [14].

Important factors determining the confidence interval width are

- The sample size. The larger the sample size the more sure you can be that the sample statistic reflects the true population parameter. This indicates that for a given confidence level, the larger your sample size, the smaller your confidence interval. However, the relationship is not linear (i.e., doubling the sample size does not halve the confidence interval).
- The proportion/percentage indicated by the sample. For instance, if 99% of the test targets are neutralised and 1% is left life then the chances of error are smaller than for the case of 51% neutralised and 49% life, irrespective of sample size. It is easier to be sure of extreme results than of middle-of-the-road ones.

Figure 2 illustrates the above factors for the CWA 15044 case. The larger the sample (number of test targets in the test lane) and/or the higher the estimated neutralisation percentage the smaller the confidence interval around the estimate and hence the more reliable the estimate is. Confidence intervals were calculated using the Clopper-Pearson or Exact Method (For calculation details see http://www.itep.ws/pdf/CWA15044/BinaryConfidenceIntervals_calc.xls).

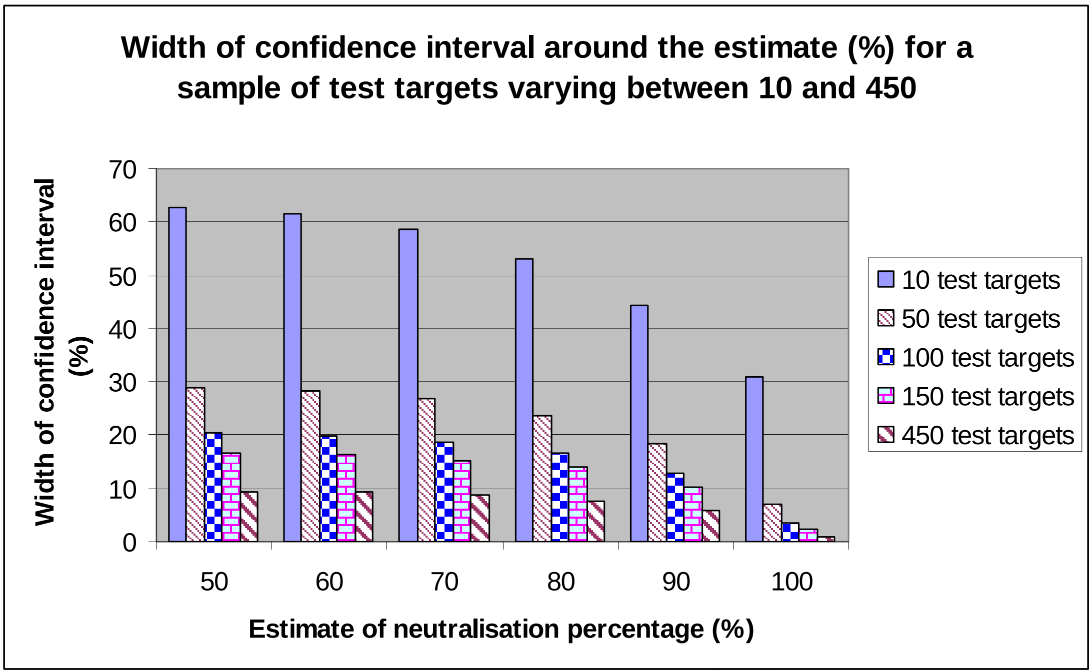


Figure 2: Confidence interval width (as a percentage of the estimate) at a 95% confidence level for samples with number of test targets varying between 10 and 450, and an estimated neutralisation capacity varying between 50 and 100 %.

Note that confidence interval calculations assume you have a genuine random sample of the relevant population. If your sample is not truly random, you cannot rely on the intervals [9].

The CWA 15044 recommends to use the 95% confidence level (5% level of significance) and to calculate confidence intervals for the neutralisation percentage according to the Clopper-Pearson or Exact Method. The Workshop which drafted the CWA 15044 further agreed that a number of 50 test targets would provide a satisfactory confidence interval around the obtained estimate at an acceptable trial cost.

The graph in Figure 3 shows that when only 10 mine targets are used and a neutralisation percentage of 80% is obtained, we can be 95% sure that the actual capability of the machine for the tested conditions (soil type, mine burial depth) is somewhere between 44% and 79%. With 50 mine targets used we can be 95% sure that the actual capability of the machine for the tested conditions is somewhere between 66% and 90% (For calculation details see http://www.itep.ws/pdf/CWA15044/BinaryConfidenceIntervals_calc.xls).

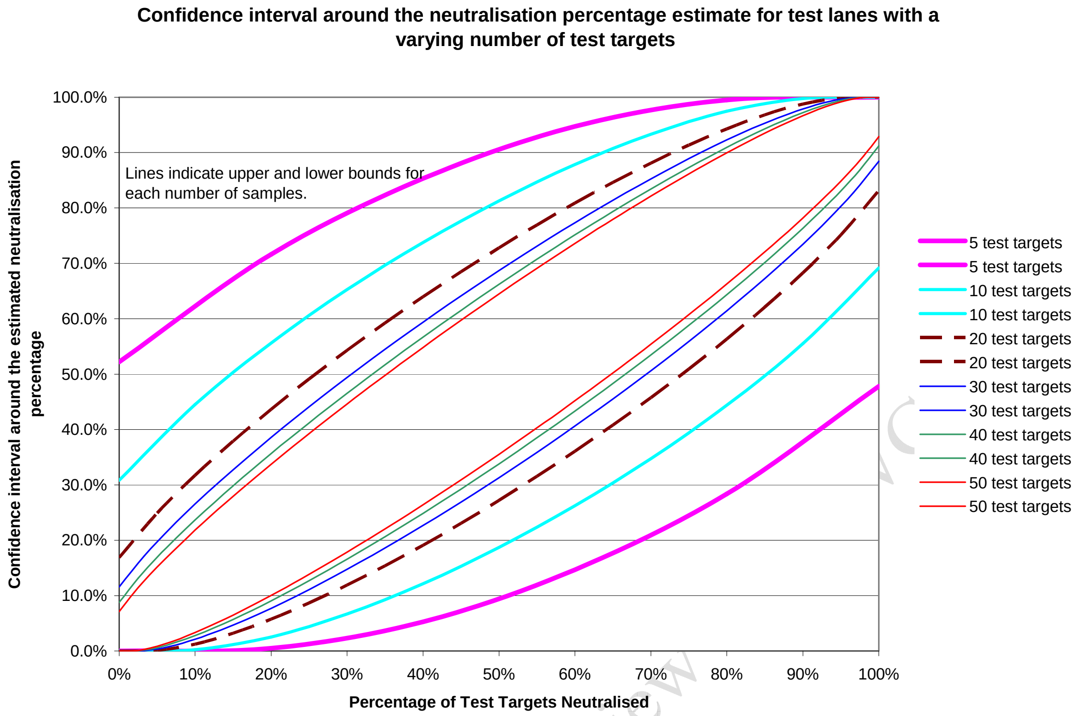


Figure 3: Confidence interval width at a 95% confidence level for test lanes with a varying number of test targets.

Note that for the hypothetical case that the neutralization percentages obtained for all conditions tested (3 soil types x 3 test target burial depths) were not significantly different the neutralization data could be pooled which would result in a sample of 450 test targets and hence an even more reliable estimate of the mine neutralization capability for the tested machine. For the previous example this would mean that you could be 95% sure that the actual capability of the machine would be somewhere between 76% and 84%.

2.3. Hypothesis testing

An important aspect of using statistical methods to estimate population parameters is that there exist standard statistical procedures to test hypotheses about the population parameters being estimated. In **hypothesis testing** one decides whether the data show a “real” effect or could have happened by change, i.e. are purely due to the sampling error [15].

A **statistical hypothesis** is an assumption about a population parameter. This assumption may or may not be true. If the sample data (observations) are consistent with the statistical hypothesis, the hypothesis is accepted; if not, it is rejected. There are two types of statistical hypotheses [15, 16]:

- **Null hypothesis, H_0**. The null hypothesis is usually the hypothesis that sample observations result purely from chance, i.e. there is no “real effect” in the population and the effect in the sample data is just due to the sampling error (chance).
- **Alternative hypothesis, H_1**. The alternative hypothesis is the hypothesis that sample observations are influenced by some non-random cause, i.e. a variable. Hence, the effect observed in the sample reflects a real effect in the population.

For the CWA 15044 case a possible H_0 could for instance be that the machine performs equally well in sand and topsoil for flush buried mines while the corresponding alternative hypothesis H_1 would be that the machine performance for flush buried mines is different in sand and topsoil. Another possible H_0 could be that machine A performs equally well as machine B for flush buried mines in sand while the corresponding H_1 would be that the machines perform differently.

In hypothesis testing the following logistic is followed [15]:

- State the hypotheses. This involves stating the null (H_0) and alternative (H_1) hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
- Assume the H_0 is true.
- Calculate the probability of getting the results observed in your data if the H_0 were true.
- If that probability is low (for instance lower than 5%), then reject the H_0.
- If H_0 is rejected, that leaves only the H_1.

In hypothesis testing, to calculate the probability that the effect observed in the data would happen by chance if the H_0 were true, a single **test statistic** is calculated and evaluated. For hypothesis testing of a binomially distributed statistic, such as a proportion (percentage), the **Chi-Square test** is used [15]. The **Chi-Square test** is based on the **Chi-Square test statistic** (χ^2) and investigates whether the proportions of certain categories are different in different groups [18]. The Chi-Square test is either a **One –Way Chi-Square test, (**also called **Test of Goodness of Fit)** or **a Two –Way Chi-Square test** (also called **Test of Independence).** In a One-way Chi-Square test, the observations are compared to a population for which the characteristics are known, while in a Two-Way Chi-Square test the frequencies of occurrence in two or more categories between two or more groups is compared. In order to calculate the Chi-Square test statistic, the observed proportion data are presented in a table, called a **contingency table**.

For CWA 15044 hypothesis testing a **Two –Way Chi-Square test** is used because the proportions of life and neutralised mines are compared for two or more groups. The latter groups can for example be:

- two different machines processing flush buried test targets in sand, or
- one single machine processing flush buried targets in sand and gravel, or
- one single machine processing test targets buried in sand at three different depths,
- etc.

The contingency tables that can be constructed for the given examples are as follows:

(a)

	Flush buried targets, sand		
	M1	M2	
Neutr.	48	40	88
Life	2	10	12
	50	50	100

(b)

	M1, flush buried targets		
	sand	gravel	
Neutr.	48	46	94
Life	2	4	6
	50	50	100

(c)

	M1, sand,			
	targets buried at 0 cm	targets buried at 10 cm	Targets buried at 20 cm	
Neutr.	48	47	39	133
Life	2	3	11	17
	50	50	50	150

The **calculation of the Chi-Square Test statistic** (χ^2) is based on the comparison of observed and expected values. In a Two-Way Chi-Square test, it is unknown how the distribution of the data over the different categories should be like. Rather, the expected values are calculated based on the observed values (table row totals, table column totals and table total) and the sum of the differences between the calculated expected values and the observed values, is then used to calculate the Chi-Square test statistic.

More information on the χ^2 **test statistic** and its calculation can be found in [17], [21], [22], [23], [24]. The below paragraphs illustrate the calculation for the CWA 15044 examples given above.

The first step in calculating the χ^2 test statistic is generating the expected value (E) for each cell of the table which contains the observed data. The expected value for each cell of the table (E_{ij}) is then calculated using the following formula:

Row total x Column total / Table total

Expected values for each cell of the example tables are given below in parenthesis and italics.

(a)

	Flush buried targets, sand		
	M1	M2	
Neutr.	48 (*44*)	40 (*44*)	88
Life	2 (*6*)	10 (*6*)	12
	50	50	100

(b)

	M1, flush buried targets		
	sand	gravel	
Neutr.	48 (*47*)	46 (*47*)	94
Life	2 (*3*)	4 (*3*)	6
	50	50	100

(c)

	M1, sand,			
	targets buried at 0 cm	targets buried at 10 cm	Targets buried at 20 cm	
Neutr.	48 (*45*)	47 (*45*)	39 (*45*)	133
Life	2 (5)	3 (5)	11 (5)	17
	50	50	50	150

The next step is to calculate the χ^2 test statistic which incorporates the differences between the expected value (E_{ij}) and the observed value (O_{ij}) of each cell according to the following formula:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Table (a): **χ^2 = 4.64**

Table (b): **χ^2 = 0.71**

Table (c): **χ^2 = 10.21**

Note that according to Chi-Square theory, the application of **Yates' correction** is recommended when calculating the χ^2 test statistic for two by two tables with one or more cells with frequencies less than five. Some apply the correction to all two by two tables. **Yates' correction** is an arbitrary, conservative adjustment [24]. The latter means that the correction will make it more difficult to establish differences, i.e. the observed differences will have to be greater for the test to indicate that the null hypothesis should be rejected. In the corrected χ^2 each (O-E) absolute value is deduced by 0.5. All other calculations remain the same.

In CWA 15044 hypothesis testing two by two contingency tables containing cells with less than five frequently occur, and hence Yates' correction should be applied.

Table (a): **χ^2 corrected= 6.06**

Table (b): **χ^2 corrected= 0.18**

In hypothesis testing the value of the calculated test statistics is then compared to a threshold value, also called the **critical value.** The critical value for any hypothesis test depends on **the significance level (α)** at which the test is carried out, and whether the test is **one-sided or two-sided**.

The Chi-Square test evaluates if the χ^2 value obtained from the observed data is a likely value to obtain for a variable which follows the Chi-Square distribution and for a certain level of significance, i.e. it is compared to a critical value which determines the limit between 1) the range of values that are likely to be obtained (region of acceptance) assuming the null hypothesis is true and 2) the range of values which are unlikely to obtain (region of rejection) assuming the null hypothesis is true. If the test statistic falls within the region of acceptance, the null hypothesis is accepted. On the other hand, if the χ^2 value obtained from the observed data falls within the range of rejection the null hypothesis $\mathbf{H_0}$ is rejected and the alternative hypothesis $\mathbf{H_1}$ is accepted. The region of acceptance (and hence the critical value) is determined so that the chance of rejecting the null hypothesis when it is true is equal to the significance level **α**. The significance level **α** is related to the confidence level chosen by the user according to the following formula: confidence level (%) = (**1- α**) x 100. Hence a confidence level of 95% chosen by the user implies the use of a 0.05 significance level in a hypothesis test.

For the CWA 15044 case the critical value of the Chi-Square distribution is determined so that the probability of accepting the null hypothesis when it is true is 95% (confidence level = 95%) and the chance of rejecting the null hypothesis when it is true is 5% (significance level α = 0.05). For the examples given above this means that when we the reject the null hypothesis and therefore conclude that: (a) the two tested machines have a different performance when clearing flush buried mines in sand, (b) the machine performs differently when clearing flush buried mines in sand and gravel, (c) the machine performs differently clearing mines in sand depending on the burial depth, then there is a 5% change that the conclusion we have drawn is wrong.

The Chi-Square distribution is a mathematical distribution of which the shape is determined by the number of **degrees of freedom df** [25]. Hence, the critical value to which the calculated value of the test statistic is compared also depends on the degrees of freedom of the Chi-Square distribution. The degrees of freedom (df) amounts to the number of independent pieces of information that go into the estimate of a parameter. In general, the degrees of freedom of an estimate is equal to the number of independent scores that go into the estimate minus the number of parameters estimated as intermediate steps in the estimation of the parameter itself [26]. A simple rule for a test comparing the frequencies of occurrence in two or more categories between two or more groups is that the degrees of freedom equal (number of columns minus one) x (number of rows minus one) not counting the rows or columns for the totals.

When the degrees of freedom are known, then the critical values of the Chi-Square distribution for different significance levels can be determined from critical value tables available on the web (example at [29]) or from the corresponding spreadsheet function[4]

In hypothesis testing the usual null hypothesis is that there is no difference between the populations from which the data come. If the null hypothesis is not true the alternative hypothesis must be true, i.e. there is a difference. Since the null hypothesis specifies no direction for the difference nor does the alternative hypothesis, it is considered a **two- sided test**. In a **one-sided test** the alternative hypothesis does specify a direction - for example, in medicine that an active treatment is better than a placebo [19]. If a two-sided test is executed the region of acceptance is delineated by two critical values. When the obtained value of the test statistic is greater than the upper critical value or less than the lower critical value, the null hypothesis is rejected. For a one-sided test, the region of acceptance is delineated by one critical value only.

Two sided tests should be used unless there is a very good reason for doing otherwise. If one sided tests are to be used, the direction of the test must be specified in advance, i.e. before collecting any data [19], [20]. A one-sided test is appropriate when you can state with certainty (and before collecting any data) that there either will be no difference or that the difference will go in a direction you can specify in advance. If you cannot specify the direction of any difference before collecting data, then a two-sided test is more appropriate [20].

For CWA 15044 hypothesis testing a two-sided test is the better choice because we can usually not state with certainty before collecting any data that for instance one particular machine will le better than the other to neutralise test targets, or that a particular machine will be better at neutralising test targets in sand than in topsoil, etc. A one-sided test might be appropriate, for instance, when an upgrade of a machine is compared to the previous version and one can state with certainty that the machine will not be worse at clearing test targets but equal or better than the previous version.

[4] In ExCel: CHIINV(α,df)

When a two-sided Chi-Square test is carried out at a significance level of 5 % (α = 0.05), the obtained value for the Chi-Square statistic is compared to the two critical values corresponding to **α/2 = 0.025 and 1-α/2=0.975** When the test statistic's value is greater than the upper critical value (corresponding to **α/2 = 0.025)** or less than the lower critical value (corresponding to **1-α/2=0.975**), the null hypothesis is rejected.
When a one-sided Chi-Square test is carried out at a significance level of 5 % (α = 0.05), the obtained value for the Chi-Square statistic is compared to the upper critical value corresponding to **α = 0.05**. When the test statistic's value is greater than this upper critical value the null hypothesis is rejected [29]

For the CWA 15044 examples given above we consider 1) a two-sided test, i.e. the null hypothesis is that the performances of the machine(s) are not different while the alternative hypothesis is that they are different and 2) a level of significance of 0.05. The critical values are as follows (read from the tables in [29])

(a)
df=1
upper critical value $\alpha/2 = 0.025 = 5.024$
lower critical value $1-\alpha/2= .975 = 0.001$

(a)
df=1
upper critical value $\alpha/2 = 0.025 = 5.024$
lower critical value $1-\alpha/2=0.975 = 0.001$

(c)
df=2
upper critical value $\alpha/2 = 0.025 = 7.378$
lower critical value $1-\alpha/2=0.975 = 0.051$

The results of the tests can be interpreted as follows:

(a)
χ^2 corrected= 6.06 > 5.024
Reject H_0
Accept H_1
The tested machines perform differently for flush buried mines in sand

(b)
χ^2 corrected= 0.18 < 5.024
χ^2 corrected= 0.18 > 0.001
Accept H_0
The machine performs equally well in sand and gravel for flush buried mines

(c)
χ^2 = **10.21** > 7.378
Reject H_0
Accept H_1
The performance of the machines for flush buried mines is dependent on the target burial depth

The above calculations illustrated with a CWA 15044 data set, can be done on any data set provided that **a sufficiently large sample size** is assumed. There is no accepted cutoff. Some set the minimum sample size at 50, while others would allow as few as 20. Note that Chi-square must be calculated on actual count data (not substituting percentages) **and adequate cell sizes** are also assumed. Some require cell sizes of 5 or more and others require 10 or more. Use of the Chi-Square test, however, is inappropriate if any frequency is below 1 or if the frequency is less than 5 in more than 20% of your cells. In the two by two case of the Chi-Square test of independence, expected frequencies less than 5 are usually considered acceptable if Yates' correction is employed [24]. The calculator given at [24] allows for the interactive calculation of the Chi-Square test statistic and indicates if the Chi-Square statistic is appropriate for the given data set.

The curve in Figure 5 of the CWA 15044 summarises the significance calculations for differences between two samples, with each sample containing 50 test targets. The curve indicates the cut-off point for which the difference in neutralisation percentage observed during the tests can be viewed as a significant difference from a statistical point of view. Annex 1 provides additional details on the assumptions used to obtain The curve in Figure 5 of the CWA 15044 and the alternative assumptions and hence calculations that could be used. It further illustrates the effect of the different assumptions on the trial conclusions with some practical examples.
Note that from a pure statistical point of view CWA 15044 figure 5 allows us to conclude that machine performances are different but we cannot state that one machine performs better than the other.

3. References

[1] Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition, D. J. Sheskin, Chapman & Hall/CRC, 2004

[2] Elementary Statistical Methods for Foresters, Agricultural Handbook 317, F. Freese, U.S. Department of Agriculture, Forest Service, 1980

[3]Introductory Statistics: Concepts, Models, and Applications. The Sampling Distribution, http://www.psychstat.missouristate.edu/Introbook/sbk19.htm

[4] Hyperstat Online, Sampling Distribution, http://davidmlane.com/hyperstat/A11150.html

[5] Web Center for Social Research Methods, Statistical Terms in Sampling, http://www.socialresearchmethods.net/kb/sampstat.php

[6] Yale University Department of Statistics, Sampling in Statistical Inference, http://stat.yale.edu/Courses/1997-98/101/sampinf.htm

[7] Yale University Department of Statistics, The Binomial Distribution, http://stat.yale.edu/Courses/1997-98/101/binom.htm

[8] Engineering Statistics Handbook, What are confidence intervals? http://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm

[9] University of Connecticut, Confidence Interval, http://www.gifted.uconn.edu/siegle/research/Samples/ConfidenceInterval.htm

[10] Statistical Tutorial: Binomial Distribution, StatTrek Teach Yourself Statistics, http://stattrek.com/Lesson2/Binomial.aspx?Tutorial=Stat

[11] Estimating Completion Rates from Small Samples using Binomial Confidence Intervals: Comparisons and Recommendations, J. Sauro, J.R. Lewis, Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, 2005 http://www.measuringusability.com/papers/sauro-lewisHFES.pdf

[12] Understanding Binomial Confidence Intervals, P. Mayfield, SigmaZone.com, 2008 http://www.sigmazone.com/binomial_confidence_interval.htm

[13] Southwest Oncology Gropup Statistical Center, Binomlial confidence interval calculator, http://www.swogstat.org/stat/public/binomial_conf.htm

[14] Measuring Usability, Confidence Interval Calculator for a Completion Rate, J. Sauro, 2005, http://www.measuringusability.com/wald.htm

[15] DePaul University, Hypothesis testing and Comparing Two Proportions, http://condor.depaul.edu/~dallbrit/extra/psy241/psy241-lec10a-chi-square.ppt#257

[16] Statistical Tutorial: Hypothesis Tests, StatTrek Teach Yourself Statistics, http://stattrek.com/Lesson5/HypothesisTesting.aspx

[17] Research Design in Occupational Education, Module S7 – Chi Square, J.P. Key, 1997,
http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage28.htm

[18] MicrobiologyBytes: Maths & Computers for Biologists: Inferential Statistics – Comparing Groups II, chi-squared test,
http://www.microbiologybytes.com/maths/1011-21.html

[19] Statistics Notes: One and Two Sided Tests of Significance, J.M. Bland, D.G. Bland , Medical Statistics Laboratory, Imperial Cancer Research Fund, London, 1994, http://www.bmj.com/cgi/content/short/309/6949/248

[20] Intuitive Biostatistics: Choosing a statistical test, GraphPad.com,
http://www.graphpad.com/www/Book/Choose.htm

[21] QMSS e-Lessons, The Chi-Square Test
http://ccnmtl.columbia.edu/projects/qmss/the_chisquare_test/about_the_chisquare_test.html

[22] Connections, What is the chi-square statistic? M. Mamahlodi, 2006
http://cnx.org/content/m13487/latest/

[23] Department of Mathematics and Computer Science, Hobart and William Smith Colleges, The Chi Square Statistic,
http://math.hws.edu/javamath/ryan/ChiSquare.html

[24] An interactive calculation tool for chi-square tests of goodness of fit and independence, K. J. Preacher, 2001, Available from http://www.quantpsy.org,
http://people.ku.edu/~preacher/chisq/chisq.htm

[25] MathWorld – A Wolfram Web Resource, Chi-Squared Distribution, E. W. Weisstein, http://mathworld.wolfram.com/Chi-SquaredDistribution.html

[26] Hyperstat Online, Degrees of Freedom,
http://davidmlane.com/hyperstat/A42408.html

[29] Engineering Statistics Handbook, Critical Values of the Chi-Square Distribution,
http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

[30] NC State University, CHASS, Chi-Square Significance Tests,
http://faculty.chass.ncsu.edu/garson/PA765/chisq.htm

[31] What's this fuss about true random? Random generator, http://www.random.org/

[32] Random Number Generator and Checker,
http://www.psychicscience.org/random.aspx

[33] Concepts & Applications of Inferential Statistic, L. Lowry,shttp://faculty.vassar.edu/lowry/webtext.html

[34] Engineering Statistics Handbook, Gallery of Distributions,
http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

[35] Hyperstat Online, Standard Deviation and Variance,
http://davidmlane.com/hyperstat/A16252.html

[36] Pearson Education, Binomial Probabilities Table, http://wps.aw.com/wps/media/objects/384/394213/Binomial%20Table.pdf

4. Annex 1: statistical approaches to obtain the observed neutralisation fraction at which observed differences in neutralisation percentage become statistically significant

The graph in Figure 5 of the CWA 15044 showing when the observed differences in CWA 15044 performance test runs are considered statistically different are based on the statistical principles explained in the main body of this LL. The formulas used were those for a One-sided Chi-Square test and without Yates' correction.

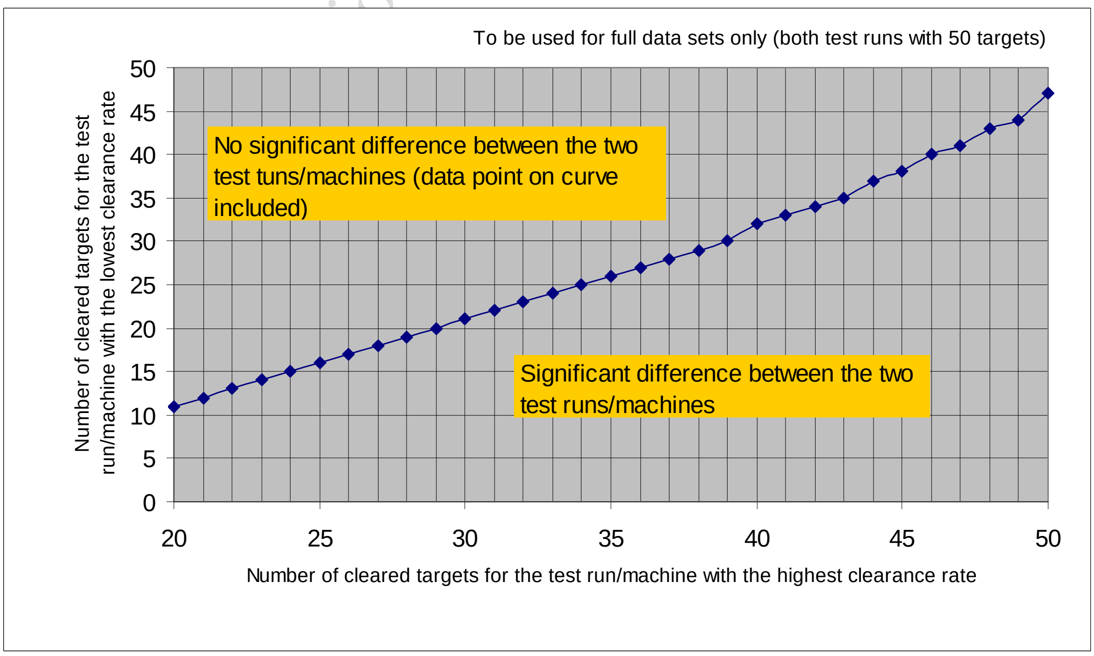
If statistical theory is applied correctly to the CWA 15044 performance test a Two-sided Chi-Square test is probably the better choice and a Yates' correction could be applied (see main body of the text for further explanation).

This annex shows the different conclusions one would draw when using the different approaches (One-sided Chi-Squared test without Yates' correction, Two-sided Chi-Squared test without Yates' correction, One-sided Chi-Squared test with Yates' correction and Two-sided Chi-Squared test with Yates' correction). It is clear from the below that the graph included in the CWA 15044 is the least conservative one, i.e. smaller differences in neutralisation percentage are considered statistically different. Based on the graphs and explanation in this LL document, users should be able to choose the best approach for their purposes.

4.1. Graphs

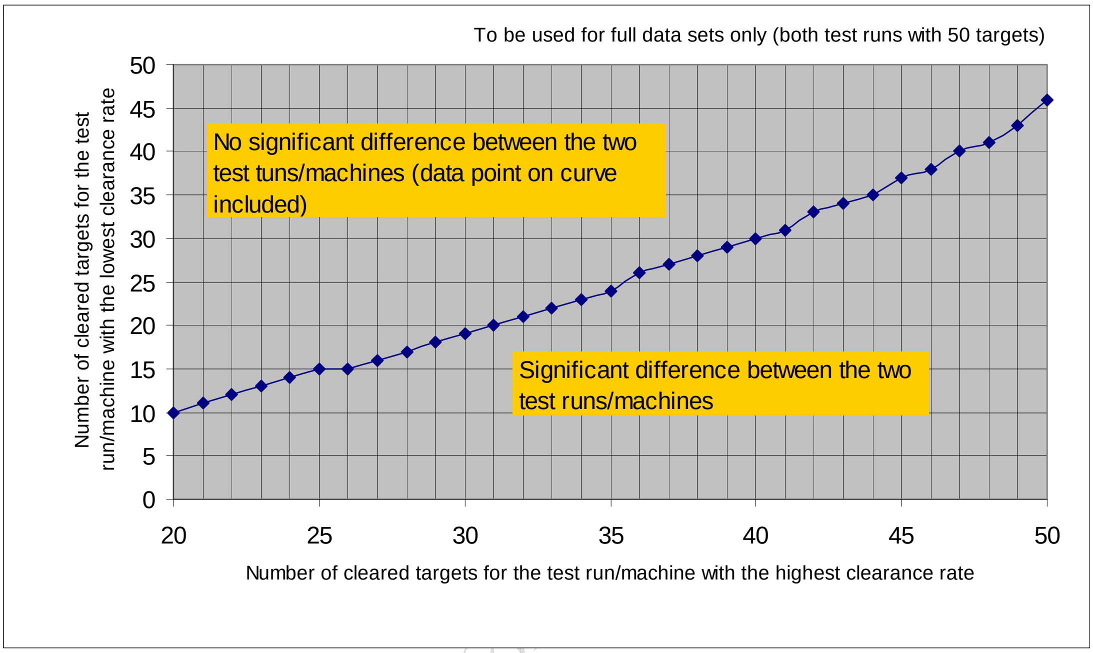
The calculations for the below graphs are available at
http://www.itep.ws/pdf/CWA15044/Statistics_CWA15044.xls

Approach 1, One-sided Chi-Squared test without Yates' correction



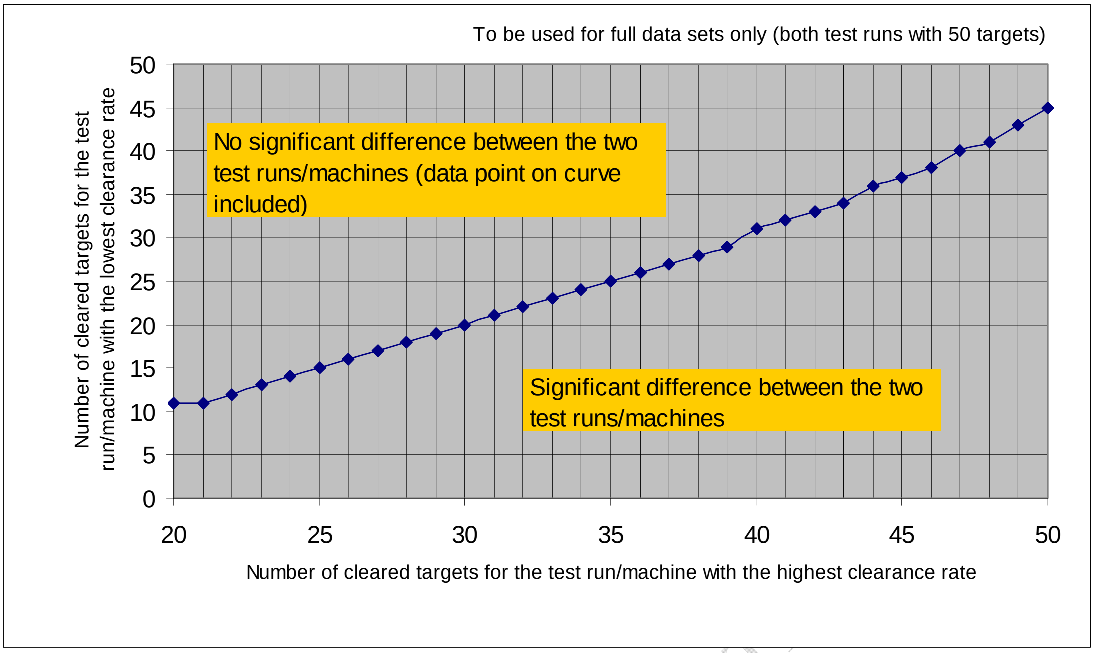
Example: for two test runs, each using a test lane with 50 test targets, and for one of the test runs showing all 50 targets neutralised the other test run needs to leave more than 3 test targets life (less than 47 neutralised) for the observed difference to be statistically different.

Approach 2, Two-sided Chi-Squared test without Yates' correction



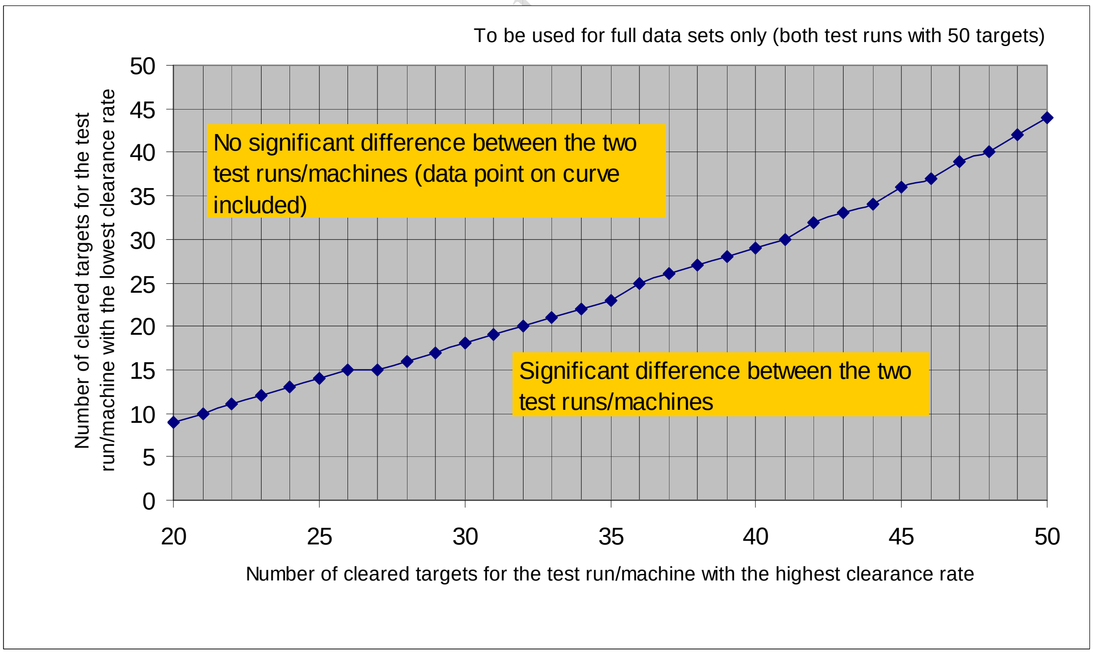
Example: for two test runs, each using a test lane with 50 test targets, and for one of the test runs showing all 50 targets neutralised the other test run needs to leave more than 4 test targets life (less than 46 neutralised) for the observed difference to be statistically different.

Approach 3, One-sided Chi-Squared test with Yates' correction



Example: for two test runs, each using a test lane with 50 test targets, and for one of the test runs showing all 50 targets neutralised the other test run needs to leave more than 5 test targets life (less than 45 neutralised) for the observed difference to be statistically different.

Approach 4, Two-sided Chi-Squared test with Yates' correction



Example: for two test runs, each using a test lane with 50 test targets, and for one of the test runs showing all 50 targets neutralised the other test run needs to leave more than 6 test targets life (less than 44 neutralised) for the observed difference to be statistically different.

4.2. Application of the different statistical approaches to CWA 15044 performance test data

4.2.1. CWA 15044 performance test results for the Mini MineWolf flail and tiller tools

Data taken from test report (ITEP Project 3.2.44, 2007)
http://www.itep.ws/pdf/FinalReportMiniMineWolf2007.pdf

	Mini MineWolf flail			Mini MineWolf tiller		
	Sand	Gravel	Topsoil	Sand	Gravel	Topsoil
Flush	50/50	50/50	49/50	47/50	50/50	50/50
10 cm	50/50	49/50	50/50	48/50	49/50	49/50
15 cm	50/50	50/50	50/50	49/50	50/50	50/50

Using the graph corresponding to approach 1 and the data in the above table it can be concluded that there are no significant differences in flail performance for the different soils and the different burial depths, i.e. the flail works as well in sand, gravel and topsoil and for mines buried at different depths up to 15 cm. The same can be stated for the tiller performance. Furthermore, the data also show that there is no significant difference in mine target neutralisation performance between the flail and tiller for all conditions tested.

The above conclusion implies that the estimated AP mine target neutralisation performance for the flail as well as for the tiller can be calculated using a sample of 450 test targets instead of a sample of 50 test targets producing a smaller confidence interval on the estimate and hence a more reliable estimate of the neutralisation percentage.

	Mini MineWolf flail	Mini MineWolf tiller
Neutralisation performance	448/450 95% sure that the real neutralisation percentage of the Mini MineWolf flail lies between 98.4% and 99.9%	442/450 95% sure that the real neutralisation percentage of the Mini MineWolf tiller lies between 96.5% and 99.2% (see confidence limit table)

(For details on the confidence interval limits see main body of the LL)

4.2.2. CWA 15044 performance test results for the Bozena-4 flail and Bozena-5 flail

Data taken from test reports Bozena-4 flail (ITEP Project 3.2.22, 2004) http://www.suffield.drdc-rddc.gc.ca/reports/English/DRDC_Suffield_TR_2005-138.pdf and Bozena-5 flail (ITEP Project 3.2.33, 2006) http://www.itep.ws/pdf/Bozena5_DRDC_2007.pdf

	Bozena-4 flail			Bozena-5 flail[5]		
	Sand	Gravel	Topsoil	Sand	Gravel	Topsoil
Flush	47/50	46/50	49/50	49/49	47/50	45/47
10 cm	48/50	47/50	48/50	50/50	48/50	49/50
15 cm	46/50	46/50	47/50	50/50	49/50	42/46

Using the graph corresponding to approach 1[6] and the data in the above table it can be concluded that the Bozena-4 flail as well as the Bozena-5 flail work as well in sand, gravel and topsoil for all mine burial depths up to 15 cm. When looking at the difference in performance between the Bozena-4 and Bozena-5 flail the only performance difference observed which is statistically different according to approach 1 is for test targets buried at 15 cm depth in sand. All other differences are not statistically significant.

	Bozena-4 flail	Bozena-5 flail
Neutralisation performance	424/450 95% sure that the real neutralisation percentage of the Bozena-4 flail lies between 91.6% and 96.2 %	429/442 95% sure that the real neutralisation percentage of the Bozena-5 flail lies between 95% and 98.4%
	Bozena-4 flail, sand, mines buried at 15 cm	Bozena-5 flail, sand, mines buried at 15 cm
	46/50 95% sure that the real neutralisation percentage of the Bozena-4 flail lies between 80.8% and 97.8%	50/50 95% sure that the real neutralisation percentage of the Bozena-4 flail lies between 92.9% and 100%

(For details on the confidence interval limits see main body of the LL)

However, using the other approaches the observed difference for mines buried at 15 cm depth in sand is not statistically different. The conclusion would therefore be that the Bozena-4 and Bozena-5 flail do not perform differently for all condition tested.

4.2.3. Comparing CWA 15044 performance test results for the Bozena flails and the Mini MineWolf flail

	Mini MineWolf flail			Bozena-4 flail		
	Sand	Gravel	Topsoil	Sand	Gravel	Topsoil
Flush	50/50	50/50	49/50	47/50	46/50	49/50

[5] Note that one or several test targets could not be recovered after the tests. It is therefore not known if it/they had been neutralized and hence it/they have been eliminated from the dataset in this example. In the test report, however, the missing targets were listed as life.
[6] The graph was used for the Bozena-4 flail but for the Bozena-5 flail the Chi-Square calculator at http://people.ku.edu/~preacher/chisq/chisq.htm was used as the data set is not complete, i.e. the total number of test targets was less than 50 for some of the test runs.

10 cm	50/50	49/50	50/50	48/50	47/50	48/50
15 cm	50/50	50/50	50/50	46/50	46/50	47/50

Using the graph corresponding to approach 1 and the above data it can be concluded that the observed differences in performance are statistically significant for flush buried targets in gravel, and for targets at 15 cm depth in sand and gravel. However, when the other approaches are used the latter differences are not considered to be significant.

	Mini MineWolf flail			Bozena-5 flail[7]		
	Sand	Sand	Sand	Sand	Gravel	Topsoil
Flush	50/50	50/50	49/50	49/49	47/50	45/47
10 cm	50/50	49/50	50/50	50/50	48/50	49/50
15 cm	50/50	50/50	50/50	50/50	49/50	42/46

Using the graph corresponding to approach 1 it can be concluded from the above data that the only observed performance difference which is statistically significant is the one for targets buried at 15 cm depth in topsoil. This difference becomes not significant however if the other approaches are used.

[7] Note that one or several test targets could not be recovered after the tests. It is therefore not known if it/they had been neutralized and hence it/they have been eliminated from the dataset in this example. In the test report, however, the missing targets were listed as life.