Environment Protection Authority

Sampling design part 2 - interpretation

Contaminated Land Guidelines

Acknowledgements

The NSW Environment Protection Authority acknowledges the contributions of the following organisations in the preparation of these guidelines:

Easterly Point Environmental Pty Ltd, JBS&G Australia Pty Ltd, Enviroview Pty Ltd, GHD Pty Ltd, Envirocene Pty Ltd, WSP Australia Pty Ltd, JK Environments Pty Ltd.

Cover: Prepared sample bottles and jars, sealed with EPA legal tape. Photo: EPA.

Published by:
NSW Environment Protection Authority
4 Parramatta Square
12 Darcy Street, Parramatta NSW 2150
Locked Bag 5022, Parramatta NSW 2124

Phone: +61 2 9995 5000 (switchboard)
Phone: 131 555 (NSW only – environment information and publications requests)
Fax: +61 2 9995 5999
TTY users: phone 133 677, then ask
for 131 555
Speak and listen users:
phone 1300 555 727, then ask for 131 555

Email: info@epa.nsw.gov.au
Website: www.epa.nsw.gov.au

Report pollution and environmental incidents
Environment Line: 131 555 (NSW only) or info@epa.nsw.gov.au
See also www.epa.nsw.gov.au

ISBN 978 1 922778 40 6
EPA 2022P3916
August 2022

Contents

Figures

Tables

1. Introduction

The NSW Environment Protection Authority (EPA) has prepared these guidelines to assist contaminated-land consultants, site auditors, regulators, landholders and developers, and inform members of the public who have an interest in the outcomes of the assessment and management of contaminated land. They will help consultants to design sampling for contaminated sites, regarding where samples are collected, how many samples are collected, and ways the data is compared to relevant criteria.

The guidelines will help users obtain data that is appropriately representative for the purposes of the sampling and the media being sampled, and analyse and interpret the collected data.

As when following any guidance, users should justify the approaches they use, and demonstrate that they are appropriate and fit for purpose.

The guidelines are in two parts. The first part describes the application of sampling design; the second part (this document) provides guidance on interpreting the results. This second part is not a stand-alone document and should be read in conjunction with *Sampling design part 1 – application*.

These guidelines have been made in accordance with the *Contaminated Land Management Act 1997* (CLM Act). They should be read in conjunction with the CLM Act, the Contaminated Land Management Regulation 2013 (CLM Regulation), and any guidelines made or approved by the EPA under the CLM Act.

The guidelines complement other guidelines made by the EPA, and several national guidance documents that have been approved by the EPA. Those guideline documents are listed in References and cited in the text where appropriate.

1.1. Scope of these interpretation guidelines

Section 2

Information on comparing sampling data to action levels. Appendix A includes a summary of common descriptive statistics, and Appendices F to L show associated procedures and worked examples.

Section 3

Summary of the main statistical distributions and information on associated data transformations and data analysis.

Section 4

Introduction of the concepts of hypothesis testing, including decision errors and methods for conducting hypothesis tests. Procedures for common methods of hypothesis testing, along with worked examples, are shown in Appendices F to L.

Section 5

Information on confidence intervals for use in estimation problems, along with the use of upper confidence limits of the mean ($UCL_{\bar{x}s}$) as another means of hypothesis testing. Appendices I to L provide procedures and worked examples for use of confidence intervals and upper confidence limits (UCLs), based on common distributions.

Section 6

Discusses trend analysis for temporal series of site contamination assessment data, including use of linear regression and the Mann–Kendall statistic.

Section 7

Discusses drawing conclusions from the data as part of the Data Quality Assessment process.

Section 8

Includes abbreviations and a glossary of technical terms.

Section 9

References for guidance and technical documents used in these guidelines.

1.2. Environmental media

These guidelines address sampling soil and solid media, as these are the most common targets when assessing site contamination. Information is also provided for other media, including groundwater, surface water, sediments and air.

Some statistical procedures described in these guidelines can be applied to these other media, although the EPA recommends that the following references are consulted when designing sampling programs for other media:

- NEPC 2013 – soil, groundwater and soil vapour
- ANZG 2018 – surface water
- DEC 2007 – groundwater
- EPA 2020a, DECCW 2010 and CRC Care 2013 – soil vapour
- Simpson and Batley 2016 – sediments.

This document does not specifically address biota sampling and ecotoxicity testing. For these areas, see the following references:

- ANZG 2018
- DES Queensland 2018
- DEC 2004.

2. Comparing data results to action levels

Schedule B1 of the *National Environment Protection (Assessment of Site Contamination) Measure 1999* (NEPC 2013) discusses the application of investigation and screening levels for Tier 1 assessments for soil results.

2.1. Use of statistics in the assessment of site contamination

Statistics can be broadly categorised as either **descriptive** statistics, which describe the sample, or **inferential** statistics, which relate the sample information to characteristics of the population. When assessing site contamination, both descriptive and inferential statistics are used to characterise sites and decision areas.

Descriptive statistics are discussed further in Appendix A. See the Glossary for more definitions of statistical concepts.

For **inferential statistics**, tests can be **parametric** or **non-parametric**. Parametric statistical tests make assumptions about the parameters of the population distribution, whereas non-parametric tests (sometimes called distribution-free statistics) make no assumptions about the distribution although they may make assumptions about the data.

All parametric statistical tests assume that the data are drawn from a particular probability distribution – normal, log-normal, gamma, or some other known statistical model. Parametric tests generally have non-parametric counterparts which can be used when the assumptions of the parametric test cannot be met. As non-parametric tests do not make assumptions about the distribution, they typically have lower statistical power than parametric tests in cases where the assumptions hold. However, non-parametric tests are often more accurate and powerful than parametric tests for even modest departures from parametric test assumptions.

Two assumptions that apply to many forms of inferential statistics are, first, that the sampling data are unbiased and, second, that each member of the population has an equal chance of being included in a sample. Consequently, the data points are an independent and identically distributed sequence of observations. **Independent** means that each observation is not controlled by the value of any other observation. Independence can generally be assumed for random samples if the sample consists of less than 10% of the population. **Identically distributed** means that the samples have been taken from a parent population whose mean and variance is stationary over the space and time of collection.

Biased sampling can be judgmental (also known as targeted) and arbitrary where certain observations are included or excluded because of some feature: this leads to members of the population having an unequal opportunity of being sampled. Bias can arise from a subconscious decision by the sampler.

Basic statistical tests can be validly applied only to unbiased sample data; data from judgmental or arbitrary sampling should not be used for statistical tests. For this reason, it is recommended that data obtained using a combination of judgmental and random (probabilistic) sampling approaches is collated and considered separately, and that the formal use of statistical techniques is confined to probabilistic sample data only. This means that results from judgmental sampling – for example, validation of an excavation, or investigation of a contaminating feature such as a leaking pipeline – should be removed from a dataset before statistical analysis is performed on the remaining data.

2.2. Descriptive statistics

Terms used in statistics are referred to as **statistical descriptors**. Commonly used statistical descriptors are the sample **range**, sample measures of central tendency (**mean**, **median** and **mode**), sample **percentiles**, and sample variability (**variance**, **standard deviation** and **coefficient of variation**). A preliminary data review could include basic graphical representations of the data, such as spatial plots,

box and whisker plots, frequency plots, histograms, ranked data plots, quantile–quantile (Q–Q) plots, two variable scatterplots, and temporal plots[1].

Common statistical descriptors can be used to summarise the basic quantitative characteristics of the sampling data, allowing them to be presented in tables or illustrated graphically. Where there are multiple populations or decision areas, it is useful to separate the data for analysis and comparison to reduce the variability of the individual datasets.

Reviewing the data numerically and graphically leads to a better understanding of the structure of the data, reveals patterns in distribution and relationships, and identifies potential anomalies. Data should be verified and validated before it is reviewed.

Descriptive statistics are further summarised in Appendix A, and specific procedures for determination and worked examples are included in Appendices B to D.

2.2.1. Software tools and packages

Statistical software tools and packages are available in spreadsheets, commercial software and open-access freeware. These can be used to determine both descriptive and inferential statistics. Freeware is particularly recommended as it allows other stakeholders including auditors and regulators easy access for checking outputs, without the associated financial costs and licence restrictions associated with commercial products.

A detailed review and summary of widely available statistical software packages can be found in Appendix D of ITRC 2013. This review covers both general statistical packages for broad applications and packages specifically designed for statistical analysis of environmental data, and includes both commercial and open-source freeware.

2.2.2. Data presentation

Spreadsheets and statistical software tools and packages can create sophisticated outputs to represent the sampling data and associated statistical information. For a preliminary data review, for example, they can present data in plan and cross section, both spatially and temporally, and as graphics.

As noted by DoE 1998:

> *While reporting of minima, maxima, mean, median, standard deviation, upper confidence limits etc. provides necessary information, such data may not be sufficient to characterise a site. The use of histograms or frequency distributions should also be considered to illustrate the distribution of results.*

Appendix E gives examples of graphical presentations that can be easily developed.

2.3. Maximums

The maximum observed value in a dataset is important in assessing site contamination, as a site or decision area is generally considered suitable for the intended land use if the maximum observed value is below the criterion or action level. However, such a condition may be misleading. The maximum observed value of the contaminant of interest is unlikely to be the maximum value present in the population, and the relationship between the two cannot be determined in the absence of statistical analysis.

Where sampling data is highly variable or based on small sample sizes, it may not be representative of the underlying population's variability and decision errors can arise. The recommended approach to control decision errors is to conduct appropriate tests that allow statistical inference. Appropriate tests include hypothesis tests, such as one-sample t-tests and UCL$\bar{x}$s. Section 4 and Appendix F discuss hypothesis testing; Section 5 and Appendices I to L for skewed distributions discuss UCLs.

[1] See Sections 14.3–14.6 of Schedule B2 of NEPC 2013, USEPA 2006a G-9S and USEPA 2006c G-9R for more details.

When comparing sample results to criteria and action levels, the sampling data needs to show that no single value exceeds 250% of the relevant investigation or screening level (schedule B1, NEPC 2013).

2.4. Outliers

In statistics, **outliers** are data points that do not fall into the expected range of a defined probability distribution function. In the context of site contamination assessment however, the characteristics of a probability distribution function of a contaminant can be difficult to define. Complex historical site uses can result in the superposition of multiple probability distribution functions. **Hotspots** – small areas of relatively high concentration – may also be present, with their own probability distribution function.

Discarding an outlier from a dataset should be done with extreme caution as environmental datasets often include legitimate extreme values (USEPA 2006b). A more thorough examination of the reasons for any unexpectedly high values may lead to new insights into the data such as the presence of an unsuspected hotspot of contamination, or to reconsidering underlying assumptions about the data and its distribution.

All data resulting from probability-based sampling must be included in the subsequent inferential statistical analysis, unless:

- it can be demonstrated with a high level of confidence that the individual data points are invalid due to transcription errors, data coding errors, or measurement errors in the laboratory analysis

  or
- the individual data points are subsequently identified – again, with a high level of confidence – as part of a hotspot, and the hotspot is appropriately remediated or managed and thereby effectively removed from the population.

In either case, a determination is then needed as to whether further data needs to be generated through additional investigations, or if sufficient data is available to support the required decisions. These determinations should include appropriate statistical analysis of the remaining dataset.

2.5. Non-detects

As part of the assessment of site contamination, where the concentration of an analyte ranges between zero and the limits of reporting (LORs) of the laboratory method, the results are reported as less than the LORs. This is referred to as **left-censored** data. In some instances, the data below the LORs may represent another **population**, and the data, including geological logs and field notes, should be reviewed to determine if a more appropriate grouping of data is relevant. For determining mean values, mixing a large number of results below the LORs with a limited number of detected results can lead to estimation problems if simplistic methods are used.

There are various imputation methods to replace these censored values. **Direct substitution** is the easiest but least satisfactory. Generally, substitution should only be adopted where the fraction of the sample that is censored is relatively small. With substitution, a constant value is assigned to the non-detects by one of the following:

- assuming the non-detects are equal to zero
- assuming the non-detects are equal to the LORs
- assuming the non-detects are equal to some fraction of the LOR, usually one half.

The proxy value is then used as though it were the value for that measurement. However, the uncertainty associated with the substitution method increases as the proportion of non-detects in the dataset increases. Statistical determinations and inferences associated with censored data become increasingly problematic, because of errors in the estimates of parameters such as the mean, which becomes biased down. The direction and extent of the bias in variance is highly dependent on the data and substituted value.

Some statistical software packages have methods that enable a user to enter data and indicate that it is a non-detect; the software calculates statistical parameters such as 95% UCL and standard deviations for the dataset, and the output of the statistical package provides guidance on which method is recommended (USEPA 2015a). A worked example is provided in Appendix M.

If statistical software is unavailable, Section 4.7 of USEPA 2006a provides more detailed guidance for analysing data with non-detects. If the direct substitution method described above is used, results for the three substitutions listed above – zero, equal to LOR and assumed fraction of LOR – should be reported.

Where non-detects below LORs exist::

- always report detection limits for non-detects
- do not convert non-detects to zeros without specific justification
- consider using non-parametric methods if further statistical analysis is required.

Other methods of imputation – replacing data with substituted values – include **multiple imputation**, **fractional imputation** and **Bayesian modelling**. The appropriate imputation method depends on the size of the dataset and the proportion of measurements reported as non-detects. If the proportion of non-detects is high (> 50%) or the number of samples is small (n < 5), analysis may be challenging.

The method of **maximum likelihood** by first principles can be used to estimate the parameters of a probability distribution even where there is censored data: for censored data points, summing is replaced by integration between limits (zero and LOR). In general, the point of maximum likelihood cannot be determined algebraically but must be solved numerically (for instance, with the hill climbing technique or Newton Rapson technique), though this is no longer an issue for those with access to desktop computing.

Additionally, there are various statistical packages dealing with censored data that are suitable for laboratory measurements.

Refer to ITRC 2013 and USEPA 2015a for details of specific methods for managing non-detects in statistical analyses.

Wendelberger and Campbell 1994 note that:

> *[t]he manner in which the nondetect values are handled should depend on the type of decision to be made and the magnitude and frequency of the nondetect values. If the nondetects are small in magnitude or low in frequency, the method of handling the nondetects will probably have minimal impact on the final outcome of the analysis. However, if the detection limits are close to important decision values, or if the frequency of nondetects is high, the treatment of the nondetect values can greatly influence resulting decisions.*

Whichever statistical approach is adopted, the conceptual site model (CSM) should be re-developed to reflect the proportion of the dataset samples that are non-detects. For instance, if a site has a few detections and many non-detections, the source of the contamination should be carefully considered when refining the CSM. An option might be to stratify the site so areas where there are widespread non-detections are assessed separately from areas with detections, especially if investigation levels are being exceeded.

2.6. Pseudoreplication

In site contamination assessment, the collection and analysis of duplicate and triplicate samples is conducted as part of quality assurance/quality control (QA/QC) programs. Although this is important for determining the data's usability, these replicate sample results must not be treated as an independent sample. Doing so is known as **pseudoreplication** because the duplicates and triplicates are not independent of the primary sample. Pseudoreplication increases the number of samples while providing another data point similar to the primary sample, resulting in bias and distortion of any statistical analysis being undertaken.

2.7. Contaminant distribution

The variation of contaminant concentrations over a site or decision area means that individual measurements cannot be used to fully describe the distribution of a contaminant. If the contaminant concentrations are plotted against their respective frequency of occurrence, the resulting curve or histogram represents the concentration distribution of that contaminant over the site or decision area.

While histograms inform the characterisation of the site or decision area, they do not represent spatial information across the site or decision area. They show the range, central tendency, variation and

distribution of the variables being considered. Under the **multiple lines of evidence and weight of evidence** approach, these parameters should be considered when interpreting the data and comparing it to the criteria or action levels.

For example, comparing the sampling results to 250% of the relevant investigation or screening level can lead to identifying apparent hotspots and the recommendation that removing these areas will make the site or decision area suitable for the proposed land use. However, closer examination of the data may show that the apparent hotspots relate to heterogeneity of the soil or fill and any subsequent validation would result in identifying other 'hotspots'. In these situations, more characterisation to confirm the variability of the soil or fill may provide better information for decision making on remediation or management. In such situations, the use of statistical tools can assist, particularly in relation to decision errors and determining a suitable number of samples.

Schedule B1 of NEPC 2013 requires that sampling results should be checked to ensure the standard deviation of the variable is less than 50% of the relevant investigation or screening level. Although 50% is an arbitrary value, it provides a warning that the variance is potentially excessive, prompting further review of the contaminant distribution.

In these cases, further segregating the data by, for example, depth, soil type or spatial distribution may demonstrate that multiple populations are inappropriately being considered as a single population. Alternatively, the data may indeed represent a highly variable population and indicate more sampling is needed.

3. Distributions, transformations and data analysis

The sampling distribution is the frequency or probability of occurrence of measured values. In the assessment of site contamination, data can be analysed using parametric (distribution based) methods, or non-parametric methods where the population is not assumed to fit a specific population distribution. Statistical software packages provide more complex calculations of UCL$\bar{x}$ using a number of parametric and non-parametric distributions, however a brief review of the predominant distributions used is warranted.

Where the sampling data has a normal (or more strictly, nearly normal), log-normal or gamma distribution, parametric methods can be applied. Where the sampling data does not have such a distribution, non-parametric methods should be used. Non-normal datasets can have a transform applied to essentially normalise the data, which aids analysis.

3.1. Parametric methods

Population parameters are estimated from samples. Different random samples will produce different estimates of each parameter; for instance, each sample will produce a different estimate (using $\bar{x}$) of the population mean, μ. These estimates themselves have a distribution, known as the **sampling distribution**. Many common statistical methods are based on a knowledge of, or the assumed characteristics of, the sampling distributions of population parameter estimates.

3.1.1. Normal distribution

The most commonly used distribution in parametric statistics is the normal. The central limit theorem (CLT) states that the sampling distribution of the mean for **n** independent random samples approaches a normal distribution as **n** increases. This holds for all population distributions with finite mean and variance. With normal distribution, the mean, median and mode are equal.

Based on the CLT, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution when the sample size n is sufficiently large (> 30), irrespective of the shape of the population distribution. The larger the value of n, the better the approximation (Devore and Farnum 2005).

Appendix J demonstrates the one-sided Student's t-test method used with normal distribution.

3.1.2. Log-normal distribution

Log transformations convert samples to natural log values to allow the use of log-normal or exponential distributions for analysis. The log-normal is a continuous distribution in which the logarithm of a variable has a normal distribution. Thus, if the random variable x is log-normally distributed, then y = ln(x) has a normal distribution. Likewise, if y has a normal distribution, then the exponential function of y (that is, x = exp(y)) has a log-normal distribution.

In log-normal distributions, the mean, median and mode are not equal. The difference between mean and mode depends on the skewness of the population, while the median is independent of skewness.

Appendix K gives a worked example of the Land's H-statistic method used with log-normal distribution.

3.1.3. Gamma distribution

Gamma distribution is more flexible for fitting data than normal and log-normal distribution. Gamma distribution is a rank-order transformation where contaminant concentration data is sorted into ascending order and converted to an integer ranked list. This process eliminates the scale effects in contaminant concentrations commonly found in site contamination datasets and reduces the effect of large differences between results in a dataset.

This distribution type is relevant to the assessment of contaminated sites due to its relationship to exponential and normal distributions. The **gamma distribution** is a two-parameter family of continuous

probability distributions. The exponential distribution and the chi-squared distribution are special cases of the gamma distribution.

The three common parametrisations for gamma distributions are:

- a shape parameter k and a scale parameter θ
- a shape parameter α = k and an inverse scale parameter β = 1/θ, called a rate parameter
- a shape parameter k and a mean parameter μ = kθ = α/β.

In each of these, both parameters are positive real numbers and control the shape and skewness of the distribution.

3.1.4. Parametric methods in the analysis of site contamination assessment data

When using distributions to assess site contamination data, the limitations imposed by each distribution must be accounted for as these determine how well the distribution can provide a reliable interpretation of the actual population.

Data for assessing contaminated sites is rarely normally distributed, due to the kind of processes that lead to site contamination. When the mean, median and mode are not equal, or the coefficient of variance is > 0.5, consider carefully before using the normal distribution for analysis. Similarly, be cautious when applying the log-normal distribution, as the data for assessing contaminated sites is often not truly log-normally distributed. The application of either distribution needs to the be verified by testing that the data is approximately normally distributed, or normally distributed after the log transform is applied. The distribution can be tested by using a statistical software package to construct quantile–quantile (Q–Q) plots, which graph the quantiles of the dataset against the quantiles of a specific probability distribution.

It is generally recommended that skewed datasets are assessed using a gamma distribution rather than a log-normal distribution, as this produces more reliable results. A log-normal transformation disguises the effect of high values that may not represent background and exaggerates the apparent standard deviation of the modelled log-normal distribution. This increases the risk of making an incorrect decision in relation the population distribution and associated statistical parameters. Therefore, for assessing skewed site contamination datasets, the gamma distribution should be used when performing parametric analysis, particularly if the sample size is less than 20 and/or contains outliers. Because of the gamma function's flexibility in accommodating a wide range of symmetric and asymmetric (skewed) distributions, it can represent log-normally distributed datasets without the risk of masking the effects of outliers.

When the site assessment data is highly skewed by extreme values or a significant number of non-detect values, it may be hard to determine an appropriate distribution for parametric analysis. In such cases non-parametric methods may give more reliable results.

3.2. Non-parametric methods

Non-parametric statistics are analysis methods that either make no assumption about the distribution of the data or the population, or, where a specific distribution is assumed, do not specify the distribution's parameters. Commonly used non-parametric methods for making inferences in the assessment of site contamination data are the **bootstrap**, **jackknife** and **Chebyshev** methods.

Compared with non-parametric methods, analogous parametric methods are usually more effective when the assumptions of parametric methods hold. Where a population departs from these assumptions, the non-parametric tests can be superior.

3.2.1. Bootstrapping

Bootstrapping involves estimating properties of a statistical parameter by measuring those properties through randomly re-sampling the dataset with replacement data. Data points need to be independently and identically distributed. This 'new' dataset is then used to estimate the statistical parameters such as mean, median, mode and standard deviation. Bootstrapping can also be used for constructing hypothesis tests, as an alternative to statistical inference based on the assumption of a parametric model, when that assumption is in doubt.

Bootstrapping, like any non-parametric resampling method, offers a useful means of reducing the influence of extreme outliers on the overall statistical parameters of the underlying sampled population. However, caution is needed to avoid diminishing the importance of outliers in relation to the overall decision: where the outlier represents a hotspot, a non-parametric re-sampling method such as bootstrapping may not be appropriate. The use of this method must therefore be justified in the context of the importance of the outliers to the overall decision being made.

3.2.2. Jackknifing

Jackknifing is similar to bootstrapping, in that the method re-samples the dataset and generally produces similar results, although instead of making random replacements, it randomly removes a sample in each resampling step. The re-sampled dataset can then be analysed with the same methods as those used for bootstrapping. Jackknifing is subject to the same limitations and cautions as bootstrapping.

3.2.3. Chebyshev

The Chebyshev method is a non-parametric method that does not involve resampling the dataset but instead relies on use of the Chebyshev's inequality. This specifies that, for all distributions with finite mean and variance, only a certain fraction of values can be more than a certain distance from the mean, that is, no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean with 100% certainty.

This inequality can be applied to any probability distribution in which the mean and variance are defined. When applied to datasets for assessing site contamination, the Chebyshev method can determine statistical parameters, particularly the mean, for highly skewed datasets or ones that contain significant outliers. In most applications, the Chebyshev method gives a more conservative result than other parametric and non-parametric methods.

Appendix L gives a worked example of the Chebyshev method.

4. Hypothesis testing

Decision problems can be addressed as statistical hypothesis tests, which are recommended under the USEPA's DQOs (data quality objectives) process. The **null hypothesis significance testing** (NHST) framework, derived from approaches for testing data, is a method of statistical inference used to determine if a null hypothesis (H_0) should be rejected in favour of an alternative hypothesis (H_A) at a specified level of confidence. In the assessment of site contamination, H_0 is that the site or decision area is not suitable for the specified use, i.e. that the site or decision area is contaminated.

The basis of the hypothesis test is that H_0 can be rejected where the findings are incompatible with H_0 being true, in which case H_A is more likely. Alternatively, H_0 can fail to be rejected, which does not necessarily mean that H_0 is true but that there is not enough evidence to reject the site being contaminated.

Before testing, an environmentally significant difference from the criterion level should be established. For H_0 to be rejected, the data must show, with given confidence, that the population parameter is at or below this level. This environmentally significant difference is greater or equal to zero to provide an environmental buffer.

The most common form of hypothesis testing is for nearly-normally distributed populations, where estimated population means are tested using the Student's t test (t-test) which is used to test for differences in population means. This test can be:

- a **one-sample t-test**, to test whether the mean of a single population is different from a target value, such as a specified health investigation level (HIL)
- a **two-sample t-test**, to compare the means of two groups, such as site data and background data
- a **paired sample t-test**, to compare the means from the same group at different times, such as before and after remediation.

If there are non-detects, more work is required to estimate the mean and variance.

Worked examples are shown in Appendix F (a one-sample t-test) and Appendix G (a two-sample t-test).

There are also comparable parametric methods for non-normal distributions, and non-parametric methods for testing differences in means and/or medians in unknown distributions (see USEPA 2006a, G-9S).

4.1. Sampling uncertainty and decision errors

Uncertainty in estimates is unavoidable due to, for example, inherent variability in the characteristics of interest of the target population, the limits on the number of samples that can be collected and imperfect measurements. Statistical methods provide quantitative tools for characterising an estimate's uncertainty, and help in designing an investigation that will generate probabilistic data of a sufficient type, quality and quantity.

One can never be 'certain' about an answer derived from sampling, so the uncertainty must be specified for a statistical statement to have meaning. In statistics, uncertainty is technically referred to as **risk** or **confidence level**. The risk of incorrectly rejecting H_0 is denoted by α (alpha) and has a magnitude of between 0 and 1. The risk of incorrectly accepting H_0 is denoted by β (beta), which is also between 0 and 1. For example, if a particular statistical statement is quoted as having a 95% confidence level, ($\alpha = 0.05$), this implies that at least 95 out of I00 repeats of the sampling will correctly accept a true H_0. A power of 80% ($\beta = 0.2$) means an 80% chance of correctly rejecting a false H_0.

In the assessment of site contamination, α risk is the risk of deciding that the site or decision area is suitable for the proposed use when in fact it is not, and the confidence level is always equal to 1 - α. The probabilities generally used in the assessment of site contamination are $\alpha = 0.05$ and $\beta = 0.2$, or a 95% confidence level and a statistical power of 80%, although higher probabilities can be used, such as $\alpha = 0.01$ and $\beta = 0.1$, or a 99% confidence level and a statistical power of 90%.

Changing one probability inevitably changes the other. One way to obtain both a high confidence level and high statistical power is to increase the number of samples. More sophisticated sampling designs and associated analysis can also be used to increase the power – see Section 4.

Within hypothesis testing, **decision errors** refer to the incorrect decisions that can be made about a site or decision area, based on the data collected. They arise from using data that are not sufficiently representative of the site or decision area due to sampling errors, measurement errors or more commonly, both. Such errors can lead to decisions that assess contaminated land as uncontaminated when it is contaminated, or that determine that remediation is required when it is not. The combination of errors from all sources is referred to as the **total study error**, and directly affects the probability of making decision errors. The statistical theory behind hypothesis testing allows the probability of making a decision error to be quantified, given the data collected and the specified level of significance.

Decision errors result from:

- **sampling errors**, which arise from using information from a sample instead of measuring the whole population
- **sampling design errors**, which arise when the sampling design does not validly capture the structure of the population – they include sampling frame selection, sampling unit definition, selection probabilities and the number of samples collected
- **measurement errors**, which arise from the variability inherent in sample collection, handling, preparation, analysis and data reduction.

Study error is managed by correctly choosing suitable sampling designs and measurement systems.

See Appendix H for more information on the types of decision errors.

4.2. Use of hypothesis tests

Formal statistical methods can quantify the uncertainty associated with decisions. ITRC 2103 notes common decision errors when assessing site contamination, and hypothesis tests that can control them:

- **concluding that a site or decision area is suitable when the sample maximum is less than the criterion or action level.** For some distributions and sample sizes, the population mean of the site or decision area may be greater than the criterion or action level, even though a particular sample maximum is less than the criterion or action level.

  This is a Type I decision error, and a one-sample hypothesis test will allow statistical inference and control of decision errors.
- **concluding that a site or decision area is not suitable when the sample maximum is more than the criterion or action level.** The population mean of the site or decision area may be less than the criterion or action level when the sample maximum is more than the decision criterion, depending on the nature of the distribution and the sample size.

  This is a Type II decision error, and a one-sample hypothesis test will allow statistical inference and control of decision errors. The systematic planning for the investigation should describe how maximum values will be treated – for instance, what further data analysis or investigation will be carried out if the maximum value exceeds 250% of criterion or action level.
- **concluding that the failure to reject the null hypothesis 'proves' the null hypothesis (i.e. that the site is too contaminated to be acceptable).** As environmental data typically shows large random variability, the sample could include a preponderance of elevated concentrations, particularly if the sample size was small and so the statistical test is of insufficient power.

  The power of the test ($1 - \beta$) should be determined and compared to the decision criteria, and/or the number of samples required to achieve the specified decision criteria determined using the combined risk value (CRV) method discussed in *Sampling design part 1 – application*. If not enough samples were collected (i.e. the test was conducted with insufficient power) further data analysis or investigation may be required
- **directly comparing the maximum value of a site or decision area with a background maximum or mean, without considering potential decision errors.** The maxima from the two datasets should not be compared to make inferences about the means of the datasets, as decision errors are

not controlled for and this can result in Type I errors. A two-sample hypothesis test is recommended to allow statistical inferences and control decision errors.

5. Confidence intervals and upper confidence limits

A **confidence interval** estimates a population parameter from sample data and is composed of two parts: an interval calculated from the data and a confidence level associated with the interval. In the assessment of site contamination, the confidence interval is generally expressed as a point estimate, usually the mean plus and minus (±) the margin of error. Because confidence intervals are expressed in this way, they are determined using two-sided intervals for the t critical values.

Upper confidence limits (UCLs) are the upper component of the confidence interval and are determined using the one-sided interval for the t critical values.

Hypothesis tests and confidence intervals are related, as they are determined using variations of the same formula, and often a confidence interval can be used to test a hypothesis, making it unnecessary to perform the entire hypothesis test. In assessing site contamination, a decision is generally only required on whether the estimated population parameter exceeds the criterion or action level, so UCLs can be used by themselves as a form of hypothesis testing.

5.1. Confidence intervals

Performance criteria are needed to estimate an unknown parameter to within a specified amount, with a given confidence level: they specify the maximum width of the confidence interval. The width of a confidence interval depends on the number of samples used to calculate the interval, the precision or variability of the dataset and the specified confidence level. Placing limits on the maximum width of a confidence interval enables the precision and the number of samples needed to calculate it to be determined. As the variability of the population being studied is generally fixed, only the confidence level and number of samples can be controlled.

For independent samples from an approximately normal distribution, or where the sample size is large ($n \geq 30$), confidence intervals for mean values are determined by using the one-sample Student's t-test. This test is reasonably robust if the population distribution deviates only moderately from normality; however, for highly skewed datasets with significant outliers, or where significant non-detects are included in the dataset, other distributions or non-parametric methods should be used.

Appendix F shows how to determine confidence intervals using the one-sided Student's t-test and gives a worked example. For other distributions not discussed below, or for non-parametric methods, see USEPA 2006a, G-9S.

5.2. Upper confidence limits

When assessing site contamination, the main way to determine if sites or decision areas are suitable for their proposed uses is to employ UCLs as one-sided hypothesis tests for comparing the sample mean to the action levels or criteria. The appropriate method is determined by the population distribution, as indicated by the sampling data.

- For normal distributions, Appendix J demonstrates the one-sided Student's t-test method.
- For log-normal distributions, Appendix K demonstrates the Land's H-statistic method.
- For skewed distributions, Appendix L demonstrates the Chebyshev inequality method.

6. Trend analysis

Trend analyses are used in the assessment of site contamination to determine if a contaminant's concentrations are increasing, decreasing or remaining constant over time. The objective of a trend analysis is to determine if the changes of a contaminant concentration can be statistically correlated to time and, if so, how significant the correlation is. The two trend analysis methods described below are generally applied to datasets for water or air, although they can be used for assessing remediation, such as the bioremediation of soil.

6.1. Linear regression

The calculation of a linear regression, or line of best fit, is a common way to measure the relationship between two variables. In the assessment of site contamination, a linear regression analysis is often used to assess if there is a trend between a contaminant concentration and time, for example, is the concentration of benzene in monitoring well two decreasing over time? The need for temporal trend analysis and the minimum number of data points required depends on the CSM, which should be developed based on site specific characteristics.

Data should be presented on a time plot to determine, visually, if a trend is likely, then the r-value (or Pearson Correlation Coefficient) should be calculated. This is a measure of the strength of the linear relationship between the two variables. The r-value can be a value between 1 and -1, with 1 indicating a strong positive relationship between the two variables, -1 indicating a strong negative relationship, and 0 indicating no relationship at all.

While the calculation of an r-value of 0.98 may indicate a strong positive relationship between the contaminant concentration and time, other factors could be affecting this relationship. Simple linear regressions can be affected by seasonality, the distribution of the data and the number of samples below the LORs. USEPA 2006a, G-9S states that due to these limitations, linear regressions are not generally recommended for estimating and detecting trends but can be used as an informal and quick screening tool to detect if a strong linear trend is present.

6.2. Mann–Kendall

The Mann–Kendall test is used to assess trends in datasets, and being a non-parametric test, it makes no assumption regarding data distribution and is unaffected by missing data or values below the LORs.

The test compares each data point against the next data point, and a score of 1 or -1 is given for each comparison, according to whether there is an increase or decrease in concentration. The test is not affected by the magnitude of the change.

The individual scores are tallied to provide the Mann–Kendall statistic (S): a positive S indicates an upward trend whilst a negative S indicates a downward trend. The value of S is then compared to an S-critical value. A p-value is then calculated for comparison with the adopted significance level, which determines if the null hypothesis (of no trend) is rejected or accepted.

The Mann–Kendall test is also affected by seasonality, and only data from similar months each year should be compared if this is likely to be important. Where high seasonality effects can be expected, to calculate a meaningful result means collecting data for at least four years.

The output of the Mann–Kendall test will be:

- the concentrations are increasing, or
- the concentrations are decreasing, or
- there is no trend.

Following this test, a linear regression analysis can be performed to determine the strength of the trend, providing the potential limitations of the linear regression are considered. Further information on the use of the Mann–Kendall test to assess trends can be found in Gilbert 1987, USEPA 2006a, G-9S and IRTC 2013.

7. Drawing conclusions from the data

Once the investigation has been conducted and validation data have been collected for a project, the consultant should consider if the data quality objectives have been met by referring to the data quality indicators determined during the DQO process.

The consultant should document any statistical calculations clearly, evaluate the results and draw conclusions. If warranted, the CSM for the site should be updated to incorporate any analytical data. These should be presented in an assessment report prepared in accordance with EPA 2020b.

Further details on drawing conclusions in accordance with the DQO process are provided in USEPA 2006a.

Guidance on assessment for media other than soil or fill is provided in:

- NEPC 2013 – soil, groundwater and soil vapour
- ANZG 2018 –surface water
- DEC 2007 – groundwater
- EPA 2020a, DECCW 2010 and CRC Care 2013 – soil vapour
- Simpson and Batley 2016 – sediments.

See EPA 2020b for advice on preparing reports.

8. Abbreviations and glossary

8.1. Acronyms and abbreviations

ABC	Ambient background concentration
ANZG	Australian and New Zealand water quality guidelines
CECs	Contaminants of emerging concern
CLM	Contaminated land management
CLT	Central limit theorem
CoPC	Contaminants of potential concern
CRV	Combined risk value
CSM	Conceptual site model
CV	Coefficient of variation
DNAPLs	Dense non-aqueous phase liquids
DQIs	Data quality indicators
DQOs	Data quality objectives
DSI	Detailed site investigation
DUs	Decision units
EPA	Environment Protection Authority
HIL	Health-based investigation level
HSL	Health screening level
ISM	Incremental sampling methods
LNAPLs	Light non-aqueous phase liquids
LOR	Limits of reporting
Metals	Arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), lead (Pb), mercury (Hg), nickel (Ni) and zinc (Zn)
MoE	Margin of error
MPE	Maximum probable error
MQOs	Measurement quality objectives
NEPM	National Environmental Protection Measure
NHST	Null-hypothesis significance testing
NOW	New South Wales Office of Water
OEH	New South Wales Office of Environment and Heritage
PAHs	Polycyclic aromatic hydrocarbons
PFAS	Per- and poly-fluorinated alkyl substances
PFHxS	Perfluorohexane sulfonate
PFOS	Perfluorooctane sulfonate
PFOA	Perfluorooctanoic acid
PSH	Phase-separated hydrocarbon
PSI	Preliminary site investigation

PID	Photoionisation detector
PTFE	Polytetrafluoroethylene
QAPP	Quality assurance project plan
QA/QC	Quality assurance/quality control
Q–Q	Quantile–quantile
RAP	Remediation action plan
RSD	Relative standard deviation
SAQP	Sampling and analysis quality plan
SOPs	Standard operating procedures
STP	Sewage treatment plant
SWL	Standing water level
TOFA	Total organic fluorine assay
TOPA	Total oxidisable precursor assay
TRHs	Total recoverable hydrocarbons, including volatile C6–C10 fractions and semi- and non-volatile C11–C40 fractions
UCLs	Upper confidence limits
UCL$\bar{x}$	Upper confidence limits of means
UPSS	Underground petroleum storage system
USEPA	United States Environmental Protection Agency
UST	Underground storage tank
VOCs	Volatile organic compounds

8.2. Statistical notations

$1 - \alpha$	Confidence level
α	Type I error rate (see Glossary)
β	Type II error rate (see Glossary)
c	Criterion/action level
df	Degrees of freedom
exp	Exponential function
H_A	Alternative hypothesis
H_0	Null hypothesis
n	Number of samples or measurements in a sample (see sample definition)
θ	Scale parameter of the gamma distribution
σ	The population standard deviation, which is generally not known
σ^2	The population variance, which is generally not known
p-value	Probability value
Δ	Uppercase Greek letter delta, denoting the width of the grey region associated with hypothesis testing
s	The sample standard deviation, which is determined from the measurements taken
s^2	The sample variance, which is determined from the measurements taken
δ_0	Difference (delta) of zero

$t\alpha$	Critical value
t_0	Test statistic
μ	The population mean, which is generally not known
$UCL\bar{x}$	Upper confidence limit of mean
$\bar{x}$	The sample mean, which is determined from the measurements taken
x_i	The i^{th} measurement in the dataset

8.3. Glossary

α risk

The probability, expressed as a decimal, of making a Type I error when a hypothesis is tested statistically. A Type I error wrongly rejects a null hypothesis when in fact the null hypothesis is true. In this document, the null hypothesis always assumes that the site is contaminated and thus the α risk refers to the probability of a site being validated as uncontaminated when it is contaminated.

β risk

The probability, expressed as a decimal, of making a Type II error when a hypothesis is tested statistically. A Type II error wrongly accepts a null hypothesis when in fact the null hypothesis is false. In this document, the null hypothesis always assumes that the site is contaminated and thus the β risk refers to the probability that a site is determined as contaminated when it is uncontaminated.

Acceptable limit

A threshold concentration value below which the level of contamination is regarded as acceptable. An acceptable limit can either be adopted from the appropriate guidelines or it can be derived on a site-specific basis using risk assessment. Where site remediation is involved, acceptable limits are often referred to as ‘clean-up standards’ or ‘remediation standards’.

Acceptance criteria

A statistical statement specifying how a contaminant distribution will be compared with an acceptable limit (see above definition) to determine whether a site should be evaluated as ‘contaminated’ or ‘uncontaminated’. Contaminant concentrations can vary over orders of magnitude in a sampling area. All site assessments must state the appropriate acceptance criteria, as well as the appropriate acceptable limits.

Ambient air

External air environment, not including the air inside buildings or structures.

Arithmetic mean

The arithmetic mean is commonly referred to as the average and is used to describe the centre of the data distribution. It is obtained by adding all the values and dividing the result by the number of values.

Central tendency

The central or typical value for a probability distribution, and may be considered the average value in a dataset. It is generally described as the mode, median, or, more commonly, the mean, and describes where a sample distribution is centred.

Chi-squared distribution

A type of cumulative probability distribution that varies depending on the degrees of freedom (df). It is used to test relationships between categorical variables in the same population.

Coefficient of variation (CV)

CV is the measurement of the relative homogeneity of a distribution. Low CV values, for example 0.5 or less, indicate fairly homogeneous contaminant distribution, while CVs with values of more than 1–1.2 imply that the concentration distribution of a contaminant is heterogeneous and probably highly skewed to the right.

Composite sample

The bulking and thorough mixing of soil samples collected from more than one sampling location to form a single soil sample for chemical analyses.

Conceptual site model (CSM)

Provides a three-dimensional overview of the contamination at sites and their surroundings, highlighting the sources, receptors and exposure pathways between the sources and receptors.

Confidence level

The probability, expressed as a percentage, that a statistical statement is correct. Confidence level is the opposite expression of 'risk' (see definitions of α and β risks). In this document when a risk that needs to be regulated, the confidence level is always equal to I - α.

Contaminated

For the purpose of this document and depending on the context, 'contaminated' can have slightly different meanings. If a site or a sampling area is evaluated as contaminated, it means that it has not met the acceptance criteria (see definition of acceptance criteria). Contaminated can also describe a localised area or soil that has contaminant concentrations exceeding an acceptable limit (see definition of acceptable limit). Note: depending on what the acceptance criteria are, an entire site could be considered uncontaminated even though a certain percentage of it is expected to be contaminated.

Contamination

NEPC 2013 defines contamination as 'the condition of land or water where any chemical substance or waste has been added as a direct or indirect result of human activity at above background level and represents, or potentially represents, an adverse health or environmental impact'.

Data quality objectives (DQOs)

A systematic planning process that defines the type, quantity and quality of data needed to support decisions on the environmental condition of a site or a specific decision area.

Decision area

A specific area or medium on a site, or offsite, about which data is being gathered so a decision can be made. For example, a decision can be made on part of a site, soil, a stockpile, soil gas, groundwater, surface waters or sediments.

Estimate

An estimate is a value that is inferred for a population based on data collected from a sample of units from that population. For example, the measured data from a sampling event used to calculate the sample mean ($\bar{x}$) is then used to estimate the population mean (μ).

Estimation

A technique that systematically adjusts the sample data to determine an estimated value for the population.

Geometric mean

This is similar to the **arithmetic mean** (described above), in that it is also a measure of the central tendency of the distribution of a population or sample. It is sensible to calculate geometric means only on

populations or samples that contain positive values. The geometric mean is obtained by multiplying n values from the dataset together, then taking the nth root of the product.

Grab samples

Samples collected from different locations that will be analysed individually.

Hotspot

A localised area where the level of contamination is noticeably greater than in surrounding areas. Note that a hotspot is only **relatively** high in contamination.

Inter well

Comparison between two groundwater monitoring wells that are separated spatially.

Intra well

Comparison of measurements over time at one groundwater monitoring well.

Maximum

The maximum observed value in a data. It generally provides a conservative estimate of the potential exposure risks so if the maximum is below the action level, the site should be suitable for its proposed land use.

Median

The middle value of the distribution. Half the data values are less than the median and half are greater.

Minimum size effect

The acceptable magnitude of the difference between the populations or groups being studied.

Mode

The value that occurs most frequently. It is determined by counting the number of times each value occurs.

Modules

A series of discrete DQOs outputs, based on logical categories, that addresses selected components of a site investigation. Modules can be selected for contaminant types, media, decision areas, or a workable combination of these.

Neyman–Pearson method

A method of statistical inference used to determine if a null hypothesis (H_0) should be rejected in favour of an alternative hypothesis (H_A), at a specified level of confidence.

Outlier

A data point that sits outside the expected range of the data. An outlier can have either a high or low value. Outliers must be retained in sample datasets unless there is a demonstratable reason for rejecting them such as a coding error, sample contamination or equipment failure.

Parameters

Numerical measures of the characteristic of interest in the population being sampled. Typical parameters are the population mean (μ), variance (σ^2) and standard deviation (σ). Parameter values are usually unknown.

Percentiles and quartiles

These are descriptive values used to equally split a dataset into 100 parts. A percentile is the value that a percentage of observations in a dataset is equal to or less than, for example, 80% of observations in a dataset are at or below the 80th percentile, while 20% are above.

Quartiles are commonly used to break the dataset up into four equal parts, providing an indication of the distribution and variance of the data.

- First quartile – the 0^{th} percentile up to (and including) the 25^{th} percentile.
- Second quartile – from the 25^{th} percentile up to (and including) the 50^{th} percentile.
- Third quartile – from the 50^{th} percentile up to (and including) the 75^{th} percentile.
- Fourth quartile – from the 75^{th} percentile up to (and including) the 100^{th} percentile.

Population

Any large collection of objects, things or individuals with some characteristics in common, that is being studied and for which information is sought. The population under consideration must be clearly and succinctly defined to allow effective sampling design and subsequent reporting.

The population can be further defined as the **target population** and the **sampled population**, and ideally these should be the same. The target population is the set of all units that comprise the items of interest, that is the population about which a decision is required, and the sampled population is that part of the target population that is accessible and available for sampling. If the two diverge significantly, the target population should be redefined.

Probabilistic sampling

Probabilistic sampling occurs when each member of the population has a given probability (greater than zero and less than one) of being included in the sample. If the probability is the same for all population members the sample will be unbiased. Because inclusion in the sample is based on probability, subsequent samples will not necessarily include the same members.

Range

The range of a dataset measures the spread between the highest and lowest values in it. Other measures such as the standard deviation and the interquartile range are required to provide an understanding of the data's distribution.

Residual soil

The soil at a site that is not contaminated by industrial, commercial, or agricultural activities, consistent with the term 'ambient background concentration' (ABC) from the NEPM. Residual soils can include natural soils, reworked natural soils and historically imported material. Residual soils may have naturally occurring background levels of contaminants, contaminants that have been introduced from diffuse or non-point sources by general anthropogenic activity, and only low levels of contaminants attributed to industrial, commercial, or agricultural activities.

Sample

'Sample' has several meanings including:

- as more broadly used in statistics, a representative group drawn from a population for description or measurement
- a physical amount of a material such as soil, water or air or an aliquot, taken for testing or chemical analysis
- a sampling point or sample location, being the location in plan at which a sample is collected, including description, for example, geological logs and field screening, for example, PID or XRF.

Sample size

The number of samples or sampling points in a sampling program.

Sampling, analysis and quality plan (SAQP)

Incorporates the CSM and the DQO outputs, to provide the context of and justification for the selected sampling and analysis. The methods, procedures and QC samples associated with the DQIs, including the frequency and MQOs and any associated contingencies, are also documented. The SAQP ensures that the data collected is representative and provides a robust basis for site assessment (NEPC 2013).

Sampling pattern

The locational pattern of sampling points in a sampling area.

Sampling point

The location at which a sample is collected.

Site characterisation

The assessment of the nature, level and extent of contamination. A typical site characterisation involves a preliminary site investigation (PSI), followed by a detailed site investigation (DSI), where warranted.

Site validation

The process of showing that a site is successfully remediated.

Standard deviation

Calculated by taking the square root of the variance (described below). It provides an indication of a population or sample data's typical deviation from its mean.

Statistic

Any summary number that describes the sample, such as an average or percentage. For example, the mean of a sample is described as $\bar{x}$ (x-bar) and the standard deviation as **s**. When describing the population from which the sample is drawn, a summary number is called a **parameter**.

Statistical power

The probability of correctly determining a positive result based on sample data, for example, a change or difference in the population.

Sub-sample

A sample that will be combined with other sub-samples to form a composite for chemical analyses.

Systematic planning

A planning process based on a scientific method which helps the project to unfold logically. Systematic planning includes established management and scientific elements. In the assessment of site contamination, it includes the application of the **DQOs** process and development of a **CSM** and **SAQP**.

Variable

A characteristic, number or quantity that is the subject of the inquiry. In the assessment of site contamination, it is usually continuous numerical variables that are being assessed, for example the concentration of a contaminant in soil, soil gas or water. Discrete or discontinuous variables are at times considered, such as the number of fish in a waterbody. These are both quantitative variables in that they are derived by measurements.

Qualitative or categorical variables include ordinal or ranked variables and nominal variables. Ordinal variables are observations that take a value that can logically be ordered or ranked, such as first, second, third, whereas nominal observations take a value that cannot be organised in a logical sequence, such as presence or absence. Categorical variables are not commonly used in the assessment of site contamination.

Variance

The average squared distance of population or sample data points from the associated mean.

Weight of evidence/lines of evidence

‘Weight of evidence’ describes the process of collecting, analysing and evaluating a combination of different qualitative, semi-quantitative or quantitative lines of evidence to make an overall assessment of contamination.

Applying a weight-of-evidence process incorporates judgements about the quality, quantity, relevance and congruence of the data contained in the different lines of evidence (ANZG 2018).

9. References

Australian and New Zealand Environment and Conservation Council (ANZECC) and Agriculture and Resource Management Council of Australia and New Zealand (ARMCANZ) 2000, *Australian and New Zealand guidelines for fresh and marine water quality*, paper no. 4, ANZECC and ARMCANZ, Canberra.

Australian and New Zealand Governments and Australian State and Territory Governments (ANZG) 2018, *Australian and New Zealand guidelines for fresh and marine water quality*, Water Quality Australia, Canberra ACT, www.waterquality.gov.au/anz-guidelines.

British Standards Institution (BSI) 2013, *Investigation of potentially contaminated sites: code of practice*, BS 10175:2011+A1:2013, BSI Standards Limited.

Clements L, Palaia T & Davis J 2009, *Characterisation of sites impacted by petroleum hydrocarbons: national guideline document*, CRC Care Technical Report no. 11, CRC for Contamination Assessment and Remediation of the Environment, Adelaide.

Contaminated Land: Applications in Real Environments (CL:AIRE) 2008, *Guidance on Comparing Soil Contamination Data with a Critical Concentration*, The Chartered Institute of Environmental Health and CL:AIRE, London.

CRC Care 2013*, Petroleum hydrocarbon vapour intrusion assessment: Australian guidance*, CRC CARE Technical Report no. 23, CRC for Contamination Assessment and Remediation of the Environment, Adelaide.

Crumbling DM 2002, In search of representativeness: evolving the environmental data quality model, *Quality Assurance*, vol.9, pp179–90, http://cluin.org/download/char/dataquality/dcrumbling.pdf.

Davis GB, Wright J & Patterson BM 2009, *Field Assessment of vapours,* CRC CARE Technical Report no. 13, CRC for Contamination Assessment and Remediation of the Environment, Adelaide.

Department of Agriculture and Water Resources 2018*, Australian & New Zealand guidelines for fresh and marine water quality*, Department of Agriculture and Water Resources, Canberra, www.waterquality.gov.au/anz-guidelines.

Department of Environment and Conservation (DEC) 2004, New South Wales (NSW) *Australian river assessment system (AUSRIVAS) sampling and processing manual*, DEC NSW, Sydney.

Department of Environment and Conservation (DEC) 2005a, *Contaminated sites: Guidelines for assessing former orchards and market gardens*, DEC 2005/195, DECCW NSW, Sydney.

Department of Environment and Conservation (DEC) 2005b, *Information for the assessment of former gasworks sites*, DEC 2005/237, DECCW NSW, Sydney.

Department of Environment and Conservation (DEC) 2007, *Contaminated sites: Guidelines for the assessment and management of groundwater contamination*, DEC 2007/144, DEC NSW, Sydney.

Department of Environment, Climate Change and Water (DECCW) 2009, *Guidelines for implementing the Protection of the Environment Operations (Underground Petroleum Storage Systems) Regulation 2008*, DECCW 2009/653, DECCW NSW, Sydney.

Department of Environment, Climate Change and Water (DECCW) 2010, *Vapour intrusion: technical practice note*, DECCW 2010/774, DECCW NSW, Sydney.

Department of Environment (DoE) Queensland 1998, *Draft guideline for the assessment & management of contaminated land in Queensland*, DoE, Brisbane.

Department of Environment and Science (DES) Queensland 2018, *Monitoring and sampling manual: environmental protection (water) policy 2009* [sic], DES, Brisbane.

Department of Health and Ageing and EnHealth Council 2012, *Environmental health risk assessment: guidelines for assessing human health risks from environmental hazards*, Department of Health and Ageing, Canberra.

Devore J & Farnum N 2005, *Applied statistics for engineers and scientists*, 2nd Edition, Brooks/Cole, Cengage Learning, Belmont CA.

Environment Protection Authority (EPA) 1995, *Contaminated sites: Guidelines for the vertical mixing of soil on former broad-acre agricultural land*, EPA 2003/28, NSW EPA, Sydney.

Environment Protection Authority (EPA) 1997, *Contaminated sites: guidelines for assessing banana plantation sites*, EPA 97/37, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2014a, *Resource Recovery Order under Part 9, Clause 93 of the Protection of the Environment Operations (Waste) Regulation 2014: the excavated natural material order 2014.*

Environment Protection Authority (EPA) 2014b, *Best practice note: landfarming*, EPA 2014/0323, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2015a, *Guidelines on the duty to report contamination under the Contaminated Land Management Act 1997*, EPA 2015/0164, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2015b, *Technical note: light non-aqueous phase liquid assessment and remediation*, EPA 2015/0553, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2017, *Contaminated land management: guidelines for the NSW site auditor scheme (3rd edition)*, EPA 2017P0269, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2018, *Guidelines on resource recovery orders and exemptions: for the land application of waste materials as fill*, EPA 2017/P0392, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2020a, *Assessment and management of hazardous ground gases: contaminated land guidelines*, EPA 2019P2047, NSW EPA, Sydney.

Environment Protection Authority (EPA) 2020b, *Consultants reporting on contaminated land: contaminated land guidelines*, EPA 2020P2233, NSW EPA, Parramatta.

Environment Protection Authority (EPA) 2020c, *Guidelines for implementing the Protection of the Environment Operations (Underground Petroleum Storage Systems) Regulation 2019*, EPA 2020/P2700, NSW EPA, Parramatta.

Environment Protection Authority (EPA) South Australia 2005, *Composite soil sampling in site contamination assessment and management*, Government of South Australia, Adelaide.

Environment Protection Authority (EPA) Victoria 2009, *Industrial waste resource guidelines: soil sampling*, IWRG702.

Ferguson CC 1992, The statistical basis for spatial sampling of contaminated land*, Ground Engineering*, vol. 25, no. 6, pp 34–38.

Gilbert RO 1987, *Statistical methods for environmental pollution monitoring*, John Wiley & Sons Inc., Brisbane.

Gray JM & Murphy BW 1999, *Parent material and soils: A guide to the influence of parent material on soil distribution in eastern Australia*, Technical Report No. 45, NSW Department of Land and Water Conservation, Sydney.

Hamon RE, Mclaughlin MJ, Gilkes RJ, Rate AW, Zarcinas B, Robertson A, Cozens G, Radford N & Bettenay L 2004, Geochemical indices allow estimation of heavy metal background concentrations in soils, *Global Biogeochemical Cycles*, vol. 18, GB1014.

Harr ME 1987, *Reliability-based design in engineering*, McGraw-Hill, New York.

HEPA 2018, *PFAS national environmental management plan version 2.0*, Heads of EPAs Australia and New Zealand.

Interstate Technology and Regulatory Council (ITRC) 2007, *Vapor intrusion pathway: a practical guideline*, VI-1, ITRC Vapor Intrusion Team, Washington DC, USA, www.itrcweb.org/Documents/VI-1.pdf.

Interstate Technology and Regulatory Council (ITRC) 2012, *Incremental sampling methodology*, (ISM-1), ITRC, Washington DC, USA, https://itrcweb.org/teams/projects/incremental-sampling-methodology.

Interstate Technology and Regulatory Council (ITRC) 2013, *Groundwater statistics and monitoring compliance: statistical tools for the project life cycle*, GSMC-1, ITRC, Washington DC, USA, http://www.itrcweb.org/gsmc-1/.

Interstate Technology and Regulatory Council (ITRC) 2017, *Naming conventions and physical and chemical properties per- and polyfluoroalkyl substances (PFAS)*, ITRC, Washington DC.

Lock WH 1996, Composite sampling, in *National Environmental Health Forum Monographs, Soil Series No. 3*, South Australian Health Commission, Adelaide.

McDougall KW & Macoun TW 1996, *Guidelines for the assessment and clean up of cattle tick dip sites for residential purposes*, NSW Agricultural in conjunction with CMPS&F Environmental, Wollongbar NSW.

Naidu R, Jit J, Kennedy B & Arias V 2016, Emerging contaminant uncertainties and policy: The chicken or the egg conundrum, *Chemosphere*, vol. 154, pp 385–390.

National Environment Protection Council (NEPC) 2013, *National environment protection (assessment of site contamination) amendment measure 2013 (no. 1)*, Schedule A and Schedules B(1)–B(9), National Environment Protection Council, Canberra.

Nickerson RS 2000, Null hypothesis significance testing: A review of an old and continuing controversy, *Psychological Methods*, vol. 5, no. 2, pp 241–301.

New Jersey Department of Environmental Protection (NJDEP) 2005, *Vapor intrusion guidance*.

Northern Territory Environment Protection Authority (EPA) 2013, *Guidelines on conceptual site models*, NT EPA, Darwin.

Office of Environment and Heritage 2011, *Guidelines for consultants reporting on contaminated sites*, OEH 2011/0650, NSW OEH, Sydney.

Parkhurst DF 1998, Arithmetic versus geometric means for environmental concentration data, *Environmental Science & Technology*, vol. 32 (3), pp 92A–98A.

Perezgonzalez JD 2015, Fisher, Neyman–Pearson or NHST? A tutorial for teaching data testing, *Frontiers in Psychology, vol. 6*, article 223, p. 1.

Provost LP 1984, Statistical methods in environmental sampling, in Schweitzer GE and Santolucito JA (eds), *Environmental sampling for hazardous wastes*, American Chemical Society, Washington DC.

Reinhart A 2015, *Statistics done wrong: The woefully complete guide*, No Starch Press, San Francisco CA.

Simpson S and Batley G (eds) 2016, *Sediment quality assessment: a practical guide*, CSIRO Publishing, Melbourne.

South Australian Health Commission (SAHC) 1995, *Guidelines for the composite sampling of soils*, SAHC, Adelaide.

US Environmental Protection Agency (USEPA) 1996, *Soil screening guidance: user's guide* (2nd edition), Attachment B, Soil Screening DQOs for Surface Soils and Subsurface Soils, EPA/540/R-96/018, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2000, *Data quality objectives process for hazardous waste site investigations (QA/G-4HW)*, EPA/600/R-00/007, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2001, *Guidance on data quality indicators (QA/G-5i)*, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2002a, *Guidance on environmental data verification and data validation (QA/G-8)*, EPA/240/R-02/004, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2002b, *Guidance on choosing a sampling design for environmental data collection for use in developing a quality assurance project plan (QA/G-5S)*, EPA/240/R-02/005, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2002c, *Guidance for quality assurance project plans (QA/G-5)*, EPA/240/R-02/009, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2002d, *Calculating upper confidence limits for exposure point concentrations at hazardous waste sites*, OSWER 9285.6-10, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2006a, *Data quality assessment: statistical methods for practitioners (QA/G-9S)*, EPA/240/B-06/003, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2006b, *Guidance on systematic planning using the data quality objectives process (QA/G-4)*, EPA/240/B-06/001, Appendix: Derivation of sample size formula for testing mean of normal distribution versus an action level, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2006c, *Data quality assessment: a reviewer's guide (QA/G-9R)*, EPA/240/B-06/002, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2007, *Guidance for preparing standard operating procedures (SOPs) (QA/G-6),* EPA/600/B-07/001, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2009, *Statistical analysis of groundwater monitoring data at RCRA facilities: Unified Guidance*, EPA 530/R-09-007, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2014, *Test methods for evaluating solid waste: physical/chemical methods compendium (SW-846)*, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2015a, *ProUCL version 5.1.002: technical guide: statistical software for environmental applications for datasets with and without nondetect observations*, EPA/600/R-07/041, USEPA, Washington DC.

US Environmental Protection Agency (USEPA) 2015b, *ProUCL version 5.1.002: User guide: Statistical Software for environmental applications for datasets with and without nondetect observations, EPA/600/R-07/041*, USEPA, Washington DC.

Viveros R 1997, Inference about the mean in log-regression with environmental applications, *Environmetrics*, vol. 8(5), pp. 569–582.

Wendelberger J & Campbell K 1994, *Non-detect data in environmental investigations*, American Statistical Association, Toronto, Canada.

Wilson S, Card G & Haines S 2009, *Ground gas handbook*, Whittles Publishing, Dunbeath, UK.

Appendix A: Descriptive statistics

This Appendix briefly reviews the descriptive statistics commonly used for summarising data.

Appendices B to D show how they are used, giving specific procedures and worked examples.

Range and percentiles

The range of a dataset measures the spread between the highest and lowest observed values in the dataset. It can be expressed as an interval, such as a–b, where a is the lowest value and b is the highest, or it can be expressed as an interval width, such as b - a = c. While either approach covers the range of the observed values, the maximum value is of particular concern in the assessment of site contamination, and the range is generally more informative as an interval, as it shows the spread and the extremes of the data.

As the range only measures the spread between highest and lowest values, other measures such as the **standard deviation** or the **interquartile range** (IQR) are needed to more fully describe data distribution.

Maximum

The maximum observed value in a dataset is important when assessing site contamination as it generally conservatively estimates the potential exposure risks. It is usually assumed that if the maximum is below the action level, the site will be suitable for the associated land use. However, this assumption is true only if there is enough data and the data is representative. If this is not the case, the maximum observed value may overestimate or underestimate the risk.

Where the consequences of decision error will be severe, or there are not enough samples to estimate the population mean from the sample mean, the maximum value can be used as an estimate of the population mean and termed **max test** for statistical analysis. This is often done for judgmental samples, such as with soil gas or groundwater data. Where this approach is used, it should be appropriately documented and justified.

Percentiles and quartiles

Percentiles, as suggested by the name, are descriptive values used to equally split a dataset into 100 parts. The Xth percentile in a dataset has a value greater than or equal to X% of the data – for example, the 80th percentile has a value greater than or equal to 80% of the data.

Percentiles can be used as the statistical parameter of interest, for instance, for comparing to criteria or action levels. For example, ANZG 2018 states that '[f]or toxicants, it is recommended that action is triggered if the 95th percentile of the test data exceeds the guideline value'.

Quartiles are used to break up the dataset into four equal parts, providing an indication of the distribution and variance of the data. When observations are placed in ascending order by value:

- the first quartile, Q1, also called the lower quartile, is the value of the observation at or below which a quarter (25%) of observations lie, and is the 25th percentile
- the second quartile, Q2, is the median value at or below which half (50%) of observations lie, and is the 50th percentile
- the third quartile, Q3, also called the upper quartile, is the value of the observation at or below which three-quarters (75%) of the observations lie and is the 75th percentile.

The interquartile range (IQR) is used as a measure of the spread of the dataset which also indicates its dispersion. It is the difference between the upper and lower quartiles (Q3 - Q1 = IQR), that is, it measures the spread between the 25th and 75th percentiles. The IQR spans 50% of a dataset and eliminates the influence of outliers as it excludes the highest and lowest quarters.

Percentiles and quartiles can be used for datasets with limited observations, and for all types of data collection, as their use requires no assumptions about the underlying distribution or whether the samples were judgmental or probabilistic. However, ANZG 2018 notes that the precision with which percentiles are estimated depends heavily on the sample size, with at least 13 samples need to estimate the 25th

and 75th percentiles with an associated 95% confidence interval, and a minimum of 36 samples needed to estimate the 10th and 90th percentiles. Even larger sample sizes are required to estimate extreme percentiles, i.e. the 5th and 95th.

As the IQR does not depend on extreme values, it can be used when a dataset includes non-detects, at least where < 25% of the data is below the limits of reporting (LORs). For datasets that are not nearly-normal, or which contain extreme values, the IQR may be more representative of the dispersion of the data than the standard deviation. The IQR is therefore described as a robust estimate.

Appendix B explains how to determine quartiles.

Central tendency

Central tendency is the central or typical value for a probability distribution, and may be considered the average value in a dataset. It is generally described by the mode, median, or, most commonly, the mean, and indicates where a sample distribution is centred. While these estimates can generally be regarded as being representative or typical of the data, for small or highly skewed datasets they should be considered as approximations only.

Appendix C explains how to determine measures of central tendency.

Mode

The mode is the value that occurs most frequently and is determined by counting the number of times each value occurs. Since a sample mode may not exist or may not be unique, for example, the distribution may be bimodal, it is rarely used as a measure of central tendency, although it can be useful for qualitative data such as categories.

Median

The median is the middle value of the distribution: half the data points have values greater than the median, and half have values less than it. The sample median is not influenced by extreme values so can be used when the underlying distribution is unknown: it is commonly used to describe the centre of the distribution when non-parametric methods are employed. The median can also be used if non-detects are present, although care should be taken if there are many of them. If a median is found to be a non-detect while there are locations reporting values above detection levels, stratifying the site should be considered.

A number of guidelines recommend the use of median values in certain circumstances. For example:

- NEPC 2013, B2 states that when using non-parametric approaches, the median can be used to describe the centre of the distribution
- ANZG 2018 notes that for comparing test data with guideline values for physico-chemical stressors, '[a] trigger for further investigation of the test water body will be deemed to have occurred when the median concentration of a particular measurement parameter in **n** independent samples taken at the test water body exceeds the 80th percentile (or is below the 20th percentile if "less is worse") of the same measurement parameter at the reference site'.

Arithmetic mean

The arithmetic mean is commonly referred to as the average and is used to describe the centre of the data distribution. The arithmetic mean is denoted as **μ** (lowercase Greek letter *mu*) for the population mean or as $\bar{\mathbf{x}}$ (x-bar) for the sample mean. In the assessment of site contamination, the population mean is generally not known, so the sample mean is used as an estimate of the population mean. The arithmetic mean is calculated by dividing the sum of the sample measurements by the number of samples.

Larger sample sizes tend to produce sample means that are closer to the population mean, as in theory extreme data values balance each other out. But when sample sizes are small, the arithmetic mean can be affected by outliers, and when judgmental sampling is used, the arithmetic mean is often a biased measure of central tendency.

The mean value may be more representative of site contamination than the maximum value by providing a better estimate of the contaminant concentrations that receptors would be exposed to over a period of time. However, it is important that small areas of high concentration (hotspots) are not ignored by averaging with lower values from other parts of the site or the decision area.

Geometric mean

The geometric mean is similar to the arithmetic mean in that it is also a measure of the central tendency of the distribution of a population or sample. This is also described as the arithmetic mean of the logarithmic scale of a dataset, or the nth root of the product of n numbers.

Due to the log transformation involved in the calculation, the geometric mean is not as affected by outliers as the arithmetic mean, and is commonly used when the data is skewed or log-normally distributed. However, the curvature of the logarithmic function may downplay the higher values in favour of the lower ones.

Higher values are important in the assessment of site contamination. If assumptions regarding the condition of a site are based on the geometric mean, downplaying higher values may increase the chance of a Type I error. Because of this potential bias, geometric means, including back transformation, should not be used in isolation to compare against action levels, and if they are used appropriate justification should be provided. Where log-transformed data are approximately normal or at least reasonably symmetric, back transformation may be appropriate (USEPA 2009 and Viveros 1997), but for skewed datasets that are not log-normal, the geometric mean is likely to be a poor estimator of population mean (Parkhurst 1998).

Variability

An important aspect of data analysis is determining the variability of the sampling data. Calculating variability can provide an indication of how heterogenous the variables are likely to be across a decision area, and how representative of the sampling data the measures of central tendency are. The variability of data is measured by **variance**, **standard deviation** and the **coefficient of variation**.

See Appendix D for how to determine measures of variability.

Variance

Variance is the average squared distance of each data point from the sample mean. It can be affected by extreme values and by large numbers of values below the LORs.

Standard deviation

The standard deviation is calculated by taking the square root of the variance and provides an indication of the data's typical deviation from the mean. The standard deviation of a population is denoted as $\boldsymbol{\sigma}$ (Greek lowercase sigma), and as **s** for a sample. The sample standard deviation is commonly used in site contamination assessment, as the standard deviation of the population is generally not known.

A large sample variance or standard deviation indicates that the data points are not closely clustered around the mean. Both the variance and the standard deviation are strongly influenced by the number of samples collected, and influenced by extreme values in either direction.

Coefficient of variation

The coefficient of variation (CV) or relative standard deviation (RSD) measures the relative homogeneity of a distribution. The CV is determined as the standard deviation of a distribution divided by the mean of the distribution, that is, $CV = s/\bar{x}$ for sample data. The RSD is determined in the same way, but expressed as a percentage, that is, a CV of 0.5 = an RSD of 50%.

Low CV values, for example, of 0.5 or less, indicate a fairly homogeneous contaminant distribution, while CVs with values of more than 1–1.2 imply that the concentration distribution of a contaminant is heterogeneous.

Appendix B: Determining quartiles

Percentiles are descriptive values that split a dataset into 100 equal parts, providing a representation of the sampling data that can be used for either normal or non-normal distributions. A percentile provides the value that a given percentage of observations in a dataset is less than or equal to (for example, 25% of observations in the dataset have values at or below the value of the 25th percentile).

Percentiles can be used in the statistical analysis of datasets that have limited observations. The dataset can also be divided by **quartiles**, which are the 25th, 50th and 75th percentiles.

Determination

To calculate percentiles, values are ordered from the lowest to the highest and assigned a rank, with the required percentile calculated using the formula shown below. While this procedure can be used for small datasets, it is commonly conducted using spreadsheets or statistical packages. Note that all percentiles of sample data are biased estimators of population percentiles.

The values are ranked from lowest to highest:

$$X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)} \ldots, X_{(n)}$$

The p^{th} percentile is calculated by:

$$y_p = (1 - f) \times X_i + f \times X_{(i+1)}$$

Where:

y_p	the value of the p^{th} percentile
p^{th}	the specified percentile
r	$(n - 1)p + 1$
floor(r)	calculate r and discard decimals
i	floor(r)
f	r – i
X_i	the value of the ith rank
$X_{(i + 1)}$	the value of the ith + 1 rank

The data in Table 1is used for the worked examples in this and the following appendices.

Table 1 Summary of analytical results – metals in soil (mg/kg)

Sample ID or statistic	Arsenic	Chromium	Copper	Lead	Nickel	Zinc
Limits of reporting	5	2	5	5	2	5
Analytical						
Analytical sample B2-01	103	12	34	20	18	11
Analytical sample B2-02	50	21	30	7	2	10
Analytical sample D2-01	43	26	83	17	14	35
Analytical sample D2-02	9	10	29	14	5	12
Analytical sample A4-01	203	4	260	18	12	232
Analytical sample A4-02	54	5	55	17	9	41
Analytical sample C4-01	341	19	401	133	7	543
Analytical sample C4-02	34	17	46	16	10	13
Analytical sample B6-01	71	18	24	14	5	9
Analytical sample B6-02	14	6	8	17	12	5
Analytical sample D6-01	62	11	51	15	3	36
Analytical sample D6-02	6	4	18	16	24	10
Analytical sample A8-01	27	17	61	16	4	24
Analytical sample A8-02	7	10	38	20	13	10
Analytical sample C8-01	24	15	39	12	6	8
Analytical sample C8-02	13	16	17	14	19	7
Descriptive statistics						
Number of samples	16	16	16	16	16	16
Number of detects	16	16	16	16	16	16
Percentage non detects	0%	0%	0%	0%	0%	0%
Maximum	341	26	401	133	24	543
Third quartile	64.3	17.3	56.5	17.3	13.3	35.3
Median value	38.5	13.5	38.5	16.0	9.4	11.5
First quartile	13.8	9.0	27.8	14.0	5.2	9.8
Minimum	6	4	8	7	2	5
Arithmetic average	66.3	13.2	74.6	22.9	10.2	62.9
Geometric average	35.2	11.4	43.5	17.3	8.3	20.0
Mode	-	10	-	17	12	10
Variance	7,792.2	42.4	10,988.8	872.1	39.7	19,410.1
Standard deviation	88.3	6.5	104.8	29.5	6.3	139.3
Coefficient of variation (CV)	1.3	0.5	1.4	1.3	0.6	2.2
Inferential statistics						
Standard error of the mean ($SE\bar{x}$)	22.1	1.6	26.2	7.4	1.6	34.8
Relative standard deviation (RSD)	133.1%	49.4%	140.5%	129.1%	61.9%	221.6%
Margin of error (MoE)	47.0	3.5	55.9	15.7	3.4	74.2

Sample ID or statistic	Arsenic	Chromium	Copper	Lead	Nickel	Zinc
Maximum probability error (MPE)	70.9%	26.3%	74.9%	68.8%	33.0%	118.1%
95% UCL$\bar{x}$ two-sided Student's t	113.4	16.7	130.5	38.6	13.5	137.1
95% UCL$\bar{x}$ one-sided Student's t	105.0	16.0	120.5	35.8	13.0	123.9
ProUCL determination	120.5	16.0	135.2	55.1	13.0	214.7
Method recommended*	Gamma	Student's t	H-UCL	Chebyshev	Student's t	Chebyshev
Criteria and number of samples						
HIL-A land use (NEPC 2013, B1)	100	100	6,000	300	400	7,400
Number of samples to be used (whole number) – CRV method	44	2	2	2	2	2
Number of samples – MPE method	15	18	16	16	14	15

Worked example

The metals data in mg/kg from Table 1 is used in this example.

To determine the 25th percentile of the sampling data for arsenic (As):

The values are ordered from lowest to highest and assigned a rank:

$$X_{(1)} = 6,\ X_{(2)} = 7,\ X_{(3)} = 9,\ \mathbf{X_{(4)}} = \mathbf{13},\ \mathbf{X_{(5)}} = \mathbf{14},\ X_{(6)} = 24,\ X_{(7)} = 27,\ X_{(8)} = 34,$$

$$X_{(9)} = 43,\ X_{(10)} = 50,\ X_{(11)} = 54,\ X_{(12)} = 62,\ X_{(13)} = 71,\ X_{(14)} = 103,\ X_{(15)} = 203,\ X_{(16)} = 341$$

Bolded values are $X_{(i)}$ and $X_{(i+1)}$.

The input parameters are calculated for the 25th percentile:

$$r = (n - 1)p + 1$$

$$r = (16 - 1)\ 0.25 + 1$$

$$r = 4.75$$

$$i = 4$$

$$f = r - i$$

$$f = 0.75$$

The 25th percentile is calculated as:

$$y_{(0.25)} = (1 - f) \times X_i + f \times X_{(i+1)}$$

$$y_{(0.25)} = (1 - 0.75) \times 13 + 0.75 \times 14$$

$$y_{(0.25)} = 13.8$$

The 25th percentile of the sampling data for As is 13.8 mg/kg.

Appendix C: Determining measures of central tendency

The **central tendency** is a central or typical value for a probability distribution, and may be considered the average value in a set of data. Methods for calculating the **median**, the **arithmetic mean** and the **geometric mean** are shown below.

The **mode** is the value that occurs with the greatest frequency (that is, the greatest number of times): to calculate it, simply count the number of times each value occurs. As the mode does not always exist or may not be unique, it is the value of central tendency that is least commonly used, although it can be useful for describing qualitative data.

Determination

Measures of central tendency are determined as follows.

Median with an odd number of samples

$$\text{median} = X_{(n+1)/2}$$

Median with an even number of samples

$$\text{median} = \frac{1}{2}\left[X_{(n/2)} + X_{(n/2+1)}\right]$$

Arithmetic mean

$$\text{sample arithmetic mean} = \frac{(X_1 + X_2 + \ldots X_n)}{n}$$

Geometric mean

$$\text{sample geometric mean} = \sqrt[n]{(X_1 \times X_2 \times \ldots X_n)}$$

Worked example

The metals data in mg/kg from Table 1 is used in this example to determine the measures of central tendency for the sampling data for arsenic (As):

Median

The values are ordered from lowest to highest and assigned a rank:

$$X_{(1)} = 6,\ X_{(2)} = 7,\ X_{(3)} = 9,\ X_{(4)} = 13,\ X_{(5)} = 14,\ X_{(6)} = 24,\ X_{(7)} = 27,\ \mathbf{X_{(8)}} = \mathbf{34},\ \mathbf{X_{(9)}} = \mathbf{43},$$

$$X_{(10)} = 50,\ X_{(11)} = 54,\ X_{(12)} = 62,\ X_{(13)} = 71,\ X_{(14)} = 103,\ X_{(15)} = 203,\ X_{(16)} = 341$$

Bolded values are $X_{(n/2)}$ and $X_{(n/2\ +\ 1)}$.

As n = 16, an even number, the sample median is determined as:

$$\text{sample median} = \frac{1}{2}\left[X_{\left(n/2\right)} + X_{\left(n/2+1\right)}\right]$$

$$\text{sample median} = \frac{1}{2}\left[X_{\left(16/2\right)} + X_{\left(16/2+1\right)}\right]$$

$$\text{sample median} = \frac{1}{2}\left[X_{(8)} + X_{(9)}\right]$$

$$\text{sample median} = \frac{1}{2}\left[34 + 43\right]$$

$$\text{sample median} = 38.5$$

The sample median for As is 38.5 mg/kg.

Arithmetic mean

$$\text{sample arithmetic mean} = \frac{(X_1 + X_2 + \ldots X_n)}{n}$$

$$\text{sample arithmetic mean} = \frac{(103 + 50 + \cdots 13)}{16}$$

$$\text{sample arithmetic mean} = 66.3$$

The sample arithmetic mean for As is 66.3 mg/kg.

Geometric mean

$$\text{sample geometric mean} = \sqrt[n]{(X_1 \times X_2 \times \ldots X_n)}$$

$$\text{sample geometric mean} = \sqrt[16]{(103 \times 50 \times \ldots 13)}$$

$$\text{sample geometric mean} = 35.2$$

The sample geometric mean for As is 35.2 mg/kg.

As Table 2 shows, each method provides a different result for the measure of central tendency.

Table 2 Variation in central tendency by method of calculation

Method	Result (mg/kg)
Median	38.5
Arithmetic mean	66.3
Geometric mean	35.2

For sample data that is skewed, as in this case, the median and geometric mean are similar, while the arithmetic mean is ‘dragged’ to the right because of the outliers in the dataset. For a nearly-normal dataset, the three measures would be similar.

The appropriate measure of central tendency should be chosen to represent the sampling data according to the contaminant distribution and the proposed use of the selected measure.

Appendix D: Determining measures of variability

An important aspect of data analysis is determining the variability of the data. Calculating variability can indicate how heterogenous a contaminant is likely to be across a site. The variability of the data is measured by variance, standard deviation or the coefficient of variation.

Variance, represented by $\mathbf{s^2}$, is the average squared distance of each data point from the sample mean, and can be affected by extreme values and large numbers of values below the limits of reporting (LORs). It is used to estimate the population variance $\boldsymbol{\sigma^2}$.

The **standard deviation** of a sample, represented by **s**, is calculated by taking the square root of the variance, and indicates the population's typical deviation from the mean. The standard deviation of the population, represented by **σ**, is generally unknown when assessing site contamination, so **s** is used as an estimate. Note that although s^2 is an unbiased estimate of σ^2, s is a **biased** estimate of σ.

The **coefficient of variation** (CV) or **relative standard deviation** (RSD) measures the relative homogeneity of a distribution. The CV is the standard deviation of a distribution divided by the mean of the distribution. The RSD is determined in the same way but expressed as a percentage.

Determination

The methods for determining the measures of variability are shown below.

Variance

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Standard deviation of a sample

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

Estimate of standard deviation

Where sampling data are not available, an estimate of the standard deviation can be made by dividing the expected range by six, that is, three standard deviations in each direction, as this should represent approximately 99.7% of a nearly-normal distribution.

$$\sigma_E = \frac{C_H - C_L}{6}$$

The relative standard deviation is determined in the same way, but expressed as a percentage, that is, a CV of 0.5 = an RSD of 50%.

Coefficient of variation

$$CV = \frac{s}{\bar{x}}$$

Where:

s^2	variance
x_i	the value of the sample
$\bar{x}$	the arithmetic mean (see Appendix C)
n	number of samples
s	standard deviation

σ_E estimate of population standard deviation

C_H estimate of the highest possible value in the sampling area

C_L estimate of the lowest possible value in the sampling area

CV coefficient of variation

RSD relative standard deviation

Worked example

In this example we determine the measures of variability for the sampling data for arsenic (As) in Table 1.

The values for As, shown in mg/kg, are:103, 50, 43, 9, 203, 54, 341, 34, 71, 14, 62, 6, 27, 7, 24 and 13.

The number of samples, n, is 16, and the arithmetic average of the sampling data is 66.3.

Variance

$$s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n - 1}$$

$$s^2 = \frac{(103 - 66.3)^2 + (50 - 66.3)^2 \ldots (13 - 66.3)^2}{16 - 1}$$

$$s^2 = \frac{1{,}346.9 + 265.7 + \cdots 2{,}840.9}{15}$$

$$s^2 = 7{,}792.2$$

Standard deviation

$$s = \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n\text{-}1}}$$

$$s = \sqrt{7{,}792.2}$$

$$s = 88.3$$

Estimate of standard deviation

$$\sigma_E = \frac{C_H - C_L}{6}$$

$$\sigma_E = \frac{341 - 6}{6}$$

$$\sigma_E = 55.8$$

In this example, the standard deviation calculated using the sampling data is much greater than the estimate of the standard deviation. This is because the sampling data is skewed to the right and does not follow a nearly-normal distribution.

This example shows that, while estimates of standard deviation can be determined when sampling data are not available, they should always be used with caution. If required sample numbers were calculated using an estimated value such as the one in this example, the result would be too low. Accordingly, the sampling data should be used to refine the assumptions made as part of systematic planning.

Coefficient of variation (CV)

$$CV = \frac{s}{\bar{x}}$$

$$CV = \frac{88.3}{66.3}$$

$$CV = 1.3$$

Relative standard deviation (RSD)

$$RSD = 133.1\%$$

In this example, the CV of 1.3 (equivalent to an RSD of 133.1%) shows a distribution not nearly-normal and expected to be skewed to the right. Any statistical inference should assume a log-normal or other non-normal distribution, and use log-normal or non-parametric methods for analysis.

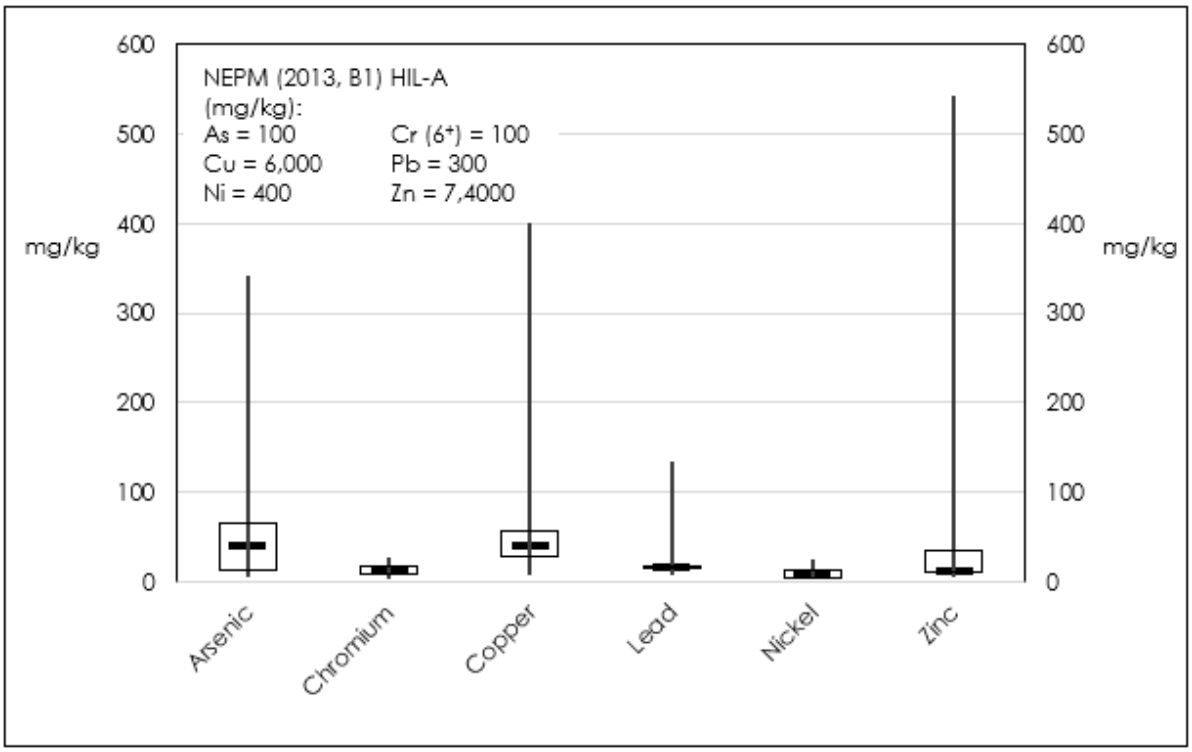
Appendix E: Assessing contaminant distribution

This appendix explains how contaminant distribution can be assessed using commonly available spreadsheet and statistical software, as discussed in Section 2.7.. The sampling data in this example is sourced from Table 1.

Table 3 Graphical presentations of example contamination data

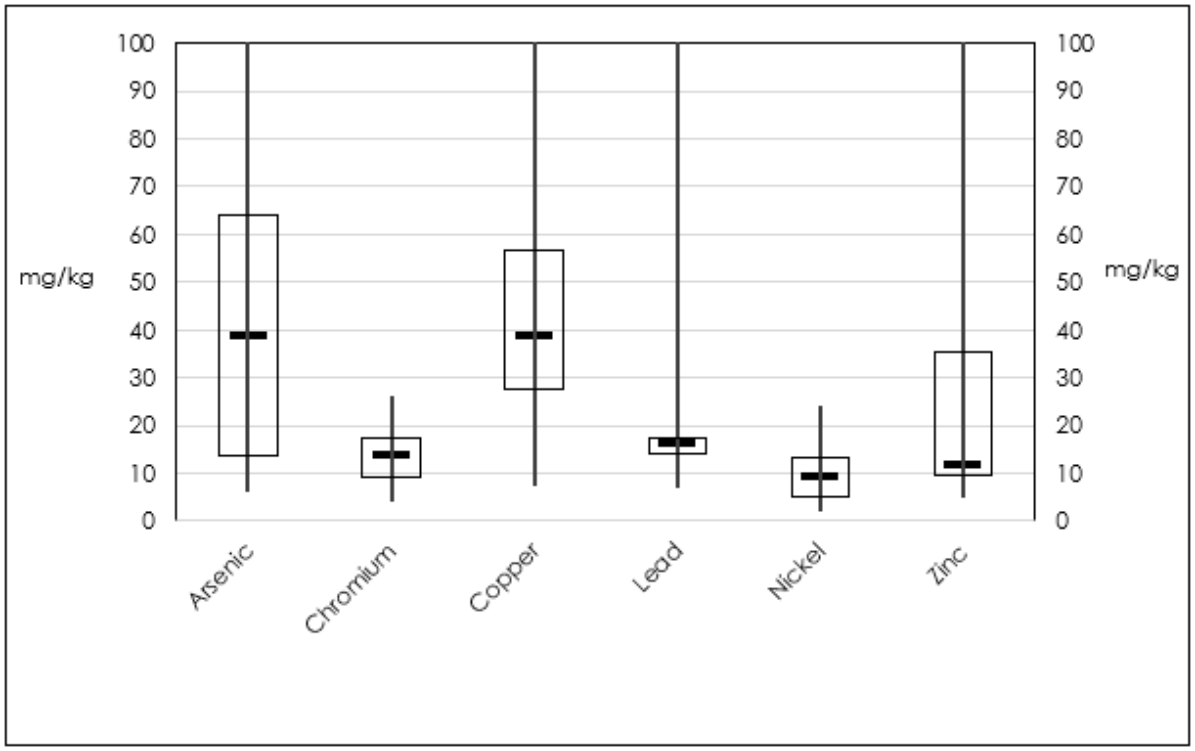
Figure	Description
Figure 1	Summary statistics: metals in fill (mg/kg) as box-and-whiskers plots showing minimum, first quartile, median, third quartile and maximum
Figure 2	Summary statistics: metals in fill (mg/kg) as box-and-whiskers plots showing minimum, first quartile, median, third quartile and maximum, with adjusted scale
Figure 3	Standardised summary statistics (values/criteria): metals in fill (%) as box-and-whiskers plots showing minimum, first quartile, median, third quartile and maximum
Figure 4	Standardised summary statistics (values/criteria): metals in fill (%) as box-and-whiskers plots showing minimum, first quartile, median, third quartile and maximum, with adjusted scale
Figure 5	Multiple histograms for metals in fill (mg/kg)
Figure 6	Q–Q plot for arsenic (mg/kg)
Figure 7	Q–Q plot for chromium (mg/kg)
Figure 8	Q–Q plot for copper (mg/kg)
Figure 9	Q–Q plot for lead (mg/kg)
Figure 10	Q–Q plot for nickel (mg/kg)
Figure 11	Q–Q plot for zinc (mg/kg).

Figure 1 Summary statistics, metals in soil (mg/kg) – minimum, first quartile, median, third quartile, maximum



Source: Easterly Point Environmental Pty Ltd

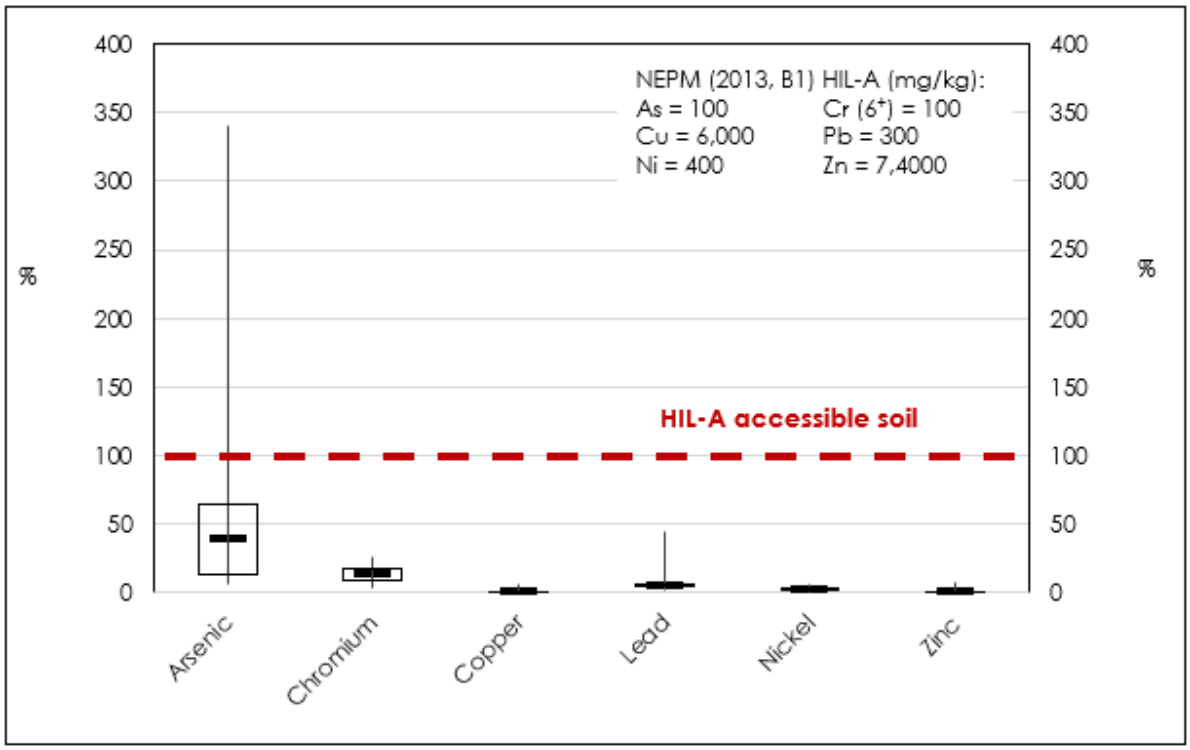
Figure 2 Summary statistics, metals in soil (mg/kg) – minimum, first quartile, median, third quartile, maximum – scale adjusted



Source: Easterly Point Environmental Pty Ltd

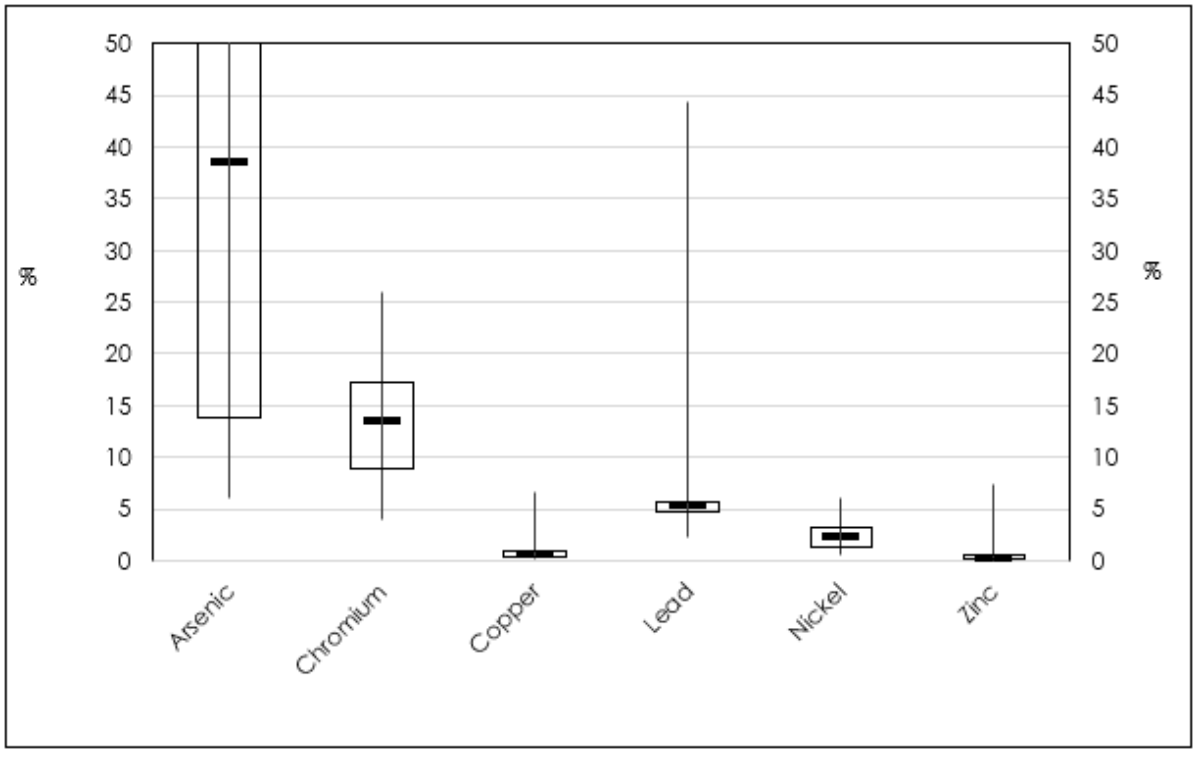
Figure 1 and Figure 2 show the data is generally skewed to the right in the cases of As, Cu, Pb and Zn, as a result of extreme values in the dataset. Cr and Ni look generally symmetrically distributed, suggesting a nearly-normal distribution.

Figure 3 Standardised summary statistics, metals in soil (%) – metals data relative to acceptance criteria



Source: Easterly Point Environmental Pty Ltd

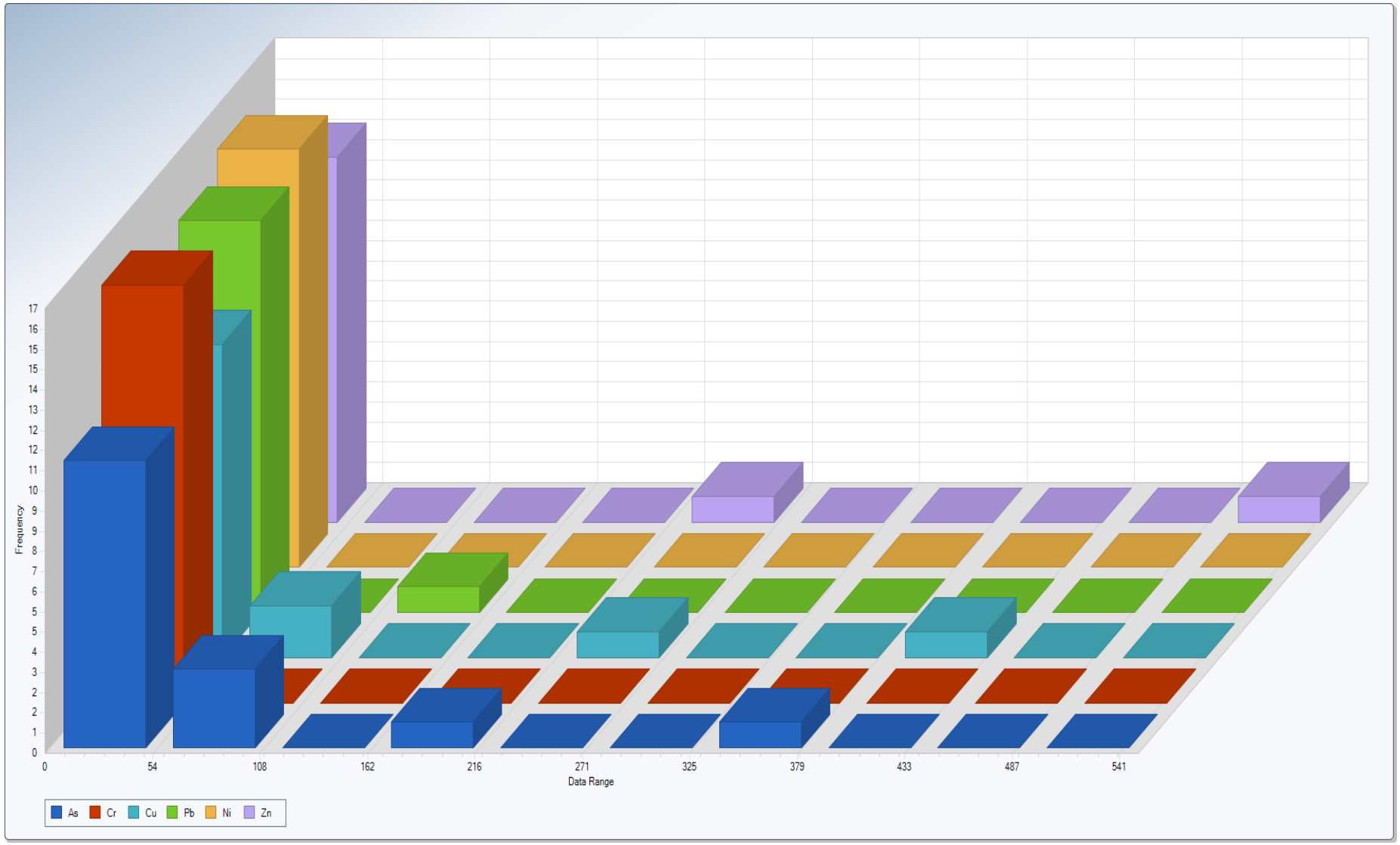
Figure 4 Standardised summary statistics, metals in soil (%) – metals data relative to acceptance criteria – scale adjusted



Source: Easterly Point Environmental Pty Ltd

Figures 3 and 4 show that in relation to the criteria for HIL-A residential with accessible soil, only As exceeds 50% of its criterion, with the maximum As value exceeding the criterion by 341%, that is, more than 250% of the criterion. Cu, Pb and Zn are elevated, but are below HIL-A.

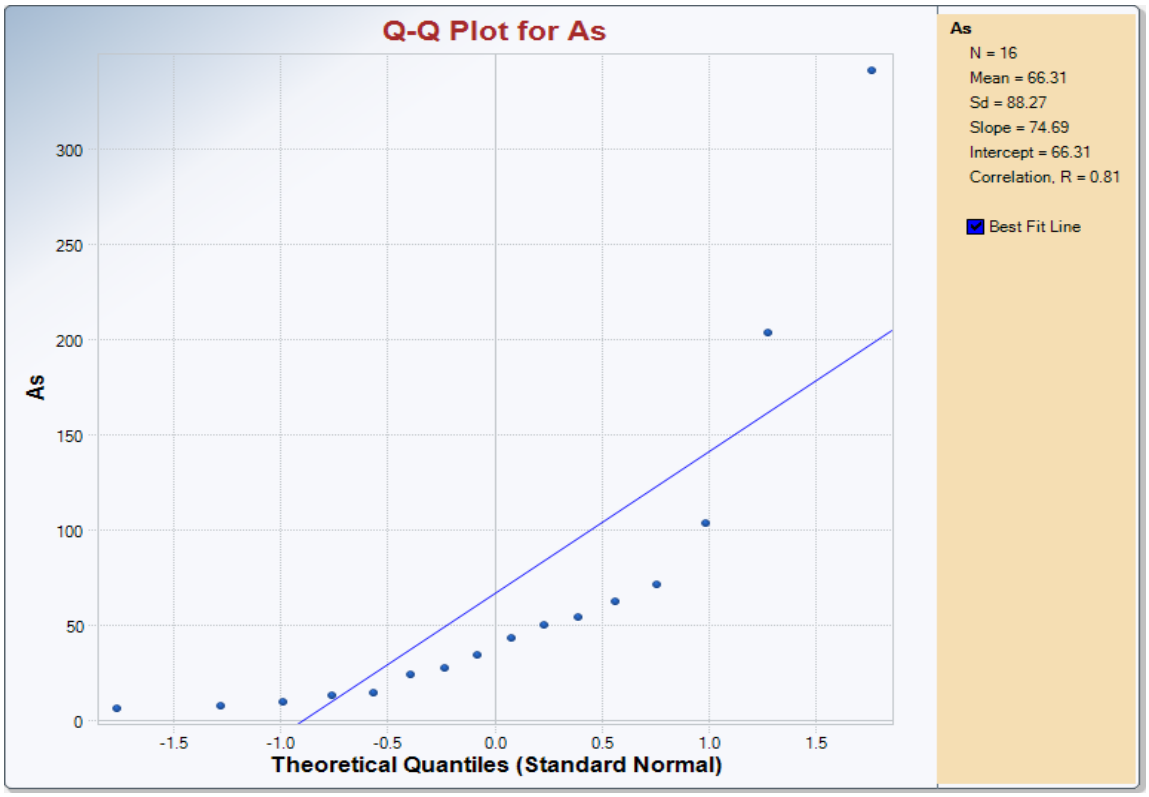
Figure 5 Multiple histograms for metals in soil (mg/kg) – data from Table 1



The x-axis shows concentration of the metal (in mg/kg) and the y-axis shows the number of samples.
Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

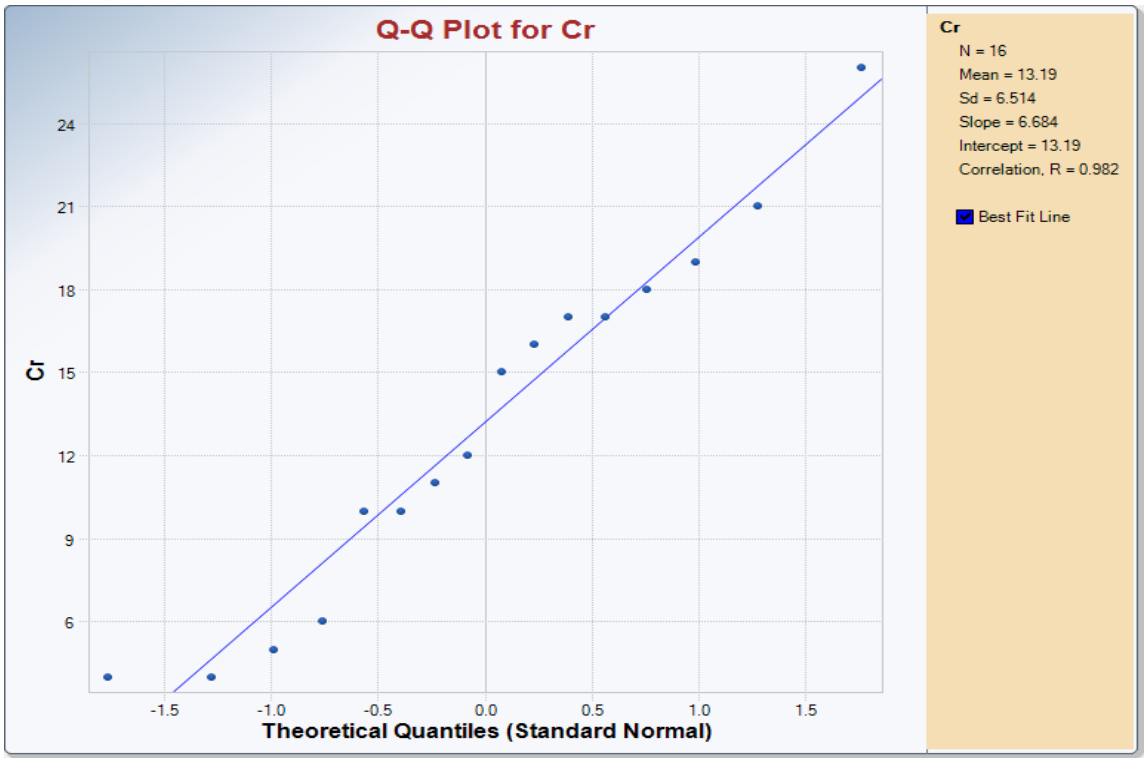
Figure 5 shows that As, Cu, Pb and Zn are right-skewed because of extreme values. As the sample size was small (less than 30 samples), the normality of the distribution cannot be confirmed using histograms.

Figure 6 Q–Q plot for arsenic (mg/kg)



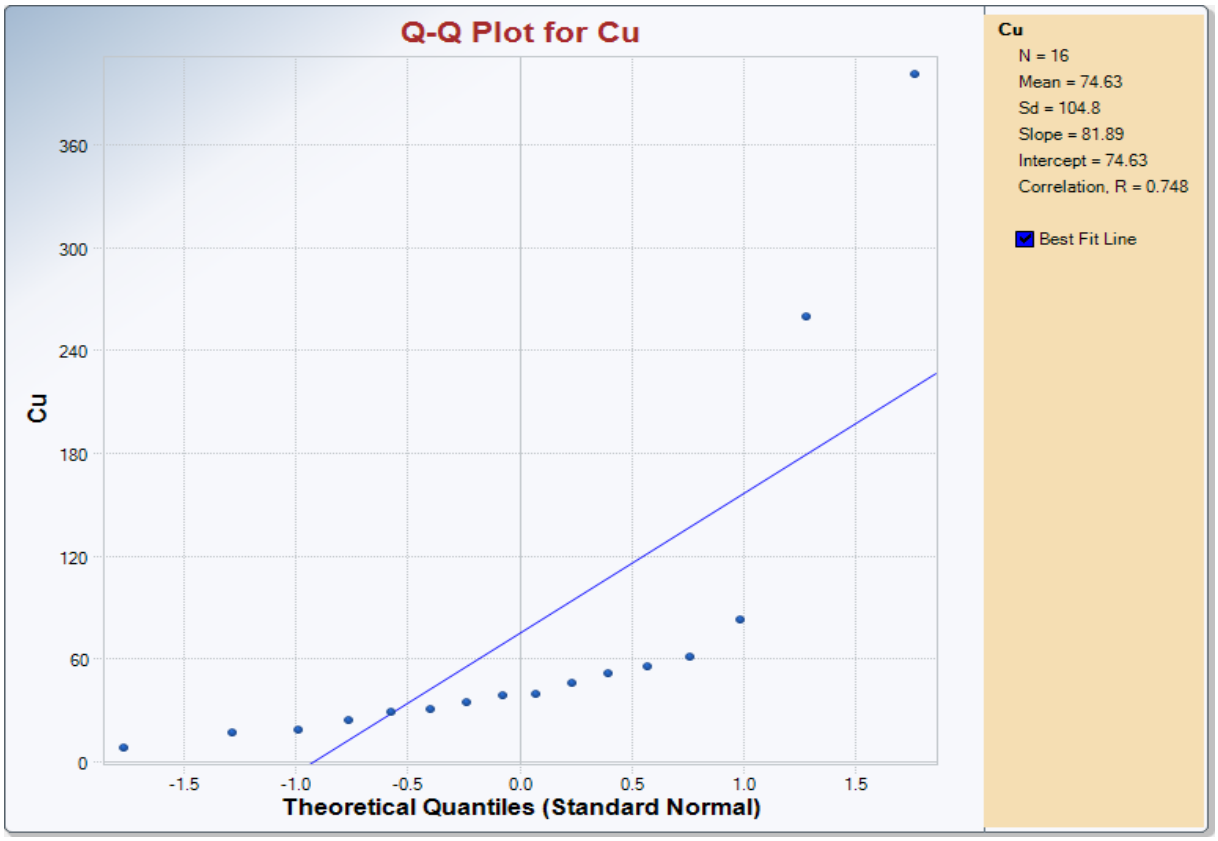
Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

Figure 7 Q–Q plot for chromium (mg/kg)



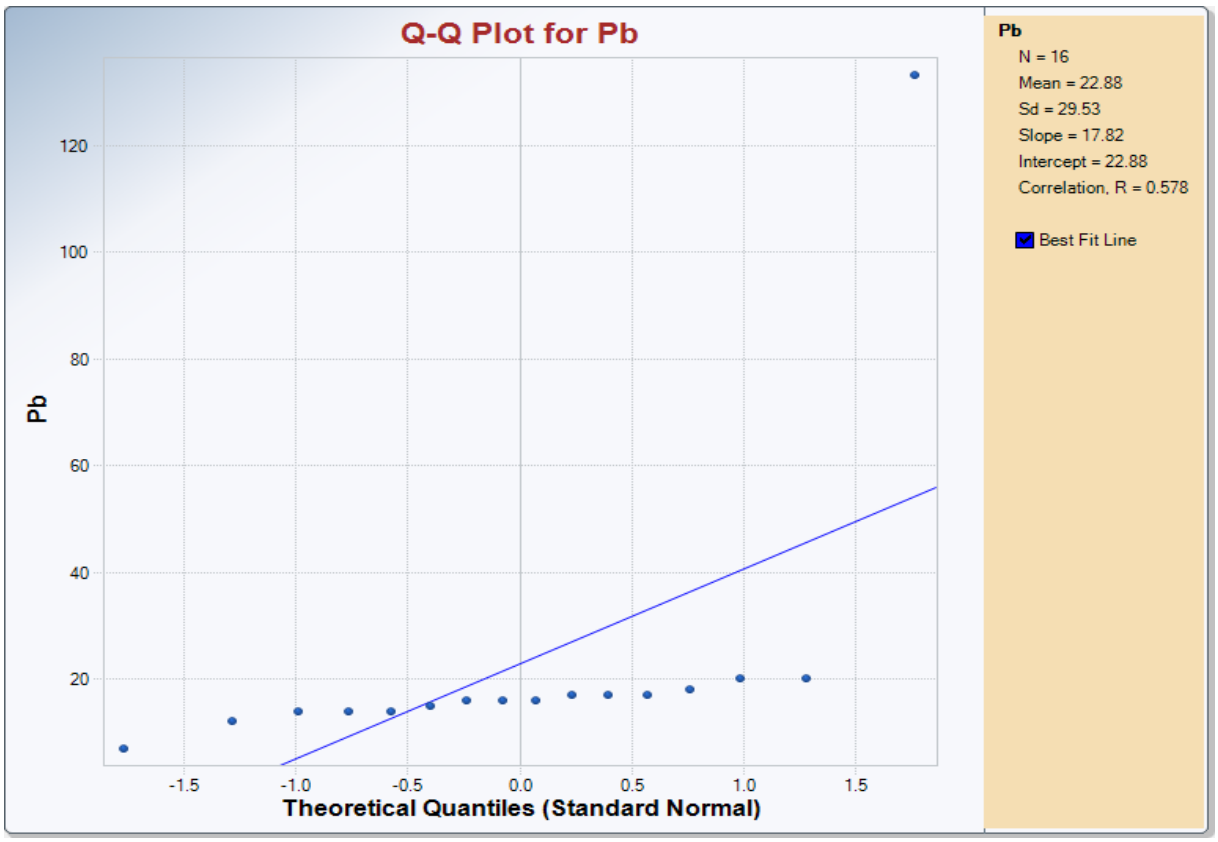
Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

Figure 8 Q–Q plot for copper (mg/kg)



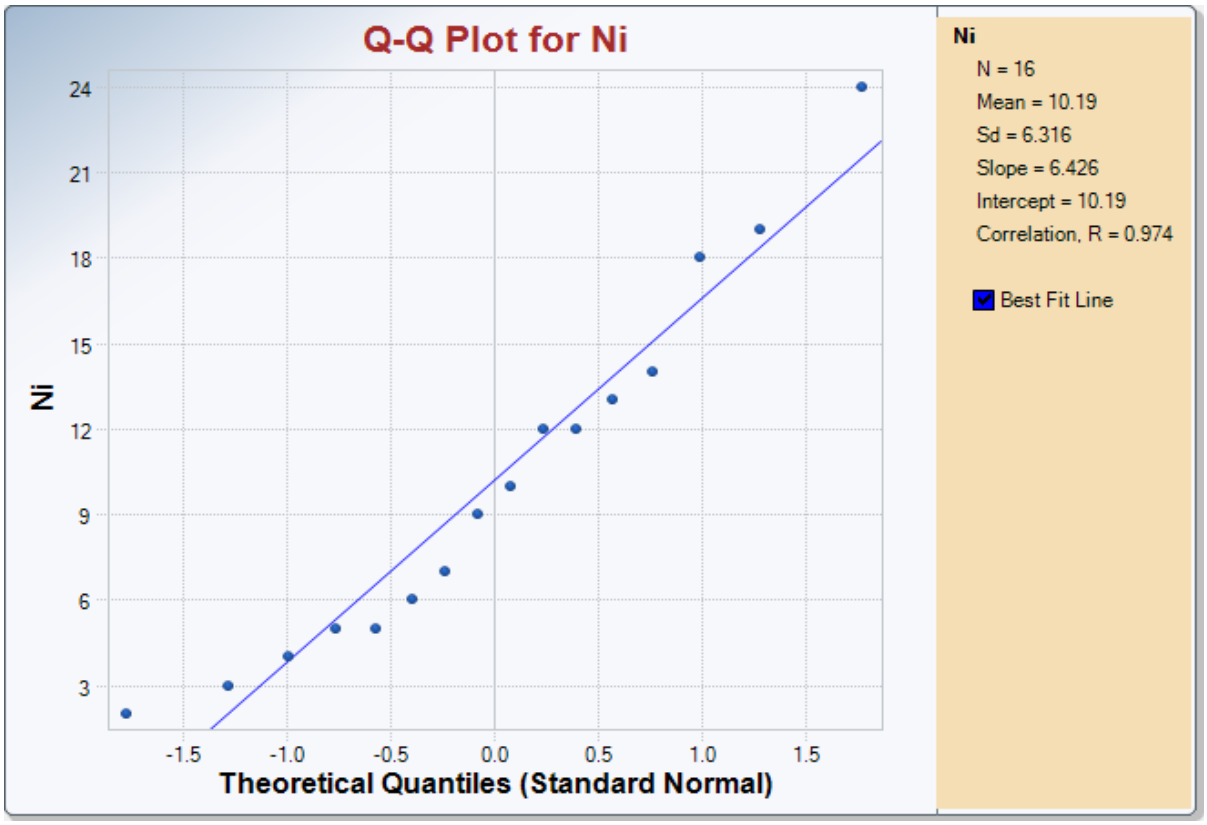
Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

Figure 9 Q–Q plot for lead (mg/kg)



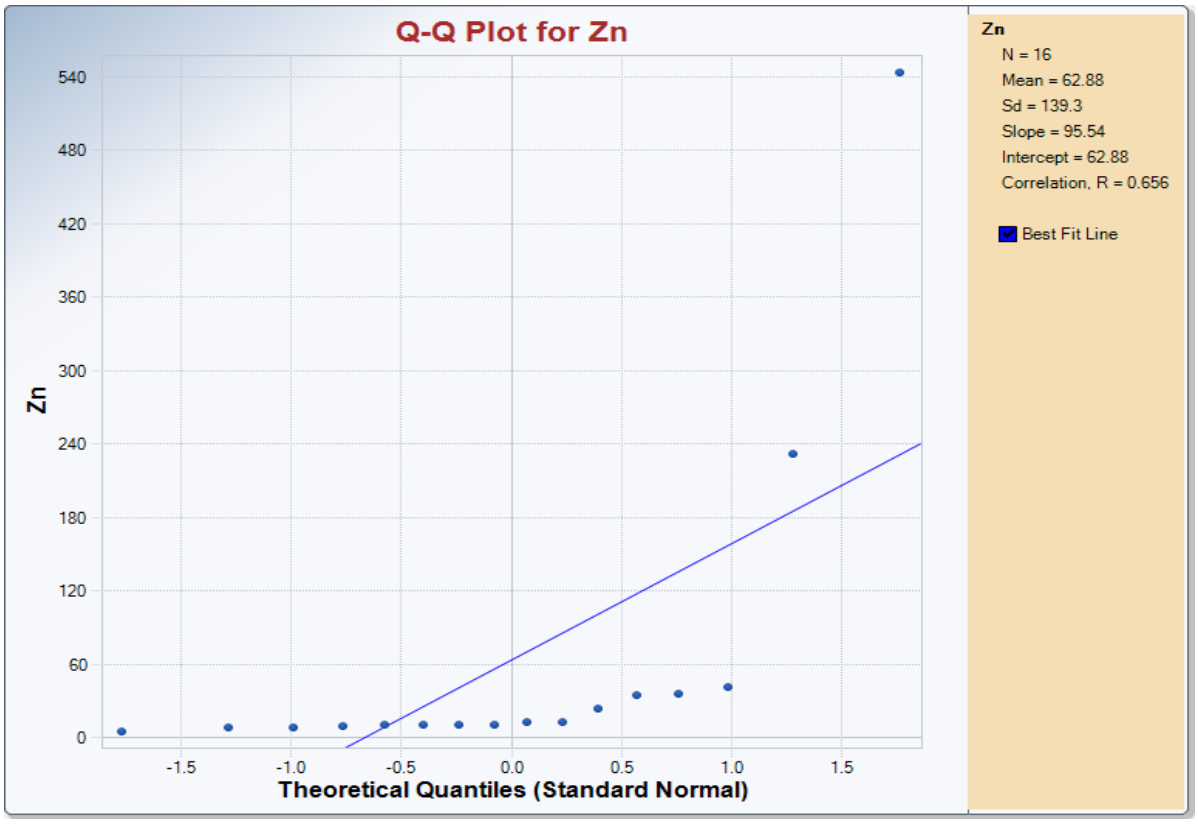
Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

Figure 10 Q–Q plot for nickel (mg/kg)



Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

Figure 11 Q–Q plot for zinc (mg/kg)



Outputs from USEPA's ProUCL, created by Easterly Point Environmental Pty Ltd

The Q–Q plots in Figures 6–11 show that As, Cu, Pb and Zn are unlikely to be nearly-normally distributed, so parametric methods that assume near-normality cannot be used for statistical inference.

Instead, some form of transformation or another distribution type should be used. For Cr and Ni, the Q–Q plots suggest a nearly-normal distribution, so parametric methods that assume near-normality may be appropriate for analysis.

Appendix F: One-sample t-test hypothesis testing

In the assessment of contaminated land, when a decision requires comparing a sampled population to a target value, such as a specified health investigation level (HIL), a one-sample t-test can be used. This is a parametric method, and assumes a nearly normal distribution, at least for sample sizes of less than 30: it is not suitable for highly skewed datasets. See USEPA 2006a, G-9S for non-parametric methods.

Determination

Establish the null hypothesis (H_0) and alternative hypothesis (H_A). EPA policy is to always assume that the site or decision area is contaminated, so the null hypothesis is always written as:

$$H_0: \mu > \text{criterion or action level}$$

The alternative hypothesis for a one-sided test is then:

$$H_A: \mu \leq \text{criterion or action level}$$

The test statistic (t_0) is calculated using the t-score formula:

$$t_0 = \frac{\bar{x} - C}{s/\sqrt{n}}$$

Where:

μ	population mean
t_0	test statistic
$\bar{x}$	sample mean
C	criterion or action level
s	sample standard deviation
n	number of samples
$t\alpha$	critical value.

The critical value ($t\alpha$) is determined from a table of critical values of Student's t-distribution (see Table 4) or by using an appropriate software program. The confidence level ($1 - \alpha$) and the degrees of freedom ($n - 1$) are used to select $t\alpha$.

The test statistic is then compared to the critical value, and the following decisions made:

- if $t_0 < t\alpha$, then fail to reject the null hypothesis that the true population mean is greater than the criterion or action level
- if $t_0 > t\alpha$, then reject the null hypothesis that the true population mean is greater than the criterion or action level and accept the alternative hypothesis that the true population mean is less than or equal to the criterion or action level.

Whereas the signs of t_0 and $t\alpha$ are important in regard to whether an upper-tailed or lower-tailed test is being conducted, when comparing t_0 to $t\alpha$, it is the absolute values that are compared.

The probability or p-value is also determined, either approximately from a table of critical values of Student's t-distribution (see Table 4) or by using an appropriate software program. This is then compared to the selected value of alpha (α), with the following decisions made:

- if p-value > α, then fail to reject the null hypothesis that the true population mean is greater than the criterion or action level
- if p-value < α, then reject the null hypothesis that the true population mean is greater than the criterion or action level and accept the alternative hypothesis that the true population mean is less than or equal to the criterion or action level.

While the sign of the p-value is important in regard to whether an upper-tailed or lower-tailed test is being conducted, when comparing the p-value to α it is the absolute values that are compared.

As critical values and p-values are mathematically related, either approach will always provide the same conclusion.

Worked example

In this example we use the arsenic (As) and lead (Pb) data from Table 1 to determine whether the null hypothesis (H_0) should be rejected in favour of the alternative hypothesis (H_A). The selected criteria are the HILs for a residential land use (HILs-A), and the test is to be conducted at a confidence level of 95%, that is $\alpha = 0.05$.

The null hypothesis is:

$$H_0: \mu > \text{criterion}$$

The alternative hypothesis is then:

$$H_A: \mu \leq \text{criterion}.$$

The test statistic (t_0) is calculated using the t-score formula:

$$t_0 = \frac{\bar{x} - C}{s/\sqrt{n}}$$

For As, n = 16, $\bar{x}$ = 66.3, s = 88.3 and HIL-A = 100, so:

$$t_0 = \frac{66.3 - 100}{88.3/\sqrt{16}}$$

$$t_0 = -1.53$$

For Pb, n = 16, $\bar{x}$ = 22.9, s = 29.5 and HIL-A = 300, therefore:

$$t_0 = \frac{22.9 - 300}{29.5/\sqrt{16}}$$

$$t_0 = -37.54$$

From a table of critical values of Student's t-distribution (see Table 4), at a confidence level of 95% for 15 degrees of freedom, $t\alpha = 1.75$.

Critical value

For As, as 1.53 is less than 1.75, that is, $t_0 < t\alpha$, fail to reject the null hypothesis that the true population mean is greater than the criterion.

For Pb, as 37.54 is more than 1.75, that is, $t_0 > t\alpha$, reject the null hypothesis that the true population mean is greater than the criterion and accept the alternative hypothesis that the true population mean is less than or equal to the criterion.

P-value

For As, from a table of critical values of Student's t-distribution (see Table 4), the p-value is between 0.1 and 0.05, that is, $t\alpha$ is between 1.34 and 1.75. Using a software package, the p-value is calculated to be 0.074. As 0.074 is more than 0.05, that is, the p-value > α, fail to reject the null hypothesis that the true population mean is greater than the criterion.

For Pb, from a table of critical values of Student's t-distribution, the p-value is less than 0.005, that is, $t\alpha$ is > 2.95. Using a software package, the p-value is calculated to be 1.5×10^{-16}. As 1.5×10^{-16} is less than 0.05, that is, the p-value < α, reject the null hypothesis that the true population mean is greater than the criterion and accept the alternative hypothesis that the true population mean is less than or equal to the criterion.

Critical region

In the case of As, t_0 does not fall within the critical region (the area beyond the critical value, $t\alpha$). It is therefore unlikely that the observed test statistic is more extreme than would be expected if the null hypothesis were true. Similarly, as the p-value > α, the probability of observing a p-value as extreme as 0.074 would be high, if H_0 were true. Based on both the critical value approach and the p-value approach, there is insufficient evidence at a 95% confidence level to conclude that the population mean for As is less than HIL-A.

In the case of Pb, t_0 falls within the critical region, and it is likely that the observed test statistic is more extreme than would be expected if the null hypothesis were true. And, as the p-value < α, the probability of observing a p-value as extreme as 1.5×10^{-16} would be low, if H_0 were true. Based on both the critical value approach and the p-value approach, there is sufficient evidence at a 95% confidence level to reject the null hypothesis and to accept the alternative hypothesis that the population mean for Pb is less than HIL-A.

Table 4 Critical values of the Student's t-distribution

Degrees of freedom	Significance level for one-sided interval (α), e.g. confidence limits	15%	10%	5%	2.5%	1%	0.5%
	Confidence level for one-sided interval ($t_{1-\alpha}$), e.g. confidence limits	85%	90%	95%	97.5%	99%	99.5%
	Significance level for two-sided interval ($\alpha/2$), e.g. confidence intervals	30%	20%	10%	5%	2%	1%
	Confidence level for two-sided interval ($t_{1-\alpha/2}$), e.g. confidence intervals	70%	80%	90%	95%	98%	99%
1		1.963	3.078	6.314	12.706	31.821	63.657
2		1.386	1.886	2.920	4.303	6.965	9.925
3		1.250	1.638	2.353	3.182	4.541	5.841
4		1.190	1.533	2.132	2.776	3.747	4.604
5		1.156	1.476	2.015	2.571	3.365	4.032
6		1.134	1.440	1.943	2.447	3.143	3.707
7		1.119	1.415	1.895	2.365	2.998	3.499
8		1.108	1.397	1.860	2.306	2.896	3.355
9		1.100	1.383	1.833	2.262	2.821	3.250
10		1.093	1.372	1.812	2.228	2.764	3.169
11		1.088	1.363	1.796	2.201	2.718	3.106
12		1.083	1.356	1.782	2.179	2.681	3.055
13		1.079	1.350	1.771	2.160	2.650	3.012
14		1.076	1.345	1.761	2.145	2.624	2.977
15		1.074	1.341	1.753	2.131	2.602	2.947
16		1.071	1.337	1.746	2.120	2.583	2.921
17		1.069	1.333	1.740	2.110	2.567	2.898
18		1.067	1.330	1.734	2.101	2.552	2.878
19		1.066	1.328	1.729	2.093	2.539	2.861
20		1.064	1.325	1.725	2.086	2.528	2.845
21		1.063	1.323	1.721	2.080	2.518	2.831
22		1.061	1.321	1.717	2.074	2.508	2.819
23		1.060	1.319	1.714	2.069	2.500	2.807
24		1.059	1.318	1.711	2.064	2.492	2.797
25		1.058	1.316	1.708	2.060	2.485	2.787
26		1.058	1.315	1.706	2.056	2.479	2.779
27		1.057	1.314	1.703	2.052	2.473	2.771
28		1.056	1.313	1.701	2.048	2.467	2.763
29		1.055	1.311	1.699	2.045	2.462	2.756

Degrees of freedom	Significance level for one-sided interval (α), e.g. confidence limits	15%	10%	5%	2.5%	1%	0.5%
	Confidence level for one-sided interval ($t_{1-\alpha}$), e.g. confidence limits	85%	90%	95%	97.5%	99%	99.5%
	Significance level for two-sided interval ($\alpha/2$), e.g. confidence intervals	30%	20%	10%	5%	2%	1%
	Confidence level for two-sided interval ($t_{1-\alpha/2}$), e.g. confidence intervals	70%	80%	90%	95%	98%	99%
30		1.055	1.310	1.697	2.042	2.457	2.750
40		1.050	1.303	1.684	2.021	2.423	2.704
60		1.046	1.296	1.671	2.000	2.390	2.660
120		1.041	1.289	1.658	1.980	2.358	2.617
∞		1.036	1.282	1.645	1.960	2.326	2.576

Modified from USEPA 2006a G-9S.

Appendix G: Two-sample t-test hypothesis testing

When assessing contaminated land, a decision may require two independent populations to be compared – for example, a potentially contaminated area and a background area, or concentration levels from up-gradient monitoring wells and downgradient monitoring wells. In such cases, a two-sample t-test can be used.

This is a parametric method, so the assumption of normality should be checked; see USEPA 2006a, G-9S for non-parametric methods, if those are required. Two-sample t-tests can also be used for paired populations, such as concentrations before and after remediation; again, see USEPA 2006a, G-9S for parametric and non-parametric methods for paired data.

The method used for conducting a two-sample t-test varies depending on whether the variances (s^2) of the two samples are equal or unequal. For environmental data, the variances are generally unequal, and this method is used in the following determination.

Determination

Establish the null hypothesis (H_0) and alternative hypothesis (H_A). As the objective is to compare two populations, the null hypothesis is set to be that the two populations are equal:

$$H_0: \mu_1 - \mu_2 = \delta_0$$

The alternative hypothesis for a one-sided test is then:

$$H_A: \mu_1 - \mu_2 > \delta_0$$

To calculate the test statistics (t_0) for unequal variance, it is first necessary to determine the degrees of freedom (df) using the Welch–Satterthwaite equation:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2)^2}{n_1^2(n_1\text{-}1)} + \frac{(s_2^2)^2}{n_2^2(n_2\text{-}1)}}$$

The test statistic, t_0, is then calculated using the Welch's t-test formula, which a modification of the Student's t-test formula:

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

μ_1 population 1

μ_2 population 2

df degrees of freedom

s_1^2 sample variance from population 1

s_2^2 sample variance from population 2

n_1 number of samples from population 1

n_2 number of samples from population 2

t_0 test statistic

$t\alpha$ critical value

$\bar{x}_1$ sample mean from population 1

$\bar{x}_2$ sample mean from population 2

δ_0 difference (delta) of zero

Critical value

The critical value ($t\alpha$) is determined from a table of critical values of Student's t-distribution (see Table 4) or using an appropriate software program. The confidence level ($1 - \alpha$) and the degrees of freedom are used to select $t\alpha$.

The test statistic is then compared to the critical value, and the following decisions made:

- if $t_0 < t\alpha$, fail to reject the null hypothesis that the difference between the population means is zero
- if $t_0 > t\alpha$, reject the null hypothesis that the difference between the population means is zero and accept the alternative hypothesis that the mean of population 1 is greater than the mean of population 2.

While the signs of t_0 and $t\alpha$ are important regarding whether an upper-tailed or lower-tailed test is being conducted, when comparing t_0 to $t\alpha$, it is the absolute values that are compared.

p-value

The probability or p-value is also determined, either approximately from a table of critical values of Student's t-distribution (see Table 4) or using an appropriate software program. The p-value is then compared to the selected value of alpha (α) and the following decisions made:

- if p-value > α, fail to reject the null hypothesis that the difference between the population means is zero
- if p-value < α, reject the null hypothesis that the difference between the population means is zero and accept the alternative hypothesis that the mean of population 1 is greater than the mean of population 2.

While the sign of the p-value is important regarding whether an upper-tailed or lower-tailed test is being conducted, when comparing the p-value to α, it is the absolute values that are compared.

As critical values and p-values are mathematically related, either approach will always provide the same conclusion.

Worked example

In this example we use the arsenic (As) data from Table 1 to determine whether contamination is limited only to the surficial soils (population 1), and therefore if the deeper soils (population 2) can be considered separately. The descriptive statistics for the two datasets, and the original combined dataset for comparison, are shown in Table 5.

Table 5 Arsenic summary statistics by population (mg/kg) – simulated data from Table 1

Statistic	Surface population 1	Depth population 2	Combined
Maximum	341	54	341
Mean	109.3	23.4	66.3
Medium	66.5	13.5	38.5
Minimum	24	6	6
Variance	12,093.4	390.3	7,792.2
Standard deviation	110.0	19.8	88.3

The test is to be conducted at a confidence level of 95%, i.e. $\alpha = 0.05$.

The null hypothesis is:

$$H_0: \mu_1 - \mu_2 = \delta_0$$

The alternative hypothesis for a one-sided test is then:

$$H_A: \mu_1 - \mu_2 > \delta_0$$

The degrees of freedom is first calculated using the Welch–Satterthwaite equation:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2)^2}{n_1^2(n_1 - 1)} + \frac{(s_2^2)^2}{n_2^2(n_2 - 1)}}$$

$$df = \frac{\left(\frac{12{,}093.4}{8} + \frac{390.3}{8}\right)^2}{\frac{(12{,}093.4)^2}{8^2(8 - 1)} + \frac{(390.3)^2}{8^2(8 - 1)}}$$

$$df = \frac{1{,}560.5^{\,2}}{3.3 \times 10^5 + 340}$$

$$df = 7.45$$

Rounded down to the next integer, the degrees of freedom is seven (7). A conservative approach is to estimate the degrees of freedom by using the smaller of $n_1 - 1$ or $n_2 - 1$: in this case, that number is also seven.

The test statistic, t_0, is then calculated using Welch's t-test formula:

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_0 = \frac{(109.3 - 23.4) - 0}{\sqrt{\left(\frac{12{,}093.4}{8}\right) + \frac{390.3}{8}}}$$

$$t_0 = 2.174$$

Critical value

From a table of critical values of Student's t-distribution (see Table 4), at a confidence level of 95% for seven degrees of freedom, $t\alpha = 1.895$.

As 2.174 is more than 1.895, that is, $t_0 > t\alpha$, the null hypothesis that the population means are equal is rejected, and the alternative hypothesis H_A (that the mean of population 1 is greater than the mean of population 2) is accepted, i.e. $\mu_1 - \mu_2 > \delta_0$.

p-value

From a table of critical values of Student's t-distribution, the p-value is between 0.025 and 0.05, that is, $t\alpha$ is between 2.365 and 1.895. Using a software package, the p-value is calculated to be 0.033.
As 0.033 is less than 0.05, that is, the p-value < α, the null hypothesis H_0 (that the population means are equal) is rejected, and the alternative hypothesis H_A (that the mean of population 1 is greater than the mean of population 2) is accepted, i.e. $\mu_1 - \mu_2 > \delta_0$.

Critical region

As t_0 falls within the critical region, it is likely that the observed test statistic is more extreme than would be expected if the null hypothesis were true. And, as the p-value < α, the probability of observing a p-value as extreme as 0.033 would be low, if H_0 were true. Both the critical-value approach and the p-value approach give sufficient evidence at a 95% confidence level to reject the null hypothesis and accept the alternative hypothesis that the mean of population 1 is greater than the mean of population 2.

Based on review of the summary data, relative to a HIL-A of 100 mg/kg, and the results of the two-sample t-test, it appears that significant impacts relate to the surficial soils rather than the deeper soils. Accordingly, for the design of further investigations and consideration of remedial options, the surficial soils and deeper soils should be considered as separate decision areas. The actual depths which these two populations encompass will need to be determined by further investigations.

Appendix H: Decision errors

Statistical hypothesis testing using a null hypothesis significance testing (NHST) framework – the testing of the null hypothesis (H_0) against an alternative hypothesis (H_A) – can lead to the following four outcomes:

- accepting H_0 when H_0 is true – this is a correct decision for the confidence level of the test ($1 - \alpha$), for example, $\alpha = 0.05$ and confidence level = 95%
- rejecting H_0 when H_0 is true – this is a Type I or α decision error and results in the false rejection of H_0
- accepting H_0 when H_0 is false – this is a Type II or β decision error and results in the false acceptance of H_0
- rejecting H_0 when H_0 is false – this is a correct decision – the power of the test is ($1 - \beta$), for example, $\beta = 0.20$ and power = 80%.

These outcomes are summarised in Table 6.

Table 6 Decision errors in hypothesis testing

Decision made	Actual condition – H_0 is true	Actual condition – H_0 is false (H_A is true)
Accept H_0 (fail to reject H_0)	Correct decision $1 - \alpha$ = confidence level	Decision error (Type II error) **False acceptance**
Reject H_0 (accept H_A)	Decision error (Type 1 error) **False rejection**	Correct decision $1 - \beta$ = power of test

In this instance, the null hypothesis is that the site or decision area is contaminated.

Decision errors are therefore generally defined as follows:

- The site or decision area is considered not to be contaminated when it actually is – a Type I error. Type I errors can lead to unacceptable risks to human health and the environment, so the regulatory framework is established to protect against Type I errors.
- The site or decision area is considered to be contaminated when it actually is not – a Type II error. Type II errors can lead to sites or decision areas being remediated unnecessarily, or land being used for a less-sensitive land use, or unwarranted restrictions on the surrounding environment, such as water-use restrictions or fishing bans.

Appendix I: 95% confidence intervals

Confidence intervals can be used as an indicator of uncertainty around a point estimate, in this case the mean. By choosing a method for expressing uncertainty, a performance metric that quantifies uncertainty can be specified, allowing limits to be established against which the quality and quantity of the data can be compared (USEPA 2006b, G-4).

A method for determining the 95% confidence interval (CI) of the mean for a nearly-normal distribution is shown in this appendix, using the Student's t formula. For mildly skewed datasets, the Student's t-statistic should be used, but for moderate to highly skewed datasets, the confidence interval based on the t-statistic can fail to cover the population mean, especially for small sample sizes (USEPA 2006a, G-9S). It is therefore important to test the data for normality. This is most easily done by constructing normal Q–Q plots, using appropriate statistical software packages. For other distributions or non-parametric methods, refer to USEPA 2006a, G-9S.

Determination

The test statistic is calculated using the two-sided Student's t-UCL formula:

$$95\%\ \text{confidence interval} = \bar{x} \pm t_{\alpha/2,\, n-1}\ \frac{s}{\sqrt{n}}$$

Where:

$\bar{x}$	sample mean
$t_{\alpha/2,n-1}$	critical value
s	sample standard deviation
n	number of samples
$s/\sqrt{n}$	standard error of the mean ($SE_{\bar{x}}$).

The standard error of the mean ($SE_{\bar{x}}$) describes the variability in the sampling distribution, that is, the distribution of means from multiple sampling events of the same population, not the variability in the underlying population. One key feature of the $SE_{\bar{x}}$ is that it decreases as the sample size increases (Devore and Farnum 2005).

The $SE_{\bar{x}}$ multiplied by the critical value gives the margin of error (MoE), which can be defined as the radius, or half the width, of a confidence interval for a particular statistic at a specified confidence level (in the equation above, at a 95% confidence level). The MoE also decreases as the number of samples increases.

The critical value is determined from a table of critical values of Student's t-distribution (Table 4 in Appendix F) or using an appropriate statistical software package. The confidence level ($1 - \alpha$) and the degrees of freedom ($n - 1$) are used to select $t_{\alpha/2,n-1}$ for a two-sided interval.

Worked example

In this example we use the metals data from Table 1 to determine the 95% confidence interval for chromium (Cr) for surface fill ($n = 8$) and all fill ($n = 16$), at a confidence level of 95% ($\alpha = 0.05$).

The 95% confidence interval is calculated using the Student's t-UCL formula:

$$95\%\ \text{confidence interval} = \bar{x} \pm t_{\alpha/2,\, n-1}\ \frac{s}{\sqrt{n}}$$

Surface material

The critical value is selected for a two-sided interval from a table of critical values of Student's t-distribution (Table 4 in Appendix F). At seven (7) degrees of freedom the critical value is 2.365.

For surface soil, $\bar{x} = 15.3$, $s = 6.5$ and $n = 8$:

$$95\%\ \text{confidence interval} = 15.3 \pm 2.365 * \frac{6.5}{\sqrt{8}}$$

$$95\%\ \text{confidence interval} = 15.3 \pm 5.4$$

$$95\%\ \text{confidence interval} = 9.8 \text{ to } 20.7\ \text{mg/kg}$$

All material

The critical value is selected for a two-sided interval from a table of critical values of Student's t-distribution (Table 4 in Appendix F). At 15 degrees of freedom the critical value is 2.131.

For all fill, $\bar{x} = 13.2$, $s = 6.5$ and $n = 16$:

$$95\%\ \text{confidence interval} = 15.3 \pm 2.131 * \frac{6.5}{\sqrt{16}}$$

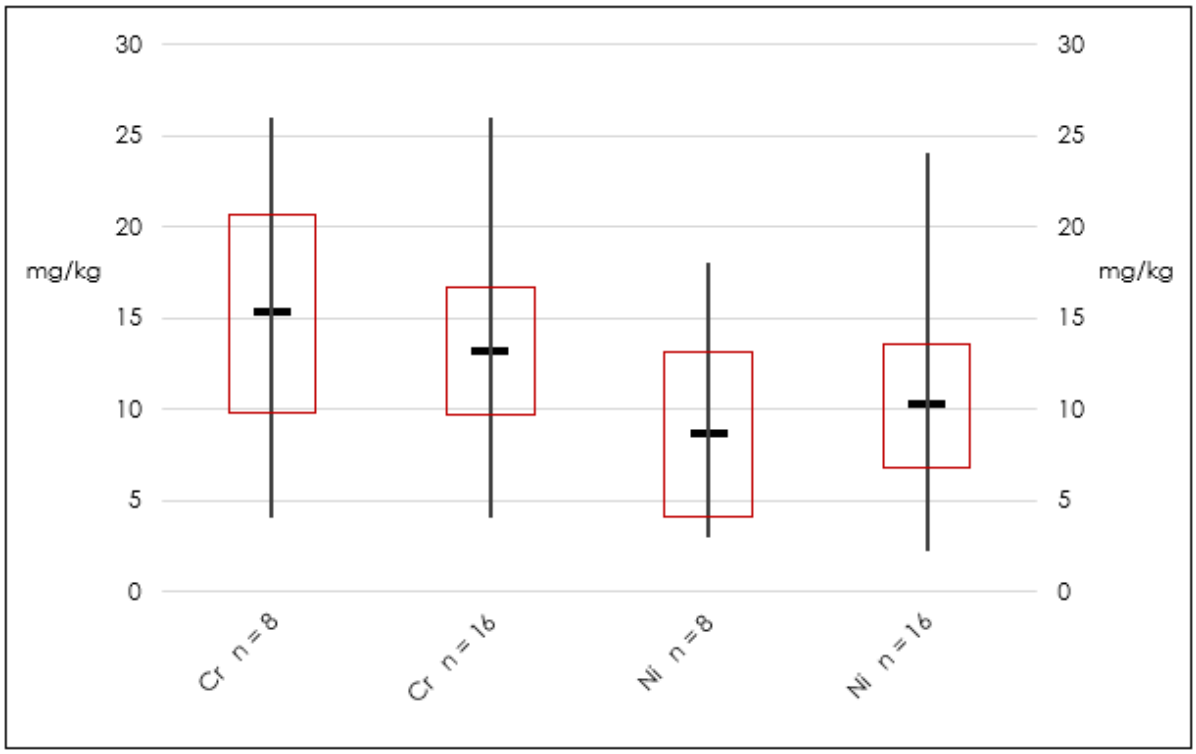
$$95\%\ \text{confidence interval} = 15.3 \pm 3.5$$

$$95\%\ \text{confidence interval} = 9.7 \text{ to } 16.7\ \text{mg/kg}$$

Based on similar datasets, the greater number of samples used in the analysis for all soil samples (16) results in a smaller MoE, and therefore a narrower confidence interval, than does the fewer samples used in analysing the surficial soil (8 samples). Figure 12 illustrates this for both Cr and nickel (Ni); Table 7 and Table 8 show the associated summary statistics.

The maximum probable error (MPE), which is a relative measure based on the MoE divided by the mean (MPE = MoE/$\bar{x}$), can be used to specify the required statistical precision for data collection. For example, for Ni, Table 7 and Table 8 show that the MPE for eight (8) samples is 52.6%, while 16 samples are required to achieve an MPE of 33.0%.

Figure 12 Summary statistics for Cr and Ni data with variable n (mg/kg) – minimum, 95% LCL, mean, 95% UCL, maximum



Source: Easterly Point Environmental Pty Ltd

Table 7 Summary statistics for Cr and Ni data (mg/kg) – surface locations

Surface data	Chromium	Nickel
Number of samples	8	8
Sample mean	15.3	8.6
Standard deviation	6.5	5.4
Standard error of the mean (SE$\bar{x}$)	2.3	1.9
Relative standard deviation (RSD)	42.6%	62.9%
Margin of error (MoE)	5.4	4.5
Maximum probable error (MPE)	35.6%	52.6%

Table 8 Summary statistics for Cr and Ni data (mg/kg) – all locations

All data	Chromium	Nickel
Number of samples	16	16
Sample mean	13.2	10.2
Standard deviation	6.5	6.3
Standard error of the mean (SE$\bar{x}$)	1.6	1.6
Relative standard deviation (RSD)	49.4%	61.9%
Margin of error (MoE)	3.5	3.4
Maximum probable error (MPE)	26.3%	33.0%

Appendix J: 95% UCL$\bar{x}$ for normal distributions

Here we show a method for determining the 95% upper confidence limit of the mean (UCL$\bar{x}$) for a nearly-normal distribution, using the Student's t formula.

For mildly skewed datasets, the Student's t-statistic should be used, but for moderate to highly skewed datasets, the 95% UCL$\bar{x}$ based on the t-statistic may not cover the population mean, especially for small sample sizes. It is therefore important to test the data for normality. This is most easily done by constructing normal Q–Q plots, using appropriate statistical software packages.

Determination

The test statistic is calculated using the one-sided Student's t-UCL formula:

$$95\%\,\mathrm{UCL}\bar{x} = \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

Where:

95% UCL$\bar{x}$	test statistic
$\bar{x}$	sample mean
$t_{\alpha,n-1}$	critical value
s	sample standard deviation
n	number of samples

The critical value is determined from a table of critical values of Student's t-distribution (Table 4 in Appendix F), or using an appropriate statistical software package. The confidence level $(1 - \alpha)$ and the degrees of freedom (n – 1) are used to select $t_{\alpha,n-1}$.

Worked example

Here we use the metals data from Table 1 to determine the 95% UCL$\bar{x}$ for arsenic (As) and chromium (Cr) at a confidence level of 95% ($\alpha = 0.05$), at 15 degrees of freedom (16 – 1 = 15).

The 95% UCL$\bar{x}$ is calculated using the Student's t-UCL formula:

$$95\%\,\mathrm{UCL}\bar{x} = \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

The critical value is selected from a table of critical values of Student's t-distribution (Table 4 in Appendix F). In this instance it is 1.753.

Arsenic

For As, $\bar{x} = 66.3$, $s = 88.3$ and $n = 16$:

$$95\%\,\mathrm{UCL}\bar{x} = 66.3 + 1.753\frac{88.3}{\sqrt{16}}$$

$$95\%\,\mathrm{UCL}\bar{x} = 66.3 + 38.7$$

$$95\%\,\mathrm{UCL}\bar{x} = 105.0$$

Chromium

For Cr, $\bar{x} = 13.2$, $s = 6.5$ and $n = 16$:

$$95\%\,\mathrm{UCL}\bar{x} = 13.2 + 1.753\frac{6.5}{\sqrt{16}}$$

$$95\%\ \mathrm{UCL}_{\bar{x}} = 13.2 + 2.85$$

$$95\%\ \mathrm{UCL}_{\bar{x}} = 16.05$$

The coefficient of variation (CV) for As is 1.3, suggesting a distribution that is not nearly-normal: this is confirmed by the Q–Q plot for As (Figure 6 in Appendix E). Figures 1, 2 and 5 show that the dataset is skewed to the right, indicating that it cannot be appropriately analysed with a Student's t-distribution. Running the data through a statistical software package gives the same conclusion: the software recommends the use of a gamma distribution and calculates a 95% UCL$_{\bar{x}}$ of 120.5 mg/kg.

The CV for Cr is 0.5, suggesting a distribution that is nearly-normal: this is confirmed by the distribution shown in Figures 1 and 2 and the Q–Q plot for Cr in Figure 7. Cr appears to be normally and symmetrically distributed, and therefore the calculated value is likely to be an accurate estimate of the 95% UCL$_{\bar{x}}$. A statistical software package confirmed this: it recommended use of a Student's t-distribution and calculated a 95% UCL for the mean of 16.04 mg/kg.

Based on use of the Student's t-UCL formula to calculate these 95% UCL$_{\bar{x}}$, there is a 95% probability that the mean concentration of Cr will not exceed 16.05 mg/kg. The As dataset needs to be analysed further by another method.

Appendix K: 95% UCL$\bar{x}$ for log-normal distributions

Here we show a method for determining the 95% upper confidence limit of the mean (UCL$\bar{x}$) for a log-normal distribution, using the Land's H-statistic.

This method assumes log-normality, and it is very important to test this assumption. The easiest way to do this is to construct log-normal Q–Q plots using an appropriate statistical software package.

Determination

The test statistic is calculated using the one-sided Land's H-statistic:

$$95\%\ \text{H-UCL}\bar{x} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}}\right)$$

Where:

95% H-UCL$\bar{x}$	test statistic
exp	exponential function, that is, 2.7183 to the power of the value inside the brackets
$\bar{x}$	mean of the log-transformed sample measurements
s_y^2	variance of the log-transformed sample measurements
s_y	standard deviation of the log-transformed sample measurements
$H_{1-\alpha}$	H-statistic critical value, at the stated confidence level ($1 - \alpha$), which depends on the values of s_y and n
n	number of samples

The sample data is transformed using the natural logarithm, that is, a logarithm to the base **e** (2.7183), so $y_i = \ln x_i$, and the descriptive statistics $\bar{y}$, s_y^2 and s_y are determined from the transformed data.

The value of $H_{1-\alpha}$ is selected from Table 9 for a 95% confidence level, based on the values for s_y and n. For other confidence levels, refer to USEPA 2006a, G-9S, and for values of s_y and n not listed in Table 9, use interpolation.

Worked example

Here we use the metals data from Table 1 to determine the 95% H-UCL$\bar{x}$ for arsenic (As) and copper (Cu) at a confidence level of 95% ($\alpha = 0.05$).

Arsenic

The sample data is transformed using the natural logarithm, and for As, $\bar{y} = 3.561$, $s_y^2 = 1.347$, $s_y = 1.160$ and $n = 16$.

The value of H is selected from Table 9. Based on s_y and n, H is between 2.564 and 3.163.
By interpolation, $H = 2.958$.

The test statistic is calculated from:

$$95\%\ \text{H-UCL}\bar{x} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}}\right)$$

$$95\%\ \text{H-UCL}\bar{x} = \exp\left(3.561 + \frac{1.347}{2} + \frac{1.160 * 2.958}{\sqrt{16-1}}\right)$$

$$95\%\ \text{H-UCL}\bar{x} = \exp(5.120)$$

$$95\%\ \text{H-UCL}\bar{x} = 167.4$$

The coefficient of variation (CV) for As is 1.3, suggesting a distribution that is not nearly-normal: this is confirmed by the Q–Q plot for As (Figure 6 in Appendix E). Figures 1, 2 and 5 show that the dataset is skewed to the right. While this suggests that an H-UCL$\bar{x}$ may be appropriate, when a statistical software package was used to generate a range of distributions for calculating the 95% UCL$\bar{x}$, it recommended a gamma distribution.

Although the As data appear log-normal, the Land's H-statistic is sensitive to deviations from log-normality, and produces very high values for large variance or skewness, or where n is small (< 30) (USEPA 2002d). Accordingly, USEPA 2015a recommends that positively skewed datasets should first be tested for a gamma distribution. If the dataset follows a gamma distribution, the UCL$\bar{x}$ should then be computed using a gamma distribution.

Assuming a gamma distribution for the As data, the software package determined a 95% UCL$\bar{x}$ of 120.5 mg/kg – markedly different from the 95% H-UCL$\bar{x}$ of 167.4 mg/kg. As both exceed the HIL-A for As of 100 mg/kg, further data analysis or investigations would be recommended.

Copper

The sample data is transformed using the natural logarithm, and for Cu, $\bar{y} = 3.773$, $s_y^2 = 0.950$, $s_y = 0.974$ and $n = 16$.

The value of H is selected from Table 9. Based on s_y and n, H is between 2.432 and 2.744. By interpolation, H = 2.619.

The test statistic is calculated from:

$$95\%\ \text{H-UCL}\bar{x} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}}\right)$$

$$95\%\ \text{H-UCL}\bar{x} = \exp\left(3.773 + \frac{0.950}{2} + \frac{0.974 * 2.619}{\sqrt{16-1}}\right)$$

$$95\%\ \text{H-UCL}\bar{x} = \exp(4.907)$$

$$95\%\ \text{H-UCL}\bar{x} = 135.2$$

The CV for Cu is 1.4, suggesting a distribution that is not nearly-normal. This is confirmed by the Q–Q plot for Cu in Figure 8. Figure 1 and Figure 5 show that the dataset is skewed to the right, suggesting that an H-UCL$\bar{x}$ may be appropriate. This was confirmed by using a statistical software package to generate a range of distributions for calculating the 95% UCL$\bar{x}$. In both cases the 95% UCL$\bar{x}$ was 135.2 mg/kg.

Table 9 Values of H for one-sided 95% confidence level for computing H-UCL on a log-normal mean

S_y	n = 3	n = 5	n = 7	n = 10	n = 12	n = 15	n = 21	n = 31	n = 51	n = 101
0.10	2.750	2.035	1.886	1.802	1.775	1.749	1.722	1.701	1.684	1.670
0.20	3.295	2.198	1.992	1.881	1.843	1.809	1.771	1.742	1.718	1.697
0.30	4.109	2.402	2.125	1.977	1.927	1.882	1.833	1.793	1.761	1.733
0.40	5.220	2.651	2.282	2.089	2.026	1.968	1.905	1.856	1.813	1.777
0.50	6.495	2.947	2.465	2.220	2.141	2.068	1.989	1.928	1.876	1.830
0.60	7.807	3.287	2.673	2.368	2.271	2.181	2.085	2.010	1.946	1.891
0.70	9.120	3.662	2.904	2.532	2.414	2.306	2.191	2.102	2.025	1.960
0.80	10.43	4.062	3.155	2.710	2.570	2.443	2.307	2.202	2.112	2.035
0.90	11.74	4.478	3.420	2.902	2.738	2.589	2.432	2.310	2.206	2.117
1.00	13.05	4.905	3.698	3.103	2.915	2.744	2.564	2.423	2.306	2.205
1.25	16.33	6.001	4.426	3.639	3.389	3.163	2.923	2.737	2.580	2.447
1.50	19.60	7.120	5.184	4.207	3.896	3.612	3.311	3.077	2.881	2.713
1.75	22.87	8.250	5.960	4.795	4.422	4.081	3.719	3.437	3.200	2.997
2.00	26.14	9.387	6.747	5.396	4.962	4.564	4.141	3.812	3.533	3.295
2.50	32.69	11.67	8.339	6.621	6.067	5.557	5.013	4.588	4.228	3.920
3.00	39.23	13.97	9.945	7.864	7.191	6.570	5.907	5.388	4.947	4.569
3.50	45.77	16.27	11.56	9.118	8.326	7.596	6.815	6.201	5.681	5.233
4.00	52.31	18.58	13.18	10.38	9.469	8.630	7.731	7.024	6.424	5.908
4.50	58.85	20.88	14.80	11.64	10.62	9.669	8.652	7.854	7.174	6.590
5.00	65.39	23.19	16.43	12.91	11.77	10.71	9.579	8.688	7.929	7.277
6.00	78.47	27.81	19.68	15.45	14.08	12.81	11.44	10.36	9.449	8.661
7.00	91.55	32.43	22.94	18.00	16.39	14.90	13.31	12.05	10.98	10.05
8.00	104.6	37.06	26.20	20.55	18.71	17.01	15.18	13.74	12.51	11.45
9.00	117.7	41.68	29.46	23.10	21.03	19.11	17.05	15.43	14.05	12.85
10.00	130.8	46.31	32.73	25.66	23.35	21.22	18.93	17.13	15.59	14.26

From Gilbert 1987

For values of s_y and n not listed, use interpolation.

For other confidence levels, refer to USEPA 2006a, G-9S.

Appendix L: 95% UCL$\bar{x}$ for skewed distributions

Here we give a method for determining the 95% upper confidence limit of the mean (UCL$\bar{x}$) when the distribution cannot be identified. It is based on the non-parametric Chebyshev inequality formula.

The Chebyshev inequality formula makes no assumptions about distribution. For moderately skewed datasets, it yields conservative but realistic values for UCL$\bar{x}$. For highly skewed datasets, it can substantially underestimate the UCL$\bar{x}$, especially for small sample sizes, because it assumes that the standard deviation of the underlying distribution is known. In such cases you can use higher confidence limits (USEPA 2015a): statistical software packages will usually recommend these.

Determination

For unknown distributions, the test statistic is calculated using the one-sided Chebyshev inequality formula:

$$95\%\ UCL\bar{x} = \bar{x} + k_{(1-\alpha)} \frac{s}{\sqrt{n}}$$

Where:

95% UCL$\bar{x}$	test statistic
$\bar{x}$	sample mean
$k_{(1-\alpha)}$	critical value
s	sample standard deviation
n	number of samples

The critical value, k, which is based on the one-sided Chebyshev inequality, is selected from Table 10. It is determined as:

$$k = \sqrt{\frac{1}{\alpha} - 1}$$

Table 10 Critical values based on the Chebyshev theorem

Confidence level %	alpha (α)	k
99	0.01	9.95
95	0.05	4.36
90	0.10	3.00
85	0.15	2.38
80	0.20	2.00
75	0.25	1.73

Adapted from CL:AIRE (2008).

Worked example

Here the metals data from Table 1 is used to determine the 95% UCL of the mean for arsenic (As) and zinc (Zn), at a confidence level of 95% (α = 0.05).

The test statistic is calculated using the Chebyshev inequality formula:

$$95\%\ UCL\bar{x} = \bar{x} + k_{(1-\alpha)} \frac{s}{\sqrt{n}}$$

The critical value is selected from Table 10. For $\alpha = 0.05$, $k_{(1-\alpha)} = 4.36$.

Arsenic

For As, $\bar{x} = 66.3$, $s = 88.3$ and $n = 16$:

$$95\%\ UCL_{\bar{x}} = 66.3 + 4.36\ \frac{88.3}{\sqrt{16}}$$

$$95\%\ UCL_{\bar{x}} = 66.3 + 96.2$$

$$95\%\ UCL_{\bar{x}} = 162.5$$

Zinc

For Zn, with $\bar{x} = 62.9$, $s = 139.3$ and $n = 16$:

$$95\%\ UCL_{\bar{x}} = 62.9 + 4.36\ \frac{139.3}{\sqrt{16}}$$

$$95\%\ UCL_{\bar{x}} = 62.9 + 151.9$$

$$95\%\ UCL_{\bar{x}} = 214.7$$

Table 1 shows that the coefficient of variation (CV) for As is 1.3, suggesting a distribution that is not nearly-normal. This is confirmed by the Q–Q plot for As in Figure 6. Figures 1 and 2, and the histogram in Figure 5, show that the dataset is skewed to the right, implying that a Student's t-distribution is not appropriate for this dataset. A statistical software package confirmed this, and also determined that the Chebyshev inequality method produced an overly conservative $UCL_{\bar{x}}$ for this dataset. The package recommended the use of a gamma distribution; this led to a calculated value for the 95% $UCL_{\bar{x}}$ of 120.5 mg/kg.

The dataset for Zn has a CV of 2.2 and is highly skewed to the right, as can be seen from Figures 1 and 2, the Q–Q plot for Zn in Figure 11, and the histogram in Figure 5. The skewness suggests that a Student's t-distribution is not appropriate for this dataset. A statistical software package confirmed this, also finding that the dataset does not follow a discernible distribution. The package therefore recommended the use of the Chebyshev inequality method, which calculated a 95% $UCL_{\bar{x}}$ of 214.7 mg/kg.

Appendix M: Worked example including results which are non-detects

Background

A site is proposed to be re-developed for recreational purposes as a park. An intrusive investigation has been carried out on a systematic square grid and all results were found to be less than the health investigation levels (HIL) and ecological investigation levels (EIL) for the proposed recreational use (HIL-C) except for cadmium which has a HIL-C criterion of 90 mg/kg. The underlying natural material was found to have no detections of cadmium greater than the limit of report (LOR). A QA/QC assessment has been performed and all data were found to be acceptable.

Results

The dataset shown below consists only of samples from the fill. Eight sample locations were assessed via testpits (testpit 1: TP01_0.2 and TP01_0.5, etc.). Samples were collected from different depths (reflected in the suffix of the sample identification so TP01_0.2 is the sample collected from 0.2 m below ground level at TP01). The field notes observed that the fill was highly heterogenous. The results are shown in Table 11.

Table 11 Results for cadmium

Sample ID	Cadmium (mg/kg)
TP01_0.2	<LOR
TP01_0.5	**95**
TP02_0.2	<LOR
TP02_0.5	<LOR
TP03_0.2	<LOR
TP03_0.5	**119**
TP04_0.2	<LOR
TP04_0.5	<LOR
TP05_0.2	<LOR
TP05_0.5	**93**
TP06_0.2	<LOR
TP06_0.5	<LOR
TP07_0.2	87
TP07_0.5	81
TP08_0.2	<LOR
TP08_0.05	<LOR

The data that is shown as <LOR is less than the limit of reporting (0.4 mg/kg).

There are 16 results in the dataset, of which 10 are less than the limit of reporting. Three results are greater than the assessment criteria. The consultant needs to decide if the material is suitable for the proposed site use or if they should recommend further sampling, a site-specific risk assessment, remediation or management.

Calculations and discussion

The NEPM (NEPC 2013) requires that the 95% UCL is less than the assessment criteria (90 mg/kg) AND the standard deviation is less than 50% of the assessment criteria (45 mg/kg) AND none are greater than 250% of the assessment criteria (225 mg/kg).

Initial calculations for 95% UCL and the standard deviation were performed using the 'substitution method' where zero, half the limit of reporting and the limit of reporting were each used, to substitute when the results were <LOR. The results are shown in Table 12.

Table 12 Calculations of 95% UCL and standard deviation for cadmium using the substitution method

Substitution value (mg/kg)	Percent of the limit of reporting	95%UCL (calculated using statistical software) (mg/kg)	Standard deviation (calculated using statistical software) (mg/kg)
0	0%	79.9	**46.1**
0.2	50%	**144**	**46.0**
0.4	100%	**144**	**45.9**

When considering the substituted results, there is a range of results that the consultant could report. The 0% substitution found that the 95% UCL was less than the criteria, but the standard deviation was greater than 50% of the assessment criterion. As well, the substitution with zero for the limit of reporting potentially underestimated the risk. The 95% UCL calculated for the 50% and the 100% substitutions exceeded the assessment criterion, as did the maximum result. The standard deviation for all substitutions was greater than 50% of the assessment criterion. After considering these results, the consultant might conclude that further sampling, a site-specific risk assessment, remediation or management under an EMP is required. A more sophisticated statistical treatment of the non-detects can be used to support decision making.

Using the same statistical software package (USEPA 2015a), the consultant entered the results so non-detects could be identified, rather than substitute a value of 0%, 50% or 100% of the limit of reporting. The software then calculated the 95% UCL and standard deviation, using a variety of statistical methods (USEPA 2006a) and recommended which values to use, based on considerations of the dataset's statistical characteristics. In this instance, values of 52 mg/kg and 44 mg/kg were obtained for the 95% UCL and the standard deviation, respectively.

This analysis was then used as one strand in a multiple line of evidence approach, where consideration was also given to the depth of the exceedances of the assessment criteria and the likelihood of receptors being exposed to the high cadmium-containing material. The consultant could rely on a more statistically defensible approach and conclude that the fill material was suitable to remain on-site without requiring remediation or management.