# MISUSE OF THE NULL HYPOTHESIS IN DATA REPORTING AND INTERPRETATION

**DECEMBER 1992**


Environment

Environnement

ISBN 0-7778-0607-X

# MISUSE OF THE NULL HYPOTHESIS IN DATA REPORTING AND INTERPRETATION

DECEMBER 1992

Cette publication technique
n'est disponible qu'en anglais.



Log 92-2726-009
PIBS 2201

# MISUSE OF THE NULL HYPOTHESIS IN DATA REPORTING AND INTERPRETATION

Report prepared by:

Donald E. King

Laboratory Services Branch
Ontario Ministry of the Environment

DECEMBER 1992

Log 92-2726-009
PIBS 2201

# MISUSE OF THE NULL HYPOTHESIS IN DATA REPORTING AND INTERPRETATION


## ABSTRACT

The concept of analytical method detection capability (MDC), has been adopted for several decades as the basis for determining measurement reporting limits. One outcome has been inadequate control of method performance at low-levels. As a result, the frequently poor comparability of such data among laboratories, has led to the introduction of progressively higher data reporting limits and the consequent loss or degradation of potentially valuable low-level environmental data. This paper proposes that the statistical process behind this concept of detection has been misunderstood and misapplied. It reviews the application of the 'null' hypothesis, in particular the requirement that this hypothesis be the opposite of what one suspects to be true. The traditional approach applies to the case where an analyte is known to be very probably present (conventional parameters, nutrients, and major ions). But in other cases, (ultra-trace contaminants in drinking water), we know that the analyte is present at levels below our analytical capability to measure. Therefore the statistical process and logic must be inverted. This paper affirms, on a statistical basis, the need to report low-level data, and disavows the application of reporting limits at levels any higher than 3 times the method repeatability as estimated by the within-run standard deviation (Sw). It supports the reporting of low-level estimates, and the adoption of four generic reference points for data interpretation: W (between Sw/2 and Sw), CD (= 3 Sw), DL (= 6 Sw), and QL (=12 Sw).



**Keywords:** detection, data reporting, data interpretation, null hypothesis.

# MISUSE OF THE NULL HYPOTHESIS
# IN DATA REPORTING AND INTERPRETATION

## INTRODUCTION

The concepts of analytical detection and quantitation are critical to the way in which data is reported and interpreted. In the 1960's, A.L. Wilson[1] and others promoted the use of statistical protocols to assess the performance of analytical methods. He proposed a `criteria for detection' (**CD**) and a `detection limit' (**DL**) based on a within-run estimate of standard deviation (Sw). But the multiplicity of terms, definitions, and estimation procedures, introduced over the past thirty years have failed to clarify their application in every day life. The present paper will not review the frequently contradictory terms and definitions currently encountered in the literature. (See Currie[2] for a comprehensive examination of the topic.) Instead, it proposes that the widespread misunderstanding and misapplication of these CD and DL concepts to the issue of data reporting, particularly data censoring, and the subsequent introduction of even higher criteria to restrict reporting of low-level data, has been a disservice to the goal of Quality Improvement in environmental analytical data quality. To many, these various terms and concept convey a negative sense of analytical performance and data quality at low-levels. Low-level measurements are perceived to be inherently invalid, biased, and non-defensible.

Detection, quantitation and higher reporting limits are used to limit the obligation of less competent laboratories to produce well-controlled data. When they are applied in regulatory situations, this censoring of low-level estimates promotes the perception that effluents are free of contaminants, and restricts the ability of a discharger to determine and demonstrate that their effluent is perhaps much cleaner than required. 'Detection' terminology varies widely with respect to function, intent, the type of data to be used, the statistical concepts involved, and application. And they vary in magnitude from 1.64 to many times higher than the within-laboratory analytical standard deviation. Since environmental concerns are focused primarily on those analytes which are only present, if at all, at levels below current quantitation limits, the absence of low-level estimates inhibits our ability to develop appropriate environmental strategies. This paper demonstrates how the statistical logic must be inverted when the target analyte is strongly suspected to be 'absent' (no analytical response at a level above Sw). It demonstrates the need to re-examine data reporting practices and to improve data quality management objectives for low-level data.

Recently Keith[3] has proposed the adoption of three generic levels for assessing low-level data quality. His terms MDL, RDL (2 MDL), and RQL (4 MDL) derive from the US-EPA definitions and exactly parallel the terms CD, DL, and QL (to be discussed below). He proposes the term `Level' rather than 'Limit' to stress their use as criteria for data

interpretation rather than as criteria for data reporting. To further reduce the confusion surrounding 'DL' terminology, the following discussion proposes adoption of a generic term 'method detection capability' (MDC) to replace 'method detection limit' (MDL). This separates analytical performance from data reporting and data interpretation. MDC represents any statistical concept of detection based on an estimate of total method repeatability (Sw), without specifying the level of confidence required or implied. For general comparison among methods and laboratories the factor 3 will provide sufficient confidence and promote comparability of estimates. [Statistical confidence depends on the degrees of freedom (i.e: number of replicate measurements) used to estimate Sw.] MDC will incorporate the impact of the total method on analytical variability for a typical sample matrix. MDC will not incorporate the impact of particularly complex or atypical matrices. This parallels the US-EPA term MDL but avoids the confusion surrounding the term 'detection limit'.

**BACKGROUND**

In Wilson's papers, CD (later $L_c$) represented the level above which a measurement could be considered to be a firm indication of the presence of analyte in a sample. DL (later $L_d$) was set at 2 CD and represented the minimum quantity needed in the sample to ensure that most measurements would exceed CD. These 'decision points' were based on an estimate of the analytical within-run standard deviation (Sw). They included the impact of baseline/blank corrections. CD and DL were to be used to describe the capability or the relative suitability of alternative methods.

But many analysts and data users, concerned that relative error increases dramatically for measurements near CD and DL introduced the concept of a 'quantitation level' (QL). Thus if the uncertainty of a measurement is estimated by ± CD, then results above a QL set at 4 CD would have a relative uncertainty of less than 25%. (Results between DL and CD would be considered to be semi-quantitative.) Finally, since these levels were based on single analyst estimates of repeatability, and it was known that repeatability (and bias) varied among laboratories, the concept of a 'practical quantitation level' (PQL) was introduced. It would be based on among-laboratory estimates of standard deviation (S). In a reasonably comparable group of laboratories the ratio of S/Sw would not generally exceed about 1.3 to 1.5 based on statistical considerations (the F-test). So there is some justification for setting a PQL at 5 to 6 times CD. In principle then, we had the basis for an internally consistent sequence of meaningful criteria for interpreting data.

During the 1970's, instrument manufacturers, professional analytical associations, and regulatory agencies flooded the literature with a multitude of conflicting definitions and terms to describe instrument 'noise' and method detection capability. As an example, Wilson's CD concept has a similar basis as terms such as 'lower limit of detection' and 'method detection limit', while his DL concept is comparable to terms such as 'limit of

quantitation' LOQ. The term PQL appears to have been developed and interpreted in a number of ways. Thus, it represents: a regulatory upper limit for laboratory MDL values, a surrogate estimate for the individual MDL estimates in a multi-parameter scan, the level above which some percentage of participants in an inter-laboratory study can estimate an unknown to within some percentage of the target value, etc.

The lack of data quality, and the general misuse of analytical data during the 1970's to support opposing viewpoints on environmental and health protection issues, required laboratories and data users to be particularly wary when releasing any data which was not absolutely beyond reproach. The concepts of DL, QL, and PQL were adopted as successively higher criteria for restricting the reporting and use of low-level measurements. There is still a lack of consensus among analysts about the validity of such data. And the confusion associated with the concepts of detection has made it difficult to discuss mechanisms for improving the quality of such data.

In environmental studies, relative change is often less important than absolute trends and differences. At lower levels, absolute performance is described by the equation:

$$Sw = S_o + f \bullet C$$

where: Sw is the total repeatability of the procedure
$S_o$ is the effect of methodology,
f is a factor such as 0.05, and
C represents analyte concentration.

The term $S_o$ actually incorporates both the inherent variability of the instrument and the methodology (analytical technique). Thus:

$$S_0^2 = S_i^2 + S_m^2 + S_c^2$$

Typically, for the analytical sample preparation procedures required to analyze complex environmental matrices, $S_m$ is the dominant term in the equation. [The term $S_c$ has been appended to cover the impact of applying corrections for interference from sample matrix constituents. It can be quite critical for some tests, for example, the correction for chlorine interference in the estimation of vanadium by ICP/MS systems which depends on the level of chlorine be corrected.]

It should be apparent that Sw does not vary greatly within the usual operating range. Thus, given a range 1 to 200 µg/L: if Sw is 2 at 5 µg/L, then it will be about 3 at 20 µg/L, and about 7 at 100 µg/L. We must recollect that, statistically, a change or difference of 2 Sw between two measurements is significant (risk of <5%), whether it occurs at DL or 10 DL.

When we are interested in spatial/temporal changes and can read to the nearest 1, the results 2 and 6 indicate a significant change. A similar level of statistical confidence at 100 µg/L requires a change of about 14 µg/L. So although the low-level data may barely meet criteria for 'quantitative' based on relative error, **it is still very meaningful in absolute terms.**

The quality, and hence reportability and interpretation, of low-level data is always a concern. Environmental analysts have had ongoing conceptual problems with the adoption of a purely statistical basis to decide the reportability of data. There are real-life problems in determining the 'zero' point (sources of bias), establishing specific analyte identification criteria, and reducing the impact of the sample matrix (interference). Given that any suggestion of the presence of a hazardous contaminant would have serious public repercussions, many analysts simply refuse to report low-level data. Inadequate control protocols within the laboratory continue to have serious repercussions on the comparability of data from different laboratories and the acceptability of data for regulatory purposes. Analytical and regulatory personnel resort to between-run and among-laboratory estimates of variability to defend their application of high data reporting limits. The incapability of some analytical facilities to produce reliable data has been used to set limits which allows the remaining facilities to neglect the issues of low-level measurement control. Finally, the issue of data quality is aggravated because many environmental regulations resort to method specification rather than performance specification. This approach depends on the assumption that all laboratories have equally competent analysts with expertise in the increasingly technological instrumentations required, and that they apply more than the minimum level of quality assurance and quality control practices. This approach also favours the retention of obsolete analytical technologies which are inadequate for reliable identification of trace pollutants, and inhibits (because of the need for 'equivalency' the introduction of better (more specific, selective, precise, and accurate) methodology.

Ultimately, for a specific methodology, the uncertainty of measurement, the difficulty of setting `zero', and problems in identifying and confirming the presence of a specific analyte, will forestall the generation of reliable low-level measurements pending a significant improvement in analytical procedure. But the commonly accepted practice of rejecting low-level data because of the potential for bias, misidentification, etc., impedes the achievement of data comparability among laboratories.

The primary laboratory data quality issues at low-levels are proper control of baseline, correction for the method blank. Other factors include laboratory contamination, analyte recovery and identification. It is general practice to ignore low-level blank estimates because they are typically below the routine reporting limits. When samples contain reasonable amounts of the target analyte this is no problem. But this practice contributes to measurement bias, affects the quality of all data, and will definitely impair the comparison

of inter-laboratory findings. On the other hand, spatial and temporal patterns and other knowledge about the environment being studied, can be used to affirm the internal consistency of data from a particular laboratory. Even if data sets produced by two agencies are biased relative to each other, they may still affirm any trends present. Initial estimates can always be confirmed by replicate measurements, multiple sampling, or by pattern analysis (contour or time sequence plots) of many single estimates at various points or times. Project planning should recognize and incorporate a strategy to compensate for the variability and potential bias of measurement and sampling activities. Re-sampling is rarely a satisfactory approach for replacing missing information or improving data consistency.

## DETECTION AND PERFORMANCE CAPABILITY

Detection capability is a **performance** characteristic which indicates confidence that the analyte response was sufficiently different from the appropriate 'zero' response to conclude that the response is not an effect of measurement noise or variability. The variability estimate will reflect one or more of the following:

a) the sensitivity of the detector ($S_i$);
b) the variability of the method blank under standard conditions ($S_m$);
c) the impact of other sample constituents on variability and bias ($S_c$).

Part of the confusion in the literature arises from lack of consensus as to what type of data should be used in estimating capability and 'decision points'. Alternatives include various combinations of direct injection or total method replication, standards or natural samples, single analyst or inter-analyst, within or between batch, the level of confidence required (95% or 99%), and the nature of the 'null' hypothesis to be adopted. The most commonly encountered terms among working analysts are the instrument detection limit (IDL), the method detection limit (MDL), and the sample detection limit (SDL). IDL tends to reflect and relate to instrumental readout signal/noise ratios. MDL as described by US-EPA tends to reflect the entire method but only for standard solutions or 'typical' matrix samples. SDL tends to be used when the analytical determination cannot be performed due to severe matrix effects (e.g., foaming, emulsions, severe background effects, in which case no measurement can be performed), or when the result is particularly uncertain because of the uncertainty induced by attempts to correct for determinate sources of error (e.g., chlorine in the vanadium test by ICP/MS). Again, to eliminate the confusion surrounding the terms 'detection limit' and 'detection level', the term detection capability is preferred. Thus these terms should be changed to IDC, MDC and SDC respectively

The most commonly encountered definitions for detection capability reflect the variability of in-run blank estimates. This approach to the question of low-level method reliability depends on the within-batch standard deviation (Sw), which is an estimate of within-batch

single-analyst total method repeatability. Typically, in general conversation among analysts, MDC is expressed as 3 Sw. This factor of 3 has been established by tradition and is traced by them to generally poorly understood statistical principles. Of course the actual factor would vary based on the number of replicates used to determine Sw and the level of risk one is willing to take that a decision may turn out to be wrong. Although this statistical basis is generally not well understood by many bench analysts, most understand that it provides 99% confidence that the analyte is present.

**Detection capability is independent of any bias introduced by correcting for (or ignoring) day-to-day variation in the blank, or within-run drift of the baseline, matrix interference, or inadequate analyte recovery or identification.** Therefore, although bench analysts accept and use the generic concept of MDC, they will still express some lack of confidence in the quantitative nature of such low measurements (as discussed previously). Many laboratories and analysts are well aware that blank estimates can change dramatically but are unable or unwilling to develop adequate control strategies to limit the bias effect.

## STATISTICAL CONCEPTS FOR DATA INTERPRETATION

We as analysts tend to think of detection as the 'appearance' of something that wasn't there. And this is certainly valid. BUT for waste management program managers and data users, decisions are required about the 'disappearance' of an environmental constituent (e.g., as we move from the centre to the edge of a contaminated site). They know it is there. At what point does it merge with the surrounding background and how far has it spread? Ultimately, some of these contaminants become unmeasurable. Detection limits and data reporting practices were developed at a time when we measured only things which were high enough to measure. On the other hand, drinking water treatment engineers are required to provide the 'purest' possible product. They are often required to test for contaminants which previous experience suggests are unlikely to be found at measurable levels. (As measurement analysts we are expected to have tried diligently to measure at levels below those considered to represent a potential health impact.)

Statistical decisions are based on awareness that well-controlled measurements follow the normal distribution (figure 1). It is more likely to observe results close to the average than far away from the average. Given a result R one can predict that the average will be closer rather than farther from the result, and one can predict how different the average might be from the observed result (figure 2). These confidence limits and confidence intervals are fairly straight forward and easily appreciated. Statistical decisions are another matter: they introduce the concept of the null hypothesis and the logic of 'double negatives'. If the test fails to meet a specific criterion then we can't reject the preliminary assumption without an increased risk of making an incorrect decision.

Decisions involve risk: statistical techniques provide a basis for estimating the risk. In statistical decision making, when we know or suspect something to be true, then we MUST choose the opposite possibility as the 'null' hypothesis. In this way the anticipated positive measured result will normally lead us to reject (correctly) the (unlikely) null hypothesis. There are two undesirable outcomes.

Type I decision error: we accept our understanding of the truth when it was actually not true.
Type II decision error: we reject the truth because we were mislead by our data.

A Type I 'false positive' decision error occurs when we reject a 'null' hypothesis (the 'lie') when it was really true, (i.e., we become convinced that what we thought was true has been demonstrated to be true when in reality it was never true to start with). Alpha is the risk of making this decision error. We reduce this risk of a 'false positive decision' by setting alpha at 0.05 or 0.01 . BUT by reducing this risk we delay recognizing that our previous understanding of the 'truth' has been substantiated by the data. A Type II error occurs when we fail to reject the 'null' hypothesis. Beta is the risk, given that our previous understanding of the 'truth' was correct, that our data will fail to substantiate that 'truth' (because we set alpha too high). **Beta represents the risk of rejecting the truth.**

If the construction of a multi-million dollar waste treatment facility depends on our single measurement being above or below some critical operational value (e.g., an effluent quality limit), we prefer to hold off until we get more data. On the other hand, one could argue that the potential presence of a hazardous contaminant, revealed by an improvement in measurement technology, should not be arbitrarily rejected without additional confirmation. But analyst concern about the misuse of low-level data has lead to the practice of using statistical principles to defend the 'censoring' of low-level data. While there might be some justification for setting reporting limits (RL) at the laboratory's MDC value or even higher when we are only concerned about exceeding some measurable health or water quality guideline, the following discussion will demonstrate that the practice of automatically setting RL at 2 to 5 times MDC is statistically indefensible.

Statistical protocols provide a tool for assessing the risk in drawing a particular conclusion. They can not be applied if the data is not reported, or if the performance capability expected or observed is unavailable to the data user.

## ANALYTICAL DECISION POINTS: CD and DL

Our present use of the terms CD and DL will be somewhat derived from A.L.Wilson. Given an estimate of low-level standard deviation (Sw), based on replicate measurements performed within the same batch/run by a single analyst, Wilson recommended setting CD at 1.64 Sw assuming that:

- Sw is estimated from a large number of replicate measurements;
- a Type I error risk of 5% was acceptable (see discussion later).

Wilson set DL at 2 CD (i.e. 3.29 Sw). The basis for this is discussed below. Wilson also discussed the incorporation of variability due to blank estimates. Since this blank issue complicates the discussion, and since most analytical results already incorporate the blank estimate, the following assumes that Sw already includes the impact of the 'blank' correction.

In North America, the corresponding terms typically encountered are MDL (Method Detection Limit) and LOQ = 2 MDL (Limit of Quantitation). MDL is set at 3 Sw assuming that:

- Sw is based on 8 replicates measurements i.e: 7 degrees of freedom (df);
- a Type I error risk of <1% is necessary (see discussion later).

Thus, based on Wilson's definitions, MDL is actually a CD, and LOQ is actually a DL. In the following discussion we will assume that CD = 3 Sw, although any other factor could be used.

- CD is the level above which it is unlikely to observe a result unless analyte is actually measurably present in the sample;
- CD is a factor t times the repeatability Sw. (t depends on df and preselected risk alpha);
- a reliable estimate of Sw requires that measurements be made in increments less than Sw;
- a measurement exceeding Sw/4, Sw/2, or Sw suggests presence of analyte, (the alpha risk is <40%, <30%, <15% respectively.);
- a measured result of about Sw/2 or higher represents a `measurable' result indicating possible 'presence', (but it is certainly not a quantitative estimate).

## ALTERNATIVE NULL HYPOTHESES

The traditional environmental analytical literature uses the concepts of CD (MDL) and DL (LOQ) to distinguish the two levels at which statistical decision points have been set. The

first (CD) is presumably set to minimize the risk (alpha) of deciding that something has happened when it has not, (e.g: to conclude that analyte is present in the sample when it is actually not present). The second (DL) is described in the literature as being selected (based on a preset CD) to minimize the risk (beta) of deciding that something has not happened when it has, (e.g: that analyte is not present in the sample because the analyst has reported `less-than CD' although analyte was actually present). In traditional detection contexts, beta indicates the likelihood of obtaining a result below CD when the sample contains a sample level of DL.

When a certain fact is 'very likely to be true', statistical theory requires us to adopt the hypothesis that it is 'not true'. This ensures that we will normally reject the null hypothesis, and therefore accept the 'truth'. Traditional detection concepts have been developed on the basis that we know and fully expect the analyte is present at easily measurable levels. Therefore:

a) The analyst's perspective always the following 'null' hypothesis (a lie): the sample does not contain a measurable level of analyte.

If this hypothesis were true, levels above CD are statistically unexpected. In fact, based on previous or alternative sources of information, we know that measurements above CD are **extremely likely** (e.g, calcium in water, phosphate in wastewater).

b) The client's perspective adopts the so-called 'alternative' hypothesis (the truth): the sample does contain the analyte, preferably at a level above DL (2 CD).

If this is true then measurements below CD will be avoided. The client wishes to avoid data in this range because it is typically censored. If there is concern that the actual level of analyte will fall below the analyst's reporting limit, the client has the option to expand the sampling design to incorporate replicate samples, or replicate measurements, (or both). Of course, if the data is not going to be reported, all this extra effort will be wasted.

In principle CD must be determined before we can set DL. To estimate CD we estimate the within-run standard deviation Sw and select a risk alpha and then look up the critical Student's t-factor. Thus, the MDL as defined by the US-EPA can be set at 2.998 times Sw based on 8 replicate measurements and a risk level alpha of <1%.

In principle DL and CD can be set anywhere in relation to each other. Thus:

- If DL is set exactly equal to CD then the risk beta of a result below CD is 50%. If the reporting limit is then set at CD, a Type II decision error can occur 50% of the time.

- If DL is set at 2 CD then the risk of a result below CD is <1%.
  A Type II error will occur less than 1% of the time, so long as results at or above CD are reported. But if the reporting limit is adjusted to DL = 2 CD then beta remains at 50%.

In general practice DL is set at 2 CD. But 3 Sw can be interpreted to be either a CD or a DL. And some of the confusion in terminology arises from the fact that the value 3 Sw can be obtained by preselecting various combinations of alpha and beta risk, taking care to estimate Sw from the corresponding number of replicate measurements. The following optional mechanisms for obtaining a factor of 3 should demonstrate that the common statement 'my detection limit is 3 Sw' carries no specific statistical interpretation. The following are only a few of the possible optional interpretations of the value 3 Sw:

- a CD with a risk level of alpha <1% (df = 7);
  (null hypothesis that sample mean is located at 'zero', tail facing upwards)

- a CD with a risk level of alpha <0.1% (df = large, e.g. >60 ); (null hypothesis that sample mean is located at 'zero', tail facing upwards)

- a DL with a risk level of alpha <1% and beta=50% (7 df) (* see note below); (null hypothesis that sample mean is located at 'zero', with the alternate hypothesis that the sample mean is located at CD, tail facing downwards to `zero').

- a DL with alpha <7% and beta <7% (large degrees of freedom), or

- any other combination of alpha and beta risk levels at appropriate degrees of freedom which yield a factor equal or close to 3!

Note: **Beta varies depending on how data is reported/censored**. The proper interpretation of low-level data requires that all data be reported and that the CD value be available to the client. And keep in mind the analyst's warning that all bets are off if the sources of determinate error are not controlled, or the identity of the analyte can not be confirmed.

The above statements were developed in the context that the 'null' hypothesis is that analyte is `absent'. But there are two alternatives in environmental sample measurement:

a) The analyte is generally known to be present at measurable levels, (e.g: major ions, conventional parameters, some trace metals, etc.).

b) The analyte is generally known or strongly suspected to be absent, or only present at barely measurable levels (e.g: mercury in water, many organic pollutants, ultra-trace elements, etc.).

It is only relatively recently that we have become interested in the quality and interpretation of so-called 'less-than' data for analytes we 'very much suspect' to be absent (i.e: not measurable). [For some analysts the term 'measurable' implies the amount present to be above the quantitation level.] For statistical correctness: when we suspect that analyte is 'absent', we must hypothesize it to be 'present'. The anticipated 'zero' result will allow us to reject the 'null' hypothesis, and conclude that the analyte is 'absent'. (Of course this hypothesis requires us to state a level such as CD to represent 'present'. The statement 'absent' is then true in the context that the 'null' hypothesis was 'present at a level of CD'.)

## DETECTION HYPOTHESIS

Case a) The analyte is known or strongly suspected to be present:

NOTE: We MUST first assume the 'null' hypothesis = absence (sample content of analyte is less than about Sw/4).

[Given that we must be able to measure in units smaller than the value of Sw in order to estimate it reliably, it is certainly feasible to measure at levels below some factor times Sw.]

Therefore, given results in the vicinity of Sw or higher, then:

- given a normal distribution and a one-tailed test (facing upscale)
- we choose a critical level or decision point CD (e.g., 3 Sw), and
- if the result R is at or above CD we will reject the null hypothesis,
- so we conclude that analyte is 'present' (no surprise!).
- the statistical risk that analyte is not present is 'alpha'
  (when CD = 3 Sw, alpha is set for 0.01 or 1% risk or less)
  (CD could be set at Sw, for a statistical risk of <15%)
- for analytes 'known' to be present the real risk is negligible

**********

In theory, when the level CD is set lower, the risk (alpha) of Type I (false positive) decision error increases, (i.e the decision to reject the null hypothesis has a higher risk of being wrong). BUT, given that we already 'know' that the analyte is likely present (that's why we chose the null hypothesis of 'absent'), it is quite likely that results R below CD will be obtained from samples which contain between Sw/4 (not zero) and R + CD. This fact leads to the 'alternative to null hypothesis', namely that the sample contains CD (knowing that it

probably contains more).

If we examine the lower tail of a normal distribution centred on CD, we note that:

- a result below CD will be observed 50% of the time, (i.e. beta = 0.5 or 50%)
- a result below about Sw/4 will be observed less than (100 alpha)% of the time.

**BETA RISK is THE RISK OF REJECTING THE TRUTH**

Failure to report results below CD WILL cause the uninformed data user to conclude that the analyte was not present. Given previous knowledge of probable presence, this decision has a high probability (50% risk) of Type II (false negative) decision error. When we set the reporting limit at an even higher level (as generally practised) we INCREASE this beta risk to virtual certainty.

**DETECTION HYPOTHESIS**

Case b) The analyte is known or strongly suspected to be absent:

NOTE: We MUST first assume the 'null' hypothesis = presence.

[Given that we cannot infer presence (when the null hypothesis is 'absent') until results R exceed CD, we can also appreciate that samples containing less than CD will rarely give a result greater than 2 CD.]

Therefore we can decide that absent in the sample will mean that the sample actually contains not more than DL = 2 CD, and:

- given a normal distribution and a one-tailed test (facing toward zero).
- if the result R is below CD we will reject the null hypothesis.
- so we conclude that analyte is not present at levels above DL.
- the statistical risk that analyte is 'present and >DL' is 'alpha'.
- for analytes 'known' to be absent or certainly <CD the real risk is negligible.

Here, Type I error (false positive decision) arises from results below CD but obtained from a sample which contained more than DL = 2 CD (which must be demonstrated by subsequent replicate measurements). BUT, given that we already know that the analyte is likely absent (that's why we chose the null hypothesis of 'present' it is likely that results below CD will only be obtained from samples that do not contain more than R + CD. This leads to the alternative hypothesis to null hypothesis', namely that the sample contains 'zero' (i.e. below about Sw/4).

Type II error occurs when R is at or above CD. BUT this is unlikely unless the analyte was present initially, in which case it would be inappropriate to adopt this as the 'null' hypothesis. Therefore it is important that we have prior knowledge that the sample really does not contain the analyte at an appreciable level.

If we examine the upper tail of a normal distribution centred on 'zero', we. note:

- a result above 'zero' will be observed 50% of the time'.
- a result above CD (3 Sw) will be observed < 1% of the time.

Given a result R below CD, we have the option to report: either the result R, or the conclusion that sample contains less than R + CD.

**BETA RISK is THE RISK OF REJECTING THE TRUTH**

Failure to report results below CD WILL cause the uninformed data user to conclude that the analyte was not present. Since this is likely true the real risk of a 'false negative' conclusion is extremely low.

The importance of selecting the proper 'null' hypothesis seems to have been poorly understood by many who have promoted the use of statistical procedures for evaluating presence or absence of a target analyte in environmental sample analysis. Text boxes 1 and 2, and figures 3 and 4, review the logic associated with these two contradictory hypotheses. In as much as an improper `null' hypothesis is applied when interpreting low-level data, it is proposed that the detection definitions and related conclusions as described in the environmental analytical literature are frequently improper.

**DATA REPORTING AND INTERPRETATION**

Traditionally the bench analyst equates trace positive measurements with a high risk of 'false positive' decisions, and has preferred to withhold results that fall any where near CD. This practice is justified on the analyst's awareness of the potential for a wide variety of sources of bias or misidentification in analytical measurements, particularly at low levels. When the 'null' hypothesis is 'absent', a 'positive' result below CD carries a higher risk of a 'false positive' decision by the client. Therefore, there has been a strong proscription against the reporting of such data. 'False positive' and 'false negative' decision errors are inverted when the initial 'null' hypothesis should be 'present'. Thus, a `positive' result between CD and DL carries a risk of a 'false negative' decision by the client, in which case the client decides that the analyte is present when it was not. This need to invert the logic provides a strong argument for prohibiting the use of the terms 'false positive' and 'false negative' by bench analysts when discussing the validity of their measurement results. These terms

apply to **decisions** not to measurements. When the analyte is expected, low positive values are probably not false positive values.

Based on the previous discussion, it appears to be difficult from a statistical perspective to defend the 'censoring' of low-level data. But the analyst must still retain responsibility for ensuring that the client is fully aware of the quality of data being reported. When the client's needs are well known, and the analyst is fully capable of providing measurements which meet that need, there may be no need to report low results. But when analytes are expected to be 'absent', and assuming that the client has good reason to request analysis at or near the current limits of detection capability, and assuming that there are no alternative better methodologies available, then (based on the logic in text boxes 1 and 2), it should be clear that there is no logical support for using DL (LOQ) or any value higher than CD (MDL) as the basis for setting a data reporting limit. Hunt and Wilson come to the same conclusions in the section on data reporting in their text "The Chemical Analysis of Water" [4].

When decisions must be made based on low-level data it is essential that reported results must be adequately qualified. There is some support for even changing the way in which it is reported. Thus, it is appropriate to report that "on the basis of controlled laboratory measurements, the sample is deemed to contain":

a) an amount R falling in a range between R - CI and R + CI;
b) an amount not exceeding R + CI; or,
c) an amount not below R - CI.

where CI is a confidence interval which can be approximated by CD at low levels.

Recall that two-thirds of the time the sample level is expected to lie within CD/3 (Sw) of the observed value. [Although we would not normally be concerned until a difference exceeds 3 Sw, one should keep in mind that results which differ by more than Sw may well indicate that the two samples on which the measurements were made may have different amounts of analyte. At the bench level this may indicate non-representative aliquotting or problems with sample homogeneity.]

We must be careful to distinguish between 'trace measurements', and 'trace sample levels'. The former are represented by results R below CD. The latter are represented by results R between CD and 2 CD. When a measurement is only one of several to be performed on the same or similar specimens, each measurement result must be recorded and reported for future statistical summarization. Simple doubt, or fear of consequences if the result is incorrect, is not sufficient reason for withholding measurement data. If there are real concerns about data quality, the measurements should be repeated to either gain confidence or to verify that the previous value was incorrect before withdrawing it.

In addition to the question of 'censoring', care should be taken when `rounding-off' or `truncating' (again see Hunt). For instance: the practice of rounding results which start with a `1' to the nearest two figures, is inappropriate since it can introduce bias of as much as 10 to 20%. For example, both results 106 and 115 are typically rounded to 110 units. But the 115 may be an under-estimate from a sample containing more than 120, whereas the 106 might be an over-estimate from a sample containing less than 100 units. In general results should be recorded in increments that are smaller than the analytical repeatability Sw. If an estimate of Sw is smaller than the reporting increment, it should be considered to be an invalid estimate of Sw, and the readout system should be enhanced to provide finer reading increments.

**MOE AND LOW-LEVEL DATA**

Since the early 1970's the Ontario Ministry of the Environment (MOE) laboratories have been intimately involved in the generation, control, and reporting of low-level data. This position arose from our participation in a series of International Joint Commission (IJC) studies of Great Lakes Water Quality, to record the background levels of environmental major ions, nutrients, and trace pollutants. In a series of interlaboratory performance studies, it was observed that some analysts, who were using comparable analytical systems, had censored low-level data at levels well above the statistical detection limit while others reported down to 'zero'. This censored data often covered the range of environmental concentrations being observed by the laboratories that reported down to 'zero'. In particular, interlaboratory bias appeared to be better controlled in those laboratories which reported low-level data.

To facilitate laboratory performance evaluation, the Data Quality Subcommittee of the IJC Water Quality Board requested that all measurements be reported down to 'analytical zero'. They identified two remark codes W and T to replace the inconsistently applied ‘less-than' (<) code. To report a 'zero' the analyst would use the less-than format (e.g: < 2) with the remark code W. Any value coded W represented the smallest reading increment, which is necessarily smaller than the within-run standard deviation Sw. The value reported with W provided a means for distinguishing the relative analytical capability of the laboratories. (Some laboratories were set up for wastewater analysis while others were set up for near-shore or open-water investigations.) Laboratories were then free to use the code T to qualify any reported low-level data (values at or above their W value which they would otherwise have censored).

This W and T protocol was adopted later as an ASTM standard[5] for reporting low-level data. But it sets T at 1.64 Sw (Wilson's CD). If T suggests 'trace' or 'tentative' measurements, the ASTM practice sets T too low and is not really appropriate for use in the 'real world', where data is typically considered to be of inadequate quality and is typically censored at much higher levels. MOE laboratories did not introduce the routine use of the T qualifier at that

time since they already reported low-level measurements for Great Lakes water quality programs. But in the early 1980's these laboratories uniformly adopted the remark codes <W and <T to qualify 'zero' and 'trace' measurements[6] and began reporting low-level data for all Ministry programs.

More recently, in its Municipal and Industrial Strategy for Abatement (MISA) regulation[7], Ontario established a regulated method detection limit (RMDL) criteria for acceptable low-level performance. Each laboratory was required to determine its method detection limit (MDL) by a particular protocol[8] based on the US-EPA definition of MDL[9]. The laboratory MDL was then required to be at or below the RMDL value. Private laboratories were required to report results down to their calculated MDL, and were encouraged to report results below this MDL[10]. They were provided with the remark codes <W, <DL, and <T to qualify reported results. In this regulation <T covered the range from the laboratory's MDL up to the regulated MDL criteria (RMDL). The code < was reserved for the case where no measurement was possible due to sample matrix effects. Any value accompanied by <W, <DL or <T was used as reported. Values accompanied by < were considered as non-measurements and were excluded. Evaluation of this low-level data provided a quality assurance tool for assessing data at or near the RMDL value. In as much as the MISA program was specifically directed at establishing effluent quality limits for those parameters above RMDL, it was found that this low-level data was critical to the evaluation of the likely presence/absence of analytes when results were observed in the vicinity of RMDL. And the need to report such data appears to have had a positive impact on the management of method blank and related baseline data.

## LOW-LEVEL DATA QUALIFIERS

The Ontario Ministry of the Environment laboratories currently provide data qualifiers (remark codes) for use when reporting low-level data. These include:

<W: measured value is 'zero' i.e below W (where W is between Sw/2 and Sw);
<DL: measured value is below 3Sw (MDL) (=CD in the above discussion);
<T: measured value is trace or tentative.

For consistency one should consider changing <DL to <CD. One could also consider the codes:

<S measured value is below the sample specific SDC as determined for this sample,
<C Confidence statement: sample is deemed to contain less than C = R + CD.

As a general practice, the readout system should be sensitive enough to ensure that estimates of Sw are larger than the readout increment (RI) used (see figure 1). In MOE laboratories the W value is determined as follows:

a) if Sw is in the range 5 to 10 units*, W = 5;
if Sw is in the range 2 to 5 units, W = 2;
if Sw is in the range 1 to 2 units, W = 1.
(* the location of the decimal point is irrelevant)

b) if Sw is smaller than RI, then W can be set at RI provided the typical sample contains easily measurable levels of the analyte, otherwise one should attempt to improve the RI.

When using <W, <T, etc., the accompanying reported value must always include the impact of dilution\concentration factors. Since W (the maximum allowed reading increment) is usually set between Sw/2 and Sw, and given the uncertainty attached to any estimate of Sw, it is valid (see figure 1) to state that:

- 5 W is essentially equivalent to 3 Sw (=CD) (=MDL),
- 10 W is essentially equivalent to 6 Sw (=DL) (=LOQ).

In the MOE laboratories the following practice has been used for <T in qualifying low-level data:

- for conventional analytes, major ions, nutrients, T = 5W = approx. CD (MDL);
- for metals and trace organics, T = 10W = approx. DL (LOQ);
- for the MISA program, T was set at RMDL and is unrelated to a given laboratory's W.

In manual systems, reading increments may reflect chart divisions on a strip chart manual readout system, the nearest 0.5 division on a burette, the nearest even value (0.2, 0.4, etc.). Note that for electronic readout systems, every retained reading increment is 'significant' since it reflects the true output of the device. The number of digits to be retained depends on method performance.

The 'significance' of analytical data is based on the estimate of variability and bias, NOT on the number of digits reported. When we report data we should probably avoid the phrase 'significant figures'. This is a mathematical concept used to determine the number of digits in a final calculated result depending on how the constituent values are multiplied or added, on the assumption that the constituent values were previously rounded or truncated in a particular way. It is generally better to report too many rather than too few digits, since the

unbiased statistical summarization of data requires the extra digits. In fact, in internal and interlaboratory performance studies initiated by staff of the MOE Quality Management Office of the Laboratory Services Branch, MOE, it has been observed that laboratories which report extra digits tend to demonstrate better control of determinate error (bias).

The data user (client) must always be reminded that estimates of Sw are statistically predicted to vary within a factor of about 1.5 to 2 (about 95 to 99% confidence interval depending on d.f.). Therefore any values calculated from Sw (such as CD, DL, QL, etc.) will vary by that much. It strikes me as odd that statisticians are so specific about the confidence level, given such uncertainty in the estimate of Sw. Student's t-factors may be known to 4 significant figures (e.g., 2.998), but Sw certainly is not.

## CONCLUSIONS

As long as method blank and baseline estimates are reasonably well controlled, and other sample matrix effects are negligible, the concept of Method Detection Capability (MDC) provides a useful tool for determining whether analyte might be present or absent. The tool works particularly well when we already 'know' which case is likely true, since we must choose the opposite as our 'null' hypothesis. Use of this criterion to support the censoring of low-level data is statistically incorrect. One approach to data interpretation would use multiples of MDC to express the increasing quantitative validity of low-level measurements. Given that low results (including those below CD) should be reported, there are four levels at which we can note a change in the interpretation of the result. Thus:

| | |
|---|---|
| Measurable: | above W (Sw/2 to Sw) suggests 'presence' of analyte (alpha risk < 30%); |
| Present: | above CD (5 W or 3 Sw) affirms 'presence' of analyte (alpha risk < 1%); |
| Semi-quantitative: | above DL (10 W or 6 Sw) confidence interval <50%; |
| Quantitative: | above QL (20 W or 12 Sw) i.e. confidence interval <25%. |

Since Sw estimates vary somewhat from analyst to analyst, and since ratios of two Sw estimates are rarely significant till they exceed a factor of two, W values are preferred as a more stable basis for the reference points above. The comparability of data among laboratories, or the lack of it, is not a proper basis for deciding detectability. So-called PQL's may have a place (like the MISA RMDL) for setting an analytical performance criteria for laboratories wishing to support particular regulatory programs. But none of these terms (except W) should be automatically promoted as a reporting limit. When the analyte is suspected to be 'absent' one must report the measurement R to defend the statement that analyte concentration is lower than R + CD. Statements such as 'non-detected' (ND) or 'trace' have developed indeterminate meanings and should be avoided.

The concept of a sample-specific SDC (sample detection capability) must be retained for use by those who can estimate the impact on measurement variability of correcting for bias induced by other sample constituents. But the term SDC should not be used for situations when measurement was inhibited by matrix effects. The symbol ‘less-than’ (<) should be reserved for this case (since usually no measurement was performed and the reported value is simply an analysts estimate of the amount that would have had to be present). The < value would include the effects of dilution factors or other analytical considerations.

CD, DL, and QL correspond directly with Keith's recent proposals for MDL, RDL, and RQL. If CD is unacceptable, the term MDC is preferred since MDL has a specific definition, and has been presented as a CD rather than a DL concept, and is therefore confusing. In any event any statement of method detection capability (MDC) should be accompanied by a description of the data base upon which it is derived (instrument or total method?), (standards, or simple\complex sample matrices?). Efforts should be initiated to eliminate all other terms because of their inconsistent use of the word 'detection'.

CD, DL, and QL have value for data reporting, qualification and interpretation purposes. The implementation of these criteria, and the coincident requirement either to report low-level data, or to properly state the probable upper limit for sample content given the result R, will resolve many of the data quality and data interpretation problems associated with the 'censoring' of such data, and will promote more defensible practices for data reporting, qualification and interpretation. Incidently, proper consideration of low-level measurements will improve the use of baseline, blank, and other related correction factors, and will lead to a general improvement in inter-laboratory data comparability.

**ACKNOWLEDGEMENTS**

I wish to recognize the ongoing input of many of the scientific staff of MOE Laboratory Staff over the past many years. This is not a topic that is readily understood, so many analysts prefer to use the terminology without appreciating the principles. These extremely complex and interrelated issues of measurement, detection, data reportability, and data interpretation have generated much discussion. The outcome must reflect all aspects including analytical, statistical, and program objectives. Hopefully this document addresses the concerns of all in a holistic fashion, and will provide the basis for broader understanding of the topic.

## REFERENCES


1. Wilson, A.L.; "The Performance Characteristics of Analytical Methods"; Talanta, 1970, 17, pp.21 (part 1) and pp.31 (part 2).

2. Currie, L.A. (editor); "Detection in Analytical Chemistry, Importance, Theory, and Practice", ACS Symposium Series 361, ACS, Washington, DC, 1988.
   This text includes the following Panel Discussion:
   Brossman, M.W., et al.; "Reporting Low-Level Data for Computerized Data Bases" including King, D.E.; "A Rethink of the Factors Involved in Reporting Results Below the Method Detection Limit", pp 319.

3. Keith, L.H.; "Environmental Sampling and Analysis - A Practical Guide", Lewis Publishers, 1991.

4. Hunt, D.T.E and Wilson, A.L; "The Chemical Analysis of Water", 2nd Ed., pg 298, Royal Society of Chemistry, 1986.

5. ASTM D-19 Committee D4210-83; "Intralaboratory Quality Control and a Discussion on Reporting Low-Level Data, A Standard Practice", 1983.

6. King, D.E.; Environment Ontario, Laboratory Services Branch Reports "Quality Assurance Policy and Guidelines 1986" "Code of Practice for Environmental Laboratories", Apr. 1989 (ISBN 0-7729-5874-2).

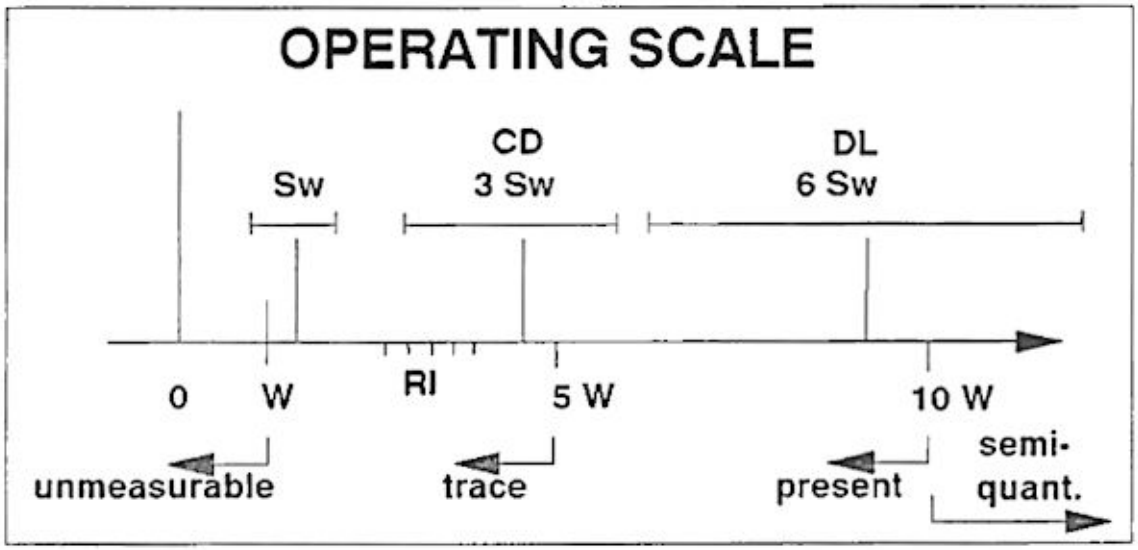
7. Ontario Regulation 695/88 as amended to Regulation 533/89 under the Environmental Protection Act; "Effluent Monitoring - General".

8. Crawford, G.; "Estimation of Analytical Method Detection Limits (MDL)", Environment Ontario MISA Reports, revised June 1991, (ISBN 0-7729-8817-X).

9. Federal Register, Vol. 49, No. 209, (Oct. 26, 1984), Appendix B to Part 136, Revision 1.11.

10. Crawford, G. and King, D.E.; "Protocol for the Reporting of Analytical Data", Oct. 1991, Environment Ontario MISA Report (Draft), (ISBN 0-7729-8971-0).

**Figure 1**




RI = INSTRUMENTAL READING INCREMENT

W = MAXIMUM RECOMMENDED REPORTING INCREMENT

W IS AN INTERVAL OF 1, 2, 5, 10, 20, 50, ... RI

BASED ON THE METHOD VARIABILITY Sw

CHOSEN TO BE JUST SMALLER THAN Sw

## CD, DL, and QL

CD = Criterion for detection = 3 Sw
DL = Detection Level = 6 Sw
QL = Quantitation Level = 12 Sw

SINCE ESTIMATES OF Sw VARY WITHIN A FACTOR OF 2

CD, DL, and QL ARE NOT FIXED POINTS

ESTIMATES OF Sw VARY DEPENDING ON:

operating range period of time
amount & quality of data
analyst skill, laboratory

**Figure 2**

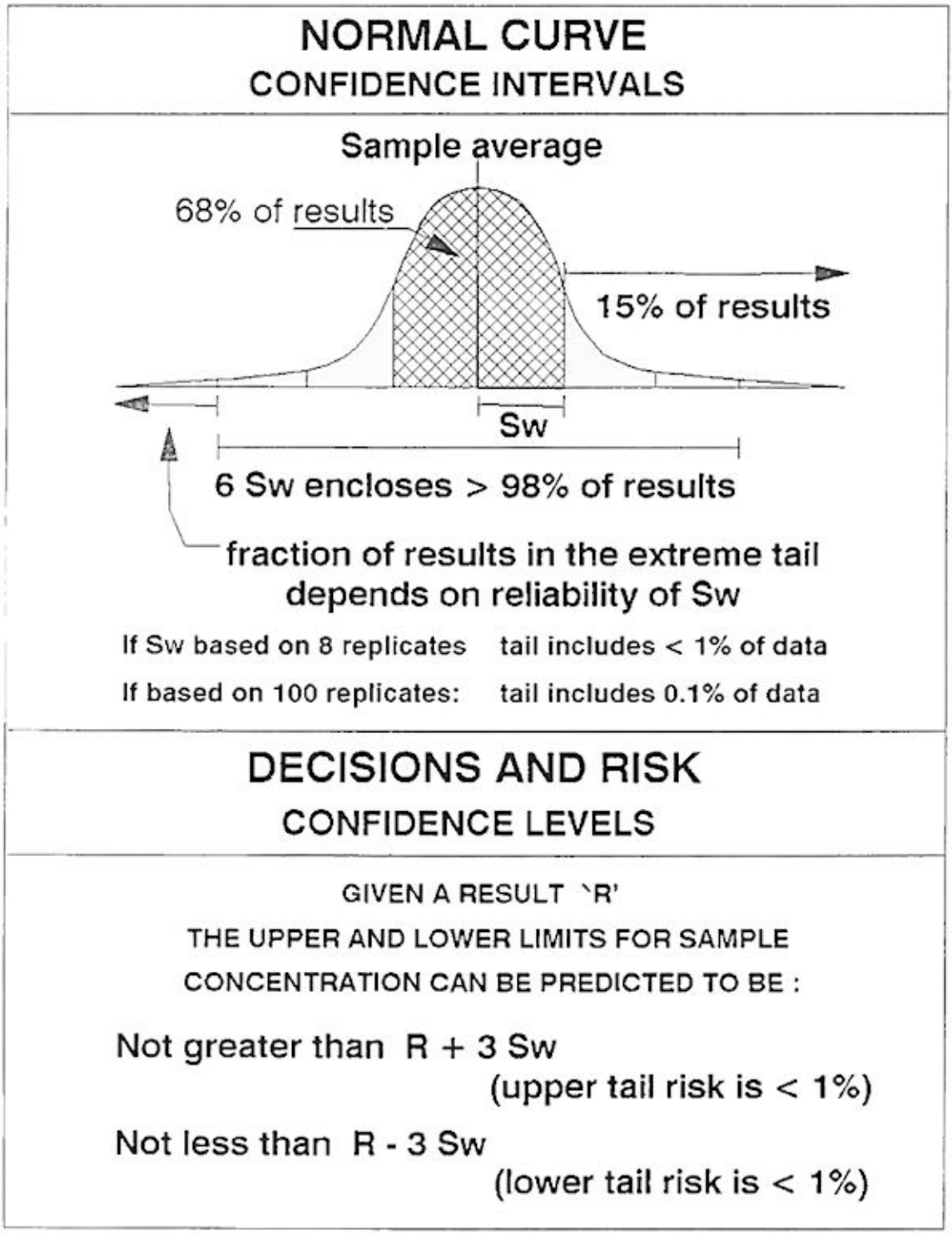

NORMAL CURVE
CONFIDENCE INTERVALS
Sample average
68% of results
15% of results
Sw
6 Sw encloses > 98% of results
fraction of results in the extreme tail
depends on reliability of Sw
If Sw based on 8 replicates tail includes < 1% of data
If based on 100 replicates: tail includes 0.1% of data
DECISIONS AND RISK
CONFIDENCE LEVELS
GIVEN A RESULT `R'
THE UPPER AND LOWER LIMITS FOR SAMPLE
CONCENTRATION CAN BE PREDICTED TO BE :
Not greater than R + 3 Sw
(upper tail risk is < 1%)
Not less than R - 3 Sw
(lower tail risk is < 1%)

**Figure 3**

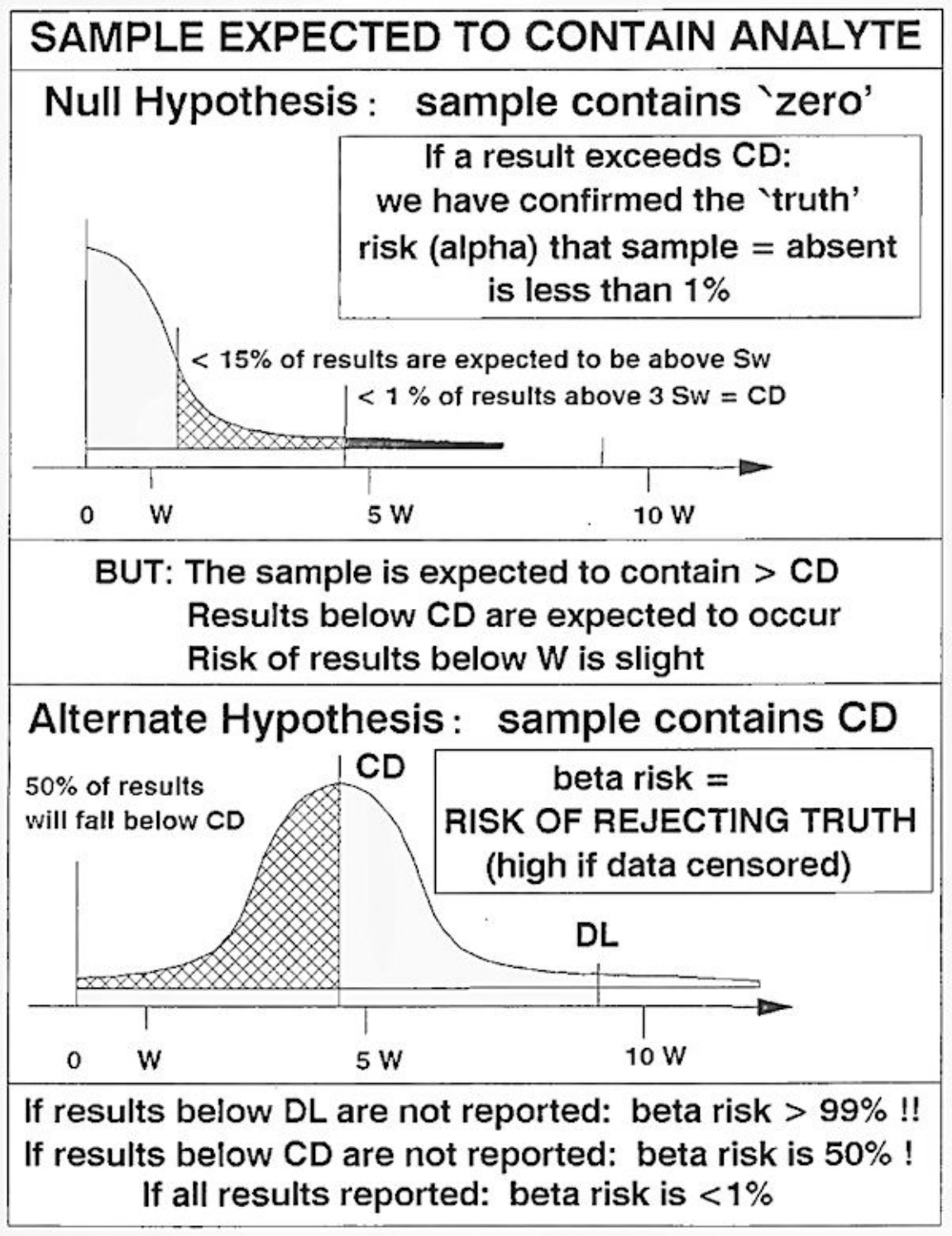
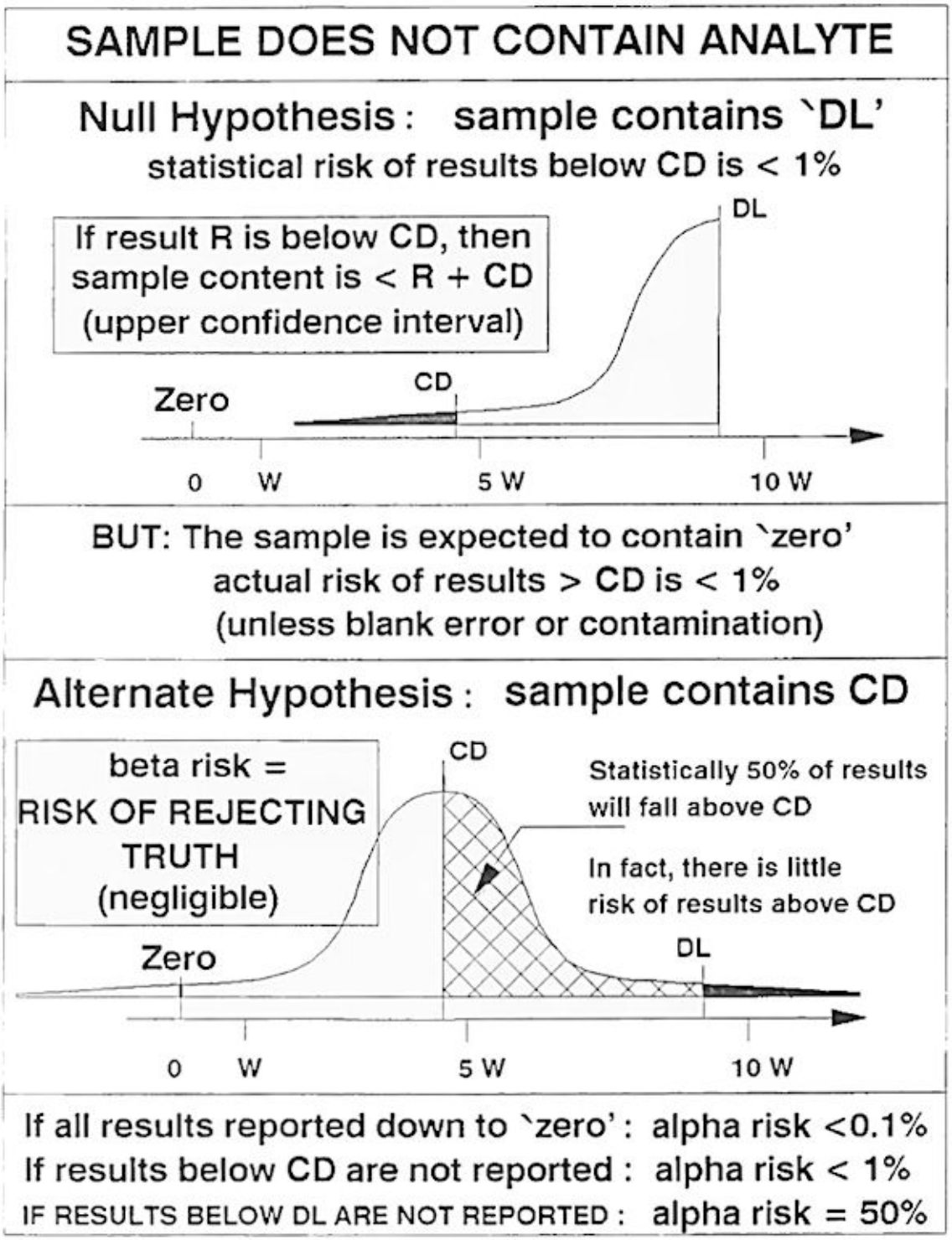

SAMPLE EXPECTED TO CONTAIN ANALYTE
Null Hypothesis : sample contains `zero'
If a result exceeds CD:
we have confirmed the `truth'
risk (alpha) that sample = absent
is less than 1%
< 15% of results are expected to be above Sw
< 1 % of results above 3 Sw = CD
0
W
5 W
10 W
BUT: The sample is expected to contain > CD
Results below CD are expected to occur
Risk of results below W is slight
Alternate Hypothesis : sample contains CD
50% of results
will fall below CD
CD
beta risk =
RISK OF REJECTING TRUTH
(high if data censored)
DL
0
W
5 W
10 W
If results below DL are not reported: beta risk > 99% !!
If results below CD are not reported: beta risk is 50% !
If all results reported: beta risk is <1%

**Figure 4**


SAMPLE DOES NOT CONTAIN ANALYTE
Null Hypothesis : sample contains `DL'
statistical risk of results below CD is < 1%
If result R is below CD, then sample content is < R + CD (upper confidence interval)
DL
CD
Zero
0
W
5 W
10 W
BUT: The sample is expected to contain `zero'
actual risk of results > CD is < 1%
(unless blank error or contamination)
Alternate Hypothesis : sample contains CD
beta risk = RISK OF REJECTING TRUTH (negligible)
CD
Statistically 50% of results will fall above CD
In fact, there is little risk of results above CD
Zero
DL
0
W
5 W
10 W
If all results reported down to `zero' : alpha risk <0.1%
If results below CD are not reported : alpha risk < 1%
IF RESULTS BELOW DL ARE NOT REPORTED : alpha risk = 50%