

**Statistical Analysis of the SSD Approach for
Development of Canadian Water Quality
Guidelines
Project # 354-2005**

Prepared by:
Zajdlik & Associates Inc.
July 21st, 2005
Last Updated:
September 23rd, 2005

PN 1414

This report was prepared by Zajdlik & Associates Inc., under contract to the Canadian Council of Ministers of the Environment (CCME). It contains information which has been prepared for, but not approved by, CCME. CCME is not responsible for the accuracy of the information contained herein and does not warrant, or necessarily share or affirm, in any way, any opinions expressed therein.

Table of Contents

1	Introduction.....	1
2	Data Review.....	4
2.1	Data Issues	4
2.1.1	Multiple Values from the Same Species.....	4
2.1.2	Defining “Chronic”.....	4
2.1.3	Endpoints to Use.....	4
2.1.4	Range of Concentrations.....	5
2.2	Data Sets	6
2.3	Methods.....	8
2.3.1	Parameter and Confidence Limit Estimation.....	8
2.3.2	Effect of Sample Sizes.....	9
2.4	Results.....	10
2.4.1	2,4-D	10
2.4.2	Atrazine.....	12
2.4.3	Copper.....	13
2.4.4	Nitrate	15
2.4.5	Uranium	16
2.4.6	Zinc	17
2.4.7	Estimation Results Summary.....	18
2.4.8	Exploration of Sample Size Effect Results.....	19
2.4.8.1	Weibull Distribution	19
2.4.8.2	Lognormal Distribution	21
2.4.9	Oscillations in HC ₅ Estimates.....	23
3	Conclusions.....	24
3.1	Treatment of Replicates	24
3.2	Choosing a Distribution	24
3.2.1	On Estimation Methods	25
3.2.1.1	Parametric versus Nonparametric Methods.....	25
3.2.1.2	Choice of Parametric Methods.....	26
3.2.1.3	Goodness of Fit Tests.....	26
3.2.2	Treatment of Replicate Species Data.....	26
3.3	The SSD Approach and Small Datasets.....	27
3.4	Problems with Current Implementations of the SSD Approach to Estimating Guidelines	28
3.5	Recommendations.....	30
4	Citations	32
5	Appendix 1: Quantile-Quantile Plots.....	34
5.1	2,4-D	35
5.2	Copper.....	36
5.3	Uranium	37
5.4	Zinc	38

List of Tables

Table 1: Table of Acronyms.....	i
Table 2: Summary of Data Sets Used to Estimate HC ₅ 's.....	6
Table 3: Parameter Estimates for Weibull Distribution fit to the 2,4-D SSD.....	11
Table 4: Parameter Estimates for Weibull Distribution fit to the Uranium SSD.....	16
Table 5: Parameter Estimates for Weibull Distribution fit to the Zinc SSD.....	18
Table 6: Summary of HC ₅ Estimates and Lower 95% Confidence Limits.....	18
Table 7: Oscillations in HC ₅ Estimates for the Zn Dataset.....	23
Table 8: Summary of Treatment of Replicate Species Data.....	26

List of Figures

Figure 1: Exploratory Graphics for 2,4-D (µg/l) SSD.....	10
Figure 2: Exploratory Graphics for Atrazine (µg /l) SSD.....	12
Figure 3: Exploratory Graphics for Copper (µg /l) SSD.....	13
Figure 4: Exploratory Graphics for Ln(Copper) (ln(µg /l)) SSD.....	14
Figure 5: Exploratory Graphics for Nitrate (µg active ingredient /l) SSD.....	15
Figure 6: Exploratory Graphics for Uranium (µg active ingredient /l) SSD.....	16
Figure 7: Exploratory Graphics for Zinc (µg /l) SSD.....	17
Figure 8: Sample Size Simulation Results for HC ₅ – Weibull Distribution, Zinc Dataset	19
Figure 9: Sample Size Simulation Results for Lower Confidence Limits– Weibull Distribution, Zinc Dataset.....	20
Figure 10: Sample Size Simulation Results for HC ₅ – Log Normal Distribution, Copper Dataset.....	21
Figure 11: Sample Size Simulation Results for HC ₅ – Log Normal Distribution, Copper Dataset.....	22

Table 1: Table of Acronyms

Acronym	Definition
ECDF	empirical cumulative density function
EPDF	empirical probability density function
IC25	concentration resulting in 25% inhibition in a response relative to a control response.
MLE	maximum likelihood estimation
NEC	no effect concentration
NOEC	no observed effect concentration
SSD	species sensitivity distribution

1 Introduction

Species sensitivity distribution approaches to deriving water quality criteria have been used in one form or another for decades. The earliest methods of setting a guideline were based on examining the available toxicity test results for a contaminant and using expert judgment to derive a guideline. Expert judgment usually defaulted to selecting the most sensitive toxicity test result. This *ad hoc* examination of available data was in fact an unwitting “species sensitivity distribution” approach to deriving water quality criteria since a set of toxicity test results from different species was used to derive a criterion to protect other species. This *ad hoc* approach was subject to criticism on the basis of:

- subjectivity in how data were used;
- uncertainty regarding the level of ecosystem-wide protection afforded; and,
- strong reliance on the most sensitive species tested, etc.

The *ad hoc* approach to developing water quality criteria was improved by creating procedures for developing water quality criteria. These procedures could be referred to before derivation of a guideline, thereby forestalling the criticism of subjectivity in how data were used. In the ensuing decades, the procedures for development of water quality criteria have evolved to address issues such as:

- the percent of species in an ecosystem that must be protected;
- definition of the term “protected”;
- the use of safety factors;
- the minimum number of species, taxonomic diversity and number of observations required;
- the use and relative merits of “acute” and “chronic” toxicity test results; and,
- trophic diversity in the species sensitivity distribution to afford ecosystem-relevant protection, etc.

Each of these issues has been treated differently in various jurisdictions due to differing beliefs regarding the scientific literature and the jurisdiction-specific balance between science and environmental policy.

Current approaches embrace the concept of a distribution in sensitivity of species to a contaminant. This parallels the tolerance distribution concept embedded in estimation of a toxicity test endpoint for a single experiment, where individuals have differing levels of sensitivity. In the aggregate sense, the cumulative response (if mortality could be cumulated over one organism) of the individuals exposed is described by a cumulative normal or Gaussian distribution. This distribution of individual tolerances describes the sensitivity of the sample of exposed organisms to the contaminant under the prescribed conditions. The distribution is used to make inferences regarding the population of potentially exposed organisms from the same species.

One endpoint that is commonly estimated when the cumulative distribution represents mortality is the LC50, or concentration that results in 50% mortality. The utility of this endpoint has gradually decreased over the last few decades as awareness of environmental effects has led to interest in lower levels of mortality and/or estimation of smaller fractions of toxicity test organisms exhibiting non-lethal responses. Commonly estimated endpoints for single toxicity tests include IC25s and sometimes IC10's. With respect to species tolerance or sensitivity distributions, interest centers on even lower percentiles such as the 5th percentile. The estimation of a toxicity test endpoint and an endpoint from a species sensitivity distribution share in common:

- methods for model selection;
- to a large extent, the suite of potentially useful models;
- optimal mathematical/statistical methods for estimating endpoints, i.e. the underlying algorithm;
- sensitivity of endpoints to model selection; and,
- choice of relevant percentile of organisms (IC25 versus IC10 or LC20 versus LC50, etc.).

Estimation of a toxicity test endpoint and an endpoint from a species sensitivity distribution differ in that:

- the treatment of multiple observations for a single species is not relevant when estimating a toxicity test endpoint; and,
- there is more controversy regarding the percentile to estimate in the case of multiple species tolerance or sensitivity distributions.

Statistically, the problem for either the single species or multiple species case is seemingly straightforward; estimate a quantile and confidence interval from a sample distribution. Aside from the ecological problems (relevance of species collected to a given ecosystem, relevance of "acute" versus "chronic measurements", problem of coverage of trophic levels and critical trophic levels, keystone species, etc.) and policy problems (choice of quantile and by extension, degree of environmental protection to estimate, degree of precision required for quantile estimate etc.) the following statistical issues arise.

Choosing the General Estimation Approach: In order to estimate a percentile from a data set one need only rank the data and choose the observation corresponding to the desired percentile or use an interpolation method when the available observation ranks do not coincide with the desired percentile. This approach is extremely sensitive to the sample size and the toxicity test results in the vicinity of the desired percentile. However the most severe condemnation of this method is that any guideline derived using such an approach can only be one of the observed values. Another approach that does not suffer from this latter shortcoming is to model the observed data in the same way that the tolerance distribution generated by a single toxicity test is modelled. The parameters of the model are estimated and then desired percentile is predicted.

Choosing a “Model” or Tolerance Distribution: When estimating a “middle” percentile such as the median, the choice of model (tolerance distribution) will often not substantively affect the endpoint estimated. For example, LC50s estimated after assuming logistic, normal, Weibull or Gompertz tolerance distributions are virtually identical. However, when estimating an extreme percentile such as 5 or 95%, the choice of model may greatly affect the estimated endpoint. Therefore it is critical that objective tools be developed and applied for choosing the most appropriate model. Note that the phrase “correct model” is not used, in keeping with the statement made by a famous statistician, George Box: “All models are incorrect but some are useful.” We do not necessarily believe that one model is correct and all others are incorrect but rather that one model is “less incorrect” than another.

Sample Sizes: The small-sample behaviour of extreme percentiles may vary from model to model. Therefore not only should a model be the most appropriate but it should also have desirable small-sample behaviour. Small-sample behaviour is largely concerned with the convergence of variance terms to asymptotic results.

Other issues not covered but still relevant to using the SSD approach to derivation of water quality guidelines are:

- availability of suitable software and ease of estimation by scientific consultants;
- selection of optimal numerical algorithms for estimation;
- congruence with other methods; and.
- congruence with other jurisdictions, etc.

This report examines several data sets provided by the Ontario Ministry of Environment in the context of estimating CCME-endorsed water quality guidelines. Issues that are addressed are:

- Select candidate models for possible use as descriptors of SSDs using the available data sets.
- Examine the small-sample properties for two of the selected distributions.
- Evaluate oscillations in HC_5 values with changes in data set.
- Discuss the ease of use of these models given the target audience (scientific consultants).
- Provide a review of the most suitable goodness of fit test for each posited distribution.
- Comment on the SSD approach and small datasets.

2 Data Review

2.1 Data Issues

2.1.1 Multiple Values from the Same Species

One project deliverable was to assess the use of the median or geometric mean to represent multiple endpoints from the same species. Here replicate refers to replication at the level of the estimated endpoint, say for example an EC20. Note that information on the biological response measured was not usually available. Therefore it is almost certain that biological responses were incorrectly collapsed over toxicity test endpoints.

2.1.2 Defining “Chronic”

The following “working definition” of the term “chronic” was provided by T. Fletcher (MOE, pers. comm.)

- Fish - 96 h or longer
- Long-lived invertebrate (mollusks and decapods)- 96-h or longer
- Short-lived invertebrate - 48-h or longer
- Algae - by definition all responses are chronic

When a test duration was reported as $> X$, the value of X was used in assignment to the “chronic” or “acute” groups and in subsequent calculations.

2.1.3 Endpoints to Use

The following endpoints appear in the SSD data sets:

- LC50 – lethal concentration for 50% of organisms
- EC50 – concentration causing a response in 50% of organisms
- LOEC – lowest observed effect concentration
- TLm – concentration that 50% of organisms can tolerate – should be equivalent to LC50

The following endpoints were excluded from the SSD data sets (T. Fletcher, pers. comm.):

- EC100 – concentration causing a response in 100% of organisms;
- NOEC – no observed effect concentration;
- LC01 - lethal concentration for 1% of organisms;
- records where no endpoint was entered; and
- records where the endpoint was entered as “NR”.

2.1.4 Range of Concentrations

When a range of concentrations was presented for a single record, the lowest value was retained.

2.2 Data Sets

Table 2: Summary of Data Sets Used to Estimate HC5's

Contaminant	Data Set Type ¹	Sample Size	Effect of Averaging on Data Set Size	Source	Comments
2,4-D	unknown	51	from n = 116 to n = 51 or 56% reduction	Acquire download provided by T. Fletcher.	
ammonium nitrate	chronic	3		Worksheet "ammonium nitrate" in file "ammonium nitrate chronic data(reformatted)1.xls."	<ul style="list-style-type: none"> No modelling attempted.
atrazine	unknown		from n = 367 to n = 119 or 68% reduction	Acquire download provided by T. Fletcher.	<ul style="list-style-type: none"> Assumed "dph" under test duration was converted to hours
copper	chronic		from n = 940 to n = 125 or 87% reduction	Acquire download provided by T. Fletcher.	<ul style="list-style-type: none"> Organisms labeled as "aquatic community" deleted from data set as (although intriguing) they do not represent a species. Removed record for <i>Ischnochiton hakodadensis</i> as this is a marine species. Asterisks attached to endpoints were removed. Euryhaline copepods retained.
formaldehyde	chronic	3	from n = 3 to n = 3 or 0% reduction	Worksheet "Formaldehyde - Chronic" in file "Tableaux	<ul style="list-style-type: none"> No modelling attempted.

¹ Most data sets received from the MOE were designated as "acute" or "chronic" although some were not. Despite the label, the criteria discussed in section 2.1.2 were applied to all datasets.

Contaminant	Data Set Type ¹	Sample Size	Effect of Averaging on Data Set Size	Source	Comments
				formaldéhyde eau douce 1999(reformatted)4.xls”	
nitrate	chronic	32	from n = 59 to n = 32 or 46% reduction		<ul style="list-style-type: none"> • <i>Macrobrachium</i> sp. – a decapod labeled as “chronic”
uranium	chronic	12	from n = 20 to n = 12 or 40% reduction	Worksheet “chronic” in file “U data set for B Zdadlik(reformatted)2.xls”	<ul style="list-style-type: none"> • Endpoint “MDEC” entered as “LOEC” following comment “MDEC (LOEC equivalent)” • 3 records with rank = “NA” were retained
zinc	unknown	101	from n = 435 to n = 101 or 76.8% reduction	Acquire download provided by T. Fletcher.	<ul style="list-style-type: none"> • Removed record for <i>Ischnochiton hakodadensis</i> as this is a marine species. Assumed rice (<i>Oryza sativa</i>) is an aquatic plant and retained it in the data set. • The euryhaline harpacticoid copepod <i>Nitocra spinipes</i> was retained. • Asterisks attached to endpoints were removed.

2.3 Methods

2.3.1 Parameter and Confidence Limit Estimation

Ten potential² SSD distributions were selected by matching exploratory graphics (presented herein) with known distributional shapes. The exploratory graphics include empirical probability density functions (EPDF) and empirical cumulative density functions (ECDF).

A probability density function describes the probability that a random variable falls between two specified values. The familiar bell-shaped normal distribution is an example of a probability density. If we do not know the statistical distribution of a data set we can generate a frequency or probability histogram (recall that a probability histogram is identical in shape to the frequency histogram but the ordinate or y-axis has units of probability rather than frequency). A probability histogram is an empirical description of the probability density function.

Similarly a probability distribution function or cumulative density function, describes the probability that a random variable is \leq some specified value. A familiar example is the bell shaped normal distribution. For the normal distribution, if the random variable is equal to 1.645, the probability of being less than 1.645 is 95%. If we cumulate the probabilities in the EPDF, we obtain the ECDF.

Once potential models were selected, the parameters were estimated using maximum likelihood methods. For all distributions other than the normal, the Anderson-Darling test statistic was used to assess goodness of fit as its known sensitivity to values in the tail of distribution (D'Agostino and Stephens, 1986) is in keeping with our interest in the lower tail of the SSDs. The Shapiro-Wilk test was used to assess goodness of fit to the normal distribution following D'Agostino and Wilks, (1986). P-values for the Anderson-Darling test statistic were obtained using Monte Carlo simulation. The p-values obtained from either the Shapiro-Wilks test or by Monte Carlo simulation were used in conjunction with quantile-quantile plots to choose an SSD.

Following selection of the SSD, the quantile corresponding to a cumulative probability of 5% (known in the ecotoxicological literature as the HC₅) was estimated following two general approaches.

1) Parametric Methods:

- All distributions except the normal - The HC₅ is estimated following the usual MLE parametric approach. The one-sided lower 95% confidence limit for the HC₅ is estimated using the distribution-free approach advocated by Stuart *et al*, (1999, section 19.40).

² Distributions considered include normal, lognormal, logistic, log-logistic, inverse Gaussian, gamma, Weibull, exponential, extreme value and Laplace.

- The normal distribution - The HC_5 is estimated using a prediction expectation approach following Han and Meeker (1995). The exact one-sided lower 95% confidence limit for the HC_5 is estimated using the non-central t-distribution approach (Johnson et al, 1994b)

2) Nonparametric Method: A nonparametric quantile estimate with standard error following Harrell and Davis, (1982). The one-sided lower 95% confidence limit for the HC_5 is estimated assuming asymptotic normality of the quantile.

2.3.2 Effect of Sample Sizes

The effect of sample sizes was evaluated using two data sets: 1) The Zn data following a Weibull distribution as discussed in section 2.4.6 and 2) the Cu data following a lognormal distribution as discussed in section 2.4.3.

Using the estimated parameters, random variables from the distribution were generated for samples size of 5, 10, 15 ... 50. The HC_5 was estimated from each of the 10 different samples. This procedure was repeated 100 times. From each vector of 100 HC_5 estimates, the lower one-sided 95% and 99% confidence limits were estimated following the method recommended by Hyndman and Fan (1996). Convergence plots illustrate how the change in sample size affects the stability of the HC_x .

2.4 Results

The empirical or observed probability density function (PDF) superimposed over the probability histogram and the cumulative density function are plotted for each dataset.

2.4.1 2,4-D

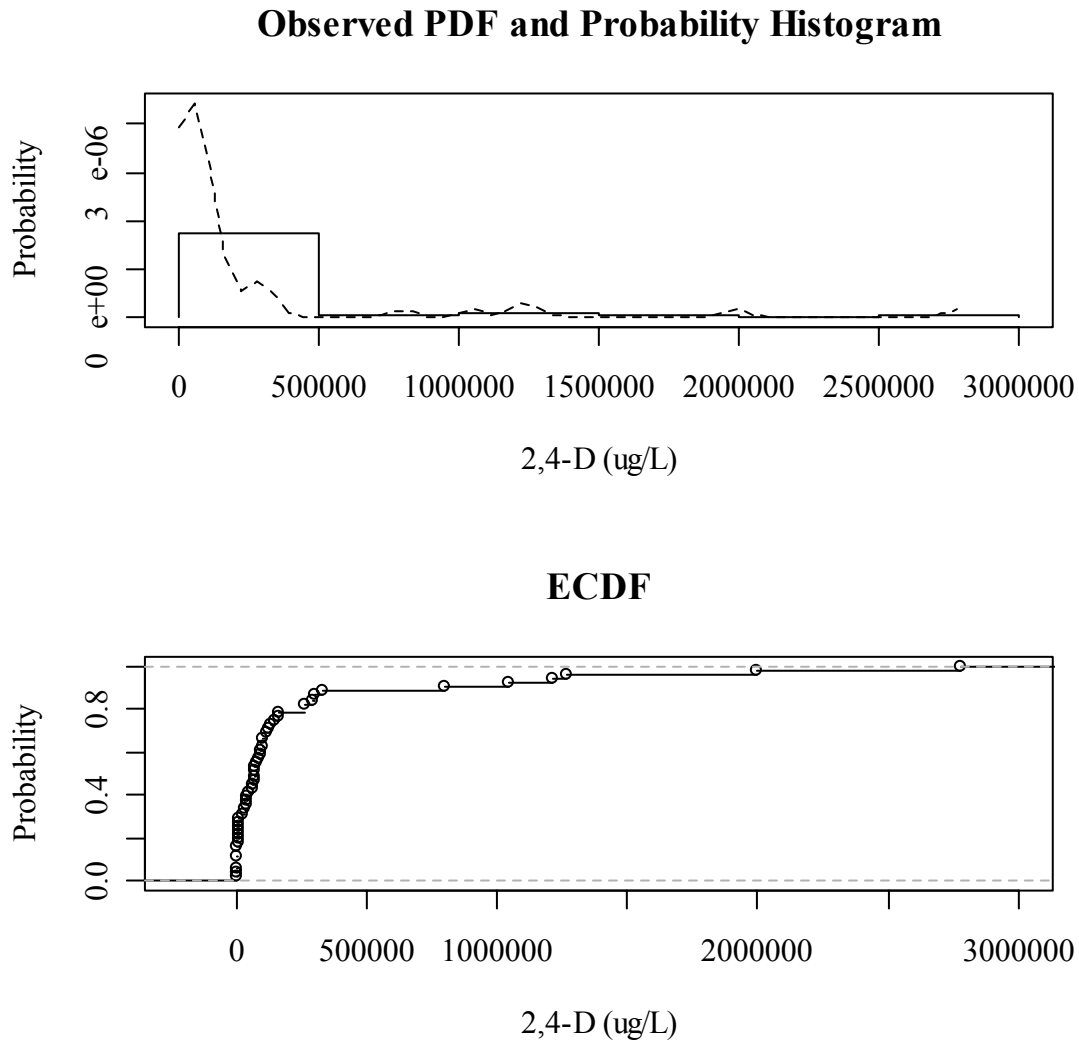


Figure 1: Exploratory Graphics for 2,4-D ($\mu\text{g/l}$) SSD

A variety of distributions were fit to the 2, 4-D data set. The Weibull distribution was selected with an Anderson-Darling test statistic = 0.6842 and a p-value = 0.07380 using 50,000 Monte Carlo simulations.

Table 3: Parameter Estimates for Weibull Distribution fit to the 2,4-D SSD

	Scale	Shape
Estimate	128728.8610	0.5294
Variance	1.2853D+009	0.003341

Using the parametric approach, the HC₅ is estimated as 471.1 µg/l with a lower one-sided 95% confidence interval of 106.5µg/l. Using the nonparametric approach, we obtain the values 1688.1980 and 1073.5 µg/l for the HC₅ and lower one-sided 95% confidence interval, respectively.

2.4.2 Atrazine

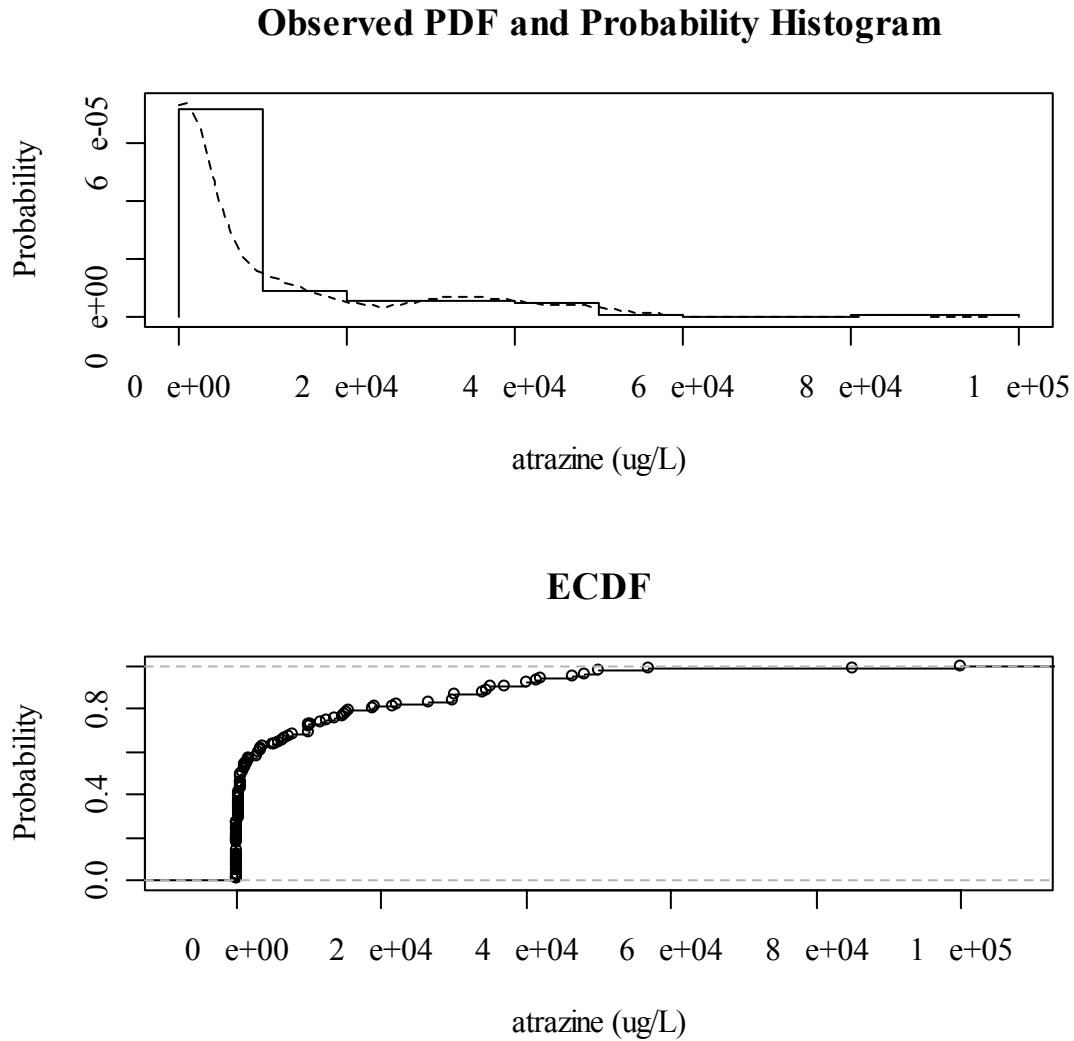


Figure 2: Exploratory Graphics for Atrazine ($\mu\text{g/l}$) SSD

None of the distributions listed in section 2.3 adequately described the data although the extreme value distribution came close with a Kolmogorov-Smirnov p -value³ of 0.0326. Note that the Kolmogorov-Smirnov test places less emphasis on the tails of the distribution. This test chosen for the atrazine test as quantile-quantile plots (not shown) indicated a poor fit between the 60th and 80th percentiles.

Using the nonparametric method we obtain estimates of 24.6 and 15.1 $\mu\text{g/l}$ for the HC_5 and lower one-sided 95% confidence interval, respectively.

³ Estimated using 50,000 Monte Carlo simulations.

2.4.3 Copper

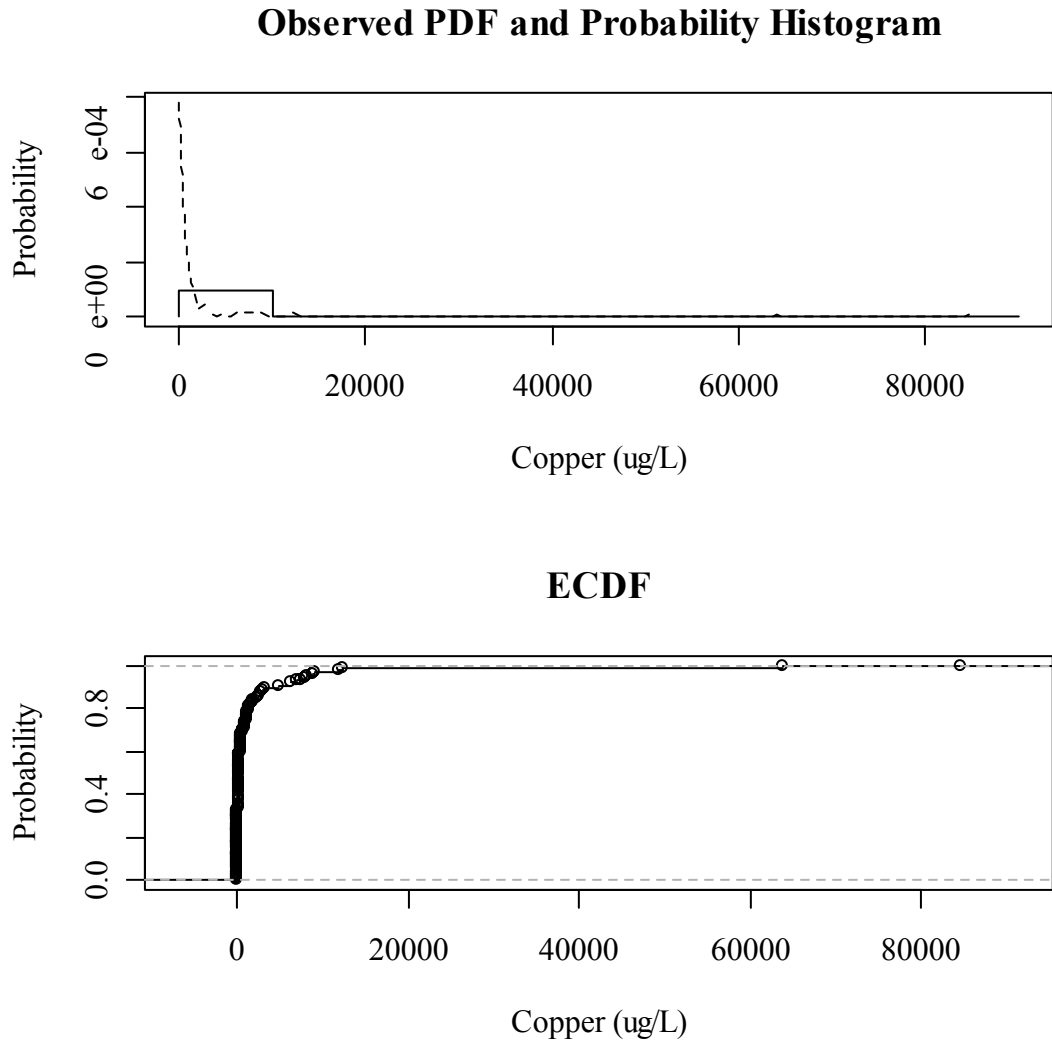


Figure 3: Exploratory Graphics for Copper ($\mu\text{g/l}$) SSD

Note the two extremely insensitive species. Given the graphic above, we attempt a logarithmic transformation.

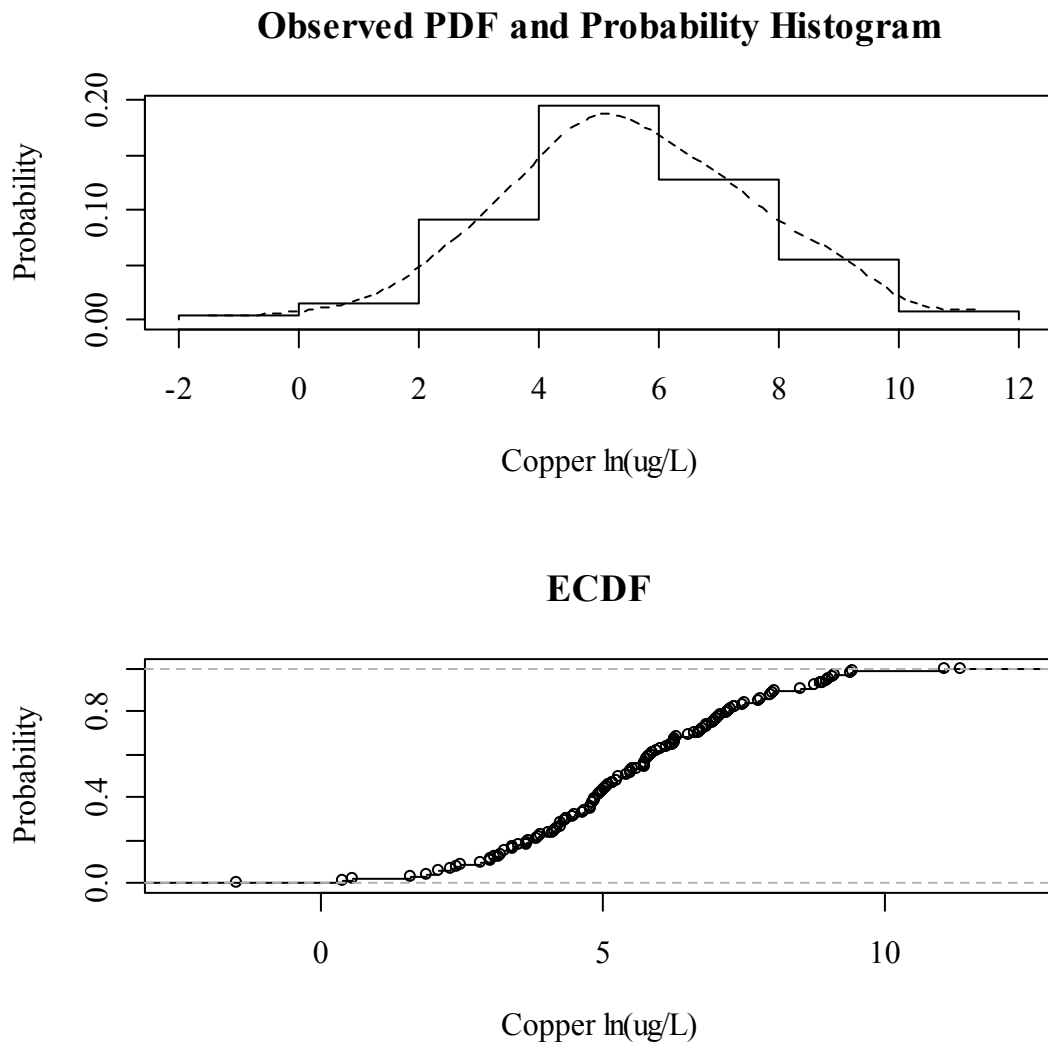


Figure 4: Exploratory Graphics for Ln(Copper) ($\ln(\mu\text{g/l})$) SSD

Using the methods described in section 2.3 we choose the log-normal as a useful descriptor of the copper SSD data set. The HC_5 is estimated as $6.5 \mu\text{g/l}$ with a lower one-sided 95% confidence interval of $4.0 \mu\text{g/l}$.

The nonparametric method produced estimates of 8.2 and $3.6 \mu\text{g/l}$, for the HC_5 and lower one-sided 95% confidence interval, respectively.

2.4.4 Nitrate

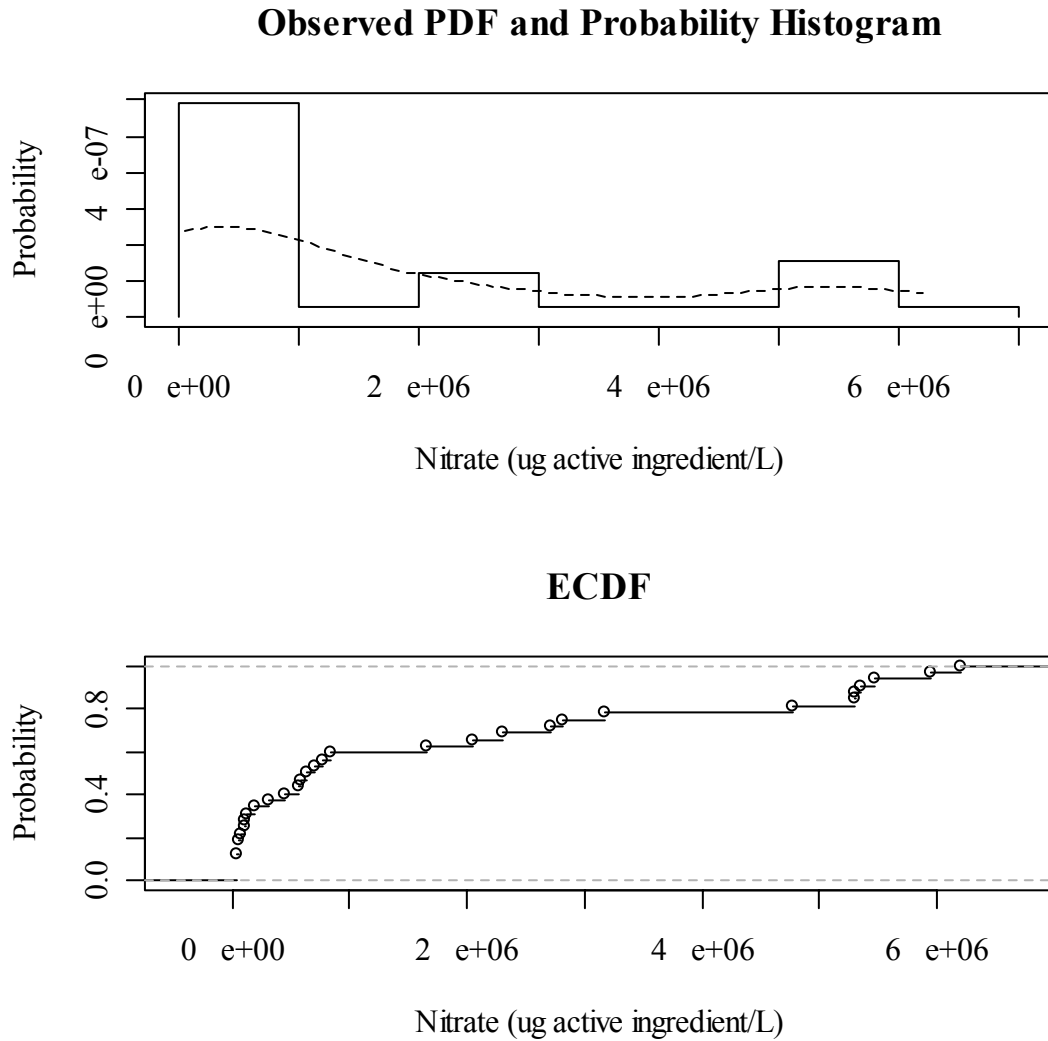


Figure 5: Exploratory Graphics for Nitrate (μg active ingredient /l) SSD

The distributions described in section 2.3 were used to describe the nitrate SSD data set without success. Consequently only the nonparametric method of Harrell and Davis (1982) are employed herein. Using this approach, the estimated HC_5 is 41,833.7 μg active ingredient /l with a one-sided lower 95% confidence interval of 37,964.6 μg active ingredient /l.

2.4.5 Uranium

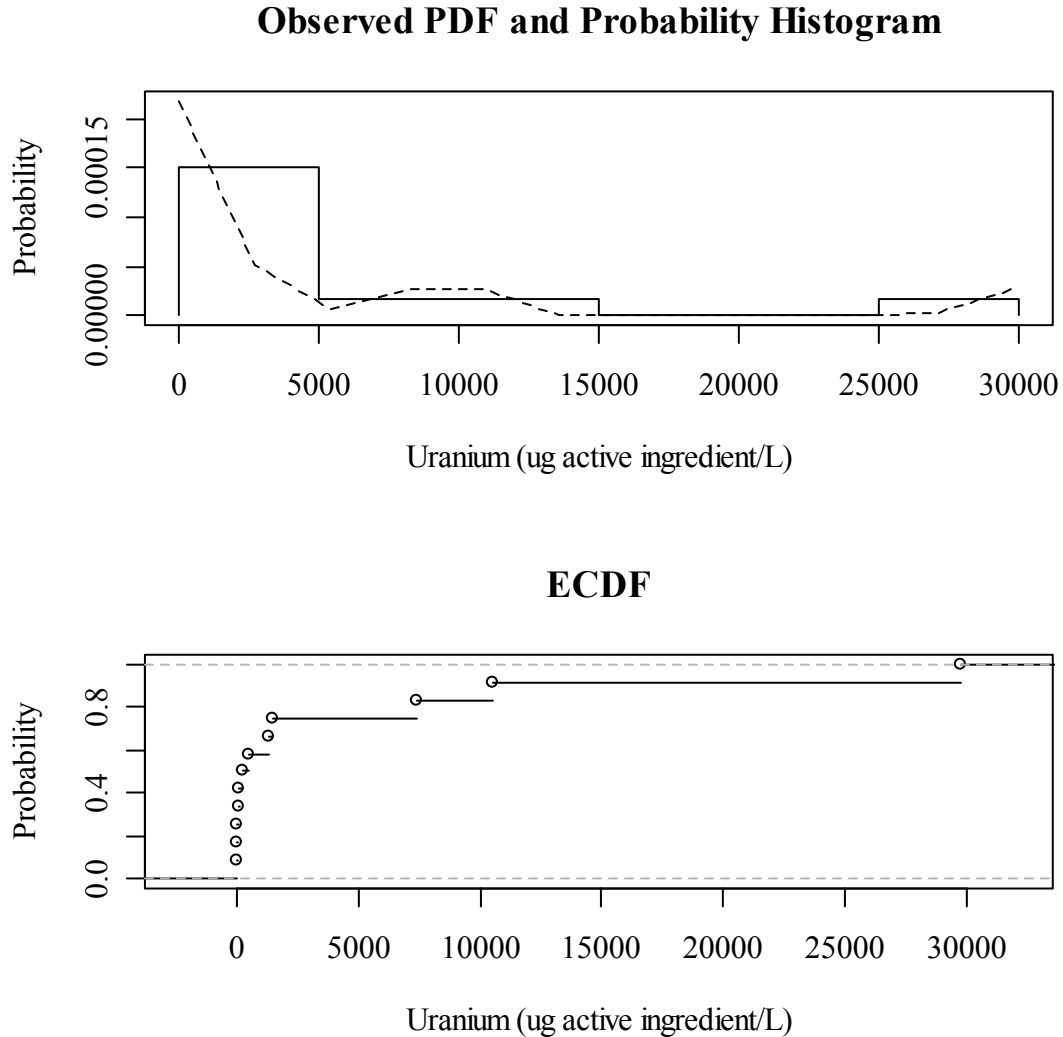


Figure 6: Exploratory Graphics for Uranium (μg active ingredient /l) SSD

A variety of distributions were fit to the uranium data set. The Weibull distribution was selected with an Anderson-Darling test statistic = 0.3240 and a p-value = 0.5517 using 50,000 Monte Carlo simulations.

Table 4: Parameter Estimates for Weibull Distribution fit to the Uranium SSD

	Scale	Shape
Estimate	1.4756	0.4233
Variance	1.1228	0.009077

Using the parametric approach, the HC₅ is estimated as 1.3 µg active ingredient/l with a lower one-sided 95% confidence interval of 7.60E-003 µg active ingredient/l. Using the nonparametric approach, we obtain the values 12 and 0.0 µg active ingredient/l for the HC₅ and lower one-sided 95% confidence interval, respectively.

A discussion of guidelines in the context of a lower 95% confidence limit around the HC₅ equal to 0 is provided in section 3.4, pg. 29.

2.4.6 Zinc

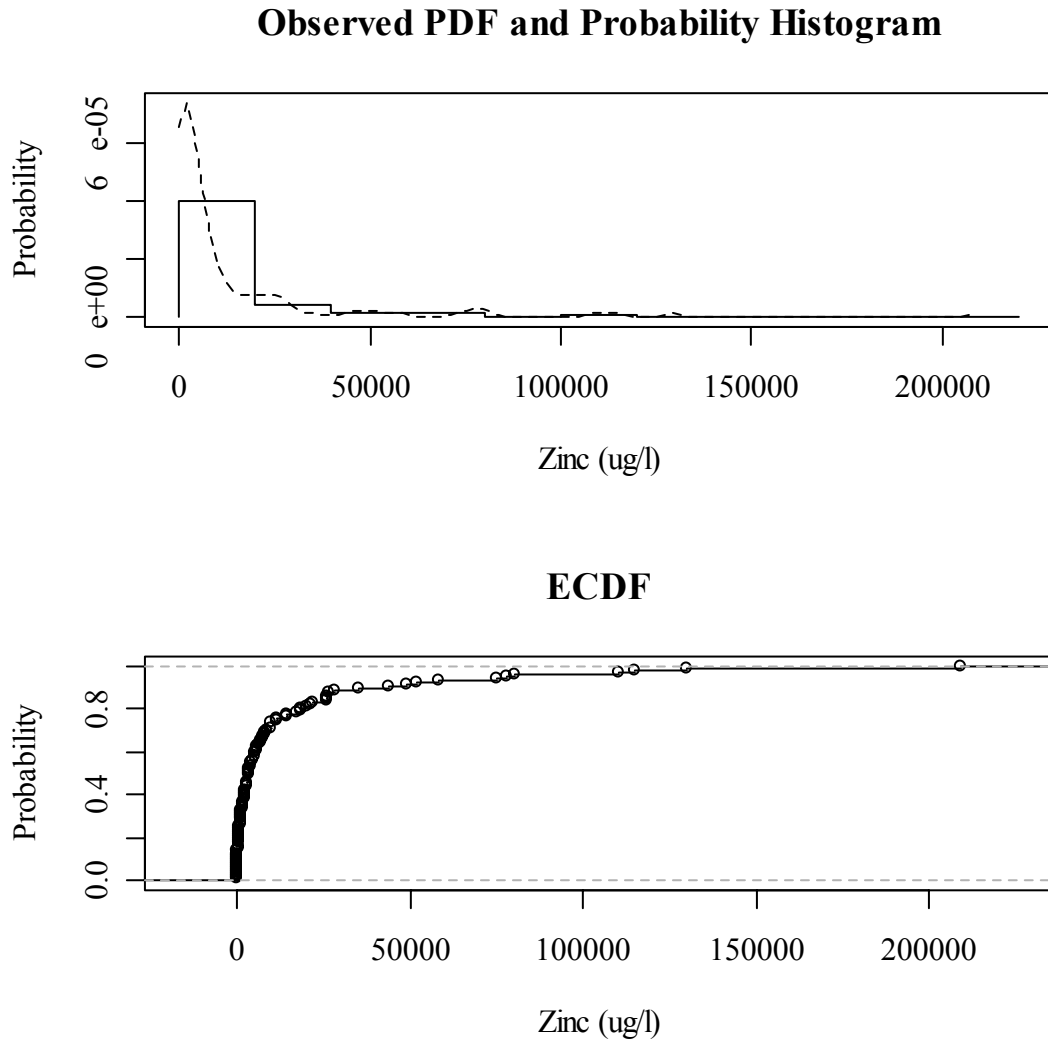


Figure 7: Exploratory Graphics for Zinc (µg /l) SSD

A variety of distributions were fit to the zinc data set. The Weibull distribution was selected with an Anderson-Darling test statistic = 0.3203 and a p-value = 0.5536 using 50,000 Monte Carlo simulations.

Table 5: Parameter Estimates for Weibull Distribution fit to the Zinc SSD

	Scale	Shape
Estimate	7922.6218	0.5219
Variance	2529410.04420	0.001640

Using the parametric approach, the HC₅ is estimated as 26.7 µg /l with a lower one-sided 95% confidence interval of 9.7 µg /l. Using the nonparametric approach, we obtain the values 40.9 and 2.4 µg /l for the HC₅ and lower one-sided 95% confidence interval, respectively.

2.4.7 Estimation Results Summary

Table 6: Summary of HC5 Estimates and Lower 95% Confidence Limits

Substance	Current CCME Guideline (µg/L)	Parametric		Non-Parametric	
		HC ₅	1 – sided 95% LCL	HC ₅	1 – sided 95% LCL
2,4-D (µg/l)	4.0	471.1	106.5	1,688.2	1,073.5
atrazine(µg/l)	1.8	NA	NA	24.6	15.1
copper (µg/l)	2.0, 3.0, 4.0 ⁴	6.5	4.0	8.2	3.6
nitrate (µg active ingredient /l)	13,000	NA	NA	41,833.7	37,964.6
uranium (µg active ingredient/l)	NA	1.3	7.6E-003	12	0.0
zinc (µg/l)	30	26.7	9.7	40.9	2.4

Comments on **Error! Reference source not found.**:

The nonparametric HC₅ estimates are always larger than their parametric counterparts. A lack of congruence between current CCME guidelines may reflect differences between the data sets used, rather than a methodological effect.

⁴ (For hardness ranges of 0-120, 120-180 and > 180 mg/l for hardness as $CaCo_3^{2-}$, respectively.)

2.4.8 Exploration of Sample Size Effect Results

2.4.8.1 Weibull Distribution

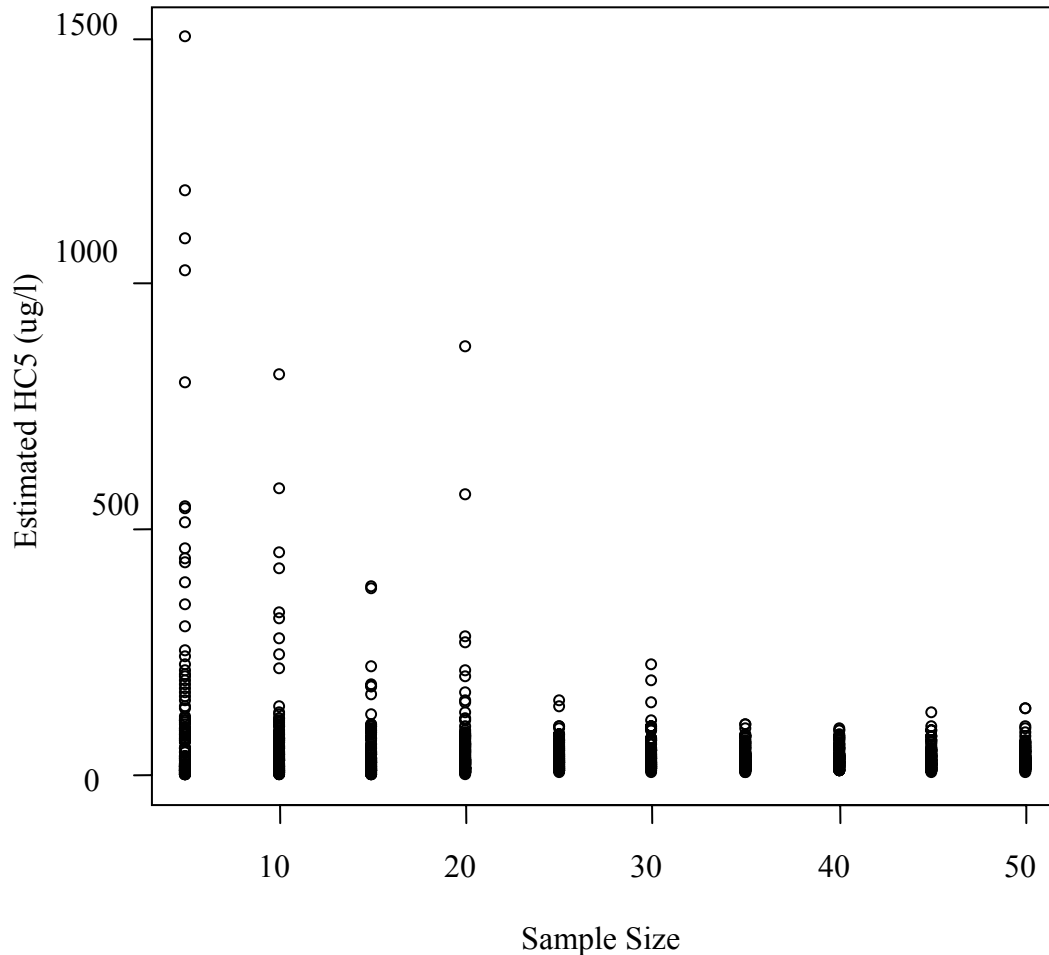


Figure 8: Sample Size Simulation Results for HC₅– Weibull Distribution, Zinc Dataset

In Figure 9, above we note that the variability in HC₅ is extremely high for sample sizes of 10 or less. Variability decreases with increasing sample size and is reasonably stable for sample sizes of 25 or more. The observation that the HC₅ stabilizes for sample sizes greater than 25 should not be over generalized.

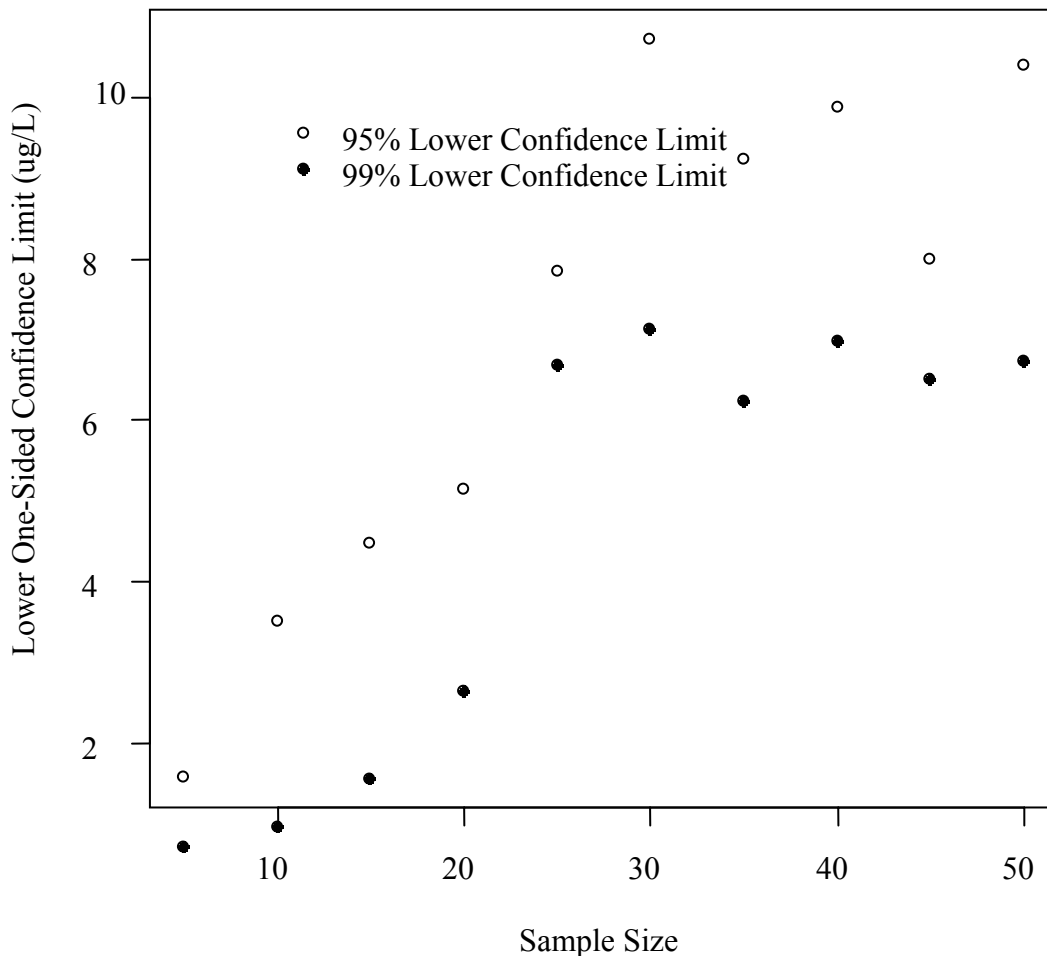


Figure 9: Sample Size Simulation Results for Lower Confidence Limits– Weibull Distribution, Zinc Dataset

In Figure 9, above we note that the lower one-sided 95% confidence for the HC_5 continues to increase over the range of sample sizes examined. The lower one-sided 99% confidence for the HC_5 seems to stabilize with a sample size greater than 25. Note that these results apply to a specific distribution, the Weibull, with a specific set of parameters. The observation that the lower one-sided 99% confidence limit stabilizes for sample sizes greater than 25 should not be over generalized.

Using the parametric approach with the original data set with sample size = 101, the HC₅ was estimated as 26.7 µg /l with a lower one-sided 95% confidence interval of 9.7 µg /l concurring with our simulation above.

2.4.8.2 Lognormal Distribution

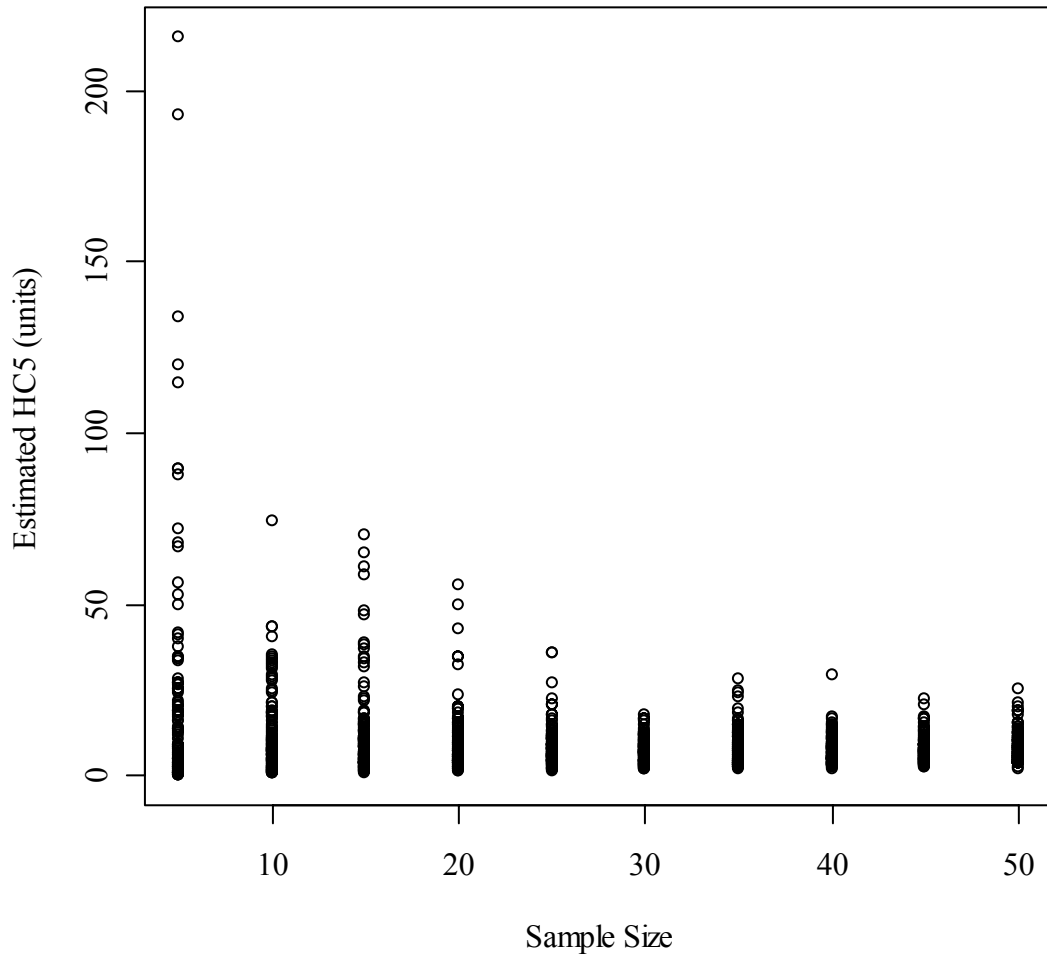


Figure 10: Sample Size Simulation Results for HC₅– Log Normal Distribution, Copper Dataset

In Figure 10 above, we note that the variability in HC₅ is extremely high for sample sizes of 20 or less. Variability decreases with increasing sample size and is reasonably stable for sample sizes of 25 or more. The observation that the HC₅ stabilizes for sample sizes greater than 25 should not be over generalized.

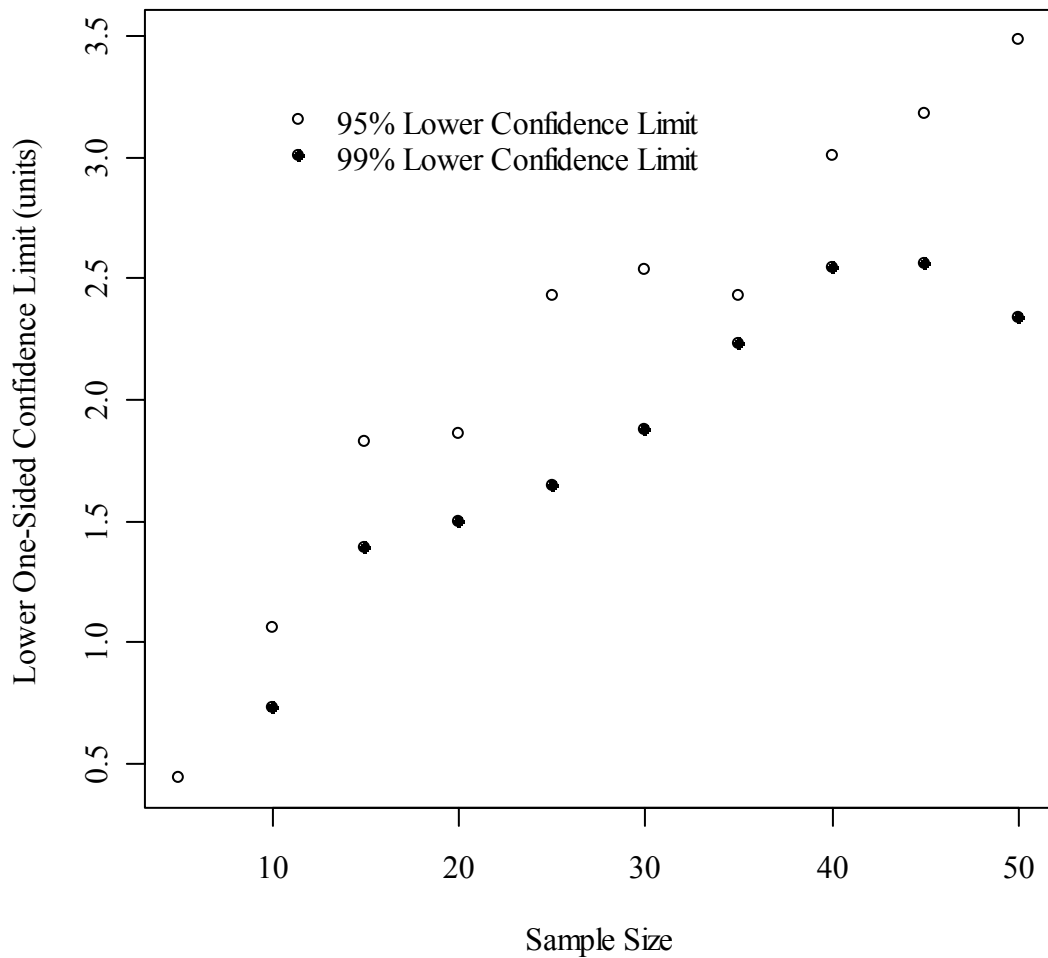


Figure 11: Sample Size Simulation Results for HC₅ – Log Normal Distribution, Copper Dataset

We observe that lower 99% confidence limit stabilizes after the sample sizes reaches 30. The lower 95% confidence limit continues to climb over the range of simulated sample sizes.

The HC₅ was estimated as 6.5356 µg /l with a lower one-sided 95% confidence interval of 3.9633 µg /l using the raw data with a sample size of 125. This concurs with the results of the simulation, above.

2.4.9 Oscillations in HC₅ Estimates

One of the project deliverables was to use one of the test datasets (zinc) broken up into three sub-sets (i) all data older than 1985, (ii) all data older than 1995 and (iii) the entire dataset, and run the SSD process to calculate a proposed target level (the HC₅) and the lower confidence limit. The lower confidence limit was not calculated as oscillations are due to both changes in sample sizes and different sets of observations.

As the Weibull distribution was used to model the full data set, it was also used to model subsets of the data with the assumption that truncating the data set by year would not affect the distribution best describing the data set.

Table 7: Oscillations in HC₅ Estimates for the Zn Dataset

Data Set Publication Date	HC ₅ (µg /l)	Sample Size	Lower one-sided 95% Confidence Limit (µg /l)
pre – 1985	48.3	57	16.1
pre – 1995	30.3	94	12.7
all data	26.7	101	9.7

Although it is tempting to posit a decrease in HC₅ as the sample size increases, one should be cautious. The decrease may be due⁵ to a preponderance of endpoints reflecting interest in lower percentiles of effects in latter years (i.e. more or a higher proportion of EC₂₀s following 1995 versus more or a higher proportion of EC₅₀'s prior to 1985).

We note that the HC₅ and the lower one-sided confidence limit vary by less than a factor of two for the subdivisions of data examined. Despite the increase in sample size the lower one-sided 95% confidence did not increase as expected. This is because the lower one-sided 95% confidence tracks the HC₅.

⁵ Attributes of the data set were not examined at this time. If the question of oscillations in HC₅ estimates requires further attention we suggest more comprehensive simulation studies.

3 Conclusions

3.1 Treatment of Replicates

Given the effect of averaging on data set size (shown in **Error! Reference source not found.**), it is likely very much worth the effort to determine what response an endpoint represents. For example consider the following data extracted from the “chronic” worksheet of the formaldehyde data set “... eau douce 1999(reformatted)1.xls”.

General Type	Common Name	Genus	Species	Duration	Endpoint	Effect Concentration
plant	brown algae	<i>Phyllospora</i>	<i>comosa</i>	96	LOEC	10
plant	brown algae	<i>Phyllospora</i>	<i>comosa</i>	96	LOEC	0.1

If the biological response represents the same measurement for each of the LOECs we should end up with 1 number $(10+0.1)/2$. However if the response is different for each entry (i.e. one LOEC represents root length and the other shoot length or anything different from root length) we should have two different LOECs. The lack of information on the response greatly reduced the apparent amount of available information.

3.2 Choosing a Distribution

When a sample size is large, distributional tests (where available) can provide an objective method for the non-statistician to choose a statistical distribution that describes the empirical sensitivity distribution.

However many of the available toxicity data sets are quite small where small is operationally defined as < 15 . Choosing a distribution when the sample size is small requires judgment that in our opinion can only be obtained through a combination of experience in fitting distributions, some knowledge of contaminant-specific modes of toxic action and an understanding of basic distribution theory.

If an objective of the CCME is to allow SSD estimates to be updated frequently by non-statisticians, we therefore recommend provision of software similar to Burrlioz (CSIRO, 2000). In this scenario, one or more distributions pre-selected for their utility in describing SSDs would be assessed and the best-fitting distribution would be selected using a numerical criterion.

There are a large number of statistical distributions that could describe SSDs. Only some of these were fit to the SSD data sets. The choice of potential distributions was restricted to those where routines for estimation of parameters and random number generators necessary for quantile-quantile plots were available. Of these the Weibull, lognormal and extreme value are worth further consideration based upon the data sets examined.

Note that the log-logistic distribution although advocated by Aldenberg and Slob (1993) and adopted by OECD (1995), was never the best fitting distribution for the data sets examined.

There are two distributions that were not examined due to the lack of available functions but that should be investigated. The first of these is the Burr Type III distribution as described in Shao, (2000) and implemented in the software distributed by CSIRO (2000). This distribution has as a limiting case⁶, the log-logistic distribution and therefore allows for congruence with OECD (1995) when appropriate.

The four parameter generalized F-distribution should also be examined as it includes, as limiting cases the following distributions: the generalized gamma, two parameter log-logistic, 3 parameter Burr and generalized Gumbel distributions. Estimating the 4 generalized F-distribution parameters likely implies stronger data requirements than other distributions with fewer parameters. However the ability to model asymmetric tails in the SSDs suggests that the utility of this distribution be examined.

From an implementation perspective if CCME proceeds with software development programming the generalized F-distribution might⁷ be less costly than programming for four separate distributions. Other “supra” distributions such as the generalized exponential might also be considered for similar reasons.

3.2.1 On Estimation Methods

3.2.1.1 Parametric versus Nonparametric Methods

Nonparametric methods use only the data collected with the maximum and minimum values estimated from the sample. As our interest is on the 5th percentile, the sampling properties of the minimum are to a large extent relevant.

⁶ The phrase “limiting case” refers to the case when a parameter taking on a specific value causes the model to simplify. A trite example is the simple linear regression equation $y = a + bx$, where x and y are the independent and dependent variables respectively. If the slope parameter “ b ” takes on the value 0, the model simplifies to $y=a$. This simpler model is the limiting case as b approaches 0.

⁷ Johnson et al (1994b) discuss maximum likelihood estimation via a generalized reduced gradient method (loc. cit. Lasdon, L.S., A.D. Waren, A. Jain and M. Ratner. 1978. Design and testing of a generalized reduced gradient code for nonlinear programming. Assoc. for Computing Machinery Transactions on Mathematical Software. 4:34-50.) The necessity of using such a specialized optimization method is unknown.

The minimum of a data set is a notoriously biased estimate of the population minimum and is greatly affected by sampling effort. As the nonparametric methods use only the data collected, they cannot “extrapolate” beyond the sample. Thus the 5th percentile must always lie within the sample. If we examine **Error! Reference source not found.**, we can see that the nonparametric HC₅ is always greater than the parametric HC₅.

Therefore we suggest that one criterion for sufficiency of a data set is that it can be fit by a parametric distribution. This also partially⁸ addresses the issue of data quality, as data of insufficient quality will not be describable by a statistical distribution.

3.2.1.2 Choice of Parametric Methods

Although not explicitly studied herein, the widely accepted maximum likelihood method was chosen as maximum likelihood parameter estimates are “sufficient”, asymptotically “consistent” and “best asymptotically” normal (Stuart *et al.*, 1999). The words and phrases in quotations have specialized statistical meanings beyond the scope of this document.

The important point is that in general, parameters estimated using maximum likelihood estimates are the best possible estimates.

3.2.1.3 Goodness of Fit Tests

Also not explicitly studied herein, but adopted based on emphasis on observations in the tail of a distribution is the Anderson-Darling goodness of fit test for all distributions except the normal. The Shapiro-Wilk test as recommended by D’Agostino and Stephens (1986) was chosen to evaluate the goodness of fit of the normal and lognormal distributions.

3.2.2 Treatment of Replicate Species Data

Table 8: Summary of Treatment of Replicate Species Data

Substance	Median		Geometric Mean	
	HC ₅	1 – sided 95% LCL	HC ₅	1 – sided 95% LCL
2,4-D (µg/l)	471.1012	106.5297	472.9441	107.8109
copper (µg/l)	6.5356	3.9633	6.2937	3.6603
uranium (mg active ingredient/l)	0.001323	7.6270E-006	0.001046	5.7334E-006
zinc (µg/l)	26.7494	9.7091	25.2001	9.1581

⁸ Other criteria for data quality include taxonomic representation, quality of individual data sets, etc.

Replicates (as defined in section 2.1.1) were aggregated using medians and geometric means. For the available datasets we observe little difference between the two approaches. Given the stability of the median relative to aberrant values we suggest that the use of the median be adopted.

This suggestion implies that the sensitivity corresponding to the largest proportion of a taxonomic group (for a given biological response) is of primary interest; not the most sensitive measurement that can be found.

There is also a sound statistical reason for using the median to aggregate replicate data. Any single observation in an SSD data set (the intersection of a toxicity test endpoint, biological response and taxa) represents the expected value⁹ for the conditions (exposure conditions, culture history, etc.) under which the toxicity test was conducted. This comprises one type of data – expected values.

If several “replicate” observations are available, choosing the minimum introduces another type of observation into the SSD dataset. This comprises another type of data – extreme values.

Empirically, the introduction of minima into a data set comprised largely of expected values may not be detectable, but it is theoretically undesirable as there would be two types of data within the data set.

3.3 The SSD Approach and Small Datasets

Small datasets can be fit to a parametric distribution. One consequence of fitting a distribution to a small data set, is that if a repeated (small) sample from the population were drawn (i.e. one obtained n additional endpoints) the new HC_x would vary slightly. The amount of variability varies inversely with sample size and the absolute distance between x and 50%.

Pederson *et al* (1994) found that for the Danish approach (HC₅ estimation following the log-normal distribution), the HC₅ did not change appreciably when more than 5 data points are used. ANZECC (2000) has not evaluated the minimum data requirements for their approach.

Our limited simulation using the Weibull and log normal distributions with specific sets of parameters estimated from the Zn and Cu datasets respectively, showed that the HC₅ stabilized for sample sizes greater than 25. Note that the median HC₅ estimates stabilize fairly rapidly. It may be this stabilization that Pederson *et al* (1994) are referring to.

Another consequence of estimating an HC_x from ever smaller datasets is that the lower confidence limit or guideline value will move further from the HC_x. This is a desirable change as it reflects decreasing certainty. OECD (1995), states that “although 5

⁹ The expected value for many distributions is well approximated by the median. Note that in this footnote “distribution” refers to the distribution of the replicates.

observations from at least 4 species may be sufficient to estimate the HC₅, the confidence interval will be so wide as to be unrealistically low”.

Our simulations showed that the lower one-sided confidence limits for the HC₅ increased as the sample size increased reflecting increasing certainty with respect to the HC₅. This is in keeping with the general result that as sample sizes increase, the width of confidence intervals, decreases. The lower 99% one sided confidence interval stabilized with samples sizes larger than 25 and 35 for the Weibull and log normal distributions, respectively. Note that these findings should not be over generalized as they apply to only a single set of parameters in each instance.

3.4 Problems with Current Implementations of the SSD Approach to Estimating Guidelines

This section discusses problems with current implementations of the SSD approach to guideline derivation. Some of these problems are theoretical and have not yet been addressed by any jurisdiction; another is being addressed by the Canadian approach.

1. The SSD approach as generally implemented (as well as herein) ignores the assumption that the data are measured without error. All toxicity test endpoints may be considered as random variables and hence have an associated variance. The effect of ignoring the variability around each endpoint comprising the SSD is that the variability around the parameters of the SSD is overestimated (part of the variability should be subsumed by a variance term for the measurement error). It is not possible to state the magnitude of the effect of ignoring this variability on derived guidelines. However we can say that the derived guidelines will be conservative estimates; that is removing this extraneous variability will only increase the guideline value.

Note that this bias is not an inherent characteristic of the SSD approach. We can only speculate as to why this variability is ignored by other jurisdictions; possible reasons include ignorance or unavailability of variance estimates for the endpoints.

2. All jurisdictions limit the choice of potential statistical distributions to one, although the Australian/New Zealand distribution (Burr Type III) includes another distribution as a limiting case. Proponents of the bootstrapping approach to HC_x estimation report on the number of times an SSD cannot be fit to the single distribution allowed by a jurisdiction (Newman *et al.*, 2000, Grist *et al.*, 2002) as a rationalization for the bootstrap approach. Examination of only 6 datasets revealed that at least two distributions were required to describe the datasets. We do not believe that any single distribution is sufficient to describe all datasets and our (limited) empirical results and the findings of the authors cited above, confirm this belief.

3. The proportions of primary producers, invertebrates and fish in the environment are 64, 26, and 10%, respectively (Forbes and Calow, 2002). However the distribution of species in a SSD database does not necessarily reflect this ratio. Forbes and Calow, (2002) found that the proportion of these organisms in an SSD database used by Versteeg *et al*¹⁰ (1999) was 28, 35 and 39%.

Taxonomic representation will likely affect the HC_x and associated lower confidence interval. Duboudin *et al* (2002) found that among methods for accounting for within-species variation and weighting to account for taxonomic composition, weighting most significantly affected the HC_x.

We suggest that the discussion regarding species composition for acute and chronic fresh and marine water guidelines held on June 21-22 by the CCME SSD sub-committee be re-visited in light of this information.

4. When data are excessively variable the lower confidence limit of the HC₅ may be less than zero (and consequently set to 0). This is partially a consequence of the method used but also partially a consequence of variability in the data set and proximity of the HC₅ to zero. Various options to cover this scenario are discussed below:
 - a. State that a guideline cannot be estimated due to the insufficiency of the data set.
 - b. Use an alternative confidence limit such as 50% and declare such a guideline as “provisional”. The Dutch use two degrees of confidence, 95 and 50%. Guidelines derived using the 95% confidence limit are designated as “primary” and those using a lower 50% confidence limit (around the HC₅) are designated as “secondary” (Aldenberg and Slob, 1993 and MHSPE, 1994). The difference between 50 and 95% levels of secondary protection may be as much as 1 order of magnitude (Warne, 1998). Such a large difference might encourage stakeholders to generate additional data.

¹⁰ *loc. cit.* Versteeg D.J, S.E. Belanger and G.J. Carr GJ. 1999. Understanding single species and model ecosystem sensitivity: Data-based comparison. *Environ. Toxicol. Chem.* 18:1329–1346.

3.5 Recommendations

Recommendations made elsewhere in this document are summarized within this section. Note that some of the recommendations fall outside the scope of project deliverables. These recommendations are based upon ancillary literature review, participation in working group meetings and experience with the application of statistics to environmental data.

When constructing databases for SSD estimation:

- Attempt to include the biological response measured so that toxicity test endpoints representing different biological responses are not incorrect¹¹ly synthesized (through an average, median, geometric mean, minimum etc.)
- Obtain the standard error or variance of the toxicity test endpoint so that this variability can be accounted for when estimating the parameters of the SSD.

When rationalizing the CCME protocol:

- More thoroughly investigate the effects of small sample sizes on the lower one-sided 95% confidence limit of the HC₅.

When developing the CCME protocol:

- Revisit the discussion regarding species composition for acute and chronic fresh and marine water guidelines held on June 21-22 by the CCME SSD sub-committee in light of the discussion in section 3.4.
- Consider examination of the Burr Type III distribution as a descriptor of SSDs.
- Consider examination of the generalized F distribution as a descriptor of SSDs.

¹¹ If raw data are available the standard error of endpoints other than the NOEC and LOEC may be estimated while estimating the statistical endpoint. Standard errors may also be backcalculated from confidence intervals when the method for generating the confidence interval is known. As most non-statisticians almost always use a large sample approximation using a Wald-type variance estimate, the standard error is straightforward to estimate. One problem with this suggestion is that there is no variance for NOECs and LOECs using the currently advocated hypothesis testing approaches.

If variance estimates are available, likelihood functions incorporating this extra source of variability would be optimized. The outcome will be a reduction in the width of the confidence around the HC_x which is currently overinflated.

When implementing the CCME protocol:

- Consider writing software that follows (either completely¹² or partially) the CCME SSD protocol.
- Review the software Burrlioz for potential use in Canada.
- Consider adopting the pragmatic criterion¹³ that if a data set it can be fit by a parametric distribution the data set is sufficient to develop a CCME guideline.

¹² The cost of writing software is a function of how many decisions the software must make on behalf of the user.

¹³ Note that this will not be the sole criterion for adequacy of a data set for generating CCME guidelines.

4 Citations

- Aldenberg, T. and W. Slob. 1993. Confidence limits for hazardous concentrations based on logistically distributed NOEC data. *Ecot. Env. Safety*. 25:48-63.
- ANZECC, 2000. Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Paper No. 4.
- CSIRO, 2000. Burrlioz, ver. 1.0.14.
- D'Agostino, R. B. and M.A. Stephens. 1986. Goodness-of-fit techniques. Marcel Dekker, New York.
- Duboudin, C., P. Ciffroy and H. Magaud. 2004. Effects of data manipulation and statistical methods on species sensitivity distributions. *Env. Tox. Chem.* 23(2):489-499.
- Forbes V.E, and P. Calow. 2002. Species sensitivity distributions revisited: A critical appraisal. *Hum Ecol Risk Assess* 8:3:473–492.
- Grist, E.P.M., K. M.Y. Leung, J. R. Wheeler, and M. Crane. 2002. Better bootstrap estimation of hazardous concentration thresholds for aquatic assemblages. *Env. Tox. Chem.* 21(7):1515-1524.
- Hahn, G.H. and W. Q. Meeker. 1991. *Statistical Intervals: A Guide for Practitioners*. Wiley & Sons, New York.
- Harrell, F.E. and C.E. Davis. 1982. A new distribution-free quantile estimator. *Biometrika* 69:635-640.
- Hyndman, R. J. and Y. Fan. 1996. Sample quantiles in statistical packages. *Amer. Stat.* 50:361-365.
- Johnson, N.J., S. Kotz and N. Balakrishnan. 1994a. *Continuous Univariate Distributions Volume 1 – 2nd Edition*. Wiley Series in Probability and Mathematical Statistics, New York.
- Johnson, N.J., S. Kotz and N. Balakrishnan. 1994b. *Continuous Univariate Distributions Volume 2 – 2nd Edition*. Wiley Series in Probability and Mathematical Statistics, New York.
- MHSPE. 1994. Environmental quality objectives in the Netherlands. Ministry for Housing, Spatial Planning and Environment, The Hague, The Netherlands.

- Newman, M.C., D.R. Ownby, L.C.A. Mezin, D.C., Powell, T.R.L. Christensen, S.B. Lerberg, S.B. and B.A. Anderson. 2000. Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient numbers of species. *Env. Tox. Chem.* 19 (2):508–515.
- OECD. 1995. Guidance document for aquatic effects assessment. OECD Environment Monographs No 92, Organisation for Economic Co-operation and Development, Paris.
- Pedersen F, P. Kristensen, A. Damborg and H.W. Christensen. 1994. Ecotoxicological evaluation of industrial wastewater. Miljøprojekt nr. 254. Danish Environmental Protection Agency, Ministry of Environment, Copenhagen.
- Shao, Q. 2000. Estimation for hazardous concentrations based on NOEC data: an alternative approach *Environmetrics*. 11:583-595.
- Stewart, A., J.K. Ord and S. Arnold. 1999. *Kendall's Advanced Theory of Statistics: Volume 2A Classical Inference and the Linear Model*. Arnold Publishing London.
- Warne, M. St.J. 1998. Critical review of methods to derive water quality guidelines for toxicants and a proposal for a new framework. Supervising Scientist Report 135, Supervising Scientist, Canberra.

5 Appendix 1: Quantile-Quantile Plots

A quantile-quantile plot is a scatter plot used to compare two distributions to one another, or one distribution to a theoretical distribution. When determining whether a random variable has a specific distribution, quantiles from the observed data are plotted against the quantiles expected from the posited distribution. The scatter plot of observed versus theoretical quantiles will produce a straight line if the observed data follows the posited distribution.

Notes:

- A quantile or percentile is an observation below which, the stated proportion of observations lie. For example a median is the 50th percentile of a data set.
- Theoretical quantiles are often referred to as order statistics.

The following graphics contain quantile-quantile plots. The observed or sample quantiles are on the vertical axis and the theoretical quantiles or order statistics corresponding to the chosen distribution are on the horizontal axis although reversal of the axis is inconsequential.

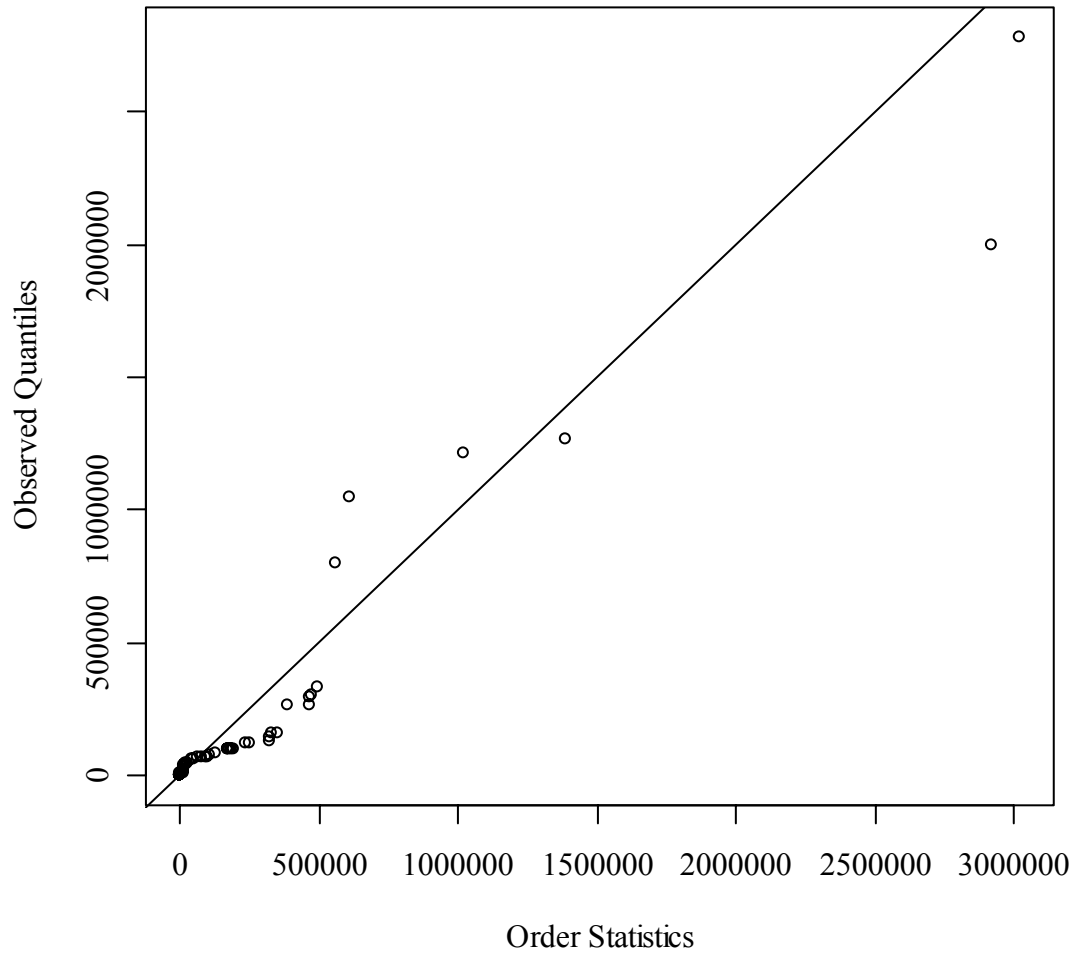
Two points are should be considered by the reader examining these graphics:

- 1) How well do the pairs of observations fall along a straight line? A strong straight linear relationship implies that the data fit the putative distribution well.
- 2) Are one or two extreme pairs of observations responsible¹⁴ for an apparently good fit? A few influential pairs of observations indicate that the fit to the observed distribution may not be driven by the majority of the data set. In this case quality assurance of the unusual observations is important to make sure that distribution fit is not spurious.

¹⁴ Such observations are known as “influential” observations and have a very particular statistical definition. Measures of influence were not estimated when performing tests of goodness of fit.

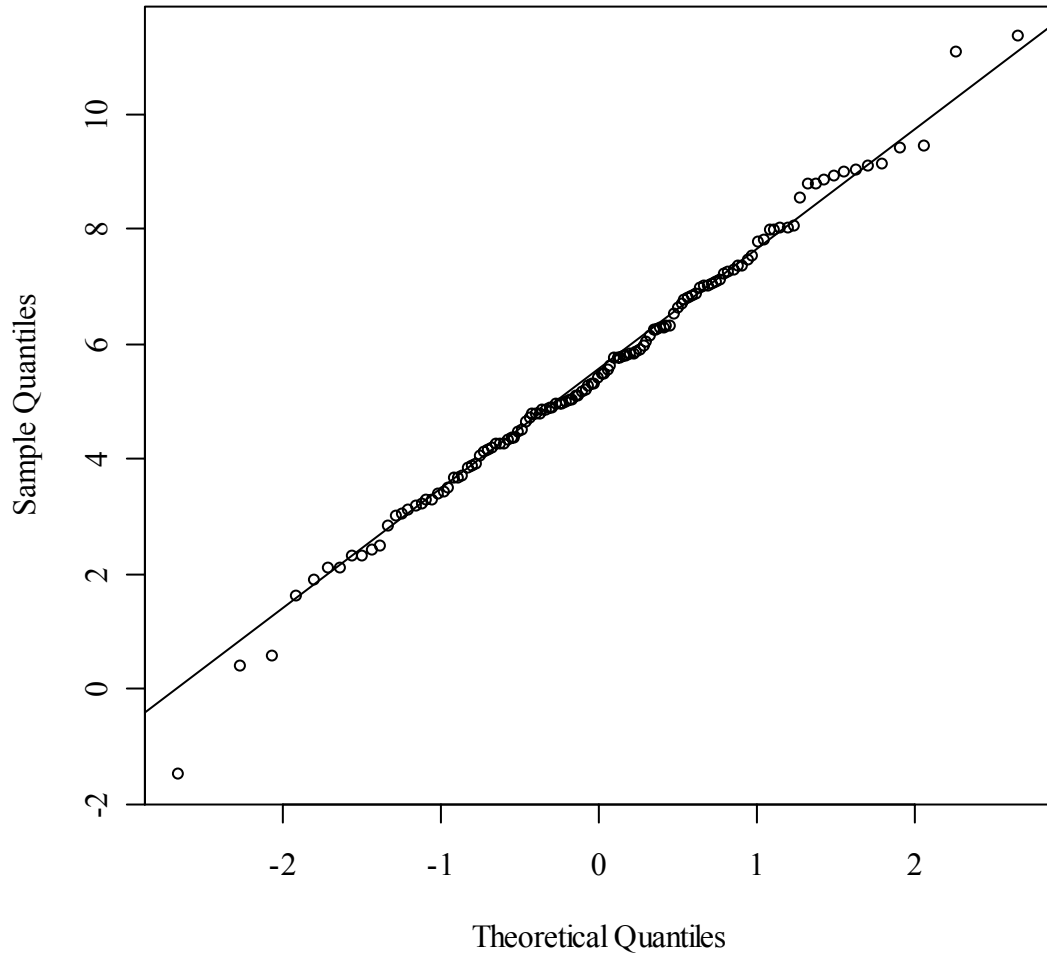
5.1 2,4-D

QQ-plot for Weibull Distribution



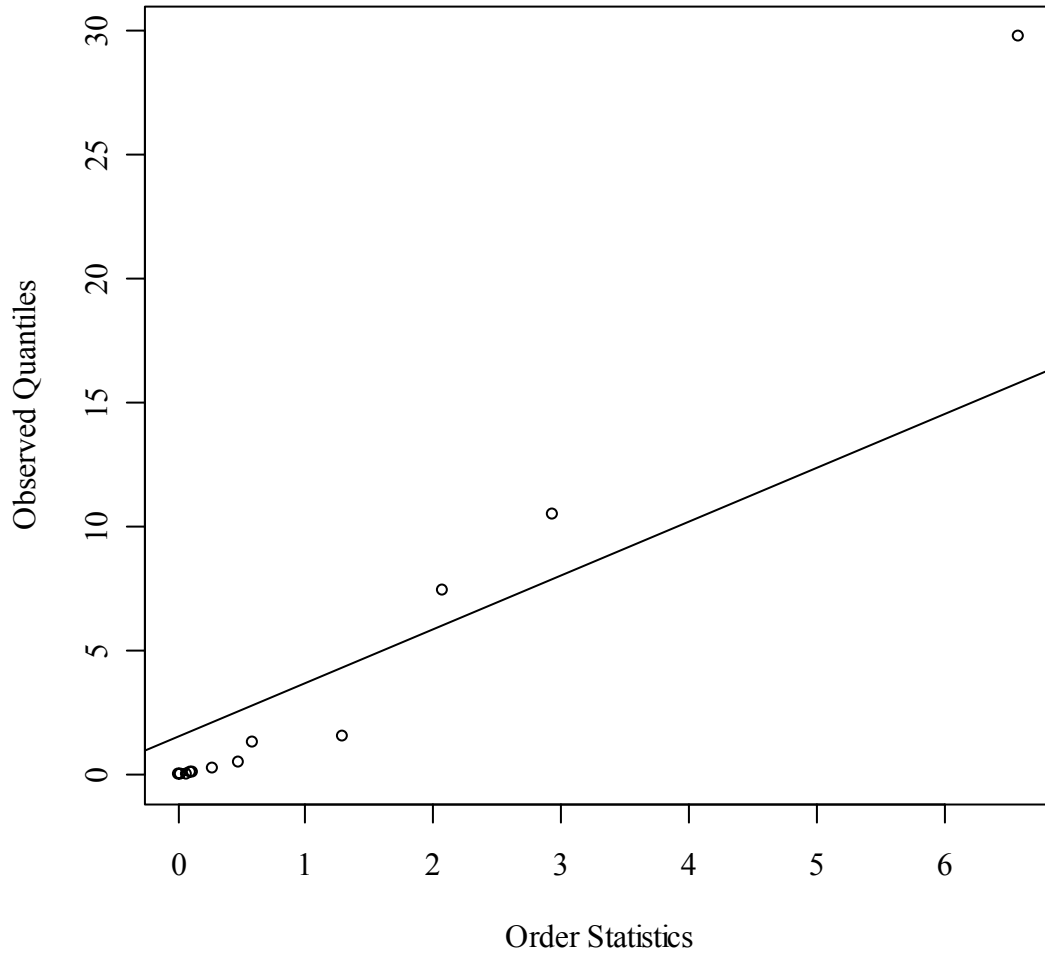
5.2 Copper

Normal Q-Q Plot



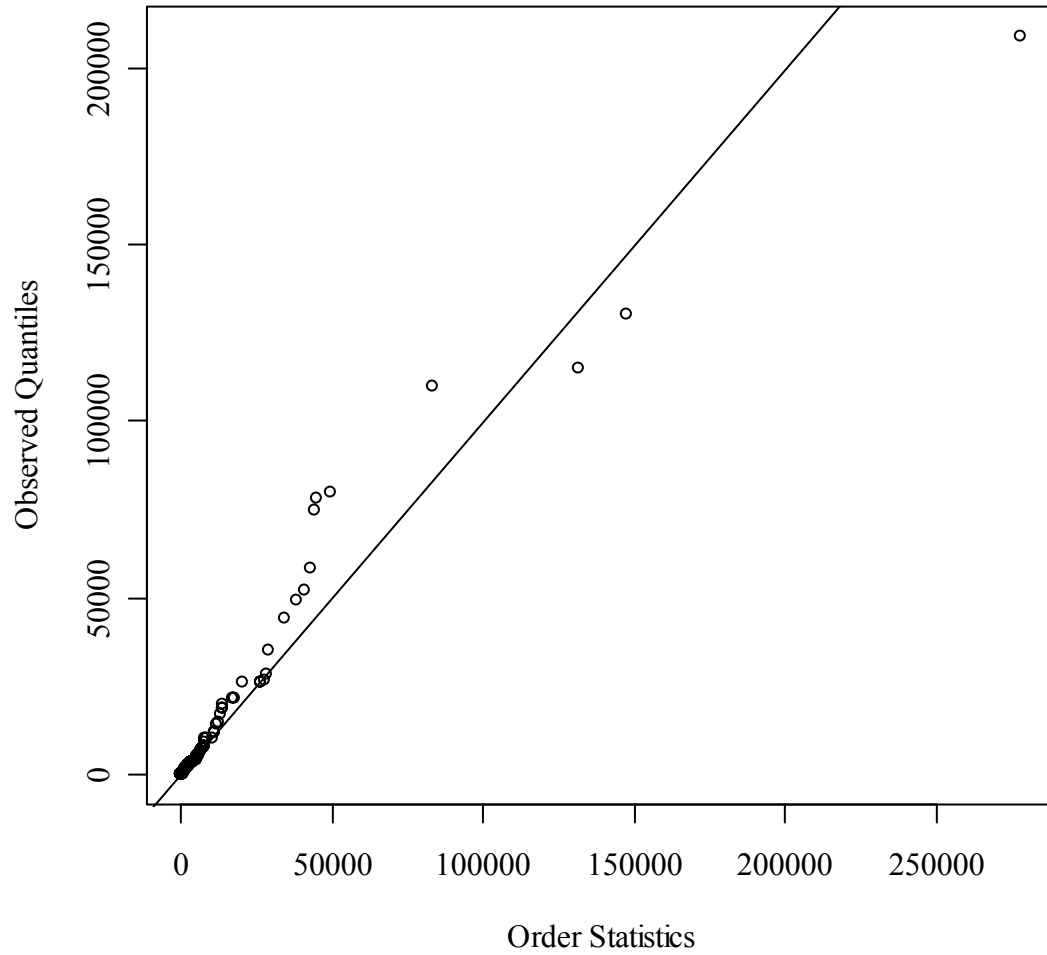
5.3 Uranium

QQ-plot for Weibull Distribution



5.4 Zinc

QQ-plot for Weibull Distribution



6 Appendix 2: Probability Distribution Functions

A frequency histogram may be used to illustrate some useful concepts regarding statistical distributions.

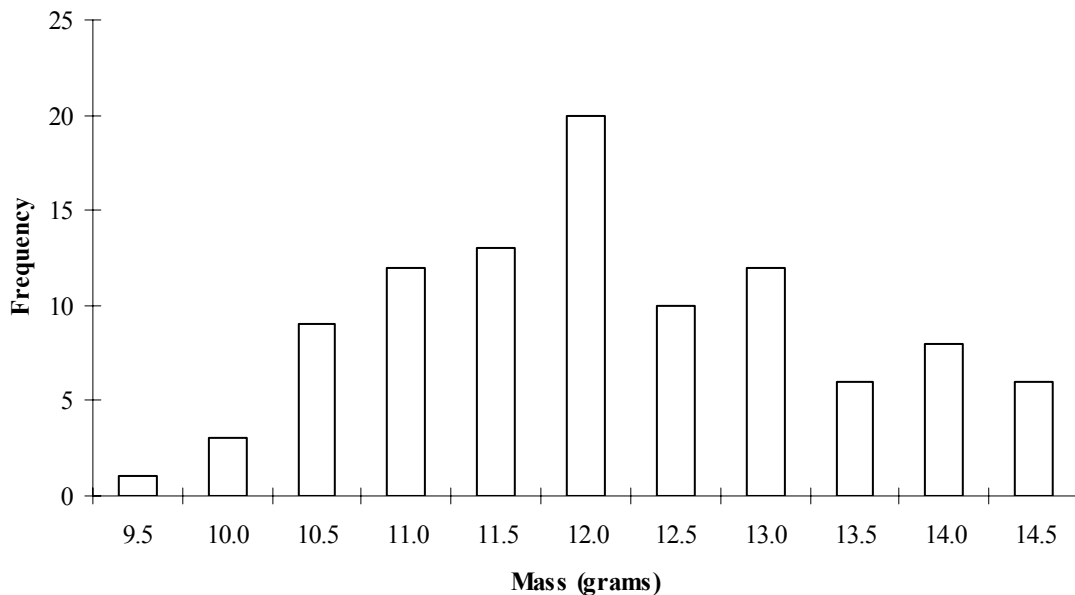


Figure 12: Frequency Histogram of Fish Weights

Each box in the frequency histogram represents a range of fish weights. The first box represents the number of fish with weights between 9.25 and 9.75 grams with a mid-point of 9.5 grams. If this number is divided by the total number of fish, the results are expressed as a proportion. Figure 12, above represents 100 fish, therefore 1/100 fish fall into the first size class. Given these size classes, and this histogram a potential question is: “How likely would it be to randomly select a fish (from the same group of fish from which the original 100 fish were randomly selected) with a weight greater than 13.75 grams?” The answer is 14/100 or 14% (8 fish in the 13.75-14.25g size class + 6 fish in the 14.25 -14.75g size class). The answer 14%, is a function of the distribution of the fish weights across the size classes. Another question is “How likely is it that a randomly selected fish weighs between 11.75 and 12.25 grams? The answer is 12%. Note that if the shape of the distribution changes so do our answers.

Fish weight can take on any value within a certain minimum and maximum. Thus it is an example of a continuous random variable. The probability of observing any single value is zero for continuous data because in theory we can always increase the precision of the measurement to the point where the probability of observing that specific value becomes vanishingly small. However the probability that a random variable falls between two

values can be estimated from a frequency histogram as shown above. A mathematical function describing the distribution of probability is known as a probability distribution function or pdf. The pdf for the normal distribution is given below.

Equation 1: The Normal Probability Distribution Function

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \text{ for } \sigma > 0, -\infty < x < \infty, \mu < \infty,$$

where:

- μ is the population mean
- σ is the population standard deviation; and
- x is the observed value.

This pdf produces the familiar bell shaped curve encountered in many introductory statistics courses. Other pdfs or statistical distributions have other characteristic shapes. How can we use this information to determine what distribution might represent a given data set?

As shown above, the frequency histogram is an empirical representation of a pdf. If we plot a pdf and it matches the pdf from a known distribution we have empirical evidence that the observed data corresponds to the putative distribution. Formal tests are also available to determine whether a dataset follows a specific distribution.

We conclude this section with two definitions:

probability distribution - A function describing the probability that a random variable is \leq some specified value. A familiar example is the sigmoidally shaped normal distribution. If the random variable is equal to 1.645, the probability of being less than 1.645 is 95%.

probability density: A function describing the probability that a random variable falls between two specified values. The familiar bell-shaped normal distribution is an example of a probability density.