Unclassified **ENV/JM/MONO(2006)18**

Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development

09-May-2006

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND
THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

ENV/JM/MONO(2006)18
Unclassified

OECD SERIES ON TESTING AND ASSESSMENT
Number 54

CURRENT APPROACHES IN THE STATISTICAL ANALYSIS OF ECOTOXICITY DATA: A GUIDANCE TO APPLICATION

Ms. Laurence MUSSET
Tel: +33 (0)1 45 24 16 76; Fax: +33 (0)1 45 24 16 75; Email: laurence.musset@oecd.org

JT03208537

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

English - Or. English

OECD Environment Health and Safety Publications

Series on Testing and Assessment

No. 54

CURRENT APPROACHES IN THE STATISTICAL ANALYSIS OF ECOTOXICITY DATA: A GUIDANCE TO APPLICATION

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2006

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (1993; reformatted 1995)*

No. 2, *Detailed Review Paper on Biodegradability Testing (1995)*

No. 3, *Guidance Document for Aquatic Effects Assessment (1995)*

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment (1995)*

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing (1996)*

No. 6, *Report of the Final Ring-test of the Daphnia magna Reproduction Test (1997)*

No. 7, *Guidance Document on Direct Phototransformation of Chemicals in Water (1997)*

No. 8, *Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)*

No. 9, *Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides during Agricultural Application (1997)*

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (2004)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals Which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (draft)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (2005)*

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances Which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (in preparation)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures Which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (2003)*

No. 43, *Draft Guidance Document on Reproductive Toxicity Testing and Assessment (in preparation)*

No. 44, *Description of Selected Key Generic Terms Used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

No. 46, *Detailed Review Paper on Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances (2004)*

No. 47, *Detailed Review Paper on Fish Screening Assays for the Detection of Endocrine Active Substances (2004)*

No. 48, *New Chemical Assessment Comparisons and Implications for Work Sharing (2004)*

No. 49, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)Sars] on the Principles for the Validation of (Q)Sars (2004)*

No. 50, *Report of the OECD/IPCS Workshop on Toxicogenomics (2005)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 52, *Comparison of emission estimation methods used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case study of pulp and paper and textile sectors (2006)*

No. 53, *Guidance Document on Simulated Freshwater Lentic Field Tests (Outdoor Microcosms and Mesocosms) (2006)*

No. 54, *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (2006)*

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in ten different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides and Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and the Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/ehs/).

This publication was produced within the framework of the Inter-Organisation Programme for the Sound Management of Chemicals (IOMC).

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The participating organisations are FAO, ILO, OECD, UNEP, UNIDO, UNITAR and WHO. The World Bank and UNDP are observers. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

**For this and many other Environment,
Health and Safety publications, consult the OECD's
World Wide Web site (www.oecd.org/ehs/)**

or contact:

**OECD Environment Directorate,
Environment, Health and Safety Division**

**2 rue André-Pascal
75775 Paris Cedex 16
France**

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

In May 2000, the 12th Meeting of the Working Group of the National Coordinators of the Test Guidelines Programme (WNT) agreed that a Guidance Document describing the main statistical methods used for analysis of data from ecotoxicological studies should be developed. In 2000, an Expert Group was established for this project, consisting of experts nominated from seven member countries led by France (Institut National de l'Environnement Industriel et des Risques), as France was already taking the lead in an ISO Working Group that had been established earlier in the same year to develop a guidance document on the same subject. The Expert Group met in May 2001, and agreed that the Group should work together with the ISO Working Group to avoid duplication of work.

The OECD Expert Group and the ISO Working Group, meeting jointly in September 2001, May 2002 and February 2003, developed a draft Guidance Document. In May 2003, the draft was circulated to the National Co-ordinators for their review. The same document was circulated to the ISO member bodies as a Committee Draft (ISO/CD20281) for comment. A number of member countries and stakeholders made comments.

The Expert Group had its last meeting in October 2003 to review these comments and discuss the revisions to the draft. After reviewing all the comments at the meeting, the Expert Group revised the draft and continued discussion through e-mail communications following an agreed schedule. All the issues raised in this communication were taken into account. Best efforts were made to accommodate late comments. In March 2004, the chairperson of the Expert Group submitted the revised Draft Guidance Document for discussions at the 16th WNT in May 2004.

The WNT, while appreciating the efforts by the Expert Group, identified some rather fundamental issues relating to the document and the expectations of some of the member countries. Different countries expressed different opinions on how guidance on statistical analysis could be included in this document. The discussion at the WNT resulted in an agreement that the proposed document was valuable as an overview of current approaches for statistical analysis, but should not be called a Guidance Document, because this document did not provide specific guidance on the statistical methods that should be used for specific purposes or in particular circumstances. The 16th WNT agreed to have a further commenting period to finalise the document. France kindly agreed to receive further comments until the middle of July 2004.

No further comments were received except for the expression of approval of the document. In line with the agreement at the 16th WNT meeting, the 17th WNT approved the document under the new title: "Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application".

This document is published on the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology.

TABLE OF CONTENTS

1. INTRODUCTION

1. Ecotoxicity tests are biological experiments performed to examine if either a potentially toxic compound, or an environmental sample (e.g. effluent, sediment or soil sample) causes a biologically important response in test organisms. If so, the goal is to determine the concentration that produces a given level of effects or produces an effect that cannot be distinguished from background variation.

2. In a test, organisms are exposed to different concentrations or doses of a test substance or a test substrate (e.g. waste water, sludge, or a contaminated soil or sediment), sometimes diluted in a test medium. Typically, at least one group of test organisms (the control group) is not exposed to the test substance or substrate, but is otherwise treated in the same way as the exposed organisms

3. The endpoint(s) observed or measured in the different batches may be the number of surviving organisms, size or growth of organisms, number of eggs or offspring produced or any relevant biochemical or physiological variable that can be reliably quantified. Observations are made after one or several predefined exposure times. The endpoint's relationship with the concentration of the tested chemical or substrate is examined. The way statistics are applied may have a considerable impact on the results and conclusions from ecotoxicity tests, and consequently on the associated policy decisions. Various documents (Williams, 1971, Piegorsch and Bailer, 1997; Tukey et al., 1985, Pack, 1993; Chapman et al., 1995; Hoekstra, 1993; Kooijman & Bedaux, 1996; Laskowkj, 1995; Chapman, 1996; OECD, 1998; ASTM, 2000) exist on the use of available statistical methods, the limitations of these methods and how to cope with specific problematic data. Discussions of statistical principles and commonly used techniques are found in general references as Armitage and Berry (1987) [basic information on hypothesis testing and regression, transformations], Finney (1978) [analysis of quantal data, especially probit models], Hochberg and Tamhane (1987) [thorough treatment of multiple comparison methods], Newman (1994) [information related to biology based models, ECx], and Sparks (2000) [a collection of articles covering field and laboratory experiments, multivariate techniques, risk assessment, and environmental monitoring] .

4. When problematic data are encountered or critical decisions depend upon inferences from ecotoxicity tests, consultation with a qualified statistician is useful. [Note that statisticians should be consulted before beginning the experiment to ensure proper design, sample size, limitations, and to be sure that the study will actually be able to answer the research question that the experimenter poses. Once bad data have been collected, there is little a statistician can do to rectify the problem.]

5. Chapter 8 contains a table listing all the existing ISO and OECD ecotoxicity standards/guidelines that could be analysed using this guidance document. For each standard/guideline, reference is made to the adapted chapters of this document.

6. Chapter 4 details the different statistical approaches that can be used for the analysis of ecotoxicity data, depending on the aim. In particular, it gives the assumptions made when using hypothesis testing methods, concentration -response modelling methods or biology based methods and their limitations. It also gives some indication on experimental design issues. Some general principles and advice are also given for the process of data analysis.

7. Chapter 5 deals with hypothesis testing, chapter 6 with dose response modelling and chapter 7 with biology based methods.

8. There was an ISO resolution (ISO TC147/SC5/WG10 Antalya 3) as well as an OECD workshop recommendation (OECD, 1998) that the NOEC should be phased out from international standard.

9. However, the NOEC is still required in many regulatory standards from many countries and in some cases where a detailed determination of an ECx is not relevant and the alteration of the study design is too costly to fulfil the requirements for regression models. Therefore guidance will be provided on the statistical methods for the determination of the NOEC.

10. In the annexes, examples of analyses with the three main methods (hypothesis testing for NOEC estimation, dose response modelling and biology base modelling) of 4 different data sets are given. They concern:

- Acute toxicity on *Daphnia magna*
- Inhibition of algae growth
- Reproduction of *Daphnia magna*
- Fish growth

2. SCOPE

11. This document is a description of statistical methods for the analysis of data of standardised ecotoxicity tests. It focuses on statistical methods for obtaining statistical estimates of parameters in current and future use, e.g. ECx (LCx), NOEC, NEC etc.

12. The methods described here are intended to cover laboratory ecotoxicity tests: aquatic, sediment or terrestrial tests, and may also be relevant for other toxicity tests.

13. The main objective of this document is to provide practical guidance on how to analyse the observations from ecotoxicity tests.

14. Hypothesis testing, concentration response modelling and biology based modelling will be discussed for the different data types (quantal, continuous and discrete data corresponding to mortality, growth or reproduction).

15. In addition, some guidance on experimental design will be given. Although the main focus is on giving assistance to the experimentalist, a secondary aim is to help those who are responsible for evaluating toxicity tests. And, finally, the document may be helpful in developing new toxicity tests guidelines by giving information on experimental design and statistical analysis issues.

3. DEFINITIONS

16. In this guidance document, the following definitions, sorted by alphabetical order, apply:

Accuracy and Precision

The quality of an estimated parameter from a set of data has two aspects: accuracy and precision. Accuracy is a measure of how close the estimate is to the 'true value' of the parameter (this true value is unknown). Precision is a measure of the amount of variability in the estimate (quantified by the standard error or the confidence interval of the estimate). Precision may be increased by using larger sample sizes or by reducing the experimental variation. However, as Figure 3.1 illustrates, an estimate being precise does not imply that it is also accurate.

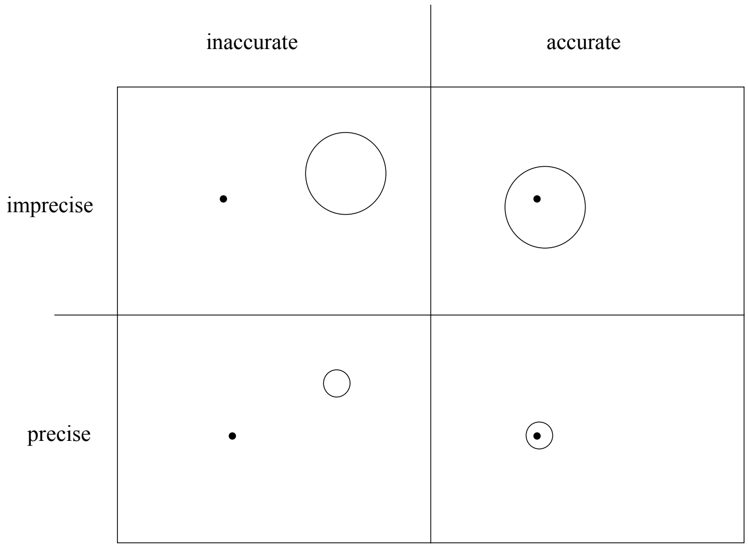


Figure 3.1 Conceptual illustration of accuracy and precision.

The dot represents the true parameter value, the circle represents the confidence interval of the estimate, small circles indicate high precision and large circles indicate low precision.

Concentration and Dose

Concentration and dose both refer to the amount of test material to which the test organism is subjected. Concentrations are used to describe the amount of test material in the testing environment (e.g., mg/L in water, mg/kg in soil or mg/kg in food). Doses are used to describe the amount of test material administered to a subject (e.g., mg/kg-bodyweight in an avian bolus study). Statistical methods for both types of studies are identical; however, interpretations are different. Although "concentration" is used throughout this document, all the statistical methods presented here also apply to studies in which a dose is used.

Confidence interval

A rough definition could be: An x% confidence interval for a parameter is an interval of values that theoretically covers the true value of the estimated parameter with x% confidence. Standard confidence intervals are based on the assumption that the underlying mathematical model is correct. It does not capture model uncertainty.

A more precise definition is the following: interval estimator (T_0 , T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that P[$T_0 < \theta < T_1$] ≥ 1 - α [ISO 3534-1[1)]].

NOTE 1: Associated with this confidence interval is the confidence level 100(1 – α)% where α is generally a small number. The confidence level is typically 90% or 95%. The inequality P[$T_0 < \theta < T_1$] ≥ 1 - α holds for any specific but unknown population value of θ.

NOTE 2: A confidence interval does not reflect the probability that the observed interval contains the true value of the parameter (it either does or does not contain it). The confidence reflects the proportion of cases that the confidence interval would contain the true parameter value in a long series of repeated random samples under identical conditions.

Data types

Quantal/binary data

Quantal (binary) data arise when a particular property is recorded to be present or absent in each individual (e.g. an individual shows an effect or it does not show an effect). Therefore, these data can exhibit only two states. Typically, quantal data are presented as the number of individuals showing the property (e.g., mortality) out of a total number of individuals observed in each experimental unit. Although this can be expressed as a fraction, it should be noted that the total number of individuals cannot be omitted.

Continuous data

Data are continuous when they can (theoretically) take any value in an open interval, for instance any positive number. Examples include measurements of length, body weight, etc. Due to practical reasons the measured resolution depends on the quality of the measurement device. For example, if test units are observed once per day then ‘time to hatch’ can only be recorded in whole days; however, the underlying distribution of ‘time to hatch’ is continuous. Typically, continuous data have a dimension (e.g. grams, moles/litre).

Discrete data

Discrete data are data that have a finite or countable number of values. There are three classes of discrete data: nominal, ordinal and interval. *Nominal data* express qualitative attributes that do not form a natural order (e.g. colours). *Ordinal data* reflect the relative magnitude from low to high (e.g. an individual shows no effect, minimal effect, moderate effect or high effect). These data cannot be interpreted with regard to relative scale (i.e., an ordinal variable with a value of ‘4’ can be interpreted as being higher then the value of ‘2’, but not twice as high). Ordinal data can often be reduced to quantal data. *Interval* data (e.g., number of eggs or offspring per parent) allows the ranking of the items that are measured, and the differences between individuals and groups can be quantified. Often, interval data can be analysed as if the data were continuous. The analyses for interval discrete data are presented in this document; analyses of nominal and ordinal data are not included but will be addressed in a future revision

Effect

[1] In preparation

An effect is a change in the response variable under consideration compared to a control. For quantal endpoints, an effect is usually described in terms of a change in the percentage of individuals affected. For continuous endpoints, it is typically described in terms of a percent change in the mean values of the endpoint, but it can also be described in terms of absolute change.

Effect Concentrations

Quantal LCx/ECx (LDx/EDx)

The quantal 'Effective Concentration' or 'Effective Dose' is the concentration of test material in water, soil, or sediment (e.g., mg/L or mg/kg) or dose of test material (e.g., mg/kg-bodyweight in an avian bolus study) that causes x% change in response (e.g., mortality, immobility) during a specified time interval. This corresponds to an effect predicted on x% of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling. An example of a concentration-response relationship and its associated estimates of EC_{10} and EC_{50}, are illustrated in Figure 3.2. When the effect is mortality, LCx or LDx are the abbreviations used.

Continuous ECx (EDx)

The continuous 'Effective Concentration' or 'Effective Dose' is the concentration of test material in water, soil, or sediment (e.g., mg/L or mg/kg) or dose of test material (e.g., mg/kg-bodyweight in an avian bolus study) that causes x% in the *size* of the endpoint during a specified time interval. This parameter is also estimated by dose-response modelling.

Endpoint (or Response variable)

The endpoint is the biological parameter observed, e.g. survival, number of eggs, size or growth, enzyme level. An ecotoxicological study can have one or many endpoints.

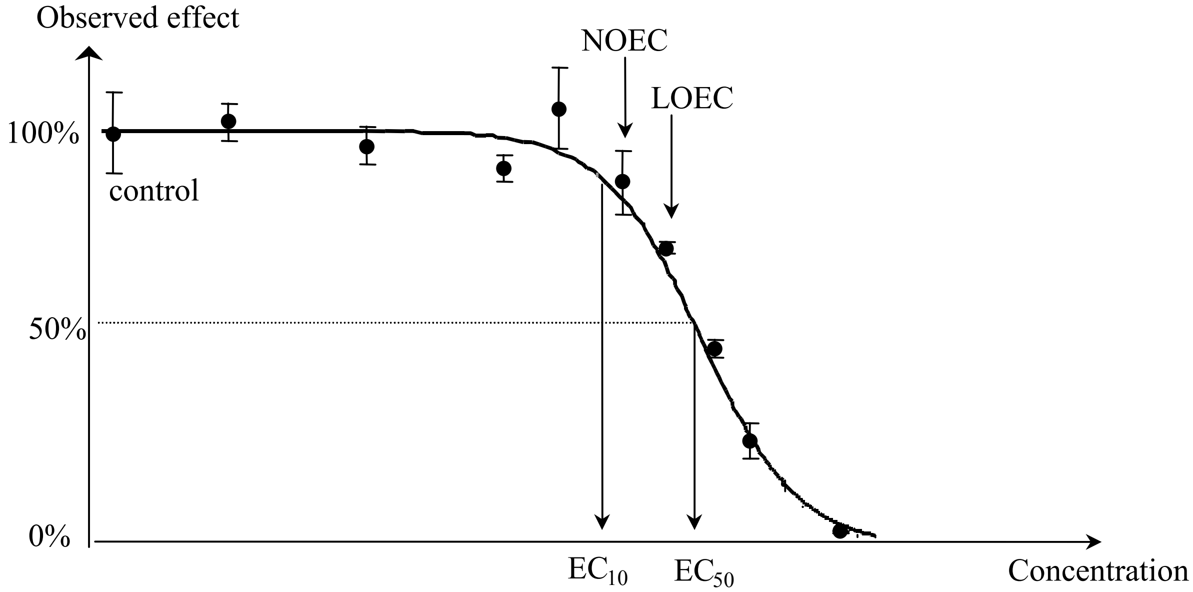


Figure 3.2 Illustration of a concentration-response relationship and of the estimates of the ECx and NOEC/LOEC.

The order of the parameters given in this figure has been taken at random.

ETx (LTx)

The ETx ‘Effective Time’ or LTx ‘Lethal Time’ is the time at which an effect of x% is expected at a specified test concentration when the test organisms are exposed to a given concentration of material in water or sediment or soil). An LTx is estimated when the response of interest is mortality. ETx (LTx) is estimated by modelling a time-response relationship.

Experimental unit/replicate/sampling unit

The experimental unit (replicate) is the smallest unit of experimental material to which a treatment can be allocated independently of all other units. By definition, experimental units (e.g., aquariums, beakers, or plant pots) must be able to receive different treatments. Each experimental unit may contain multiple sampling units (e.g. fish, daphnia or plants) on which measurements are taken. Within each experimental unit, sampling units may not be independent. However, in some special case situations, individual organisms (housed in common units) can be treated as the experimental units: these special cases require some proof or strong argument of independence of organisms.

Extrapolation / Interpolation

Extrapolation refers to predicting the value of variates outside the range of observations. Extrapolation may not lead to a reliable estimate (see e.g. section 6.4).

Interpolation refers to predicting the value of variates within the range of observations. For example, when an ECx estimated from a fitted concentration-response function is lower than the lowest nonzero concentration tested in the study or higher than the highest concentration tested in the study, it is obtained by extrapolation. When the ECx is between two consecutive nonzero test concentrations, it is said to be obtained by interpolation.

Hormesis

Hormesis is an effect where the tested substance is a stimulant in small concentrations, but it is inhibitory in large concentrations. The result is a biphasic (or U-shaped) concentration-response relationship. This observed stimulatory effect may be due to the tested substance, but it could also be due to an experimental artefact (e.g., solvent effect, non-random allocation of treatments to experimental units, experimental error). Models incorporating hormesis are not detailed in this document; analysis approaches will be addressed more fully in future documents. Two issues of Critical Reviews in Toxicology (2001, volume 31, issues 4 and 5, pages 351-694) and other journal articles discuss to the issues concerning hormesis. Some discussion can be found in Environment Canada (2003)

LOEC and NOEC

The Lowest Observed Effect Concentration is the lowest concentration out of the tested concentrations at which a statistically significant difference from the control group is observed. The No Observed Effect Concentration is the tested concentration just below the LOEC. They are obtained by hypothesis testing. . An example of NOEC and LOEC are illustrated in Figure 3.2.

Monotonic and Non-monotonic concentration-responses

In a monotonic concentration-response relationship, the true, underlying concentration-response relationship exhibits an increase or a decrease over the range of concentrations in the study. If the concentration-response is monotone and non-increasing, the location parameters (mean or median) would exhibit the following relationship: $\gamma_0 \geq \gamma_1 \geq \gamma_2 \geq \gamma_3 \geq. . . \geq \gamma_k$, where γ is the location parameter and 0, 1, 2,

…, k are the concentration groups. If the monotone relationship is non-decreasing, the inequalities are reversed. In a non-monotonic concentration-response relationship, the inequalities are not consistent across the concentrations.

NEC

The No Effect Concentration is a parameter in some concentration-response model. When these models are used, it has the interpretation of being the highest concentration in which the compound does not affect the endpoint, even after very long exposure to the compound. So the NEC equals the EC0 at infinite time.

Response

A response corresponds to an observed value of any endpoint.

Parametric and Non-parametric methods

Parametric methods assume that all the properties of the model are specified, except for the values of the parameters. For example, in classical analysis of variance (ANOVA) the residuals are assumed to follow a normal distribution with a mean of zero and some unknown variance (will be estimated). In Poisson regression, the response variable is assumed to follow a Poisson distribution (parameters to be estimated in the fitting process). Non-parametric methods make weaker assumptions2 about the shape of the distribution of the residuals, and the analysis is often based on ranks of the observations. For example, the non-parametric analogue to a two-sample t-test is the Mann-Whitney test, and non-parametric regression is often conducted using a variety of smoothing techniques.

Parsimony principle

The parsimony principle says that data should be described with as few parameters as possible. A common decision criterion of including more parameters in the model is the observation that such leads to a significantly better description of these data.

Systematic errors

The term systematic error is used for the situation that a single concentration (dose) groups differs from the others not only with respect to the intended treatment (i.e. the concentration or dose) but also with respect to some unintended experimental factor. For instance, containers housing the animals may differ by themselves, and in a design with few or only one container per dose group a deviating container may lead to a systematic error in that group. The factor of time may underly systematic errors in various ways, e.g. time of feeding, time of observation. The problem of systematic errors is that thay may be wrongly interpreted as an effect of the intended treatment.

Statistical significance

In hypothesis testing, a result is statistically significant at the chosen level α if the test statistic falls in the rejection region. The finding of statistical significance implies that the observed deviation from what was expected under the null hypothesis is unlikely to be attributable to chance variation. In this document, the α-level will be 0.05 unless otherwise stated.

[2] with fewer constraints

4. GENERAL STATISTICAL PRINCIPLES

4.1. Different statistical approaches

17. For each of the three analysis methods introduced below, it is necessary to obtain data from a designed experiment with replications of controls and concentration groups. All three classes of analysis methods (hypothesis testing, concentration-response modelling, and biology-based methods) are suitable for data from toxicity tests as currently standardised by several OECD and ISO guidelines. However, designs for each of these studies can be optimised with respect to cost-effectiveness and the selected analysis approach. The number and spacing of the concentrations will depend on the study being conducted and the type of data analysis to be utilised.

18. For each of the three approaches introduced below, the following is provided:

- A brief description of the use of each method in ecotoxicity tests.
- A brief outline of specific analysis methods presented in the later Chapters of this document.
- A listing of some major assumptions and limitations for each approach.

4.1.1. Hypothesis-testing methods

19. Hypothesis testing is a statistical inference technique used to compare the responses among two or more test groups. Hypothesis testing has many uses in ecotoxicology, ranging from detecting whether there is a significant difference in the measured response between the control and a given concentration, to establishing a LOEC and a NOEC. Discussion in this document focuses on use in determining LOECs and NOECs, the most frequent use of hypothesis testing in OECD guidelines.

20. Methods discussed in Chapter 5 include analyses for quantal data and continuous data. For both type of data, parametric approaches (when an underlying distribution e.g.: normal, lognormal is characterised) and non-parametric approaches (when weaker assumptions are made regarding the distribution) are presented. In chapter 5, assessment is limited to conducting data analysis separately at each time point, though this is not a limitation of the method. Three terms often used when discussing hypothesis tests are Type I errors, Type II errors, and power (Table 4.1). *Type I errors* (false positives) occur when the null hypothesis is the truth but the hypothesis test results in a rejection of the null hypothesis in favour of the alternative hypothesis. The probability of a making a Type I error is often referred to as α and is usually specified by the data analyst – often at 0.05, or 5%. *Type II errors* (false negatives) occur when the alternative hypothesis is true but the test fails to reject the null hypothesis (i.e., there is insufficient evidence to support the alternative hypothesis). The probability of a making a Type II error is often referred to as β (1 – power). *Power* is the probability of rejecting the null hypothesis (H_O) in favour of the alternative hypothesis (H_A), given that the alternative hypothesis is the true. Power of a test varies with sample size, variance of the measured response, the size of an effect that it is of interest to detect, and the choice of statistical test. Power to detect differences can be increased by increasing the sample size and/or reducing variation in the measured responses. Thus, if a test has low power to detect an effect of a given size, this is equivalent to saying that the test has a low probability of detecting an effect of that size.

Table 4.1 Probabilities of finding a significant or non-significant test outcome, given that the null hypothesis is true or not.

		State of the world	
		H_O true	**H_A true**
Result of hypothesis test	**not significant**	1- α	Type II error β
	significant	Type I error α	1 - β = power

21. Several assumptions made when conducting hypothesis tests to determine the NOEC are:

- Concentration-response relationship may or may not be assumed depending on the specific statistical tests used.
- This approach makes only weak assumptions about the mechanisms of the toxicant or the biology of the organism.

22. Several limitations of using hypothesis testing to determine the NOEC are:

- Since the NOEC (or NOEL) does not estimate a model parameter, a confidence interval cannot be assessed.
- The value of the NOEC is limited to being one of the tested concentrations (i.e., if different values were chosen for the tested concentrations, the value of the NOEC would be different).
- If power is low (due to high variability in the measured response and/or small sample size), the biologically important differences between the control and treatment groups may not be identified as significantly different. If power is high, it may occur that biologically unimportant differences are found to be statistically significantly different.

4.1.2. Concentration-response modelling methods

23. Regression methods are used to determine the relationship between a set of independent variables and a dependent variable. For designed experiments in ecotoxicology, the main independent variable is the concentration of the test substance and the dependent variable is the measured response (e.g., percent survival, fish length, growth rate). Regression methods fit a concentration-response curve to the data and use this curve to estimate an Effective Concentration (ECx) at a given time point. The mathematical model used may be any convenient function that is able to describe the data; however, some models are more frequently used and accepted within the ecotoxicity testing literature. Several methods are available for model fitting and parameter estimation.

24. Methods discussed in Chapter 6 include analyses for quantal data and continuous data. Parametric approaches (when a specific underlying distribution is assumed) are presented. Although non-parametric methods have been developed for fitting concentration-response curves and estimating an ECx, they are not presented in this document. Sources on non-parametric regression include Green and Silverman (1994), Easton and Peto, Fan and Gibjels (1996), Hardle, W. (1991), Azzalini and Bowman (1997), Silverman, B. (1985), Akritas and. Van Keilegom (2001), Carroll, *et al* (1999), Smith-Warner *et al* (1998). Software for non-parametric regression can be found, for example, at http://wwwstat.mathematik.uni-essen.de/~kovac/ftnonpar.html.

25. The effect of exposure time is also considered in chapter 6.

26. Although power is typically only discussed when hypothesis tests are conducted, both sample size and variation in the response variable within groups affect the inferences of concentration-response models as well. Small sample sizes and high variability in the response within groups will increase the width of the confidence interval of the parameters of interest (e.g., ECx), and the fitted model may not reflect the true concentration-response relationship. To increase the level of confidence in the parameter estimate, the number of replicates can be increased and measures to minimise unexplained variability could be taken. The width of the confidence interval also depends on the experimental design (i.e., the location and number of concentrations chosen). Finally, in addition to precision, accuracy of the estimated parameter is just as important. To enhance accuracy, concentrations should be chosen such that various different response levels are observed.

27. These specific properties of the experimental design, the number and spacing of doses and the number of replicates, are related to the value of X of interest in the particular experiment. Different designs may be employed to estimate an EC50, as opposed to an EC05. Further guidance for the design of experiments of this type is discussed in Chapter 6.

28. Several assumptions of concentration-response modelling are:

- The models discussed in this document assume the response have a monotonic concentration-response relationship.
- The fitted curve is close to the true concentration-response relationship.
- This is an empirical model and does not make strong assumptions about the mechanisms of the toxicant or the biology of the organism.

29. Several limitations of concentration-response modelling are:

- Estimation of ECx values outside the concentration range introduces a great deal of uncertainty (i.e., extrapolation outside the range of the data).
- Once the experiment has been performed, the resulting concentration-response data may not be suitable for the estimation of parameters of a concentration-response model. In particular, when the gaps between consecutive response levels are so large that many different concentration-response models would fit equally well to the observed data, interpolation would not be warranted.

4.1.3. Biology-based methods

30. The biology-based methods presented in this guidance provide models for exploring the effect of the test chemical over time as well as incorporating a toxicokinetic model for the behaviour of the chemical. By modelling concentration and exposure time simultaneously, these methods fit response surfaces to response data to estimate an ECx as a function of exposure time, rather than fitting separate response curves at each time point.

31. Methods discussed in Chapter 7 include analyses for quantal data and continuous data for several aquatic toxicity tests (acute and chronic tests on survival/immobility for daphnids and fish, fish growth test, daphnia reproduction test, and alga growth inhibition test). The models presented in this document utilise dynamic energy budget theory (see Chapter 7 for details and associated references). This theory provides a quantitative description for the processes of feeding, digestion, storage, maintenance, growth, development, reproduction and ageing and their interrelationships. As with concentration-response

modelling, the level of confidence in the parameter estimates (as evidenced by the width of the confidence interval) is a function of sample size and inherent variation in the response.

32. Because of additional assumptions regarding the toxicokinetic behaviour of the chemical and the biological behaviour of the organism in the system, it is sometimes possible to carry out additional extrapolation from the toxicity test. The assumptions are endpoint-specific; therefore, for each type of test, these assumptions need to be defined. The definition of these assumptions usually involves eco-physiological background-research prior to the specification of the test. However, if these additional assumptions can be made, an example of additional outcomes this method can predict are: chronic responses from acute responses, responses to time-varying concentrations using responses to constant concentrations, and responses by a species using responses to a conspecific or physiologically related species of a different body size for given test compound.

33. Several general assumptions made when using biology-based methods are:

- The models discussed in this document assume the response has a monotonic concentration-response relationship.
- This analysis method incorporates mechanistic models for toxicokinetics and physiology.

34. Several limitations of biology-based methods are:

- Estimation of parameter values (e.g., ECx and NEC) outside the concentration range introduces a great deal of uncertainty (i.e., extrapolation outside the range of the data).
- When the gaps between consecutive response levels are so large that different biology-based models would fit equally well to the observed data, NEC estimation would not be warranted, if they differ substantially between the models.
- To date, models have been developed for some of the common aquatic toxicity tests (acute and chronic tests on survival/immobility for daphnids and fish, fish growth test, daphnia reproduction test, and alga growth inhibition test). Nevertheless, these models can be applied to any test species.

4.2. Experimental design issues

35. The usual factors (independent variables) studied in ecotoxicity tests are concentration of the tested substance and duration of exposure. For the estimation of the effect at a given condition it is necessary to replicate these conditions, to control experimental variation (see section 4.2.2)

36. The experimental design will, amongst others, specify the tested concentrations of the substance, the number of replicates and number of containers per tested concentration as well as the times of observation.

NOEC, NEC or ECx: implications for design.

37. The estimation of an ECx puts different demands on the study design than does the assessment of a NOEC. When the aim is to assess a NOEC, an important demand is that the study warrants sufficient statistical power. To that end, the concentration (dose) groups need a sufficient number of replicates (possibly at the expense of the number of dose groups). Many test guidelines are based on this principle. When the aim is, however, to provide an estimate of an ECx, the primary demand on the study design is to have a sufficient number of concentration (dose) groups. This may be at the expense of the number of replicates per group (e.g. keeping the total size of the experiment the same), since the precision of the estimated ECx depends more on the number and spacing of concentrations rather than on the sample size

per concentration or dose group. The demands for study designs aimed at estimating a NOEC or an ECx are further discussed in sections 5.1.6 and 6.5, respectively.

38. Therefore, the choice between assessing a NOEC and estimating an ECx should actually be made before designing the study. If one wishes (or is required) to assess both, a compromise between the two opposing demands must be made, i.e. a design with both a sufficient number of dose groups and a sufficient number of replicates in each group. The number of replications per group needed for assessing a NOEC depends on the desired power of the statistical test involved (see section 4.1.1**)**. For assessing an ECx three concentration groups, next to the control group, is an absolute (theoretical) minimum. However, when just one dose group appeared to have been unluckily chosen, the assessment of an ECx would probably fail, and more concentration groups are therefore required in practice. Design recommendations for experiments using a biologically based model include those for ECx. Additional recommendations are discussed in section 7.8.3.

4.2.1. Randomisation

39. Variability is inherent in any biological data set. This variability is directly visible in continuous and discrete data. Although the following discussion holds for any type of data, it is easiest to use continuous data as an example. In analysing concentration-response data by statistical methods, the observed scatter is sometimes called noise or variation, but when designing experiments and interpreting results it is good to understand the reasons for the noise. The following factors may play a role:

- the variation between the individual animals, due to genetic differences,
- the differences in the conditions under which the animals grew up prior to the experiment, resulting in epigenetic differences between animals,
- the heterogeneity of the experimental conditions among the animals during the experiment,
- variation within subjects (i.e., fluctuations in time, such as female hormones, which may be substantial for some endpoints), and
- measurement errors.

40. Randomisation processes are used in designed experiments to eliminate bias in estimates of treatment effects, and to ensure independence of error terms in statistical models. Ideally, randomisation should be used at every stage of the experimental process, from selection of experimental material and application of treatments, to measurement of responses. To minimise the effects of the first two factors, animals need to be randomly distributed into concentration groups. To minimise the effects of the third factor (both intended and unintended, such as location in the room), application of treatments should be randomised as much as possible. To minimise the effects of the fourth factor, the measurement of responses should be randomised in time (e.g., although all responses will be recorded at 24 hrs, the order in which the experimental units are measured should be randomised). With good scientific methods, measurement errors can be minimised.

41. In some circumstances it may be difficult, or expensive, to randomise at every stage in an experiment. If any experimental processes are carried out in a non-random way, then statistical analysis of the experimental data should include a phase in which the potential effect of not randomising on the experimental results is examined.

4.2.2. Replication

42. As discussed above, noise cannot be avoided, and therefore it is necessary to assign a certain number of replicates (experimental units) to each treatment group and control group. The number of

replicates influences the power in hypothesis testing and the confidence limits of parameter estimates. A standard assumption of all methods is that replicates are independent. Treating observations as independent replicates whereas in fact they are not, represents an error called pseudoreplication (Hurlbert , 1984). This issue becomes important when animals are housed together, as in a tank or beaker. There are two types of housing effects:

- containers may differ from each other in some (usually unknown) sense
- the organisms within a container affect each other's responses.

43. Both effects result in non-independence (or pseudoreplication) of the individual organisms' responses. The first effect may result in different mean responses between containers (at a given concentration). This type of non-independence can be addressed by taking the variation between containers into account in the statistical model. For instance, with continuous data this may be done using a nested ANOVA, where the individual observations are nested within the container. The second effect might distort the distribution of the observations related to the individual organisms. For instance, with quantal data the assumption of binomial distribution may not hold. In an example with continuous data when there is competition among individuals in the same container, the responses of the individual organisms may appear bimodal. See Chapter 5 and 6 for a more detailed discussion.

4.2.3. Multiple controls included in the experimental design

44. It is common in aquatic and certain other types of experiments that the chemical under investigation cannot be administered successfully without the addition of a solvent or vehicle. In such experiments, it is customary to include two control groups. One of these control groups receives only what is in the natural laboratory environment (e.g., dilution water), while the other group receives the dilution water with added solvent but no test chemical. In ecotoxicity experiments, these are often termed negative or dilution water (non-solvent) and solvent controls. OECD recommends limiting the use of solvents (OECD, 2000); however, appropriate use of solvents should be evaluated on a case-by-case basis. Details regarding the use of solvents (e.g., recommended chemicals, maximum concentrations) are discussed in the relevant guideline documents for a specific ecotoxicity test. In addition, regulatory guidelines must be followed by both controls with regard to the range of acceptable values (e.g., minimum acceptable percent survival or mean oyster shell deposition rate). Multiple control groups can be utilised regardless of whether the experiment was intended for hypothesis testing (i.e., determination of a NOEC), regression analysis (i.e., determination of an ECx), or biology-based methods. The focus of this section is to present data analysis methodology for experiments in which a solvent is used.

45. Data from the two control groups are analysed to determine if the solvent had a statistically significant effect on the measured response. If there was a statistically significant difference between the negative and solvent control groups, any conclusions and inferences based on this study could be impacted due to presence of a solvent effect. If there are no significant differences in the means (or proportions for quantal data or medians for non-parametric data) between the negative and the solvent controls, then it is concluded that there is insufficient evidence to detect a difference between the controls.

46. The solvent control group is the appropriate control group for comparisons with treated groups. Each group must have the same solvent concentration as the control. For a toxicity test in which a solvent is used in conjunction with the test chemical, the assumptions are that the solvent had no effect on the responses of interest and there was no interaction between the test chemical and the solvent. With the addition of a negative control (as is required in all experiments using a solvent), the assumption regarding a solvent effect can be tested. However, unless the chemical is also tested in absence of a solvent, the assumption of no interaction between the solvent and the test chemical cannot be evaluated.

47. Some practitioners consider combining the data into one ‘pooled control’ for comparison to the treated groups when no statistically significant differences between the solvent and negative control were identified. However, this does not take into account the fact that the differences between the two controls might not have been detected with a statistical test because the sample sizes are too small (i.e., low power) or that it combines two sources of variability.

48. The methods used for statistical comparison of negative and solvent controls vary depending on the type of data and the assumptions regarding distribution of the data. Methods and mathematical details for carrying out these tests are found in Chapter 5 and its associated Annexes.

4.3. Process of data analysis

49. A typical data analysis more or less follows a general pattern, usually constituting the following steps. First, the data are plotted and visually inspected. Then, a suitable type of analysis is chosen, based on particular assumptions that appear reasonable for the data at hand. After the analysis the underlying assumptions are checked. If necessary, an adjusted analysis is performed. And finally, the results are reported by making plots and/or tables.

4.3.1. Data inspection and outliers

50. A useful first step in analysing dose-response data is to visually inspect the data. For continuous data, the individual responses (together with the group means) may be plotted as a function of dose. For quantal data, one may plot the observed frequencies of response as a function of dose. These plots are useful to assess whether the data show a dose-response relationship. Further, these plots may indicate any peculiarities in the data. In particular, the observed data may show outliers, i.e. data points far away from intuitive expectation, or from the general pattern shown by the data. In continuous data one may detect both outliers that relate to the individual organism (or, more generally, the biological system serving as the experimental unit), and outliers that relate to a whole treatment group. In quantal data, outliers always relate to a treatment group, since a deviating individual cannot be detected based on a “yes” or “no” response.

51. Outliers that relate to a whole treatment group may arise due to the fact that a treatment group differed systematically from the other groups by some (usually unknown) experimental factor(s). For instance, the organisms in the various dose groups were held in different aquaria, and one of them contained an infection. Or, the organisms in the different dose groups were treated in a specific order (with respect to feeding, time of observation, etc). Detection of this type of outliers typically cannot be enforced by any formal statistical method, and one has to rely on visual inspection, judgement and experience.

52. Obviously, treatment group outliers are highly undesirable, since they directly interfere with the effect that one wishes to measure, thereby increasing the probability of both false positive and false negative results. For example, a NOEC may be assessed at a level where substantial effects do occur, or a LOEC may be assessed at a level without real effects (i.e. from the chemical). The only way to prevent outliers at the group level is a design that is perfectly randomised with respect to all experimental actions that may potentially influence the (observed response of) the biological system. In practical biological studies, however, perfect randomisation is hard to realise, and it is not feasible to reduce the probability of getting group outliers to nil. Therefore, it is paramount to make the study design relatively insensitive to potential outliers, i.e., by randomised replicated dose groups, and/or by increasing the number of different doses (followed by dose-response modelling, see chapter 6).

53. Outliers at the individual level can only be detected in continuous data. When a particular distribution is assumed for the scatter in the data, the judgement of outliers may be based on a specific,

small probability that any single data point could occur. This implies that the judgement of outliers can depend quite strongly on the assumed distribution. For example, values that appear to be extremely high when assuming a normal distribution may be judged as non-extreme when assuming a lognormal distribution. Vice versa, low values may be judged as extremes when assuming a lognormal distribution, but not so when a normal distribution is assumed.

54. The statistical analysis of the data is sensitive to individual outliers, although less dramatically than to group outliers. On the one hand, individual outliers may result in biased estimates of the effect (either too small or too large). On the other hand, the estimate of the residual variance (the "noise") will be increased, implying that statistical tests tend to be less powerful, and estimated parameters (e.g. ECx) less precise. Therefore, if reasons can be found explaining the outliers, it is favourable to delete them from the analysis.

55. Although non-detectable, individual outliers can also occur in quantal data and affect the analysis. For example, when just one of the individuals in the controls shows a response, but is in fact an outlier, this outlier may have quite an impact on the statistical analysis. Being non-detectable, individual outliers are a larger problem in quantal than in continuous data.

56. In conclusion, outliers can have dramatic effects on the statistical analysis and the conclusions drawn. Therefore, it is very important to reduce their impact by using designs that are relatively insensitive to them, i.e. by utilising replicated dose groups and/or multiple dose groups. More information can be found in Atkinson (1985), Belsey et al. (1980) and Cook and Weisberg (1982).

4.3.2. Data inspection and assumptions

57. Visual inspection may also be used to explore the general pattern of the scatter around (continuous) data. Thus, one may find out if the scatter around the mean response appears to be symmetrical or skew, and if the scatter is more or less homogenous. Heterogeneity of variance (scatter) may have a biological basis i.e. the individual organisms (units) respond differently to the chemical. However, an apparent increase of decrease in the scatter may also be related to the statistical distribution of the data, e.g. the scatter increases with the mean response. This distinction is important, both for the analysis, and for the interpretation of results, and some further clarification is given below.

Heterogeneous variances and distribution

58. When the plotted data show scatter that is correlated with the mean response, such pattern may be related to the underlying distribution of the data. Some examples will illustrate this.

59. In lognormally distributed data, it may be theoretically expected that the standard deviation increases proportionally with the mean (or, equivalently, the Coefficient of Variation, CV, is homogenous). Also, for the gamma distribution, the CV is expected to be homogenous. When a particular dataset (such as weights, concentrations) show scatter that increases with the mean, one may plot such data on the log-scale, which usually makes the scatter independent from the means. In addition, when the scatter is relatively large (say, CV larger than 20%), the scatter may be skewed on the original scale, but not on the log-scale. The latter would confirm that the pattern in the scatter is a result from the underlying distribution.

60. As another example, counts may follow a Poisson distribution. Here, the variances are expected to be equal to the means (or, proportional to them). Such a pattern should vanish when the data are plotted on square root scale.

61. Finally, in quantal data with replicated dose groups, it can be also be expected *a priori* that the scatter between the replicates depends on the mean response (this follows directly from the properties of frequencies). Here, one may plot the frequencies after the transformation arcsin($\sqrt{p}$), where p is the observed frequency (fraction). This transformation is able to remove the (theoretical) relationship between the variance and the mean frequency (assuming a binomial distribution).

Heterogeneous variances and true variation in response

62. Heterogeneity in the scatter might also be caused by the treatment (the applied chemical) itself, i.e. some individuals respond stronger to the chemical than others. This could happen when genetically heterogeneous organisms are used, e.g. subject to genetic polymorphism. In many toxicity tests, however, the organisms used are genetically homogenous, and real (biological) heterogeneity in response to the chemical will, in those cases, not be very likely.

Consequences for the analysis

63. Heterogeneity of variances may be a matter of scaling that can be removed by the right transformation. Usually such a transformation also tends to make the data more normally distributed. Thus, one may apply standard methods based on normality (e.g. t-test, ANOVA, linear regression) to the transformed data. Another approach is to omit the transformation, and use methods that are directly based on the assumed distribution (i.e. generalised linear models). When a particular transformation is found that results in homogenous variances, only one variance parameter needs to be estimated. Thus, all the data contribute in the variance estimate, which is in statistical terms reflected by a larger number of degrees of freedom[3].

64. However, when the heterogeneity of variances appears to be due to real biological heterogeneity in responses among individual organisms, one should carefully consider if further analysis is meaningful. For example, when the organisms (or experimental units) consist of two distinct subpopulations, one responding, the other not, any estimated change in mean response has no useful meaning. When such two subpopulations can be distinguished from observable features (e.g. sex), the appropriate way to proceed is to analyse both subpopulations separately, or by using the observable feature as a covariate (see, e.g. section 6.3.2, and Fig. 6.9)

4.3.3. Transformation of data

65. Many standard parametric methods (e.g. ANOVA, t-tests, linear regression analysis) assume normally distributed data and homogenous variances. In practice, the data often deviate from these assumptions, and if so, the inferences resulting from these standard methods may be disturbed. A variance-stabilising transformation is often applied to the data, and then the statistical analysis is performed on the transformed data. Examination of residual plots (plot of the residuals vs. the predicted values) and tests of heterogeneity of variance (e.g., Levene, Bartlett, Hartley's F-max, or Cochran's Q) can help to identify instances when the variances among the concentration groups are unequal. References on this topic include Box and Cox (1964), Box and Hill (1974), Box and Tidwell (1962), Draper and Cox (1969).

66. For a variance-stabilising transformation to exist, there must be a relationship between the population means and variances. In many cases, the theoretical distribution of the response variable can guide the choice of a transformation. For example, if the underlying distribution is assumed to be Poisson,

[3] In general, it is favourable to include as few parameters (that need to be estimated from the data) as possible in the analysis, and yet describe the data accurately. Too few parameters will probably result in biased estimates, too many parameters tend to result in too wide confidence intervals. This is also referred to as the parsimony principle.

the square root transformation, $y_i' = (y_i^{½})$ or $y_i' = ((y_i + 1)^{½})$, is used. If the underlying data are lognormal, the log-transformation, $y_i' = \log(y_i)$, is often used. For proportions with binomial distributions, the arc-sin square-root $[y_i'=\arcsin(y_i^{½})]$ and Freeman-Tukey $[y_i'=(y_i+1)^{½}+ (y_i^{½})]$ transformations are often used. If the underlying theoretical distribution is unknown, a data-based procedure (Box-Cox transformation) can be used (Box and Cox ,1964).

67. The use of transformations will often simplify the data analysis, in that the more familiar and traditional data analysis methods can be used, but care must be taken in interpreting the results of this data analysis. Several aspects are discussed below.

68. If a transformation is used, it is also necessary to back-transform the means and confidence intervals to the original scale, when reporting results. It is not correct to back-transform the standard errors. It is important to understand that the back-transformed means differ from the arithmetic means of the original data. These back-transformed means should be interpreted as estimates of the median of the underlying data distribution, if the transformed data are normally (or at least symmetrically) distributed. In the special case of a log-transformation, the back-transformed mean is the geometric mean of the original data, and this value estimates the median of the underlying lognormal distribution.

69. When a transformation is not used in the data analysis, the difference in the means is a logical measure for the size of an effect. This difference is interpreted as an absolute change in the original units (e.g., a decrease of 1.2 grams). The back-transformed difference in means (of the transformed data) however has another, usually more difficult, interpretation. In the special case of a log-transformation, the difference between the back-transformed means does allow a simple interpretation: it estimates the ratio (or percent change) of the median responses.

70. In addition, transformations may not maintain additivity of effects (interactions among factors, e.g., test substance, sex, age, in the experiment). Other possible consequences of using transformations are that they change the interpretation of outliers and that they affect the value of r (Pearson correlation coefficient) and R^2. Not all data problems can be fixed by transformation of the response. For example, if a large percentage of the responses have the same measured value (ties), no transformation will address that issue.

4.3.4. Parametric and non-parametric methods

71. A visual inspection of the data may have indicated that the scatter is more or less symmetric and homogeneous, possibly after a particular transformation. In that case, one may analyse the data by the standard parametric methods based on normality. Or, one may choose to analyse the data based on a particular distribution other than the normal. Here, some basic aspects of parametric and nonparametric methods are discussed.

Parametric Methods

72. When the data are assumed to follow a particular statistical distribution, they can be summarised by the parameters of that distribution. For example, data that are normally distributed can be summarised by just two parameters, the mean and the variance. Therefore, methods that are based on an assumed distribution are called “parametric methods”. Obviously, these methods intend to estimate the parameters of the (assumed) distribution, such as the mean and the variance, or any derived parameters, such as the ECx.

73. If one is interested in the value of some entity (such as the ECx), rather than a “categorical” answer (significant or nonsignificant), parametric methods are the natural approach of analysis. In addition, in hypothesis testing parametric methods such as ANOVA are also widely used.

74. Whatever distribution is assumed, parametric methods are based on the general principle of fitting the data to the model. In hypothesis testing this may be the ANOVA model, in dose-response modelling this may be a particular dose-response model. In applying parametric hypothesis tests, one must examine the data for outliers, deviations from normality and homogeneity, assessment of monotonicity of the dose-response (for some approaches), and do a general assessment of whether the proposed model adequately describes the data. These points are discussed in depth in 5.1, 5.1.3, 5.1.5, 5.2.2.2, 5.2.2.4, 5.3.1, 5.3.1.2, 5.3.1.2, 5.3.1.5, and 5.3.1.6. In dose-response modelling, the process of model fitting is eminent and indeed the focus of the analysis. Therefore, in any dose-response analysis (as discussed in Chapters 5, 6 and 7) the user should understand the general principles of model fitting. These are discussed in section 4.3.5, 5.1.3, 5.2.2, 5.3.1, and 6.7.

GLMs

75. Generalised linear models are an alternative approach to use parametric methods when the normality assumption is violated. In this approach, the analysis of the (untransformed) data is based on another (than normal) distribution, for example, a Poisson distribution (for counts), or a binomial distribution for frequencies. GLMs are not discussed in this document, and the reader is referred to the literature (Mc Cullagh and Nelder, 1983; Kerr and Meador, 1996).

Nonparametric methods

76. Nonparametric methods have been developed for those cases where one is not willing to assume any distribution at all. These methods can be used to test the null hypothesis that the observations in two (or more) treatment groups do not differ (i.e., they stem from the same, but unknown distribution). These methods are based on the rank order of the observations. Therefore, significantly different treatment groups are supposed to differ in the medians (since the median can be defined in terms of rank order). To prevent misunderstanding, the medians should always be reported when nonparametric methods were used (differences between means may not be consistent with differences between medians; i.e. means are more sensitive to outliers than medians are).

How to choose?

77. Parametric analyses have various advantages over nonparametric methods. They are typically simpler to conduct (wide availability of software), methods have been developed for a wider array of designs (e.g. designs with replicated dose groups), confidence intervals are more easily computed, the methods are more universally used, and interpretation of results is often easier. The advantage of non-parametric methods is that they are based on very weak assumptions. Further, since nonparametric analyses are based on the rank order of the data, they are less sensitive to outliers than parametric analyses.

78. When the data appear to comply with the assumptions (after a visual inspection) of a particular parametric analysis, parametric is the obvious method to choose. The assumptions can be further checked as part of the analysis (e.g. by examining the residuals, see below). It may be noted that parametric analysis based on normal assumptions is reasonably robust to mild violations against the assumptions. When a data transformation results in a (better) compliance with the normality assumptions, one should be reminded that transformations other than the log-transformation may impair the interpretation of the results. This is because the log-transformation is naturally linked to the intuitive notion that biological effects are proportional (or multiplicative) rather than additive (compare definition of ECx). Thus, when omitting or applying a log-transformation does not make a large difference for complying with the assumptions, one might yet choose to apply it for reasons of interpretation.

79. In situations where specific distributions are natural candidates for the data type at hand, one may consider the use of GLMs. When not any regular distribution can be assumed, as e.g. in the case of tied observations4, one may resort to nonparametric analysis.

4.3.5. Pre-treatment of data

80. In general, pre-treatment of data (other than data transformation) is not a favourable strategy of data analysis. A few practical examples will be discussed.

81. Some methods (e.g. probit and logit model for quantal dose-response analysis) use a log-transformation for concentration. It is not appropriate to add a small positive constant to the zero-concentration (or to all concentrations) to avoid taking the log of zero (see chapter 6 for more details): the shape of the concentration-response curve is very sensitive to the constant and a biological basis for choosing one constant over another is very unlikely to ever exist.

82. A current habit in analysing continuous data is to divide the observed responses by the (mean) observed response in the controls. These corrected observations then reflect the percent change compared to the controls, which is usually the entity of interest. However, such a pre-treatment of the data is improper: Among other problems it assumes that the (mean) response in the controls is known without error, which is not the case. Therefore, this should be avoided, and instead the background response should be estimated from the data by fitting the model to the untreated data. Thus, the estimation error in the controls is treated in the same way as the estimation errors in the other concentration groups. (see e.g. chapter 6.2.2 and 6.3.2).

4.3.6. Model fitting

83. All parametric methods employ the general principle of model fitting. The particular assumptions they are based on can be regarded as a particular model. The model contains specific parameters, and the goal of the data analysis is to estimate these parameters. The parameters are estimated by fitting the model to the data.

84. As a very simple example, consider a single sample of data. If it is assumed that these data follow a normal distribution, than the model is simply the normal distribution. The model contains two parameters, the mean and the variance. Depending on what values are chosen for the parameters, the agreement between the distribution and the data will be better or worse. The question now is what parameter values give the best agreement between the model and the data, i.e. gives the best fit of the model to the data.

85. To be able to answer the latter question, we have to define a measure for the “distance” between data and model, to be used as the fit criteria. A very general criterion is the likelihood. This measure directly follows from the assumed distribution, and is applicable to whatever distribution is assumed. The likelihood criteria should be maximised, and when this is achieved, the associated parameter values are called maximum likelihood estimates.

86. Another much used fit criterion is the residual Sum of Squares (SS). This measure is defined as the sum of the squared residuals, i.e. the differences of each separate observed response with its associated expected response (according to the model). The best fit is found by minimising the SS. In the simple example of fitting a normal distribution to a single dataset, the residuals are simply the differences of the observations to the mean. By changing the value of the mean, the SS will vary.

[4] Tied data are two or more observations of the same value. Parametric methods do exist for tied data, but these are beyond the scope of this document.

87. The value of the mean resulting in the best fit, is exactly the value of the (arithmetic) sample mean. Put another way, the sample mean is the estimate of the mean of a normal distribution that results in the best fit according to the SS criteria. In the special case of a normal distribution, the sample mean is at the same time the maximum likelihood estimate. In other distributions however, maximum likelihood or minimising the SS results in different estimates of the parameters. For instance, for quantal dose-response data, the sum of squares is not appropriate, and the likelihood is the usual fit criteria.

88. The same principle of model fitting holds for more complicated models than a single dataset. For example, by replacing the mean of the normal distribution by a function of the dose we obtain a dose-response model. Here, fitting the model by minimising the SS or by maximising the likelihood results in the same fit (because of the normality assumption).

89. A general method of finding the best fit is by trial and error, i.e. in an iterative search one tries to improve the likelihood by changing the parameter values, until an improvement cannot be found anymore. General algorithms exist that perform such an iterative search in an efficient way. In particular models ("linear" models) the maximum likelihood estimates can be derived from explicit formulae, and search algorithms are not required (for that reason linear models used to be popular before the availability of computers). In nonlinear models search algorithms can hardly be avoided. Although the user need not worry about the calculations underlying these algorithms, fitting nonlinear models does require some basic understanding of the general principles of search algorithms (see section 6.7).

4.3.7. Model checking

Analysis of residuals

90. After a model has been fitted to the data, a final check for the appropriateness of the fitted model may be performed. Do the data indeed comply with the model assumptions? For instance, do the data comply with the assumed distribution (in parametric analyses), are the variances homogenous (e.g. in ANOVA), and is the dose-response model suitable for the dose-response data at hand (in dose-response analysis).

91. A general approach for checking such assumptions is the analysis of residuals: the differences between the observations and the value predicted by model. For instance, in ANOVA the predicted value is the associated group mean, while in dose-response modelling it is the value of the model at the relevant dose.

92. To check the distribution, the residuals can be taken together and be plotted in a single histogram, or in a (distribution-specific) QQ-plot[5]. Visual inspection of such plots may reveal deviations from the assumed distribution, in particular when inspecting a QQ-plot, which should be linear if the data comply with the assumed distribution. Formal tests exist as well (see chapter 5), but it should be noted that mild violation of the assumptions is no reason for concern, and tests do not measure the degree of violation.

93. Various other plots of the residuals can be made, e.g.

- against the predicted value (i.e. the group means, usually), to check if the variances are homogenous (if such were assumed)
- against the model prediction, to check for systematic deviations from the fitted model

[5] QQ-plot corresponds to plots of observed quantiles versus expected quantiles.

- against other experimental factors, if relevant, for instance the order in time by which the observations were made. Such a plot may show if the pertinent factor influenced the response systematically.

94. Finally, one may perform a formal goodness of fit test. This test is sensitive for all the assumptions simultaneously. A significant overall test of goodness of fit may indicate that one of the assumptions is not met, but this does not necessarily imply that the model is not useful for the particular purpose of the analysis. Here, one should judge the nature of the violated assumption and its potential impact on the results one is interested in. On the other hand, a nonsignificant goodness of fit test does not imply that the model used may be regarded as reliable. The test is more easily passed when the data contain relatively little information, and, as a result, various models may pass the goodness of fit test, but lead to different conclusions. In other words, not only the model, but also the data should be "validated", by asking the question of whether they contain the information needed for answering the question of interest. The evaluation of a dose-response model for describing a particular dataset is more fully discussed in section 6.4.

Validation of fitted dose-response model

95. In dose-reponse modelling, it may happen that the data appear to be unsuitable for that approach. This would happen if the dose-response information is too weak to have faith in any fitted dose-response model (see section 6.4). For example, there may be large gaps between response levels or too few dose levels in areas where the response changes rapidly. Therefore, the estimation of an ECx is only warranted if the dose-response data contain sufficient information on the shape of the dose-response relationship.

4.3.8. Reporting the results

96. The final step in a statistical analysis is reporting the results. Basically, two types of information should be given: the results of the analysis, and the justification of the methods (assumptions) used.

97. The results of the analysis typically exist of summarising statistics, in current practice these are usually the means and standard deviations (or standard errors) per dose group. This may not generally be the best way of reporting results, however. When a parametric analysis assuming homogenous variances is applied, it is more informative to report the estimate of the common variance (residual Mean Square), together with a justification of the homogeneity assumption (e.g. a plot of the individual data or of the residuals against dose). When a log-transformation is applied before the analysis, it is more adequate to report the geometric means, and the (possible common) geometric standard deviation (GSD) or Coefficient of Variation (CV).

98. When a NOEC is assessed, the associated test used should be reported, along with the test outcomes.

99. In the case where an ECx is assessed, the fitted model should be reported, as well as the justification that the model was acceptable for assessing the ECx (see section 6.4).

100. More specific guidance for reporting results are given at the end of chapters 5, 6 and 7.

5. HYPOTHESIS TESTING

5.1. Introduction

101. This chapter provides an overview of both hypothesis testing and methodological issues specific to determining NOECs under various experimental scenarios. It is divided into three major parts. The first part includes flow charts summarising possible schemes for analysing quantal (Fig. 5.1) and continuous data (Fig. 5.2 and 5.3), along with some basic concepts that are important to the understanding of hypothesis testing and its use in the determination of NOECs. Special attention is given to the choice of the hypothesis to be tested, as this choice may vary depending on whether or not a simple dose-response trend is expected, and on whether increases, or decreases (or both) in response are of concern. The remainder of the chapter is divided into two major sections that discuss statistical issues related to the determination of NOECs for quantal and continuous data (Sections 5.2 and 5.3 respectively) and provide further details on the methods listed in Figures 5.1 and 5.2.). This division reflects the fact that different statistical methods are required for each type of data, and that problems arise that are unique to the analysis of each type of data. An attempt has been made to mention the most widely used statistical methods, but to focus on a set of methods that combine desirable statistical properties with reasonable simplicity. For a given set of circumstances, more than one statistical approach may be acceptable, and in such cases the methods are described, the limitations and advantages of each are given, and the choice is left to the reader. The flow charts in Figures 5.1 and 5.2 indicate a possible choice of methods. Examples of the application of many of these methods, mathematical details and properties of the methods are presented in Annex 5.1.

102. The most commonly used methods for determining the NOEC are not necessarily the best. Relatively modest changes in current procedures for determining NOECs (e.g., selection of more powerful or biologically more plausible statistical methods) can improve the scientific basis for conclusions, and result in conclusions that are more protective of both the environment and business interests. Thus, some of the methods recommended may be unfamiliar to some readers, but all of the recommended methods should be compatible with current ISO and OECD guidelines that require the determination of NOECs.

103. A basic principle in selecting statistical methods is to attempt to use underlying statistical models that are consistent with the actual experimental design and underlying biology. This principle has historically been tempered by widely adopted conventions. For example, it is traditional in ecotoxicological studies to analyse the same response measured at different time points separately by time point, although in many cases unified analysis methods may be available. It is not the purpose of this section to explore this issue. Instead, discussion will be restricted to the most appropriate analysis of a response at a single time point and, usually, for a single sex.

104. NOECs, as defined and discussed in this document, are based on a concept sometimes called "proof of hazard". In essence, the test substance is presumed non-toxic unless the data presents sufficient evidence to conclude toxicity. Alternative approaches to assessing toxicity through hypothesis testing exist. For example Tamhane *et al* (2001) and Hothorn and Hauschke (2000) develop an approach based on proof of non-hazard. Specifically, if an acceptable threshold of effect is specified, such as a 20% decrease in mean, then the maximum safe dose (MAXSD in Tamhane *et al* (2001)) is the highest concentration for which there is significant evidence that the mean effect is less than 20%. These are relatively new approaches that have not been thoroughly tested in a practical setting and for few endpoints is there agreement on what level of effect is biologically important to detect. All current guidelines regarding NOEC are based on the proof of hazard concept. For these reasons, this alternative approach will not be presented in this chapter, though they do hold some promise for the future. The only common exception to this is in regard to limit tests, where in addition to determining whether there is a statistically significant

effect in the single test concentration, one also tests for whether the effect in the test concentration is less than 50%. A simple t-test can be used for that purpose.

105. It should also be realized that statistics and statistical significance cannot be solely viewed as representative of biological significance. There can be no argument that statistical significance (or lack thereof) depends on many factors in addition to the magnitude of effect at a given concentration. Statistics is a tool that is used to aid in the determination of what is biologically significant. If an observed effect is not statistically significant, the basis for deciding it is nonetheless biologically significant is, obviously, not statistical. Lack of statistical significance may be because of a low power test. On the other hand, a judgment of biological significance without sufficient data to back it up is questionable.

106. The flow-charts and methodology presented indicate preliminary assessment of data to help guide the analysis. For example, assessments of normality, variance homogeneity, and dose-response monotonicity are advocated routinely. Such preliminary assessments do affect the power characteristics of the subsequent tests. The alternative to making these assessments is to ignore the characteristics of the data to be analyzed. Such an approach can be motivated on the perceived general characteristics of each endpoint. However, this does not avoid the penalty of sometimes using a low power or inappropriate method when the data do not conform to expectation. A bias of this chapter is to examine the data to be analyzed and use this examination to guide the selection of formal test to be applied. The preliminary assessment can be through formal tests or informed by expert judgment or some combination of the two. Certainly expert judgment should be employed whenever feasible, and when used, is invaluable to sound statistical analysis. These charts provide guidance, but sound statistical judgment will sometimes lead to departures from the flowcharts.

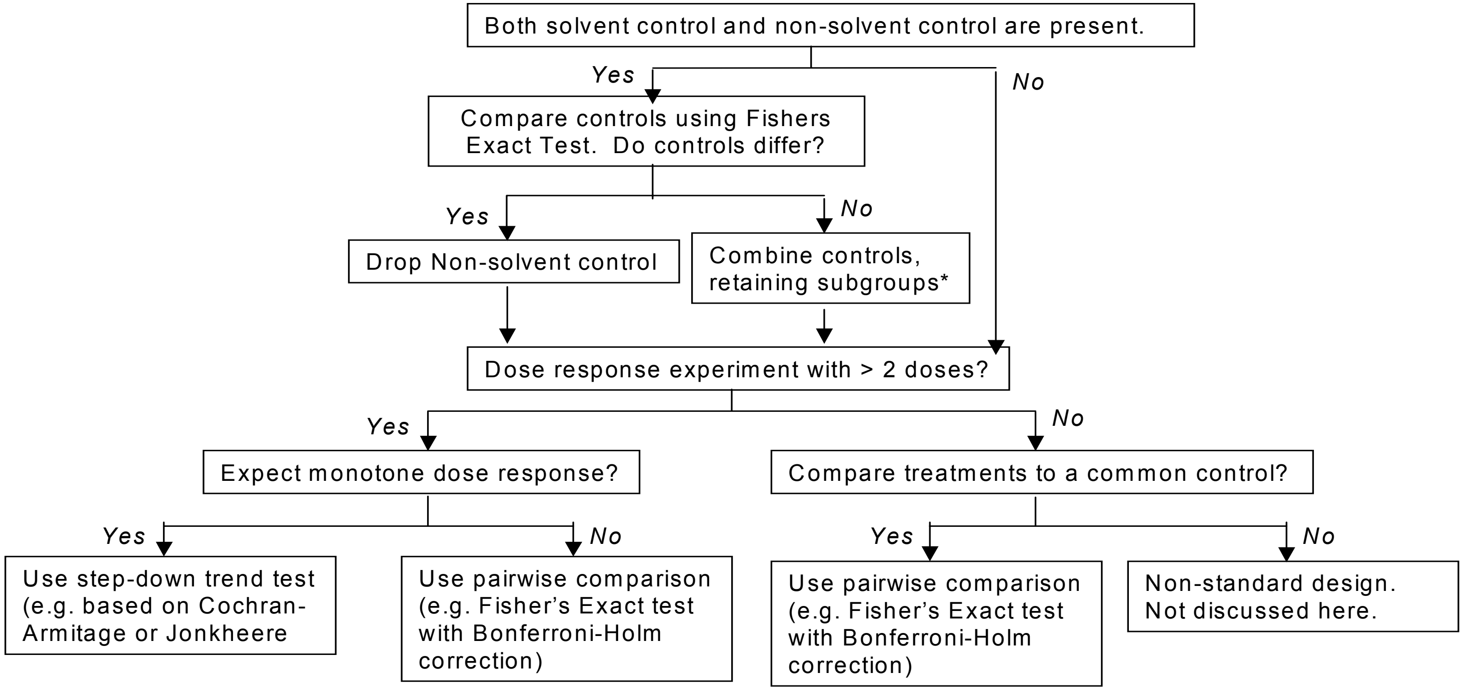
107. The flow charts (Figures 5.1 and 5.2) are intended to include experiments which contain only two concentrations (control and one test concentration). Such experiments are generally referred to as limit tests and the methods described are applicable to these tests.

108. It should be noted that tests of hypotheses might also be required for various special-case assessments of study results (e.g., use of a contingency table to assess the significance of male-female differences in frequency of responses at some dose). These types of analyses are beyond the scope of this document.

109. The terms “dose” and “concentration” are used interchangeably in this chapter and the control is a zero dose or zero concentration group. Consistent with this, the terms “doses” and “concentrations” include the control, so that, for example, an experiment with only two concentrations has one control group and one positive concentration group.

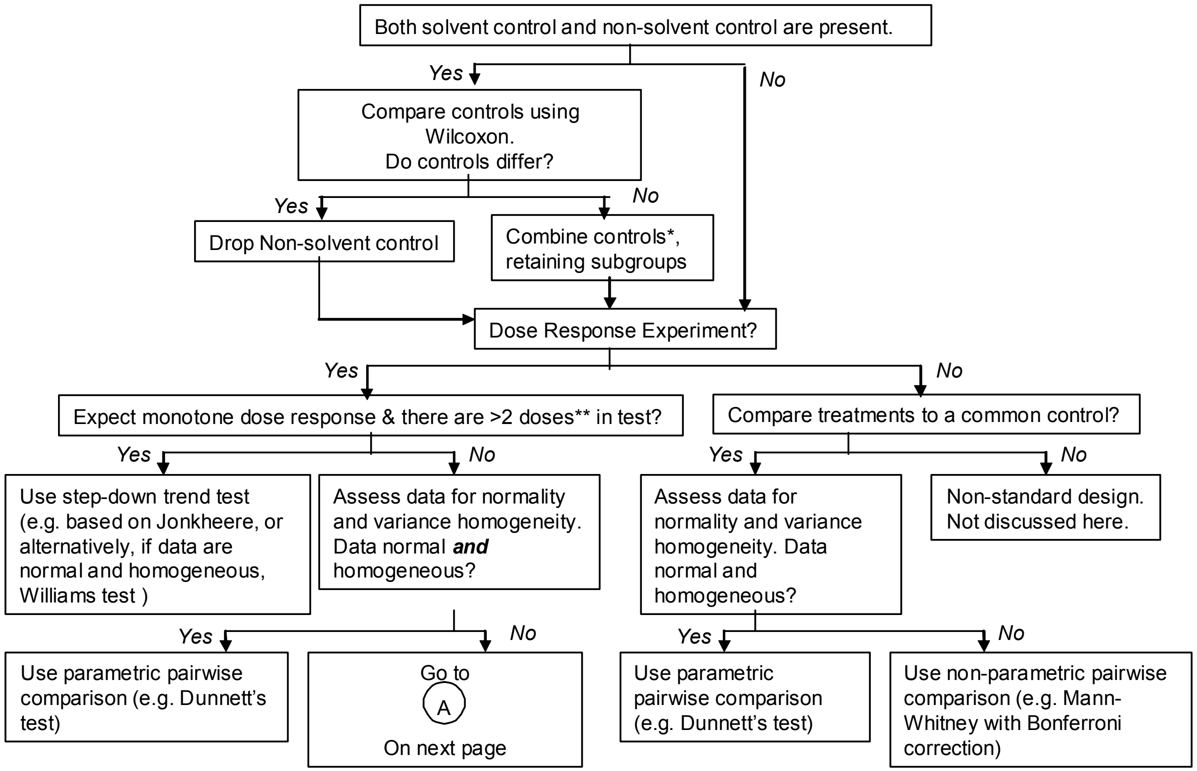
110. The tests discussed in this chapter, with the exception of the Tamhane-Dunnett and Dunn tests, are all available in commercial software. For example, they are available in SAS version 8 and higher. The two-sided Tamhane-Dunnett test (though not called such) is available in SAS through the studentized maximum modulus distribution provided by the probmc function. Where these tests are discussed, alternatives are provided, so that the reader can follow the general guidance of this chapter without being forced to develop special programs.

111. It will be observed that there is no special flow chart for the exact Jonckheere-Terpstra and exact Wilcoxon tests. One of the appealing features of these two tests is that there are both asymptotic and exact versions and the same logic applies to both.


Both solvent control and non-solvent control are present.
Yes
No
Compare controls using Fishers Exact Test. Do controls differ?
Yes
No
Drop Non-solvent control
Combine controls, retaining subgroups*
Dose response experiment with > 2 doses?
Yes
No
Expect monotone dose response?
Compare treatments to a common control?
Yes
No
Yes
No
Use step-down trend test (e.g. based on Cochran-Armitage or Jonkheere
Use pairwise comparison (e.g. Fisher's Exact test with Bonferroni-Holm correction)
Use pairwise comparison (e.g. Fisher's Exact test with Bonferroni-Holm correction)
Non-standard design. Not discussed here.

* Both scientific judgment and regulatory guidance must be considered in deciding whether to pool non-solvent and solvent controls. The flow chart depicts appropriate actions if pooling is permissible given these constraints.

Figure 5.1. Analysis of Quantal Data: Methods for determining the NOEC. Note that the dose count in '>2' includes the control.

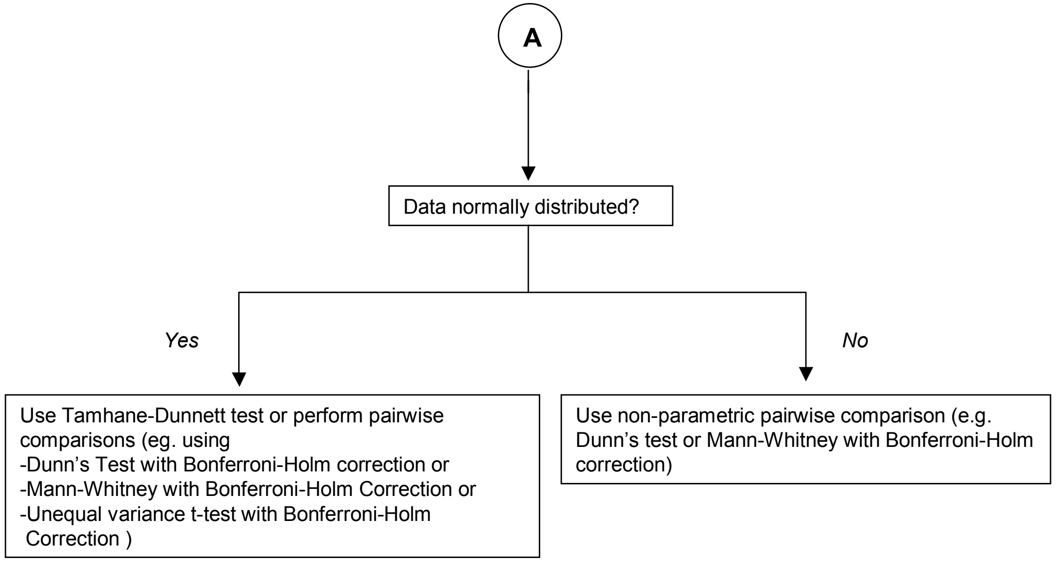

Both solvent control and non-solvent control are present.
Yes
No
Compare controls using Wilcoxon. Do controls differ?
Yes
No
Drop Non-solvent control
Combine controls*, retaining subgroups
Dose Response Experiment?
Yes
No
Expect monotone dose response & there are >2 doses** in test?
Compare treatments to a common control?
Yes
No
Use step-down trend test (e.g. based on Jonkheere, or alternatively, if data are normal and homogeneous, Williams test)
Assess data for normality and variance homogeneity. Data normal ***and*** homogeneous?
Yes
No
Assess data for normality and variance homogeneity. Data normal and homogeneous?
Non-standard design. Not discussed here.
Yes
No
Use parametric pairwise comparison (e.g. Dunnett's test)
Go to
A
On next page
Yes
No
Use parametric pairwise comparison (e.g. Dunnett's test)
Use non-parametric pairwise comparison (e.g. Mann-Whitney with Bonferroni correction)

* Both scientific judgment and regulatory guidance must be considered in deciding whether to pool non-solvent and solvent controls. The flow chart depicts appropriate actions if pooling is permissible given these constraints.

** Doses include 0-dose control

Note: If there are <5 experimental units per treatment, or there are massive ties (see text) then exact trend or pairwise tests should be used if possible.

Figure 5.2. Analysis of Continuous Data: Methods for determining the NOEC



Data normally distributed?

Use Tamhane-Dunnett test or perform pairwise comparisons (eg. using
-Dunn’s Test with Bonferroni-Holm correction or
-Mann-Whitney with Bonferroni-Holm Correction or
-Unequal variance t-test with Bonferroni-Holm Correction)

Use non-parametric pairwise comparison (e.g. Dunn’s test or Mann-Whitney with Bonferroni-Holm correction)

Note: If there are <5 experimental units per treatment, or there are massive ties (see text) then exact trend or pairwise tests should be used if possible.

Figure 5.3. Analysis of Continuous Data: Methods for determining the NOEC.

5.1.1. The NOEC: What it is, and what it is not.

112. The NOEC is defined as the test concentration below the lowest concentration that did result in a significant effect in the specific experiment, i.e. the NOEC is the tested concentration next below the LOEC.

113. A significant effect is generally meant to be a statistically significant effect, as resulting from a hypothesis test. Obviously, no claim can be made that the condition of organisms exposed to toxicants at the NOEC is the same as the condition of organisms in the control group, or that the NOEC is an estimate of the threshold of toxicity (if such exists). Rather, no effect could be detected in this particular experiment. The detectability of an effect depends on the quality and the size of the experiment and the statistical procedure used. Of course, zero effects are never detectable. The relationship between the detectability of effects and the quality of the experiment can be quantified by the concept of statistical power. For a given null and alternative hypothesis, sample size and variance, statistical power is the probability that a particular magnitude of effect will result in a significant test outcome. In large experiments (i.e., many replicates) smaller sized effects are detectable as compared to small experiments. Thus, one may consider the detectable effect size of a particular experiment as an analogue of the detection limit of a particular chemical analysis. The detectable effect size can be increased not only by using larger sample sizes, but also by taking measures to make the experimental (residual) error smaller and by selecting more powerful statistical tests.

114. Power calculations are useful for the purpose of designing experiments in such a way that effect sizes that are considered relevant are likely to be (statistically) detected. Care must be taken when using information on the power for interpreting a NOEC. If the test was designed to detect a difference of x% and an observed treatment effect is not found statistically significant this does not allow one to conclude with a specified level of confidence that the true effect in the population is less than x%.

115. Meaningful confidence intervals for the effect size at a given concentration are sometimes possible. An application of this is discussed in section 5.1.3 and methods for doing this are developed in Annex 5.3. For some techniques, obtaining meaningful confidence intervals is very difficult and this is discussed in greater detail in that annex.

5.1.2. Hypothesis Used to determine NOEC

116. The hypothesis that is tested in determining the NOEC for a toxicological experiment reflects the risk assessment question and the assumptions that are made concerning the underlying characteristics, or statistical model, of the responses being analysed (e.g., does the response increase in an orderly (i.e., monotone) way with increasing toxicant concentration?). The statistical test that is used depends on the hypothesis tested (e.g., are responses in all groups equal?), the associated statistical model, and the distribution of the values (e.g., are data normally distributed?). Thus, it is necessary to understand the question to be answered and to translate this question into appropriate null and alternative hypotheses before selecting the test procedure.

117. The need to select a statistical model for assessing the results of toxicity tests is not unique to the hypothesis testing approach. All methods of assessment assume a statistical model. The hypothesis testing approach to evaluation of toxicity data is based in part on keeping to a reasonable number the untestable or difficult-to-test assumptions, particularly those regarding the statistical model that will be used in reaching conclusions. The models used in regression and biologically based methods use stronger assumptions than the models used in the hypothesis testing approach.

118. The simplest statistical model generally used in hypothesis testing assumes only that the distributions of responses within these populations are identical except for a location parameter (e.g., the mean or median of the distribution of values from each group). Another statistical model that is often used assumes that there is a trend in the response that is associated with increasing exposure. Each of these models suggests a set of hypotheses that can be tested to determine whether the model is consistent with the data. These two types of hypotheses can further be expressed as 1-sided or 2-sided. The discussion below is developed in terms of population means, but applies equally to hypotheses concerning population medians. The most basic hypothesis (in 1-sided form) can be stated as follows:

$H_0 : \mu_0=\mu_1=\mu_2=\ldots=\mu_k$ vs. $H_1 : \mu_0>\mu_i$ for at least one i, (model 1)

where $\mu_{i,}$ i=0, 1, 2, 3, …, k denote the means of the control and test *populations*, respectively.

119. Thus, one tests the null hypothesis of no differences among the population means against the alternative that at least one population mean is smaller than the control mean. There is no investigation of differences among the treatment means, only whether treatment means differ from the control mean. The one-sided hypothesis is appropriate when an effect in only one direction is a concern. The direction of the inequality in the above alternative hypothesis (i.e. in $H_1 : \mu_0>\mu_i$) would be appropriate if a decrease in the endpoint was a concern but an increase was not (for instance, if an exposure was expected to induce infertility and reduce number of offspring). If an increase in the endpoint was the only concern, then the direction of the inequality would be reversed.

Two-sided Trend Test

120. In the two-sided form of the hypothesis, the alternative hypothesis is :

$H_1 : \mu_0\neq\mu_i$ for at least one i.

Trend or Pairwise test

121. If no assumption is made about the relationships among the treatment groups and control (e.g., no trend is assumed), the test statistics will be based on comparing each treatment to the control, independent of the other treatments. Many tests have been developed for this approach, some of which will be discussed below. Most such tests were developed for experiments in which treatments are qualitatively different, as, for example, in comparing various new therapies or drug formulations to a standard.

122. In toxicology, the treatment groups generally differ only in the exposure concentration (or dose) of a single chemical. It is further often true that biology suggests that if the chemical is toxic, then as the level of exposure is increased, the magnitude effect will tend to increase. Depending on what response is measured, the effect of increasing exposure may show up as an increase or as a decrease in the measured response, but not both. The statistical model underlying this biological expectation is what will be called a trend model or a model assuming monotonicity of the population means:

$\mu_0 \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq \mu_k$ (or with inequalities reversed) (Model 2)

The null and alternative hypotheses can then be stated as

$H_{02} : \mu_0=\mu_1=\mu_2=\ldots=\mu_k$ vs $H_{12} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq \mu_k$, with $\mu_0 > \mu_k$.

Note that $\mu_0 > \mu_k$ is equivalent, under the alternative, to $\mu_0 > \mu_i$ for at least one *i*. If this monotone model is accepted as representing the true responses of test organisms to exposure to toxicants, it is not possible for, say, μ_3 to be smaller than μ_0 and μ_6 not to be smaller.

123. Under the trend model and tests designed for that model, if tests of hypotheses H_{02} vs. H_{12} reveal that μ_3 is different from μ_0, but μ_2 is not, the NOEC has been determined (i.e. it is the test concentration associated with $\mu_{2)}$, and there is no need to test whether μ_1 differs from $\mu_{0.}$ Also, finding that μ_3 differs from μ_0 implies that a significant trend exists across the span of doses including μ_0 and μ_3, the span including μ_0 and μ_4, and so on. For the majority of toxicological studies, a test of the trend hypothesis based on model (2) is consistent with the basic expectations for a model for dose-response. In addition, statistical tests for trend tend to be more powerful than alternative non-trend tests, and should be the preferred tests if they are applicable. Thus, a necessary early step in the analysis of results from a study is to consider each endpoint, decide whether a trend model is appropriate, and then choose the initial statistical test based on that decision. Only after it is concluded trend is not appropriate do specific pairwise comparisons make sense to illuminate sources of variability.

124. Toxicologists sometimes do not know whether a compound will cause measurements of continuous variables such as growth or weight to increase or decrease, but they are confident it will act in only one direction. For such endpoints, the 2-sided trend test is appropriate, described in 5.1.6. One difference between implementing step-down procedures for quantal data and continuous data is that two-sided tests are much more likely to be of interest for continuous variables. Such a model is rarely appropriate for quantal data, as only increased incidence rate above background (control) incidence are of interest in toxicology.

125. The two-sided version of the step-down procedure is based on the underlying model:

$$\mu_0 \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq . . . \geq \mu_k$$

or

$$\mu_0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq . . . \leq \mu_{k.}$$

126. Under this model, in testing the hypothesis that all population means are equal against the alternative that at least one inequality is strict, one first tests separately each 1-sided alternative at the 0.025-level of significance with all doses present. If neither of these tests is significant, the NOEC is higher than the highest concentration. If both of these tests are significant, a trend-based procedure should not be used, as the direction of the trend is unclear. If exactly one of these tests with all the data is significant, then the direction of all further tests is in the direction of the significant test with all groups. Thereafter, the procedure is as in the 1-sided test, except all tests are at the 0.025 significance level to maintain the overall 0.05 false positive rate.

127. Where it is biologically sensible, it is preferable to test the one-sided hypothesis, because random variation in one direction can be ignored, and as a result, statistical tests of the one-sided hypothesis are more powerful than tests of the two-sided hypothesis.

128. Note that a hypothesis test based on model 2 assumes only a monotone dose-response rather than a precise mathematical form, such as is required for regression methods (Chapter 6) or the biologically based models (Chapter 7).

5.1.3. Comparisons of single-step (pairwise comparisons) or step-down trend tests to determine the NOEC

129. In general, determining the NOEC for a study involves multiple tests of hypotheses (i.e., a family of hypotheses is tested), either pairwise comparisons of treatment groups, or a sequence of tests of the significance of trend. For that reasons, statisticians have developed tests to control the family-wise error rate, FWE, (the probability that one or more of the null hypotheses in the family will be rejected incorrectly) in the multiple comparisons performed to identify the NOEC. For example, suppose one compares each of ten treatments to a common control using a simple t-test with a false positive error rate of 5% for each comparison. Suppose further that none of the treatments has an effect, i.e., all of the treatment and control population means are equal. For each comparison, there is a 5% chance of finding a significant difference between that sample treatment mean and the control. The chance that at least one of the ten comparisons is wrongly declared significant is much higher, possibly as high as $1\text{-}.95^{10}$ =0.4 or 40%. The method of controlling the family-wise error rate has important implications for the power of the test. There are two approaches that will be discussed: single-step procedures and step-down procedures. There are numerous variations within each of these two classes of procedures that are suited for specific data types, experimental designs and data distributions.

130. A factor that must be considered in selecting the methods for analysing the results from a study is whether the study is a dose-response experiment. In this context, a dose-response experiment is one in which treatments consist of a series of increasing doses of the same test material. Monotone responses from a dose-response experiment are best analysed using step-down procedures based on trend tests (e.g., the Cochran-Armitage, Williams, or Jonckheere-Terpstra trend test), whereas non-monotone responses must be analysed by pairwise comparisons to the control (e.g., Fisher's exact test or Dunnett's test). This section will discuss when to use each of these two approaches.

131. *Single-step procedures* amount to performing all possible comparisons of treatment groups to the control. Multiple comparisons to the control may be made, but there is no ordered set of hypotheses to test, and no use of the sequence of outcomes in deciding which comparisons to make. Examples of the single-step approach include the use of the Fisher's exact test, the Mann-Whitney, Dunnett and Dunn tests. Since many comparisons to the control are made, some adjustment must be made for the number of such comparisons to keep the family-wise error (FWE) rate at a fixed level, generally 0.05. With tests that are inherently single comparison tests, such as Fisher's exact and Mann-Whitney, a Bonferroni adjustment can be made: a study with k treatment levels would be analysed by performing the pair-wise comparisons of each of the treatment groups to the control group, each performed at a significance level of α/k instead of α. (This is the Bonferroni adjustment.) Equivalently, the calcutaed p-value ignoring multiplicities is multiplied by k. That is, $p^b_i = k{*}p_i$ The Bonferroni adjustment is generally overly conservative, especially for large k. Modifications reduce the conservatism while preserving the FWE at 0.05 or less.

132. For the Holm modification of the Bonferroni adjustment, arrange the k unadjusted p-values for all comparisons of treatments to control in rank order, i.e., $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \ldots \leq p_{(k)}$. Beginning with $p_{(1)}$, compare $p_{(i)}$ with $\alpha/(k-i+1)$, stopping at the first non-significant comparison. If the smallest i for which $p_{(i)}$ exceeds $\alpha/(k-i+1)$ is $i=j$, then all comparisons with $i>j$ are judged non-significant without further comparisons. It is helpful (Wright (1992)) to report adjusted p-values rather than the above comparisons. Thus, report $p^*_{(1)} = p_{(1)}{*}(k\text{-}i+1)$ and then compare each adjusted p-value to α. Table 5.1 illustrates the advantage of the Bonferroni-Holm method. In this hypothetical example, only the comparison of treatment 4 with the control would be significant if the Bonferroni adjustment is used, whereas all comparisons except the comparison of the Control with treatment 1 would be significant if the Bonferroni-Holm adjustment is used.

Comparison	Unadjusted *p* value	Bonferroni-Holm Adjusted *p* value $p*_{(i)}$	Bonferroni Adjusted p-values $P^b{}_i$
Control – Treatment 4	$p_{(1)}$=0.002	0.002*4=0.008	0.002*4=0.008
Control – Treatment 2	$p_{(2)}$=0.013	0.013*3=0.039	0.013*4=0.052
Control – Treatment 3	$p_{(3)}$=0.020	0.020*2=0.040	0.02*4=0.08
Control – Treatment 1	$p_{(4)}$=0.310	0.310*1=0.310	0.310*4=1.

Table 5.1 Comparison of Adjusted and Unadjusted P-Values

133. Alternatives based on the Sidak inequality (each comparison at level $1\text{-}(1\text{-}\alpha)^k$) are also available. The Bonferroni and Bonferroni-Holm adjustment guarantee that the family-wise error rate is less than α, but they are conservative. Other tests, such as Dunnett's, have a "built-in" adjustment for the number of comparisons made and are less conservative (hence, more powerful). For completeness, it should be understood that if only one comparison is made, the Bonferroni and Bonferroni-Holm adjustments leave the p-value unchanged. Of course, there is no need to refer to an adjustment in this simple case, but the discussion becomes needlessly complicated if special reference is always made to the case of only one comparison.

134. *Step-down procedures* are generally preferred where they are applicable. All step-down procedures discussed are based on a sequential process consisting of testing an ordered set of hypotheses concerning means, ranks, or trend. A step-down procedure based on trend (for example) works as follows: First, the hypothesis that there is no trend in response with increasing dose is tested when the control and all dose groups are included in the test. Then, if the test for trend is significant, the high dose group is dropped from the data set, and the hypothesis that there is no trend in the reduced data set is tested. This process of dropping treatment groups and testing is continued until the first time the trend test is non-significant. The highest dose in the reduced data set at that stage is then declared to be the NOEC. Distinguishing features of step-down procedures are that the tests of hypothesis must be performed in a given order, and that the outcome of each hypothesis test is evaluated before deciding whether to test the next hypothesis in the ordered sequence of hypotheses. It is these two aspects of these procedures that account for controlling the family-wise error (FWE) rate.

135. A step-down method typically uses a critical level larger than that used in single-step procedures, and seeks to limit the number of comparisons that need to be made. Indeed, the special class of "fixed-sequence" tests described below fix the critical level at 0.05 for each comparison but bound the FWE rate at 0.05. Thus, step-down methods are generally preferable to the single-step methods as long as the response means are monotonic.

136. Tests based on trend are logically consistent with the anticipated monotone pattern of responses in toxicity tests. Step-down procedures make use of this ordered alternative by ordering the tests of hypotheses. This minimises the number of comparisons that need to be made, and in all the methods discussed here, a trend model is explicitly assumed (and tested) as a part of the procedure.

137. Procedures that employ step-down trend tests have more power than procedures that rely on multiple pairwise comparisons when there is a monotone dose-response because they make more use of the biology and experimental design being analysed. When there is a monotone dose-response, procedures that compare single treatment means or medians against the control, independent of the results in other

treatments (i.e. single-step procedures), ignore important and relevant information, and suffer power loss as a result.

138. The trend models used in the step-down procedures do not assume a particular precise mathematical relationship between dose and response, but rather use only monotonicity of the dose-response relationship. The underlying statistical model assumes a monotone dose-response in the *population* means, not the *observed* means.

139. Rejection of the null hypothesis (i.e., rejecting the hypothesis that all group means, or medians, or distributions are equal) in favour of the stated alternative implies that the high dose is significantly different from the control. The same logic applies at each stage in the step-down application of the test to imply, whenever the test is significant, that the high dose remaining at that stage is significantly different from the control. These tests are all applied in a 1-sided manner with the direction of the alternative hypothesis always the same. Moreover, this methodology is general, and applies to any legitimate test of the stated hypotheses under the stated model. That is, one can use this fixed-sequence approach with the Cochran-Armitage test on quantal data, the Jonckheere-Terpstra or Williams or Brown-Forsythe tests of trend on continuous data. Other tests of trend can also be used in this manner.

140. *Deciding between the two approaches* Bauer (1997) has shown that certain tests based on a monotone dose-response can have poor power properties or error rates when the monotone assumption is wrong. For example, departures from monotonicity in non-target plant data are common, where they arise from low dose stimulation. Davis and Svendsgaard (1990) suggest that departures from monotonicity may be more common than previously thought.. These results suggest that a need for caution exists. There are two testing philosophies used to determine whether a monotone dose-response is appropriate. Some recommend assessing in a general way for an endpoint or class of endpoints, whether a monotone dose-response is to be expected biologically. If a monotone trend is expected, then trend methods are used. This procedure should be augmented, at a minimum, by adding that, if a cursory examination of the data shows strong evidence of departure from monotonicity (i.e., large, consistent departures), then pairwise methods should be used instead.

141. A second philosophy recommends formal tests to determine if there is significant monotonicity or significant departure from monotonicity. With continuous data, one can use either a positive test for monotonicity (such as Bartholomew's test) and proceed only if there is evidence of monotonicity, or use a "negative" test for departure from monotonicity (such as sets of orthogonal contrasts for continuous responses and a decomposition of the chi-square test of independence for quantal responses) and proceed unless there is evidence of non-monotonicity. Details on these procedures are given in Annexes 5.1 and 5.3. Either philosophy is acceptable. The second approach is grounded in the idea that monotonicity is the rule and that it should take strong evidence to depart from this rule. Both approaches reduce the likelihood of having to explain a significant effect at a low or intermediate concentration when higher concentrations show no such effect. The "negative" testing approach is more consistent with the way tests for normality and variance homogeneity are used and is more likely to result in a trend test than a method that requires a significant trend test to proceed. This is what is shown in the flow diagrams presented below.

142. Formal tests for monotonicity are especially desirable in a highly automated test environment. One simple procedure that can be used in this situation for continuous responses is to construct linear and quadratic contrasts of normalised rank statistics (to avoid the complications that can arise from non-normal or heterogeneous data). If the linear contrast is not significant and the quadratic contrast is significant, there is evidence of possible non-monotonicity that calls for closer examination of the data or pairwise comparison methods. Otherwise, a trend-based analysis is used. A less simple, but more elegant procedure would be to construct simultaneous confidence intervals for the mean responses assuming monotonicity (i.e., isotonic estimators based on maximum likelihood criteria – see Annex 5.3) and use a trend approach

unless one or more sample (i.e., non-isotonic) means fall outside the associated confidence interval. For quantal data using the Cochran-Armitage test, there is a built-in test for lack of monotonicity.

143. Where expert judgement is used, formal tests for monotonicity or its lack may be replaced by visual inspection of the data, especially of the mean or median responses. The same concept applies to assessing normality and variance homogeneity.

5.1.4. Dose metric in trend tests

144. Various authors have evaluated the influence on trend tests of the different ways of expressing dose (i.e. dose metrics), including actual dose-values, log(dose), and equally-spaced scores (i.e., rank-order of doses). Lagakos and Lewis (1985) discuss various dose metrics and prefer the rank-order as a general rule. Weller and Ryan (1998) likewise prefer rank ordering of doses for some trend tests.

145. When dose values are approximately equally spaced on a log scale, there is little difference between using log(dose) and rank-order, but use of actual dose values can have the unintended effect of turning a trend test into a comparison of high dose to control, eliminating the value of the trend approach and compromising its power properties. This is not an issue with some tests, such as the Jonckheere-Terpstra test discussed below, since rank-order of treatment groups is built into the procedure. With others, such as Cochran-Armitage and contrast-based tests, it is an important consideration.

146. Extensive computer simulations have been done (J. W. Green, in preparation) to compare the use of rank-order to dose-value in the Cochran-Armitage test. One simulation study involved over 88,000 sets of dose-response scenarios for 4- and 5-dose experiments found 12-17% of the experiments where the rank-order scoring found lower NOEC than dose-value did and only 1% of the experiments where dose-value scores lead to lower NOEC than when rank-order scores were used. In the remaining cases, the two methods established the same NOEC. While these simulations results do not, by themselves, justify the use of rank-order over actual dose levels or their logarithms, they do suggest that use of rank-order will not lessen the power of statistical tests. All trend based tests discussed in this document, including contrast tests for monotonicity, are based on rank ordering of doses.

5.1.5. The Role of Power in Toxicity Experiments

147. The adequacy of an experimental design and the statistical test used to analyse study results are often evaluated in terms of the power of the statistical test. Power is defined as the probability that a false null hypothesis will be rejected by the statistical test in favour of a true alternative. That power depends on the alternative hypothesis. In the context of toxicology, the larger the effect, the higher the power to detect that effect. So, if a toxicant has had some effect on the organisms in a toxicity test, power is the probability that a difference between treatment groups and the control will be detected. The power of a test can be calculated if we know the size of the effect to be detected, the variability of the endpoint measured, the number of treatment groups, and the number of replicates in each treatment group. (Detailed discussions are given in sections 5.2 and 5.3 and Annexes 5.1 and 5.3).

148. It should be understood that the goal of selecting a method for determining a NOEC is not to find the most powerful method. Rather, the focus should be on selecting methods most appropriate for the data and end result. Power is certainly an ingredient in this selection process. As discussed below, power can be used in designing experiments and selecting statistical tests to reduce animal use without loss of statistical power. This can be accomplished by selecting an inherently more powerful test applied to fewer animals, so that the result is to retain the power of more traditional tests but use fewer animals.

149. The primary use of power analysis in toxicity studies is in the design stage. By demonstrating that a study design and test method have adequate power to detect effects that are large enough to be deemed

important, if we then find that, at a given dose, there is no statistically significant effect, we can have some confidence that there is no effect of concern at that dose. However, power does not quantify this confidence. Failure to adequately design or control an experiment so that statistical tests have adequate power can result in large effects being found to be statistically insignificant. On the other hand, it is also true that a test can be so powerful that it will find statistically significant effects of little importance.

150. Deciding on what effect size should be considered to be large enough to be important is difficult, and may depend on both biological and regulatory factors. In some cases, the effect size may be selected by regulatory agencies or specified in guidelines.

151. A requirement to demonstrate an adequate power to detect effects of importance will remove any perceived reward for poor experimental design or technique, as poor experimental design will be shown to have low power to detect important effects, and will lead to the selection of more powerful statistical tests and better designs. The latter will be preferable to the alternative of increasing sample sizes. Indeed, it is sometimes possible to find statistical procedures with greater power to detect important differences or provide improved estimates and simultaneously decrease sample sizes.

152. For design purposes, the background variance can be taken to be the pooled within-experiment variance from a moving frame of reference from a sufficiently long period of historical control data with the same species and experimental conditions. The time-window covered by the moving frame of reference should be long enough to average out noise without being so long that undetected experimental drift is reflected in the current average. If available, a three-to-five year moving frame of reference might be appropriate. When experiments must be designed using more limited information on variance, it may be prudent to assume a slightly higher value than what has been observed. Power calculations used in design for quantal endpoints must take the expected background incidence rate into account for the given endpoint, as both the Fisher Exact and Cochran-Armitage test are sensitive to this background rate, with highest power achieved for a zero background incidence rate. The background incidence rate can be taken to be the incidence rate in the same moving frame of reference already mentioned.

153. While at the design stage, power must, of necessity, be based on historical control data for initial variance estimates, it may also be worthwhile to do a post-hoc power analysis as well to determine whether the actual experiment is consistent with the criteria used at the design stage. Care must be taken in evaluating post-hoc power against design power. Experiment-to-experiment variation is expected and variance estimates are more variable than means. The power determination based on historical control data for the species and endpoint being studied should be reported.

154. Alternatively, for experimental designs constructed to give an acceptable power based on an assumed variance rather than on historical control data, a post-hoc test can be done to compare the observed variance to the variance used in designing the experiment. If this test finds significantly higher observed variance (e.g., based on a chi-square or F-test) than that used in planning, then the assumptions made at design time may need to be reassessed.

5.1.6. Experimental design

155. Factors that must be considered when developing experimental designs include the number and spacing of doses or exposure levels, the number of subjects per dose group, and the nature and number of subgroups within dose groups. Decisions concerning these factors are made so as to provide adequate power to detect effects that are of a magnitude deemed biologically important.

156. The choice of test substance concentrations is one aspect of experimental design that must be evaluated for each individual study. The goal is to bracket the NOEC with concentrations that are as

closely spaced as practical. If limited information on the toxicity of a test material is available, test concentrations or doses can be selected to cover a range somewhat greater than the range of exposure levels expected to be encountered in the field and should include at least one concentration expected not to have a biologically important effect. If more information is available this range may be reduced, so that doses can be more closely spaced. Where effects are expected to increase approximately in proportion to the log of concentration, concentrations should be approximately equally spaced on a log scale. Three to seven concentrations plus concomitant controls are suggested, with the smaller experiment size typical for acute tests and larger experiment sizes most appropriate when preliminary dose-finding information is skimpy.

157. The trade-off between number of subjects per subgroup and number of subgroups per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among- subgroup variation and correlation. If there are no subgroups, then there is no way to distinguish housing effects from concentration effects and neither between- and within-group variances or nor correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to subgroup means other than the Jonckheere-Terpstra test. Thus, a minimum of two subgroups per concentration is recommended; three subgroups are much better than two; four subgroups are better than three. The improvement in modelling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modelling is improved if we get better estimates of both among- and within-subgroup variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance will have, at least approximately, a chi-squared distribution. The quality of a variance estimate can be measured by the width of its confidence interval and a look at a chi-squared table will verify the statements made.) The precise needs for a given experiment will depend on factors such as the relative and absolute size of the between- and within-replicate variances. Examples 1 and 2 in Annex 5.3 illustrate the trade-offs between replicates per concentration and subjects per replicate.

158. In any event, the number of subgroups per concentration and subjects per subgroup should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.

159. Since the control group is used in every comparison of treatment to control, consideration should be given to allocating more subjects to the control group than to the treatment groups in order to optimise power for a given total number of subjects. The optimum allocation depends on the statistical test to be used. A widely used allocation rule was given by Dunnett (1955), which states that for a total of N subjects and k treatments to be compared to a common control, if the same number, n, of subjects are allocated to every treatment group, then the number, n_0, to allocate to the control to optimise power is determined by the so-called square-root rule. By this rule, the value of n is (the integer part of) the solution of the equation $N= kn + n\surd k$, and $n_0 = N - kn$. [It is almost equivalent to say $n_0 = n\surd k$.] This has been shown to optimise power for Dunnett's test. It is used, often without formal justification, for other pairwise tests, such as the Mann-Whitney and Fisher exact test. Williams (1972) showed that the square-root rule may be somewhat sub-optimal for his test and optimum power is achieved when $\surd k$ in the above equation is replaced by something between $1.1\surd k$ and $1.4\surd k$.

160. The optimality of the square-root rule to other tests, such as Jonckheere-Terpstra and Cochran-Armitage has not been published in definitive form, but simulations (manuscript in preparation by J. W. Green) show that for the step-down Jonckheere-Terpstra test, power gains of up to 25% are common under this rule compared to results from equal sample sizes. In all cases examined, the power is greater following this rule compared to equal sample sizes, where the total sample size is held constant In the absence of definitive information on the Jonckheere-Terpstra and other tests, it is probably prudent to follow the

square-root rule for pairwise, Jonckheere-Terpstra and Cochran-Armitage tests and either that or Williams' modification of the rule for other step-down procedures.

161. The selection of an allocation rule is further complicated in experiments where two controls are used, since if the controls are combined for further testing, a doubling of the control sample size is already achieved. Since experience suggests that most experiments will find no significant difference between the two controls, the optimum strategy for allocating subjects is not necessarily immediately clear. This of course would not apply if a practice of pooling of controls is not followed.

162. The reported power increases from allocating subjects to the control group according to the square root rule do not consider the effect of any increase in variance as concentration increases. One alternative, not without consequences in terms of resources and treatment of animals, is to add additional subjects to the control group without subtracting from treatment groups. There are practical reasons for considering this, since a study is much more likely to be considered invalid when there is loss of information in the controls than in treatment groups.

5.1.7. Treatment of Covariates and Other Adjustments to Analysis

163. It is sometimes necessary to adjust the analysis of toxicity data by taking into account some restriction on randomisation, compartmentalisation (housing) or by taking into account one or more covariates that might affect the conclusions. Examples of potential covariates include: initial body weights, initial plant heights, and age at start of test. While a thorough treatment of this topic will not be presented, some attention to this topic is in order.

164. For continuous, normally distributed responses with homogeneous variances, analysis of covariance (ANCOVA) is well developed. Hocking (1985) and Milliken and Johnson (1984) are among the many references on this topic. For continuous responses that do not meet the normality or homogeneity requirements, non-parametric ANCOVA is available.

165. Shirley (1981) indicates why nonparametric methods are needed in some situations. Stephenson and Jacobson (1988) contain a review of papers on the subject up to 1988. Subsequent papers include Wilcox (1991) and Knoke (1991). Stephenson and Jacobson recommend a procedure that replaces the dependent variable with ranks but retains the actual values of the independent variable(s). This has proved useful in toxicity studies. Seaman et al (1985) discuss power characteristics of some non-parametric ANCOVA procedures.

166. When the response variable is quantal and is assumed to follow the binomial distribution, ANCOVA can be accomplished through logistic regression techniques. In this case, the covariate is a continuous regressor variable and the dose groups are coded as 'dummy variables.' This approach can be more generally described in the Generalized Linear Model (GLM) framework (McCullagh and Nelder (1989)). For quantal data, Koch et al (1998), Thall and Vail (1990), Harwell and Serlin (1988), Tangen and Koch (1999a, 1999b) consider some relevant issues.

167. Adjustments must be made to statistical methods when there are restrictions on randomisation of subjects such as housing of subjects together. This is discussed for both quantal and continuous data in sections 5.2.2.6, 5.2.3, and 5.3.2.7, where the possibility of correlations among subjects housed together is considered, as are strategies for handling this problem. In the simple dose-response designs being discussed in this chapter, other types of restrictions on randomisation are less common. However, there is a large body of literature on the treatment of blocking and other issues that can be consulted. Hocking (1985) and Milliken and Johnson (1984) contain discussions and additional references.

168. Transformation of the doses (i.e. *not* response measures) in hypothesis testing is restricted, in this chapter, to the use of rank order of the doses. For many tests, the way that dose values (actual or rank order) are expressed has no effect on the results of analysis. An exception is the Cochran-Armitage test. (See Annex 5.1)

5.2. Quantal data (e.g., Mortality, Survival)

5.2.1. Hypothesis testing with quantal data to determining NOEC values

169. Selection of methods and experimental designs in this chapter for determining NOEC values focuses on identifying the tests most appropriate for detecting effects. The appropriateness of a given method hinges on the design of the experiment and the pattern of responses of the experimental units. Figure 5.1 illustrates an appropriate scheme for method selection, and identifies several statistical methods that are described in detail below. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

170. If there are two negative controls (i.e., solvent and non-solvent) Fisher's exact test applied just to the two controls is used to determine whether the two groups differ wherever it is appropriate to analyse individual sampling units. Where replicate means or medians are the unit for analysis, the Mann-Whitney rank sum test can be used. Further discussion of when each approach is appropriate is given in sections 5.2.2 and 5.2.2.3. Section 4.2.3 contains discussions of issues regarding multiple controls in an ecotoxicity study.

171. Figure 5.1 identifies a number of powerful methods for the analysis of quantal data. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

172. The methods used for determining NOEC values on quantal data can be categorised according to whether the tests involved are parametric or non-parametric and whether the methods are single-step or step-down. Table 5.2 lists methods that can be used to determine NOEC values. Some of these methods are applicable only under certain circumstances, and some methods are preferred over the others.

173. Except for the two Poisson tests, those tests listed in the column "Parametric" can be performed only when the study design allows proportion of organisms responding in replicated experimental units to be calculated (i.e. there are multiple organisms within each of multiple test vessels within each treatment group). Such a situation yields multiple responses, namely proportions, for each concentration, and these proportions can often be analysed as continuous. For very small samples, such a practice is inappropriate.

174. Typically, if responses increase or remain constant with increasing dosage, the trend-based methods perform better than pairwise methods, and for most quantal data, a step-down approach based on the Cochran-Armitage test is the most appropriate of the listed techniques. The strengths and weaknesses of most listed methods are discussed in more detail below.

	Parametric	Non-Parametric
Single-Step (Pair-wise)	Dunnett Poisson comparisons	Mann-Whitney with Bonferroni-Holm adjustments. Chi-squared with Bonferroni-Holm adjustment Steel's Many-to-One Fisher's exact test with Bonferroni-Holm adjustment.
Step-down (Trend based)	Poisson Trend Williams Bartholomew Welsch Brown-Forsythe Sequences of linear contrasts	Cochran-Armitage Jonckheere-Terpstra test Mantel-Haenszel

Table 5.2 Methods used for determining NOEC values with quantal data.

All listed single-step methods are based on pair-wise comparisons, and all step-down methods are based on trend-tests. The tests listed in Table 5.2 are well established as tests of the stated hypothesis in the statistics literature. *Note:* (The Mann-Whitney test is identical to the Wilcoxon rank-sum test.)

5.2.2. Parametric versus non-parametric tests

175. Parametric tests are based on assumptions that the responses being analysed follow some given theoretical distribution. Except for the Poisson methods, the tests listed in Table 5.2. as parametric all require that the data be approximately normally distributed (possibly after a transformation).The normality assumption can be met for quantal data only if the experimental design includes treatment groups that are divided into subgroups, the quantal responses are used to calculate proportions responding in each of the subgroups, and these proportions are the observations analysed. These proportions are usually subjected to a normalising transformation (see sections 4.32, 4.33, and 4.34), and a weighted ANOVA is performed, perhaps with weights proportional to subgroup sizes (Cochran (1943)). (It is noteworthy that some statistical packages, such as SAS version 6, do not always perform multiple comparisons within a weighted ANOVA correctly.) This approach limits the possibilities of doing trend tests to those based on contrasts, including Welsch and Brown-Forsythe tests (Roth (1983); Brown and Forsythe (1974)). Non-trend tests include versions of Dunnett's test for pairwise comparisons allowing for unequal variances (Dunnett (1980); Tamhane (1979)). These methods may not perform satisfactorily for quantal data, partly due to a loss of power in analysing subgroup proportions. An example is given on Annex 5.1.

176. The Cochran-Armitage test is listed as non-parametric even though it makes explicit use of a presumed binomial distribution of incidence within treatment groups. Some reasons for this are given in Annex 5.1. Fisher's Exact test is likewise listed as non-parametric, even though it is based on the geometric distribution. The Jonckheere-Terpstra test applied to subgroup proportions is certainly non-parametric. An advantage of Jonckheere-Terpstra over the cited parametric tests is that the presence of many zeros poses no problem for the analysis and it provides a powerful step-down procedure in both large- and small-sample problems, provided the number of subgroups per concentration is not too small. An example in Annex 5.3 will illustrate this concern.

5.2.2.1. Single-step procedures

177. Suitable single-step approaches for quantal data are Fisher's exact test and the Mann-Whitney test to compare each treatment group to the control, independently of other treatment groups, with Bonferroni-Holm adjustment. Details of these tests are given in annex 5.1.

5.2.2.2. Step-Down Procedures

178. Suitable step-down procedures for quantal data are based on the Cochran-Armitage and Poisson trend tests. First, a biological determination is made whether or not to expect a monotone dose-response. If that judgement is to expect monotonicity, then the step-down procedure described below is followed unless the data strongly indicates non-monotonicity. If the judgement is not to expect monotonicity, then Fisher's exact test is used.

179. An analysis of quantal data is based on the relationships between the response (binary) variable and factors. In such cases, the Pearson Chi-Square (χ^2) test for independence can be used to find if any relationships exist.

180. Test for monotone dose-response: If one believes on biological grounds that there will be a monotone dose-response, then the expected course of action is to use a trend test. However, statistical procedures should not be followed mindlessly. Rather, one should examine the data to determine whether it is consistent with the plan of action. There is a simple and natural way to check whether the dose-response is monotone. The *k-1* df Pearson Chi-Square statistic decomposes into a test for linear trend in the dose-response and a measure of lack of fit or lack of trend, $\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$ where $\chi^2_{(1)}$ is the calculated Cochran-Armitage linear trend statistic and $\chi^2_{(k-2)}$ is the Chi-Square statistic for lack of fit. The details of the computations are provided in annex 5.1.

181. If the trend test is significant when all doses are included in the test, then proceed with a trend-based step-down procedure. If the trend test with all doses included is not significant but the test for lack of fit is significant, then this indicates that there are differences among the dose groups but the dose-response is not monotone. In this event, even if we expected a monotone dose-response biologically, it would be unwise to ignore the contrary evidence and one should proceed with a pairwise analysis.

182. The Cochran-Armitage trend test is available in several standard statistical packages including SAS and StatXact. StatXact also provides exact power calculations for the Cochran-Armitage trend test with equally spaced or arbitrary doses.

183. The step-down procedure: A suitable approach to analysing monotonic response for quantal data is as follows. Perform a Cochran-Armitage test for trend on responses from all treatment groups including the control. If the Cochran-Armitage test is significant at the 0.05 level, omit the high dose group, re-compute the Cochran-Armitage and Chi-Squared tests with the remaining dose groups. Continue this procedure until the Cochran-Armitage test is first non-significant at the 0.05 level. The highest concentration remaining at this stage is the NOEC.

184. Possible Modifications of the Step-Down Procedure: There are two possible modifications to consider to the above. First, as noted by Cochran (1943), Fisher's Exact test is more powerful for comparing two groups than the Cochran-Armitage test when the total number of subjects in the two groups is less than 20 and also when that total is less than 40 and the expected frequency of any cell is less than 5. This will include most laboratory ecotoxicology experiments. For this reason, if the step-down procedure described above reaches the last possible stage, where all doses above the lowest tested dose are significant, then we can substitute Fisher's exact test for Cochran-Armitage for the final comparison on the

grounds that it is a better procedure for this single comparison. Such substitution does not alter the power characteristics or theoretical justification of the Cochran-Armitage test for doses above the lowest dose, but it does improve the power of the last comparison.

185. Second, if the step-down procedure terminates at some higher dose because of a non-significant Cochran-Armitage test, but there is at this stage a significant test for lack of monotonicity, one should consider investigating the lower doses further. This can be done by using Fisher's exact test to compare the remaining dose groups to the control, with a Bonferroni-Holm adjustment. The Bonferroni-Holm adjustment would take into account only the number of comparisons actually made using Fisher's exact test. The inclusion of a method within the step-down procedure to handle non-monotonic results at lower doses is suggested for quantal data (but not for continuous data) for two reasons. First, there is a sound procedure built into the decomposition of the Chi-squared test for assessing monotonicity that is directly related to the Cochran-Armitage test. Secondly, experience suggests that quantal responses are more prone to unexpected changes in incidence rates at lower doses than continuous responses, so that a strict adherence to a pure step-down process may miss some adverse effects of concern.

5.2.2.3. Alternative Procedures

186. These following parametric and nonparametric procedures are discussed because under some conditions, a parametric analysis of subgroup proportions may be the only viable procedure. This is especially true if there are also significant differences in the number of subjects within each subgroup, making analysis of means or medians problematic by other methods.

187. Pairwise ANOVA (weighted by subgroup size) based methods performed on proportion affected have sometimes been used to determine NOEC values. While there can be problems with these proportion data meeting some of the assumptions of ANOVA (e.g., variance homogeneity), performing the analysis on proportion affected opens up the gamut of ANOVA type methods, such as Dunnett's test and methods based on contrasts. Failure of data to satisfy the assumption of homogeneity of variances can often be corrected by the use of an arcsine-square-root or other normalising and variance stabilising transformation. However, this approach tends to have less power than step-down methods designed for quantal data that are described above, and is especially problematic for very small samples. These ANOVA based methods may not be very powerful and are not available if there are not distinct subgroups of multiple subjects each within each concentration. Williams' test is a trend alternative that can be used, when data are normally distributed with homogeneous variance.

188. A nonparametric trend test that can be used to analyse proportion data is the Jonckheere-Terpstra trend test, which is intended for use when the underlying response on each subject is continuous and the measurement scale is at least ordinal. The most common application in a toxicological setting is for measures such as size, fecundity, and time to an event. The details of this and other tests that are intended for use with continuous responses are given in section 5.3. A disadvantage of the use of the Jonckheere-Terpstra trend test for analysing subgroup proportions where sample sizes are unequal is that it does not take sample size into account. It is not proper to treat a proportion based on 2 animals with the same weight as one based on 10, for example. For most toxicology experiments where survival is the endpoint, the sample sizes are equal, except for a rare lost subject, so this limitation is often of little importance. Where a sub lethal effect on surviving subjects is the endpoint, then this is a more serious concern.

189. The methods described in Table 5.2 are sometimes used but tend to be less powerful than one designed for quantal data, such as those so indicated in Table 5.2. They are appropriate only if responses of organisms tested are independent, and there is not significant heterogeneity of variances among groups (i.e., within-group variance does not vary significantly among groups). If there is a lack of independence or significant heterogeneity of variances, then modifications are needed. Some such modifications are

discussed below. In the ANOVA context, a robust ANOVA (e.g., Welch's variance-weighted one-way ANOVA) that does not assume variance homogeneity can be used.

190. Poisson tests can be used as alternatives in both non-trend and trend approaches. (See annex 5.1) A robust Poisson approach (Weller and Ryan (1998)) using dummy variables for groups, or multiple Mann-Whitney tests using subgroup proportions as the responses could be used. In each case, an adjustment for number of comparisons should be made. For the robust Poisson model, this would be of the Bonferroni-Holm type. For the Mann-Whitney test, the Bonferroni-Holm adjustment could be used or these pairwise comparisons could be "protected" by requiring a prior significant Kruskal-Wallis test (i.e. an overall rank-based test of whether any group differs from any other). It should be noted that the Mann-Whitney approach does not take subgroup size into account, but this will usually not be an issue for survival data.

5.2.2.4. Assumptions of methods for determining NOEC values

191. The assumptions that must be met for the listed methods for determining NOEC values vary according to the methods. Assumptions common to all methods are given below, while others apply only to specific methods. The details on the latter are given in annex 5.1.

192. Assumption: Responses are independent. All methods listed in Table 5.2.1 are based on the assumption that responses are independent observations. Failure to meet this assumption can lead to highly biased results. If organisms in a test respond independently, they can be treated as binomially distributed in the analysis.(See section 4.2.2 for further discussion.) It is not uncommon in toxicology experiments for treatment groups to be divided into subgroups. For example, an aquatic experiment may have subjects exposed to the same nominal concentration but grouped in several different tanks or beakers. It sometimes happens that the survival rate within these subgroups varies more from subgroup to subgroup than would be expected if the chance of dying were the same in all subgroups. This added variability is known as extra-binomial (or extra-Poisson) variation, and is an indication that organisms in the subgroups are responding to different levels of an uncontrolled experimental factor (e.g., subgroups are exposed to differing light levels or are being held at differing temperatures) and are not responding independently. In this situation, correlations among subjects must be taken into account. For quantal responses, an appropriate way to handle this is to analyse the subgroup responses; that is, the subgroups are considered to be the experimental unit (replicate) for statistical analysis. Note that lack of independence can arise from at least two sources: differences in conditions among the tanks and interactions among organisms.

193. With mortality data, extra-binomial variation (heterogeneity) is not a common problem, but it is still advisable to do a formal or visual check. Two formal tests are suggested: a simple Chi-Squared test and an improved test of Potthoff and Whittinghill (1966). Both tests are applied to the subgroups of each treatment group, in separate tests for each treatment group. While these authors do not suggest one, an adjustment for the number of such tests (e.g., Bonferroni) is advisable. It should be noted also that the Chi-squared test can become undependable when the number of expected mortalities in a Chi -squared cell is less than five. In this event, an exact permutation version of the Chi-squared test is advised and is available in commercially available software, such as StatXact and SAS.

194. If organisms are not divided into subgroups, lack of independence cannot be detected easily, and the burden for establishing independence falls to biological argument. If there is a high likelihood of aggression or competition between organisms during the test, responses may not be independent, and this possibility should be considered before assigning all organisms in a test level to a single test chamber.

195. It should be noted that even if subgroup information is entered separately, a simple application of the Cochran-Armitage test ignores the between-subgroup (i.e., within-group) variation and treats the data as though there were no subgrouping. This is inappropriate if heterogeneity among subgroups is

significant. The same is true of simple Poisson modelling. Thus, if significant heterogeneity is found, an alternative analysis is advised. One in particular deserves mention. This is a modification of the Cochran-Armitage test developed by Rao and Scott (1992) that is simple to use and is appropriate when there is extra-binomial variation. The beta-binomial model of Williams (1975) is another modification of the Cochran-Armitage tests that allows for extra-binomial variation. If the Jonckheere-Terpstra test is used, there is no adjustment (or any need to adjust) for extra-binomial variation, as that method makes direct use of the between-subgroup variation in observed proportions. However, as pointed out above, if there is considerable variation in subgroup sizes, this approach suffers by ignoring sample size.

Treatment of multiple controls

196. A preliminary test can be done comparing just the two controls as a step in deciding how to interpret the experimental data. For quantal (e.g., mortality) data, Fisher's exact test is appropriate. The decision of how to proceed after this comparison of controls is given in section 4.2.3.

5.2.3. Additional Information

197. Annex 5.1 contains details of the principle methods discussed in this section, including examples. Annex 5.2 contains a discussion of the power characteristics of the step-down Cochran-Armitage and Fisher exact tests. Section 5.3 and Annex 5.3 contain a discussion of the methods for continuous responses that can be used to analyse subgroup proportions, as discussed above.

5.2.4. Statistical Items to be Included in the Study Report

198. The report describing quantal study results and the outcome of the NOEC determination should contain the following items:

- Test endpoint assessed
- Number of Test Groups
- Number of subgroups within each group (if applicable)
- Identification of the experimental unit
- Nominal and measured concentrations (if available) for each test group
- Number exposed in each treatment group (or subgroup if appropriate)
- Number affected in each treatment group (or subgroup if appropriate)
- Proportion affected in each treatment group (or subgroup if appropriate)
- Confidence interval for the percent effect at the NOEC, provided that the basis for the calculation is consistent with the distribution of observed responses. (See Annex 5.3).
- P value for test of homogeneity if performed
- Name of the statistical method used to determine the NOEC
- The dose metric used
- The NOEC
- P value at the LOEC (if applicable)
- Design power of the test to detect an effect of biological importance (and what that effect is) based on historical control background and variability.

- Actual power achieved in the study.
- Plot of response data versus concentration.

5.3. Continuous data (e.g., Weight, Length, Growth Rate)

5.3.1. Hypothesis testing with continuous data to determine NOEC

199. Figure 5.2 provides a scheme for determining NOEC values for continuous data, and identifies several statistical methods that are described in detail below. As reflected in this flow chart, continuous monotone dose-response data are best analysed using a step-down test based on the Jonckheere trend test or Williams test (the former applicable regardless of the distribution of the data, the latter applicable only if data are normally distributed and variances of the treatment groups are homogeneous).

200. Non-monotonic dose-response data should be assessed using an appropriate pairwise comparison procedure. Several such are described below. They can be categorized according whether the data are normally distributed or homogeneous. Dunnett's test is appropriate if the data are normally distributed with homogeneous variance. For normally distributed but heterogeneous data, the Tamhane-Dunnett (T3) method (Hochberg and Tamhane, 1987) can be used. Alternatively, such data can be analysed by the Dunn, Mann-Whitney, or unequal variance t-tests with Bonferroni-Holm adjustment. Non-normal data can be analysed by using Dunn or Mann-Whitney tests with Bonferroni-Holm adjustment. Normality can be formally assessed using the Shapiro-Wilk test (Shapiro and Wilk 1965) while homogeneity of variance is assessed by Levene's test (Box, 1953). Dunn's test, if used, should be configured only to compare groups to control. All of these procedures are discussed in detail below. Alternatives exist to these if software used does not include these more desirable tests. For normality, the Anderson-Darling, Kolmogorov-Smirnov, Cramér-von Mises, Martinez-Iglewicz and D'Agostino Omnibus test are available. For variance homogeneity, Cochran's Q, Bartlett's and the Maximum F test can be used. The tests described in detailed in this chapter are recommended where available, based on desirable statistical properties.

201. There are, of course, a number of statistical procedures that are not listed in Figure 5.2 that might also be applied to continuous data. The following discussion identifies many of the procedures that might be used, and attempts to provide some insight into the strengths and weaknesses of each..

202. Table 5.3.1 lists methods that are sometimes used to determine NOEC values. Some of these methods are applicable only under certain circumstances, and some methods are preferred over the others. Parametric tests listed are performed only when the distribution of the data to be analysed is approximately normally distributed. Some parametric methods also require that the variances of the treatment groups be approximately equal.

	Parametric	Non-Parametric
Single-Step (Pair-wise)	Dunnett Tamhane-Dunnett	Dunn Mann-Whitney with Bonferroni correction
Step-down (Trend based)	Williams Bartholomew Welch trend Brown-Forsythe trend Sequences of linear contrasts	Jonckheere-Terpstra Shirley

Table 5.3.1. Methods used for determining NOEC values with continuous data.

All listed single step methods are based on pair-wise comparisons, and all step-down methods are based on trend-tests.

5.3.1.1. Parametric versus non-parametric tests

203. The parametric tests listed in Table 5.3.1, all require that the data be approximately normally distributed. Many also require that the variances of the treatment groups are equal (exceptions are the Tamhane-Dunnett, Welch and Brown-Forsythe tests). Parametric tests are desirable when these assumptions can be met. The failure of the data to meet assumptions can sometimes be corrected by transforming the data. (Section 5.1.10) Some non-parametric tests are almost as powerful as their parametric counterparts when the assumptions of normality and homogeneity of variances are met. The non-parametric tests may be much more powerful if the assumptions are not met. Furthermore, a test based on trend is generally more powerful than a pairwise test. A decision to use a parametric or non-parametric test should be based on which best describes the physical, biological and statistical properties of a given experiment.

204. Piegorsch and Bailer (1997), referenced in the document, warns that use of the Jonckheere-Terpstra test requires that shapes of distributions or the response variable be equivalent and in many cases, this translates to requiring that the response variable have a common variance. They conclude the applicability of the Jonckheere-Terpstra test is brought into question when there are large disparities in variances. While the Jonckheere-Terpstra test discussed in detail below is a distribution-free trend test, that fact alone does not mean that its results are not susceptible to heterogeneity of variance. While most people who have investigated the usual nonparametric methods find them less sensitive to these problems than the usual parametric procedures, they are not impervious to these problems. To address this question, a large power simulation study has been carried out (J. W. Green, manuscript in preparation) comparing the effects of variance heterogeneity on the Jonckheere, Dunnett, and Tamhane-Dunnett tests. These simulations have shown the Jonckheere test to be much less affected by heterogeneity than the alternatives indicated and to lose little of its good power properties.

205. Heterogeneity and non-normality are inherent in some endpoints, such as first or last day of hatch or swim up. There will be observed zero within-group variance in the control and lower concentrations quite often and non-zero variance in higher concentrations. No transformation will make the data normal or

homogeneous. It may be possible to apply some generalized linear model with a discrete distribution to such data, but that is not addressed in this chapter.

5.3.1.2. Single-step (pairwise) procedures

206. These tests are used when there is convincing evidence (statistical or biological) that the dose-response is not monotone. This evidence can be through formal tests or through visual inspection of the data, as discussed in section 5.3.2.3. Pairwise procedures are also appropriate when there are differences among the treatments other than dose, such as different chemicals or formulations. These tests are described briefly here. Details of each test, including mathematical description, power, assumptions, advantages and disadvantages, relevant confidence intervals, and examples are discussed in Annex 5.3.

207. *Dunnett's test:* Dunnett's test is based on simple t-tests from ANOVA but uses a different critical value that controls the family-wise error (FWE) rate for the $k-1$ comparisons of interest at exactly α. Each treatment mean is compared to the control mean. This test is appropriate for responses that are normally distributed with homogeneous variances and is widely available.

208. *Tamhane-Dunnett Test:* Also known as the T3 test, this is similar in intent to Dunnett's test but uses a different critical value and the test statistic for each comparison uses only the variance estimates from those groups. It is appropriate when the within-group variances are heterogeneous. It still requires within-group responses to be normally distributed and controls the FWE rate at exactly α.

209. *Dunn's Test:* This non-parametric test is based on contrasts of mean ranks. In toxicity testing, it is used to compare the mean rank of each treatment group to the control. To control the FWE rate at α or less, the Bonferroni-Holm correction (or comparable alternative) should be applied. Dunn's test is appropriate when the populations have identical continuous distributions, except possibly for a location parameter (e.g., the group medians differ), and observations within samples are independent. It is used primarily for non-normally distributed responses.

210. *Mann-Whitney test:* This is also a non-parametric test and can be applied under the same circumstances as Dunn's test. The Mann-Whitney rank sum test compares the ranks of measurements in two independent random samples and has the aim of detecting if the distribution of values from one group is shifted with respect to the distribution of values from the other. It can be used to compare each treatment group to the control. When more than one comparison to the control is made, a Bonferroni-Holm adjustment is used.

5.3.1.3. Step-down trend procedures

211. For continuous data, two trend tests are described for use in step down procedures, namely the Jonckheere-Terpstra and Williams' Test (described below) that are appropriate provided there is a monotone dose-response. Where expert judgement is available, the assessment of monotonicity can be through visual inspection. For such an assessment, plots of treatment means, subgroup means, and raw responses versus concentration will be helpful. An inspection of treatment means alone may miss the influence of outliers. However, a visual procedure cannot be automated, and some automation may be necessary in a high-volume toxicology facility. Although not discussed here in detail, the same methodology can be applied to the Welsch, Brown-Forsythe or Bartholomew trend tests.

212. A general step-down procedure is described in the next section. Where the term "trend test" is used, one may substitute either "Jonckheere-Terpstra test" or "Williams' test." Details of these, as well as advantages and disadvantages, examples, power properties, and related confidence intervals for each are given in Annex 5.3.

5.3.1.4. Determining the NOEC using a step-down procedure based on a trend test

213. This section describes a generalised step-down procedure for determining the NOEC for a continuous response from a dose response study. It is appropriate whenever the treatment means are expected to follow a monotone dose-response and there is no problem evident in the data that precludes monotonicity.

214. *Preliminaries:* The procedure described is suitable if the experiment being analysed is a dose response study with at least two dose groups (Fig. 62). For clarity, the term "dose group" includes the zero-dose control. Before entering the step-down procedure, two preliminary actions must be taken. First, the data are assessed for monotonicity (as discussed in section 5.1.4). A step-down procedure based on trend tests is used if a monotonic response is evident. Pairwise comparisons (e.g., Dunnett's, Tamhane-Dunnett, Dunn's test or Mann-Whitney with Bonferroni-Holm correction, as appropriate) instead of a trend-based test should be used where there is strong evidence of departure from monotonicity. Next, examine the number of responses and number of ties (as discussed in section 5.3.2.1). Small samples and data sets with massive ties should be analysed using exact statistical methods if possible. Finally, if a parametric procedure (e.g. Dunnett's or Williams' test) is to be used, then an assessment of normality and variance homogeneity should be made. These are described elsewhere.

215. *The Step-Down Procedure: The preferred approach to analysing monotonic response patterns is as follows.* Perform a test for trend (Williams or Jonckheere) on responses from all dose groups including the control. If the trend test is significant at the 0.05 level, omit the high dose group, and re-compute the trend statistic with the remaining dose groups. Continue this procedure until the trend test is first non-significant at the 0.05 level, then stop. The NOEC is the highest dose remaining at this stage. If this test is significant when only the lowest dose and control remain, then a NOEC cannot be established from the data.

216. *Williams' test:* Williams' test is a parametric procedure that is applied in the same way the Jonckheere-Terpstra test is applied. This procedure, described in detail in Annex 5.3, assumes data within concentrations are normally distributed and homogeneous. In addition to the requirement of monotonicity rather than linearity in the dose-response, an appealing feature of this procedure is that maximum likelihood methods are used to estimate the means (as well as the variance) based on the assumed monotone dose-response of the population means. The resulting estimates are monotone. An advantage of this method is that it can also be adapted to handle both between- and within-subgroup variances. This is important when there is greater variability between subgroups than chance alone would indicate. Williams' test must be supplemented by a non-parametric procedure to cover non-normal or heterogeneous cases. Either Shirley's (1979) non-parametric version of Williams' test or the Jonckheere-Terpstra test can be used, but if these alternative tests are used, one loses the ability to incorporate multiple sources of variances. Limited power comparisons suggest similar power characteristics for Williams' and the Jonckheere-Terpstra tests.

217. *Jonckheere-Terpstra Test:* The Jonckheere-Terpstra trend test is intended for use when the underlying response of each experimental unit is continuous and the measurement scale is at least ordinal. The Jonckheere-Terpstra test statistic is based on joint rankings (also known as Mann-Whitney counts) of observations from the experimental treatment groups. These Mann-Whitney counts are a numerical expression of the differences between the distributions of observations in the groups in terms of ranks. The Mann-Whitney counts are used to calculate a test statistic that is used in conjunction with standard statistical tables to determine the significance of a trend. Annex 5.3 gives details of computations. The Jonckheere-Terpstra test reduces to the Mann-Whitney test when only one group is being compared to the control.

218. The Jonckheere-Terpstra test has many appealing properties. Among them is the requirement of monotonicity rather than linearity in the dose-response. Another advantage is that an exact permutation version of this test is available to meet special needs (as discussed below) in standard statistical analysis packages, including SAS and StatXact. If subgroup means or medians are to be analysed, the Jonckheere-Terpstra test has the disadvantage of failing to take the number of individuals in each subgroup into account.

219. Extensive power simulations of the step-down application of the Jonckheere-Terpstra test compared to Dunnett's test have demonstrated in almost every case considered where there is a monotone dose-response, that the Jonckheere-Terpstra test is more powerful than Dunnett's test (Green, J. W., in preparation for publication). The only situation investigated in which Dunnett's test is *sometimes* slightly more powerful than the Jonckheere-Terpstra is when the dose-response is everywhere flat except for a single shift. These simulations followed the step-down process to the NOEC determination by the rules given above and covered a range of dose-response shapes, thresholds, number of groups, within-group distributions, and sample sizes.

5.3.1.5. Assumptions for methods for determining NOEC values

Small Samples / Massive Ties

220. Many standard statistical tests are based on large sample or asymptotic theory. If a design calls for fewer than 5 experimental units per concentration, such large sample statistical methods may not be appropriate. In addition, if the measurement is sufficiently crude, then a large proportion of the measured responses have the same value, or are very restricted in the range of values, so that tests based on a presumed continuous distribution may not be accurate. In these situations, an exact permutation-based methodology may be appropriate. While universally appropriate criteria are difficult to formulate, a simple rule that should flag most cases of concern is to use exact methods when any of the following conditions exists: (1) at least 30% of the responses have the same value; (2) at least 50% of the responses have one of two values; (3) at least 65% of the responses have one of three values. StatXact and SAS are readily available software packages that provide exact versions of many useful tests, such as the Jonckheere-Terpstra and Mann-Whitney tests.

Normality

221. When parametric tests are being considered for use, then a Shapiro-Wilk test (Shapiro and Wilk 1965) of normality should be performed. If the data are not normally distributed, then either a normalising transformation (section 5.1.10) should be sought or a non-parametric analysis should be done. Assessment of non-normality can be done at the 0.05 significance level, though a 0.01 level might be justified on the grounds that ANOVA is robust against mild non-normality. The data to be checked for normality are the residuals after differences in group means are removed; for example, from an ANOVA with concentration, and, where necessary, subgroup, as class (i.e., non-numeric) variables.

Variance Homogeneity

222. If parametric tests are being considered for use and the data are normally distributed, then a check of variance homogeneity should be performed. Levene's test (Box, 1953) is reasonably robust against marginal violations of normality. If there are multiple subgroups within concentrations, the variances used in Levene's test are based on the subgroup means. If there are no subgroups the variances based on individual measurements within each treatment group would be used. It should be noted that ANOVA is robust to moderate violations of assumptions, especially if the experimental design is balanced (equal n in the treatment groups), and that some tests for homogeneity are less robust than the ANOVA itself. Small

departures from homogeneity (even though they may be statistically significant by some test) can be tolerated without adversely affecting the power characteristics of ANOVA based tests. For example, it is well known that Bartlett's test is very sensitive to non-normality. It is customary to use a much smaller significance level, (e.g., 0.001) if this test is used. Levene's test, on the other hand, is designed to test for the very departures from homogeneity that cause problems with ANOVA, so that a higher level significance (0.01 or 0.05) in conjunction with this test can be justified. Where software is available to carry out Levene's test, it is recommended over Bartlett's.

223. For pairwise (single-step) procedures, if the data are normally distributed but heterogeneous, then a robust version of Dunnett's test (called Tamhane-Dunnett in this document) is available. Such a procedure is discussed in Hochberg and Tamhane (1987). Alternatives include the robust pairwise tests of Welch and Brown-Forsythe. If the data are normally distributed and homogeneous, then Dunnett's test is used. Specific assumptions and characteristics of many of the tests referenced in this section are given in Annex 5.3.

224. Of course, expert judgement should be used in assessing whether a significant formal test for normality or variance homogeneity reveals a problem that calls for alternative procedures to be used.

5.3.1.6. Operational considerations for statistical analyses

Treatment of Experimental Units

225. A decision that must often be made is whether the individual animals or plants can be used as the experimental unit for analysis, or whether subgroups should be the experimental unit. The consequences of this choice should be carefully considered. If there are subgroups in each concentration, such as multiple tanks or beakers or pots, each with multiple specimens, then the possibility exists of within- and among-subgroup variation, neither of which should be ignored. If subjects within subgroups are correlated, that does not mean that individual subject responses should not be analysed. It does mean that these correlations should be explicitly modelled or else analysis should be based on subgroup means. Methods for modelling replicated dose groups (e.g., nested ANOVA) are available. For example, Hocking (1985), Searle (1987, especially section 13.5), Milliken and Johnson (1984, esp. chapter 23), John (1971), Littell (2002) and many additional references contain treatments of this.

226. Technical note: If both within-subgroup and between-subgroup variation exist and neither is negligible, then the step-down trend test should either be the Jonckheere-Terpstra test with mean or median subgroup response as the observation, or else an alternative trend test such as Williams' or Brown-Forsythe with the variance used being the correct combination of the within- and among-subgroups variances as described in the discussion on the Tamhane-Dunnett test in Appendix 5.3.1.

227. Given the possibility of varying subgroup sample sizes at the time of measurement, it may not be appropriate to treat all subgroup means or medians equally. For parametric comparisons, this requires only the use of the correct combination of variance components, again as described as Appendix 5.3.1. For non-parametric methods, including Jonckheere's test, there are no readily available methods for combining the two sources of variability. The choices are between ignoring the differences in sample sizes and ignoring the subgroupings. If the differences in sample sizes are relatively small, they can be ignored. If the differences among subgroups are relatively small, they can be ignored. If both differences are relatively large, then there is no universally best method. A choice can be made based on what has been observed historically in a given lab or for a given type of response and built into the decision tree.

Identification and Meaning of Outliers

228. The data should be checked for outliers that might have undue influence on the outcome of statistical analyses. There are numerous outlier rules that can be used. Generally, an outlier rule such as Tukey's (Tukey, 1977) that is not itself sensitive to the effects of outliers is preferable to methods based on standard deviations, which are quite sensitive to the effects of outliers. Tukey's outlier rule can be used as a formal test with outliers being assessed from residuals (results of subtracting treatment means from individual values) to avoid confounding outliers and treatment effects.

229. Any response more than 1.5 times the interquartile range above the third quartile (75th percentile) or below the first quartile (25th percentile) is considered an outlier by Tukey's rule. Such outliers should be reported with the results of the analysis. The entire analysis of a given endpoint can be repeated with outliers omitted to determine whether the outliers affected the conclusion. While it is true that nonparametric analyses are less sensitive to outliers than parametric analyses, omission of outliers can still change conclusions, especially when sample sizes are small or outliers are numerous.

230. Conclusions that can be attributed to the effect of outliers should be carefully assessed. If the conclusions are different in the two analyses, a final analysis using non-parametric methods may be appropriate, as they are less influenced than parametric methods by distributional or outlier issues.

231. It is not appropriate to omit outliers in the final analysis unless this can be justified on biological grounds. The mere observation that a particular value is an outlier on statistical grounds does not mean it is an erroneous data point.

Multiple Controls

232. To avoid complex decision rules for comparing a water and solvent control, it is recommended that a non-parametric Mann-Whitney (or, equivalently, Wilcoxon) comparison of the two controls be performed, using only the control data. This comparison can be either a standard or an exact test, according as the preliminary test for exact methods is negative or positive. If a procedure for comparing controls using parametric tests were to be employed, then another layer of complexity can result, where one has to assess normality and variance homogeneity twice (once for controls and again later, for all groups) and one must also consider the possibility of using transformations in both assessments.

General

233. Outliers, normality, variance homogeneity and checks of monotonicity should be done only on the full data set, not repeated at each stage of the step-down trend test, if used. Diagnostic tools for determining influential observations can also be very helpful in evaluating the sensitivity of an analysis to the effects of a few unusual observations.

5.3.2. Statistical Items to be Included in the Study Report.

234. The report describing continuous study results and the outcome of the NOEC determination should contain the following items:

- Description of the statistical methods used
- Test endpoint assessed
- Number of Test Groups
- Number of subgroups within each group and how handled (if applicable)

- Identification of the experimental unit
- Nominal and measured concentrations (if available) for each test group
- The dose metric used.
- Number exposed in each treatment group (or subgroup if appropriate)
- Group means (and median, if a non-parametric test was used) and standard deviations
- Confidence interval for the percent effect at the NOEC, provided that the basis for the calculation is consistent with the distribution of observed responses. (See Annex 5.3).
- The NOEC
- P value at the LOEC (if applicable)
- Results of power analysis
- Plot of response versus concentration

6. DOSE-RESPONSE MODELLING

6.1. Introduction

235. The main regulatory use of dose-response modeling in toxicity studies is to estimate an *ECx*, the exposure concentration that causes an x% effect in the biological response variable of interest, and its associated confidence bounds. The value of *x*, the percent effect, may be specified in advance, based on biological (or regulatory) considerations. Guidelines may specify for which value(s) of *x* the *ECx* is required. This chapter discusses how an *ECx* may be estimated, as well as how it may be judged that the available data are sufficient to do so.

236. Dose-response (or concentration-response) modelling aims at describing the dose-response data as a whole, by means of a dose-response model. In general terms, it is assumed that the response, *y*, can be described as a function of concentration (or dose), *x* :

$$y = f(x)$$

where *f* can be any function that is potentially suitable for describing a particular dataset. Since *y* is considered as a function of *x,* the response variable *y* is also called the dependent variable, and the concentration *x,* the independent variable. As an example, consider the linear function

$$y = a + b x$$

where the response changes linearly with the concentration. Here, *a* and *b* are called the model parameters. By changing parameter a one may shift the line upwards or downwards, while by changing the parameter b one may rotate the line. Fitting a line to a dataset is the process of finding those values of a and b that result in “the best fit”, i.e., making the distances of the data points to the line as small as possible. Similarly, for any other dose-response model, or function *f,* the best fit may be achieved by adjusting the model parameters.

237. This example illustrates that the data determine the values of the parameters *a* and *b,* and thereby the location and angle of the line. However, whatever the data, the result of the fitting process will, for this model, always be a straight line, so the flexibility of the dose-response model in following the dose-response data is limited. In general, the flexibility of a dose-response model tends to be larger when it includes more parameters. For example, the model

$$y = a + bx + cx^2 + dx^3$$

has four parameters (*a*, *b*, *c*, and *d*), which can all be varied in the fitting process. Therefore, this model is more flexible compared to the linear model, and can take on various shapes other than a straight line. One might conclude here: "the more parameters, the better", but that is not the case. It only makes sense to include more parameters in a model when the data contain the information to estimate them (also referred to as the parsimony principle), or when including the parameter in the model leads to a significantly better fit.

238. The fit of the model to the data may be defined in various ways. One measure for the fit is the sum of squares of the residuals, where the residuals are simply the distances (differences) between the data and the model value at the pertinent concentration. The best fit is then found by minimising the sum of squared residuals, or briefly the Sum of Squares (SS). Another measure for the fit is the likelihood, which is based on a particular distribution that is assumed for the data (e.g. a normal or lognormal distribution for continuous data, a binomial distribution for quantal data, or a Poisson distribution for count data). In that case the best fit is found by maximising the likelihood (or the log-likelihood, which amounts to the same thing). See section 4.3.5. for a general discussion of model fitting.

239. In this chapter a dose-response model is generally written as $y = f(x)$, where x may denote either concentration or dose. Indeed, a concentration-response and a dose-response model are not different from a statistical point of view. The response y may refer to data of various types. The type of the response data, either quantal or continuous (see chapter 3), does make an important difference, not only for the statistical analysis, but also for the interpretation of the results. In this chapter dose-response modelling is separately discussed for quantal (6.2) and for continuous (6.3) data, since the statistical analysis is completely different. The flow chart given in Fig. 6.1 summarizes the main lines of a dose-response modeling approach.

240. Of course, the response in biological test systems not only depends on the concentration (dose) but also on the exposure duration. Yet, most ecotoxicity tests only vary the concentration (dose), at a single exposure duration. Therefore, the larger part of this chapter addresses how to model the concentration-response relationship, ignoring the exposure duration. Obviously, any results from the statistical analysis then only hold for that particular exposure duration.

241. For data sets where the exposure duration varied as well, one may apply models where the response is a function of both concentration and exposure duration. The inclusion of exposure duration is discussed in section 6.6, for both quantal and continuous data. In the other sections of this chapter, time is considered as fixed. In chapter 7 the role of time and exposure duration in describing the response is further discussed from the perspective of biologically-based modelling.

6.2. Modelling quantal dose-response data (for a single exposure duration)

242. A quantal response y is defined as $y = k/n$, where k is the number of responding organisms (or experimental units) out of a total of n. A quantal response may also be expressed as a percentage, but the total number of observed units, n, cannot be omitted. For example, 2 responses out of 4 is not the same information as 50 out of 100. See also section 4.

243. The purpose of dose-response modelling of quantal data is to estimate an ECx (LCx), where x is a given percentage usually equal to or lower than 50%. When the dose-response data relate to a particular (single) exposure duration, the estimated parameters (EC50 or ECx) obviously only hold for that particular exposure duration (or to, e.g., a single acute oral dose).

244. In this chapter the terms ED50 / EC50 / LD50 / LC50 are used interchangeably, as well as EDx / ECx / LDx / LCx, where x denotes a particular response level (usually smaller than 50). Note that *x* in model expressions denotes the concentration (or dose). In human risk assessment the term Benchmark dose (BMD)[6] is used, and is, for the case of quantal data, equivalent to the ECx (but not so for continuous responses, see section 6.3).

245. The terms dose and concentration are used interchangeably, as well as dose-response relationship and concentration-response relationship.

[6] Originally the BMD was introduced by Crump (1984) as the lower confidence bound of the point estimate. More recently this is often indicated as BMDL.

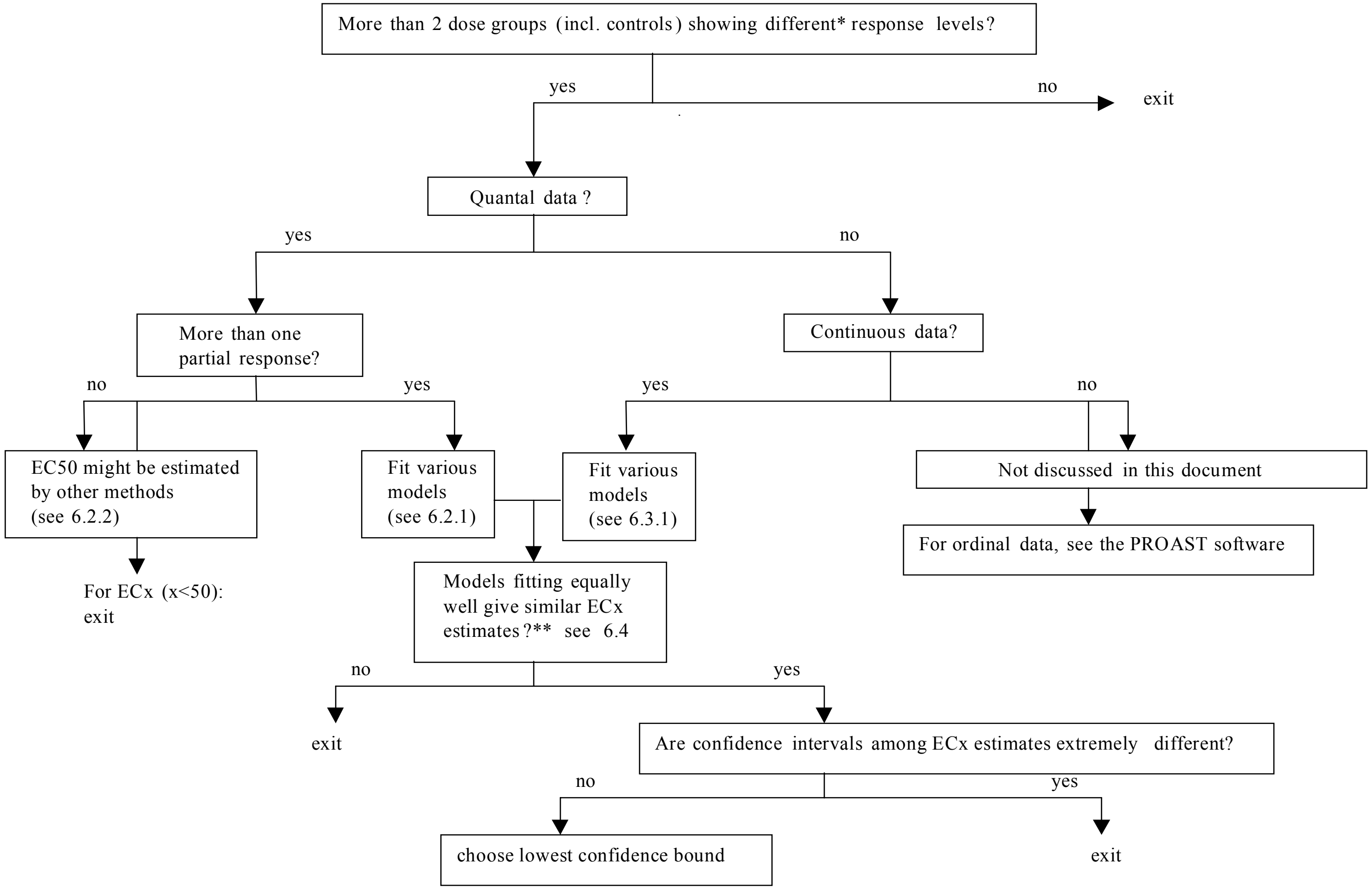

More than 2 dose groups (incl. controls) showing different* response levels?
yes
no
exit
Quantal data ?
yes
no
More than one partial response?
Continuous data?
no
yes
yes
no
EC50 might be estimated by other methods (see 6.2.2)
Fit various models (see 6.2.1)
Fit various models (see 6.3.1)
Not discussed in this document
For ordinal data, see the PROAST software
For ECx (x<50): exit
Models fitting equally well give similar ECx estimates ?** see 6.4
no
yes
exit
Are confidence intervals among ECx estimates extremely different?
no
yes
choose lowest confidence bound
exit

Fig. 6.0 Flow chart for dose -response modeling. Doses = concentrations
Exit: ECx cannot be assessed from the data at hand; repeat the experiment with more (adequate) doses, or go to Chapter 5.
* i.e. apparently different, given the noise in the data
** in addition, it should be assessed by visual inspection that the fitted model is sufficiently supported by the data.

6.2.1. Choice of model

246. A (statistical) dose-response model serves to express the observed response as a function of dose, to provide for a tool to estimate the parameters of interest, (in particular the ECx) and assess confidenceintervals for those estimates. A statistical regression model itself does not have any meaning, and the choice of the model (expression) is largely arbitrary. It is the data, not the model, that determines the dose-response, and thereby the ECx. Of course, an improper choice of the model can lead to an inappropriate estimate of the ECx, but the choice of the model is in most ecotoxicity studies governed by the data.

247. Numerous dose-response models are theoretically possible, but in practice only a limited number is applied, mostly determined by historical habits in the field of application. Only the more frequently applied models will be discussed here. See section 6.4 for a discussion on model selection.

248. For quantal data an obvious property for a dose-response function is that it ranges between 0 and 1 (0% and 100%). Further, one would normally expect the response to be monotone, i.e., it only increases (or decreases). Cumulative distribution functions (e.g., normal, logistic, Weibull) obey that property, and are therefore candidates for dose-response modelling of quantal data.

249. The use of cumulative distribution functions for quantal dose-response modelling can also be considered from the idea of tolerance distributions. By assuming that each individual in the population observed has its own tolerance for the chemical, a tolerance distribution expresses the variability between the individuals. Plotting the tolerance distribution cumulatively results in the quantal dose-response relationship, where the fraction of responding individuals (at a given concentration) is viewed as all individuals having a tolerance lower than that concentration. For example, a predicted response of 25% at concentration 10 ppm is interpreted as 25% of the individuals having a tolerance lower than 10 ppm. Given this interpretation, the slope of a quantal dose-response relationship is a reflection of the variability between the individuals, with steeper slopes meaning smaller variability in tolerances.

250. In light of the preceding, the choice of a quantal concentration-response model may be based on an assumed tolerance distribution. For several reasons one may expect a tolerance distribution to be approximately lognormal, or, equivalently, to be approximately normal for the log-concentrations. Indeed, a long history of experience has confirmed this, and it has become standard that models that are based on symmetrical tolerance distributions (e.g. the probit and logit model, see below) are fitted against the logarithms of the concentrations.

251. In general, a dose-response model for quantal data is a function of the concentration or dose x:

$$y = f(x)$$

where y is the quantal response. It is important to keep in mind that in the model, y represents the true response, which may be thought of as the fraction of responding individuals in the *infinite* population, or as the probability of response for any individual. The function *f(x)* is chosen such that it equals zero at concentration zero (and unity for infinite concentration). However, theoretically the probability of response in the unexposed population might be very small, but it cannot be (strictly) zero. Therefore, it is theoretically more appropriate to extend the model and include a background incidence parameter by putting

$$y = f(\mathrm{x}) = \mathrm{a} + (1 - a)\, g(x)$$

where *a* denotes the true background probability of response, and *g*(*x*) is a function increasing from 0 to 1 for *x* increasing from zero to infinity. In this formulation the response at infinite concentrations remains unity (since *g*(*x*) = 1 for infinite *x*).

252. Some of the more commonly used models are discussed below. For a more extensive list of models, see Scholze et al. 2001.

The probit model

253. The probit model is the cumulative normal distribution function. In practice it is usually applied to the log-concentrations, implying that a lognormal tolerance distribution is assumed. The probit model (without the background mortality parameter *a*) can be expressed as:

$$z(y) = b\{\log(x) - \log(ED50)\} = b\log(\frac{x}{ED50}) \qquad (1)$$

where *z* is the standard normal deviate associated with probability *y*. At first sight, the use of log-concentration in this model appears to present a problem for dose zero. Note, however, that for *x* = 0, *z* = -∞, and the associated probability *y* is zero. In other words, model (1) assumes that the probability of ever observing a response in the control group is strictly zero. Therefore, when model (1) is fitted to the data, the control observations can just as well be deleted[7]. They only provide information to the model when a background parameter is included in the model (see expression (5)).

254. The standard normal deviate cannot be calculated from an explicit expression, as opposed to the logit model (see below). Common statistical software packages use standard algorithms; therefore this should not concern the user.

255. The probit model has two parameters: the ED50 and the slope (*b*).

256. The ED50 is the median of the (lognormal) tolerance distribution, and the slope is the inverse of the standard deviation of that distribution.

257. Figure 6.1 shows an application of the probit model to mortality data.

The logit model

258. The logit model is the cumulative logistic distribution function. The logistic distribution has wider tails than the normal distribution, but is similar otherwise. Just as with the probit model, the logit model is usually applied to the log-concentrations.

259. The logit model (without the background mortality parameter *a*) can be expressed as:

$$y = \frac{1}{1 + \exp[b\log(ED50/x)]} \qquad (2)$$

[7] Adequate software simply sets the log-likelihood score for observations in the control group at zero (whatever the observations). When a background response parameter is included in the model (see expression (5)), the log-likelihood score associated with the observations in the control group only depends on the value of the background parameter.

where y is the probability of response.Just as in the probit model, the logarithm of dose does not present any problem for dose zero. It can be seen immediately that, in the limit, y equals zero for x approaching zero. In fitting the model, the control observations can be simply deleted, as they do not provide any information, unless a background parameter is included (see expression (5)).

260. The logit model has two parameters: the ED50 and the slope (*b*).

261. The ED50 is the median of the (log-logistic) tolerance distribution, and the slope is related to the standard deviation by:

$$SD_{tolerance\,distribtuion} = \frac{\pi}{b\sqrt{3}}$$

262. Figure 6.2 illustrates the logit model applied to the same mortality data as Figure 6.1.

The Weibull model

263. The Weibull distribution is not necessarily symmetrical, and is usually applied to the concentrations themselves (not their logs). The Weibull model (without the background mortality parameter *a*) may be expressed as

$$y = 1 - \exp[-(x/b)^c] \qquad (3)$$

264. It has two parameters, a location parameter *b*, and a parameter *c* (high values of *c* give steep slope). The ED50 is related to *b* and *c* by

$$\text{ED50} = b\ln(2)^{1/c}$$

265. Fig. 6.3 illustrates the Weibull model applied to the same mortality data as fig. 6.1 and 6.2.

Multi-stage models

266. The multi-stage model (see e.g., Crump et al. 1976) is often used for describing tumour dose-response data. It is usually applied in a simplified version (the linearized multi-stage model, briefly LMS)

$$y = 1 - \exp\{-a - bx - cx^2 - dx^3 - ....\} \qquad (4)$$

where the number of parameters is also called the number of stages. It includes the one-stage model

$$y = 1 - \exp\{-a - bx\},$$

also referred to as the one-hit model. Note that in the multistage model background mortality is included, and equals 1 – exp(-*a*).

267. The multi-stage model can be regarded as a family of nested models. For example, by setting the parameter *d* in the three-stage model equal to zero, one obtains the two-stage model. Thus, one can let the number of stages depend on the data (see below for a further discussion of nested models).

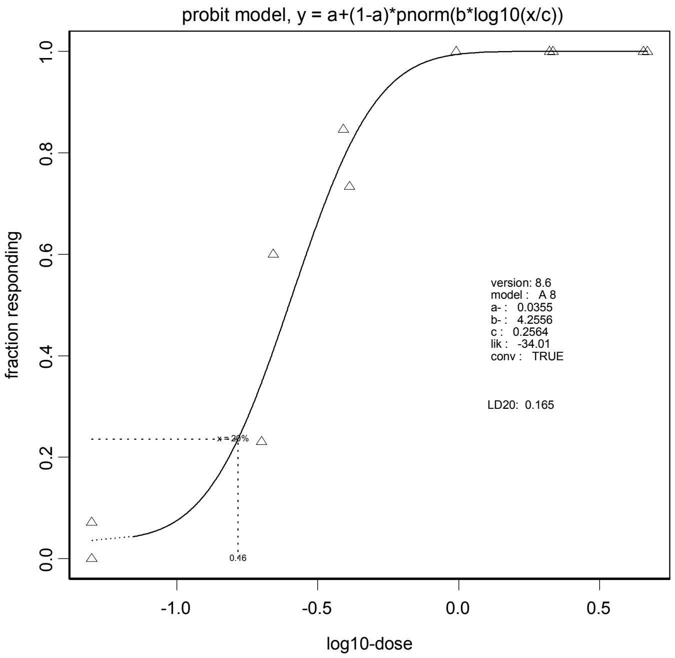


Figure 6.1 Probit model fitted to observed mortality frequencies (triangles) as a function of log-dose.

Note that, on log-scale, the zero concentration is minus infinity.

a = background mortality, b = slope, c = LD50, dashed lines indicate the LD20, pnorm = cumulative standard normal distribution function.

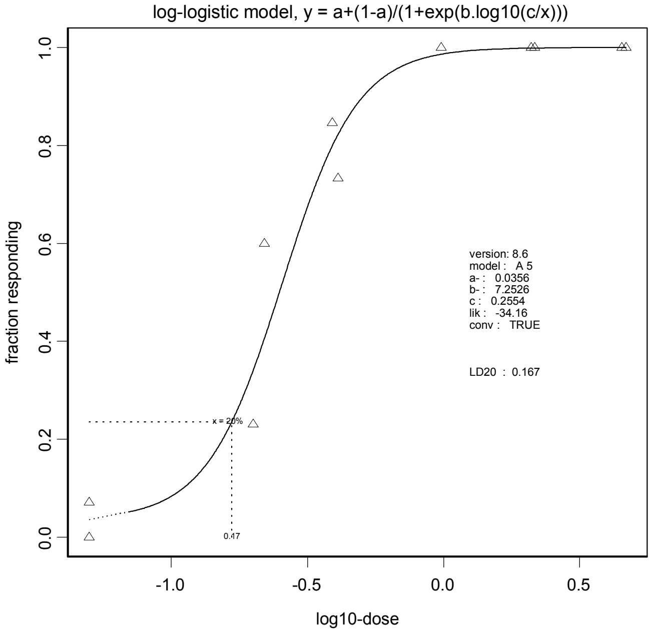


Figure 6.2 Logit model fitted to mortality dose-response data (triangles).

c = LD50, b = slope, a = background mortality, dashed lines indicate the LD20.

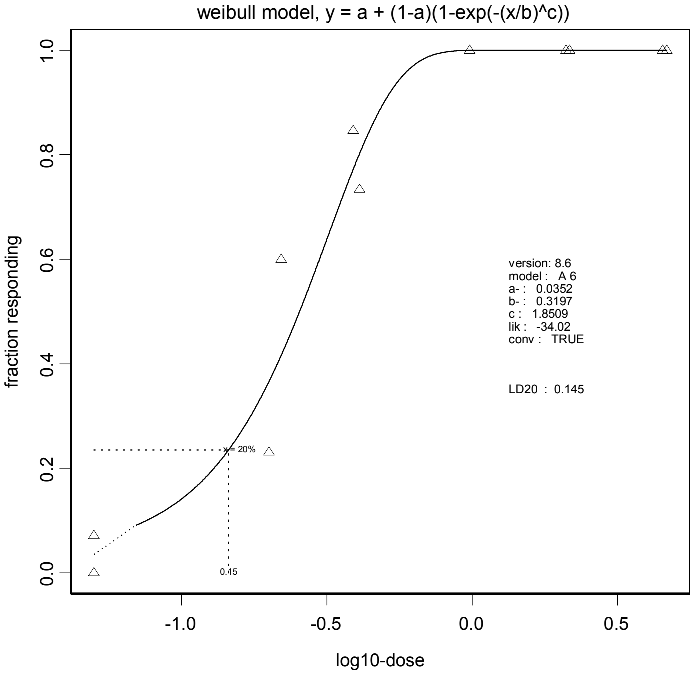


Figure 6.3 Weibull model fitted to mortality dose-response data (triangles).

b = “location” parameter, c = “slope” parameter, a = background mortality, dashed lines indicate the LD20.

268. The LD50 equals b ln(2) $^{1/c}$ = 0.145. Note that the data and the model in Figs. 6.1 – 6.3 are plotted against log-dose with the purpose of improving the readability of the plots. However, the Weibull model was fitted as a function of dose, while the probit and logit models were fitted as a function of log-dose.

Definitions of EC50 and ECx

269. The ECx is defined as the concentration associated with x% response, with the EC50 as a special case[8]. The situation of nonzero background response complicates the definition of the EC50 and of the ECx, since the background response may be taken into account in various ways.

270. The EC50 is defined as the concentration associated with 50% response. However, a 50% response (i.e. incidence) can relate to the whole population, irrespective of the background response, or only to that part of the population that did not respond at concentration zero. Consider the general quantal dose-response model where the background response (incidence) *a* is included as a model parameter:

$$y = f(x) = a + (1\text{-}a)\, g(x) \qquad (5)$$

271. Here $g(x)$ may be any cumulative tolerance distribution, ranging from zero to one. It reflects the dose-response relationship for the fraction of the population that did not show a response at concentration zero. The background-corrected EC50 then simply is the EC50 as given by $g(x)$. For example, when $g(x)$ denotes the log-logistic model, then parameter *c* is the background-corrected EC50. This definition of the EC50 (LD50) is used in Fig 6.1 to 6.3.

272. For response levels x% smaller than 50%, the ECx may be defined in various ways, e.g.,

$$x\%/100\% = f(ECx) - a \qquad \text{(additional risk)},$$
$$x\%/100\% = [f(ECx) - a\,] / (1\text{-}a) = g(\text{ECx}) \qquad \text{(extra risk)}.$$

273. For example, when the background response *a* amounts to 3%, then the EC10 according to the additional risk definition corresponds to a response in the population of 13% (since 13%-3%=10%). In the extra risk definition the EC10 would correspond to a response of 12.7% (since [12.7%-3%] / 97% = 10%).

274. Note that the background-corrected ECx according to the extra risk concept is equal to the (uncorrected) ECx of $g(x)$ in expression (5). Therefore, extra risk appears favourable, but the numerical difference for the ECx based on additional or extra risk is usually small. The illustrative examples in Fig. 6.1-6.3 used the additional risk concept.

275. In ecotoxicity testing the additional risk is common for the ECx when x < 50%. However, in the case of the EC50 the background response will usually be taken into account according to the extra risk concept (as in Figs. 6.1-6.3). It may be noted that in other disciplines, still other risk concepts are used. For instance, in epidemiology more common measures are relative risk (response of exposed subjects divided by response in non-exposed subjects), and derived concepts such as attributable proportion, and odds ratio.

[8] In human risk assessment the term Benchmark dose (BMD) is used, defined as the dose associated with a certain Benchmark response (=x%). Originally the Benchmark dose was defined as the lower confidence limit of the point estimate (Crump, 1984), also indicated as BMDL. See also Table in section 6.3.

6.2.2. Model fitting and estimation of parameters

276. Fitting a model to dose-response data may be done by using any suitable software , e.g. SAS (www.sas.com), SPSS (www.spss.com), splus (www.insighful.com), and PROAST[9] (Slob, 2003).

277. The user does not need to be aware of the computational details, but some understanding of the basic principles in nonlinear regression is required to be able to interpret the results properly. These principles are discussed in 6.7. Furthermore, the user should be aware of the assumptions underlying the fit algorithm. For quantal data it is usually assumed that the data follow a binomial distribution, and the common fit algorithm is based on maximising the binomial likelihood (see section 6.2.3 for a discussion of the assumptions). The parameter values produced by this algorithm are the values associated with the maximum likelihood, and are also called the Maximum Likelihood Estimates (MLEs).

278. Maximum likelihood can only be applied for data including at least two concentrations with partial responses, otherwise the MLE of the slope will tend to infinity. When the data only include 0% and 100% responses, or only a single concentration with partial response, the slope of the dose-response can therefore not be estimated. But there are several methods available for estimating the EC50 in those situations. These methods include procedures for assessing the precision of the estimated EC50 (Hoekstra, 1993).

Response in controls

279. Instead of estimating the background response (incidence) as a parameter in the dose-response model, Abbott's correction is often used in situations where dose-response data show nonzero observed response in the controls. In this correction, each observed response p_i is replaced by $(p_i - p_0) / (1 - p_0)$ where p_0 denotes the observed background response. However, this is inappropriate, since the observed background response p_0 contains error, which is not taken into account in this way. Instead the background response should be treated as an estimate containing error, just like the observed responses in the other dose groups. By incorporating the background response as a parameter in the model, it is estimated from the data, and estimation errors are accounted for, e.g. in calculating confidence intervals. As already discussed, it is theoretically impossible that the probability of response in the controls equals (strictly) zero. Therefore, the background response should be regarded as an unknown value, and be estimated from the data, even if the observed background response is zero (the fact that all observed control individuals did not respond does not imply that a response is impossible). Nonetheless, as Figure 6.4 illustrates, the background response may be estimated to be (virtually) zero, and in such situations fixing the background response at zero versus estimating it as a free parameter in the model does not make much difference (although the confidence intervals could be different, but probably not too much). Of course, one may always compare both ways of analyses in any practical situation. It should be noted that omitting the background parameter from the model has the advantage of one less parameter to be estimated (parsimony principle), but at the same time the observations in the control group are made worthless in that way.

280. In practice it may happen that the best fit of the model results in a negative estimate of the background response. To prevent this, the model should be fitted under the constraint that the background response must be nonnegative (i.e. $a \geq 0$). Instead of a negative estimate, the background response a associated with the best fit will, in those situations, then be estimated to be zero.

Analysis of data with various observed fractions at each dose group

281. Ecotoxicological (quantal) dose-response data often show replicated observed fractions at each concentration or dose group. For example, the individual organisms in each dose group may be housed in

[9] Available upon request (Wout.slob@rivm.nl)

different containers, each container resulting in an observed fraction of responding organisms. As another example, the fraction of fertile eggs may be observed in individual female birds, where each dose group consists of various female birds.

282. In more general terms, these designs have various experimental units per dose group, and in each experimental unit the fraction of responding sampling units is counted. Of course, a dose-response model can be fitted to such data by simply regarding the various observed fractions at each dose group as true replicates. In that case, it is assumed that the experimental units themselves (e.g. aquaria, of female birds) do not differ from each other.

283. If this cannot be assumed, the variability between experimental units must be taken into account in the statistical analysis. Here, two approaches are briefly mentioned. One approach is to apply a normalising (e.g. the square-root arcsine) transformation to the observed fractions related to each experimental unit. The transformed data can then be analysed as continuous data, as discussed in section 6.3. However, this approach is problematic for data with 0% and 100% responses. Another approach is to account for the among-container variation by adjusting the binomial distribution. For example, the parameter reflecting the probability of response in the binomial distribution may be assumed to follow a beta distribution (reflecting the variability among containers). This implies that the observed response is beta-binomial distributed rather than binomial, and the associated likelihood may be maximised (see e.g. Teunis and Slob, 1999). Section 5.2.2.4. gives a description of two methods for deciding whether extra biromila variation is present.

Analysis of data with one observed fraction at each dose group

284. When the study design has only one container per dose group, the analysis appears at first sight simpler as compared to the situation of replicated containers at each dose. However, this is apparent only. If the containers differ by themselves, this between-container variation will result in extrabinomial variation just as well. Theoretically, the variation among containers could be taken into account by the approaches mentioned above. However, experience with how this works in practical ecotoxicity data appears to be lacking.

Extrapolation and ECx

285. Because of the fact that a fitted statistical model only reflects the information in the data, extrapolation outside the range of observation is usually unwarranted. Consequently, an ECx that is estimated to be below the lowest applied (nonzero) dose should not be trusted.

Confidence intervals

286. Whatever definition for the ECx is used, it is estimated from the point estimates of the parameters in the model. When these point estimates are obtained by maximum likelihood, these are Maximum Likelihood Estimates (MLEs). The ECx is also a MLE when it is (indirectly) calculated from these values .

287. The MLE for the EC50, or any other ECx, is a point estimate only, and may, to a larger or smaller extent, be imprecise. The imprecision may be quantified by the standard error of the estimate, but it is more informative to calculate a confidence interval. A confidence interval indicates the plausible range for the parameter, e.g., a 95%-confidence interval is supposed to contain the true value of the parameter with probability 95%. Confidence intervals may be assessed in various ways:

- plus or minus twice the parameter's standard error (provided by most dose-response software), which is estimated by the second derivative of the likelihood function (Hessian or information matrix), possibly with Fieller's correction (Fieller 1954),

- based on the profile of the log-likelihood function, using the Chi-square approximation of the log-likelihood,
- value(s),
- bootstrap methods (see e.g., Efron 1987, Efron and Tibshirani, 1993),
- Bayesian methods, in particular if one has some preliminary knowledge on the plausible range of the parameter.

Various studies have compared the first three methods (see e.g. Moerbeek et al. 2004).

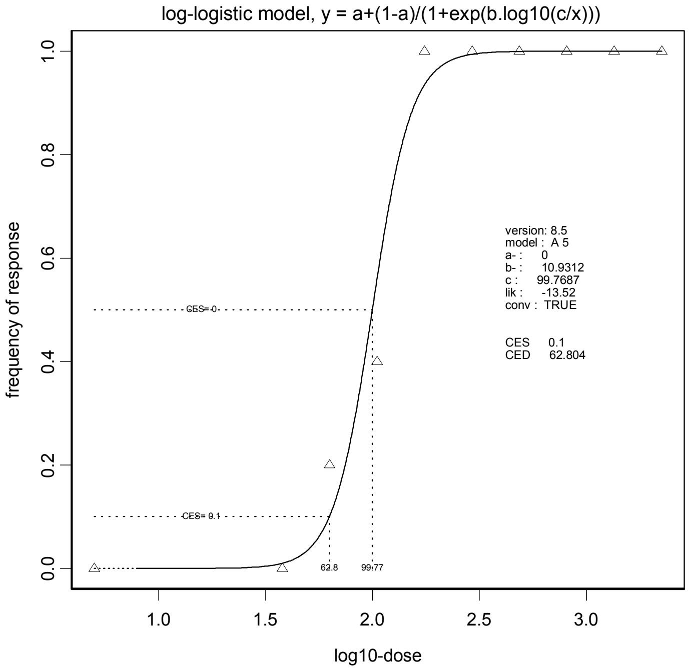


Figure 6.4 Logit model fitted to mortality dose-response data (triangles).

Here background mortality (parameter *a*) was included as a free parameter in the model, and estimated to be close to zero. The dashed lines indicate the LD50, and the LD10.

6.2.3. Assumptions

288. A dose-response model consists of a deterministic part (the predicted dose-response relationship) and a stochastic part (describing the noise). The assumptions made in the statistical part are analogous to those in hypothesis testing, and will only be briefly mentioned here. The focus in this chapter is on the additional assumption, that of the (deterministic) dose-response model.

Statistical assumptions

The assumptions for hypothesis testing equally hold for dose-response modelling:

- Binomial distribution for observations per experimental unit, i.e. independence between the animals in the same experimental unit (e.g. container). When the experimental unit is not accounted for in the statistical model, it is additionally assumed that experimental units do not vary among each other by themselves (i.e. at the same dose).

- No systematic differences (caused by unintended experimental factors) between dose groups (the latter is particularly relevant for unreplicated designs, i.e. one container per dose-group).
- The values of the concentrations/doses are assumed to be known without error, or, in situations where they are measured, the measurement errors are assumed to be negligible.

Additional assumption:

- The fitted model has a shape that is close to the true dose-response relationship

Evaluation of assumptions

Basic assumptions:

289. In designs with sample units (e.g. organisms, eggs) within experimental units (e.g. containers, female birds) the assumption of binomially distributed data may not be met, due to variation among the experimental units themselves. One way to check this is by fitting a model based on a betabinomial distribution, and comparing the associated log-likelihood with that obtained from a fit based on a binomial distribution. This comparison can be done by a likelihood ratio test, since the binomial and betabinomial distributions are nested.

Additional assumption:

290. Fulfilment of the assumption that the shape of the fitted model is close to the true dose-response relationship depends not only on the choice of a proper dose-response model, but also on the quality of the dose-response data. Therefore, one not only needs to consider if the model is suitable to describe the data, but also if the data are good enough to sufficiently guide the model in obtaining the right shape. For a fuller discussion of evaluating the shape of the fitted dose-response model, see section 6.4.

Consequences of violating the assumptions

Basic assumptions:

291. When the assumption of binomial distribution is not met, due to variation between experimental units, a fitted quantal dose-response model may result in a biased estimate of the ECx, as well as in too narrow confidence intervals.

Additional assumption:

292. Given that the data include both (close to) zero and (close to) 100% responses, violation of the assumption that the fitted model indeed reflects the true underlying dose-response relationship is less serious for an EC50 than for an ECx (the more so for lower values of x). The (point) estimate of the ECx may be inaccurate (biased), and the associated confidence interval may in extreme cases not even include the true value of the ECx. Therefore, it is not recommended to estimate an ECx if the fitted model appears not sufficiently confined by the data from visual inspection, or if it is found that various models fitting equally well result in different ECx estimates. In the latter case one might consider to construct an overall confidence interval for the ECx based on various models that fit the data equally well (if repeating the experiment, aimed at more concentrations with partial responses, is no option).

6.3. Dose-response modelling of continuous data (for a single exposure duration)

293. While a quantal response is based on the observation of whether or not each single organism (biological system) has a particular property (e.g. death, clinical signs, immobilisation), a continuous response is a quantitative measure of some biological property (e.g. body weight, concentration of enzyme). Such continuous response is measured in each experimental unit, and since organisms (biological systems) are never identical by themselves or not observed under identical conditions, the resulting data show a certain amount of scatter, depending on the homogeneity of the treatment group. This scatter may be assumed to follow a certain distribution, e.g. a normal, a lognormal, or a Poisson distribution.

294. Continuous data do not only differ from quantal data in a purely statistical sense (i.e. the underlying distribution). A more fundamental difference is that changes in response are interpreted in a completely different way. While the ECx in quantal responses relates to a change in response rate, an ECx in continuous responses relates to a change in the degree of the effect, as occurring in the average individual (of the population observed). For example, an IC10 in a fish test is associated with a 10% inhibition of the growth rate in the "average" fish (under the average experimental conditions).

295. The purpose of dose-response modelling of continuous data is to estimate the ECx, where x is any given percentage. When the dose-response data relate to a single exposure particular duration, the estimated ECx obviously only hold for that particular exposure duration (or to, e.g., a single acute oral dose).

Terms and notation

296. In this section the following terms and notations are used. The continuous response y is related to the dose (or concentration) x by function f:

$$y = f(x) \, .$$

297. In ecotoxicology the term ECx is defined as the concentration (or dose) associated with an effect x[10], where x is defined as:

$$x\% \;=\; 100\left(\frac{y(ECx)}{y(0)} \;-\; 1\right)\% \quad ,$$

i.e., x is defined as a percent change in the (average) level of the endpoint considered, *e.g*., a 10% decrease in weight.

298. In human toxicology different terms exist for the ECx. The equivalent terms are CED (Critical Effect Dose), which is equivalent to the ECx, and the CES (Critical Effect Size) which is equivalent to x in ECx (see e.g. Slob and Pieters, 1998). However, in human toxicology another approach has been proposed, which is based on a change in response rather than on a change in the degree of effect. In that approach (also called the hybrid approach) the terms BMD and BMR are used (e.g. Crump 1995, Gaylor and Slikker 1990), but these terms are not comparable to the ECx in continuous responses in ecotoxicology. The following table summarises the terms.

[10] Note that x is used for both concentration (dose) and the degree of effect.

	Ecotoxicology	human toxicology
Quantal response (x in terms of response)	x Ecx	BMR (benchmark response) BMD (benchmark dose)
Continuous response (x in terms of degree of effect)	x ECx (ICx)	CES (critical effect size) CED (critical effect dose)
Continuous response (BMR in terms of response)	- -	BMR BMD

6.3.1. Choice of model

299. A (statistical) dose-response model only serves to smooth the observed dose-response, to estimate an ECx by interpolating between applied doses, and to provide for a tool to assess confidence intervals. A statistical regression model itself does not have any meaning, and the choice of the model (mathematical expression) is largely arbitrary. Numerous dose-response models are theoretically possible, but in practice only a limited number is applied, mostly determined by historical habits in the field of application. A number of useful (families of) models will be discussed here.

300. A first distinction that can be made is linear versus nonlinear regression models. This distinction is made as the type of calculations is different between these two classes of models. In linear models the calculations are relatively simple, and could be done without a computer, which is hardly possible for nonlinear models. Clearly, given the widespread use of computers, this advantage has become more and more irrelevant, and nonlinear models are gaining attention, as they may be considered to more realistic for reflecting a dose-response relationship (see below). Yet, linear models will be briefly discussed, for the sake of completeness. After that, a number of other models (or family of models) will be discussed, most of which are nonlinear.

Linear models

301. Linear regression models are defined as models that are linear with respect to their parameters. They can be nonlinear with respect to the independent variable and thus not only include the straight line, but also quadratic, or higher order polynomials:

$$y = a + bx$$
$$y = a + bx + cx^2$$
$$y = a + bx + cx^2 + dx^3$$
$$etc.$$

302. These models have the property that the parameters (*a*, *b*, etc) in the model can be estimated by evaluating a single (explicit) formula (as opposed to nonlinear models, see below), which makes them relatively easy to apply. Another advantage is that these models are nested. For example, the quadratic model can be turned into a linear model by taking *c* = 0. Inversely, a linear model can be turned into a quadratic model by incorporating an additional parameter (here: *c*). It can be statistically tested if the addition of parameters leads to a significant improvement of the fit (e.g. by an F-test).

303. Linear models may be incorporated in the framework of GLM (generalized linear models), see e.g., Bailer and Oris (1997).

304. A disadvantage of linear models is that they are not necessarily strictly positive, while biological endpoints typically are (if the data are not pre-treated), which makes them theoretically implausible. Further, they are not necessarily monotone, which can result in doubtful results, especially in the situation of a limited number of dose groups.

Threshold models

305. A threshold model is a model that contains a parameter reflecting a dose-threshold, i.e. a dose below which the change in the endpoint is (mathematically) zero. In general, a threshold model is given by

$$y = a \quad \text{if } x < c$$
$$\mathrm{y} = \mathrm{a} + \mathrm{f(x - c)} \quad \text{if x > c} \qquad (6)$$

where *c* denotes the threshold concentration and *f*(*x*) may be any function. For example, in the ("hockey stick") model

$$y = a \quad \text{if } x < c$$
$$\mathrm{y} = \mathrm{a} + \mathrm{b\,(x - c)} \quad \text{if x > c}$$

the response is linear above the threshold. The threshold concentration could be called an EC0, i.e. an ECx with x=0. At first sight the threshold concentration appears attractive, as it avoids the discussion of what value of x in ECx is ecologically relevant. However, various objections can be raised against the use of threshold models. One of them is that the (point) estimate of the threshold can be dependent on the dose-response relationship, i.e. the function that is chosen for f(x) in expression (6).

Additive vs. multiplicative models

306. Strict continuous data (e.g. weights, concentrations) observed in toxicity studies usually have nonzero values in unexposed conditions, and the question then is to what extent the compound changes that level. Clearly, the compound interacts with that background level, by whatever biological mechanisms. This idea may be expressed in simple mathematical terms by incorporating the background level (*a*) in the dose-response model in a multiplicative way:

$$f(x) = a \cdot \mathrm{g}(x) \quad (7)$$

rather than in an additive way:

$$f(x) = a + g(x) \quad (8)$$

as is more common in models discussed in statistical textbooks. (Note that the models based on quantal models discussed in the previous section are also additive). Of course, the whole idea of defining the ECx as a given *percent* change compared to the background level, is in concordance with the multiplicative interaction between compound and background level, as expressed in (7). A further convenience of the multiplicative model is that two populations (e.g. species, sexes) showing different background levels but equally sensitive to the compound are, in this way, characterised by the same *g*(*x*). This implies that in the multiplicative model two equally sensitive populations (but possibly with different background levels) are defined to have the same ECx.

Models based on "quantal"models

307. Continuous dose-response data from ecotoxicity tests have often been described by dose-response models that are derived from the models used for quantal data, i.e. models whose predicted values range from zero to one. To make these models applicable to continuous models, they are usually adjusted as follows:

$$y = y(0) + [y(\propto) - y(0)]\,f(x)$$

for increasing dose-responses, and

$$y = y(\propto) + [y(0) - y(\propto)]\ [1 - f(x)]$$

for decreasing responses (see, e.g., Bruce and Versteeg 1992; Scholze et al., 2001). Here, $y(0)$ is the (predicted) background value, $y(\propto)$ is the (predicted) value at infinite dose, and $f(x)$ is any quantal dose-response model. Note that these models are multiplicative (with respect to the background response), while their shape is typically sigmoidal. As an example, when the logit model is chosen for $f(x)$, the associated model for the continuous data becomes

$$y = y(0) + \frac{y(\infty) - y(0)}{1 + \exp(b\ln(ED50/x))}$$

308. In current practice it is common to correct the data for the background response, and fit the model without a background parameter. As discussed in section 4.3.4, this procedure of pre-treatment of the data ignores the estimation error in the observed background, and is therefore unsound. By incorporating the parameter $y(0)$ in the model to be fitted, the estimation error is taken into account, and therefore this approach should always be taken.

Nested nonlinear models

309. Slob (2002) proposed to use the following nested family of multiplicative nonlinear models for general use in dose-response modelling.

model 1: $y = a$ *with* $a > 0$
model 2: $y = a\exp(x/b)$ *with* $a > 0$
model 3: $y = a\exp(\pm(x/b)^d)$ *with* $a > 0, b > 0, d \geq 1$
model 4: $y = a\,[c - (c - 1)\exp(-x/b)]$ *with* $a > 0, b > 0, c > 0$
model 5: $y = a\,[c - (c - 1)\exp(-(x/b)^d)]$ *with* $a > 0, b > 0, c > 0, d \geq 1$

where y is any continuous endpoint, and x denotes the dose (or concentration) . In all models the parameter a represents the level of the endpoint at dose zero, and b can be considered as the parameter reflecting the efficacy of the chemical (or the sensitivity of the subject). At high doses models 4 and 5 level off to the value ac, so the parameter c can be interpreted as the maximum relative change. Models 3 and 5 have the flexibility to mimic threshold-like responses (i.e. slowly changing at low doses, and more rapidly at higher doses). All these models are nested to each other, except models 3 and 4, which both have three parameters.

310. In all models the parameter a is constrained to being positive for obvious reasons (it denotes the value of the endpoint at dose zero). The parameter d is constrained to values larger than (or equal to) one, to prevent the slope of the function at dose zero being infinite, which seems biologically implausible. The parameter b is constrained to be positive in all models. Parameter c in models 4 and 5 determines whether

the function increases or decreases, by being larger or smaller than unity, respectively. To make model 3 a decreasing function, a minus sign has to be inserted in the exponent.

311. These models have the following properties:

- the predicted response is strictly positive
- they are monotone (i.e., either decreasing or increasing)
- they do not contain a threshold, but they are sufficiently flexible to show strong curvature at low doses, so as to mimic threshold-like responses
- they can describe responses that level off at high dose
- two populations that differ in background level but are equally sensitive can be described by the same model, with only parameter *a* being different between the populations
- it can be easily tested if two populations differ in sensitivity (by the likelihood ratio test)
- when two populations differing in sensitivity can be described by the same model from this family, with only parameter *b* (and possibly *a*) being different between the two populations, the difference in sensitivity can be quantified as the ratio of the value of *b*. This way of expressing differences in sensitivity is analogous to the relative potency factor, and to the extrapolation factors used in risk assessment.

312. For all five models, the ECx can be derived by evaluating an explicit formula:

$$ECx = \{ -\frac{\ln[(x+1-c)/(1-c)]}{b} \}^{1/d}$$

where x is defined as $x = y(ECx)/a - 1$, and where $c = 0$ for models 2 and 3, and $d = 1$ for models 2 and 4.

313. Clearly, the five multiplicative models given here only apply for those endpoints that are strictly positive and have a nonzero background value (value of *y* in unexposed conditions). For example, describing internal concentration as a function of external concentration is not possible with these models, as in that case *y* is expected to be zero for *x* = 0.

314. The procedure of selecting a model from this nested family of models, i.e., accepting additional parameters only when it results in a significantly better fit, is illustrated in fig. 6.6. In this dataset, the following log-likelihoods were found:

model 1: 277.02
model 2: 339.90
model 3: 339.90
model 4: 351.11
model 5: 351.11

315. Model 3 resulted in exactly the same fit[11] as model 2, while model 5 resulted in the same fit as model 4. But model 4 is significantly better than model 2 (critical difference is 1.92 at $\alpha = 0.05$, according

[11] Adding a parameter to a model can, by definition, not result in a lower (optimum) log-likelihood. When the log-likelihood remains the same, the additional parameter is estimated at the value that makes it disappear. In this case the parameter *d* was estimated to be one.

to the likelihood ratio test), and therefore model 4 should be selected for this dataset. (Note that model 3 and model 4 are not nested, they both have three parameters).

Hill model

316. Enzyme kinetics and receptor binding are usually described by the Hill model. It was introduced by A.V. Hill in 1910 in order to model the binding of oxygen to haemoglobin. The model is well known by enzymologists, biochemists and pharmacologists, and could be considered as one of the very few examples of a mechanistically based model. It has the form:

$$y = \frac{a\,x^c}{b + x^c}$$

where *c* is called the Hill parameter. By setting *c* = 1, it is equivalent to the Michealis-Menten expression in a strict sense, with *a* denoting the maximum level of *y* at infinite dose, and *b* the ED50 (dose resulting in half the maximum response).

317. The following formulation makes more sense for toxicology since the parameter noted b in the draft standard is actually a thermodynamic equilibrium dissociation constant Kd that can be changed as EC_{50}^{n} which is more familiar to toxicologists and is homogenous to a concentration (or dose):

$$y = y(0) + [y(\infty) - y(0)]\frac{x^n}{EC_{50}^n + x^n}$$

318. It is worth noticing that the Hill model is analytically equivalent to the logit model:

$$y = \left[1 + e^{n\ln\left(\frac{EC_{50}}{x}\right)}\right]^{-1} = \left[1 + \left(\frac{EC_{50}}{x}\right)^n\right]^{-1} = \frac{x^n}{EC_{50}^n + x^n}$$

319. It should be noted that dose-response data observed in *in vivo* studies are not the result of a single underlying receptor binding process, but of many processes acting simultaneously. Yet, it may be a very accurate model for describing particular data, see e.g. Fig. 6.15.

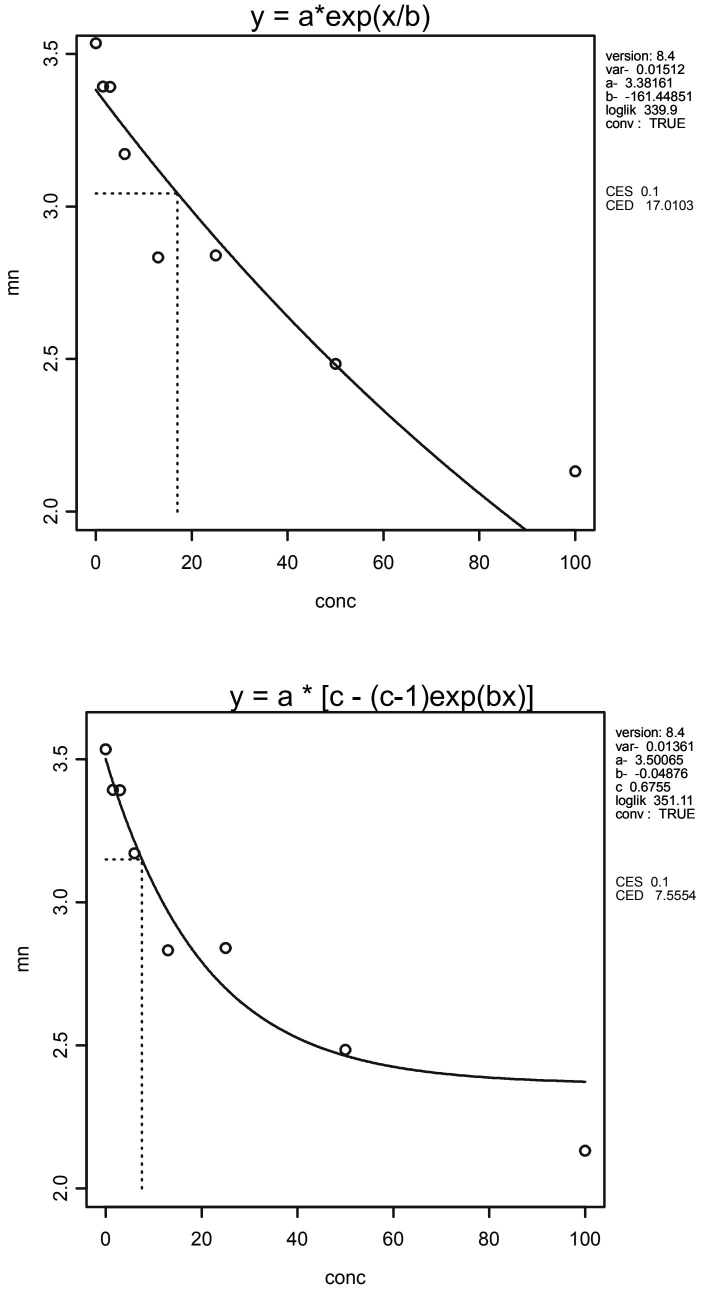


Figure 6.6 Two members from a nested family of models fitted to the same data set.

The marks indicate the observed (geometric) means of the observations. The exponential model (upper panel) is significantly improved by adding a parameter c, enabling the response to level off (lower panel).

Non monotone models

320. In some cases dose-response data appear to be non monotone. Unfortunately, it is not easy to assess if this is due to an underlying dose-response relationship that is indeed non monotone. It is not unlikely that an apparent non monotone dose-response in observed data is due to experimental artefacts, either systematic errors in unreplicated dose groups, or simply random noise. Although the latter possibility can be checked by statistical methods, the former cannot. Therefore, when the apparent monotonicity is based on a single treatment group, no unambiguous conclusion can be drawn. Only multiple dose studies with a

clear non monotone pattern, supported by various consecutive dose groups, may provide evidence of a real non monotone response.

321. When it is assumed that the data do not contain any systematic errors, the straightforward way to test for non monotonicity is by fitting a non monotone model to the data, and compare the fit with a nested model that is monotone (for an example of a nested non monotone model, see Brain and Cousens 1989, or Hoekstra 1993). If the non monotone model appears to be significantly better, it may still be doubtful if this particular model reflects the true dose-response relationship. The practical difficulty is that non monotone models are very data demanding, in particular with respect to the number of consecutive dose groups around the local maximum (or minimum) of the response. Otherwise, the location and height of the local maximum response will be highly model dependent. Therefore, fixation of the local maximum response requires the enclosure by sufficiently close adjacent dose groups. Since the location of the local maximum response is not known in advance, the study design would require a large number of dose groups. Therefore, when non-monotone dose-response relationships may be expected (as in plant grow data), a larger number of dose groups needs to be incorporated in the study design.

322. Of course dose-response models include more parameters to be estimated, and this is another reason that many dose groups are required. In most practical data sets various non monotone models would give different results, and therefore can often not be trusted.

6.3.2. Model fitting and estimation of parameters

323. Fitting a model to dose-response data may be done by using any suitable software, e.g. SAS (www.sas.com), SPSS (www.spss.com), splus (www.insighful.com), and PROAST (Slob, 2003). The user does not need to be aware of the computational details, but some understanding of the basic principles in nonlinear regression is required to be able to interpret the results properly. These principles are discussed in section 6.7. Furthermore, the user should be aware of the assumptions underlying the fit algorithm. For continuous data it is often assumed that the data follow a normal or a lognormal distribution. In the latter case, a log-transformation is used to make the data (more closely) normally distributed. When a normal distribution with homogenous variances is assumed (possibly after transformation), maximising the likelihood or minimising the Sum of Squares amounts to the same thing (see section 4.3.5). When another distribution is assumed (e.g. a Poisson for counts), the model may be fitted by maximum likelihood, based on the particular distribution assumed. The parameter values produced by maximum likelihood are also called the Maximum Likelihood Estimates (MLEs).

Response in controls

324. In all models discussed here, the background response is incorporated as a model parameter in the model. This parameter should be estimated from the data, before deriving the ECx or ICx. Pre-treatment of the data (dividing all responses by the mean background response) should be avoided (see also section 4.3.3).

Fitting the model assuming normal variation

325. When the original data are assumed to be normally distributed with homogenous variances the model may be fitted by either maximising the log-likelihood function based on the normal distribution, or by minimising the sum of squares. Both methods will result in the same estimates of the regression parameters, which are maximum likelihood estimates (MLEs) in both cases. The fitted model describes the arithmetic mean response, as a function of dose.

Fitting the model assuming normal variation after log-transformation

326. When the residual variation is assumed to be lognormal, the model may be fitted after first log-transforming both the model predictions and the data, and then either maximising the log-likelihood function based on the normal distribution, or minimising the sum of squares. Both methods will result in the same estimates of the regression parameters and the residual variance, which are maximum likelihood estimates (MLEs) in both cases. It should be noted that the resulting parameter estimates do relate to the original parameters of the (untransformed) model. Substituting the estimated regression parameters in the model results in a prediction of the median (or geometric mean) response as a function of dose. Therefore, in plotting the model together with the data, the back transformed means (which are equivalent to the geometric means) should be plotted (see e.g. fig. 6.5).

327. While the MLEs of the regression parameters relate to the model on the original scale, the MLE of the variance (s2) relates to the log-transformed data. Apart from this variance (s2) on log-scale, the variation of the scatter around the model (i.e. of the regression residuals) may be equivalently reported by the geometric standard deviation (GSD), which is the back transformed square root of s^2 , or by the coefficient of variation (*CV*), which relates to s^2 by

$$CV = \sqrt{\exp(s^2) - 1},$$

when s^2 relates to the variance of the data after natural log-transformation, or by

$$CV = \sqrt{\exp[s^2 \ln(10)] - 1}$$

when the $\log_{10}$-transformation was applied to the data.

328. At first sight, a disadvantage of taking the logarithm of the data before fitting is that the logarithm of zero does not exist. Although zero observations for continuous responses rarely occur in ecotoxicity testing, the following may be noted. Zero observations usually mean that the response is below the detection limit rather than truly zero. By regarding zero observations as truncated observations, they can be easily and accurately dealt with by incorporating the information that the observation is lower than the detection limit in the log-likelihood function.

Fitting the model assuming normal variation after other transformations

329. When another transformation is applied to the data, the same transformation should be applied to the model, before maximising the likelihood (or minimising the SS). Both the fitted model and the transformed data may be back-transformed before plotting. Again, the resulting plot relates to the predicted and observed *median* response, as a function of dose (assuming that the transformation made the scatter symmetrical).

No individual data available

330. In reported studies (published papers) individual observations are not always given. Instead, means and standard deviations (or standard errors of the mean) for each dose group are commonly reported. Since the mean and standard deviation are “sufficient” statistics for a sample from a normal distribution, a dose-response model can just as well be fitted based on these statistics without any loss of information (except possible outliers), by adjusting the log-likelihood function (Slob, 2002). In the case of an assumed lognormal distribution, sufficient statistics are provided by the geometric mean and the geometric standard deviation, or by the (arithmetic) mean and standard deviation on log-scale. These can be estimated from

the reported mean and standard deviation (Slob, 2002). Figure 6.6 exemplifies a dose-response analysis applied to the reported means and standard deviations, without knowing the individual data, but taking the reported standard deviations into account.

Fitting the model using GLM

331. Since the log-likelihood function directly derives from the postulated distribution, one may theoretically assume any distribution, and apply maximum likelihood for fitting the model based on that assumption. For a number of distributions (the so-called exponential family of distributions) one may make use of the theory of Generalised Linear Models (GLM), and use existing software without deriving and programming one's own formulae. The GLM framework is also useful for analysing data with replicated concentration groups.

332. The Poisson distribution is a member of this exponential family, and the existing GLM software can be directly used. Thus, one may assume this distribution for the analysis of counts, and check if the distribution is reasonable.

333. The gamma distribution is another example of a distribution belonging to the exponential family. This distribution can be directly dealt with by the existing (GLM) software (e.g. in SAS, SPSS, splus. The gamma distribution is very similar to the lognormal distribution regarding its behaviour of describing the variation in data. Therefore, an analysis based on either one of these two distributions may be expected to give very similar results. However, there are a few differences. While an analysis based on the lognormal distribution results in a model describing the median response (as estimated by the geometric means), an analysis based on the gamma distribution describes the response in terms of the statistical expectation (as estimated by the arithmetic means). Therefore, the latter fitted model will, on the whole, lie at a lower level than the former because the mean is larger than the median (the more so for larger experimental variation, i.e. more skewed scatter). However, both analyses may be expected to result in similar point estimates for the ECx: the difference in level will cancel as the ECx is a ratio of the two medians, or of the two mean levels, respectively. A second difference is, that the analysis based on the gamma distribution results in an estimate of the residual variation in terms of the variance on the original scale, while the analysis based on the lognormal distribution the residual variation is estimated in terms of a C.V. (or a GSD: geometric standard deviation).

Covariates

334. In many studies not only the concentration is varied systematically. Other factors also may bevaried intentionally as part of the design. For example, a chemical is studied under various conditions, e.g. temperature, pH, or soil condition. Instead of fitting a model to each subset of data, it is often possible to fit the model simultaneously to the whole data set, by letting a particular parameter (possibly more) depend on that covariate. Such an analysis is illustrated in Fig. 6.7, where AChE inhibition was measured at three points in time, i.e. at three different exposure durations. Here, a four-parameter model was fitted, two of which were allowed to depend on duration. Thus a total of nine parameters was estimated, while a separate analysis for each duration would have resulted in a total of 15 estimated parameters (three times 4 regressions plus one variance parameter). The gain of this is that the resulting confidence intervals for the ECx estimates are smaller.

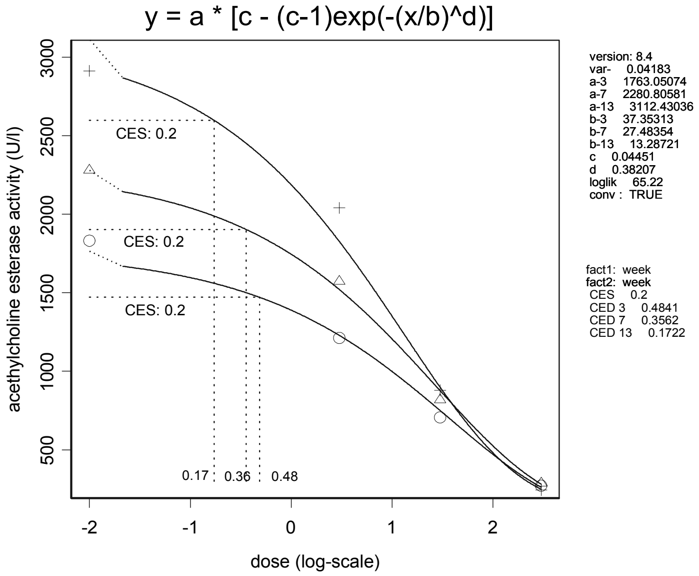


Figure 6.7 Cholinesterase inhibition as a function of dose at three exposure durations (triangles: three weeks, circles: 7 weeks, pluses: 13 weeks).

Marks denote the geometric group means, the individual observations are not plotted here. The background AChE levels increase with duration (age), while the ECx (CED for CES=0.20) decreases with exposure duration. The model used is model 5 from the nested family of models proposed by Slob(2002).

Heterogeneity and weighted analysis

335. In concordance with the principle of parsimony (as discussed in e.g. section 6.1) it is favourable to assume homogenous variances between dose groups: in this way only one single parameter for the residual variance needs to be estimated. However, it should be noted that the term "homogenous variances" is closely associated with the normal distribution. When other distributions are assumed, the variances are generally not expected to be homogenous, e.g.:

- For lognormally (or Gamma) distributed data, variances increase with the means (more specifically, CVs are predicted to be constant), and this heterogeneity should vanish when the data are log-transformed. Thus, it may be assumed that (on the original scale) the CVs are homogenous, and the statistical analysis would result in a single estimate of the CV.
- For Poisson distributed data (counts) the variances also increase with the mean. In fact they should be equal to the means, and if the data confirm this, no variance parameter needs to be estimated. In practice, this assumption is often violated, with the variances being larger than the means. This is called extra-Poisson variation, and an extra parameter may be estimated expressing the proportionality constant between mean and variance.

336. Apart from statistical reasons (the parsimony principle), the issue of homogenous variances should also be considered for biological reasons. It might be that the organisms did not respond equally to the compound due to variability in sensitivity, and this will be reflected in the variances. It is not easy to discriminate between statistical heterogeneity (distribution effects) and biological heterogeneity ("true" effects). For that reason (among others), it is important to carefully consider what distribution should be assumed, e.g. by using historical data on the same (or similar) endpoint examined for other chemicals (or treatments).

337. When the heterogeneity of variances cannot be explained by the underlying distribution, one might conclude that the responses themselves are heterogeneous. Statistically, this implies that the precision of the estimated group means is not the same among groups. This may be taken into account in the statistical analysis by using a weighted analysis, e.g. weighted least squares, where the squares are multiplied by a weight, usually the inverse of the standard deviation of the relevant group, or by using maximum likelihood where a variance is estimated for each separate group[12]. For a more extensive discussion see e.g. Scholze et al. 2001.

338. A weighted analysis should result in a more efficient estimate of the mean response (as a function of dose) in situations where the data are considered to reflect the same underlying response, and the heterogeneity is due to differences in measurement errors. The interpretation of a mean response is, however, problematic when the heterogeneity reflects that the population responds heterogeneously, in particular when this is caused by distinct subpopulations that differ in response. As an example, consider fig. 6.8, where relative liver weights are plotted on the log-scale (since for this endpoint the scatter is normally proportional to the mean level). In this particular example, the scatter first decreases, then increases with the dose. This might lead one to perform a weighted analysis (e.g. weighted least squares). However, as fig. 6.9 shows, the heterogeneity in variances is caused by different responses in males and females. Fitting the model taking sex into account results in two different dose-response relationships, each with homogenous scatter around it.

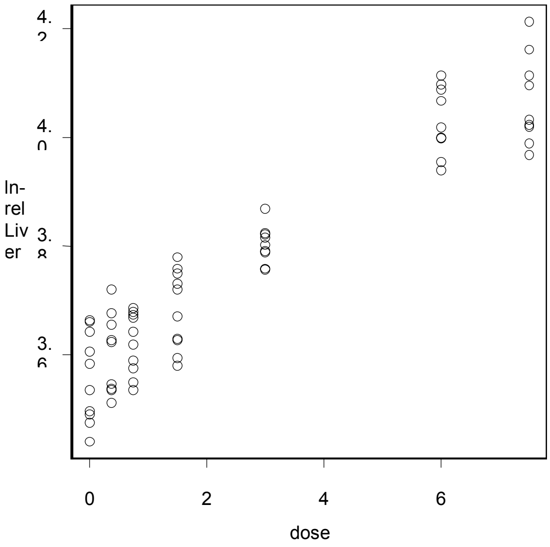


Figure 6.8 Relative liver weights against dose, plotted on log-scale.

Normally, relative liver weights show homogenous scatter in log-scale, but in these data the scatter first decreases, then increases with dose.

[12] When the heterogeneity in response changes systematically with the dose in a way that cannot be explained by the underlying distribution, one may also incorporate a dose-response relationship for the variation parameter in the likelihood function.

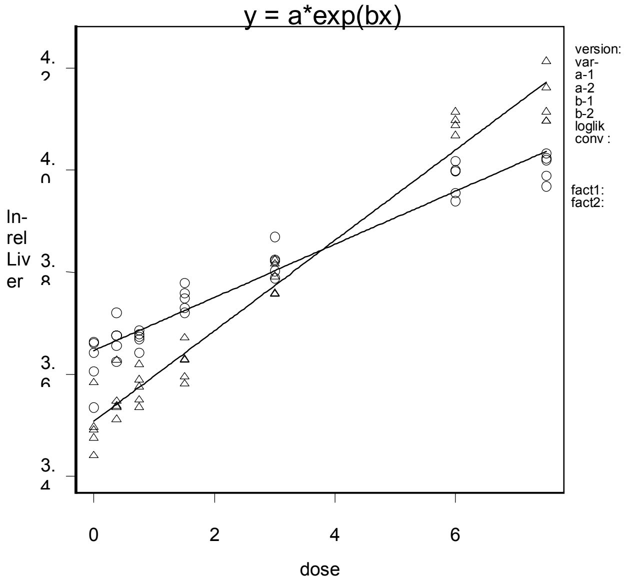


Figure 6.9 Dose response model fitted to the data of fig. 7.6a, showing that the heterogeneous variance was caused by males (triangles) and females (circles) responding differently to the chemical.

Confidence intervals

339. Confidence intervals may be assessed in various ways:

- the delta method, i.e. plus or minus twice (or the relevant standard normal deviate times) the standard error as estimated by the second derivative of the likelihood function (Hessian or information matrix); the standard errors of the parameters are provided by most dose-response software,
- based on the profile of the log-likelihood function, using the Chi-square approximation of the log-likelihood,
- bootstrap methods (see e.g., Efron 1987, Efron and Tibshirani, 1993),
- Bayesian methods, in particular if one has some preliminary knowledge on the plausible range of the parameter value(s).

340. The relative performance of the first three methods applied to a typical toxicological dataset (from a rodent study) has been examined by Moerbeek et al. (2004). In this study the second and third method resulted in similar intervals, while the first method appeared less accurate.

Extrapolation

341. Because of the fact that a fitted statistical model only reflects the information in the data, extrapolation outside the range of observation is usually unwarranted. Therefore, estimating an ECx that is much lower than the lowest applied (nonzero) dose or concentration should be avoided.

Analysis of data with replicated dose group

342. The individual organisms in each dose group may be housed in different containers. In that case, the individual observations may not be independent, due to systematic differences between the containers themselves. A straightforward and relatively simple approach for analysing such data is to follow two

steps. In the first step the model is fitted as though the data were independent (i.e. the observations from various containers at the same dose are taken together and treated as a single sample). Then, the residuals from the fitted model are calculated and these are subjected to a nested analysis of variance, resulting in an estimate for the (residual) variance within as well as among the containers. Strictly, the first step of this method is not completely valid, as it assumed independence between the observations. However, the results would normally not be much different (especially so for more or less balanced designs)

343. One may also fit a mixed model to the data, i.e. a model that contains both the (systematic) dose-response relationship and the random variation between containers.

344. These analyses will result in an estimate of the variation among containers, and the residual variation within containers.

345. In studies without replicated dose groups, the variation between containers will be incorporated into the residual variance. Theoretically, the variation between containers can still be estimated in designs with a sufficient number of dose groups, but practical experience with real toxicity data appears to be lacking.

6.3.3. Assumptions

346. A dose-response model consists of a deterministic part (the predicted dose-response relationship) and a statistical part (describing the noise). The assumptions made in the statistical part are analogous to those in hypothesis testing, and only be briefly mentioned here. The focus in this chapter is on the additional assumption, that of the (deterministic) dose-response model.

Statistical assumptions

The assumptions for hypothesis testing equally hold for dose-response modelling:

- independence between the animals in the same experimental unit (e.g. container)
- no variation between experimental units (e.g. containers) themselves, if they are not incorporated in the statistical analysis
- a particular statistical distribution and variance structure for the residual variation, e.g.,
  - normal distribution with homogenous variance
  - lognormal distribution with homogenous Coefficient of Variation (CV)
  - gamma distribution with homogenous Coefficient of Variation (CV)
  - Poisson distribution without variance parameter, or with additional parameter for extra-Poisson variation
- no systematic differences (due to unintended experimental factors) between dose groups,
- the values of the concentrations/doses are assumed to be known without error, or, in situations where they are measured, the measurement errors are assumed to be negligible.

Additional assumption:

- the shape of the fitted model is close to the true dose-response relationship.

Evaluation of assumptions

347. The statistical assumptions are similar to those in hypothesis testing, and may be further checked by plotting (analysing) the residuals (see section 4.3.5 and 5) However, the additional assumption (acceptance of the fitted dose-response model) is the most important, and the reader should first of all read and understand section 6.4.

Consequences of violating the assumptions

Basic assumptions

348. Violation of the assumptionthat containers do not vary amongst each other, while this variation is not taken into account in the statistical analysis, it does not have much impact on the point estimate of the ECx (in particular when the number of replicates is similar between dose groups). It does, however, distort the estimate of the confidence interval, which will be too narrow.

349. Systematic differences between (unreplicated) dose groups, caused by some unintended experimental factor, may have a deforming effect on the fitted model, and thereby result in a biased estimate of the ECx. However, especially for multiple dose designs, the effect may be small: systematic deviations in particular dose groups are, to a greater or lesser extent (depending on the situation) mitigated by the other dose groups in the process of fitting a single dose-response model to the complete data set. To prevent systematic errors between dose groups as much as possible, attention should be paid to applying randomisation procedures in the study protocol (see also section 4.2.1).

350. If one suspects that experimental units (e.g., containers) vary by themselves, then one should incorporate replicated dose groups in the design (e.g. various containers per dose group), or increase the number of dose groups (keeping one container per dose). In both designs the container effect can be estimated, although in the latter design this can only be done indirectly and may be difficult in practice.

351. A dose-response model is often relatively insensitive to outliers. See Fig. 6.10 for an illustration.

Additional assumption

352. Violation of the assumption that the shape of the fitted model is close to the true dose-response relationship results in a biased estimate of the ECx. There is no remedy against violation of this assumption, other than to repeat the study with an improved design. Therefore, it is not recommended to estimate an ECx if the fitted model appears not sufficiently confined by the data from visual inspection, or if it is found that various models fitting equally well result in different ECx estimates. In the latter case, one might consider to construct an overall confidence interval for the ECx based on various models that fit the data equally well (if repeating the experiment, aimed at more concentrations with different response levels, is no option).

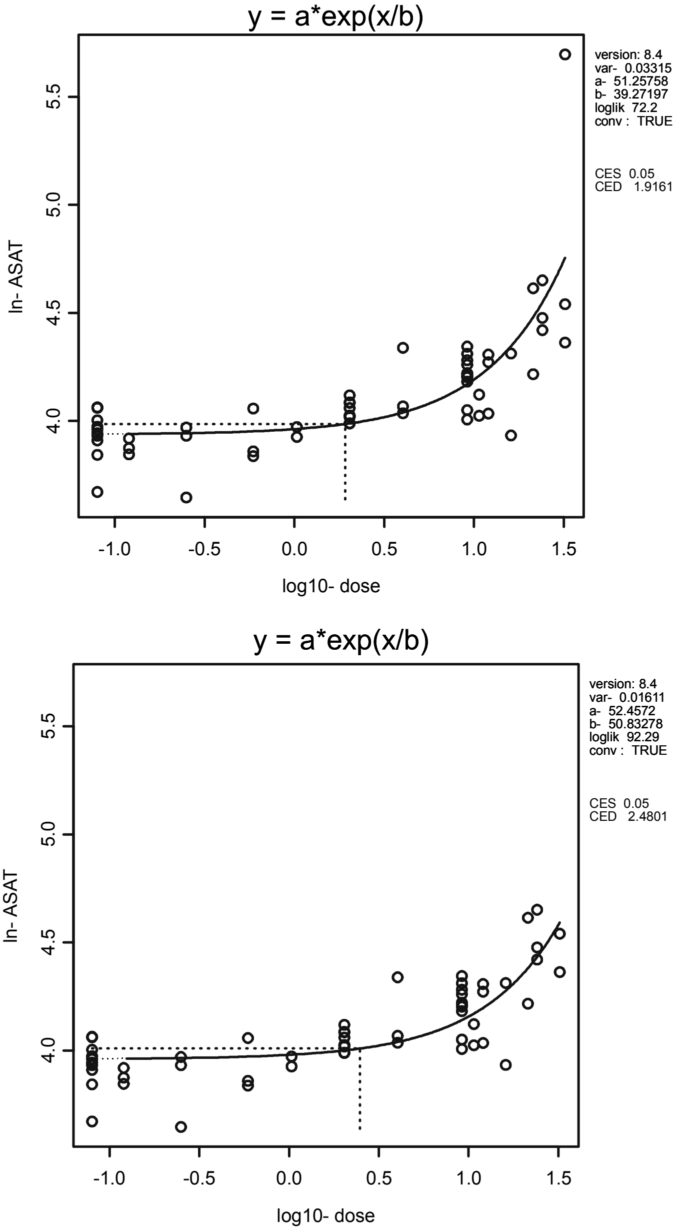


Figure 6.10 model fitted to dose-response data with and without an outlier in the top dose.

Note that the estimate of the ec05 (ced at ces=0.05) is only mildly affected, even though the outlier is in the top dose. The 90%-confidence interval was estimated at (1.63, 2.30) with the outlier included, and at (2.12, 2.93) when excluded.

6.4. To accept or not accept the fitted model?

353. A fundamental issue in dose-response modelling is the question: Can the fitted model be accepted and be used for its intended purpose (such as estimating an ECx)? The issue is not that the model used should be the "right" model, since there is no such thing (at least not for statistical models). A statistical model completely hinges on the dose-response data, and the quality of the data is in fact the crucial aspect. In the fitting process a model tries to hit the response at the observed doses. But when it is used for assessing an ECx by interpolating between observed doses, the model should also "hit" the response in the non-observed dose range in between. In other words, there are two aspects in evaluating the fitted model: one should not only assess if the model succeeded in describing the observed responses, but also if the

model can be trusted to describe the non-observed responses in between. The former aspect focuses on the quality of the model, the latter on the quality of the data. The following discussion indicates how to deal with these two aspects. It should be noticed that this discussion holds for both quantal and continuous dose-response data.

Is the model in agreement with the data?

354. This question may be addressed using the goodness-of-fit. Goodness-of fit methods can be used in an absolute or in a relative sense. In an absolute sense one may test if the data significantly deviate from a particular model. It should be noted that this test is sensitive not only to the inadequacy of the model chosen, but also to any violations of the basic assumptions (e.g. no independent observations, outliers). In particular, a single deviating concentration group (due to some unknown experimental factor) could make the model be rejected significantly even when it perfectly follows the overall trend in the data. Therefore, the (absolute) goodness-of-fit test should never be strictly applied. A visual check of the data is always needed and may overrule a goodness-of-fit test.

355. The goodness-of-fit may also be used in a relative way, i.e. to compare the fits of different models. When models are nested (as discussed in section 6.2.1 and 6.3.1), the likelihood ratio test can be applied to determine the number of parameters needed for describing the data. For non-nested models one may use the Aikaike criteria (Akaike, 1974; Bozdogan, 1987), but this test is not exact.

356. It has been suggested to focus the goodness of fit to the region of interest (around the ECx). This approach in a sense undermines the whole idea of dose-response modeling, i.e. describing the dose-response relationship as a whole. In particular, it will be more sensitive for (systematic) errors in the data that happen to occur in one of the dose groups in the range of interest. As discussed in section 6.3.3, one of the advantages of dose-response modeling is that potential systematic errors in a single dose group may be mitigated by the others.

Do the data provide sufficient information for fixing the model?

357. This question is at least as important as the previous. Therefore, the fitted dose-response model should always be visually inspected, not only to see if the data are close to the model, but also to check if the data provide sufficient information to confine the model. Here, one should ask the question: If additional data on intermediate dose groups had been available, could that substantially have changed the shape of the dose-response relationship as compared to the current fitted model? (See also section 4.3.5).

358. Another way to deal with this question is by comparing the outcomes from different fitted models. If the data do contain sufficient information to confine the shape of the dose-response relationship, different models fitting the data (nearly) equally well, will result in similar fits and similar estimates of the parameters. To illustrate this (for the case of quantal data), the results of Fig 6.1-6.3 are summarised in Table 6.13. In this case, the results are quite independent from the model chosen, and one may conclude that the data provide sufficient information to rely on dose-response modelling.

Model	*a* (background response)	LD50	LD20	confidence interval of LD20 [1)]	Log-likelihood
Probit	0.0355	0.2564	0.165	0.112 – 0.217	-34.01
Logit	0.0356	0.2554	0.167	0.121 – 0.220	-34.16
Weibull	0.0352	0.2625	0.145	0.084 – 0.218	-34.02

Table 6.13 : results of fitting three different models to the same data set (see fig. 6.1-6.3).

[1)] confidence intervals based on 1000 parametric bootstrap runs.

359. As another illustration, Fig. 6.11 shows two different models fitted to the same (continuous) data. Again, due to the good quality of the data, they result in very similar estimated concentration-response relationships, and therefore in a similar (point) estimate of any ECx. In situations where the results (in particular, the ECx) depends on the model chosen, it cannot be considered as a reliable estimate, and other methods should be considered (see section 4.1)

360. In current practice, there is a tendency to focus on the first part and a formal goodness-of-fit test is often regarded as (sufficient) evidence that the model is acceptable. This is unfortunate, since a goodness-of fit test tends to be more easily passed for data with few dose groups, and exactly in that situation the second condition is more likely not to be met . In addition, a goodness-of-fit test assumes that the experiment was carried out perfectly, i.e. perfectly randomised with respect to all potentially relevant experimental factors and actions. Clearly, this assumption is not realistic.

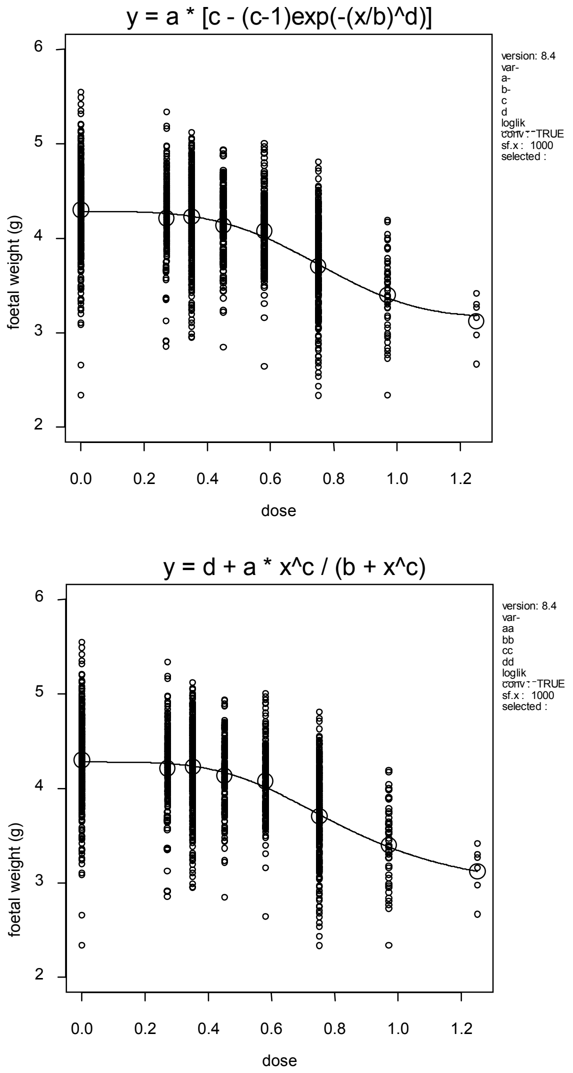


Figure 6.11 Two different models (both with four parameters) fitted to the same data set resulting in similar dose-response relationships.

Small marks indicate individual observations, large marks (geometric) means.

361. The ideas discussed here are further illustrated (theoretically) in Fig. 6.12. In the left panel, the data are insufficient to establish the dose-response relationship, leaving too much freedom to the model. In the right panel, the data are sufficiently informative to confine the shape of the dose-response relationship.

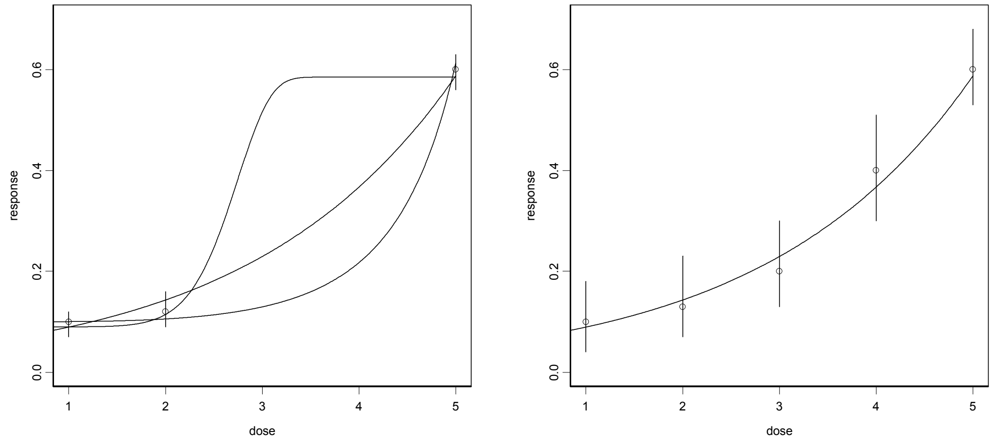


Figure. 6.12 Two data sets illustrating that passing a goodness of fit is not sufficient for accepting the model.

In the left panel the data (either quantal or continuous) do not contain sufficient information to confine the dose-response relationship, in the right panel they do. These figures also illustrate that more dose groups is more important than higher precision (indicated by vertical error bars): although the precision of the ECx estimate will be lower in the left panel, it is more likely to be biased. Note: dose group number 1, as indicated on the abscissa, may be read as the control group in these plots.

362. A number of general guidelines may be formulated in choosing and accepting a particular model for describing the dose-response data:

- When one of two nested models results in a significantly better fit, choose that model, otherwise the one with fewer parameters. One more parameter in the model can be regarded to result in a significantly better fit (at $\alpha = 0.05$) if the log-likelihood is increased by at least 1.92 (which is half the critical Chi-square value with one degree of freedom at $\alpha = 0.05$). One may also follow this procedure as a proxy for non-nested models (or use the Aikaike criteria).
- When two (or more) models have the same number of parameters, but one of them has a better goodness of fit, the choice of the better fitting model is obvious. However, if one prefers for some reason the other model, one may use Aikaike's criteria to compare the model fits (Akaike, 1974; Bozdogan, 1987).
- When two models result in a similar goodness of fit, but their shapes are very different (resulting in different estimates of the ECx) no conclusion can be made other than the data being inconclusive. In this situation it is not recommendable to derive an ECx based on dose-response modelling.
- The situation that two (or more) models show a similar goodness of fit, both being similar in shape (resulting in similar ECx estimates), can be considered as a confirmation that the data provide sufficient information to assess the dose-response relationship, and estimate the ECx.

363. It is re-emphasised that a dose-response model, as long as it is not based on the mechanism of action of the particular chemical, only serves to smooth the observed dose-response relationship, and to provide for a tool to assess confidence intervals. A statistical regression model itself does not have any biological meaning, and the choice of the model (expression) is to some extent arbitrary. It is the data, not the model, that should determine the dose-response relationship, and thereby the ECx (Fig. 6.12). When different models (with similar goodness of fit and equal number of parameters) result in different ECx estimates, the data are apparently not suitable for dose-response modelling.

364. Dose-response models that are based on the mechanism of action of the particular chemical are, as opposed to statistical models, supposed to contain information by themselves, and therefore be less sensitive to data gaps (between dose groups). However, they do contain unknown parameters that need to be estimated from the data, and it appears sensible to follow the guidelines described here in such models just as well. Mechanistic dose-response models are extremely rare, and contain some general elements at best. In the biological models discussed in chapter 7, the biological mechanisms in the models relate to the normal physiology in organisms rather than to the mechanism of action of specific chemicals.

6.5. Design issues

365. Concentration-response modelling can only be applied if the data contain sufficient information on the shape of the concentration-response relationship. Although this condition should be judged in each individual situation, experience teaches that at least four different response levels are needed (including the control group) in the case of continuous data. A similar condition holds for quantal data, e.g. two partial kills next to (almost) complete mortality and (almost) complete survival. When one actually ‘knows’ in advance that the concentration-response relationship is linear, designs with fewer concentration groups may be considered, and, as a matter of fact, they will be more efficient in terms of precision. However, it seems rare that one can be confident a priori that the concentration-response is indeed linear (usually not much is known in advance about the tested chemical’s action on the test organism) and extra concentration groups are highly recommendable.

366. A design with three concentration groups and a control may result in concentration-response data that allow for concentration-response modelling. However, it is always advisable to include more concentration groups for various reasons. If just one of the concentration groups was inadequately chosen (e.g. no observable response), concentration-response modelling will fail. Further, systematic differences between treatment (concentration) groups are not unusual in toxicity testing (e.g. caused by systematic order in handling the groups), which may result in biased estimation of the concentration-response relationship. This unfavourable effect can be diminished in designs with more concentration groups.

367. In general, it may be stated that for the purpose of estimating an ECx, it is important to have a sufficient number of dose groups, to prevent biased estimates of the ECx. The allocation of the organisms (or experimental units) to more dose groups may be done at the expense of the number of replicates in each group without much loss (if any) for the precision of the estimated ECx.

Location of dose groups

368. Concretely, for the purpose of estimating an ECx, the available number of organisms (replicates) should be allocated to at least three (excluding the controls), but preferably more concentration groups. Next to a sufficient number of concentration groups (resulting in different response levels), one needs to choose a lowest and highest concentration level.

369. For quantal data, one may aim at four concentrations showing different response levels, including (nearly) none and (nearly) complete response together with two concentrations with partial responses, as a minimum requirement. In continuous data, the low concentration is preferably chosen such that the observed response differs from the controls to a similar degree as x in the required ECx (to prevent that the ECx can only be estimated by extrapolation). Although one is usually interested in low response levels, high response levels are needed to assess the concentration-response relationship. The highest concentration would be preferably chosen such that the range between highest and lowest observed response is large enough to potentially include at least four different (in a rough statistical sense, that is, they appear detectable from the noise) response levels.

370. Interestingly, simulation studies show that the intuitive idea of concentrating dose levels around the ECx is not optimal. Designs that include sufficiently high dose levels (or rather sufficiently different response levels compared to the controls) perform better (Slob, in prep.).

Number of replicates

371. In typical quantal data (with both 0% and 100% observed response levels) the precision of the ECx declines with *x*, and the size of the experiment (total number of organisms or units) should be larger for smaller values of *x* that are considered appropriate. Thus, when only an EC50 is required, a smaller experiment is required than when an ECx is aimed for. In continuous dose-response this phenomena appears to be less prominent.

372. Theoretically, in quantal dose-response analysis the relationship between the precision of an ECx and the size of the experiment can be calculated. However, the number of organisms needed to obtain any given precision depends on the slope of the dose-response function itself, which is typically unknown before the study.

373. For the generally applicable nested family of (five) models, given in section 6.3.1, simulation studies are being performed (for continuous data), to provide an indication of the (total) number of replicates necessary to achieve a particular precision for the ECx (Slob, in prep).

Balanced vs. unbalanced designs

374. Due to the principle of leverage, observations in the extreme dose groups have more influence on the resulting model fit than the middle dose groups. This suggests that designs with larger sample sizes in the extreme dose groups may be more efficient than designs with the same sample sizes in all dose groups. Yet, preliminary simulation studies indicated that a design with twice the sample size in the controls performed only slightly better than one with equally sized dose groups. But more simulation studies are needed to give more definite answers to this question.

375. For designs with replicated experimental units (e.g. containers), where the number of replicates is small, say two, it appears wise to allocate a higher number of replicates in the controls, since a single erroneous replicate in the controls may then have a large impact on the model fit.

6.6. Exposure duration and time

376. It may be expected *a priori* that the response in biological systems is not only a function of dose, but also of the duration of the exposure. Therefore a model that describes the response as a function of both dose and duration would be more informative and give a more complete picture. Exposure duration is however a more complicated factor than dose, because it interferes with the factor time. The factor time has an impact by itself, e.g. on ageing, adaptation and repair of the processes underlying the response. Depending on the question to be answered, the study may, e.g.:

- monitor the organisms during a period of time after an acute, or a fixed period of exposure,
- monitor the organisms while they are held at various, but constant exposure levels,
- treat different groups of individuals with different exposure durations and compare the response at the end of exposure, or at a fixed point in time.

377. The second type of study is quite common in ecotoxicity testing in general (the others may be relevant for specific situations). Usually, in these studies the same (individual or groups of) organisms are followed over time. For example, the same organisms are recorded to have died or not. Or, egg production

is monitored for the same (group of) organisms over time. In other studies, however, the observations in time may relate to different experimental units. As a result, the observations may be or not be independent, and this should be taken into account in the analysis of the data. This section will briefly discuss the analysis of this type of data for both quantal and continuous data.

Quantal data

378. When a quantal response is observed at various points in time (e.g. number of additional deaths recorded each day while maintaining exposure at the same level), the statistical analysis of the dose-response data may be extended to include this extra information. Some authors have suggested fitting a dose-response model to the separate data sets, i.e. for each exposure duration separately, and plotting the ensuing EC50s as a function of time. The value to which this function levels off is called the incipient EC50, interpreted as the EC50 for "infinite" exposure. This is not a proper method and should be avoided, for various reasons. First, conceptual problems arise, e.g. an incipient LC50 does not make sense as more than 50% of the animals die without any exposure at longer exposure durations. Second, statistical problems arise, e.g. the dose-response data at different time points are not independent, which hampers the establishment of confidence intervals for the incipient EC50. And third, comparing dose-response models (such as the log-logistic) that are fitted for several time points separately may lead to inconsistent results (e.g. the fitted dose-response functions for various exposure durations intersect each other).

379. The approach of fitting a response surface to dose (concentration) and time simultaneously (multiple regression) is also improper, since the observations in time are not independent.

380. A proper way of modelling dose-time-response data where each individual is followed in time, is by assuming a relationship of dose with the hazard. The hazard[13] reflects the probability of an individual to respond (e.g. die in the case of mortality) in a very small time interval, divided by the probability that it is still alive at that age. On a population level, this reflects the incidence of response during that small time interval, divided by the fraction of the population still alive at that age. By assuming that the hazard is a function of dose, the dose-time-response data can be described in a single model. The hazard can be directly transformed into a survival (or mortality) function, or, more generally, in a quantal time-response function. This function may be used for deriving the log-likelihood given the observed frequencies of response, in the usual way. There is a vast literature on survival analysis (see, e.g. Cox and Oakes 1984; Miller 1981; Tableman and Kim 2004). For an example of dose-response modeling based on the hazard function, see section 7.2.

Continuous data

381. For many continuous endpoints observations can be (and sometimes are) made in time. For example, body weights of animals can be determined at particular time intervals during the study. Or, the growth of algae can be monitored over time. As another example, the number of eggs produced can be counted at specific time intervals. It is re-emphasised that the observations in time may relate to the same or to different units (organisms), determining if the data should be treated as dependent or independent observations.

[13] The hazard may be formally defined as $-\frac{dS(t)/dt}{S(t)}$, where S(t) denotes the survival function.

Independent observations in time

382. In some studies the observations in time relate to different units. For example, in algal growth studies, the biomass at a concentration is followed in time (e.g. day 0, 1 and 2). Suppose that once any of the algal test vessels has been measured it is removed from the test. In that case each observation relates to another vessel, and the data can be treated as independent, i.e. they can be taken together in a single analysis. As an illustration consider the data in fig 6.14, where at 9 different concentrations the biomass was measured at three consecutive days (each time with two replicates). Here a time-response model (i.e., a dose-response model with dose replaced by time) was fitted to all the data simultaneously, by assuming that the biomass at time zero was equal among the concentrations, while the growth rate differed between the concentrations. Thus, for each concentration a slope parameter b was estimated, but only a single parameter *a* and a single variance parameter. Thus, 11 parameters in total were estimated in a simultaneous fit of the model to these data.

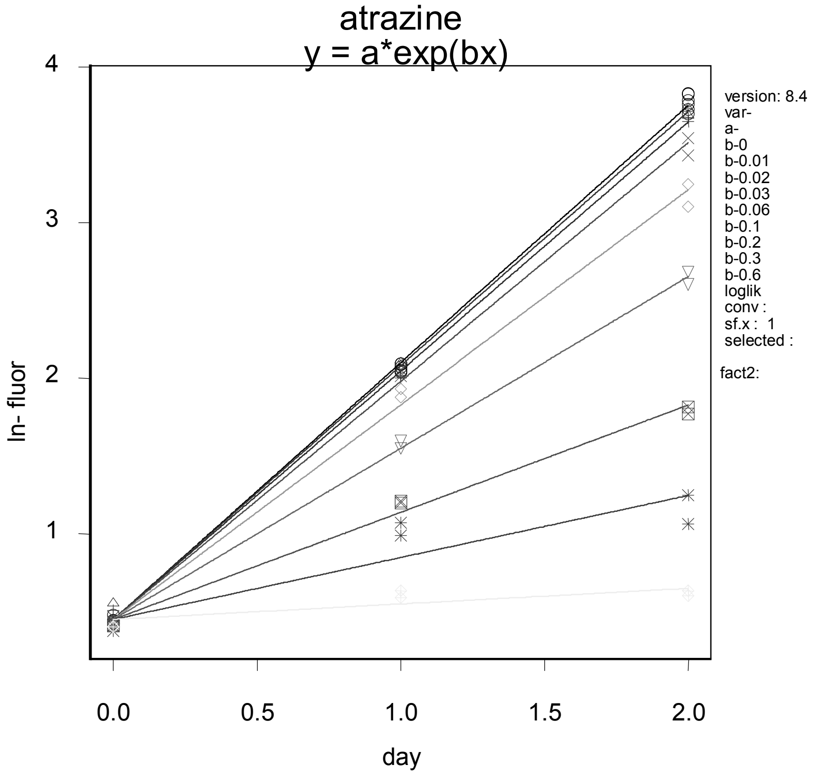


Figure.6.14 Observed biomasses (marks) as a function of time, for nine different concentrations of atrazine.

Here, an exponential growth model was fitted, thereby estimating a single background value (*a*), a separate growth rate (*b*) for each concentration, and a single residual variance (for log-biomass). Note that replicates are treated as independent observations in this analysis.

383. The estimated growth rates can subsequently be subjected to a dose-response analysis, as shown in Fig. 6.15.

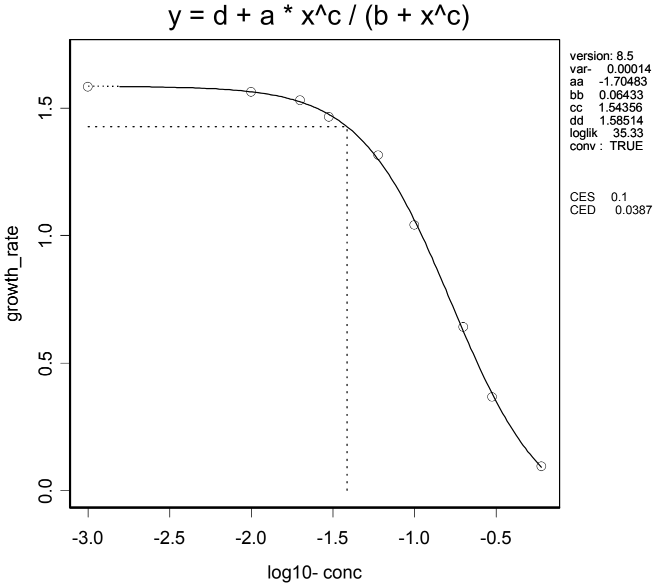


Figure 6.15 Growth rates as derived from biomasses observed in time (see fig. 6.14) at nine different concentrations (including zero), with the Hill model fitted to them.

Point estimate of EC10 (=CED): 0.0387 mg/l, with 90%-confidence interval: (0.0351, 0.0424), based on 1000 bootstrap runs.

Dependent observations in time

384. When the data in time relate to the same experimental units, the observations cannot be treated as independent data, and an analysis as in fig. 6.14 is improper. When the data show a clear trend in time, a straightforward approach is to fit the exponential growth model to the biomasses, but now allowing each experimental unit (flask) to have its own growth rate. This amounts to fitting a separate time-response model for each separate experimental unit, and subsequently subject the relevant[14] parameter estimates to a dose-response analysis. This analysis is analogous to that illustrated in fig. 6.14 and 6.15, but the concentration response data now have replicates (see fig. 6.16).

[14] The relevant parameter should follow from understanding the biological process. In algal biomass the obvious parameter is the growth rate; when the observations relate to number of eggs, supposed to level off at a constant level with age, the parameter reflecting that level or the parameter reflecting the rate at which that level is reached could both be the relevant parameter.

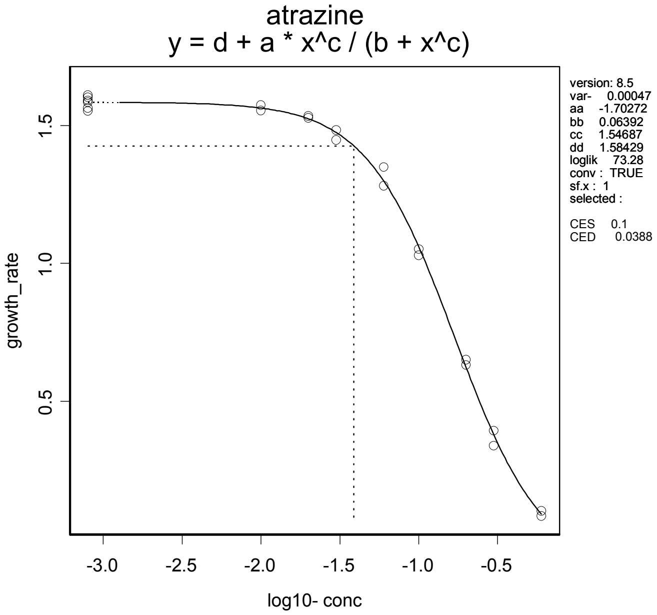


Figure. 6.16 estimated growth rates as a function of (log-)concentration atrazine.

Here, the individual flasks were taken into account, resulting in two growth rate estimates for each (nonzero) concentration, and six growth rates for concentration zero. Point estimate of ec10 (=ced): 0.0388 mg/l, with 90%-confidence interval : (0.0355, 0.0421), based on 5000 bootstrap runs.

385. It may be noted that the confidence intervals for the EC10 as derived from the data in Fig. 6.15 and Fig. 6.16 are very similar, despite the fact that in the latter case there are more data points. The reason is that the information in both data sets is in fact the same.

386. In data sets where no trend in time is apparent, one may just as well take the average over time (in each unit) and apply the dose-response analysis to the averages.

6.7. Search algorithms and nonlinear regression

387. As discussed previously, nonlinear regression models can only be fitted in an iterative "trial and error" approach. Computer software use efficient algorithms to do that, and the user does not need to worry about the exact nature of the calculations. However, some basic understanding of the search process in required in order to interpret the results. In addition, such understanding is needed to evaluate whether or not the algorithm was successful or not, and if not, what if anything can be done about that.

388. An iterative algorithm tries to find "better" parameter values in a process by evaluating if the fit can be improved by changing the parameter values. By regarding the fit criteria as a function of the parameters, the problem is in fact to find the maximum (in the case of likelihood) or minimum (in the case of Sum of Squares) of that function. Although algorithms have been developed to do this in an efficient way, one should keep in mind that the algorithm cannot see in advance where the optimum of the function is. One may compare the algorithm with a blindfolded person, who can only feel if there is a slope or not (and how steep it is). The algorithm recognises the optimum by the property that around the optimum the slope changes from increasing to decreasing (or vice versa).

389. Obviously, the algorithm can only start searching when the parameters have values to start with. Although the software often gives a reasonable first guess for the starting values, the user may have to change these. It is not unusual (in particular when the information in the data is hardly sufficient to estimate the intended parameters) that the end result depends on the starting values chosen, and the user should be aware of that.

390. The algorithm keeps on varying the parameter values until it decides to stop. There are two possible reasons for the algorithm to stop the searching process:

- The algorithm has *converged*, i.e. it has found a clear maximum in the log-likelihood function. In this case the associated parameter values can be considered as the "best" estimates (MLEs if the likelihood was maximised). However, it can happen that the log-likelihood function has not one but more (local) maxima. This means that one may get other results when running the algorithm again, but with other start values. This can be understood by remembering that the algorithm can only "feel" the slope locally, so that it usually finds the optimum that is closest to the starting point.
- The algorithm has not converged, i.e. the algorithm was not able to find a clear optimum in the likelihood function, but it stops because the maximum number of iterations (trials) is exceeded. This may occur when the starting values were poorly chosen, such that the associated model would be too far away from the data. Another reason could be that the information in the data is poor relative to the number of parameters to be estimated. For example, a concentration-response model with five unknown parameters cannot be estimated with a four-concentration-group study. As another example, the variation between the observations within concentration groups may be large compared to the overall change in the concentration-response. In these cases the likelihood function may be very flat, and the algorithm cannot find a point where the function changes between increasing and decreasing. The user may recognise such situations by high correlations between parameter estimates, i.e. changing the value of one parameter may be compensated by another, leaving the model prediction practically unchanged.

6.8. Reporting Statistics

391. Reporting statistics are as follows:

Quantal data

- Test endpoint assessed
- Number of Test Groups
- Number of subgroups within each group (if applicable)
- Identification of the experimental unit
- Nominal and measured concentrations (if available) for each test group
- Number exposed in each treatment group (or subgroup if appropriate)
- Number affected in each treatment group (or subgroup if appropriate)
- Proportion affected in each treatment group (or subgroup if appropriate)
- The dose metric used
- The model function chosen for deriving the EC50 (ECx)
- Plot of dose-response data with fitted model, including the point estimates of the model parameters and the log-likelihood (or residual SS)

- Fit criteria for other fitted models
- The EC50 together with its 90%-confidence interval.
- If required: the ECx together with its 90%-confidence interval.
- Method used for deriving confidence intervals

Continuous data

- Test endpoint assessed
- Number of Test Groups
- Number of subgroups within each group (if applicable)
- Identification of the experimental unit
- Nominal and measured concentrations (if available) for each test group
- The dose metric used.
- Number exposed in each treatment group (or subgroup if appropriate)
- Arithmetic group means and standard deviations, but geometric group means and standard deviation if lognormality was assumed
- The model function chosen for deriving the ECx
- Plot of dose-response data with fitted model, including the point estimates of the model parameters and the log-likelihood (or residual SS)
- Fit criteria for other fitted models
- The ECx (CED) together with its 90%-confidence interval
- Method used for deriving confidence intervals

7. BIOLOGY-BASED METHODS

7.1. Introduction

7.1.1. Effects as functions of concentration and exposure time

392. Biology-based methods not only aim to describe observed effects, but also to understand them in terms of underlying processes such as toxicokinetics, mortality, feeding, growth and reproduction (Kooijman 1997). This focus on dynamic aspects allows exposure time to be treated explicitly.

393. This chapter focuses on the analysis of data from a number of standardized toxicity tests on mortality, body growth (e.g. fish), reproduction (e.g. daphnia), steady-state population growth (of e.g. algae, duckweed). The guidelines for these tests prescribe that background mortality is small, while the duration of the test is short relative to the life-span of the test-organisms. Moreover the tests are done under conditions that are otherwise optimal, which excludes multiple stressors (e.g. effects of food restriction,

temperature (Heugens, 2001, 2003)), and quite a few processes that are active under field conditions (e.g. adaptation, population dynamics, species interactions, life-cycle phenomena (Sibly and Calow (1989)). The type of data that are routinely collected in these tests are very much limited, and do not include internal concentrations of test compounds. These restrictions exclude the application of quite a few potentially useful methods and models for data analysis, such as more advanced pharmacokinetic models and time series analysis, see e.g. Newman (1995). The theory behind biology-based methods can deal with dynamic environments (changing concentrations of test compounds, changing food densities), but the application in the analysis of results from toxicity tests is simplified by the assumption that organisms' local environment in the test is constant.

394. Biology-based methods make use of prior knowledge about the chemistry and biology behind the observed effects. This knowledge is used to specify a response *surface*, i.e. the effects as a function of the (constant) concentration of test compound in the medium *and* the exposure time to the test compound. This response surface is determined by a number of parameters. The first step is to estimate these parameters from data. The second step is to use these parameter values to calculate quantities of interest, such as the ECx-time curve, or the confidence interval of the No-Effect-Concentration (NEC). It is also possible to use these parameter values to predict effects at longer exposure times, or effects when the concentration in the medium is not constant. If the observed effects include those on survival and reproduction of individuals, these parameters can also be used to predict effects on growing populations (in the field) (Kooijman 1985, 1988, 1997, Hallam et al 1989).

395. It is essential to realise that ECx values decrease for increasing exposure time, as long as the exposure concentration and the organism's sensitivity remain constant. This is partly due to the fact that effects depend on internal concentrations (Kooijman 1981, Gerritsen 1997, Péry *et al* 2002), and that it takes time for the compound to penetrate the body of test organisms. (The standard is to start with organisms that were not previously exposed to the compound.) The exposure period during which the decrease is substantial depends on the properties of the test compound and of the organism and the type of effect. For test compounds with large octanol-water partition coefficients and test organisms with large body sizes this period is usually large. The LC50 for daphnids hardly decreases for a surfactant after two days, for instance, but their LC50 for cadmium still decreases substantially after three weeks. For this reason, biology-based methods fit a response *surface* to data, using all observation times simultaneously. If just a single observation time is available, however, these methods can still be used and the response surface reduces to a response curve. Obviously, such data hardly contain information about the dynamic aspect of the occurrence of effects. The parameter(s) that quantify this aspect are then likely to be poorly defined. This does not need to be problematic for all applications (such as the interpolation of responses for other concentrations at that particular observation time; this is the job of dose-response methods). It is strongly recommended, however, for a two-day test on survival, for instance, to use not only the counts at the end of the experiment, but also those at one day. Such data are usually available (and GLP even requires the reporting of those data), but these data are not always used. More recommendations are given in section 7.3.

396. In practice it is not unusual that very few, if any, concentrations exist with partial effects; survival of a cohort of individuals tends to be of the "all or nothing" type in most concentrations. High concentrations run out of surviving individuals more rapidly than lower concentrations. This can occur in ways such that for each single observation time, no, or very few, concentrations show partial mortality. This situation also occurs if each individual is exposed separately, and measured rather than nominal concentrations are used in the data analysis; one then has just a single individual per concentration because no two concentrations are exactly equal. Although such a case is generally problematic for dose-response methods, because a free slope parameter has to be estimated (Kooijman 1983), biology-based methods do not suffer from this problem, because the (maximum) slope is not a free parameter (models' slope of

concentration-survival curves increases during exposure), and the information of the complete response surface is used. An example will be given in section 7.3

397. Biology-based methods allow the use of several data sets simultaneously, such as survival data, sublethal effect data, and data on the concentration of test compound inside the bodies of the test organisms during accumulation/elimination experiments. As will be discussed below, logical relationships exist between those data, and these relationships can be used to acquire information about the value of particular parameters that occur in all these data sets. Both the statistical procedures and the computations can become somewhat more complex in this type of advanced applications, but free and downloadable software exist that can do all computations with minimum effort (see below).

7.1.2. Parameter estimation

398. The maximum likelihood (ML) method is used to estimate parameter values (the criterion of least squared deviations between data and model predictions is a special case of the ML method, where the scatter is independently normally distributed with a constant variance). If more than one data set is used (for instance, data on body size and reproduction rate and/or internal concentration), the assumption is that the stochastic deviations from the mean are independent for the different data sets. This allows the formulation of a composite likelihood function that contains all parameters for all models that are used to describe the available data sets. For effects on survival, the number of dead individuals between subsequent observation times follows a multinomial distribution (see e.g. Morgan 1992); for sublethal effects, the deviations from the mean are assumed to be independently normally distributed with a common (data-set-specific) variance. The deterministic part of the model prediction is fully specified by the theory,. F~~f~~or the stochastic part, only these straightforward assumptions are programmed in the DEBtox software (see Section 7.9.). The software package DEBtool, allows more flexibility in the stochastic model, e.g. for ML estimates in the case that the variance is proportional to the squared mean; this rarely results in substantially different estimates, however.)

399. If surviving individuals are counted in a toxicity test and tissue-concentrations are measured in another test, a composite likelihood function can be constructed that combines these multinomial and normal distributions. The elimination rate (dimension: per time) is a parameter that occurs in both types of data. In survival data it quantifies how long it takes for death to show up; if the elimination rate is high, one only has to wait a short time to see the ultimate effects. The elimination rate can, therefore, be extracted from survival data in absence of data on internal concentrations. Although it is helpful to have the concentration-in-tissue data (both for estimating the parameters and for testing model assumptions), these data are by no means required to analyse effects on survival. If one has prior knowledge about the value of the elimination rate, one can fix this parameter and estimate the other parameters (such as the NEC) from survival data.

400. Profile likelihood functions are used to obtain confidence intervals for parameters of special interest, and in particular for the NEC. This way of quantification of the uncertainty in a parameter value does not necessarily lead to a single compact interval, but sometimes leads to two, non-overlapping intervals. Therefore, they can better be indicated with the term “confidence set”. Computer simulation studies have shown that these confidence sets are valid for extremely low numbers of concentrations and of test organisms (Andersen et al, 2000).

401. Estimation procedures have been worked out (Kooijman 1983) to handle somewhat more complex experimental designs, in which living individuals are sacrificed for tissue analysis during the test. The information that they were still living at the moment of sampling is taken into account in the estimation of parameter values that quantify the toxicity of the compound. Péry *et al* (2001) discuss the estimation of

parameters in the case that the concentration in the media varies in time using hazard models; Kooijman (1981) and Reinert et al (2002) use critical body residue models.

7.1.3. Outlook

402. This document only discusses the simplest experimental designs of toxicity tests and the simplest models. The authors of this document are unaware of alternative~~s~~ models in the open literature that are applicable on a routine basis and hope that this document will stimulate research in this direction. The models can be and have been extended in many different ways; just one example is given. All individuals are assumed to have identical parameter values in the models that are discussed below. Individuals can differ, despite the standardisation efforts in tests. Such difference might relate to differences in one or more parameter values (Sprague 1995). It is mathematically not difficult to include such differences in the analysis, on the basis of assumptions about the simultaneous scatter distribution of the parameter values. Needless to say, one really does know little if anything about this distribution. This makes such assumptions inspired by convenience arguments rather than by mechanistic insight. A strong argument for refraining from such extensions is that the method becomes highly unpractical. The data simply do not allow a substantial increase in the number of parameters that must be estimated from routine data.

403. The theory covers many features, such as extrapolating from constant to pulse exposures and vice versa, and including the effects of senescence, that are not yet worked out in software support (see Section 7.9).

7.2.3. The modules of effect-models

404. Effects are described on the basis of a sequence of three steps (modules):

1. **Change in the internal concentration**: the step from a concentration in the local environment (here the medium that is used in the test) to the concentration in the test organism.
2. **Change in a physiological target parameter**: the step from a concentration in the test organism to a change in a target parameter, such as the hazard rate, the (maximum) assimilation rate, the specific maintenance rate, the energy costs per offspring, etc.
3. **Change in an endpoint**: the step from a change in a target parameter to a change in an endpoint, such as the reproduction rate, the total number of offspring during an exposure period, etc.

405. This decomposition of the description of effects into three modules calls for an eco-physiological model of the test organism that reveals all possible physiological targets. The primary interest is in small effects. A simplifying assumption is that just a single physiological process is affected at low concentrations and that this effect can be described by a single parameter. At higher concentrations, more processes might be affected simultaneously. This means that the number of possible effects (and so the number of required parameters) can rapidly increase for large effects. It is unpractical and, for our purpose not necessary, to try to describe large effects in detail.

406. The concept “most sensitive physiological process” has an intimate link with the concept “no-effect-concentration”. The general idea is that each physiological process has its own “no-effect-concentration”, and that these concentrations can be ordered. Below the lowest no-effect-concentration, the compound has no effect on the organism as a whole. Between the lowest and the second lowest no-effect concentrations, a single physiological process is affected; between the second and the third lowest no-effect concentrations, two processes are affected, etc.

407. The concept “**no-effect-concentration**” is quite natural in eco-physiology (see e.g. Chen & Selleck 1969). All methods for the analysis of toxicity data (including hypothesis testing and dose-response

methods) make use of the *concept* "no-effect-concentration". All methods assume, at least implicitly, that compounds in the medium, apart from the tested chemical, do not affect the organism's response. Hypothesis testing explicitly assumes that the tested chemical has no effect on the response at concentrations equal to, and lower than, the NOEC. Biology-based methods use the NEC as a free *parameter*.

408. Generally each compound has three domains in concentration:

1. Effects due to **shortage**. Think, for instance, of elemental copper, which is required in trace amounts for several co-enzymes of most species
2. **No-effect** range. The physiological performance of the organism seems to be independent of the concentration, provided that it remains in the no-effect range. Think, for instance, of the concentration of nitrate in phosphate limited algal populations; Liebig's famous minimum law rests on the "no-effect" concept (von Liebig 1840)
3. **Toxic** effects. Think, for instance, of glucose, which is a nutritious substrate for most bacteria in low concentrations, but inhibits growth if the concentration is as high as in jam.

409. It is essential to realise that the judgement "no-effect" is specific for the level of organisation under consideration. At the molecular level, molecules cannot be classified into one type that does not give effects, and another type that gives effects. The response of the individual as a whole is involved (Elsasser 1998). The concept "no-effect-concentration" can deal with the situation that it is possible to remove a kidney, for instance, from a human subject (so a clear effect at the sub-organism-level), without any obvious adverse effects at the level of the individual (during the limited time of a test). This example, therefore, shows that below the NEC effects can occur at the suborganismic level (e.g. enzyme induction), as well as on other endpoints that are not included in the analysis (e.g. changes in behaviour).

410. Most compounds are not required for the organisms' physiology, which means that their range of concentrations that cause effects due to shortage is zero, and the *lower* bound of the no-effect range is, therefore, zero as well. Some compounds, and especially the genotoxic ones (van der Hoeven et al 1990, de Raat et al 1985, 1987, Purchase & Auton 1995), are likely to have a no-effect range of zero as well, and the *upper* bound of the no-effect range is, therefore, also zero. This gives no theoretical problems in biology-based methods. A NEC of zero is just a special case, and a point estimate for this concentration from effect-data should (ideally) not deviate significantly from zero (apart from the Type I error ; a Type I error occurs if the null hypothesis is rejected, while it is true).

411. The model for each of the three modules for the description of effects is kept as simple as possible for practical reasons, where one usually has very little, if any, information about internal concentrations, or physiological responses of the test organisms. Each of these modules can be replaced by more realistic (and more complex) modules if adequate information is available. Some applications allow further simplification. Algal cells, for instance, are so small that the intracellular concentration can be safely assumed to be in instantaneous equilibrium with the concentration in the media that are used in the test for growth inhibition. This gives a constant ratio between the internal and external concentrations, and simplifies the model considerably. The standard modules are introduced below.

7.2.1. Toxico-kinetics model

412. The toxico-kinetic module is taken to be a first order kinetics by default; the accumulation flux is proportional to the concentration in the local environment, and the elimination flux is proportional to the concentration inside the organism. This simple two-parameter model is rarely accurate in detail, but frequently captures the main features of toxico-kinetics (Harding & Vass 1979, Kimerle et al 1981, McLeese et al 1979, Spacie & Hamelink 1979, Wong et al 1981, Janssen et al 1991, Legierse et al 1998,

Jager, 2003, Jaget et al 2003). It can be replaced by a more-compartment model, or a pharmacokinetic model, if there are sound reasons for this. Metabolic transformation, and satiation in the elimination rate can modify toxico-kinetics in ways that are sometimes simple to model (Kooijman 2000).

413. If the organism grows during exposure, or changes in lipid content occur (for instance when the test organisms are starved during exposure), predictable deviations from first order kinetics can be expected, and taken into account (Kooijman & van Haren 1990, Kooijman 2000). Dilution by growth should always be taken into account in the test for body growth and reproduction, since such a dilution affects the effect-time profiles substantially.

7.2.2. Physiological targets of toxicants

414. The specification of sublethal effects involves an eco-physiological model that reveals all potential target parameters, and allows the evaluation of the endpoints of interest. A popular endpoint is, for instance, the cumulative number of offspring of female daphnids in a three-week period. The model should specify such a number, as well as the various physiological routes that lead to a change of this number. It should also be not too complex for practical application. An example of such a model is the Dynamic Energy Budget (DEB) model. Because it is the only model for which generic applications in the analysis of toxicity data has been worked out presently, the following discussion will focus on this model.

415. The DEB model results from a theory that is described conceptually in Kooijman (2001) and Nisbet et al (2000), and discussed in detail in Kooijman (2000). Figure 7.1 gives a scheme of fluxes of material through an animal, which are specified mathematically in the DEB model, on the basis of mechanistic assumptions. The model's main features are indicated in the legend of Figure 7.1. The DEB theory is not confined to animals, however, and covers all forms of life.

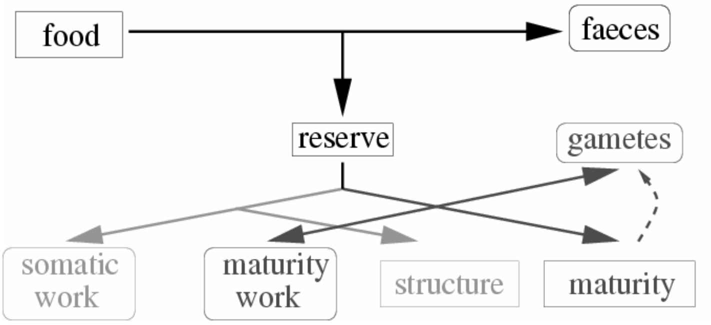


Figure 7.1 Fluxes of material and energy through an animal, as specified in the DEB model.

Assimilation, i.e. the conversion of food into reserve (plus faeces) is proportional to structure's surface area. Somatic and maturity work (involved in maintenance) are linked to structure's mass, but some components (heating in birds and mammals, osmo-regulation in freshwater organisms) are linked to structure's surface area. Allocation to structure is known as growth; to maturity as development; to gametes as reproduction. Embryos do not feed, juveniles do not reproduce, adults do not develop. Reserves and structure are both conceived as mixtures of mainly proteins, carbohydrates and lipids; they can differ in composition. The rate of use of reserve depends on the amount of reserve and structure; this rate is known as the catabolic rate. A fixed fraction of the catabolic flux is allocated to somatic maintenance plus growth, as opposed to maturity maintenance plus development (or reproduction).

416. The general philosophy behind the DEB theory is a full balance approach for food (nutrients, energy, etc): "what goes in must come out". Offspring is (indirectly) produced from food, which relates reproduction to feeding. Large individuals eat more than small ones, which links feeding to growth. Maintenance represents a drain of resources that is not linked to net synthesis of tissue or to reproduction.

An increase of maintenance, therefore, indirectly leads to a reduction of growth, so to a reduction of feeding and reproduction.

417. This reasoning shows that the model requires a minimum level of complexity to address the various modes of action of a compound. One needs to identify this route to translate effects on individuals to that on the growth of natural populations (in the field). If food conditions are good, investment in maintenance, for instance, comprises only a small fraction of the daily food budget of individuals. Small effects of a toxicant on maintenance, therefore, result in very small effects on the population growth rate. If food conditions are poor, however, maintenance comprises a large fraction of the daily food budget. Small effects on maintenance can now translate into substantial effects on the population size. This reasoning shows that effects on populations depend on food conditions, which generally vary in time (Kooijman 1985, 1988, Hallam et al 1989). The different modes of action usually result in very similar point estimates for the NEC, within the current experience. Furthermore, no effects on individuals implies no effects on populations of individuals, but the mode of action is particularly important for predicting the effects at the population level.

7.2.3. Change in target parameter

418. The value of the target parameter is assumed to be linear in the internal concentration. The argumentation for this very simple relationship is in the Taylor's Theorem- which states that any regular function can be approximated with any degree of accuracy for a limited domain by a polynomial of sufficiently large order. The interest is usually in small effects only, and routine applicability urges for maximum simplicity, so a first order polynomial (i.e. a linear relationship) is a strategic choice.

419. The biological mechanism of a linear relationship between the parameter value and internal concentration boils down to the independent action at the molecular level. Each molecule that exceeds an individual's capacity to repress effects acts independent of the other molecules. Think of the analogy where photosynthesis of a tree is just proportional to the number of leaves as long as this number is small; as soon as the number grows large, self-shading occurs and photosynthesis is likely to be less than predicted.

420. We doubtlessly require non-linear responses for larger effect levels, but then also need to include more types of effects. Interesting extensions include receptor-mediated effects. The biochemistry of receptors is rather complex. Two popular models are frequently used to model receptor-mediated effects and concentration: the Michaelis Menten model boils down to a hyperbolic relationship, rather than a linear one (which has one parameter more, Muller & Nisbet (1997)); the Hill model boils down to a log-logistic relationship (and has two parameters more than the linear model, Hill (1910), Garric et al (1990), Vindimian et al (1983)). Such extensions are particularly interesting if toxicokinetics is fast, and the internal concentration is proportional to the external one (such as in cell cultures). The assumption that the target parameter is linear in the internal concentration does *not* translate into a linear response of the endpoint; it usually translates into sigmoid concentration-endpoint relationships, which are well known from empirical results. Notice that the linear model is a special case of the hyperbolic one, which is a special case of the log-logistic one.

7.2.4. Change in endpoint

421. The DEB model specifies how changes in one or more target parameters translate into changes in a specified endpoint. Popular choices for endpoints are reproduction rates (number of offspring per time), cumulative number of offspring (in daphnia-reproduction tests), body length (in fish-growth tests) and survival probability. Survival and reproduction together determine steady state population growth, if they are known for all ages. Reproduction rates depend on age, namely, and the first few offspring contribute

much more to population growth than later offspring. This is a consequence of the principle of interest-upon-interest; early offspring start reproduction earlier than later offspring. As will be discussed below, indirect effects on reproduction come with a delay of the onset of reproduction, while direct effects on reproduction do not. The DEB model takes care of these more complex, but important, aspects of reproduction. Given the DEB model, there is no need to study all ages of the test organism once the DEB parameters are known. This application requires some basic eco-physiological knowledge about the species of test organism, but the acquisition of this knowledge does not have to be repeated for each toxicity test.

7.3. Survival

422. The effects on the survival probability of individuals are specified via the hazard rate. A hazard rate (dimension: probability per time) is also known as the instantaneous death rate. The hazard rate *h*(*t*) relates to the survival probability *q(t)* as

$$h(t) = -q(t)^{-1} \tfrac{d}{dt} q(t) \quad \text{or} \quad q(t) = \exp\{-\int_0^t h(s)ds\}$$

423. The product h times dt has the interpretation of the probability of dying in a small time increment dt given that the organism is alive at time t. If the hazard rate is constant, which is the standard assumption for the death rate in the control, the relationship between the survival probability and the hazard rate reduces to *q*(*t*) = exp{-*ht*}. Generally, the hazard rate increases with time, however. The mortality process can be modelled via the hazard rate, as is standard in survival analysis (Miller, 1981; Cox & Oakes, 1984). The hazard rate can depend on ageing and toxicity, as implied by the present model for survival, and can decrease in time, if , for instance, the concentration of a toxic compound decreases in time. If the concentration is constant, the ultimate LC50 equals the NEC.

424. The following assumptions specify the survival probability at any concentration of test compound:

- Assumptions on control behaviour
  - The hazard rate in the control is constant
  - The organisms do not grow during exposure
- Assumption on toxico-kinetics
  - The test chemical follows first order kinetics
- Assumption on effects
  - The hazard rate is linear in the internal concentration
- Assumptions on measurements/toxicity test
  - The concentrations of test-compound are constant during exposure.
  - The measured numbers of dead individuals in subsequent time intervals are independently multinomially distributed

425. In summary the model amounts to: the hazard rate is linear in the internal concentration, which follows first order kinetics. These assumptions result in sigmoidal concentration-survival relationships, not unlike the log-logistic one, with a slope that increases during exposure (see Figure 7.2).

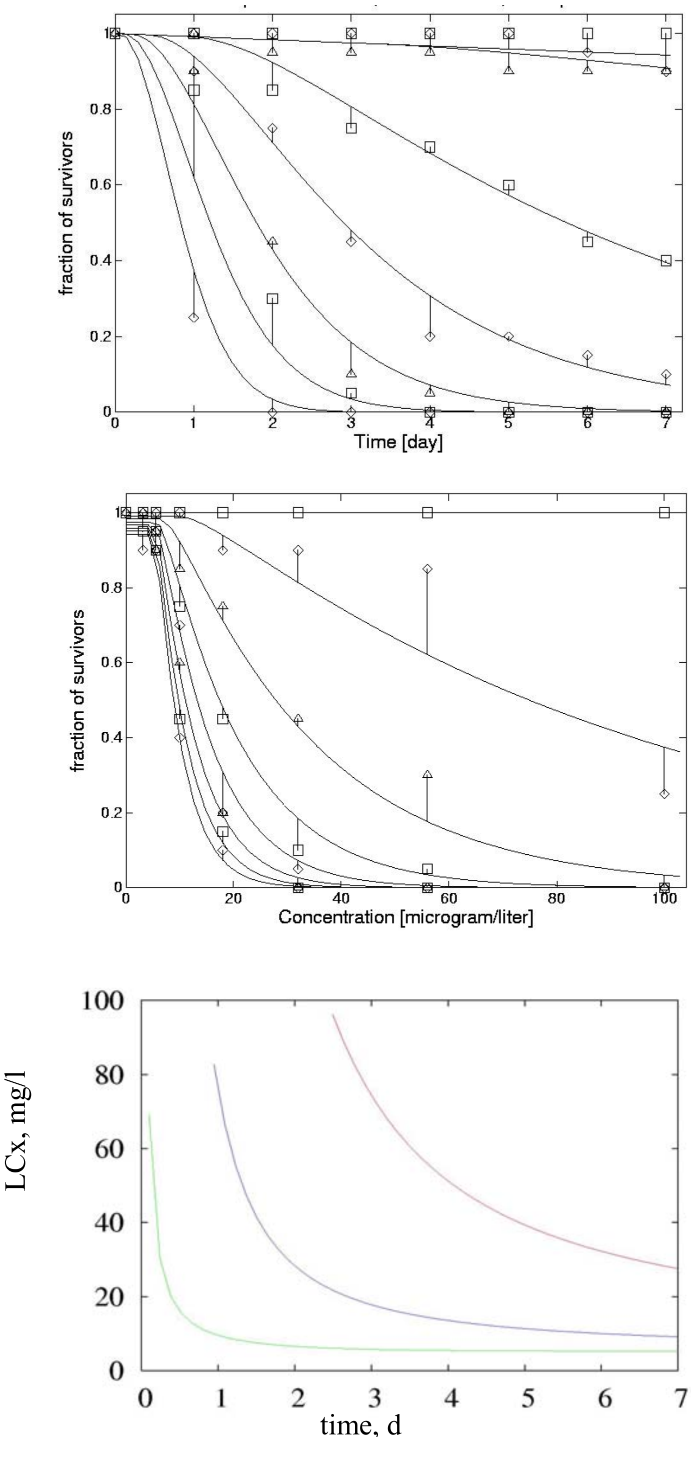


Figure 7.2 The time and concentration profiles of the hazard model, together with the data of Figure 7.7.

The resulting ML estimates are: control hazard rate = 0.0083 1/d, NEC = 5.2 µg/l, killing rate 0.037 (µg.d)$^{-1}$, elimination rate = 0.79 d^{-1}. From the last three parameters, LCx-time curves can be calculated, curves for the LC0, LC50 and LC99 are shown. (Calculated with DEBtox and DEBtool, see 7.9). For long exposure times, the LCx curves will tend towards the NEC, for all x, in absence of blank mortality.

426. As is shown, the three exposure- time-independent parameters of the hazard model completely determine the response surface, thus the LCx-time curves. It is even possible to reverse the reasoning. If

the LC50.1d = 50 mM, LC50.2d = 30 mM and LC50.3d = 25 mM, the NEC = 17.75 mM, the killing rate = 0.045 1/(mM.d), the elimination rate = 2.47 1/d. Such reconstructions are not very reliable, however, but they improve somewhat if more LC50 values are used.

427. If the observation times are very close together, the resulting huge matrix of survival-count data can be reduced to time-to-death data. Concentration-response modelling is traditionally considered to be different from time-to-death modelling, c.f. Newman et al (1989), Dixon & Newman (1991), Diamond et al (1991), but in the framework of biology-based models, these two approaches are just extreme cases of analyses of response-surfaces; their distinction vanishes and we generally deal with mixtures of both. The log likelihood function then reduces to

$$l = \sum_i \ln h(t_i) - \sum_j \int_0^{t_j} h(s)ds$$

where the first summation is across the individuals that actually died at the observed time points, excluding the ones that are taken alive out of the experiment. This can happen, for instance at the end of the experiment, or because their internal concentration is measured in a destructive way. The second one is across all individuals (the ones that died, as well as the ones that were removed alive). This sampling scheme allows that the concentrations for all individuals differ. An example of application is as follows:

Time-to-death and concentration pairs (in d and mM, respectively):

(21,1); (20,1.1); (20,0.9);(18,1.2); (16,1.3); (16,1.4); (15,1.5); (10,2); (9,1.8); (6,2.2); (5,2.5); (2,3); (2,4.3); (1,5); (1,4.5). Time-of-removal and concentration pairs: (21,0); (21,0); (21,0); (21,1). The ML estimates for this combined data set for 19 individuals in total are: control hazard rate = 0.061 d^{-1}, NEC = 1.93 mM, killing rate = 0.33 1/(mM.d), elimination rate 0.75 d^{-1}. This means, for instance, that the LC50.2d = 5.6 mM and the LC50.21d = 2.06 mM. (Calculations with DEBtool, see 7.9.2)

428. The link between the DEB theory and the survival model is in the ageing module of the DEB model, where the hazard rate, as affected by the ageing process, depends on the respiration rate in a particular way due to the action of free radicals; genotoxic compounds have a very similar mode of action and these compounds accelerate the ageing process (Kooijman, 2000). The processes of tumour induction and growth have direct links with the ageing process (van Leeuwen and Zonneveld, 2001). These effects on survival are beyond the scope of the present document, which deals with survival during (short) standardised exposure experiments.

429. On the assumption that test animals do not recover from immobilisation, the concept “death” can be replaced by “initiation of immobilisation” in this model. Due to the non-linearity that is inherent to toxico-kinetics, this model does not belong to the class of generalised linear models for survival, which has been proposed for the analysis of toxicity data (Newman 1995, McCullagh & Nelder 1989).

430. The model for effects on survival, and details about the statistical properties of parameter estimates (especially that of NECs) are discussed in Andersen et al (2000), Bedaux & Kooijman (1994), Klepper & Bedaux (1997, 1997a), Kooijman & Bedaux (1996, 1996a). Effects at time-varying concentrations are discussed in Péry et al (2001, 2002), Widianarko & van Straalen (1996).

7.4. Body growth

431. The DEB model allows for (at least) three routes for affecting body growth:

1. a decrease of the assimilation rate. Assimilation deals with the transformation from food into reserves, and can be affected by a decrease of the feeding rate, or a decrease of the digestion efficiency.

2. an increase of the somatic maintenance costs. These costs comprise protein turnover, the maintenance of intracellular and intra-organismal concentration gradients of compounds, osmo-regulation, heating of the body (mainly in birds and mammals), activity, and other drains on resources that are not linked to processes of net synthesis. Somatic maintenance costs directly compete with body growth for resources (in the DEB model). Thus an increase of maintenance costs directly results in a decrease of body growth, due to conservation of mass and energy.

3. an increase in the specific costs for growth. This is the case where the resource allocation to body growth is not affected, but the conversion of these resources to new tissue is.

432. This list does not exhaust all possibilities. An interesting alternative is in the change of the allocation to somatic maintenance plus body growth versus maturity maintenance and maturation (or reproduction). Under control conditions, the DEB model takes the relative investments in these two destinations to be constant (the absolute investments can change in time). Parasites and endocrine disrupting compounds (e.g. Andersen et al 2001, Kooijman, 2000) are found to change these relative investments. It is possible that a large number of compounds have similar effects. A practical problem in the application of a model that accounts for changes in the allocation fraction is that standardised tests for body growth do not include measurements that are necessary to quantify the effect appropriately. Detailed modelling of effects on mammalian development has been developed and applied (Setzer et al 2001, Lau et al 2000), but such approaches require adequate data and are specific for the compound as well as the test organism.

433. The following assumptions specify the effect on body growth at any concentration of test compound:

- Assumption on control behaviour
  - the test-organisms follow a von Bertalanffy growth curve in the control.
- Assumption on toxico-kinetics
  - the test chemical follows first order kinetics.
    (Dilution by growth is taken into account.)
- Assumption on effects
  One of three modes of action occur
  - the assimilation rate decreases linearly in the internal concentration.
  - the maintenance rate increases linearly in the internal concentration.
  - the costs for growth increases linearly in the internal concentration.
- Assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure.
  - the measured body lengths are independently normally distributed with a constant variance

434. The von Bertalanffy growth curve is given by $L(t) = L_\infty - (L_\infty - L_0)\exp\{-r_b t\}$, where $L(t)$ is the length at time t, L_0 is the initial length, L_∞ is the ultimate length, and r_b is the von Bertalanffy growth rate. The DEB model predicts that body growth is of the von Bertalanffy type only at constant food

densities, in the case of isomorphs (i.e., organisms that hardly change in shape during growth). An implied assumption is, therefore, that food density is constant, or high. Food intake depends hyperbolically on food density in the DEB model; variations in food density, therefore, hardly result in variations in food intake as long as food remains abundant. Examples of application of the model of effects on growth by an increase of the maintenance costs and by a decrease of assimilation are as follows:

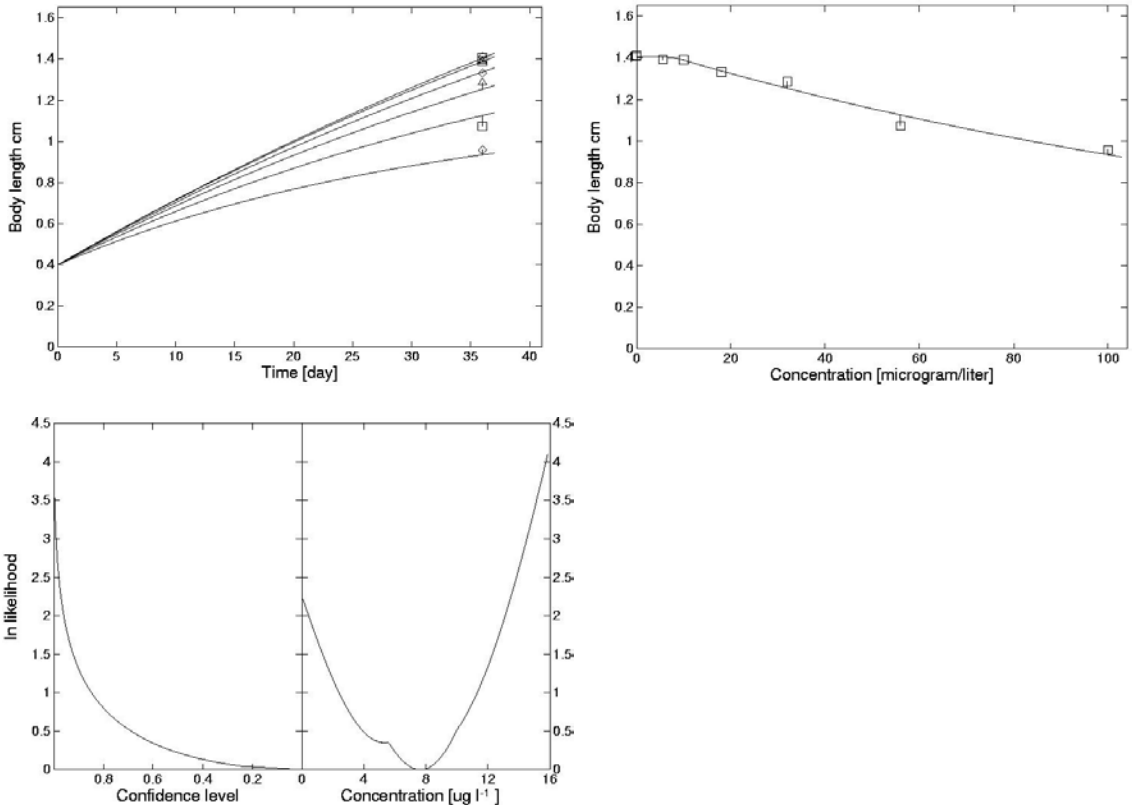


Figure 7.3 The time and concentration profiles for effects on growth of *Pimephalus promelas* via an increase of specific maintenance costs by sodium pentachlorophenate (data by Ria Hooftman, TNO-Delft).

The parameters estimates are: NEC = 7.65 g/l; control ultimate length = 37 mm; tolerance conc = 43.5 g/l; elimination rate = large; Fixed parameters are: initial length = 4 mm; von Bertalanffy growth rate = 0.01 d. The profile likelihood function for the NEC is given left. The EC0.36d = 766g/l; EC50.36d = 176 g/l. The use of the profile likelihood graphs to obtain confidence intervals is explained in the legend to Figure 7.8.

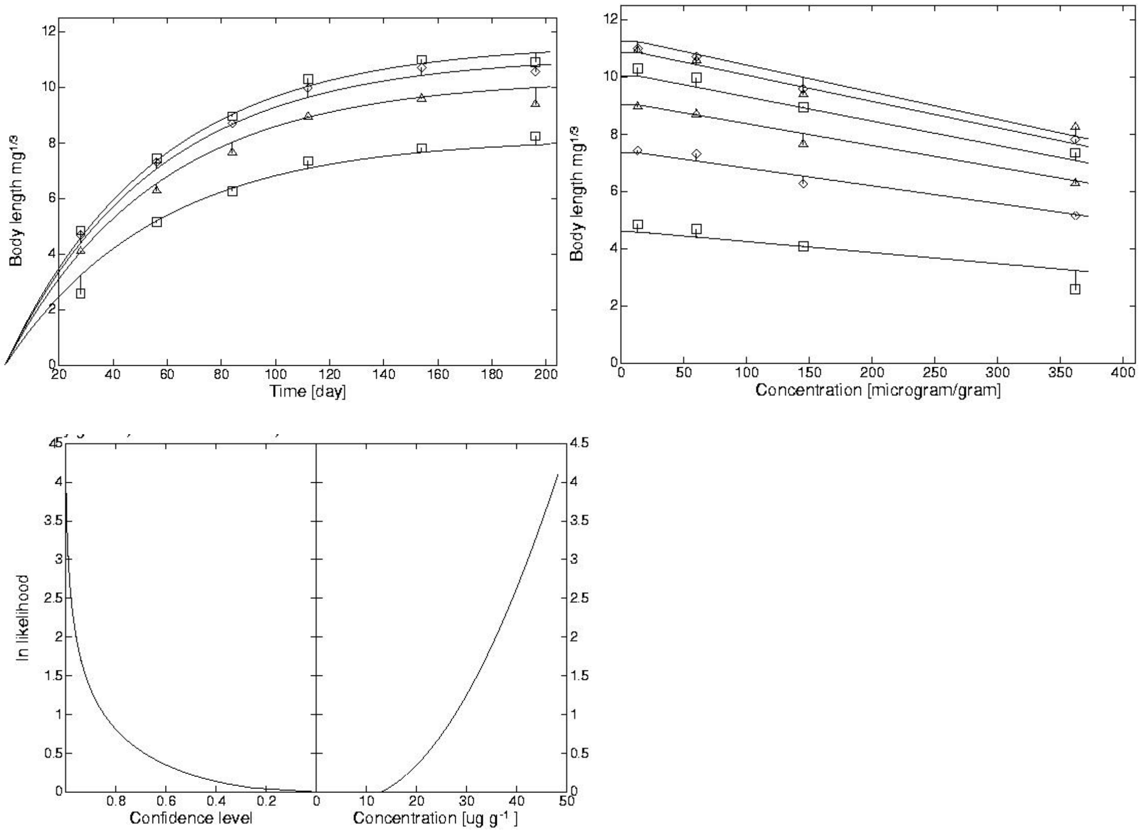


Figure 7.4 The time and concentration profiles for effects on growth of *Lumbricus rubellus* via a decrease of assimilation by copper chloride (data from Klok & de Roos 1996).

The parameter~~s~~ estimates are: NEC = 13 g/g; control ultimate length = 11.6 mm; tolerance conc = 1.2 mg/g; elimination rate = large; Fixed parameters are: initial length = 0 mm; von Bertalanffy growth rate = 0.018 d. The profile likelihood function for the NEC is given left. The EC0.100d = 13g/g; EC50.100d = 605 g/g.

435. The first example shows that it is not necessary to have observations in time; the second example shows that it is not absolutely necessary to have a control. Although inclusion of a control is always advisable, the control is treated in the same way as positive concentrations in the DEBtox method. The statistical properties of the parameter estimates and the confidence one has in them obviously improve if controls and positive concentrations are available.

436. At high concentrations, the test compound probably not only affects body growth, but usually also survival. The DEBtox software (see section 7.9) accounts for differences in number of individuals of which the body size has been measured.

437. The models for effects on body growth, and details about the statistical properties of parameter estimation (especially that of NECs) are discussed in Kooijman & Bedaux (1996, 1996a)

7.5. Reproduction

438. The DEB model allows for (at least) five routes that affect reproduction. The first three routes are identical to that for growth and are called the indirect routes. The DEB model assumes namely that food intake is proportional to surface area, so big individuals eat more than small ones. This means that if growth is affected, feeding is directly or indirectly affected as well, which leads to a change in resources

that are available for reproduction. The routes not only lead to a reduction of reproduction, but also to a delay of reproduction. In addition there are two direct routes for affecting reproduction

1. an increase in the costs per offspring, so an effect on the transformation from reserves of the mother to that of the embryo
2. death of early embryos, before they leave the mother. Dead embryos can be born, or are absorbed; only the living ones are counted.

439. These two direct routes assume that the allocation to reproduction is not affected by the compound, but that the compound affects the conversion of these resources into living embryos.

440. The following assumptions specify the effect on reproduction at any concentration of test compound:

- Assumptions on control behaviour
  - the test-organisms follow a von Bertalanffy growth curve in the control
  - reproduction depends on assimilation, maintenance and growth as specified by the Dynamic Energy Budget (DEB) theory
- Assumption on toxico-kinetics
  - the test chemical follows first order kinetics (Dilution by growth is taken into account.)
- Assumptions on effects: One of five modes of action occur
  - the assimilation rate decreases linearly in the internal concentration
  - the maintenance rate increases linearly in the internal concentration
  - the costs for growth increases linearly in the internal concentration
  - the costs for reproduction increases linearly in the internal conc.
  - the hazard rate of the neonates increases linearly in the internal conc.
- Assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure.
  - the measured cumulative numbers of young per female are independently normally distributed with a constant variance

441. An implication of the DEB theory is that indirect effects on reproduction (the first three modes of action) are a reduction of the reproduction rate as well as a delay of the start of reproduction, while direct effects (the last two modes of action) involve a reduction of reproduction only. All three indirect effects on reproduction also have effects on growth, despite the fact that just a single target parameter is affected. The delay of the onset of reproduction is, therefore, coupled to effects on growth. The measurement of body lengths at the end of the test on reproduction can be used as an easy check and as an identification aid to the mode of action. This mode of action is of importance to translate effects on individuals into those on growing populations (Kooijman 1985, Nisbet et al 2000).

442. The DEBtox software (see section 7.9) accounts for possible reductions of numbers of survivors in the reproduction test via weight coefficients; the more females contribute to the mean reproduction rate per female, the more weight that data point has in the parameter estimation. An example of application is from the OECD ring-test for effects of cadmium on Daphnia reproduction (Fig 7.5); the full results are reported in Kooijman at al (1998):

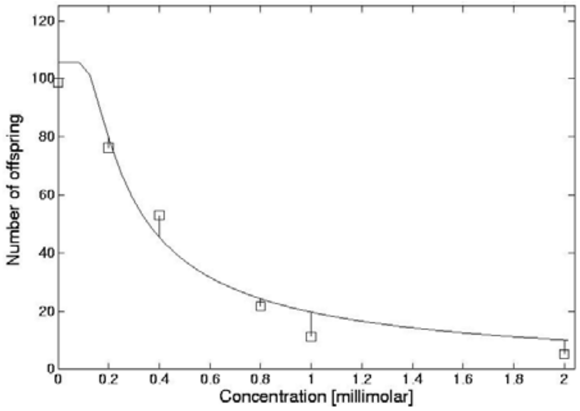


Figure 7.5 Effects of cadmium on the reproduction of Daphnia magna through an increase of the costs per offspring.

Data from the OECD ring-test. The figures show the time and concentration profiles. The Parameter estimates are: NEC = 3.85 nM, tolerance conc = 5.40 nM, max reproduction rate = 14.4 d, elimination rate = 3.0 d. Fixed parameters are: von Bertalanffy growth rate = 0.1 1/d, scaled length at birth = 0.13, scaled length at puberty = 0.42, energy investment ratio = 1. The NEC does not differ significantly from 0 on the basis of these data. If a more accurate estimate is required, lower test concentrations should be selected. These parameter values imply: EC0.21d = 0.1 mM and EC50.21d = 0.336 mM.

443. The models for effects on reproduction, and details about the statistical properties of parameter estimation (especially that of NECs) are discussed in Kooijman & Bedaux (1996b, 1996c).

7.6. Population growth

444. If individuals follow a cycle of embryo, juvenile and adult stages, one needs the context of physiologically structured population dynamics to link the behaviour of population dynamics to that of individuals. If the individuals only grow and divide, a substantial simplification is possible in the context of the DEB model. This is the case in the algal growth inhibition tests, and in tests with duckweed, for instance.

445. Three modes of action of the compound are delineated here. The following assumptions specify the model for effects on populations:

- Assumptions on control behaviour
  - the viable part of the population grows exponentially (the cultures are not nutrient or light limited during the test)
- Assumption on toxico-kinetics
  - the internal concentration is rapidly in equilibrium with the medium
- Assumptions on effects
  One of three modes of action occur
  - the costs for growth are linear in the (internal) concentration
  - the hazard rate is linear in the (internal) concentration during a short period at the start of the experiment
  - the hazard rate is linear in the (internal) concentration during the experiment

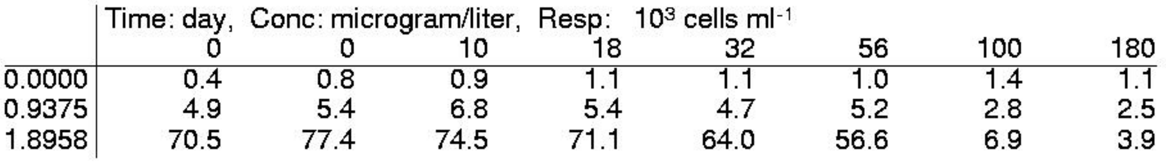
- Assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure.
  - the inoculum size is the same for all experimentally tested concentrations
  - biomass measurements include living and dead organisms
  - the measured population sizes are independently normally distributed with a constant variance

446. The rationale of the second mode of action (death only at the start of the experiment) is that effects relate to

- the transition from control culture to stressed conditions, not to the stress itself

- the position of the transition in the cell cycle; Cells are not synchronised, so the transition occurs at different moments in the cell cycle, for the different cells. If cells are more sensitive for the transition during a particular phase in the cell cycle, only those cells are affected that happen to be in that phase.

447. The ECx values for this type of test can be calculated in various ways, with different results. One way to do this is on the basis of biomass as a function of time. This should not be encouraged, however because the result depends on experimental design parameters that have nothing to do with toxicity (Nyholm 1985). Another way to do this is on the basis of specific population growth rates, which are independent of time (Kooijman et al 1996a). An example of application of the DEBtox method is as follows



	Time: day, Conc: microgram/liter, Resp: 10^3 cells ml^{-1}							
	0	0	10	18	32	56	100	180
0.0000	0.4	0.8	0.9	1.1	1.1	1.0	1.4	1.1
0.9375	4.9	5.4	6.8	5.4	4.7	5.2	2.8	2.5
1.8958	70.5	77.4	74.5	71.1	64.0	56.6	6.9	3.9

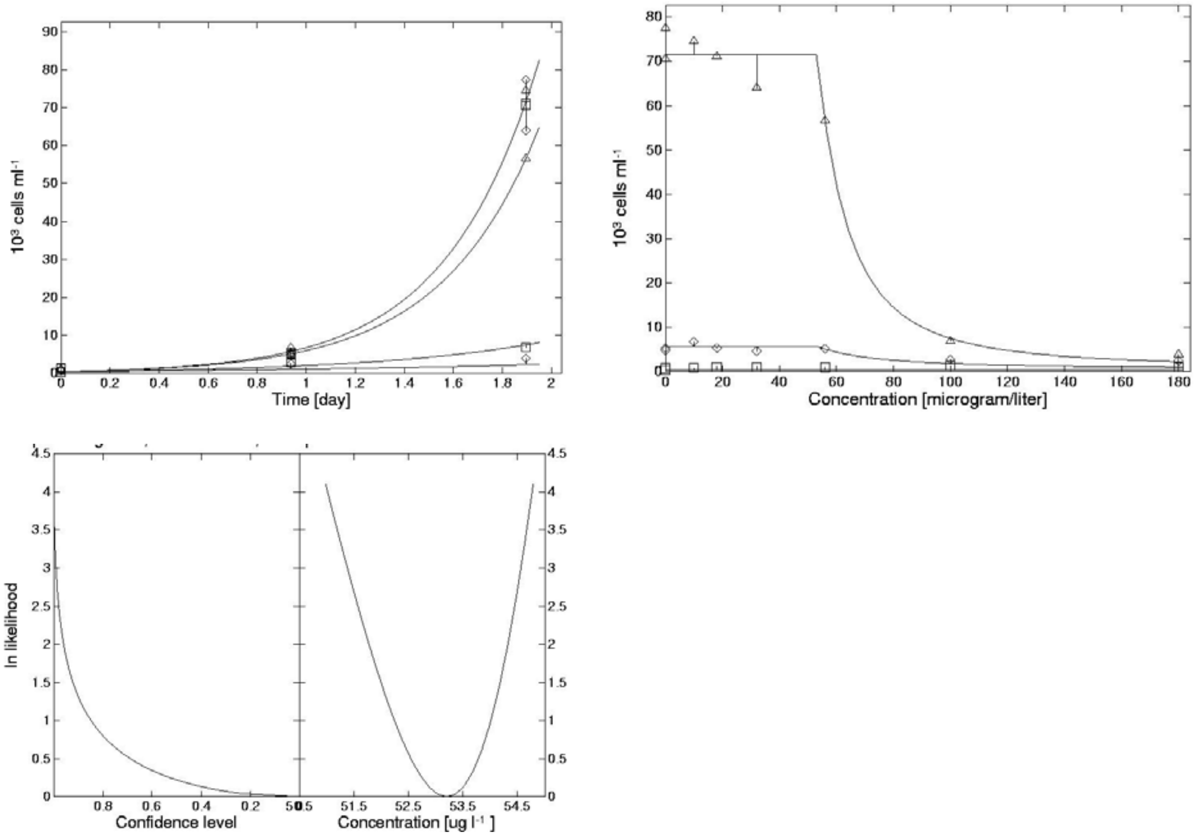


Figure 7.6. The effect of a mixture of C,N,S-compounds on the growth of *Skeletonema costatum* via an increase of the costs for growth (data from the OECD ring test).

The figures show the data, and the time and concentration profiles (note that this data set contains two blanks). The estimated parameters are: inoculum = 494 cells/ml, specific growth rate = 2.62 1/d, NEC = 0.053 mg/l, tolerance conc = 0.0567 mg/l. The profile likelihood function for the NEC is given in the figure left. The EC50 = 0.0624 mg/l. The robustness of this approach is demonstrated by the fact that removal of the highest concentration leads to the same point estimate for the NEC (but with a larger confidence interval).

448. The model for effects on population growth, and details about the statistical properties of parameter estimation (especially that of NECs) are discussed in Kooijman et al (1996a). Toxic effects on logistically growing populations in batch cultures are discussed in Kooijman et al (1983); a paper on the interference of toxic effects and nutrient limitation is in preparation.

7.7. Parameters of effect models

449. The parameters of effect models can be grouped into a set that relates directly to the effects of the test compound and a set that relates to the eco-physiological behaviour of the test organisms.

7.7.1. Effect parameters

450. The basic biology-based models have two toxicity parameters and a single dynamic parameter:

- NEC = EC0(∞): No-Effect Concentration, which is the 0% effect level at very long exposure times (dimension: external concentration).

- killing rate (for effects on survival; dimension: per external concentration per time) *or* tolerance concentration (for sublethal effects; dimension: external concentration).

- elimination rate of first order kinetics (for survival, body growth and reproduction tests; not for population growth inhibition tests. Dimension: per time). Large values mean that the internal concentration rapidly reaches equilibrium with the concentration in the medium. If the internal concentration is in equilibrium, the effects no longer change. Notice that the elimination rate has no information about the toxicity of the test compound.

451. The **killing rate** is the increase in the hazard rate per unit of concentration of test compound that exceeds the NEC:

- $$\text{hazard rate} = \text{control hazard rate} + \text{killing rate}\left(\frac{\text{internal concentration}}{\text{BCF}} - \text{NEC}\right)_+$$

where BCF = Bio-Concentration Factor and where the symbol $_+$ means that if internal conc./BCF is below NEC, then hazard rate equals control hazard rate. The BCF stands for the ratio of the internal and external concentration *in equilibrium*. No assumptions are made about its value; it can be very small for compounds that hardly penetrate the body.

452. The **tolerance concentration** quantifies the change in the target parameter per unit of concentration of test compound that exceeds the NEC:

- parameter value = control parameter value × (1 + stress value)
- $$\text{stress value} = \frac{1}{\text{tolerance concentration}}\left(\frac{\text{internal concentration}}{\text{BCF}} - \text{NEC}\right)_+$$

where BCF = Bio-Concentration Factor.

453. The target parameter value in this specification of the tolerance concentration can be the specific costs for growth, the specific maintenance costs or another physiological target parameter. This depends on the mode of action of the compound.

454. The name “tolerance concentration” refers to the fact that the higher its value, the less toxic the chemical compound. Notice that the ratio “internal concentration/ BCF” has the interpretation of an external concentration that is proportional to the internal concentration; the tolerance concentration, like the NEC, has the dimension of an external concentration. This is done because internal concentrations are generally unknown in practice. The internal concentration, and so the stress value, depends on the (constant) external concentration and the (changing) exposure time. The stress value is a dimensionless quantity, which is only introduced to simplify the specification of the change in the target parameter.

455. The NEC, the elimination rate and the tolerance concentration (or killing rate) are parameters that do NOT depend on the exposure time. This is in contrast to ECx values, which do depend on exposure time. Notice that the accumulation rate (a toxico-kinetic parameter) does not occur in the parameter set of effect models. This is because less toxic compounds that accumulate strongly cannot be distinguished from toxic compounds that hardly accumulate if only effects, and no internal concentrations, are observed. This is also the reason why NECs, killing rates and tolerance concentrations are in terms of external concentrations, while the mechanism is via internal concentrations. Effect models treat internal concentrations as hidden variables.

456. The kinetic parameters depend on the properties of the chemical compound. The elimination rate is inversely proportional to the square-root of the octanol-water partition coefficient (P_{ow}), while the uptake

rate is proportional to the square-root of this coefficient (Kooijman & Bedaux 1996, Kooijman 2000). Since effects depend on internal concentrations, so on toxico-kinetics, effect parameters depend on the partition coefficient as well; the NEC, tolerance concentration and inverse killing rate are all inversely proportional to the P_{ow} (Gerristen 1997, Kooijman & Bedaux 1996, Kooijman 2000). Such relationships can be used in practice to test parameter estimates against expectations.

457. The prediction of how the toxicity parameters depend on the octanol-water partition coefficient can be used for selecting appropriate concentrations to be tested. An example is as follows.

Suppose that compound 1 with $P_{ow} = 10^6$ has been tested for its effects on survival, which resulted in the parameter estimates: NEC = 1.3 mM; killing rate = 1.5 1/(mM.d); elimination rate = 0.5 1/d. Now have to test compound 2, with a physiologically similar mode of action and a $P_{ow} = 10^7$. Expect to find the parameter estimates NEC = 0.13 mM; killing rate = 15 1/(mM.d); elimination rate = 0.5/√10= 0.16 1/d. These three parameters imply that the LC0.2d = 0.47 mM and the LC99.2d = 1.9 mM, which gives some guidance for choosing the concentration range to be tested in a test of 2 d.

Suppose now that we tested compound 1 for effects on reproduction in Daphnia with a control max reproduction rate of 15 offspring per day. Let us assume that the compound increases the maintenance costs. This resulted in NEC = 1.3 mM, tolerance concentration = 10 mM; elimination rate = 0.5 1/d. We expect to find for compound 2: NEC = 0.13 mM, tolerance concentration = 1 mM; elimination rate = 0.16 1/d. These three parameters imply that the EC0.21d = 0.18 mM and the EC99.21d = 1.9 mM, which gives some guidance for choosing the concentration range to be tested in a reproduction test of 21 d. (Calculations with DEBtool, see 7.9.2)

458. Contrary to the more usual techniques to establish Quantitative Structure Activity Relationships (QSARs), the influence of the P_{ow} on the parameters of biology-based models can be predicted on the basis of first principles; these QSARs are not derived from regression techniques that require toxicity data for other compounds. The reason why traditional regression techniques for establishing QSARs are somewhat cumbersome is in the standardisation of the exposure period. For any fixed exposure period (usually 2d or 14d) the LC50 (or EC50) for a compound with a low P_{ow} is close to its LC50 for very long exposure times; for compounds with a large P_{ow}, however, the ultimate LC50 is much lower than the observed one. If we compare LC50s for low and high P_{ow} values, we observe complex deviations from simple relationships, which are masked in log-log plots and buried in the allometric models that are usually applied to such data. (An allometric model is a model of the type $y(x) = a\, x^b$ where a and b are parameters.)

459. Effects of modifying factors, such as pH, can be predicted, and taken into account in the analysis of toxicity data (corrections on measured or nominal concentrations, and on measured or modelled pH values). If the compound affects the pH at concentrations where small effects occur, and the NEC and/or the killing rate of the molecular and ionic forms differ, the relationships

$$b_k(pH) = \frac{b_k^m + b_k^i 10^{pH-pK}}{1+10^{pH-pK}} \quad \text{and} \quad c_0(pH) = c_0^i c_0^m \frac{1+10^{pH-pK}}{c_0^i + c_0^m 10^{pH-pK}}$$

apply, where pK is the ion-product constant, and are the NECs of the molecular and ionic forms, and are the killing rates of the molecular and ionic forms (Kooijman 2000, Könemann 1980). The pH is affected much more easily in soft than in hard water (see e.g. Segel 1976, Stumm & Morgan 1996). Compounds may affect internal pH to some extent; in that case the relationship is approximately only.

460. On the assumption that the chemical environment inside the body of the test organisms is not affected (due to homeostatic control), the observed survival pattern can be used to infer about the toxicity

of the molecular and the ionic form. The partitioning between the molecular and ionic form is fast, relative to the uptake and elimination (both in the environment and in the organism); this means that the elimination rate relates to both the molecular and the ionic form. An example is as follows.

PH	7.5	7.5	7.4	7.2	6.9	6.6	6.3	6.0
Conc	0	3.2	5.6	10	18	32	56	100
0	20	20	20	20	20	20	20	20
1	20	20	20	20	20	20	19	18
2	20	20	19	19	19	18	18	18
3	20	20	17	15	14	12	9	8
4	20	18	15	9	4	4	3	2
5	20	18	9	2	1	0	0	0
6	20	17	6	1	0	0	0	0
7	20	5	0	0	0	0	0	0

461. Suppose that we found the numbers of survivors as in the left table for a compound with ionisation product constant of 9.0. The parameter estimates are (calculations with DEBtool, see 7.9.2):

	Molecule		Ion	
	ML	sd	ML	Sd
Control mort rate	0.009	0.005		
NEC	24.9	16.9	0.17	0.03
Killing rate	0.039	0.013	2.82	2.16
Elimination rate	1.48	0.50		

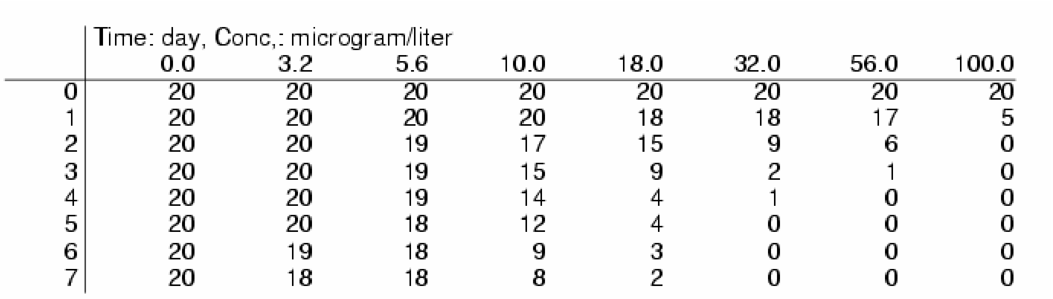
462. The elimination rate is proportional to the ratio of a surface area and a volume of the test organism, which yields an inverse length measure. This relationship implies predictable differences between elimination rates in organisms of different sizes, which have been tested against experimental data (see e.g. Gerritsen 1997). This is rather straightforward in the case of individuals of the same species, but also applies to individuals of different, but physiologically related, species. The body size scaling relationships as implied by the DEB theory suggest predictable differences in the chemical body composition, in lipid content and in elimination rate and toxicity parameters. Such relationships still wait for testing against experimental data, but are helpful in developing an expectation for parameter values; such expectations can be used in experimental design, and in checking results of parameter estimations.

463. The prediction of how the three parameters of the hazard model depend on the body size of the test organisms can also be used for selecting appropriate concentrations to be tested. An example is as follows:

Suppose that a compound has been tested using fish of a weight of 1 mg, which resulted in the parameter estimates: NEC = 1.3 mM; killing rate = 1.5 1/(mM.d); elimination rate = 0.5 1/d. Now we have to test the compound for fish of 1 g of the same species. We expect to find a difference in the elimination rate only, i.e. 0.5/10= 0.05 1/d. These three parameters imply that the LC0.2d = 1.4 mM and the LC99.2d = 5.5 mM, which gives some guidance for choosing the concentration range to be tested in a test of 2 d. (Calculations with DEBtool, see 7.9.2)

7.2.2. Eco-physiological parameters

464. The model for effects on survival has the **control mortality rate** as a parameter, which results in an exponentially decaying survival probability. This means that the model delineates two causes for death: death due to background causes (for instance manipulation during the assay) and death due to the compound. This obviously complicates the analysis of the death rate at low exposure levels, because we can never be sure about the actual cause of death in any particular case. Not only the data in the control, but all data are used to estimate the control mortality rate; if no death occurs in the control, this does not imply that the control mortality rate is zero. The profile likelihood function for the NEC quantifies the likelihoods of the two different causes of death. Figures 7.2, 7.3 and 7.4 show how background causes can be distinguished from those by the compound.



	Time: day, Conc,: microgram/liter							
	0.0	3.2	5.6	10.0	18.0	32.0	56.0	100.0
0	20	20	20	20	20	20	20	20
1	20	20	20	20	18	18	17	5
2	20	20	19	17	15	9	6	0
3	20	20	19	15	9	2	1	0
4	20	20	19	14	4	1	0	0
5	20	20	18	12	4	0	0	0
6	20	19	18	9	3	0	0	0
7	20	18	18	8	2	0	0	0

Figure 7.7 A typical table of data that serves as input for the survival model, as can be used in the software package DEBtox (Kooijman & Bedaux 1996).

The data in the body represent the number of surviving guppies. The first column specifies the observation times in days, the first row specifies the concentrations of dieldrin in g/l. Figure 7.8 shows how an answer can be found to the question whether the two deaths in the concentrations 3.2 and 5.6g/l are due to dieldrin, or to “natural” causes.

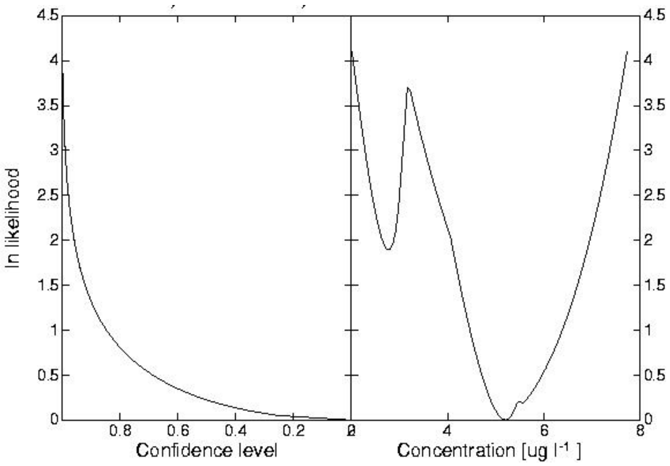


Figure 7.8 This profile likelihood function of the NEC (right panel) for the data in Figure 7.7 results from the software package DEBtox (Kooijman & Bedaux 1996).

It determines the confidence set for the NEC (first select the confidence level of your choice in the left panel, then read the ln likelihood; the concentrations in the right panel for which the ln likelihoods are below this level comprise the confidence set of the NEC; the confidence set for the NEC is a single interval for low confidence levels, but a set of two intervals for high confidence levels). The maximum likelihood estimate for the NEC is here 5.2 g/l, and corresponds to the interpretation of death in concentration 3.2g/l due to "natural" causes; the second local extreme at 2.9g/l corresponds to the interpretation of this death due to dieldrin. The figure shows that this interpretation is less likely, but the figure shows that we cannot be excluded this possibility for high confidence levels. If the lowest concentration would have no deaths in this data set, the profile likelihood function would not have a second local extreme.

465. The model for effects on growth have a single eco-physiological parameter each (the **ultimate body length**, and the **maximum reproduction rate**), that is estimated from the data, and a **scatter parameter** that stands for the standard deviation of the normally distributed deviations from the model predictions. The latter parameter also occurs in the models for effects on population growth.

466. The models for effects on body growth and reproduction have some parameter values that cannot be estimated from (routine) tests. Their values should be determined by preliminary eco-physiological experiments. These parameters are

- **von Bertalanffy growth rate** (dimension: per time). This parameter quantifies how fast the initial length approaches the ultimate length at constant food density. (The food density affects this parameter.) In principle, its value could be extracted from length measurements in the control, provided that enough observation times are included. Under standardised experimental conditions, its value should always be the same, however. Moreover, the lengths are usually only measured at the end of the test only. These data do not have information about the value of the von Bertalanffy growth rate.
- **initial body length** (dimension: length), which is the body length at the start of the test. It is assumed that this applies to all individuals in all concentrations. The DEB model for reproduction has a scaled length at birth as parameter, which is dimensionless. This scaled length is the ratio of the length at birth and the maximum length of an adult at abundant food. Since the daphnia reproduction test uses neonates, the initial body length equals the length at birth.
- **scaled length at puberty** (dimensionless). This is the body length at the start of reproduction in the control as a fraction of the maximum body length of an adult at abundant food. The DEB model takes this value to be a constant, independent of the food density. At low food density, it takes a relatively long time to reach this length. The start of reproduction, therefore, depends on

food density. The model for effects on reproduction needs the length at puberty. That on body growth does not use this parameter.

- **energy investment ratio** (dimensionless). This parameter stands for the ratio between the specific energy costs for growth and the product of the maximum energy capacity of the reserves and the fraction of the catabolic energy flux that is allocated to somatic maintenance plus growth. The maximum (energy) capacity of the reserves is reached after prolonged exposure to abundant food. The catabolic flux is the flux that is mobilised from the reserves to fuel metabolism (i.e. allocation to somatic and maturity maintenance, growth, maturation or reproduction; the relative allocation to somatic maintenance plus growth is taken to be constant in the DEB model). The value of the parameter does not affect the results in a sensitive way. The logic behind the DEB theory requires its presence, however; the parameter plays a more prominent role at varying food densities.

467. The DEBtox software (see below) fixes these parameters at appropriate default values for the standardised tests on fish growth and daphnia reproduction. The user can change these values.

468. The models for population growth have two eco-physiological parameters that are estimated from the data

- the **inoculum size** (dimension: mass or number per volume), which is taken to be equal in all concentrations
- the control **specific population growth rate** (dimension: per time)

7.8. Recommendations

7.8.1. Goodness of fit

469. As applies to all models that are fitted to data, one should always check for goodness of fit (as incorporated in DEBtox), inspect the confidence intervals of the NEC, and mistrust any conclusion from models that do not fit the data (see also Section 6.4). The routine presentation of graphs of model fits is strongly recommended. “True” models, however, do not always fit the data well, due to random errors. If deviations between data and model-fits are unacceptably large, it makes sense to make sure that the experimental results are reproducible. Problems with solubility of the test compound, pH effects, varying concentrations, varying conditions of test animals, interactions between test animals and other factors can easily invalidate model assumptions. It might be helpful to realise that one approach for solving this problem is in taking such factors into account in the model (and apply a more complex model), but another approach is to change the experimental protocol such that the problems are circumvented. The models are designed to describe small effects; if the lack of fit relates to large effects, it can be recommended to exclude the high concentration(s) from the data analysis.

470. Any model might fit data well for the wrong reasons; a good fit does not imply the “validity” of that model. This should motivate to explore all possible means for checking results from data analysis; an expectation for the value of parameters is a valuable tool.

471. The assumption of first order kinetics is not always realistic in detail. A general recommendation is to consider more elaborate alternatives only if data on toxico-kinetics are available. Depending on the given observation times, the elimination rate is not always accurately determined by the data. In such cases one might consider to fix this parameter at a value that is extracted from the literature, and/or derived from a related compound, after correction for differences in P_{ow} values.

7.8.2. Choice of modes of action

472. Experience teaches that the mode of action usually has little effect on the NEC estimates. Models for several modes of action frequently fit well to the same experimental data set; if additional type of measurements would have been available (such as feeding rate and/or respiration rate), it is much easier to choose between modes of action. These modes of action are of importance to translate effects on individuals to those on population dynamics, and how food availability interferes with toxic effects. The DEB theory deals with this translation.

473. Measurements of feeding and respiration rates, and of body size (in reproduction tests) greatly help identifying the mode of action of the compound. The proper identification of the mode of action is less relevant for estimates of the NEC.

7.8.3. Experimental design

474. DEBtox has been designed to analyse the results from toxicity tests as formulated in OECD guidelines (numbers 201, 202, 203, 204, 211, 215, 218, 219) and ISO guidelines (numbers 6341, 7346-3, 8692, 10229, 10253, 12890, 14669). The experimental design described in these guidelines is suitable for the application of DEBtox. Confidence intervals for parameter estimates are greatly reduced if not only the responses at the end of the toxicity experiments are used, but also observations during the experiment. Ideally, one should be able to observe how fast effects build up during exposure in the data, till the effect levels satiate. Note that this does not require additional animals to be tested, only that they are followed for a longer period of time.

475. Large extrapolations of effects, especially in the direction of longer exposure times, are generally not recommended; this is because, ideally, the assumptions need to be checked for all new applications. It, therefore, makes sense to let the optimal choice for the exposure period depend on the compound that is tested, and the test organisms that are used. The higher the solubility in fat of the test compound (e.g. estimated from P_{ow}), and the larger the body size of the test organisms, the longer the exposure should last.

476. As stated in the introduction, it is strongly recommended to include all available observations into the analysis; not only those at the end of the experiment, but also the observations that have been collected during the experiment (for instance when the media are refreshed). It is generally recommended that the number of observations during exposure, the concentrations of test compound and the number of used test animals are such that the model parameters can be estimated within the desired accuracy.

477. Experimental design should optimise the significance of the test; the significance of single-species tests is discussed in Anonymous (1999). From a data analysis point of view, it makes sense to extend the exposure period till no further effects show up. The length of the exposure period then relates to the physical-chemical properties of the compound.

7.8.4. Building a database for raw data

478. Since biology-based methods not only aim at a description, but also at an understanding of the processes that underlie effects, it is only realistic to assume that this understanding will evolve over the years. In the future, it might be useful to reanalyse old data in the light of new insights. In anticipation of this, it is recommended to build a raw database.

7.9. Software support

479. The models that are used by biology-based methods are fully derived and discussed in all mathematical detail in the open literature; a summary of the specification is given in the appendix of this

report. There is, therefore, no need to use any of the software that is mentioned in this section. On the other hand, fitting sets of differential equations to data (as required by the models for effects on body growth and reproduction), the calculation of profile likelihoods for NECs, and the more advanced methods of fitting several datasets simultaneously, is beyond the capacity of most standard packages. Even if packages can do the job, the optimisation of numerical procedures (such as solving initial value problems) can be somewhat laborious.

480. The computations for biology-based methods have been coded in two packages, DEBtox and DEBtool, which can be downloaded freely from the electronic DEB-laboratory at http://www.bio.vu.nl/thb/deb/. Both packages are updated at varying intervals; the user has to check the website for the latest version. These packages are used in (free) international internet-courses that are organised by the Dept Theoretical Biology at the Vrije Universiteit, Amsterdam.

481. A MS Excel macro able to estimate Hill parameters using nonlinear regression is available under the GPL license on the site: http://perso.wanadoo.fr/eric.vindimian

7.9.1. DEBtox

482. DEBtox is a load-module for Windows and Unix that is meant for routine applications.

483. The user cannot define new models. The package has many options for parameter estimation, confidence intervals and profile likelihoods (for the NEC for instance), fixation of parameters at particular values (such as NEC = 0) while estimating the other parameters, calculation of statistics (such as ECx.t and ETx.c values and their confidence intervals), hypothesis testing about parameter values (such as NEC ≠ 0), graphical representations to check goodness of fit, residual analysis, etc. Example data-files are provided for each toxicity test.

484. DEBtox is a user-friendly package, and the numerical procedures are optimised for the various models (modes of action) that can be chosen. The elimination rate, for instance, is not always accurately determined by the data, especially if a single observation time is given. DEBtox always calculates three sets of parameter estimates, corresponding with the elimination rate being a free parameter, or zero, or infinitely large. Only the best result is shown. The initial values for the parameters that are to be estimated are selected automatically. In fact many trials (some hundred) are performed, and only the best result is shown. The user does not have to bother about these computational “details”. (The likelihood function can have many local maxima, depending on the model and on the observations. The result of the numerical procedure to find a local maximum depends on the initial value; we are only interested in the global maximum, however. This problem complicates non-linear parameter estimation in practice; it is an extra reason to check the result graphically in all applications.)

485. The present version of DEBtox can handle a single endpoint only (i.e. a single table of observations of responses at the various combinations of concentration and exposure time). In the period 2002-2006, DEBtox will be extended to include multiple samples to allow the analysis of effects on survival and reproduction simultaneously, and to test hypotheses about differences of parameter values between samples.

7.9.3. DEBtool

486. DEBtool is source code (in Octave and Matlab®) for Windows and Unix that is meant for research applications. Octave is freely downloadable, Matlab is commercial. DEBtool is much more flexible than DEBtox, but requires more knowledge for proper use; it is less user-friendly than DEBtox. Initial values for parameter estimations are not automatic, for instance. DEBtool has many domains that deal with the

various applications of DEB models in eco-physiology and biotechnology; the domain "tox" deals with applications in ecotoxicology. The package can handle multiple data sets; several numerical procedures can be selected to find parameter estimates. DEBtool allows researchers to estimate parameters if the variance is proportional to the squared mean, to calculate the NEC, killing rate and elimination rate from LC50 values for three exposure times, to estimate parameters from time-to-death data, and to extract the toxicity parameters for the molecular and the ionic form when the pH is measured for each concentration, etc. Many specific models are coded, and the user can change and add models.

8. LIST OF EXISTING GUIDELINES WITH REFERENCE TO THE CHAPTERS OF THIS DOCUMENT

Guideline/standard		Test	Endpoint	Reference (NOEC)	Reference (dose response modelling)	Reference (biological based models)
OECD 201	ISO 8692: 1989 ISO 14593: 1999 ISO 13829: 2000	Alga growth inhibition	Growth rate	5.3	6.3	7.6
			Area under growth curve (biomass)	5.3	6.3	(not recommended)
OECD 202	ISO 6341: 1996	Daphnia immobilisation	Immobilisation	5.2	6.2	7.3
OECD 203	ISO 7346-1, 2, 3: 1996	Fish acute	Mortality	5.2	6.2	7.3
OECD 204		Fish prolonged	Mortality	5.2	6.2	7.3
			Body weight, length	5.3	6.3	7.4
OECD xxx		Avian acute	Mortality	5.2	6.2	7.3
OECD 205		Avian dietary	Mortality	5.2	6.2	7.3
			Body weight	5.3	6.3	7.4
			Food consumption	5.3	6.3	7.4 (theory covered, but not coded)
OECD 206 OECD xxx		Avian-1-generation	Body weight (F0, F1), organ weight, food consumption, egg-shell thickness, egg-shell strength	5.3	6.3	7.4 (body weight)
			Egg production, 14-day old survivors (counts)	5.2/5.3	6.3	7.5
			Egg abnormality rate, egg fertility, viability, hatchability, chick survival rate (proportions)	5.2	6.2	7.3 (chick survival)
OECD 207	ISO 11268-1: 1993	Earthworm acute	Mortality	5.2	6.2	7.3
			Body weight	5.3	6.3	7.4
OECD 208	ISO 11269-1: 1993 ISO 11269-2: 1995	Non-target terrestrial plant	Emergence	5.2	6.2	7.3 (as survival)
			Biomass, Root Length	5.3	6.3	theory covered, but not coded
			Visual phytotoxicity	5.3	6.3	

Guideline/standard		**Test**	**Endpoint**	**Reference (NOEC)**	**Reference (dose response modelling)**	**Reference (biological based models)**
			Mortality	5.2	6.2	7.3
	ISO 15522: 1999	Activated sludge	Microorganism cell growth	5.3	6.3	7.6
OECD 210		Fish ELS	Mortality	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Hatching success	5.2	6.2	7.3
			Days to swim-up	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Weight, length	5.3	6.3	7.4
OECD 211	ISO 10706: 2000	Daphnia reproduction	Immobilisation	5.2	6.2	7.3
			Fecundity	5.2/5.3	6.3	7.5
OECD 212	ISO 12890: 1999	Fish embryo and sac-fry stage	Mortality	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Length	5.3	6.3	7.4
OECD 213		Honeybee, acute oral	Mortality	5.2	6.2	7.3
OECD 214		Honeybee, acute contact	Mortality	5.2	6.2	7.3
OECD 215	ISO 10229: 1994	Fish juvenile growth test	Mortality	5.2	6.2	7.3
			Body weight, Length	5.3	6.3	7.4
OECD 218/219		Chironomid toxicity	Emergence	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4
			Survival	5.2	6.2	7.3
			Weight	5.3	6.3	7.4
OECD 220	ISO/CD	Enchytraeidae reproduction	Mortality	5.2	6.2	7.3
			Fecundity	5.2/5.3	6.3	7.4
OECD xxx	ISO 11268-2: 1998	Earthworm reproduction	Mortality	5.2	6.2	7.3
			Body weight	5.3	6.3	7.4
			Fecundity	5.2/5.3	6.3	7.5
	ISO 11268-3: 1999	Earthworm population size (field test)	Number of individuals (for various species)	5.2/5.3	6.3	7.6
OECD 221	ISO/CD 20079	Lemna growth inhibition	Average growth rate	5.3	6.3	7.4
			Area under growth curve	5.3	6.3	not recommended

Guideline/standard		Test	Endpoint	Reference (NOEC)	Reference (dose response modelling)	Reference (biological based models)
			Final biomass	5.3	6.3	7.4
	ISO 10253: 1995	Marine algal growth inhibition test	Growth rate	5.3	6.3	7.6
			Biomass	5.3	6.3	7.6
	ISO 14669: 1999	Marine copepods	Immobilisation	5.2	6.2	7.3
	ISO 11348-1,2,3: 1998	Light emission of Vibrio fischeri	Luminenscence	5.3	6.3	7.6 (for zero growth)
	ISO 10712: 1995	Pseudomonas putida growth inhibition	Growth rate	5.3	6.3	7.6
	ISO 13829: 2000	Genotoxicity (umu-test)	Induction rate	5.3	6.3	
	ISO 11267: 1999	Collembola reproduction inhibition	Offspring number	5.2/5.3	6.3	7.5
			Mortality	5.2	6.2	7.3

REFERENCES

References for chapter 1 to 4

Akritas, M.G. and I. Van Keilegom. (2001) - ANCOVA methods for heteroscedastic nonparametric regression models. *Journal of the American Statistical Association*. 96, 453, 220-231.

Armitage A.C. and Berry G. (1987) - Statistical methods in medical Research. Oxford, Blackwell.

Atkinson A.C. (1987) - Plots, transformations and regression. Oxford: Oxford University press.

Azzalini A. and Bowman A. W. (1997) - *Applied Smoothing Techniques for Data Analysis*, Oxford, pp. 48-85

ASTM (2000) - E1847-96 - Standard Practice for Statistical Analysis of Toxicity Tests Conducted under ASTM Guidelines. ASTM Annual Book of Standards, Vol. 11.05. ASTM, West Conshohocken, Pennsylvania.

Belsey D.A., Kuh E. and Welsch R.E. (1980) - Regression diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Box, G.E.P. and Cox, D.R. (1964) - An analysis of transformations, *J. Roy. Statist. Soc.* B-26, 211-252.

Box, G.E.P. and Hill, W.J. (1974) - Correcting inhomogeneity of variance with power transformation weighting, *Technometrics* 16, 385-389.

Box, G.E.P. and Tidwell, P.W. (1962) - Transformations of the independent variables, *Technometrics* 4, 531-550.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999), “Nonparametric regression with errors in covariates”, *Biometrika*, 86, 541–554.

Chapman P.F., Crane M., Wiles J.A., Noppert F. and McIndoe E.C. (1995) - Asking the right questions: ecotoxicology and statistics*. In: Report of a workshop held at Royal Holloway University of London, Egham, Surrey, United Kingdom*, , SETAC Eds, 32 p.

Chapman P.M., Caldwell R.S. and Chapman P.F. (1996). - A warning: NOECs are inappropriate for regulatory use. *Environ. Toxicol. Chem.* Vol. 15, No.2, pp. 77-79

Cook R.D. and Weisberg S. (1982) - Residuals and Influence in Regression. New York: Chapman Hall.

Draper, N.R. and Cox, D.R. (1969) - On distributions and their transformations to normality, *J. Roy. Statist. Soc.* B-31, 472-476.

Easton D, Peto J. (2000); Presenting statistical uncertainty in trends and dose–response relations. American Journal of Epidemiology, 152:393–394.

Environment Canada (2003) - Guidance document on statistical methods for environmental toxicity tests. Fifth draft, Environmental Protection Series, Method development and application section, Environmental technology centre, Environmental protection service, Ottawa, Ontario.

Fan J. and Gibjels I. (1996), *Local Polynomial Modelling and Its Applications,* London: Chapman& Hall.

Finney D.J. (1978) - Statistical method in biological assay. London., Griffin.

Green, P. J. and Silverman, B. W. (1994), Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman and Hall, London.

Hardle, W. (1991), *Smoothing Techniques,* London: Springer-Verlag.

Hoekstra J.A. and Van Ewijk P.H. (1993) - Alternatives for the no-observed-effect level. *Environ. Toxicol. Chem.* Vol. 12, No.2, pp. 187-194

Hochberg Y. and Tamhane A.C. (1987) - Multiple comparison procedures, Wiley, New York.

Hurlbert S.H. (1984) - Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54, 187-211.

Kerr D.R. and Meador J.P. (1996) - Modelling dose response using generalized linear models. *Environ. Toxicol. Chem.,* 15, 3, 395-401.

Kooijman S.A.L.M. and Bedaux J.J.M. (1996) - The analysis of aquatic ecotoxicity data. VU University Press, ISBN 90-5383-477-X

Laskowskj R. (1995) - Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *OIKOS* 73:1, pp.140-144

Mc Cullagh P. and Nelder J.A. (1983) - Generalized linear models. London, Chapman and Hall, p 261.

Newman M.C. (1994) - Quantitative Methods in Aquatic Ecotoxicology. Lewis Publishers.

OECD (1998) - Report on the OECD workshop on statistical analysis of aquatic toxicity data. Series on testing and assessment, N° 10. Environmental Health and Safety Publications. Series on testing and Assessment. ENV/MC/CHEM(98)18.

OECD (2000) - Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures. OECD Environmental Health & Safety Publication, Series on Testing & Assessment No. 23. Organisation for Economic Cooperation & Development (OECD), Paris. 53 pp.

Pack S. (1993) - A review of statistical data analysis and experimental design in OECD aquatic toxicology Test Guidelines.

Piegorsch W. W. and Bailer A. J. (1997) - Statistics for Environmental Biology and Toxicology, Boca Raton, FL: Chapman & Hall/CRC Press.Silverman, B. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). Journal of the Royal Statistical Society B 47: 1-52

Smith-Warner, S.A., Spiegelman, D., Yaun, S.S., van den Brandt, P.A., Folsom, A.R., Goldbohm, R.A., Graham, S., Holmberg, L., Howe, G.R., Marshall, J.R., Miller, A.B., Potter, J.D., Speizer, F.E., Willett, W.C., Wolk, A., Hunter, D.J., 1998. Alcohol and breast cancer in women: a pooled analysis of cohort studies. Journal of the American Medical Association 279, 535–540.

Sparks T. (2000) - Statistics in Ecotoxicology. John Wiley and Sons, Ltd., West Sussex, England. 320 p.

Tukey J.W., Ciminera J.L. and Heyes J.F. (1985) - Testing the statistical certainty of a response to increasing doses of drug. *Biometrics,* 41, 295-301.

Williams D.A. (1971) - A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics,* 27, 103-117.

References for chapter 5

These references serve both for quantal and continuous response hypothesis testing. Not all references included have been cited in the text. The additional references are included to provide the interested reader sources to further explore the issues and procedures that have been presented.

Agresti A. (1990) - Categorical Data Analysis, Wiley, New York.

Aitkin M., Anderson D., Francis B. and Hinde J. (1989) - Statistical analysis in GLIM, Oxford Science Publications, Clarendon Press, Oxford.

Aiyar R.J., Guillier C.L. and Albers, W. (1979) - Asymptotic relative efficiencies of rank tests for trend alternatives, J. American Statistical Association 74, 226-231.

Alldredge J. R. (1987) - Sample size for monitoring of toxic chemical sites, Environmental Monitoring and assessment 9, 143-154.

Armitage P. (1955) - Tests for linear trends in proportions and frequencies, Biometrics 11, 375-386.

Barlow R.E., Bartholomew D.J., Bremmer J.M. and Brunk, H.D. (1972) - Statistical inference under order restrictions, Wiley 1972.

Bartholomew D.J. (1961) - Ordered tests in the analysis of variance, Biometrika 48, 325-332.

Bauer P. (1997) - A note on multiple testing procedures in dose finding, Biometrics, 53, 1125–1128.

Berenson B.M. (1982a) - A comparison of several k sample tests for ordered alternatives in completely randomized designs, Psychometrika 47, 265-280.

Berenson M. L. (1982b) - A study of several useful tests for ordered alternatives in the randomized block design, Comm. Statistical (B) 11, 563-581.

Berenson M. L. (1982c) - Some useful nonparametric tests for ordered alternatives in randomized block experiments, Comm. Statistical (A) 11, 1681-1693.

Birch J.B. and Myers R.H. (1982) - Robust Analysis of Covariance, Biometrics 38, 699-713.

Bliss C.L. (1957) - Some principals of bioassay, Am. Sci. 45, 449-466.

Box G.E.P. and Cox, D.R. (1964) - An analysis of transformations, J. Roy. Statist. Soc. B-26, 211-252.

Box G.E.P. (1953) - Non-normality and tests on variances. Biometrika 40: 318-335.

Box G.E.P. and Hill W.J. (1974) - Correcting inhomogeneity of variance with power transformation weighting, Technometrics 16, 385-389.

Box G.E.P. and Tidwell P.W. (1962) - Transformations of the independent variables, Technometrics 4, 531-550.

Breslow N. (1990) - Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models, JASA 85, 565-571.

Bretz F. (1999) - Powerful modifications of Williams' test on trends, Dissertation, University of Hanover.

Bretz F. and Hothorn L.A. (2000) - A powerful alternative to Williams' test with application to toxicological dose-response relationships of normally distributed data, Environmental and Ecological Statistics 7, 135-254.

Brown M. B. and Forsythe A. B. (1974) - The small sample behavior of some statistics which test the equality of several means, Technometrics 16, 129-132.

Budde M. and Bauer P. (1989) - Multiple test procedures in clinical dose finding studies, J. American Statistical Association 84, 792-796.

Capizzi T., Oppenheimer L., Mehta H., Naimie H. and Fair J. L. (1984) - Statistical considerations in the evaluation of chronic aquatic toxicity studies, Envir. Sci. Technol. 19, 35-43.

Chase G.R. (1974) - On testing for ordered alternatives with increased sample size for a control, Biometrika 61, 569-578.

Cochran W. G. (1943) - Analysis of variance for percentages based on unequal numbers, JASA 38, 287-301.

Cochran W. G. (1954) - Some Methods for Strengthening the Common χ^2-Tests, Biometrics 10, 417-451.

Collett D. (1991) - Modelling Binary Data, Chapman and Hall, London.

Conover W. J. and Iman R.L. (1982) - Analysis of Covariance Using the Rank Transform, Biometrics 38, 715-724.

Crump K.S., Guess H.A. and Deal K.L. (1977) - Confidence intervals and tests of hypotheses concerning dose response relations inferred from animal carcinogenicity data, Biometrics 33, 437-451.

Crump K.S. (1984) - A new method for determining allowable daily intakes, Fundam. Appl. Toxicol. 4, 854-871.

Crump K. S. (1979) - Dose response problems in carcinogenesis, Biometrics 35, 57-167.

Davis J.M. and Svendsgaard D.J. (1990) - U-Shaped Dose-Response Curves: Their Occurrence and Implications for Risk Assessment, J. Toxicology and Environmental Health 30, 71-83.

Draper N.R. and Smith H. (1981) - Applied Regression Analysis, 2nd edition, Wiley, New York.

Draper N.R. and Cox D.R. (1969) - On distributions and their transformations to normality, J. Roy. Statist. Soc. B-31, 472-476.

Dunnett C. W. (1964) - New tables for multiple comparisons with a control, Biometrics 20, 482-491.

Dunnett C. W. (1955) - A multiple comparison procedure for comparing several treatments with a control, J. American Statistical Association 50, 1096-1121.

Dunnett C. W. (2000) - Power and Sample Size Determination in Treatment vs. Control Multiple Comparisons, Submitted.

Dunnett C. W. and Tamhane A. C. (1998) - New multiple test procedures for dose finding, Journal of Biopharmaceutical Statistics, 8, 353 366.

Dunnett C. W. and Tamhane A. C. (1995) - Step-up multiple testing of parameters with unequally correlated estimates, Biometrics 51, 217-227.

Dunnett C. W. and Tamhane A. C. (1992) - A step-up multiple test procedure, Journal of the American Statistical Association, 87, 162–170.

Dunnett C. W. and Tamhane A. C. (1991) - Step-down multiple tests for comparing treatments with a control in unbalanced one-way layout, Statistics in Medicine 10, 939-947.

Dunnett C.W. (1980) - Pairwise multiple comparisons in the unequal variance case, J. Amer. Statist. Assoc. 75, 796-800.

Dunn O. J. (1964) - Multiple Comparisons Using Rank Sums, Technometrics 6, 241-252.

Fairweather P.G. (1991) - Statistical power and design requirements for environmental monitoring, Aust. J. Mar. Freshwater Res. 42, 555-567.

Fleiss J.L. (1986) - The design and analysis of clinical experiments, Wiley, New York.

Freeman M.F. and Tukey J.W. (1950) - Transformations Related to the Angular and the Square Root, Annals of Mathematical Statistics 21, 607-611.

Gad S.C. and Weil C.S. (1986) - Statistics and experimental design for toxicologists, Telford Press, Caldwell, NJ, p. 86.

Gaylor D.W. (1983) - The use of safety factors for controlling risk, J. Toxicology and Environmental Health 11, 329-336.

Genz A. and Bretz F. (1999) - Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts, J. Stat. Comp. Simul. 63, 361-378

Good P. (1994) - Permutation Tests, Springer-Verlag, New York.

Harwell M.R. and Serlin R.C. (1988) - An Empirical Study of a Proposed Test of Nonparametric Analysis of Covariance, Psychology Bulletin, 268-281.

Henderson C.R. (1982) - Analysis of Covariance in the Mixed Model: Higher-Level, Nonhomogeneous, and Random Regressions, Biometrics 38, 623-640.

Hirji K. F. and Tang M.-L. (1998) - A comparison of tests for trend, Commun. Statist. – Theory Meth. 27, 943-963.

Hochberg Y. and Tamhane A.C. (1987) - Multiple comparison procedures, Wiley, New York.

Hocking R. R. (1985) - The Analysis of Linear Models, Brooks/Cole, Monterey, CA.

Hollander M. and Wolfe D.A. (1973) - Nonparametric Statistical Methods, Wiley, New York.

Holm S. (1979) - A simple sequentially rejective multiple test procedure, Scand. J. Statist., 6, 65-70.

Hoppe F.M. ed. (1993) - Multiple Comparisons, Selection, and Applications in Biometry, Marcel-Dekker, New York.

Hosmer D. W. and Lemeshow S. (1989) - Applied logistic regression, Wiley, New York.

Hothorn, L.A., and Hauschke, D. (2000): Identifying the maximum safe dose: a multiple testing approach. J. Biopharmaceutical Statistics 10 15-30.

Hsu J.C. (1992) - The Factor Analytic Approach to Simultaneous Confidence Interval for Multiple Comparisons with the Best, *Journal of Computational Statistics and Graphics*, 1, 151 -168.

Hsu J. C. and Berger R. L. (1999) - Stepwise confidence intervals without multiplicity adjustment for dose response, and toxicity studies, Journal of the American Statistical Association, 94, 468 482.

Hubert J.J., Bohidar N.R. and Peace K.E. (1988) - Assessment of pharmacological activity, 83-148 in Biopharmaceutical statistics for drug development, K.E. Peace, ed, Marcel Dekker, New York.

Hubert J.J. (1996) - Environmental Risk Assessment, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario.

John P.W.M. (1971) - Statistical Design and Analysis of Experiments, Macmillan, New York. (especially section 4.7)

Jonckheere A. R. (1954); A distribution-free *k*-sample test against ordered alternatives, Biometrika 41, 133.

Klaassen C.D. (1986) - Principals of toxicology, Chapter 2 in Casarett and Doull's Toxicology the basic science of Poisson, third ed., C.D. Klaassen, M.O. Amdur and J. Doull, editors, MacMillan, New York.

Knoke J.D. (1991) - Nonparametric Analysis of Covariance for Comparing Change in Randomized Studies with Baseline Values Subject to Error, Biometrics 47, 523-533.

Koch G.C., Tangen C.M., Jung J.-W. and Amara I. (1998) - Issues for Covariance Analysis of Dichotomous and Ordered Categorical Data from Randomized Clinical Trials and Non-Parametric Strategies for Addressing Them, Statistics in Medicine 17, 1863-1892.

Kodell R.L. and Chen J.J. (1991) - Characterization of Dose-Response Relationships Inferred by Statistically Significant Trend Tests, Biometrics 47, 139-146.

Korn E.L. (1982) - Confidence Bands for Isotonic Dose-Response Curves, Appl. Statis. 31, 59-63.

Lagakos S. W. and Lewis, T.A. (1985) - Statistical analysis of rodent tumorigenicity experiments, in Toxicological Risk Assessment, Volume I Biological and Statistical Criteria, CRC Press, Boca Raton, Florida, 149-163.

Lehmann E. I. (1975) - Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Litchfield J.T. and Wilcoxon F. (1949) - Simplified method of evaluating dose-effect experiments, J. Pharmacol. Exp. Ther. 96, 99-113.

Littell R. C. (2002) - Analysis of Unbalanced Mixed Model Data: A Case Study Comparison of ANOVA versus REML/GLS, J. Agriculture, Biological, and Environmental Statistics 7, 472-490.

Marascuilo L.A. and McSweeney M. (1967) - Non-parametric post-hoc multiple comparisons for trend, Psychol. Bull 67, 401-412.

Marcus R., Peritz E. and Gabriel, K.R. (1976) - On closed testing procedures with special reference to ordered analysis of variance, Biometrika 63, 655-660.

Marcus R. and Peritz E. (1976) - Some simultaneous confidence bounds in normal models with restricted alternatives, J. Roy. Statistical Soc., Ser. B 38, 157-165.

Marcus R. (1976) - The powers of some tests of the equality of normal means against an ordered alternative, Biometrics 63, 177-183.

Marcus R. (1982) - Some results on simultaneous confidence intervals for monotone contrasts in one-way ANOVA model, Commun. Statist., Ser. A 11, 615-622.

Maurer W., Hothorn L. A. and Lehmacher W. (1995) - Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses, Biometrie in der chemisch-pharmazeutischen Industrie 6 (Ed. J. Vollmar), 3–18, Stuttgart: Gustav Fischer Verlag.

McCullagh P. and Nelder J.A. (1989) - Generalized linear models, second edition, Chapman and Hall, London.

Mehta C. and Patel N. (1999) - LogXact 4 for Windows, Cytel Software Corporation, Cambridge, MA.

Mehta C. and Patel N. (1999) - StatXact 4 for Windows, Cytel Software Corporation, Cambridge, MA.

Miller R. J. (1981) - Simultaneous statistical inference, second edition, Springer-Verlag, New York.

Milliken G.A. and Johnson D.A. (1984) - Analysis of Messy Data Volume I: Designed Experiments, Lifetime Learning Publications, Belmont, CA.

Morgan B.J.T. (1992) - Analysis of quantal response data, Chapman and Hall, London.

Mukerjee H., Robertson T. and Wright F. T. (1987) - Comparison of several treatments with a control using multiple contrasts, Journal of the American Statistical Association, 82, 902 -910.

Odeh R.E. (1972) - On the power of Jonckheere’s k-sample test against ordered alternatives, Biometrika 59, 467-471.

Odeh R.E. (1971) - On Jonckheere's k-Sample Test Against Ordered Alternatives, Technometrics 13, 912-918.

Olejnik S.F. and Algina J. (1984) - Parametric ANCOVA and the Rank Transform ANCOVA when the Data are Conditionally Non-Normal and Heteroscedastic, Journal of Educational Statistics 9, 129-149.

Oris J. T. and Bailer A. J. (1992) - Statistical analysis of the Ceriodaphnia toxicity test: sample size determination for reproductive effects, Environmental Toxicology and Chemistry 12, 85-90.

Peritz E. (1970) - A note on multiple comparisons, unpublished manuscript, Hebrew University.

Peritz E. (1965) - On inferring order relations in analysis of variance, Biometrics 21, 337-344.

Poon A.H. (1980) - A Monte-Carlo study of the power of some k-sample tests for ordered binomial alternatives, J. Statist. Comp. Simul. 11.

Potter R.W. and Sturm G.W. (1981) - The Power of Jonckheere's Test, The American Statistician 35, 249-250.

Potthoff R. F. and Whittinghill M. (1966) - Testing for homogeneity I. The binomial and multinomial distributions, Biometrika 53, 167-182.

Puri M.L. (1965) - Some distribution-free k-sample rank tests of homogeneity against ordered alternatives, Commun. Pure Applied Math. 18, 51-63.

Puri M.K. and Sen P.K. (1985) - Nonparametric Methods in General Linear Models, Wiley, New York.

Quade D. (1982) - Nonparametric Analysis of Covariance by Matching, Biometrics 38, 597-611.

Rao J.N.K. and Scott A.J. (1992) - A simple method for the analysis of clustered binary data, Biometrics 48, 577-585.

Robertson T., Wright F.T. and Dykstra R.L. (1986) - Advances in order-restricted statistical inference, Springer-Verlag.

Robertson, T., Wright F.T. and Dykstra R.L. (1988) - Order restricted statistical inference, Wiley.

Rodda B.E., Tsianco M.C., Bolognese J.A. and Kersten M.K. (1988) - Clinical development, pp 273-328 in Biopharmaceutical statistics for drug development, K. E. Peace, ed., Marcel Dekker, New York.

Rom D. M., Costello R.J. and Connell L.T. (1994) - On closed test procedures for dose-response analysis, Statistics in Medicine, 13, 1583 -1596.

Rossini A. (1995, 1997) - Nonparametric Statistical Methods: Supplemental Text, http://software.biostat.washington.edu/~rossini/courses/intro-nonpar/text/.

Roth A.J. (1983) - Robust trend tests derived and simulated analogs of the Welch and Brown-Forsythe tests, J. American Statistical Association 78, 972-980.

Ruberg S.J. (1989) - Contrasts for identifying the minimum effective dose, Journal of the American Statistical Association, 84, 816–822.

Rothman K.J. (1978) - A show of confidence, New England Journal of Medicine 299, 1362-1363.

Salsburg D.S. (1986) - Statistics for Toxicologists, Marcel Dekker, New York, 85-86.

Sasabuchi S. and Kulatunga D.D.S. (1985) - Some approximations for the null distribution of the $\overline{E}^2$ statistic used in order-restricted inference, Biometrika 72.

Schoenfeld D. (1986) - Confidence intervals for normal means under order restrictions, with applications to dose-response curves, toxicology experiments and low dose extrapolation, J. American Statistical Association 81, 186-195.

Seaman S.L., Algina J. and Olejnik S.F. (1985) - Type I Error Probabilities and power of the Rank and Parametric ANCOVAL Procedures, Journal of Educational Statistics 10, 345-367.

Searle S. R. (1987) - Linear Models for Unbalanced Data, Wiley, New York.

Selwyn M. R. (1988) - Preclinical Safety Assessment, in K. E. Peace, ed., Biopharmaceutical Statistics for Drug Development, Marcel Dekker, New York.

Shapiro S.S. and Wilk M.B (1965) - An analysis of variance test for normality (complete samples). *Biometrika* 52: 591-611.

Shirley E. A. (1979) - The comparison of treatment to control group means in toxicology studies, Applied Statistics 28, 144-151.

Shirley E. A.C. (1981) - A Distribution-free Method for Analysis of Covariance Based on Ranked Data, Appl. Statist. 30, 158-162.

Shoemaker L.H. (1986) - A Nonparametric Method for Analysis of Variance, Comm. Statist.-Simula. 15, 609-632.

Simpson D.G. and Margolin B.H. (1986) - Recursive nonparametric testing for dose-response relationships subject to downturns at high doses, Biometrika 73.

Stephenson W.R. and Jacobson D. (1988) - A Comparison of Nonparametric Analysis of Covariance Techniques, Communications in Statistics - Simulations 17: 451-461.

Swallow William H. (1984) - Those overworked and oft-misused mean separation procedures-Duncan's, LSD, etc., Plant Disease 68, 919-921.

Tamhane A.C. (1979) - A comparison of procedures for multiple comparison of means with unequal variances, J. Amer. Statist. Assoc. 74, 471-480

Tamhane A. and Dunnett C.W. (1996) - Multiple test procedures for dose finding, Biometrics, 52, 21–37.

Tamhane A.C, Dunnett C.W., Green J.W. and Wetherington J.D. (2001) - Multiple Test Procedures for Identifying a Safe Dose, JASA 96, 835-843.

Tarone R.E. and Gart J.J. (1980) - On the robustness of combined tests for trends in proportions, JASA 75, 110-116.

Thall P.F. and Vail S.C. (1990) - Some Covariance Models for Longitudinal Count Data with Overdispersion, Biometrics 46, 657-671.

Thomas P.C. (1983) - Nonparametric estimation and tests of fit for dose response relations, Biometrics 39, 263-268.

Toft C.A. and Shea P.J. (1983) - Detecting community-wide patterns: estimating power strengthens statistical inference, The American Naturalist 122, 618-625.

Tukey J.W., Ciminera J.L. and Heyes J.F. (1985) - Testing the statistical certainty of a response to increasing doses of a drug, Biometrics 41, 295-301.

Tukey J. W. (1977) - Exploratory Data Analysis, Addison-Wesley, Reading, MA.

Tangen C.A. and Koch G.C. (1999) - Nonparametric Analysis of Covariance for Hypothesis Testing with Logrank and Wilcoxon Scores and Survival-Rate Estimation in a Randomized Clinical Trial, Journal of Biopharmaceutical Statistics 9, 307-338.

Tangen C.A. and Koch G.C. (1999) - Complementary Nonparametric Analysis of Covariance for Logistic Regression in a Randomized Clinical Trial, Journal of Biopharmaceutical Statistics 9, 45-66.

U.S.EPA (1995) - The Use of the Benchmark Dose Approach in Health Risk Assessment, Risk Assessment Forum, EPA/630/R-94/007, United States Environmental Protection Agency, Washington, DC. Principal authors: K. Crump, B. Allen, E. Faustman.

Weller E.A. and Ryan L.M. (1998) - Testing for trend with count data, Biometrics 54, 762-773.

Westfall P.H. (1999) - A course in multiple comparisons and multiple tests, Texas Tech University

Westfall, P.H., Tobias R.D., Rom D., Wolfinger R.D. and Hochberg Y. (1999) - Multiple comparisons and multiple tests, SAS Institute, Cary, North Carolina.

Westfall P.H. and Young S.S. (1993) - Resampling-Based Multiple Testing, Wiley, New York.

Wilcox R.R. (1991) - Non-parametric analysis of covariance based on predicted medians, British Journal of Mathematical and Statistical Psychology 44, 221-230.

Williams D.A. (1971) - A test for differences between treatment means when several dose levels are compared with a zero dose control, Biometrics 27, 103-117

Williams D.A. (1972) - The comparison of several dose levels with a zero dose control, Biometrics 28, 519-531.

Williams D. A. (1975) - The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratotlogy, Biometrics 31, 949-952.

Williams D.A. (1977) - Some inference procedures for monotonically ordered normal means, Biometrika 64, 9-14.

Wolfe R.A., Roi L.D. and Margosches E.H. (1981) - Monotone dichotomous regression estimates, Biometrics 37, 157-167.

Wright S.P. (1992) - Adjusted P-values for Simultaneous Inference, Biometrics 48, 1005-1013.

References for chapter 6

Akaike, H. (1974) - A New Look at the Statistical Model Identification. IEEE Transaction on Automatic Control, AC -19, 716 -723.

Bailer A.J. and Oris J.T. (1997). – Estimating inhibiting concentrations for different response scales using generalized linear models. Environ. Toxicol. Chem. 16: 1554-1559.

Box G.E.P. and Tidwell P.W. (1962) - Transformations of the independent variables, Technometrics 4, 531-550.

Bozdogan, H. (1987) - Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. Psychometrika, 52, 345 -370.

Brain P. and Cousens R. (1989). – An equation to describe dose responses where there is stimulation of growth at low doses. Weed res. 29: 93-96.

Bruce R.D. and Versteeg D.J. (1992). – A statistical procedure for modeling continuous toxicity data. Environ. Toxicol. Chem. 11: 1485-1494.

Cox D.R. and Oakes D. (1984). – Analysis of survival data. London: Chapman and Hall.

Crump K.S., Hoel D.G., Langley C.H. and Peto, R. (1976) - Fundamental carcinogenic processes and their implications to low dose risk assessment. Cancer Research 36, 2973-2979.

Crump K.S. (1984) - "A new method for determining allowable daily intakes", Fundamental and Applied Toxicology, 4, 845-871.

Crump K.S. (1995) - Calculation of Benchmark Doses from Continuous Data. Risk Anal 15, 79-89.

Efron B. (1987). - Better bootstrap confidence intervals. J Am Stat Assoc 82: 171-200.

Efron B. and Tibshirani R. (1993). – An introduction to the bootstrap. Chapman and Hall, London, UK.

Fieller E.C. (1954). - Some problems in interval estimation. J R Stat Soc B 16, 175-185.

Gaylor D.W., and Slikker W. (1990) - Risk assessment for neurotoxic effects. NeuroToxicology 11, 211-218.

Hoekstra J.A. (1993) - Statistics in Ecotoxicology. PhD thesis, Free University Amsterdam.

Miller R.G. (1981) – Survival Analysis. New York: WIley.

Moerbeek M., Piersma A.H. and Slob W. (2004) – A comparison of three methods for calculating confidence intervals for the benchmark approach. Risk Anal. 24: 31 – 40.

Scholze M., Boedeker W., Fuast M., Backhaus T., Altenburge R. and Grimme L.H. (2001) - A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. Environ. Toxicol. Chem. 20: 448-457

Teunis P.F.M. and Slob W. (1999) - The Statistical Analysis of Fractions Resulting From Microbial Counts. Quantitative Microbiology , 1: 63-88

Slob W. and Pieters M.N. (1998) - A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: general framework. Risk Analysis 18: 787-798.

Slob W. (2003. PROAST) - a general software tool for dose-response modelling. RIVM, Bilthoven.

Slob W. (2002) - Dose-response modelling of continuous endpoints. Toxicol. Sci., 66, 298-312

Tableman M. and Kim J.S. (2004). – Survival analysis using S. London: Chapman and Hall.

<u>References for chapter 7</u>

Andersen H.R., Wollenberger L, Halling-Sorensen B. and Kusk K.O. (2001) - . Development of copepod nauplii to copepodites - A parameter for chronic toxicity including endocrine disruption. Environ. Toxicol. Chem. 20: 2821 – 2829

Andersen J.S., Bedaux J.J.M., Kooijman S.A.L.M. and Holst H. (2000) - The influence of design parameters on statistical inference in non-linear estimation; a simulation study. *Journal of Agricultural, Biological and Environmental Statistics* 5: 28 - 48

Anonymous (1999) - Guidance document on application and interpretation of single-species tests in environmental toxicology. Report EPS 1/RM/34, Minister of Public Works and Government Services, ISBN 0-660-16907-X

Bedaux J.J.M. and Kooijman S.A.L.M. (1994) - Statistical analysis of bioassays, based on hazard modelling. *Environ. & Ecol. Stat.* 1: 303 - 314

Chen C.W. and Selleck R.E. (1969) - A kinetic model of fish toxicity threshold. Res. J. Water Pollut. Control Feder. 41: 294 – 308.

Cox, D. R. and Oakes, D. (1984) - Analysis of survival data. Chapman & Hall, London.

Diamond, S.A., Newman, M. C., Mulvey, M. and Guttman, S.I. (1991)- Allozyme genotype and time to death of mosquitofish, Gambusia holbrooki, during acute inorganic mercury exposure: a comparison of populations. *Aqua. Toxicol.* 21: 119-134.

Dixon, P. M. and Newman, M. C. (1991) – Analyzing toxicity data using statistical models for time-to-death. In: Newman, M. C. and McIntosh, A. W. (eds) An introduction, in metal ecotoxicology, concepts and applications. Lewis Publishers, Chelsea, MI.

Elsasser W.M. (1998) - Reflections on a theory of organisms; Holism in biology. John Hopkins University Press, Baltimore.

Gerritsen A. (1997) - The influence of body size, life stage, and sex on the toxicity of alkylphenols to Daphnia magna. PhD-thesis, University of Utrecht

Garric J, Migeon B. and Vindimian E. (1990) - Lethal effects of draining on brown trout. A predictive model based on field and laboratory studies. Water Res. 24: 59-65

Hallam T.G., Lassiter R.R. and Kooijman S.A.L.M. (1989) - Effects of toxicants on aquatic populations. In: Levin, S. A., Hallam, T. G. and Gross, L. F. (Eds), *Mathematical Ecology.* Springer, London: 352 – 382

Harding G.C.H. and Vass W.P. (1979) - Uptake from seawater and clearance of p,p'-DDT by marine planktonic crustacea. J. Fish. Res. Board Can. 36: 247 – 254.

Heugens, E. H. W., Hendriks, A. J., Dekker, T., Straalen, N. M. van and Admiraal, W. (2001) - A review of the effects of multiple stressors on aquatic organisms and analysis of uncertainty factors of use in risk assessment. *Crit. Rev Toxicol.* 31: 247-284

Heugens, E. H. W., Jager, T., Creyghton, R., Kraak, M. H. S., Hendriks, A. J., Straalen, N. M. van and Admiraal. W. (2003) - Temperature-dependent effects of cadmium on *Daphnia magna*: accumulation versus sensitivity. *Environ. Sci. Tehnol* 37: 2145-2151.

Hill H.V. (1910) - The possible effects of aggregation on the molecules of haemoglobin on its dissociation curves. J. Physiol. (London) 40: IV-VII

Hoeven N. van der, Kooijman S.A.L.M. and Raat W.K. de (1990) - Salmonella test: relation between mutagenicity and number of revertant colonies. *Mutation Res.* 234: 289 – 302.

Jager, T. (2003) - Worming your way into bioavailability; modelling the uptake of organic chemicals in earthworms. PhD thesis, Utrecht University.

Jager, T., Baerselman, R., Dijkman, E., Groot, A. de, Hogendoorn, E., Jong, A. de, Kruitbosch, J. and Peijnenburg, W. (2003) - Availability of polycyclic aromatic hydrocarbons to earthworms (*Eisenia Andrei, Oligochaeta*) in field-polluted soils and soil-sediment mixtures. *Environ. Tox Chem*. 22: 767-775.

Janssen, M. P. M., Bruins, A., Vries, T. H. de and Straalen, N. M. van (1991) Comparison of Cadmium kinetics in 4 soil arthropod species. *Arch. Environ. Cont. Toxicol* 20: 305-313

Kimerle R.A., Macek K.J., Sleight B.H. III and Burrows M.E. (1981) - . Bioconcentration of linear alkylbenzene sulfonate (LAS) in bluegill (Lepoma macrochirus). Water Res. 15: 251 – 256.

Klepper O. and Bedaux J.J.M. (1997) - Nonlinear parameter estimation for toxicological threshold models. *Ecol. Mod.* 102: 315 - 324

Klepper O. and Bedaux J.J.M. (1997a) - A robust method for nonlinear parameter estimation illustrated on a toxicological model *Nonlin. Analysis* 30: 1677 – 1686

Klok C. and de Roos A. M. (1996) - Population level consequences of toxicological influences on individual growth and reproduction of Lumbricus rubellus (Lumbricidae, Oligochaeta). Ecotoxicol. Environm. Saf. 33: 118-127

Könemann W.H. (1980) - Quantitative structure-activity relationships for kinetics and toxicity of aquatic pollutants and their mixtures for fish. PhD thesis, Utrecht University, the Netherlands.

Kooijman S.A.L.M. (1981) - Parametric analyses of mortality rates in bioassays. *Water Res.* 15: 107 - 119

Kooijman S.A.L.M. (1983) - Statistical aspects of the determination of mortality rates in bioassays. *Water Res.*: 17: 749 - 759

Kooijman S.A.L.M. (1985) - Toxicity at population level. In: Cairns, J. (ed) *Multispecies toxicity*, Pergamon Press, N.Y.: 143 - 164

Kooijman S.A.L.M. (1988) - Strategies in ecotoxicological research. *Environ. Aspects Appl. Biol.* 17 (1): 11 - 17

Kooijman S.A.L.M. (1997) - Process-oriented descriptions of toxic effects. In: Schüürmann, G. and Markert, B. (Eds) *Ecotoxicology.* Spektrum Akademischer Verlag, 483 - 519

Kooijman S.A.L.M. (2000) - Dynamic Energy and Mass Budgets in Biological Systems. Cambridge University Press

Kooijman S.A.L.M. (2001) - Quantitative aspects of metabolic organisation; a discussion of concepts. *Phil. Trans. R. Soc. B*, 356: 331 – 349

Kooijman S.A.L.M. and Bedaux J.J.M. (1996) - Some statistical properties of estimates of no-effects levels. *Water Res.* 30: 1724 - 1728

Kooijman S.A.L.M. and Bedaux J.J.M. (1996) - Analysis of toxicity tests on fish growth. Water Res. 30: 1633 - 1644

Kooijman S.A.L.M. and Bedaux J.J.M. (1996a) - Some statistical properties of estimates of no-effects concentrations. *Water Res.* 30: 1724 - 1728

Kooijman S.A.L.M. and Bedaux J.J.M. (1996b) - Analysis of toxicity tests in Daphnia survival and reproduction. *Water Res.* 30: 1711 - 1723

Kooijman S.A.L.M. and Bedaux J.J.M. (1996c) - The analysis of aquatic toxicity data. VU University Press, Amsterdam

Kooijman S.A.L.M., Bedaux J.J.M. and Slob W. (1996) - No-Effect Concentration as a basis for ecological risk assessment. *Risk Analysis* 16: 445 - 447

Kooijman S.A.L.M., Bedaux J.J.M., Gerritsen A.A.M., Oldersma H. and Hanstveit A.O. (1998) - Dynamic measures for ecotoxicity. Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data *OECD Environmental Health and Safety Publications* 10: 64-90. Also in: Newman, M.C. and Strojan, C., *Risk Assessment: Logic and Measurement.* Ann Arbor Press, 187 - 224

Kooijman S.A.L.M., Hanstveit A.O. and Nyholm N. (1996a) - No-effect concentrations in algal growth inhibition tests. *Water Res.* 30: 1625 - 1632

Kooijman S.A.L.M., Hanstveit A.O. and Oldersma H. (1983) - Parametric analyses of population growth in bioassays. *Water Res.* 17: 727 - 738

Kooijman S.A.L.M. and Haren R.J.F. van (1990) - Animal energy budgets affect the kinetics of xenobiotics. *Chemosphere* 21: 681 - 693

Lau C., Andersen M.E., Crawford-Brown D.J., Kavlock R.J., Kimmel C.A., Knudsen T.B., Muneoka K., Rogers J. M., Setzer R. W., Smith G. and Tyl R. (2000) -. Evaluation of biologically based dose-response modelling for developmental toxicology: A workshop report. Regul. Toxicol. Pharmacol. 31: 190 - 199

Leeuwen I.M.M. and Zonneveld C. (2001) - From exposure to effect: a comparison of modelling approaches to chemical carcinogenesis. *Mutation Res*. 489: 17 – 45

Legierse, K. C. H. M., Sijm, D. T. H. M. (1998) Bioconcentration kinetics of chlorobenzenes and the organophosphorus pesticide chlorthion in the pond snail *Lymnaea stagnalis* – a comparison with the guppy *Poecilia reticulata. Aquat. Toxicol.* 41: 3001-323

Liebig J. von (1840) - Chemistry in its application to agriculture and physiology. Taylor and Walton, London.

Miller, R. G. (1981) - Survival Analysis. Wiley, New York.

Morgan B.J.T. (1992) - Analysis of quantal response data. Monographs on Statistics and Applied Probability 46. Chapman & Hall, London.

McCullagh P. and Nelder J.A. (1989) - Generalised linear models. Monographs on Statistics and Applied Probability 37. Chapman & Hall, London.

McLeese D.W., Zitko V. and Sergeant D.B. (1979) - Uptake and excretion of fenitrothion by clams and mussels. Bull. Environ. Contam. Toxicol. 22: 800 - 806

Muller E.B. and Nisbet R.M. (1997) - Modelling the Effect of Toxicants on the Parameters of Dynamic Energy Budget Models. In: Dwyer, F. J., Doane, T. R. and Hinman, M. L. (Eds.): Environmental Toxicology and Risk Assessment: Modelling and Risk Assessment (Sixth Volume), p 71 - 81, American Society for Testing and Materials

Newman M.C. (1995) - Quantitative methods in aquatic ecotoxicology. Lewis Publ, Boca Raton.

Newman, M. C., Diamond, S. A., Mulvey, M. and Dixon, P. (1989) – Allozyme genotype and time to death of mosquitofish*, Gambusia affinis* (Baird and Girard) during acute toxicant exposure: A comparison of arsenate and inorganic mercury. *Aqua. Toxicol.* 21: 141-156.

Nisbet R.M., Muller E.B., Lika K. and Kooijman S.A.L.M. (2000) - From molecules to ecosystems through Dynamic Energy Budget models. *J. Anim. Ecol.* 69: 913 - 926

Nyholm N. (1985) - Response variable in algal growth inhibition tests – Biomass or growth rate? Water Res. 19: 273 – 279.

Péry A.R.R., Bedaux J.J.M., Zonneveld C. and Kooijman S.A.L.M. (2001) . Analysis of bioassays with time-varying concentrations. *Water Res.*, 35: 3825 - 3832

Péry A.R.R., Flammarion P., Vollat B., Bedaux J.J.M., Kooijman S.A.L.M. and Garric J. (2002) - Using a biology-based model (DEBtox) to analyse bioassays in ecotoxicology: Opportunities & recommendations. *Environ. Toxicol. & Chem.*, 21 (2): 459–465

Purchase I.F.H. and Auton T.R. (1995) - Thresholds in chemical carcinogenesis. Regul. Toxicol. Pharmacol. 22: 199 - 205

Raat K. de, Kooijman S.A.L.M. and Gielen J.W.J. (1987) - Concentrations of polycyclic hydrocarbons in airborne particles in the Netherlands and their correlation with mutagenicity. *Sci. Total Environ.* 66: 95 - 114

Raat K. de, Meyere F.A. de and Kooijman S.A.L.M. (1985) - Mutagenicity of ambient aerosol collected in an urban and industrial area of the Netherlands. *Sci. Total Environ.* 44: 17 – 33

Reindert, K. H., Giddings, J. M. and Judd, L. (2002) - Effects analysis of time-varying or repeated exposures in aquatic ecological risk assessment of agrochemicals. *Environ. Tox. Chem.,* 21: 1977-1992

Segel I.W. (1976) - Biochemical calculations; how to solve mathematical problems in general biochemistry. J. Wiley & Sons, New York.

Setzer R.W., Lau C., Mole M.L., Copeland M.F., Rogers J.M. and Kavlock R.J. (2001) - Toward a biologically based dose-response model for developmental toxicity of 5-fluorouracil in the rat: A mathematical construct. Toxicol. Sci. 59: 49 – 58

Sibly, R. M. and Calow, P. (1989) - A life-cycle theory of response to stress. *Biol. J. Linn. Soc.,* 37: 101-116

Spacie A. and Hamelink J.L. (1979) - Dynamics of trifluralin accumulation in river fish. Environ. Sci. Technol. 13: 817 – 822

Sprague J.B. (1995) - Factors that modify toxicity. In: Rand, G. M. (ed.) Fundamentals of aquatic toxicology. Taylor & Francis, Washington, p 1012 – 1051

Stumm W. and Morgan J.J. (1996) - Aquatic chemistry. J. Wiley & Sons, New York.

Vindimian E., Rabout C. and Fillion G. (1983) - A method of co-operative and non co-operative binding studies using non linear regression analysis on a microcomputer. J. Appl. Biochem. 5: 261-268

Widianarko B. and Straalen N. van (1996) - Toxicokinetics-based survival analysis in bioassays using nonpersistent chemicals. Environ. Toxicol. Chem. 15: 402 – 406

Wong P.T.S., Chau Y.K., Kramar O. and Bengert G.A. (1981) - Accumulation and depuration of tetramythyllead by rainbow trout. Water Res. 15: 621 – 625.