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

10-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/ANN
Unclassified

OECD SERIES ON TESTING AND ASSESSMENT
Number 54

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

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

JT03208635

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

ANNEXES

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris, April 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; revised 2006)*

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. 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.

**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

Contents

Page

ANNEX 1: ANALYSIS OF AN "ACUTE IMMOBILISATION OF *DAPHNIA MAGNA*" DATA SET (OECD GL 202 – ISO 6341) USING THE THREE PRESENTED APPROACHES 11
1.1 Examples of data analysis using hypothesis testing (NOEC determination) 12
1.2 Example of data analysis by dose response modelling 17
1.3 Example of data analysis using Debtox (biological methods) 22

ANNEX 2: ANALYSIS OF AN "ALGAE GROWTH INHIBITION" DATA SET USING THE THREE PRESENTED APPROACHES 23
Examples of data analysis using hypothesis testing (NOEC determination) 25
2.1 Example of data analysis by dose response modelling 34
2.2 Examples of data analysis using Debtox (biological methods) 40

ANNEX 3: ANALYSIS OF AN "*DAPHNIA MAGNA* REPRODUCTION" DATA SET (OECD GL 211 – ISO 10706) USING THE THREE PRESENTED APPROACHES 43
3.1 Examples of data analysis using hypothesis testing (NOEC determination) 45
3.2 Example of data analysis by dose response modelling 51
3.3 Examples of data analysis using Debtox (biological methods) 59

ANNEX 4: ANALYSIS OF A "FISH GROWTH" DATA SET (OECD GL 204/215 – ISO 10229) USING THE THREE PRESENTED APPROACHES 63
4.1 Examples of data analysis using hypothesis testing (NOEC determination) 66
4.2 Example of data analysis by dose response modelling 80
4.3 Examples of data analysis using Debtox (biological methods) 86

ANNEX 5 88
5.1 Description of Selected Methods for Use with Quantal Data 88
5.2 Power of the Cochran-Armitage Test 96
5.3 Description of Selected Tests for Use With Continuous Data 105
5.4 Power of Step-Down Jonckheere-Terpstra Test 122

ANNEX 6: ANNEXE TO CHAPTER 7 "BIOLOGY-BASED METHODS" 135

ANNEX 1: ANALYSIS OF AN "ACUTE IMMOBILISATION OF *DAPHNIA MAGNA*" DATA SET (OECD GL 202 – ISO 6341) USING THE THREE PRESENTED APPROACHES

Data set

Concentration	Number of immobilisation 0h	Number of immobilisation 24h	Number of immobilisation 48h
0	0	0	0
0.39	0	0	0
0.78	0	0	0
1.56	0	0	1
3.13	0	0	7
6.25	0	6	11
12.5	0	5	14
25	0	5	20
50	0	19	20
100	0	20	20

1.1 Examples of data analysis using hypothesis testing (NOEC determination)

Example 1a. Daphnid Acute Example: Immobility After 48 Hours Exposure
NOEC Determination by Two Methods

NOEC is 1.56 mg/L by both tests shown in the flow-chart, Figure 5.1. In this example, there was only a water control. Both Fisher's Exact test with a Bonferroni-Holm correction and the step-down Cochran-Armitage test are done to illustrate both approaches. In both cases, the overall false positive rate is controlled at 0.05 and 1-sided tests are done comparing proportion alive in each concentration to that in the control. The p-values reported for Fisher's Exact test do not include the Bonferroni-Holm correction, but the determination of significance makes that adjustment.

STATISTICAL ANALYSIS LIST FILE
DAPHNID LIVE48 OBSERVED DATA FROM DATASET ADAP_ACUTE
GROUP STATISTICS BY DOSE

dose	Number at risk	Number Responding	% Responding	Dose Score	Test Dose (mg/L)
1	20	20	100	1	0
2	20	20	100	2	0.39
3	20	20	100	3	0.78
4	20	19	95	4	1.56
5	20	13	65	5	3.13
6	20	9	45	6	6.25
7	20	6	30	7	12.5
8	20	0	0	8	25
9	20	0	0	9	50
10	20	0	0	10	100

FISHER EXACT TEST vs CONTROL FOR DAPHNID LIVE48 OBSERVED
TESTING FOR A DECREASING ALTERNATIVE HYPOTHESIS

Test Dose (mg/L)	Number at risk	Number Responding	Fisher's Exact Test P-value (Left)	Significance Rating
0.39	20	20	1.00000	
0.78	20	20	1.00000	
1.56	20	19	0.50000	
3.13	20	13	0.00416	*
6.25	20	9	0.00007	**
12.5	20	6	0.00000	**
25	20	0	0.00000	**
50	20	0	0.00000	**
100	20	0	0.00000	**

* referes to 0.01<p-value<0.05
** refer to a p-value <0.01.

Cochran-Armitage Test
Using Equally Spaced Dose Scores
Cochran-Armitage test is one-sided for DECREASE in RESPONSE
All doses included

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	141.11145	9	0	
Trend	-11.4263	1	1.545E-30	**
LOF	10.551094	8	0.2284543	

100 mg/L concentration omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	119.23185	8	0	
Trend	-10.46522	1	6.239E-26	**
LOF	9.7109205	7	0.2055557	

100 & 50 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	93.867043	7	0	
Trend	-9.126828	1	3.527E-20	**
LOF	10.568053	6	0.1026794	

100, 50 & 25 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	58.680261	6	8.341E-11	
Trend	-7.068764	1	7.816E-13	**
LOF	8.7128292	5	0.1210814	

100, 50, 25 & 12.5 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	41.584158	5	7.1493E-8	
Trend	-5.637291	1	8.6373E-9	**
LOF	9.8051068	4	0.0438418	

100, 50, 25, 12.5 & 6.25 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	25.271739	4	0.0000444	
Trend	-3.909645	1	0.0000462	**
LOF	9.986413	3	0.018682	

100, 50, 25, 12.5, 6.25 & 3.13 mg/L
concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	3.0379747	3	0.3858072	
Trend	-1.350105	1	0.0884911	
LOF	1.2151899	2	0.5446592	

As discussed in Figure 5.1 and accompanying text, The Cochran-Armitage test is first applied to the entire data set. Given that that test is significant at the 0.05 level (p-value for trend is less than 0.05), the high concentration is omitted and the test is repeated with the remaining concentrations. This procedure is repeated until the Cochran-Armitage test is first not significant. The highest concentration remaining at that step is the NOEC. In the present case, that occurs when the highest remaining concentration is 1.56 mg/L, which is thus the NOEC.

Two other terms are shown at each stage. The first is labeled 'Overall" and is the standard chi-square test with k-1 degrees of freedom for differences among the k groups represented. It will be recalled that this chi-square statistic is the sum of the 1-df chi-square statistic measuring linear trend (i.e., the square of the Cochran-Armitage test statistic) and the k-2 df chi-square statistic for depature from linearity here labelled LOF.

Example 1b. Daphnid Acute Example: Immobility After 24 Hours Exposure
NOEC Determination by Two Methods

NOEC is 3.13 mg/L by the step-down Cochran-Armitage test and 25 mg/L by Fisher's exact test. Both follow the flow-chart, Figure 5.1. In this example, there was only a water control. Both Fisher's Exact test with a Bonferroni-Holm correction and the step-down Cochran-Armitage test are done to illustrate both approaches. In both cases, the overall false positive rate is controlled at 0.05 and 1-sided tests are done comparing proportion alive in each concentration to that in the control. The p-values reported for Fisher's Exact test do not include the Bonferroni-Holm correction, but the determination of significance makes that adjustment

STATISTICAL ANALYSIS LIST FILE
DAPHNID LIVE24 OBSERVED DATA FROM DATASET ADAP_ACUTE
GROUP STATISTICS BY DOSE

dose	Number at risk	Number Responding	% Responding	Dose Score	Test Dose (mg/L)
1	20	20	100	1	0
2	20	20	100	2	0.39
3	20	20	100	3	0.78
4	20	20	100	4	1.56
5	20	20	100	5	3.13
6	20	14	70	6	6.25
7	20	15	75	7	12.5
8	20	15	75	8	25
9	20	1	5	9	50
10	20	0	0	10	100

FISHER EXACT TEST vs CONTROL FOR DAPHNID LIVE24 OBSERVED
TESTING FOR A DECREASING ALTERNATIVE HYPOTHESIS

Test Dose (mg/L)	Number at risk	Number Responding	Fisher's Exact Test P-value (Left)	Significance Rating
0.39	20	20	1.00000	
0.78	20	20	1.00000	
1.56	20	20	1.00000	
3.13	20	20	1.00000	
6.25	20	14	0.01010	
12.5	20	15	0.02356	
25	20	15	0.02356	
50	20	1	0.00000	**
100	20	0	0.00000	**

Cochran-Armitage Test
Using Equally Spaced Dose Scores
Cochran-Armitage test is one-sided for DECREASE in RESPONSE

All doses included

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	136.55172	9	0	
Trend	-9.896625	1	2.153E-23	**
LOF	38.60853	8	5.8093E-6	

100 mg/L concentration omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	99.239409	8	0	
Trend	-7.804546	1	2.986E-15	**
LOF	38.328473	7	2.6241E-6	

100 & 50 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	30	7	0.000095	
Trend	-4.485426	1	3.6384E-6	**
LOF	9.8809524	6	0.1297557	

100, 50 & 25 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	30.190275	6	0.0000362	
Trend	-4.240398	1	0.0000112	**
LOF	12.209302	5	0.0320297	

100, 50, 25 & 12.5 mg/L concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	31.578947	5	7.1979E-6	
Trend	-3.678836	1	0.0001172	**
LOF	18.045113	4	0.0012093	

When all concentrations at or above 6.25 mg/L are omitted, there is 100% survival at every remaining concentration and no test can be done, nor is a test needed. The NOEC is 3.13 mg/L.

As discussed in Figure 5.1 and accompanying text, The Cochran-Armitage test is first applied to the entire data set. Given that that test is significant at the 0.05 level (p-value for trend is less than 0.05), the high concentration is omitted and the test is repeated with the remaining concentrations. This procedure is repeated until the Cochran-Armitage test is first not significant. The highest concentration remaining at that step is the NOEC. In the present case, that occurs when the highest remaining concentration is 1.56 mg/L, which is thus the NOEC.

Two other terms are shown at each stage. The first is labeled ‘Overall” and is the standard chi-square test with k-1 degrees of freedom for differences among the k groups represented. It will be recalled that this chi-square statistic is the sum of the 1-df chi-square statistic measuring linear trend (i.e., the square of the Cochran-Armitage test statistic) and the k-2 df chi-square statistic for depature from linearity here labelled LOF.

1.2 Example of data analysis by dose response modelling

Fitting a dose-response model to data will normally be done with the aid of computer software. Therefore, the user does not need to worry about computational details. Outcomes from different software packages should be nearly identical when fitting the same model to the same data based on the same assumptions and methods. However, software packages may differ extensively in how to run them. In this section, the discussion of the examples is based on an analysis using the software package PROAST, but it is attempted to make the discussion helpful for users of other software just as well.

Bioassay on acute mortality of Daphnia magna;

It is assumed that both an EC50 and an EC10 is required after two days of exposure. Therefore, a dose-response analysis of the mortality data at day 2 will normally be sufficient. The following section discusses how such may be done.

Dose-response analysis for mortality at day 2.

The data are quantal, and include more than one partial response. According to the flow chart in chapter 6 (Fig. 6.0), various models should be fitted in a way as described in 6.2.1. The first step is to put the data in the format that is required by the software to be used. Typically, the data must be provided in a matrix form, as illustrated in Table 1. Note that the information of the sample size in each concentration (dose) group is essential. Here, the model will be fitted based on maximization of the log-likelihood assuming a binomial distribution for the observed responses.

Table 1. Data input file (as required by PROAST).

Daphnia_magna
4
1 4 0 0

Dose	Response	Sample_size	Day
100	20	20	1
50	19	20	1
25	5	20	1
12.5	5	20	1
6.25	6	20	1
3.13	0	20	1
1.56	0	20	1
0.78	0	20	1
0.39	0	20	1
0	0	20	1
100	20	20	2
50	20	20	2
25	20	20	2
12.5	14	20	2
6.25	11	20	2
3.13	7	20	2
1.56	1	20	2
0.78	0	20	2
0.39	0	20	2
0	0	20	2

Fitting the probit model

First, a probit model is fitted. As discussed in chapter 6, the background response should normally be estimated as a model parameter. However, when, as in this particular data set, fitting the model with or without a background parameter results in virtually the same outcome; the background parameter can just

as well be omitted from the model (or, equivalently, be fixed at zero). It is preferable, in this case, to avoid non-convergence of the (iterative) fit algorithm. Note that in fitting the model without a background parameter, the data in the controls are not used, and they could just as well be deleted from the data set.
Fig. 1 shows the fit of the probit model for this data set. The parameter *b* is the EC50, and the parameter *c* is the slope. It may also be interpreted as the inverse of the standard deviation (σ) of the underlying tolerance distribution (on the $\log_{10}$–scale).
Visual inspection of Fig. 1 indicates that the fitted model appears quite well supported by the data, and there is no reason to believe that the dose-response relationship could in fact be much different. To substantiate this, various other models will be fitted, in concordance with the flow chart in chapter 6.

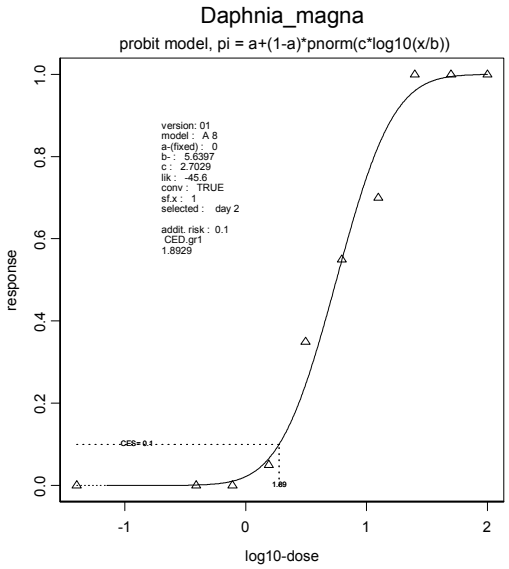


Fig. 1. Probit model fitted to mortality response at day 2. CED = EC10.

Fitting various models

Next to the probit model, three other models were fitted: the logit, the Weibull and the (two-stage) LMS model. As Table 2 shows, the resulting EC50 and EC10 estimates are reasonably similar (less then 30% difference). Also note that the EC10 is obtained by interpolation. Thus, it may be concluded that the data are suitable for deriving an EC10 by dose-response modeling.

Table 2. Summary of results regarding four models fitted to the mortality data in Fig. 1. CI = confidence interval.

Model	Log-lik	EC50		EC10	
		MLE	90%-CI*	MLE	90%-C.I*
Probit	-45.60	5.64	4.53 - 6.93	1.89	1.42 - 2.68
Logit	-46.46	5.63	4.59 - 6.94	1.90	1.36 - 2.72
Weibull	-46.25	6.64	5.36 - 8.17	1.69	1.10 - 2.68
Two-stage	-46.83	7.00	5.59 - 8.46	1.47	1.17 - 1.77

* based on 1000 bootstrap runs

Confidence intervals
The confidence intervals can be calculated by one of the ways described in chapter 6, depending on the software available. In Table 1 the 90%-confidence intervals are of the EC50 and the EC10 are given, as obtained by the bootstrap method (1000 runs). They are not dramatically different between the models, and

one may choose the lowest of these. Alternatively, one may choose the lower bounds associated with the probit model, as this model gives a somewhat higher log-likelihood value than the others. The slightly better fit can also be seen by comparing Fig. 1 with Fig. 2.

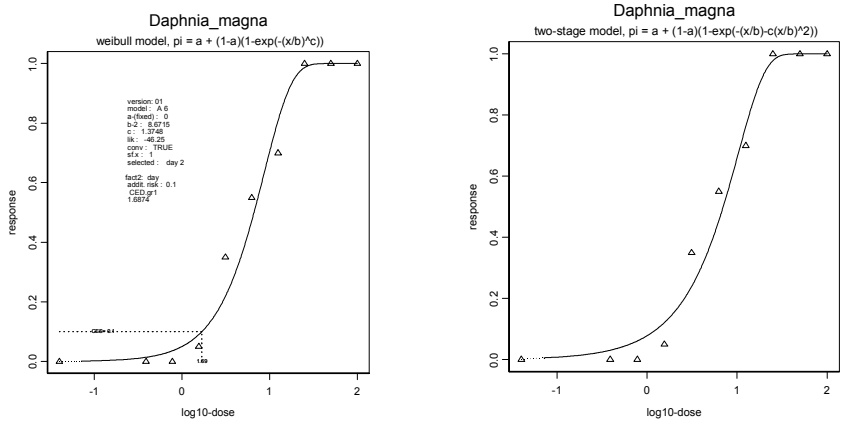


Fig. 2. The Weibull (left panel) and the two-stage LMS model fitted to the mortality data at day 2.

Mortality at both days

Although a single analysis of mortality at day 2 will normally be sufficient, some remarks are made here on using the data at day 1 as well.
First, it may be noted that a similar analysis can be done for day 1 separately (see Fig. 3, left panel). Obviously, the ECx values are higher, and for a risk assessment these results will probably not be relevant. However, from the perspective of an efficient use of the data, an analysis is which the data from both days are analyzed simultaneously would be preferable.

There are two ways of doing a simultaneous analysis. One is to fit the dose-response model to both days simultaneously, allowing for the fact that one (or some) of the parameters in the model depends on the number of days. This analysis is shown in the right panel of Fig. 3. Here, the parameter *c* (the slope) is assumed to be the same for both days, while the EC50 is allowed to differ between days. Thus, the simultaneous analysis estimated three parameters, while the two separate analyses together estimated four parameters in total. The sum of the log-likelihoods of the two separate analyses (-94.91) can be compared to the log-likelihood of the simultaneous analysis (-94.75). The former log-likelihood is associated with one more free parameter estimated from the data, and therefore can only be larger (or equal) to the latter log-likelihood. (Note: this holds in general for nested models, but not necessarily for non-nested models; two models are nested when one model can be derived from the other by reducing the number of parameters to be estimated from the data). The likelihood ratio test can be used to assess if the larger number of parameters in the separate analyses is associated with in a significantly better fit. According to this test the increase in log-likelihood should be at least 1.92 to be significant at $\alpha = 0.05$. It may therefore be concluded that the mortality data can just as well be described by a probit model with the same slope for days 1 and 2. As table 3 shows, the estimated ECx values are quite similar between a simultaneous and a separate analysis.

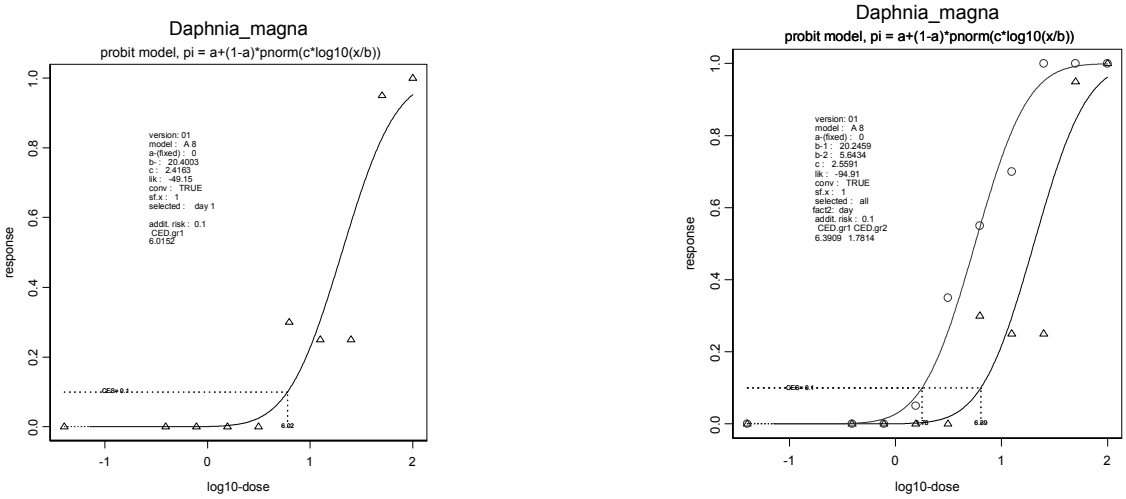


Fig. 3. Probit model fitted to mortality data on day 1 (left panel), and fitted to both day 1 and day 2 simultaneously.

Table 3. Results of fitting the probit model separately or simultaneously to day 1 and day 2.
The higher log-likelihood for the two separate analyses is not significantly higher than that for the simultaneous analysis.

	Day 1		Day 2			
	EC50	EC10	EC50	EC10	Log-lik	# parameters estimated
Separate analysis	20.40	6.02	5.64	1.89	-94.91 (n.s.)	4
Simultaneous analysis	20.25	6.39	5.64	1.78	-94.75	3

The gain of a simultaneous analysis is twofold. First, one may expect that the estimated values of the ECx are less likely to be biased (e.g. due to outliers or incidental systematic errors in the data), simply because they are based on more data points. Second, the data are described by less parameters, which complies with the parsimony principle. One of the expected consequences is that the confidence intervals are smaller. However, the analysis as now being discussed is not completely sound from a statistical point of view. The reason is that the data points for day 1 and day 2 are not independent, and therefore the confidence intervals may not be completely reliable. Nonetheless, they have been assessed for the EC50 and EC10 at day 2 (see Table 4). They are wider than those obtained from a separate analysis of day 2, which is against expectation, since a fewer number of parameters was estimated in this analysis. The reason is that in this particular case, the data at day 1 appear to show more noise than the data at day 2. Therefore, the noise is increased by including day 1 in the analysis, compared to an analysis of day 2 only. As long as there is no clear reason why the noise at day 1 may be larger than at day 2, it is hard to say if the noise for day 2 is in fact smaller, or only apparent or incidental. Thus, the increase in noise by including day 1 might be a more realistic reflection of the overall noise in the data.

Table 4. EC 50 and EC10 at day 2, with 90 confidence intervals, assessed by a simultaneous analysis of both days (see right panel of Fig. 3). It should be noted that these confidence intervals may not be completely reliable, due to dependencies in the data between day 1 and day 2.

Model	EC50 – day 2		EC10 – day 2	
	MLE	90-% CI	MLE	90%-CI
Probit	5.64	4.50 - 7.01	1.78	1.32 - 2.39

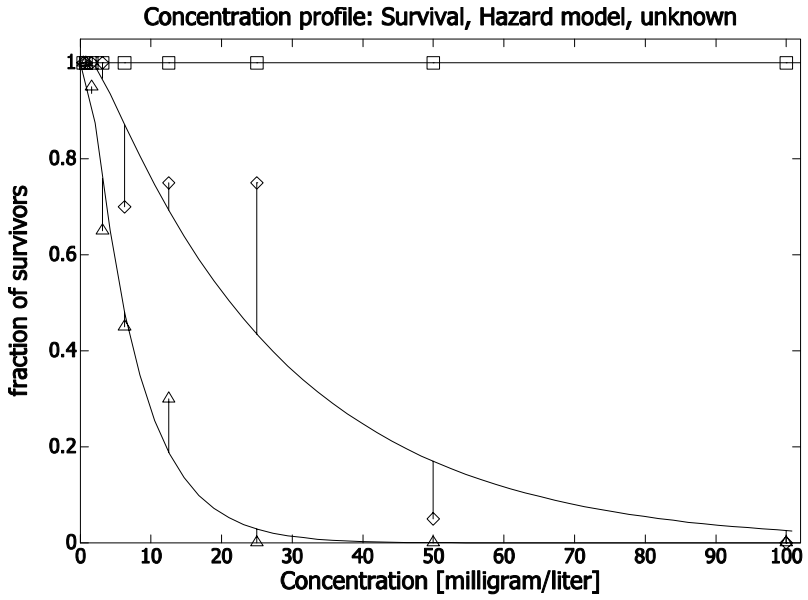
The dependencies in the data between day 1 and day 2 may be avoided by doing a simultaneous analysis the other way around. The mortality data may be regarded as a function of time for each dose group, while one of the parameters in the survival function is a function of dose. For instance, assuming a Weibull survival function, it might be assumed that the median survival time is some function of dose. Another approach is to assume a dose-response analysis for the hazard function (see e.g. Example of data analysis using Debtox (biological methods))

1.3 Example of data analysis using Debtox (biological methods)

Parameters and asymptotic standard deviations (ASD) (For a definition of all parameters, see Annexe 6 to chapter 7)

Survival, Hazard model			ASD	Correlation coefficients	
Blank mortality rate	1e-010	d^{-1}	0.000		
No-effect concentration-time	1.473	mg l^{-1} d	0.344	0.000	
Killing acceleration	0.07524	l mg^{-1} d^{-2}	0.010	0.000	0.359
Deviance	23.29				

Graphical test of model predictions against data



LCx values (derived from parameter values) in mg/l.

Day	LC0	ASD	LC50	ASD
1	1.48	0.344	21.1	2.26
2	0.738	0.172	5.94	0.574

Comments

Slow kinetics appeared to fit the data best, which means that the elimination rate was too small to be estimated reliably. This means that the model loses this parameter. The consequence is that only the killing acceleration (which is the product of the killing rate and the elimination rate) can be estimated, not the killing rate itself. Similarly the ratio of the NEC and the elimination rate, called the no-effect concentration time, could be estimated, rather than the NEC itself; the bioassay did not last long enough for this compound. The 95% confidence interval for the NEC can still be estimated, however, and was found to be (0, 0.95) mg/l, on the basis of the profile likelihood function.
The background morality rate was found to be nil. Notice that, excluding this parameter, a total of two parameters have been fitted on 18 data points.

ANNEX 2: ANALYSIS OF AN "ALGAE GROWTH INHIBITION" DATA SET USING THE THREE PRESENTED APPROACHES

The following data set has been used. The aim of the presented analysis is to show the methodology that can be applied to these kinds of data.

Data set on *Selenastrum capricornutum*

concentration	Day 0	Day 0	Day 0	Day 1	Day 1	Day 1	Day 2	Day 2	Day 2
0	1.767	1.777	1.775	7.889	7.921	7.901	46.46	46.32	46.28
0	1.748	1.762	1.76	8.179	8.178	8.188	44.35	44.31	44.27
0	1.762	1.762	1.758	7.99	7.989	8.001	41.99	41.79	41.84
0	1.784	1.796	1.794	8.322	8.335	8.328	43.2	43.2	43.29
0	1.79	1.789	1.785	8.323	8.343	8.353	46.21	46.16	45.91
0	1.84	1.846	1.85	7.977	7.979	7.981	40.63	40.73	40.75
0.01	1.738	1.741	1.747	8.229	8.23	8.25	40.4	40.27	40.22
0.01	1.989	1.955	1.959	8.376	8.372	8.38	41.97	42.03	42.03
0.02	1.892	1.89	1.892	7.662	7.666	7.653	39.82	39.75	39.79
0.02	1.785	1.789	1.783	7.765	7.773	7.775	38.78	38.84	38.79
0.03	1.779	1.773	1.78	7.737	7.733	7.748	34.79	34.82	34.84
0.03	1.759	1.758	1.763	8.035	8.045	8.039	31.2	31.16	31.23
0.06	1.776	1.775	1.775	7.122	7.129	7.136	25.92	25.91	25.96
0.06	1.754	1.751	1.76	6.775	6.776	6.785	22.42	22.5	22.5
0.1	1.774	1.776	1.77	4.947	4.937	4.941	14.83	14.89	14.83
0.1	1.71	1.716	1.722	5.162	5.165	5.17	13.76	13.71	13.69
0.2	1.746	1.741	1.741	3.593	3.59	3.593	6.389	6.391	6.394
0.2	1.736	1.738	1.739	3.554	3.547	3.555	6.108	6.11	6.127
0.3	1.828	1.822	1.827	2.92	2.925	2.92	3.731	3.708	3.714
0.3	1.688	1.681	1.688	3.157	3.147	3.151	3.13	3.124	3.122
0.6	1.72	1.722	1.72	2.109	2.117	2.122	2.125	2.122	2.099
0.6	1.733	1.735	1.731	2.03	2.04	2.038	2.053	2.066	2.057

According to the July, 2002, Draft Revised TG 201, growth rate can be determined from cell count, biomass, or fluorescence values. That guideline refers to Mayer, P., Cuhel, R. and Nyholm, N. (1997), a simple in vitro fluorescence method for biomass measurements in algal growth inhibition tests, *Water Research* 31: 2525-2531 on the use of fluorescence values for this purpose. Indeed, any of these three measures of mass (cell count, biomass, or fluorescence) can be used to obtain very similar (but not necessarily identical) measures of growth rate on each replicate by fitting the model $y = a * e^{b*time}$, using time in hours, and y the observed measure of mass. The estimate slope, b, from this fit is the sample growth rate for a given replicate. The simplest procedure is to linearize the problem by working with

logarithms to obtain the model log(y) =A + b*time, where A=log(a). In what follows, natural logarithms were used, but this choice does not affect the slope or growth-rate estimate. These growth rate values can then be analyzed by hypothesis testing or regression methods.

Examples of data analysis using hypothesis testing (NOEC determination)

Example 2a. **Atrazine Example Flourescence at Day 2**
NOEC Determination Atrazine by Two Methods

In this example, the total biomass (or its surrogate, Flourescence) is analyzed. In example 2b, growth rate is analyzed. NOEC is 0.01 mg/L by both tests. Both Dunnett's test and the step-down application of the Jonckheere test are illustrated, according to the chart in Figure 5.1

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET AATRAZINE
GROUP STATISTICS FOR Average_2 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	6	43.5278	43.5373	2.26521	0.92477
2	0.01	2	40.9206	40.9206	1.21151	0.85667
3	0.02	2	39.0623	39.0623	0.69532	0.49167
4	0.03	2	32.7739	32.7739	2.55973	1.81000
5	0.06	2	23.9689	23.9689	2.44423	1.72833
6	0.1	2	14.0523	14.0523	0.79903	0.56500
7	0.2	2	6.0204	6.0204	0.19540	0.13817
8	0.3	2	3.1888	3.1888	0.41884	0.29617
9	0.6	2	1.8543	1.8543	0.04007	0.02833

SHAPIRO-WILK TEST OF NORMALITY OF Average_2

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
1.39706	-0.099030	0.10954	0.97324	0.78479	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule is used to identify observations that may be of special interest. Outliers can have an impact on the results, as well as on the assessment of normality and variance homogeneity. A possibly valuable use of these outliers would be to re-run the analysis with outliers omitted to determine whether the NOEC is affected by these outliers. If it is, then caution should be followed in using the results. *It is important to understand that just because an observation is identified as an outlier, that does not mean the observation is "bad" or that it should not be used*. An observation should be excluded from analysis only for scientifically sound reasons and any such exclusion must be clearly stated along with its justification.

Outliers & Influential Observations

SELENASTRUM	Dose	Doseval	Group	OBSER	Pred	Resid
1	1	0	I	46.1206	43.5278	2.59278
5	1	0	I	45.8606	43.5278	2.33278
6	1	0	I	40.4706	43.5278	-3.05722

LEVENE TEST FOR Average_2 - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	8	3.36022	0.025754	**

The data was found to be normally distributed (p-value for SW test was 0.78479) but group variances were unequal (p-value for Levene's test was 0.025754). A Tamhane-Dunnett analysis is appropriate.

Tamhane-Dunnett 1-sided test for decrease in means in Average_2
Using MAXIMUM LIKELIHOOD estimates of variation on ECOTOX values.

dose	MEAN	STDERR	degfree	CONTROL	OBS_DIFF	crit	SIGNIF
2	40.9206	0.85667	3.68718	43.5278	-2.6072	5.3026	
3	39.0623	0.49167	5.87792	43.5278	-4.4656	3.5625	*
4	32.7739	1.81000	1.56884	43.5278	-10.7539	13.7828	
5	23.9689	1.72833	1.62787	43.5278	-19.5589	13.2920	*
6	14.0523	0.56500	5.55760	43.5278	-29.4756	3.7643	*
7	6.0204	0.13817	5.21273	43.5278	-37.5074	3.3204	*
8	3.1888	0.29617	5.77453	43.5278	-40.3391	3.3255	*
9	1.8543	0.02833	5.00937	43.5278	-41.6736	3.3279	*

Thus, the Tamhane-Dunnett test finds significant decreases in mean response at dose 3 = 0.02 mg/L and at 0.06 mg/L and above, but not at dose 4 = 0.03 mg/L. So, there is some departure from monotonicity in the dose-response. Whether the NOEC should be set at 0.03 or at 0.01 from this test is a matter of scientific judgement.

MONOTONICITY CHECK OF Average_2 - FULL DATA
DOSES 0, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6 mg/L

PARM	P_T	SIGNIF
DOSE TREND	0.00000	**
DOSE QUAD	0.83911	

This is a formal test for departure from monotonicity in the dose response. It is based on linear contrasts, as described in section 5.1.3. In this instance, there is evidence of a linear dose-response response (p-value for dose trend <.00001), so there is no reason, based on this test, to doubt the over-all monotonicity of the dose-response. We accordingly proceed with the step-down Jonckheere-Terpstra test. In the results presented below, JONC is the value of the Jonckhhere-Terpstra test statistics and P1DNCF is the p-value associated with this test statistic to test the hypothesis of a downward trend. P1UPCF is the test statistic for testing the significance of an upward trend and is a default printout of the software used and is not utilized in the present analysis. ZC is a standardized value of the JONC statistic, and ZCCF is the same, except that it is computed using a standard continuity correction factor. The CF in several terms refers to the use of this continuity correction factor.

```
 KEY
 ZC IS JONCKHEERE STATISTIC COMPUTED WITH TIE CORRECTION
 ZCCF IS ZC WITH CONTINUITY CORRECTION FACTOR
 P1UPCF IS P-VALUE FOR UPWARD TREND
 P1DNCF IS P-VALUE FOR DOWNWARD TREND
 P-VALUES ARE FOR TIE-CORRECTED TEST WITH CONTINUITY CORRECTION FACTOR
 SIGNIF RESULTS ARE FOR A DECREASING ALTERNATIVE HYPOTHESIS

Jonckheere Trend Test on Dose 0 + Lowest 8 Doses thru 0.6 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -5.837314   -5.8087         1 2.2331E-9 **

Jonckheere Trend Test on Dose 0 + Lowest 7 Doses thru 0.3 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
               2 -5.422661 -5.389596         1 2.4387E-8 **

Jonckheere Trend Test on Dose 0 + Lowest 6 Doses thru 0.2 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -4.972358 -4.933512 0.9999996 2.7045E-7 **

Jonckheere Trend Test on Dose 0 + Lowest 5 Doses thru 0.1 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -4.476023 -4.429398 0.9999953 3.0535E-6 **

Jonckheere Trend Test on Dose 0 + Lowest 4 Doses thru 0.06 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -3.917286 -3.859679 0.9999432 0.0000352 **

Jonckheere Trend Test on Dose 0 + Lowest 3 Doses thru 0.03 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -3.267487 -3.193226 0.9992965 0.0004163 **

Jonckheere Trend Test on Dose 0 + Lowest 2 Doses thru 0.02 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -2.466679 -2.363901 0.9909582 0.0050929 **

Jonckheere Trend Test on Dose 0 + Lowest 1 Doses thru 0.01 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -1.333333 -1.166667 0.8783275 0.0668072
```

The Jonckheere-Terpstra test is significant at the 0.05 level at each step until the highest remaining concentration used in evaluating that statistics is 0.01 mg/L. Accordingly, the NOEC is set at 0.01 mg/L. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are

accepted. That is, blind reliance on the formal test for monotonicity is not advised. Rather, an informed judgement should be made based on all the available information. The same holds for assessing normality.

Example 2b. **Atrazine Example: Growth Rate**
NOEC Determination Atrazine by Two Methods

In this section, a NOEC is obtained for growth rate, both by Dunnett's test and by the step-down application of the Jonckheere-Terpstra test. For ease of reference, a table of growth rates obtained from the fluorescence data is given below.

Conc	Replicate	G_Rate	Conc	Replicate	G_Rate
0	A	0.07082	0.03	B	0.06269
0	B	0.07010	0.06	A	0.05860
0	C	0.06886	0.06	B	0.05587
0	D	0.06911	0.1	A	0.04688
0	E	0.07050	0.1	B	0.04599
0	F	0.06714	0.2	A	0.02929
0.01	A	0.06831	0.2	B	0.02840
0.01	B	0.06628	0.3	A	0.01631
0.02	A	0.06608	0.3	B	0.01435
0.02	B	0.06692	0.6	A	0.00490
0.03	A	0.06476	0.6	B	0.00409

The NOEC is 0.01 mg/L by the step-down Jonckheere-Terpstra Test, as well as by Dunnett's test.

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET ATRZ_GRATES
GROUP STATISTICS FOR Estimate BY DOSE

dose	Conc	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	6	0.069421	.069604	.001355632	.000553435
2	0.01	2	0.067294	.067294	.001435844	.001015295
3	0.02	2	0.066500	.066500	.000598875	.000423469
4	0.03	2	0.063729	.063729	.001462545	.001034175
5	0.06	2	0.057236	.057236	.001932705	.001366629
6	0.1	2	0.046432	.046432	.000626564	.000443048
7	0.2	2	0.028843	.028843	.000627415	.000443649
8	0.3	2	0.015327	.015327	.001389178	.000982297
9	0.6	2	0.004497	.004497	.000571829	.000404344

SHAPIRO-WILK TEST OF NORMALITY OF Estimate

OBS	STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
22	.000988660	-0.42123	-0.42443	0.94515	0.25242	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The non-significant p-value (p=0.25242) for the Shapiro-Wilk test indicates no reason to reject

the normality assumption. The Tukey outlier rule identified no observations of special interest and hence, no outliers are reported.

LEVENE TEST FOR Estimate - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	8	1.38466	0.28901	

The data was found to be consistent with a normal distribution, as shown by the Shapiro-Wilk test above, and with equal variances, as shown by the Lenene test immediately above. An analysis of variance will be performed.

Obs	Class	Levels	Values
1	dose	9	1 2 3 4 5 6 7 8 9

OVERALL F-TESTS FOR ANOVA

Obs	Effect	Num DF	Den DF	FValue	ProbF
1	dose	8	13	897.85	<.0001

As discussed in Chapter 5, no use is made of this result, though it does indicate that there is significant variation among the treatment means. It is a default computer output. The reader must be prepared to make intelligent use of default output.

ESTIMATED DOSE EFFECTS & DUNNETT FOR Decreasing ALTERNATIVE
USING ALPHA=.05 FOR COMPARISONS TO CONTROL

Estimate	SIGNIF	Dunnett 1-sided p-value	Test Group Mean	N
DOSE TREND	**	0.00000	.	.
DOSE QUAD	**	0.00000	.	.
DOSE 2-1		0.16679	0.067294	2
DOSE 3-1	*	0.04470	0.066500	2
DOSE 4-1	**	0.00035	0.063729	2
DOSE 5-1	**	0.00000	0.057236	2
DOSE 6-1	**	0.00000	0.046432	2
DOSE 7-1	**	0.00000	0.028843	2
DOSE 8-1	**	0.00000	0.015327	2
DOSE 9-1	**	0.00000	0.004497	2

The significant Dose Trend result indicates that there is a significant linear trend in the dose-response and no reason, by this formal test, to question the monotonicity of the dose-response. That there is also a significant quadratic trend is generally an indication that the overall trend is not linear. It does not, in itself, indicate non-monotonicity. The plots in the regression analysis of these data are instructive in this regard. An inspection of the treatment means reveals that the means are indeed monotone. The Jonckheere-Terpstra test does not require linearity, only monotonicity. By Dunnett's test, the NOEC is 0.01 mg/L, the lowest concentration.

```
Check for ties in Estimate
Percent of all data tied at 3 most frequently observed values
Since 5 < 25%, 9 < 40% and 14 < 65%
Exact methods (StatXact) are not required on this basis.

COUNT    SUMWTS    NMISS    NOBS    RESPONSE    TIES    TIEPCT

  1        22        2       24     0.004093      1        5
  1        22        2       24     0.004901      2        9
  1        22        2       24     0.014345      3       14
```

The small number of replicates per concentration, two (except in the control) suggests an exact Jonckheere-Terpstra test may be warranted. Certainly, such a test is not wrong. Below, both the exact and asymptotic (large sample) results are given. It will be observed that they do not agree. The exact result is considered definitive and is used to declare the NOEC to be 0.01 mg/L, the same as by Dunnett's test for this example. As will be seen below, each Jonckheere-Terpstra test is significant down to the 0.02 mg/L concentration. The result for the final test, where only the lowest concentration and control remain, differ in the exact and asymptotic versions of the test.

```
      Concentrations thru Group 9

       Jonckheere-Terpstra Test
---------------------------------------
Statistic (JT)                  13.5000
Z                               -5.8373
Asymptotic Test
One-sided Pr <  Z                <.0001
Exact Test
One-sided Pr <=  JT           8.691E-15


      Concentrations thru Group 8

       Jonckheere-Terpstra Test
---------------------------------------
Statistic (JT)                  13.0000
Z                               -5.4227
Asymptotic Test
One-sided Pr <  Z                <.0001
Exact Test
One-sided Pr <=  JT           1.629E-12
```

Concentrations thru Group 7

Jonckheere-Terpstra Test

Statistic (JT)	12.5000
Z	-4.9724
Asymptotic Test	
One-sided Pr < Z	<.0001
Exact Test	
One-sided Pr <= JT	2.447E-10

Concentrations thru Group 6

Jonckheere-Terpstra Test

Statistic (JT)	12.0000
Z	-4.4760
Asymptotic Test	
One-sided Pr < Z	<.0001
Exact Test	
One-sided Pr <= JT	2.863E-08

Concentrations thru Group 5

Jonckheere-Terpstra Test

Statistic (JT)	11.5000
Z	-3.9173
Asymptotic Test	
One-sided Pr < Z	<.0001
Exact Test	
One-sided Pr <= JT	2.511E-06

Concentrations thru Group 4

Jonckheere-Terpstra Test

Statistic (JT)	11.0000
Z	-3.2675
Asymptotic Test	
One-sided Pr < Z	0.0005
Exact Test	
One-sided Pr <= JT	1.563E-04

Concentrations thru Group 3

Jonckheere-Terpstra Test

Statistic (JT)	10.5000
Z	-2.4667
Asymptotic Test	
One-sided Pr < Z	0.0068
Exact Test	
One-sided Pr <= JT	0.0063

Concentrations thru Group 2

Jonckheere-Terpstra Test

Statistic (JT)	9.0000
Z	-1.6667
Asymptotic Test	
One-sided Pr < Z	0.0478
Exact Test	
One-sided Pr <= JT	0.0714

2.1 Example of data analysis by dose response modelling

It is assumed that an EC10 is required.
The data consist of observed biomass at three consecutive days, where exposure starts at the first day (= day 0). Each flask is sampled at the three points in time, so the data are not independent in time.

One approach for dose-response analysis of this type of data is to fit a dose-response model to the biomass observations for day 1 and day 2 separately. In practice, the analysis of one day only will actually be used (the one that results in the lowest EC10).

Another approach that does make use of all data available is to consider not the biomass but the growth rate as a function of concentration. From a statistical point of view, this approach is more efficient (in using all the data in a single analysis). Further, one may argue that changes in biomass result from changes in growth rate, so that growth rate is a more relevant parameter from a biological point of view. Unfortunately, no consensus exists on this issue. This second approach will be illustrated for the atrazine data set (this analysis is also briefly discussed in chapter 6).

Fig. 1 shows the observed biomass (i.e. fluoresence) data at days 0, 1 and 2, where nine different concentrations (including zero) of atrazine have been applied. An exponential model, $y = a \exp(bx)$, was fitted to the data, where a denotes the initial biomass at day zero, and b the growth rate. Given the experimental protocol, the initial biomass cannot depend on the concentration, and therefore the parameter a in this growth model can be assumed constant. The parameter b is estimated by allowing it to be dependent on the concentration, i.e. for each concentration a particular value for b is estimated. Thus, a total of 11 parameters are estimated from this dataset: one value for a, nine values for b, and one value for var, the residual variance (the variance of the residuals[1] on the log-scale).

In the analysis underlying Fig. 1, the observations are assumed to be independent. However, as discussed above, at each concentration various flasks were used, and each flask was sampled at the three days of observation, i.e. the individual flasks were followed in time. Therefore, the data as plotted in Fig. 1 are not independent, while nonetheless independence was assumed in fitting the model. This problem may be circumvented by estimating a growth rate for each individual flask. Since 22 flasks were observed (6 in the control, and 2 in the other concentration groups), a total of 24 parameters is estimated in such an analysis (including a and var). Fig. 2 shows the results of this analysis. The log-likelihood has increased from 464.80 (see fig. 1) to 483.86 (see fig. 2). This increase of 19.06 log-likelihood units for a model with 24 – 11 = 13 more parameters is highly significant according to the likelihood ratio test (twice the difference in log-likelihood is approximately Chi-square distributed with 13 degrees of freedom; $P \approx 0.0003$). Hence, it may be concluded that individual flasks differ from each other by themselves.
Fig. 1. Exponential growth model fitted to biomass, assuming a constant initial biomass (a), and growth rate (b) dependent on concentration (0, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6 mg/l). Biomass is plotted on the log-scale, resulting in linear growth curves. Biomass was assumed to be lognormally distributed, with homogenous variance on log-scale (i.e. homogenous CV). Here, 11 parameters are estimated (in a simultaneous fit).

[1] The residuals are the deviations of the individual data points from the fitted model.

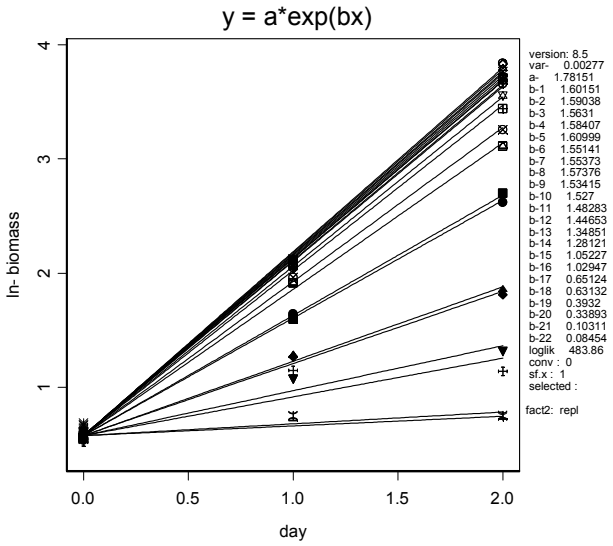


Fig. 2. Exponential growth model fitted to biomass, assuming a constant initial biomass (*a*), and growth rate (*b*) dependent on each individual flask (six for the control group, and 2 for each nonzero concentration. Thus, a total of 24 parameters are estimated (in a simultaneous fit). By comparing the log-likelihood (483.86) with that obtained in fig. 1 (loglik=464.80), it may be concluded that a significantly better fit is obtained, implying that flasks (at the same concentrations) are different with respect to growth rates.

As a second step in the analysis, the estimated growth rates are plotted against the concentration. Fig. 3 shows the growth rates from Fig. 2, as a function of concentration, and a dose-response model may be fitted to these data. Here, it will be assumed that the growth rates are normally rather than log-normally distributed (one of the reasons being that negative growth rates are possible). Fig. 3 shows the results for the Hill model fitted to the growth rates. This model fits the data extremely well, while the resulting curve is well confined by the data. Thus, estimation of an EC10 is fully warranted. It is no surprise that different models give very similar results and very narrow confidence intervals (see Table 1).

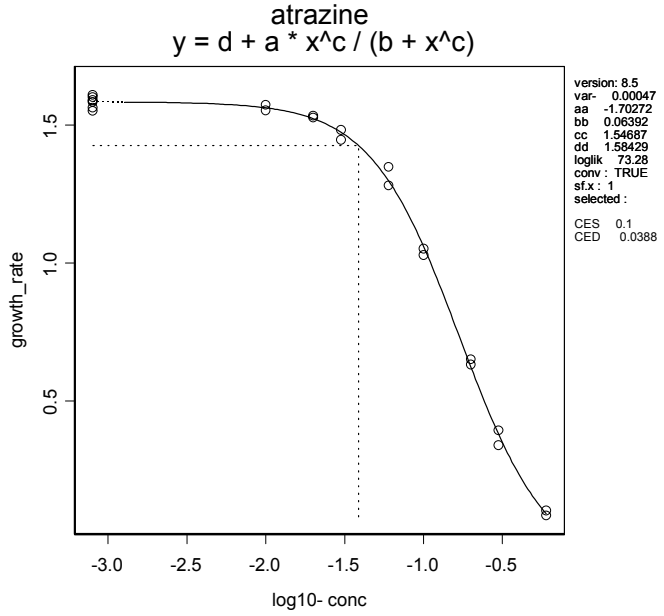


Fig. 3 Estimated growth rates (from individual flasks, see Fig. 2) as a function of concentration atrazine. The Hill model is fitted here to the data. The data are plotted against log-concentration to improve visibility.

Table 1. Summary of results of dose-response analysis for growth rate.

Model	Log-lik	EC10	90%-CI
$y = d + a\ x^c /(b^c + x^c)$	73.28	0.039	0.0355 - 0.0421 [1)]
$y = a\ (c - (c-1)\exp(-x/b)^d)$	71.17	0.035	0.0322 - 0.0386 [2)]

[1)] based on 5000 bootstrap runs
[2)] based on likelihood profile method

Assumptions

To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways. Here, two plots will be considered. One is the so-called QQ- plot, where the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with intercept equal to the mean of the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.

In interpreting a QQ-plot one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not

unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.

Biomass data
The biomass observations were assumed to be lognormally distributed, and therefore the model was fitted on log-scale. Hence, the residuals on a log-scale should be normally distributed with zero mean. From the left panel of Fig. 4 it may be concluded that the assumption of lognormally distributed biomass is reasonable. In the right panel of Fig. 4 the regression residuals are plotted for the three days separately. This plot reveals no differences in scatter between days, and it may be concluded that the assumption of homogeneous variances (on log-scale) is reasonable as well.

For the sake of illustration, the same plots are shown for the residuals resulting from an analysis without transformation. The QQ-plot shows a much less linear relationship, while the scatter is clearly not homogeneous between days. Clearly, an analysis after logarithmic transformation is more adequate.

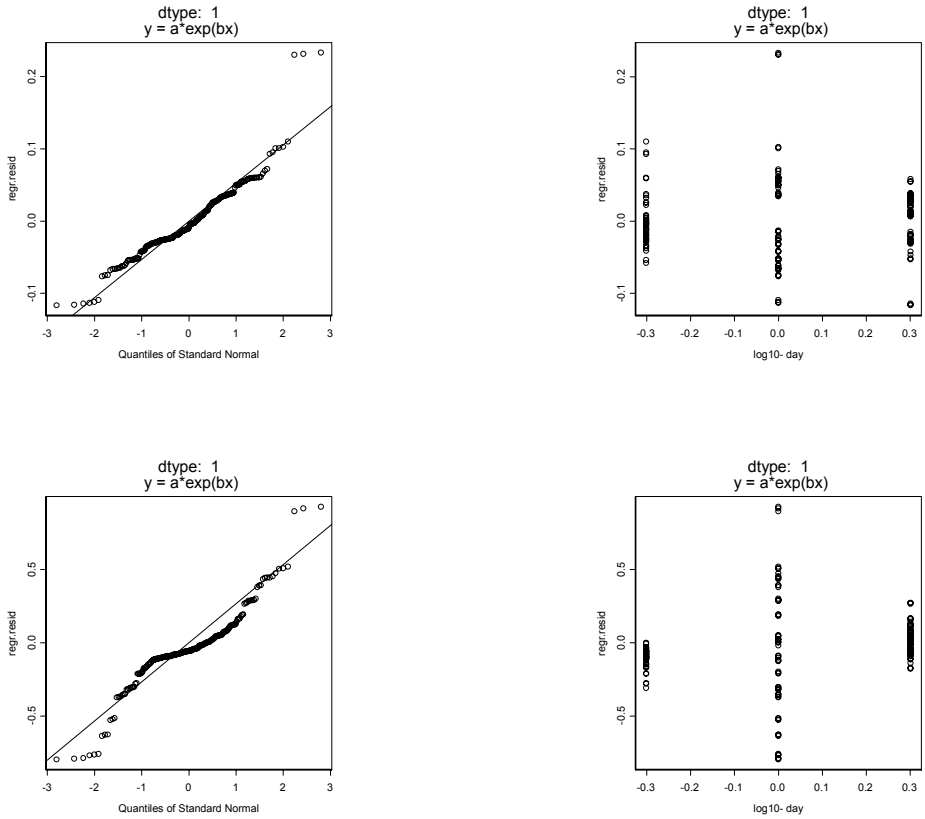


Fig. 4.Regression residuals from analysis on log-scale (upper panels) and from analysis without transformation (lower panels). The QQ-plots (left panels) show that the data comply with the assumption of log-normality, but less so with normality. The variances appear to be homogeneous for the analysis on log scale (upper right panel), with some outliers at day 1 (middle group). The same three outliers are visible in the QQ-plot (upper left panel). The analysis without transformation [lower right panel] results in a large scatter in the residuals at day 1 (middle group), and the assumption of homogeneous variances is clearly violated.

Growth rates

The growth rates were analysed without transformation, i.e. they were assumed to be normally distributed themselves, with homogeneous variance among concentration groups. As Fig. 5 shows, both assumptions appear reasonable.

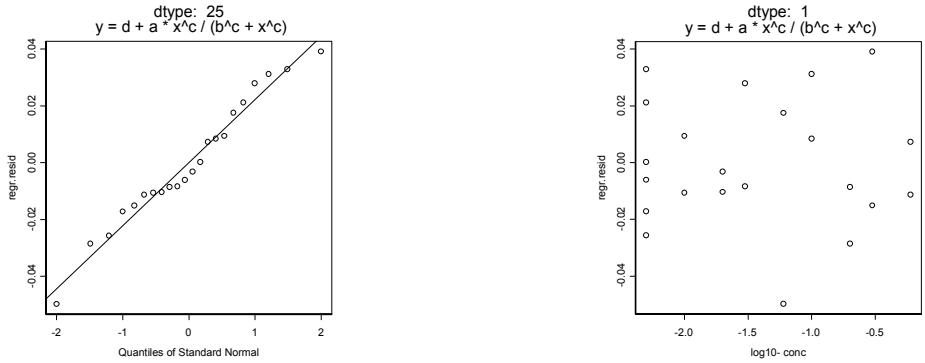


Fig. 5 Left panel: QQ-plot for regression residuals for growth rates, confirming the assumption of normality. Right panel: regression residuals plotted against (log-)concentration, confirming the assumption of homogeneous variances.

Dependencies due to individual flasks

The dose-response analysis discussed here was based on the growth rates derived for the individual flasks. As already discussed in chapter 6, an analysis that is based on an estimated growth rate for each concentration group (as in Fig. 1) will result in virtually the same estimate for the EC10 and for its confidence interval. Yet, the first analysis (separate growth rate estimate for each flask) is favourable, as it may give the information that a particular flask might be an outlier. Further, it may give information on weaknesses in the study protocol, for instance when replicate flasks in the same concentration group deviate from the general dose-response pattern. Such would indicate that the experimental protocol could be improved by better randomization. In this way the test could be made more effective.

2.2 Examples of data analysis using Debtox (biological methods)

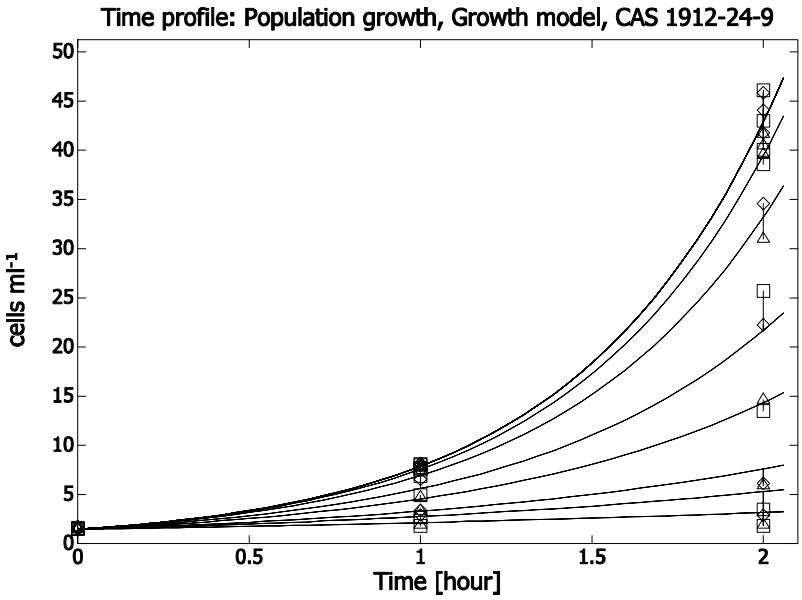
Data for effects of Atrazine in µg/l on the growth of *Selenastrum capricornutum* in cells/ml.

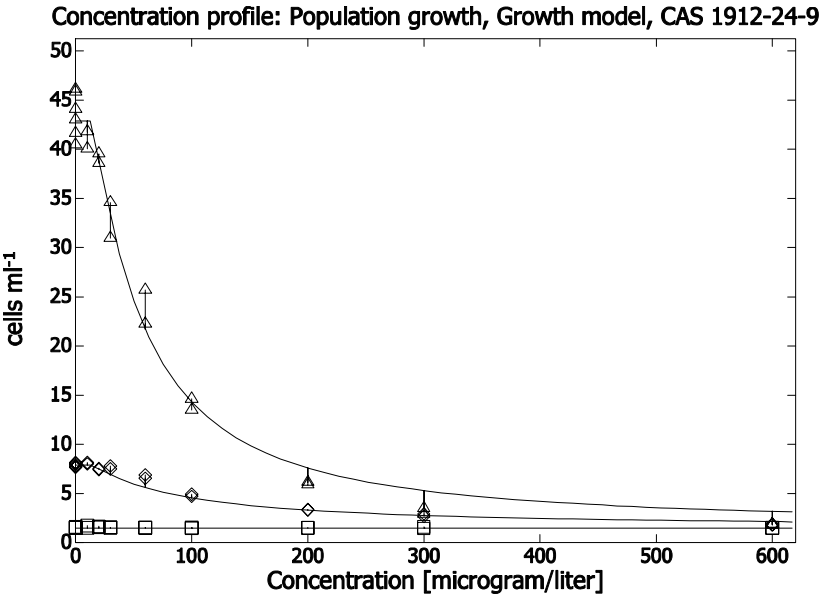
Time: hour, Conc: microgram/liter, Resp: cells ml-1																						
	0	0	0	0	0	0	10	10	20	20	30	30	60	60	100	100	200	200	300	300	600	600
0	1.5403	1.5239	1.5279	1.5586	1.5553	1.6126	1.5093	1.7349	1.6586	1.5529	1.5446	1.5273	1.5426	1.5223	1.5406	1.4833	1.5099	1.5049	1.5929	1.4529	1.4879	1.5003
1	7.6709	7.9489	7.7606	8.0956	8.1069	7.7463	8.0036	8.1433	7.4276	7.5383	7.5066	7.8069	6.8963	6.5459	4.7089	4.9329	3.3593	3.3193	2.6889	2.9189	1.8833	1.8033
2	46.1206	44.0773	41.6406	42.9973	45.8606	40.4706	40.0639	41.7773	39.5539	38.5706	34.5839	30.9639	25.6973	22.2406	14.6173	13.4873	6.1586	5.8823	3.4849	2.8926	1.8826	1.8259

Parameter estimates and Asymptotic Standard Deviations (ASD)

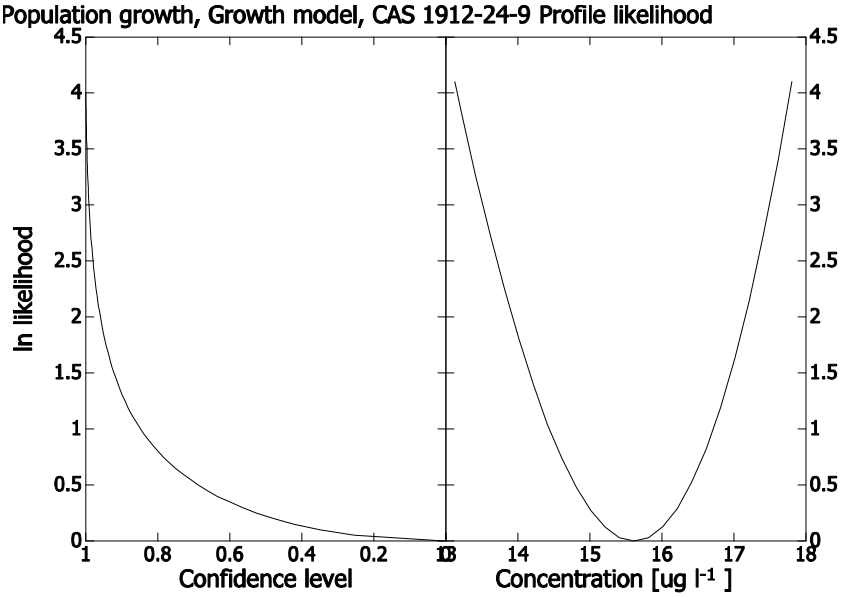
Population growth, Growth model			ASD	Correlation coefficients		
Inoculum size	1.446	·cells ml^{-1}	0.099			
Population growth rate	1.695	h^{-1}	0.035	-0.991		
No-effect concentration	15.61	ug l^{-1}	1.160	0.106	-0.161	
Tolerance concentration	176.6	ug l^{-1}	10.834	-0.570	0.559	-0.489
Mean deviation	1.13	·cells ml^{-1}				

Graphical test of model predictions against data (the different curves correspond to the different concentrations)





Profile likelihood for NEC estimate; 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



ECx values

day	EC0	ASD	EC50	ASD
1	15.6	1.16	139	5.75
2	15.6	1.16	61	2.03

Comments

The model for effects on the growth rate fits quite acceptably, but those for effects on adaptation and hazard fitted slightly better with similar NEC values (see table). The effects on growth have been selected here to improve the comparability with the concentration-response method. The 99% confidence intervals for the NEC values in µg/l for the three models are:

Hazard	5.89	14.4
Adaptation	4.46	9.59
Growth	13.2	17.4

These values are obtained by selecting the different models in the DEBtox software, and using its routine for computation of the confidence interval of the NEC.

ANNEX 3: ANALYSIS OF AN "*DAPHNIA MAGNA* REPRODUCTION" DATA SET (OECD GL 211 – ISO 10706) USING THE THREE PRESENTED APPROACHES

Data set on *Daphnia magna*

Data for the cumulative number of offspring per female as affected by an unkown compound.

		Number of Live Young Produced						
Conc (mg/L)	**Day**	**Rep A**	**Rep B**	**Rep C**	**Rep D**	**Rep E**	**Rep F**	**Rep G**
0	7	0	0	0	0	0	0	0
0	10	13	4	12	0	8	7	6
0	12	2	14	0	0	0	0	0
0	14	13	0	14	14	13	14	34
0	17	25	27	31	19	32	24	38
0	19	34	34	40	24	40	18	38
0	21	0	0	0	0	0	0	0
0.014	7	0	0	0	0	0	0	0
0.014	10	8	6	12	5	8	12	7
0.014	12	2	0	0	0	2	0	0
0.014	14	20	16	23	14	14	20	15
0.014	17	27	24	26	27	25	32	32
0.014	19	34	27	41	33	24	35	35
0.014	21	0	0	0	0	0	0	0
0.050	7	0	0	0	0	0	0	0
0.050	10	10	12	11	10	1	0	0
0.050	12	0	0	0	0	18	2	0
0.050	14	12	21	22	17	0	3	11
0.050	17	26	36	26	30	8	21	0
0.050	19	37	29	28	7	33	0	18
0.050	21	0	0	0	20	0	19	18
0.18	7	0	0	0	0	0	0	0
0.18	10	9	7	9	7	7	0	12
0.18	12	21	16	0	3	0	—	0
0.18	14	0	0	17	9	17	—	18
0.18	17	27	24	23	20	23	—	23
0.18	19	34	37	28	21	0	—	33
0.18	21	0	0	0	0	23	—	0
0.64	7	0	0	0	0	0	0	0
0.64	10	5	7	7	7	0	0	6
0.64	12	0	0	21	16	9	8	17

0.64	14	7	11	0	0	17	11	0
0.64	17	28	24	25	20	25	26	17
0.64	19	0	16	32	30	0	0	31
0.64	21	22	0	0	0	30	25	0
2.3	7	0	0	0	0	0	0	0
2.3	10	0	0	4	0	12	7	0
2.3	12	—	0	0	4	0	0	0
2.3	14	—	11	16	15	19	13	19
2.3	17	—	16	27	21	28	18	27
2.3	19	—	12	0	0	26	0	30
2.3	21	—	0	20	13	0	26	0
8.0	7	0	0	0	0	0	0	0
8.0	10	0	0	0	0	0	0	0
8.0	12	0	0	0	0	0	0	0
8.0	14	0	0	0	0	0	0	0
8.0	17	0	0	—	0	—	—	—
8.0	19	0	—	—	—	—	—	—
8.0	21	0	—	—	—	—	—	—

3.1 Examples of data analysis using hypothesis testing (NOEC determination)

Example 3a. Daphnid Chronic Reproduction Data
Total Live Young After 14 Days Exposure
NOEC Determination by Two Methods

Since TLY14 is count data, a square-root transform is used. Not reported is an analysis of untransformed values that yielded the same conclusions.

The NOEC exceeds 2.35 mg/L by both methods, the highest concentration for which there was a surviving adult. While it is possible to fit a regression model to these data, given the non-monotone and shallow nature of the dose-response, there is little point in doing so.

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET ADAP_REPRO
GROUP STATISTICS FOR TLY14 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	7	24.0000	21.0	8.4656	3.19970
2	0.015	7	26.2857	24.0	6.0198	2.27527
3	0.053	7	21.4286	22.0	10.6748	4.03471
4	0.19	6	25.3333	25.0	4.2740	1.74483
5	0.67	7	21.2857	23.0	5.4072	2.04374
6	2.35	6	20.0000	19.5	6.3875	2.60768
7	8.23	0	.	.	.	.

SHAPIRO-WILK TEST OF NORMALITY OF SQRT(TLY14)

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
0.75316	-0.43909	0.94111	0.97340	0.45806	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule is used to identify observations that may be of special interest. Outliers can have an impact on the results, as well as on the assessment of normality and variance homogeneity. A possibly valuable use of these outliers would be to re-run the analysis with outliers omitted to determine whether the NOEC is affected by these outliers. If it is, then caution should be followed in using the results. It is important to understand that just because an observation is identified as an outlier, that does not mean the observation is "bad" or that it should not be used. An observation should be excluded from analysis only for scientifically sound reasons and any such exclusion must be clearly stated along with its justification.

Outliers & Influential Observations
SQRT(TLY14) FROM FULL DATA

DAPHNID	dose	doseval	group	OBSER	Pred
20	3	0.053	III	2.23607	4.46961

SE_PRED	L95M	U95M	Resid	LB	UB
0.30488	3.85002	5.08920	-2.23354	-1.61856	1.70612

LEVENE TEST FOR SQRT(TLY14) - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	5	1.20201	0.32914	

The data was found to be normally distributed (Shapiro-Wilk p-value=0.45806) with equal variances (Levene p-value=0.32914). An analysis of variance will be performed.

Obs	Class	Levels	Values
1	dose	6	1 2 3 4 5 6

OVERALL F-TESTS FOR ANOVA

Effect	Num DF	Den DF	FValue	ProbF
dose	5	34	0.84	0.5330

The overall ANOVA F-test is not significant. However, this does not affect the remainder of the analysis. As discussed in chapter 5, a significant or non-significant F-test should not be used as a decision rule for the multiple comparisons of step-down Jonckheere-Terpstra test. The F-test can be affected by significant differences among treatments of no interest to toxicology or by the necessity to control for a large number of possible differences of no interest to toxicology.

ESTIMATED DOSE EFFECTS & DUNNETT FOR Decreasing ALTERNATIVE
USING ALPHA=.05 FOR COMPARISONS TO CONTROL

Estimate	SIGNIF	Dunnett 1-sided p-value	Test Group Mean	N
DOSE TREND		0.24959	.	.
DOSE QUAD		0.65360	.	.
DOSE 2-1		0.95623	5.09841	7
DOSE 3-1		0.49052	4.46961	7
DOSE 4-1		0.92891	5.01786	6
DOSE 5-1		0.60723	4.57826	7
DOSE 6-1		0.45937	4.42441	6

The Dose Trend and Dose Quad are formal tests for departure from monotonicity in the dose response. They are based on linear contrasts, as described in section 5.1.3. In this instance, the test for linear dose-response is not significant but neither is the test for departure from linearity (Dose Quad), so there is no reason, based on these tests, not to go on with the step-down Jonckheere-Terpstra test. However, an inspection of the treatment means does indicate some non-monotonicity in the dose-response. Thus, some caution should be used in interpreting the results of the Jonckheere-Terpstra test presented below. In the results presented below, JONC is the value of the Jonckhhere-Terpstra test statistics and P1DNCF is the p-value associated with this test statistic to test the hypothesis of a downward trend. P1UPCF is the test statistic for testing the significance of an upward trend. PIUPCF is a default printout of the software used and is not utilized in the present analysis. ZC is a standardized value of the JONC statistic, and ZCCF is the same, except that it is computed using a standard continuity correction factor. The CF in several terms refers to the use of this continuity correction factor.

```
 KEY
 ZC IS JONCKHEERE STATISTIC COMPUTED WITH TIE CORRECTION
 ZCCF IS ZC WITH CONTINUITY CORRECTION FACTOR
 P1UPCF IS P-VALUE FOR UPWARD TREND
 P1DNCF IS P-VALUE FOR DOWNWARD TREND
 P-VALUES ARE FOR TIE-CORRECTED TEST WITH CONTINUITY CORRECTION FACTOR
 SIGNIF RESULTS ARE FOR A DECREASING ALTERNATIVE HYPOTHESIS

Jonckheere Trend Test on Dose 0 + Lowest 5 Doses thru 2.35 mg/L
Analysis of TLY14

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
            281.5 -1.223426 -1.211548 0.8871572 0.1083588
```

Since the Jonckheere-Terpstra test with all concentrations present is not significant, no further testing is required. Jonckheere test results are included in the summary table. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are accepted.

It will be observed that neither the Dunnett test nor the Jonckheere-Terpstra test found a significant effect at any concentration at 14 days.

Example 3b. Daphnid Chronic Reproduction Data
Total Live Young After 21 Days Exposure
NOEC Determination by Two Methods

Since TLY21 is count data, a square-root transform is used. Not reported is an analysis of untransformed values that yielded the same conclusions.

The NOEC by Dunnett's test exceeds 2.35 mg/L, the highest tested concentration for which at least one adult daphnid survived. The NOEC by the Jonckheere-Terpstra test is 0.19 mg/L.

ECOTOX MEASUREMENTS FROM DATASET ADAP_REPRO
GROUP STATISTICS FOR TLY21 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	7	84.5714	87.0	20.3130	7.67760
2	0.015	7	86.5714	89.0	11.8583	4.48201
3	0.053	7	72.2857	84.0	21.2581	8.03479
4	0.19	6	78.0000	80.5	11.4717	4.68330
5	0.67	7	71.4286	71.0	9.5718	3.61779
6	2.35	6	64.0000	65.5	16.3707	6.68331
7	8.23	0	.	.	.	.

SHAPIRO-WILK TEST OF NORMALITY OF SQRT(TLY21)
AQUATIC DAPHNID: FULL DATA

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
0.87309	-0.34552	-0.55535	0.96350	0.22028	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as the sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule was used to identify observations that may be of special interest. For these data, no outliers were identified.

LEVENE TEST FOR SQRT(TLY21) - FULL Model
ANALYSIS OF VARIANCE ON FULL DATA SET

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	5	0.91555	0.48269	

The data was found to be consistent with a normal distribution with equal variances. An analysis of variance will be performed.

Obs	Class	Levels	Values
1	dose	6	1 2 3 4 5 6

OVERALL F-TESTS FOR ANOVA

Effect	Num DF	Den DF	FValue	ProbF
dose	5	34	1.90	0.1199

The overall ANOVA F-test is not significant. However, this does not affect the remainder of the analysis. As discussed in chapter 5, a significant or non-significant F-test should not be used as a decision rule for the multiple comparisons of step-down Jonckheere-Terpstra test. The F-test can be affected by significant differences among treatments of no interest to toxicology or by the necessity to control for a large number of possible differences of no interest to toxicology.

ESTIMATED DOSE EFFECTS & DUNNETT FOR Decreasing ALTERNATIVE
USING ALPHA=.05 FOR COMPARISONS TO CONTROL

Estimate	SIGNIF	Dunnett 1-sided p-value	Test Group Mean	N
DOSE TREND	*	0.01150	.	.
DOSE QUAD		0.71556	.	.
DOSE 2-1		0.91014	9.28556	7
DOSE 3-1		0.24321	8.41734	7
DOSE 4-1		0.59511	8.81095	6
DOSE 5-1		0.25588	8.43515	7
DOSE 6-1		0.05306	7.94130	6

The Dose Trend and Dose Quad are formal tests for departure from monotonicity in the dose response. They are based on linear contrasts, as described in section 5.1.3. In this instance, the test for linear dose-response is significant but not the test for departure from linearity (Dose Quad), so there is no reason, based on these tests, not to go on with the step-down Jonckheere-Terpstra test. As discussed in section 5.3, the only condition, based on these formal tests, to question the use of the Jonckheere-Terpstra test would be a non-significant test for dose trend and a significant test for dose quad. In addition, an inspection of the treatment means indicate some non-monotonicity in the dose-response, but the overall downward trend is quite evident. In the results presented below, JONC is the value of the Jonckhhere-Terpstra test statistics and P1DNCF is the p-value associated with this test statistic to test the hypothesis of a downward trend. P1UPCF is the test statistic for testing the significance of an upward trend. This is a default printout of the software used and is not utilized in the present analysis. ZC is a standardized value of the JONC statistic, and ZCCF is the same, except that it is computed using a standard continuity correction factor. The CF in several terms refers to the use of this continuity correction factor.

```
 KEY
 ZC IS JONCKHEERE STATISTIC COMPUTED WITH TIE CORRECTION
 ZCCF IS ZC WITH CONTINUITY CORRECTION FACTOR
 P1UPCF IS P-VALUE FOR UPWARD TREND
 P1DNCF IS P-VALUE FOR DOWNWARD TREND
 P-VALUES ARE FOR TIE-CORRECTED TEST WITH CONTINUITY CORRECTION FACTOR
 SIGNIF RESULTS ARE FOR A DECREASING ALTERNATIVE HYPOTHESIS

Jonckheere Trend Test on Dose 0 + Lowest 5 Doses thru 2.35 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
            216.5 -2.761099 -2.749249 0.9970134 0.0027775 **

Jonckheere Trend Test on Dose 0 + Lowest 4 Doses thru 0.67 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
            163.5 -2.050218 -2.035031 0.9790761 0.0194424  *

Jonckheere Trend Test on Dose 0 + Lowest 3 Doses thru 0.19 mg/L

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
            107.5 -1.255247 -1.233604 0.8913248 0.1008208
```

Since the Jonckhere-Terpstra test is not significant with all concentrations above 0.19 mg/L omitted, no further testing is required and the NOEC is 0.19. It will be observed that Dunnett's test found no significance at any concentration.
The Jonckheere test results are included in the summary table. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are accepted.

3.2 Example of data analysis by dose response modelling

Daphnia reproduction test: number of young

Total number of life young are reported at various points in time after exposure, showing a clear increase in rate of number of young produced with time. This information in time may be used, as shown at the end of this section. First, however, it is illustrated how the number of young for a particular time period may be analysed by dose-response modelling

Since the production of young only starts after a number of days, with an increasing rate in the period thereafter, the total count over the first two weeks, or that over the third week may for examplebe chosen as the response variable for dose-response analysis. As Fig. 1 illustrates, the production of young has indeed increased with age. This figure also illustrates that the dose-response relationship of the counts over the first two weeks is similar to that over the third week. Although in this way the data from the first two weeks and from the third week are used together in a single analysis, the problem is that the analysis assumes the data to be independent, which they are not (see Fig. 2). Therefore, the analysis of Fig. 1 is from a statistical point of view not valid, in particular the confidence interval may not be reliable.

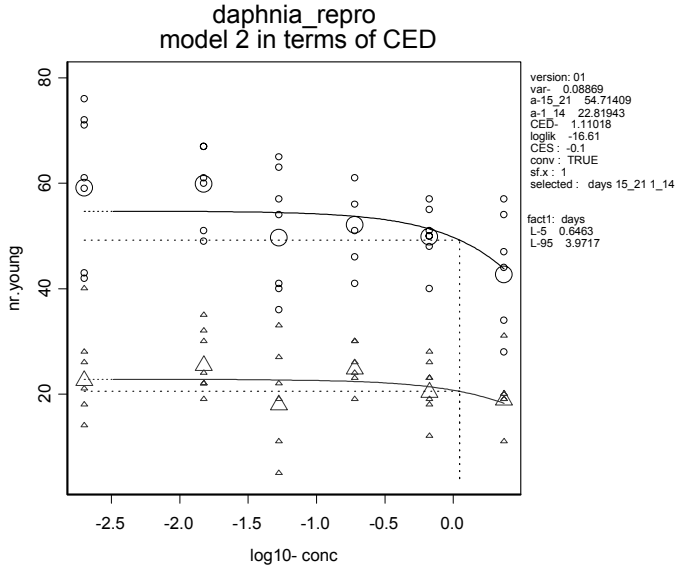


Fig. 1. Number of life young as a function of concentration (on log-scale to improve visibility), counted over the first two weeks (triangles) and over the third week (circles).

CED = EC10. Model 2 : y = a exp(bx). This model was reparameterized by substituting the parameter b by the EC10. In this way the confidence interval for the EC10 (L-5, L-95) can be estimated by the likelihood profile method. Note: the confidence interval in this analysis may not be reliable, due to violation of independence of the data (see Fig. 2).

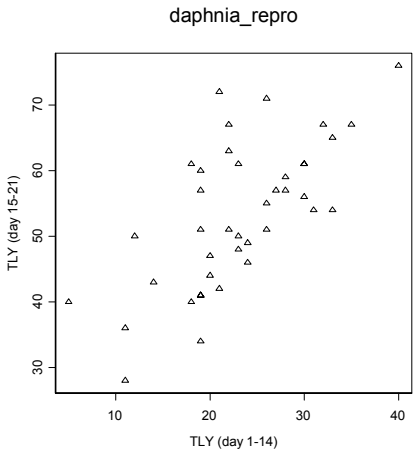


Fig. 2. Total life young (TLY) in third week plotted against TLY in first two weeks, showing the correlations between these counts.

The obvious way to avoid the problem of dependent data in the analysis of Fig. 1 is to perform the analysis on the counts over the third week only. Another argument for this selection is that it may be assumed that at this time the reproduction rate has reached a more or less stable level, whereas the counts over the first two-weeks period includes the starting-up of the reproduction (leading to more variation and problems of interpretation). According to the recommendation of chapter 6, various models will be fitted to the counts over week three.

First the nested nonlinear model proposed by Slob (2002) is fitted. This results in the following log-likelihoods:

model 1	$y = a$	loglik = 4.19
model 2	$y = a \exp(x/b)$	loglik = 8.26
model 3	$y = a \exp(\pm(x/b)^d)$	loglik = 8.26
model 4	$y = a\,[c - (c - 1) \exp(-x/b)]$	loglik = 8.26

The log-likelihoods can be compared with the likelihood-ratio test. A model with one more parameter fits the data significantly better (at $\alpha = 0.05$) than the model without that parameter when the increase in the log-likelihood is greater than 1.92. The difference in log-likelihoods between model 2 and model 1 is 4.07, so it may be concluded that the data show a significant dose-response. The models with more parameters result in the same log-likelihood, and it may be concluded that model 2 is, from this family of models, the appropriate one for describing the data. Fig. 3 shows the results for this model.

Next a polynomial model is fitted to these data. Since this is again a nested family of models, the log-likelihoods can be compared by the ratio-likelihood test:

model 1	$y = a$	loglik = 4.19
model 2	$y = a + bx$	loglik = 8.20
model 3	$y = a + bx + cx^2$	loglik = 8.50

Here, model 2 (straight line) is not significantly improved by higher order polynomials, and this model may be selected from this family of models.

Finally, the power model, $y = c + ax^b$, is fitted to the number of young. This model results in a log-likelihood value of 9.15. However, by fixing the parameter b to one, the model reduces to a straight line. Hence, the power model and the straight line are nested, and their log-likelihoods can be compared. Since the straight line resulted in a log-likelihood of 8.20, the power model does not give a significantly better fit.

It may be concluded that these data should be described by a two-parameter model, either the exponential, or the linear (straight line) model. Fig. 3 shows the fit of the exponential model. Table 1 summarizes the results for both fits, showing that the EC10 (and its confidence interval) are similar for both models.

Table 1. Number of life young: summary of results for exponential and straight-line model, both having two parameters.

Model	Log-lik	EC10	90%-CI
$y = a\exp(bx)$	8.26	0.91	0.58 – 2.05 [1)]
$y = a + bx$	8.20	1.00	0.69 – 2.15 [2)]

[1)] based on profile-likelihood method
[2)] based on 1000 bootstrap runs

Assumptions

To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways. Here, two plots will be considered. One is the so-called QQ- plot, where the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with intercept equal to the mean of the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.
In interpreting a QQ-plot one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.
As Fig. 4 (upper left panel) shows, the data did comply with the assumption of log-normality, since these residuals resulted from an analysis on the log-counts. The same residuals plotted against concentration (upper right panel) do not reveal a clear trend, and the assumption of homogeneous variances (on log-scale) appears acceptable.
Although not strictly needed, the residual plots are also shown for an analysis where the log-transformation was omitted (middle panels of Fig. 3), as well as where a square root transformation was applied (lower panels of Fig. 3). The plots for these three situations are similar. The reason of this similarity is that the scatter in these data is relatively small (CV of around 20%) A log-normal distribution gets closer to a normal distribution with smaller variation (CV). Therefore, the smaller the scatter in the data, the more data are needed to see any difference in the QQ-plots assuming normality, log-normality, or square-root-normality. For the same reason, it may be expected that applying or omitting any transformation has no large impact on the results of the analysis when the scatter in the data is relatively small**.** Indeed, re-analysing these data without transformation results in an EC10 = 0.90 mg/l, while the same analysis with log-transformation resulted in EC10 = 0.91 mg/l.

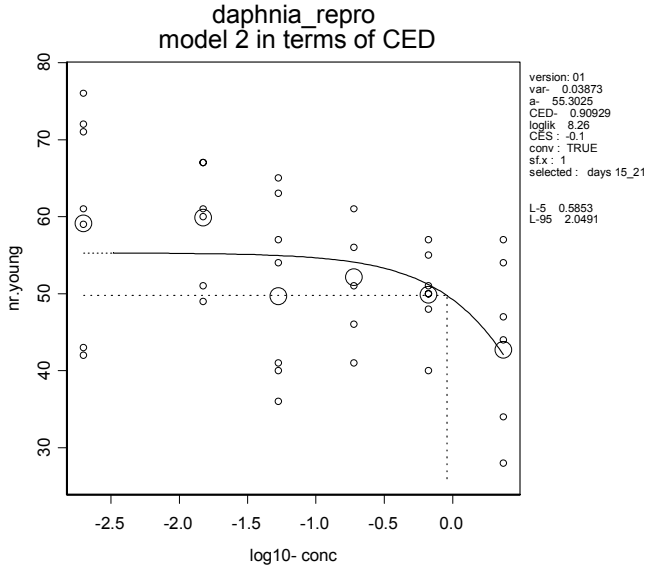


Fig. 3 Exponential model fitted to the number of life young counted over week three. This model was reparameterized by substituting the parameter *b* by the EC10. In this way the confidence interval for the EC10 (L-5, L-95) can be estimated by the likelihood profile method.

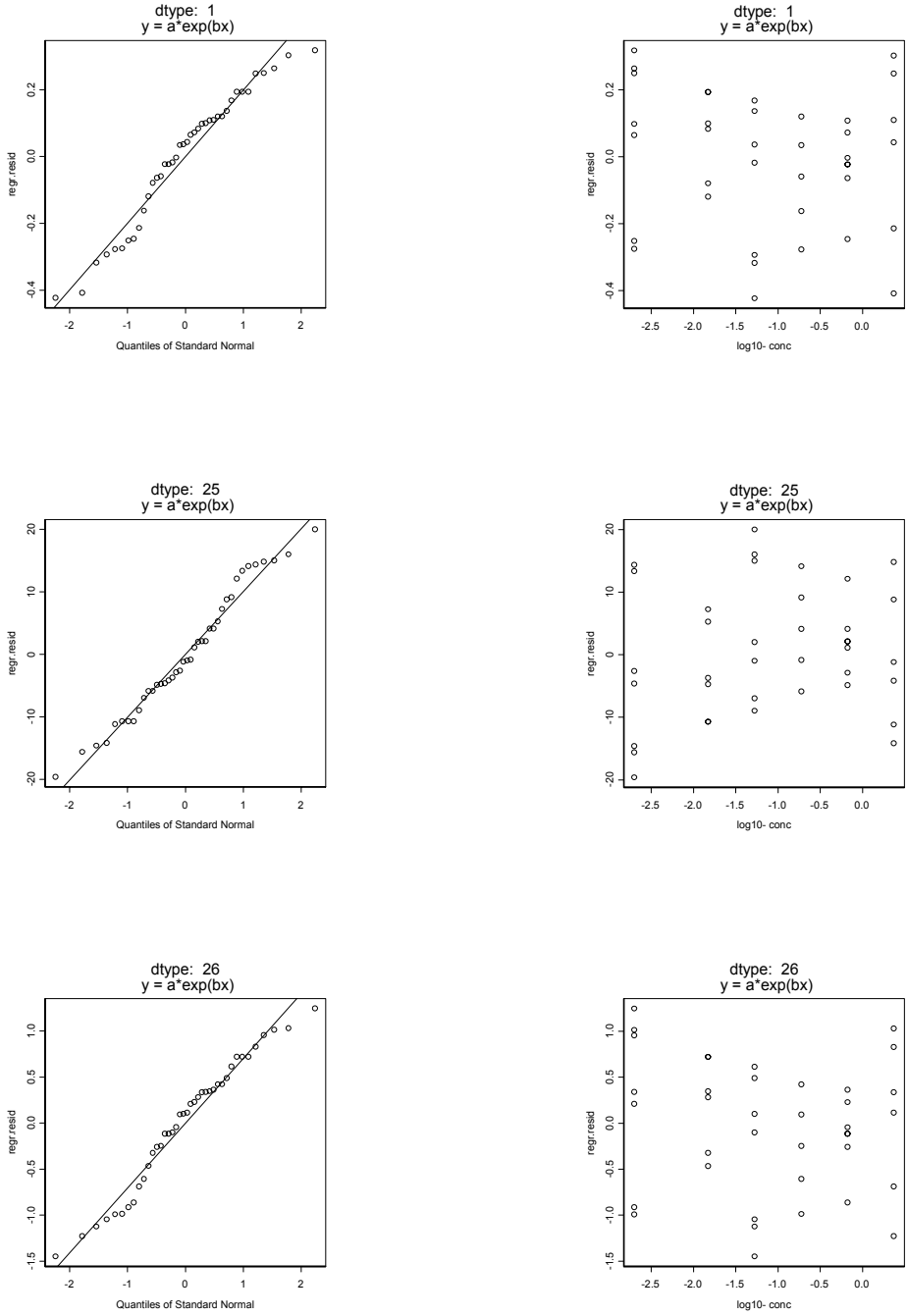


Fig. 4. Left panels: QQ-plots of the regression residuals for the exponential model. Right panels: the same residuals plotted against dose (dose on log-scale to improve visibility).

Upper panels: analysis on log-scale, middle panels: no transformation, lower panels: analysis on square root scale.

Two-step analysis: taking time into account

Similar to the algal test example, these data can be analysed in two steps, taking the information in time into account. In the first step the (cumulative) number of eggs is considered as a function of time. Fig. 5 shows the Hill model fitted to the data, where for each concentration a separate value is estimated for the parameter *b* (= ET50, time at which 50% of maximum value is achieved).
In these data of cumulative counts, the variance clearly increases with the mean (data not shown). A log-transformation resulted in the variance decreasing with the mean, but a square root transformation resulted in homogeneous variance, and compliance with the normal distribution.

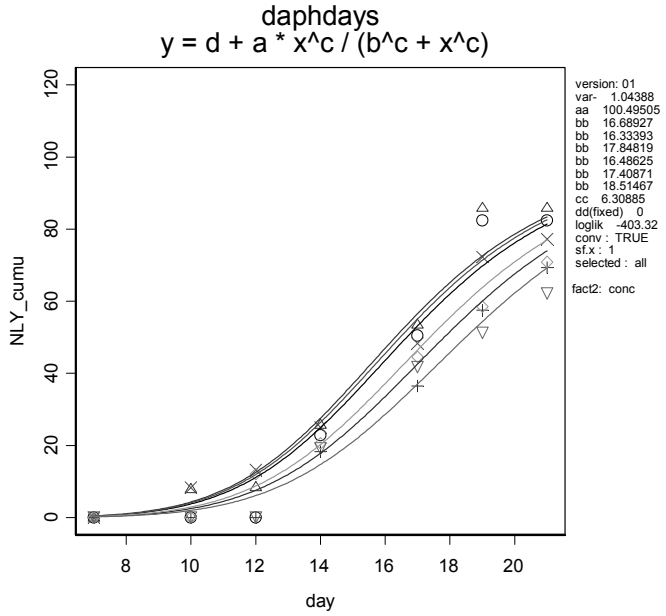


Fig. 5. Means of number of young, plotted cumulatively against time. Each symbol represents a concentration, and to each concentration the Hill model is fitted, assuming that the parameters *a* and *c* are equal amongst concentrations, while parameter *b* was assumed to be different. The parameter *d* was set at zero here (since reproduction is zero at time zero).

In the second step, the estimated values for the ET50 are considered as a function of the concentration, and a dose-response model is fitted to these data (see Fig. 6). Again, the second (nonzero) concentration appears to deviate from the general pattern (compare with Fig. 3). The EC10 for this endpoint is estimated at 2.30 mg/l, with C.I. (1.41, 6.19).
Since the number of young are followed in time for a number of replicates at each concentration, it is better to estimate an ET50 for each replicate (see discussion in algal data set). This is illustrated in Fig. 7. Next the ET50 values for each replicate may be considered as a function of the concentration (see Fig. 8). The EC10 is estimated at 2.36 mg/l, similar to the value obtained in Fig. 6, but the confidence interval is somewhat larger (1.36, 8.82).
From Fig. 8 it becomes apparent that the deviation of the second concentration group is caused by two outlier replicates. When these two outliers are removed, a more regular dose-response relationship results (see Fig. 9). This also results in a lower EC10 (1.90 mg/l), and a smaller confidence interval (1.30, 3.49).

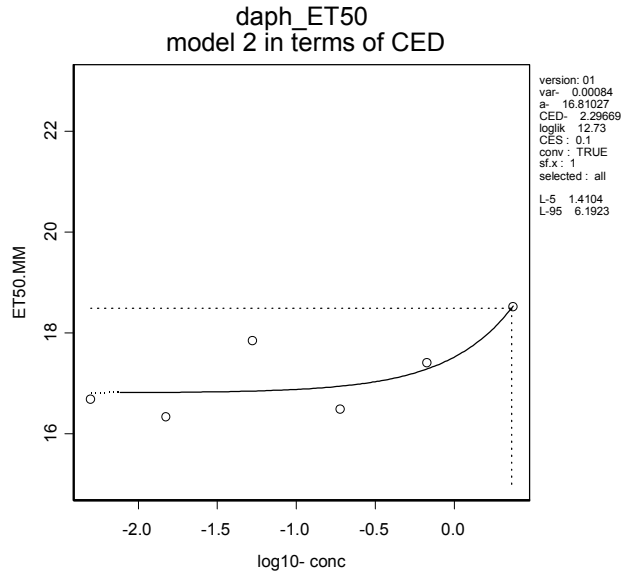


Fig. 6. Estimated ET50 values from Fig. 5, plotted against the concentration, with a fitted dose-response model.

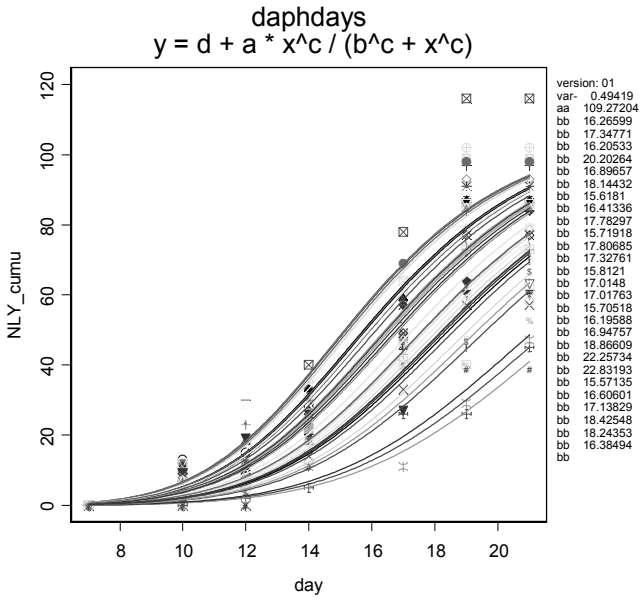


Fig. 7. Number of young, plotted cumulatively against time. Each symbol represents a replicate, and to each replicate the Hill model is fitted, assuming that the parameters *a* and *c* are equal amongst replicates, while parameter *b* was assumed to be different. The parameter *d* was set at zero here (since reproduction is zero at time zero).

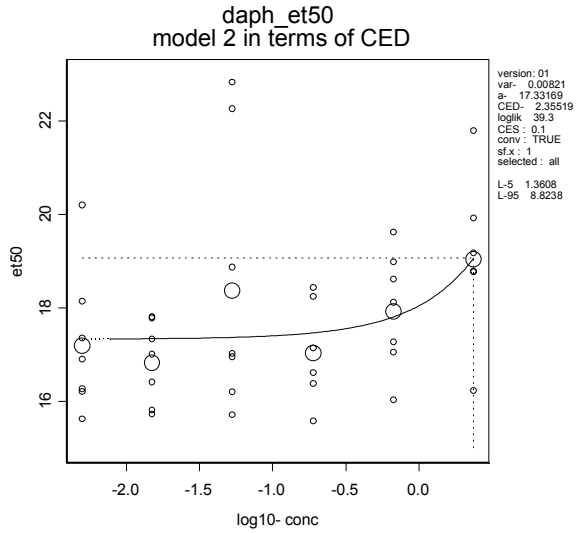


Fig. 8. ET50s estimated per replicate (see Fig. 7) as a function of the concentration with a fitted dose-response model.

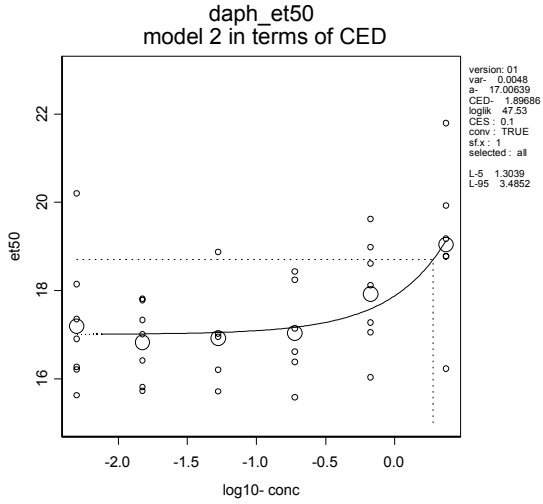


Fig. 9. ET50s estimated per replicate (see Fig. 7) as a function of the concentration, with two outliers removed.

3.3 Examples of data analysis using Debtox (biological methods)

Data for the cumulative number of offspring per female as affected by an unkown compound. The data were weighted in the estimation of parameters by the number of surviving females (data not given here).

	Time: day, Conc: milligram/liter, Resp: Number of offspring						
	0.000	0.015	0.053	0.190	0.670	2.350	8.230
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	7.14286	8.28571	6.28571	7.28571	4.57143	3.28571	0.00000
12	9.42857	8.85714	9.14286	13.95240	14.71430	3.95238	0.00000
14	24.00000	26.28570	21.42860	24.11900	21.28570	19.45240	0.00000
17	52.00000	53.85710	42.42860	47.45240	44.85710	42.28570	0.00000
19	84.57140	86.57140	64.14290	72.95240	60.42860	53.61900	0.00000
21	84.57140	86.57140	72.28570	76.78570	71.42860	63.45240	0.00000

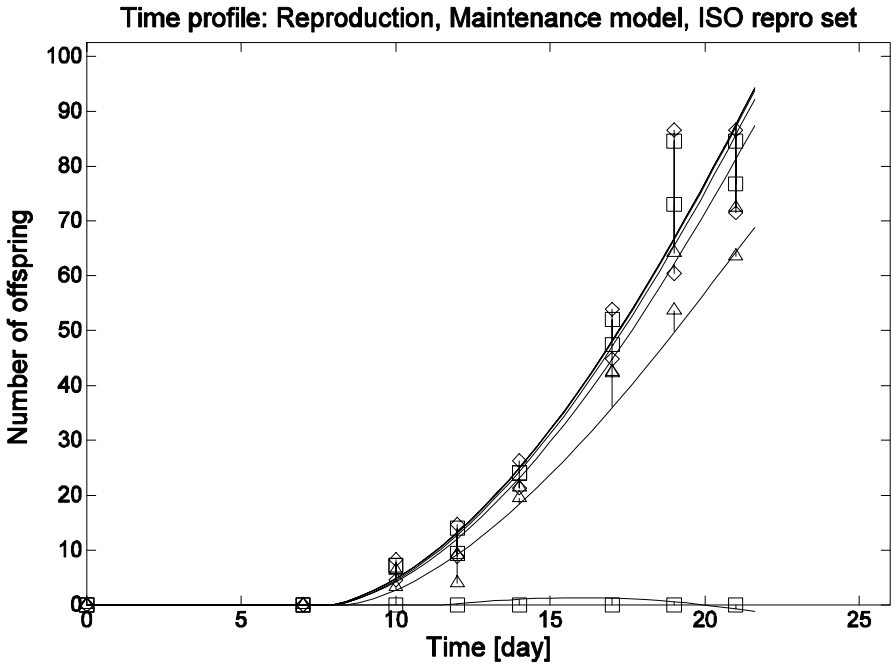
Parameter estimates and Asymptotic Standard Deviations (ASD)

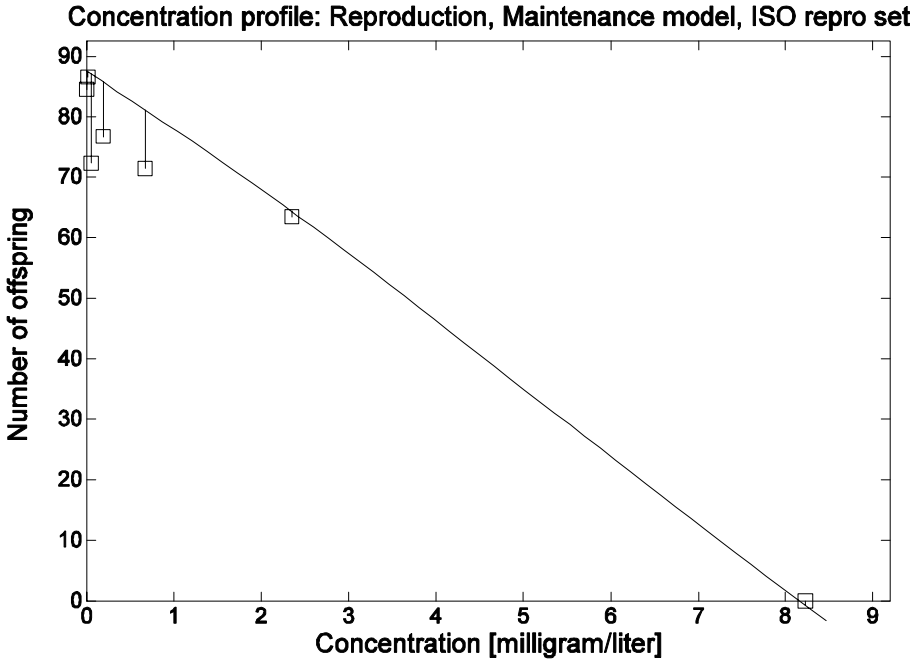
Reproduction, Maintenance model			ASD	Correlation coefficients		
No effect concentration	3.895e-009	mg l^{-1}	0.004			
Tolerance concentration	0.2265	mg l^{-1}	35.175	0.233		
Maximal reproduction rate	15.9	No d^{-1}	0.646	-0.872	-0.030	
Elimination rate	0.001268	d^{-1}	0.199	0.233	1.000	-0.031
Von Bertalanffy growth rate	0.1	d^{-1}				
Scaled length at birth	0.13					
Scaled length at puberty	0.61					
Energy investment ratio	1					
Mean deviation	5.207					

ECx values (derived from parameter values) in mg/l.

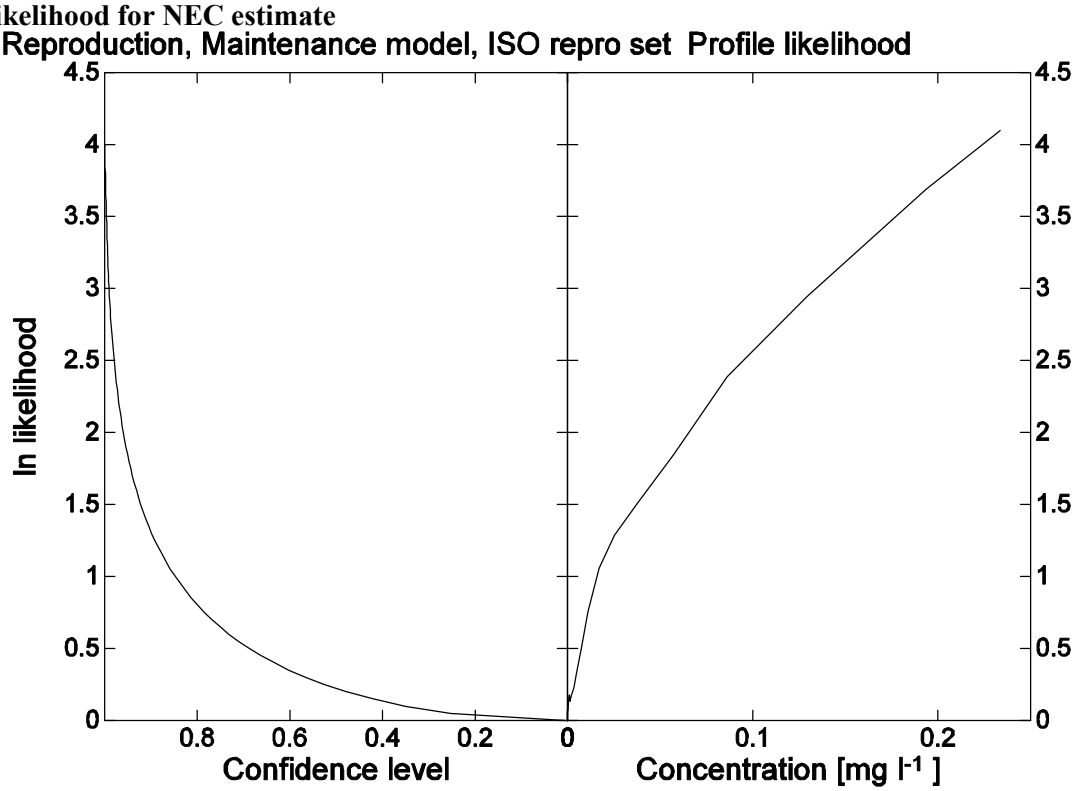
Day	EC0	ASD	EC50	ASD
21	1.10^{-5}	0.44	4.22	6.59

Graphical test of model predictions against data

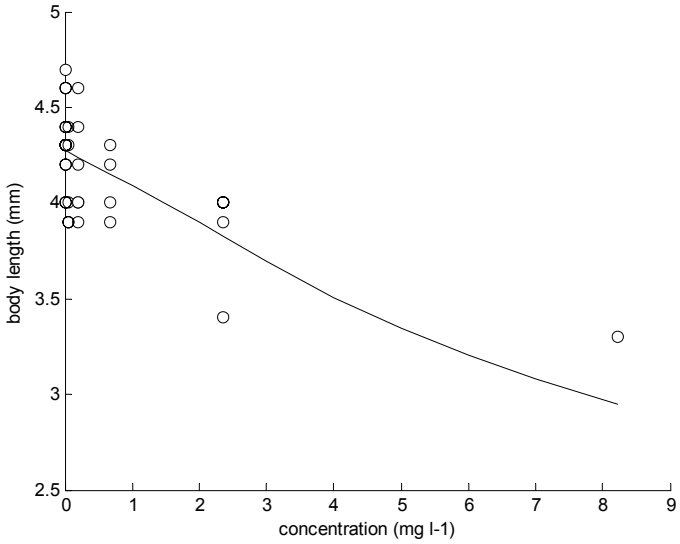

Time profile: Reproduction, Maintenance model, ISO repro set
Number of offspring
Time [day]
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25


Concentration profile: Reproduction, Maintenance model, ISO repro set
Number of offspring
Concentration [milligram/liter]
0
10
20
30
40
50
60
70
80
90
0
1
2
3
4
5
6
7
8
9

Profile likelihood for NEC estimate



Body length at 21 days



Comments

This dataset is special in several respects. We have counts of offspring at relatively few points in time, not for each day as the guideline recommends. This reduces the effectiveness of the biology-based method; the fact that these data do not give detailed information about the start of the reproduction is especially troublesome. The first graph shows that reproduction starts here later than expected on the basis of the default value of 0.42 for the scaled length at puberty. Therefore, this scaled length has been set at 0.61 to mimic this late start. The model for effects on maintenance appeared to fit the data best; an increase in maintenance costs reduces the ultimate length of the daphnids. This is confirmed by the length data, shown in the last plot; the fitted length at 21 d are calculated with DEBtool; the plotted curve involves the estimation of a single parameter: the ultimate length in the blank. All other parameters are already fixed by the reproduction data, and determine the toxicokinetics, including the dilution by growth, and the effects on growth during exposure. The length data were not used to estimate any effect parameters. The good fit of the length data, confirms the effects on growth by an increase of the maintenance costs as expected from the observed effects on reproduction. Direct effects on reproduction would not affect body size; direct effects on growth would affect the growth rate, but not the ultimate size. The data do not give information about growth, but it is likely that growth almost ceased before 21 d for daphnids, even in the stressed situation. The NEC was found to be not significantly different from zero, with a 95% confidence interval of (0, 0.082) mg/l.

ANNEX 4: ANALYSIS OF A "FISH GROWTH" DATA SET (OECD GL 204/215 – ISO 10229) USING THE THREE PRESENTED APPROACHES

Data set

The data used here have been obtained for a 21 day fish test following OECD GL204.

test organism : *Oncorhynchus mykiss*
temperature : 14-15°C

Number of dead fish

	Control	**Nominal concentration of the test item (mg/l)**					
day	0	1.0	2.2	4.6	10	22	46
1							
2							
3							
6							1
7							2
8							2
9							2
10			1				5
13			1			1	5
14			1		1	1	5
15			1		1	1	5
16			1		1	1	5
17			1		1	1	5
21	0	0	1	0	1	1	6

Length of the fish in mm

	Control	Nominal concentration of the test item (mg/l)					
	0	1.0	2.2	4.6	10	22	46
0 days	5.7	5.7	5.1	5.9	5.9	6.0	5.9
	5.3	5.8	6.0	5.9	5.7	5.6	5.6
	6.0	5.6	5.7	5.7	5.5	5.9	6.0
	5.9	6.0	5.8	5.5	5.6	5.6	5.6
	5.9	5.5	5.5	6.0	5.6	5.7	5.3
	5.6	6.0	5.8	6.0	5.9	5.8	6.0
	5.6	5.9	6.0	5.6	6.0	5.2	5.0
	5.0	5.1	5.6	5.3	5.0	5.7	5.3
	6.0	5.2	5.1	5.3	5.5	5.1	5.1
	5.4	5.4	5.6	5.1	5.1	5.5	5.5
28 days	6.5	6.2	6.5	6.5	6.7	6.3	5.3
	6.5	6.9	6.4	6.5	6.3	6.0	6.3
	6.4	6.8	6.5	6.9	5.8	6.9	5.2
	7.1	6.0	6.6	6.2	5.7	5.7	4.9
	6.8	5.8	6.5	6.0	6.4	6.7	*
	5.4	6.7	7.3	6.4	5.7	6.7	*
	6.6	6.9	6.8	5.8	6.2	5.9	*
	6.3	6.3	6.1	7.2	6.8	6.2	*
	6.1	6.3	6.0	6.0	6.1	5.8	*
	6.7	6.5	*	6.0	*	*	*

Wet weight in g

	Control	Nominal concentration of the test item (mg/l)					
	0	1.0	2.2	4.6	10	22	46
0 days	1.9	1.8	1.4	1.8	2.2	2.2	2.2
	1.7	2.0	2.2	2.0	1.9	1.7	1.9
	2.7	1.9	1.8	2.1	1.8	2.1	2.3
	2.0	2.2	2.1	1.8	2.2	1.8	1.8
	1.8	1.6	1.9	2.0	1.7	1.9	1.7
	1.8	2.2	1.9	2.3	2.0	2.2	2.0
	1.9	1.8	2.4	1.9	2.1	1.5	1.4
	1.4	1.6	1.7	1.7	1.4	1.8	1.5
	2.3	1.5	1.5	1.5	1.5	1.5	1.6
	1.5	1.5	1.7	1.5	1.4	1.7	1.6
28 days	2.8	2.7	2.9	2.9	3.4	2.7	1.6
	3.0	3.3	2.6	3.0	2.8	2.0	2.8
	2.7	2.9	2.7	3.5	2.1	3.5	1.2
	3.9	2.2	3.3	2.7	2.3	1.8	0.9
	3.1	2.0	2.7	2.3	3.1	3.1	*
	1.8	3.1	4.0	2.7	1.8	3.2	*
	2.9	3.2	3.0	2.0	2.4	2.2	*
	2.5	2.5	2.5	4.0	3.0	2.5	*
	2.2	2.5	2.2	2.2	2.3	1.8	*
	3.1	2.6	*	2.4	*	*	*

4.1 Examples of data analysis using hypothesis testing (NOEC determination)

NOEC Determination for Growth

Given that there are two measurements on each fish, one on day 0 (before exposure to compound) and one on day 28, this is a repeated-measures experiment. It is unfortunate that individual fish are not identified. This means there is no way to determine growth on an individual fish and statistical power may suffer as a result.. No replicate information is provided, so it is not possible to treat the replicate as sampling unit and do repeated measures analysis or paired difference analysis on the replicate means. The data cannot be analyzed therefore with repeated measures methodology. There are nonetheless at least two ways to proceed to determine the NOEC.

Method 1. Compute treatment means for each day and concentration. The response to be analyzed is the difference of the treatment means for each concentration, to obtain the mean growth from day 0 to day 28 for that treatment group. This leaves one observation per treatment group. ANOVA methods, such as Dunnett's test, are then not available, since there is no estimate of error. However, the Jonckheere-Terpstra test can still be applied, preferably in its exact permutation implementation. This approach ignores the reduced sample size in the 28-day high concentration group.

Method 2. Do a 2-factor ANOVA with day, concentration, and their interaction as the model terms. Then compare the growth (day 28 mean –day 0 mean) in each concentration to the growth in the control by standard ANOVA methods. This is not entirely correct, since it ignores the repeated measures nature of the data.

In the present case, the two lead to the same NOEC, namely the 22 mg/L concentration. Of the two, method 1 is theoretically soundest. Details are provided below.

Length

Method 1

The Step-Down Jonckheere-Terpstra test is applied, first with all concentrations, present then with the 46 mg/L group omitted. The variable DelatL is the difference of the day 28 mean minus the day 0 mean for each concentration. This was done using SAS Proc Freq. The default output includes both 1- and 2-sided tests for trend and both exact and asymptotic tests. These are all left for the reader to see, but only the 1-sided exact results are used in the discussion.

Jonckheere-Terpstra Test of DeltaL Through Conc=46

Statistics for Table of DeltaL by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-2.1268
Asymptotic Test	
One-sided Pr < Z	0.0167
Two-sided Pr > |Z|	0.0334
Exact Test	
One-sided Pr <= JT	0.0218
Two-sided Pr >= |JT - Mean|	0.0437

Since the Jonckheere-Terpstra test with all concentrations included is significant at the 0.05 level (p-value for exact 1-sided trend is=0.0218), the high concentration is omitted and the test is repeated with the remaining concentrations.

Jonckheere-Terpstra Test of DeltaL Through Conc=22

Statistics for Table of DeltaL by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-1.5302
Asymptotic Test	
One-sided Pr < Z	0.0630
Two-sided Pr > |Z|	0.1260
Exact Test	
One-sided Pr <= JT	0.0944
Two-sided Pr >= |JT - Mean|	0.1889

This test is not significant at the 0.05 level, so no further testing is required and the NOEC is 22 mg/L.

Method 2

Fish Growth Example: Length
Trout Size Data
FULL DATA SET

A 2-factor ANOVA on length with days and concentration and their interaction as model terms. First, the simple means are computed. It is apparent that there is a dose-response, perhaps beginning with the 2.2 mg/L concentration.

Obs	Conc	DeltaL	DeltaW
1	0	0.94000	1.20000
2	1	0.94000	1.10000
3	2.2	1.02222	1.27778
4	4.6	0.85000	1.17000
5	10	0.68889	0.97778
6	22	0.74444	0.93333
7	46	-0.07500	0.02500

Basic Statistics for length

Days	Conc	mean_ length	std_length	n_length
0	0	5.64	0.33065591	10
0	1	5.62	0.3190263	10
0	2.2	5.62	0.3190263	10
0	4.6	5.63	0.32335052	10
0	10	5.58	0.32930904	10
0	22	5.61	0.28460499	10
0	46	5.53	0.3591657	10
28	0	6.44	0.4575296	10
28	1	6.44	0.38355066	10
28	2.2	6.52	0.38005848	9
28	4.6	6.35	0.44284434	10
28	10	6.19	0.40756731	9
28	22	6.24	0.43620841	9
28	46	5.425	0.60759087	4

It will be observed that mortality in the high concentration has noticeably affected the sample size on day 28. The other differences in sample size are quite small and are likely to have insignificant impact on conclusions.

CovParm	Estimate
Residual	0.1428

Effect	Num DF	Den DF	FValue	ProbF	MSERR	SSQRS	SSERR
Days	1	117	84.86	<.0001	0.14282	12.1195	16.7102
Conc	6	117	3.78	0.0018	0.14282	3.2401	16.7102
Days*Conc	6	117	2.57	0.0224	0.14282	2.2045	16.7102

These are the overall F-Tests for ANOVA. It will be observed that both main effects and the interaction are significant at the 0.05 level. The CovParm above the F-tests is the pooled sample mean-squared error.

ANOVA SUMMARY STATISTCS

MODELSS	SSERR	TOTSS	RSQUARE
20.1477	16.7102	36.8579	0.54663

These are basic measures that can be used for model assessment. Since this is a linear model, the R-squared value indicates that the proportion of overall variation (55%) in the data accounted for by the model. For a biological response, an R-square this small is not unusual.

Class	Levels	Values
Days	2	0 28
Conc	7	0 1 2.2 4.6 10 22 46

Label	Estimate	StdErr	DF	tValue	Probt
Growth: 46 mg/l vs 0	-0.9050	0.2803	117	-3.23	0.0016
Growth: 22 mg/l vs 0	-0.1656	0.2423	117	-0.68	0.4958
Growth: 10 mg/l vs 0	-0.1911	0.2423	117	-0.79	0.4319
Growth: 4.6 mg/l vs 0	-0.0800	0.2390	117	-0.33	0.7384
Growth: 2.2 mg/l vs 0	0.1022	0.2423	117	0.42	0.6739
Growth: 1 mg/l vs 0	0.0200	0.2390	117	0.08	0.9335

The appropriateness of this ANOVA analysis is assessed partly through the following test for normality. This test is done on the residuals from the ANOVA. As pointed out above, this test ignores the correlations of measurements on the same fish (since fish are not identified individually). The only significant comparison is at the 46 mg/L concentration. Accordingly, the NOEC is 22 mg/L, the same as by the Jonckheere-Terpstra test.

SHAPIRO-WILK TEST OF NORMALITY OF length
Variable: Resid

Moments

N	131	Sum Weights	131
Mean	0	Sum Observations	0
Std Deviation	0.3585244	Variance	0.12853974
Skewness	-0.0318169	Kurtosis	-0.3018538
Uncorrected SS	16.7101667	Corrected SS	16.7101667
Coeff Variation	.	Std Error Mean	0.03132442

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.35852
Median	-0.01000	Variance	0.12854
Mode	0.38000	Range	1.91500
		Interquartile Range	0.51444

Fish Growth Example: Length
Trout Size Data
FULL DATA SET

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.98607	Pr < W	0.2041
Kolmogorov-Smirnov	D	0.065603	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.080864	Pr > W-Sq	0.2078
Anderson-Darling	A-Sq	0.59964	Pr > A-Sq	0.1199

This is default SAS Proc Univariate output, so four tests for normality are shown. It is highly recommended that each Lab select one of these tests and use it as *the* formal test of normality. Alternatively, a Lab can develop a set of rules that indicate when to use which test. In either case, care should be taken so as not to be guilty of selecting the statistical result desired. In the present case, all tests yield the same conclusion. According to the Shapiro-Wilk test, the data are consistent with normality.

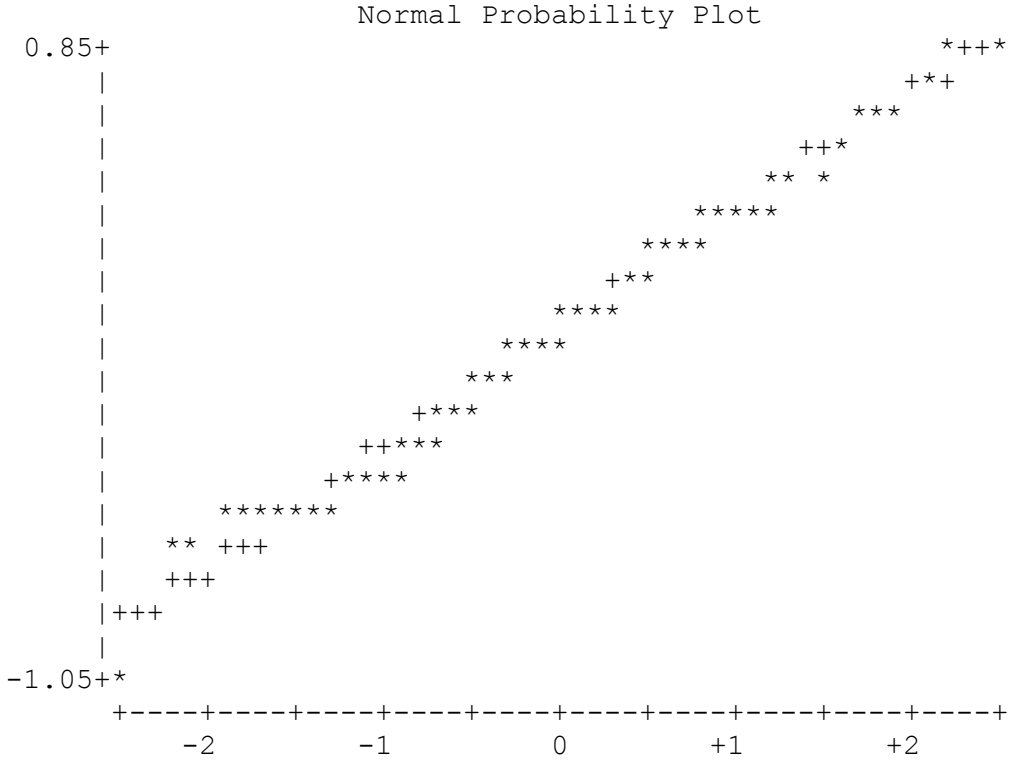
Variable: Resid

Quantile	Estimate
100% Max	0.875000
99%	0.850000
95%	0.550000
90%	0.455556
75% Q3	0.270000
50% Median	-0.010000
25% Q1	-0.244444
10%	-0.510000
5%	-0.530000
1%	-0.640000
0% Min	-1.040000

Stem	Leaf	#	Boxplot
8	58	2	
7	8	1	
6	166	3	
5	15	2	
4	2666677	7	
3	22666677788889	14	
2	1666677889	10	+-----+
1	125568889	9	| |
0	12256666677788899	17	| + |
-0	98844443322222211	17	*-----*
-1	5444322221	10	| |
-2	4443322	7	+-----+
-3	955544433	9	
-4	998443221	9	
-5	85433222221	11	
-6	44	2	
-7			
-8			
-9			
-10	4	1	0

Multiply Stem.Leaf by 10**-1

Fish Growth Example: Length
Trout Size Data
SHAPIRO-WILK TEST OF NORMALITY OF length
FULL DATA SET
Variable: Resid



POSSIBLE OUTLIERS FROM ANOVA ON length

Obs	Days	Conc	Length	Pred	Resid	LB	UB
1	28	0	5.4	6.44000	-1.04	-1.01611	1.04167

LEVENE TEST FOR length

Effect	DF	LEVENE	P_VALUE
Days*Conc	6	0.17219	0.98381

By Levene's test, there is no reason to reject the hypothesis that the within-group variances are equal. A standard ANOVA can be done.

Weight

Method 1

The Step-Down Jonckheere-Terpstra test is applied, first with all concentrations present, then with the 46 mg/L group omitted. The variable DelatL is the difference of the day 28 mean minus the day 0 mean for each concentration. This was done using SAS Proc Freq. The default output includes both 1- and 2-sided tests for trend and both exact and asymptotic tests. These are all left for the reader to see, but only the 1-sided exact results are used in the discussion. It will be observed, however, that for these data, the asymptotic and exact methods do not lead to the same NOEC.

```
Jonckheere-Terpstra Test of DeltaW Through Conc=46

      Statistics for Table of DeltaW by Conc

            Jonckheere-Terpstra Test
       ------------------------------------
       Statistic (JT)                3.0000
       Z                            -2.2528

      Asymptotic Test
       One-sided Pr <  Z             0.0121
       Two-sided Pr > |Z|            0.0243

       Exact Test
       One-sided Pr <=  JT           0.0151
       Two-sided Pr >= |JT - Mean|   0.0302
```

Since the Jonckheere-Terpstra test with all concentrations included is significant at the 0.05 level, the high concentration is omitted and the test is repeated with the remaining concentrations.

```
Jonckheere-Terpstra Test of DeltaW Through Conc=22

      Statistics for Table of DeltaW by Conc

            Jonckheere-Terpstra Test
       ------------------------------------
       Statistic (JT)                3.0000
       Z                            -1.6908

       Asymptotic Test
       One-sided Pr <  Z             0.0454
       Two-sided Pr > |Z|            0.0909

       Exact Test
       One-sided Pr <=  JT           0.0681
       Two-sided Pr >= |JT - Mean|   0.1361
```

The exact test is not significant at the 0.05 level, so no further testing is required and the NOEC is 22 mg/L. Note, however, that if standard asymptotic methods were used, the test would continue one step further, as shown below. There is no general expectation that asymptotic methods are more (or less)

sensitive or powerful than exact methods. A Lab should decide on rules for the use of exact and asymptotic methods and then use whatever conclusion follows.

Jonckheere-Terpstra Test of DeltaW Through Conc=10
Statistics for Table of DeltaW by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-0.9798
Asymptotic Test	
One-sided Pr < Z	0.1636
Two-sided Pr > |Z|	0.3272
Exact Test	
One-sided Pr <= JT	0.2417
Two-sided Pr >= |JT - Mean|	0.4833

Method 2

Fish Growth Example: Weight
Trout Size Data
FULL DATA SET

A 2-factor ANOVA on length with days and concentration and their interaction as model terms. First, the simple means are computed. It is apparent that there is a dose-response, perhaps beginning with the 2.2 mg/L concentration.

Obs	Conc	DeltaL	DeltaW
1	0	0.94000	1.20000
2	1	0.94000	1.10000
3	2.2	1.02222	1.27778
4	4.6	0.85000	1.17000
5	10	0.68889	0.97778
6	22	0.74444	0.93333
7	46	-0.07500	0.02500

Basic Statistics for length

Days	Conc	mean_ length	std_length	n_length
0	0	5.64	0.33065591	10
0	1	5.62	0.3190263	10
0	2.2	5.62	0.3190263	10
0	4.6	5.63	0.32335052	10
0	10	5.58	0.32930904	10
0	22	5.61	0.28460499	10
0	46	5.53	0.3591657	10
28	0	6.44	0.4575296	10
28	1	6.44	0.38355066	10
28	2.2	6.52	0.38005848	9
28	4.6	6.35	0.44284434	10
28	10	6.19	0.40756731	9
28	22	6.24	0.43620841	9
28	46	5.42	0.60759087	4

It will be observed that mortality in the high concentration has noticeably affected the sample size on day 28. The other differences in sample size are quite small and are likely to have insignificant impact on conclusions.

CovParm	Estimate
Residual	0.1988

Effect	Num DF	Den DF	FValue	ProbF	MSERR	SSQRS	SSERR
Days	1	117	79.24	<.0001	0.19877	15.7509	23.2566
Conc	6	117	3.29	0.0050	0.19877	3.9250	23.2566
Days*Conc	6	117	2.62	0.0203	0.19877	3.1254	23.2566

These are the overall F-Tests for ANOVA. It will be observed that both main effects and the interaction are significant at the 0.05 level. The CovParm above the F-tests is the pooled sample mean-squared error.

ANOVA SUMMARY STATISTCS

MODELSS	SSERR	TOTSS	RSQUARE
26.1387	23.2566	49.3953	0.52917

These are basic measures that can be used for model assessment. Since this is a linear model, the R-squared value indicates the proportion of overall variation in the data accounted for by the model. For a biological response, an R-square this small is not unusual.

Fish Growth Example: Weight
Trout Size Data
CLASS LEVEL INFORMATION
FULL DATA SET

Class	Levels	Values
Days	2	0 28
Conc	7	0 1 2.2 4.6 10 22 46

Fish Growth Example: Weight
Trout Size Data
TESTS OF LINEAR CONTRASTS
FULL DATA SET

Label	Estimate	StdErr	DF	tValue	Probt
Growth: 46 mg/l vs 0	-1.0750	0.3306	117	-3.25	0.0015
Growth: 22 mg/l vs 0	-0.2067	0.2859	117	-0.72	0.4712
Growth: 10 mg/l vs 0	-0.1422	0.2859	117	-0.50	0.6198
Growth: 4.6 mg/l vs 0	0.01000	0.2820	117	0.04	0.9718
Growth: 2.2 mg/l vs 0	0.1178	0.2859	117	0.41	0.6811
Growth: 1 mg/l vs 0	-0.01000	0.2820	117	-0.04	0.9718

The appropriateness of this ANOVA analysis is assessed partly through the following test for normality. This test is done on the residuals from the ANOVA. As pointed out above, this test ignores the correlations of measurements on the same fish (since fish are not identified individually). The only significant comparison is at the 46 mg/L concentration. Accordingly, the NOEC is 22 mg/L, the same as by the exact Jonckheere-Terpstra test.

Fish Growth Example: Weight
Trout Size Data
SHAPIRO-WILK TEST OF NORMALITY OF Weight
FULL DATA SET

Variable: Resid
Moments

N	131	Sum Weights	131
Mean	0	Sum Observations	0
Std Deviation	0.42296218	Variance	0.17889701
Skewness	0.50903274	Kurtosis	0.53166703
Uncorrected SS	23.2566111	Corrected SS	23.2566111
Coeff Variation	.	Std Error Mean	0.03695438

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.42296
Median	-0.03333	Variance	0.17890
Mode	-0.73333	Range	2.23000
		Interquartile Range	0.54000

Tests for Normality

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	W	0.979035	Pr < W	0.0403
Kolmogorov-Smirnov	D	0.072519	Pr > D	0.0897
Cramer-von Mises	W-Sq	0.090825	Pr > W-Sq	0.1493
Anderson-Darling	A-Sq	0.632637	Pr > A-Sq	0.0978

The Shapiro-Wilk test is significant at the 0.05 level. A examination of the QQ-plot and stem-and-leaf plot below, as well as the p-value being just below 0.05 may suggest the violation of normality is minor. Also, while it is not included here to keep the text to a minimu, if the outliers idenitified below are omitted and the resulting dataset is re-analyzed, the NOEC is the same and the significant Shapiro-Wilk test is eliminated. Altogether, this information suggests the present analysis can be accepted.

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.2300000
99%	1.1750000
95%	0.8000000
90%	0.5000000
75% Q3	0.2400000
50% Median	-0.0333333
25% Q1	-0.3000000
10%	-0.4777778
5%	-0.7000000
1%	-0.7777778
0% Min	-1.0000000

Fish Growth Example: Weight
Trout Size Data
FULL DATA SET

Variable: Resid

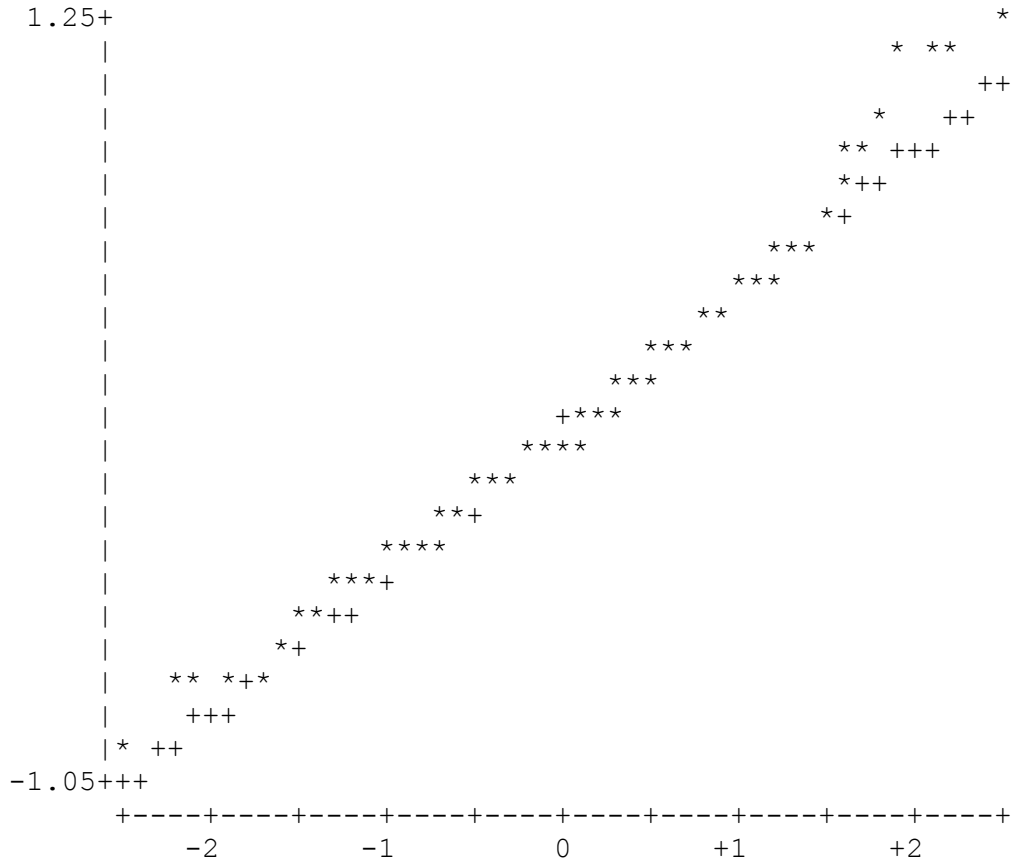
Stem	Leaf	#	Boxplot
12	3	1	0
11	028	3	0
10			
9	7	1	|
8	02	2	|
7	3	1	|
6	07	2	|
5	00247	5	|
4	000224	6	|
3	004668899	9	|
2	000234468	9	+-----+
1	0002344789	10	| |
0	000002444689	12	| + |
-0	776664432211	12	*-----*
-1	88866644200000	14	| |
-2	8881100000	10	| |
-3	8766644321100	13	+-----+
-4	87622200	8	|
-5	7300	4	|
-6	80	2	|
-7	873320	6	|
-8			|
-9			|
-10	0	1	|

Multiply Stem.Leaf by 10**-1

Fish Growth Example: Weight
Trout Size Data
SHAPIRO-WILK TEST OF NORMALITY OF Weight
FULL DATA SET

The UNIVARIATE Procedure
Variable: Resid

Normal Probability Plot



Fish Growth Example: Weight
Trout Size Data
POSSIBLE OUTLIERS FROM ANOVA ON Weight
FULL DATA SET

Obs	Days	Conc	Weight	Pred	Resid	LB	UB
1	28	0	3.9	2.80000	1.10000	-1.11	1.05
2	28	2.2	4	2.87778	1.12222	-1.11	1.05
3	28	4.6	4	2.77000	1.23000	-1.11	1.05
4	28	46	2.8	1.62500	1.17500	-1.11	1.05

LEVENE TEST FOR Weight

Effect	DF	LEVENE	P_VALUE
Days*Conc	6	0.46436	0.83347

By Levene's test, there is no reason to reject the hypothesis that the within-group variances are equal. A standard ANOVA can be done.

4.2 Example of data analysis by dose response modelling

It is assumed that an EC10 is required.
According to the flow chart of chapter 6, the dose-response data should include more than two concentration groups with different response levels. The dose-response data of this data set (for both lengths and weights) do not clearly comply with this requirement (see figures below). However, in view of the relatively large number of concentration groups, and the general trend of decreasing response with concentration, the data might nonetheless be suitable for deriving an EC10. In accordance with the general recommendation of chapter 6, various models will be fitted to the data.

Dose-response analysis for fish weight
In fitting the dose-response model, the weight data will be assumed to be log-normally distributed. Hence, the model is fitted to the data on log-scale, i.e. both the data and the model prediction is log-transformed. Note, however, that the models used describe the response variable on the original scale as a function of dose.

First the nested nonlinear model proposed by Slob (2002) is fitted. This results in the following log-likelihoods:

model 1	$y = a$	loglik = -5.21
model 2	$y = a \exp(x/b)$	loglik = 4.85
model 3	$y = a \exp(\pm(x/b)^d)$	loglik = 6.53
model 4	$y = a\,[c - (c - 1) \exp(-x/b)]$	loglik = 4.85
model 5	$y = a\,[c - (c - 1) \exp(-(x/b)^d)]$	loglik = 6.53

The log-likelihoods can be compared with the likelihood-ratio test. A model with one more parameter fits the data significantly better (at $\alpha = 0.05$) than the model without that parameter when the increase in the log-likelihood is greater than 1.92. The difference in log-likelihoods between model 2 and model 1 is 10.06, so there is no doubt that the data show a significant dose-response. While model 3 results in a higher log-likelihood than model 2, the difference is not significant, and it may be concluded that model 2 is from this family of models the appropriate one for describing the data. Fig. 1 shows the results for this model.

Next a polynomial model is fitted to these data. Since this is again a nested family of models, the log-likelihoods can be compared by the ratio-likelihood test:

model 1	$y = a$	loglik = -5.21
model 2	$y = a + bx$	loglik = 5.55
model 3	$y = a + bx + cx^2$	loglik = 6.58
model 4	$y = a + bx + cx^2 + dx^3$	loglik = 6.69

Here, model 2 (straight line) is not significantly improved by higher order polynomials, and this model may be selected from this family of models.

Finally, the power model, $y = c + ax^b$, is fitted to the weights. This model results in a log-likelihood value of 6.57. However, by fixing the parameter b to unity, the model reduces to a straight line. Hence, the power model and the straight line are nested, and their log-likelihoods can be compared. Since the straight line resulted in a log-likelihood of 5.55, the power model does not give a significantly better fit.

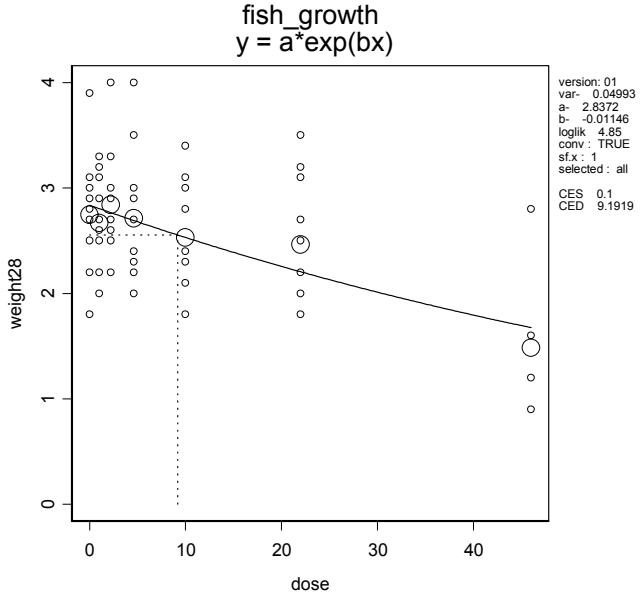
None of the three-parameter models gives a significantly better fit than its two-parameter counterpart. As Table 1 shows, both two-parameter models give similar results. The table also shows the results for two models with three parameters, even though they did not result in a significantly better fit than the associated two-parameter model. As discussed in chapter 6, a model that contains too many parameters

(i.e. in relation to the data) may results in wider confidence intervals for the EC10. This is clearly illustrated in this case.

It may be concluded that both two-parameter models give similar results. Also, the confidence intervals are similar. Hence, although this particular data set only shows two clearly different response levels, the data apparently contain sufficient information to warrant the estimation of the EC10. This may be explained by the relatively large number of concentrations tested, most of which show no or only a small response. Thus, the estimated EC10 is sufficiently supported by surrounding concentrations that have been tested.

Table 1. Summary of results for exponential model and straight-line model (in bold), both having two parameters. Also included are two models with three parameters (not in bold). Note that both the first two and the second two models form a nested couple.

Model	Log-lik	EC10	90%-CI
$\boldsymbol{y = a \exp(x/b)}$	**4.85**	**9.19**	**6.86 – 13.94**
$y = a \exp(x/b^c)$	6.53	21.08	9.89 – 41.39
$\boldsymbol{y = a + bx}$	**5.55**	**10.55**	**8.56 – 14.25**
$y = a + bx^c$	6.57	20.70	8.65 – 34.34



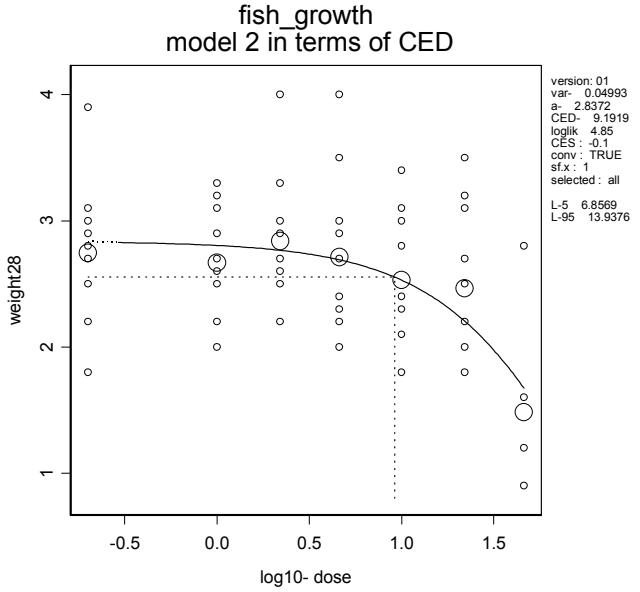


Fig. 1. Exponential model, y=a exp(bx), fitted to weights at 28 days. Upper panel: concentration on x-axis, lower panel: log10-concentration on x-axis (to improve visibility). CED = EC10. Note: the model was re-parameterized here to make the EC10 a parameter in the model (instead of *b*). In this way the confidence interval (L-5, L-95) for the EC10 can be calculated by the likelihood profile method. Large marks: geometric means.

Assumptions

To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways. Here, two plots will be considered. One is the so-called QQ- plot, where the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with intercept equal to the mean of the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.
In interpreting a QQ-plot one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.

In fig. 2 the regression residuals resulting from the analysis on log-scale are plotted. As the left panel shows, the data did comply with the assumption of log-normality. The residuals plotted against concentration do not reveal a clear trend, and the assumption of homogeneous variances (on log-scale) appears acceptable (see right panel of Fig. 2).

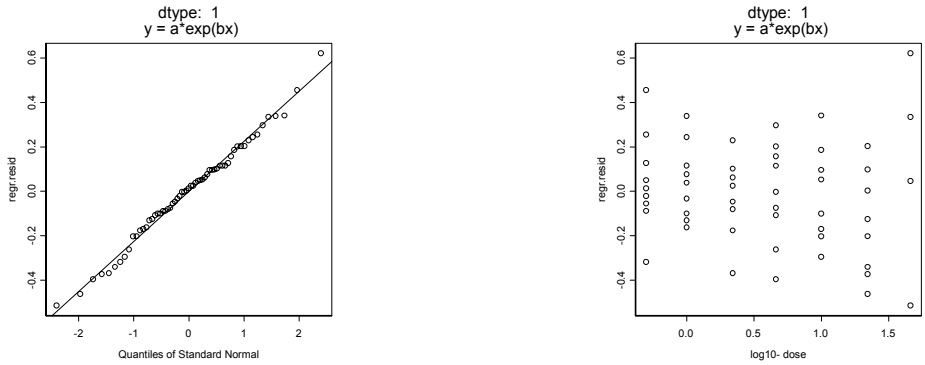


Fig. 2. Left panel: QQ-plot (for normal distribution) of the regression residuals for the analysis of Fig. 1, where the model was fitted on log-scale. Right panel: the same residuals plotted against dose (dose on log-scale to improve visibility).

Although not strictly needed, the residual plots are also shown for an analysis where the log-transformation was omitted (see Fig. 3). The QQ-plot now shows a slight curvature, which confirms the use of a log-transformation. It should be noted that the scatter in these data is relatively mild (CV = 23%), and that a log-normal distribution gets closer to a normal distribution with smaller variation (CV). Therefore, the smaller the scatter in the data, the more data are needed to see any difference in the QQ-plots assuming normality or log-normality. For the same reason, it may be expected that applying or omitting the log-transformation has no large impact on the results of the analysis when the scatter in the data is relatively small.

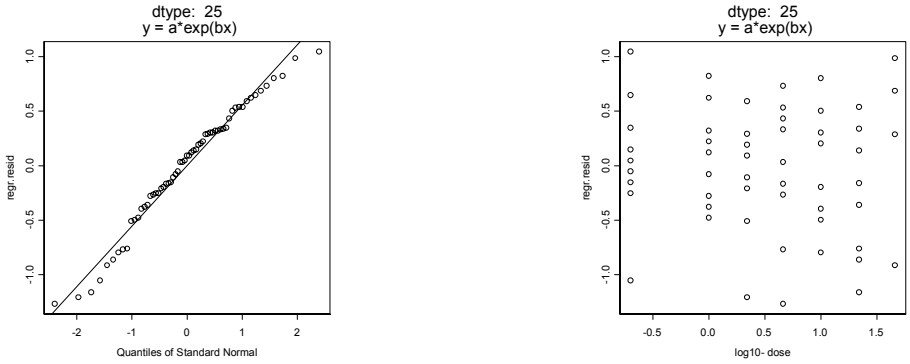


Fig. 3. Left panel: QQ-plot (normal distribution) of the regression residuals for an analysis as in Fig. 1, where the model was fitted without transformation. Right panel: the same residuals plotted against dose (dose on log-scale to improve visibility).

The analysis shown in Fig. 1 was repeated but now omitting the log-transformation (see Fig. 4). The results are somewhat different, especially with regard to the upper confidence limit for the EC10 (see Table 2).

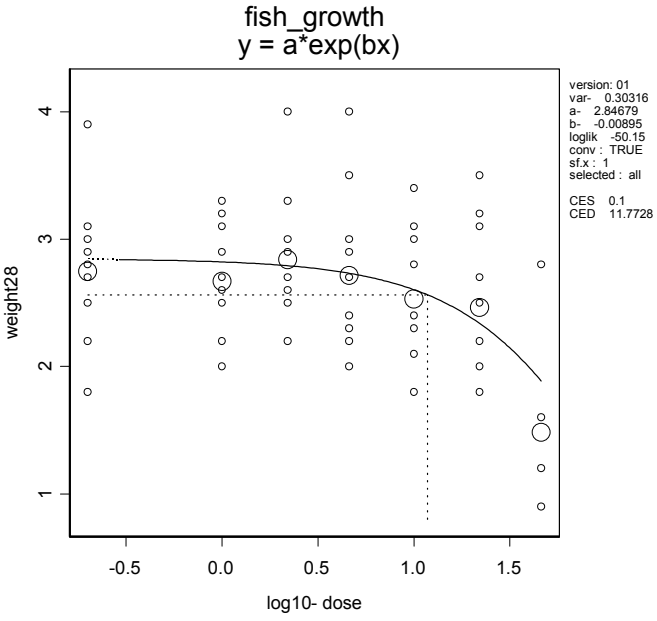


Fig. 4. Dose-response analysis of the fish weights, but without log-transformation. Large marks: arithmetic means.

Table 2. Summary of results for the exponential model fitted to the log-tranformed weights, and the untransformed weights.

	EC10	90%-CI
Analysis with log-transformation	9.19	6.86 – 13.94 [1)]
Analysis without transformation	11.78	7.79 – 21.94 [1)]

1) Obtained by the likelihood profile method

Dose-response analysis for fish length

The analysis of the length data is very similar to that of the weight data. Fig. 5 shows the exponential model fitted to the lengths (assuming lognormality). The EC10 is estimated at 31.90 mg/l, as compared to 9.19 mg/l for the weights (same model fitted). Therefore, the weight data would most likely be used for deriving an EC10. However, one might argue that a decrease in body length by 10% is not equivalent to a decrease in body weight by 10% from a biological point of view.

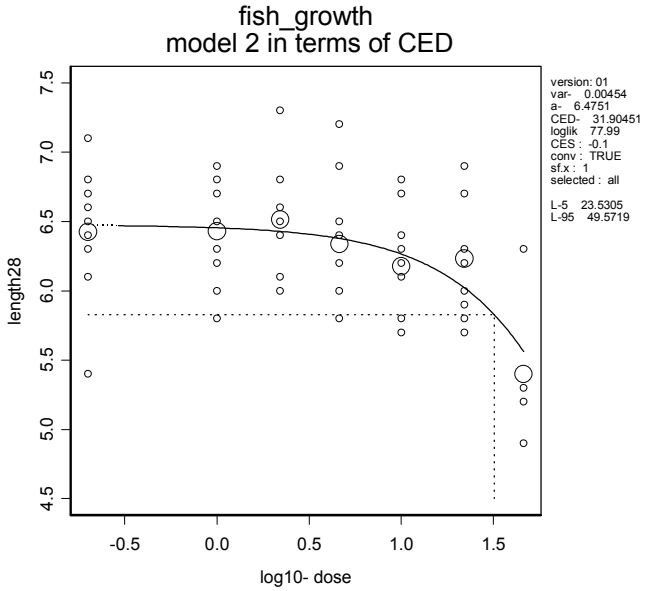


Fig. 5. Exponential model fitted to the body lengths. Large marks: geometric means.

4.3 Examples of data analysis using Debtox (biological methods)

Data: Mean initial volumetric length of *Oncorhynchus mykiss* is 1.222 $g^{1/3}$

Time: d;	Conc: mg/l; Response: mean volumetric length, $g^{1/3}$						
	0	1	2.2	4.6	10	22	46
21	1.403	1.389	1.418	1.398	1.365	1.355	1.152

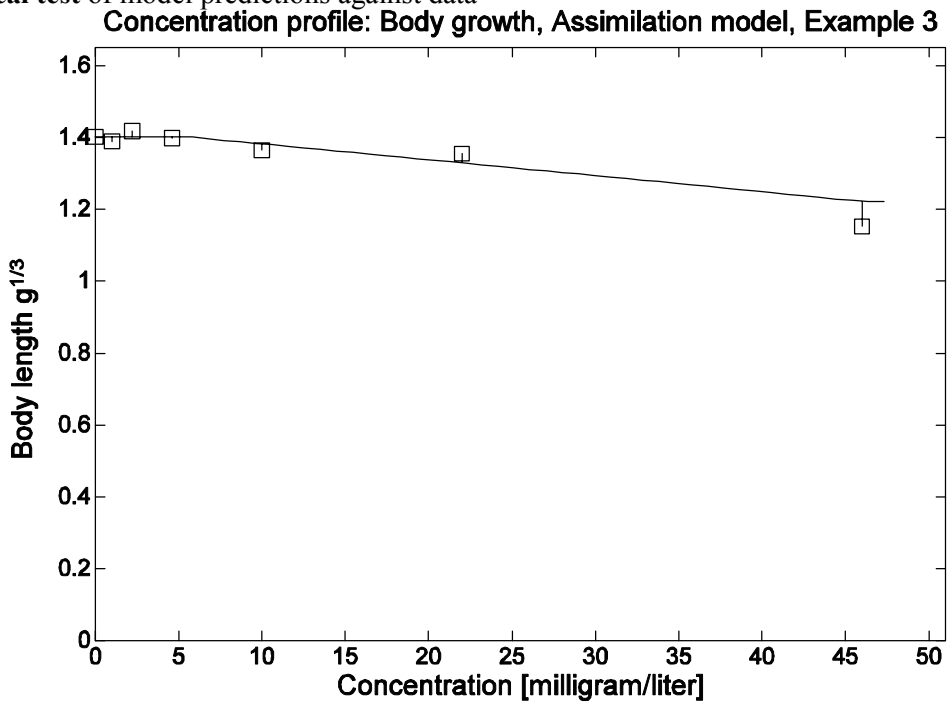
Parameter estimates and Asymptotic Standard Deviations (ASD)

No effect concentration	5.597	mg l^{-1}	7.399		
Blank ultimate length	15.84	$g^{1/3}$	1.221	-0.456	
Tolerance concentration	43.78	mg l^{-1}	11.884	-0.818	0.284
Elimination rate	Infinity	d^{-1}			
Initial length	1.222	$g^{1/3}$			
Von Bertalanffy growth rate	0.00059	d^{-1}			
Energy investment ratio	1				
Mean deviation	0.03006	$g^{1/3}$			

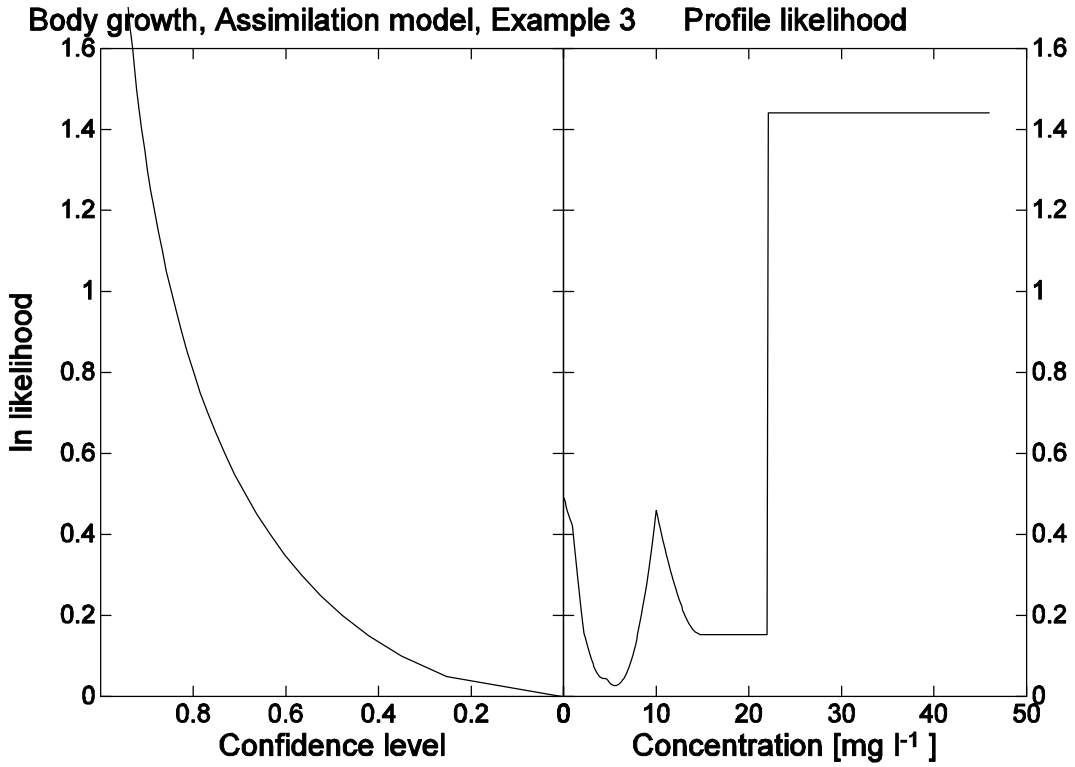
ECx values (derived from parameter values) in mg/l.

Day	EC0	ASD	EC10	ASD
21		7.4	37.4	5.71

Graphical test of model predictions against data



Profile likelihood for NEC estimate



Comments

The mean body size at the highest concentration was lower than the initial one; the increase in weight during 21 d in the blank was very small. The model for effects on assimilation fitted slightly better than for effects on maintenance or growth costs. The von Bertalanffy growth rate was fixed at 5.9 10^{-4} d (from Kooijman, 2000), and the initial size at 1.222 $g^{1/3}$ (measured mean value). The EC50 is on body length, which is not meaningful in this case since 50% of the blank length is far below the initial length. The NEC value does not differ significantly from zero, with a 95% confidence interval of 0-22 mg/l.

ANNEX 5

5.1 Description of Selected Methods for Use with Quantal Data

The Cochran-Armitage trend test - Quantal (binary) data can be collected and categorized by explanatory factors (such as dosage or treatment level). An analysis of such data usually tries to indicate relationships between the response (binary) variable and factors such as dose level. In such cases, the Pearson Chi-Square (χ^2) test for independence can be used to find if any relationships exist. The Cochran-Armitage test decomposes the Pearson Chi-Square test into a test for linear trend for the dose-response and a measure of lack of monotonicity, $\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$ where $\chi^2_{(1)}$ is the *1* df calculated Cochran-Armitage linear trend statistic and $\chi^2_{(k-2)}$ is *k-2* df Chi-Square test statistic for lack of monotonicity.
Suppose the number of affected individuals is Y_i within a group of N_i animals exposed to dose X_i. The proportion affected by dosage X_i is $p_i = Y_i/N_i$. The model is $p_i = H(\alpha + \beta z_i)$, where z_i is the dose-metric (e.g., log-dose or dose rank), and H is some twice-differentiable, monotone link function such as the logistic. The test for linear trend is a test for $\beta=0$. Standard weighted regression gives the estimate of β as

$$b = \frac{\sum_{i=1}^{k} n_i (p_i - \bar{p})(z_i - \bar{z})}{\sum_{i=1}^{k} n_i (z_i - \bar{z})^2},$$

where $\bar{z}$ is the weighted mean of the z_i, $\bar{p}$ is the weighted sum of the p_i, and $\bar{q}$ is the weighted sum of the $q_i = 1- p_i$. The test statistic for $\beta = 0$ is $b^2/ \mathrm{Var}(b)$ which is approximately distributed as $\chi^2_{(1)}$. Formally written, the Cochran-Armitage Chi-square is:

$$\chi_1^2 = \frac{\left(\sum_{i=1}^{k} y_i z_i - \frac{T_y \left(\sum_{i=1}^{k} N_i z_i \right)}{T} \right)^2}{\bar{p}\bar{q} \left(\sum_{i=1}^{k} N_i z_i^2 - \frac{\left(\sum_{i=1}^{k} N_i z_i \right)^2}{T} \right)},$$

where $T_y = \Sigma Y_i$ and $T = \Sigma N_i$.
The Cochran-Armitage Chi-square can also be expressed as a z-statistic in order to take account of the direction of the trend. The z-statistic is obtained from the formula given by removing the exponent 2 in the numerator and taking the square-root of the denominator. This z-test has a standard normal distribution under the null hypothesis of no trend, and the probability of the z –statistic can be obtained from a table of areas under the standard normal distribution. Only the z-statistic is appropriate for 1-sided tests. Unlike the 1-sided test, the χ^2 version of the Cochran-Armitage test can remain significant in a step-down application even when there is a change in direction of the trend. To avoid this situation when doing a 2-sided test, one

applies both 1-sided z-tests with all doses present at the $\alpha/2$ level. At most one of these can be significant. If one is significant, this determines the direction of the trend and all further tests are done with the z-statistic for that same direction at the $\alpha/2$ level.

A general linear trend model for quantal data is

$$p_i = H(a + bd_i),$$

where a and b are parameters to be estimated, d_i is some metric (measure) of the exposure level (e.g., dose, concentration, log concentration), p_i is the probability of response, and H is some monotone function, (referred to as a link function), e.g.,

logistic, $H_1(u) = e^u/(1 + e^u)$,
probit, $H_2(u) = M(u)$,
extreme value, $H_3(u) = 1 - \exp(e^u)$,
one-hit, $H_4(u) = 1 - e^{-u}$.

For example, the trend model for the one-hit link function could be written as

$p_i = 1 - \exp(-(a + bd_i))$.

Tarone and Gart (1980) showed that for any link function likely to be of practical use, the same test statistic for significance of trend always arises from likelihood-type considerations, namely, the Cochran-Armitage test. Thus, it is not necessary to postulate the particular form of the link function. They also showed that this test is in general the most efficient test of trend for any monotone model. This is one of the reasons some regard the Cochran-Armitage test as inherently non-parametric. Hirji and Tang (1998) discuss the favourable power properties of the Cochran-Armitage test compared to various alternatives. They offer evidence that this test can be used for small samples and sparse data without undue concern that it is an asymptotic test.

Assumptions: Subjects are independent within and among groups and subgroups (if present). Group proportions are monotonic with respect to the dose score. While formally, this is a test for linear trend in the response in relation to the dose score used, it is generally the most powerful test against any monotone dose-response alternative. Furthermore, rank-order of doses (or equally-spaced dose scores) is used to reduce dose-dependence in the test. Note: Robust versions of this test are available which allow for extra-binomial variation, notably one based on a beta-binomial model.

Power: Extensive power simulations of the step-down Cochran-Armitage test have demonstrated that in every instance considered where there is a monotone dose-response, the step-down application of the Cochran-Armitage test is more powerful than Fisher's exact test. These simulations followed the step-down process to the NOEC determination, and covered a range of dose-response shapes, thresholds, background rates, number of treatment groups and number of subjects per group. These simulation results are useful in design of experiments and are being prepared for publication by J. W. Green. A small sample giving an idea of the results will be found in Annex 5.2. These should be useful in experimental design.

Confidence Intervals for Incidence Rate: Given the discussion above of Tarone and Gart (1980), confidence intervals for the true incidence rate at a given tested concentration can be calculated from one of the regression models described in chapter 6 for quantal data. There is no direct link between the Cochran-Armitage test statistic and these confidence intervals.

Example: The data are from a trout early life stage experiment. Initially, there were 20 eggs placed in each of four replicate subgroups per concentration. Of those eggs that hatched (shown as the *Number at risk* value), some larvae did not survive to the time of thinning. The analysis is of the larval survival to time of thinning. There were two control groups, one water-only and the other including a solvent that is also present in all positive dose groups. There were no mortalities in either control, so they were combined for further analysis.

Dose (PPM)	Number at risk	Number Responding	% Responding	Dose Score
0	125	0	0.0	1
1	62	1	1.6	2
2	62	1	1.6	3
4	60	2	3.3	4
8	65	0	0.0	5
16	72	10	13.9	6
32	65	29	44.6	7

Cochran-Armitage Test Using Equally Spaced Dose Scores
Cochran-Armitage test is one-sided for INCREASE in RESPONSE

All doses included

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	140.23874	6	0	
Trend	8.7567722	1	0	**
LOF	63.55768	5	2.231E-12	

32 PPM concentration omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	34.479817	5	1.9107E-6	
Trend	4.1393854	1	0.0000174	**
LOF	17.345306	4	0.001656	

32 & 16 PPM concentrations omitted

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	5.3062133	4	0.2572958	
Trend	0.7720901	1	0.2200305	
LOF	4.7100902	3	0.1942989	

No further testing is required. NOEL is current high concentration, 8 mg/L.
There were actually four replicate subgroups in each concentration in this experiment, which this analysis did not take into account. The data, broken down by replicate subgroup, are as follows. With the possible exception of the high concentration, there is no significant evidence of extra binomial variation in these

data. While it is clear that by any reasonable analysis, the high dose group will be an effects level, and there is no reason evident in the data to suggest a need to model replicate subgroups, it may be illustrative to do such an analysis. For this purpose, we will analyse subgroup proportions, which are reported below. With the large number of zeros, an analysis based on the normality of the proportion dead, even after an arc-sine square-root or Freeman-Tukey transformation is not justified. Rather, a Jonckheere-Terpstra test will be done. While this ignores differences in sample sizes among the subgroups, these differences are small, so little precision will be lost by such an analysis. Both the Jonckheere-Terpstra and Dunn tests find the NOEC to be dose 5, or 8 mg/L, the same as the Cochran-Armitage test.

```
Dunn's Multiple Comparisons (Increasing) on DEADPROP

DOSE COUNT N0  MRANK ABS_DIFF  CRIT05  CRIT01  SIGNIF P_VAL

 0     8    8  10.500   0.000   9.7600  11.9665       .
 1     4    8  13.250   2.750  11.9535  14.6559      1.000
 2     4    8  13.750   3.250  11.9535  14.6559      1.000
 4     4    8  16.500   6.000  11.9535  14.6559      0.688
 8     4    8  10.500   0.000  11.9535  14.6559      1.000
16     4    8  26.625  16.125  11.9535  14.6559  **  0.004
32     4    8  30.375  19.875  11.9535  14.6559  **  0.000


Step-Down Jonckheere-Terpstra test

          MONOTONICITY CHECK OF DEADPROP

              PARM            P_T    SIGNIF


            DOSE TREND      0.0001      **
            DOSE QUAD       0.0001      **


KEY
ZC IS JONCKHEERE STATISTIC COMPUTED WITH TIE CORRECTION
ZCCF IS ZC WITH CONTINUITY CORRECTION FACTOR
P1UPCF IS P-VALUE FOR UPWARD TREND
P1DNCF IS P-VALUE FOR DOWNWARD TREND
P-VALUES ARE FOR TIE-CORRECTED TEST WITH CONTINUITY CORRECTION FACTOR
SIGNIF RESULTS ARE FOR AN INCREASING ALTERNATIVE HYPOTHESIS

Hi_Dose JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
32     332   4.4091617 4.4281667 4.7519E-6 0.9999943 **
16     220.5 3.0749658 3.1003788 0.0009664 0.9988541 **
 8     124.5 0.9908917 1.0305274 0.1513813 0.8292628
```

Raw Data for Dunn and Jonckheere-Terpstra Examples

TOTRISK is the number of larvae exposed. DEADLARV is the number of exposed larvae found dead by the end of the experimental period. PROPDEAD = DEADLARV/TOTRISK. REP is an identifier for a replicate subgroup within a given concentration (CONC).

CONC	REP	TOTRISK	DEADLARV	PROPDEAD
0	1	16	0	0
0	2	16	0	0
0	3	16	0	0
0	4	15	0	0
0	5	16	0	0
0	6	15	0	0
0	7	15	0	0
0	8	16	0	0
1	1	16	1	.0625
1	2	15	0	0
1	3	15	0	0
1	4	16	0	0
2	1	16	0	0
2	2	15	1	.065
2	3	16	0	0
2	4	15	0	0
4	1	14	0	0
4	2	15	1	.065
4	3	15	0	0
4	4	16	1	.065
8	1	15	0	0
8	2	16	0	0
8	3	17	0	0
8	4	17	0	0
16	1	17	2	.1176
16	2	17	3	.1765
16	3	18	2	.1111
16	4	20	3	.15
32	1	15	8	.5333
32	2	17	3	.1765
32	3	15	8	.5333
32	4	18	10	.5556

Fishers exact test - Fisher's Exact test is based on a 2x2 contingency table where control and a single treatment group are compared according to their prospective counts (Affected/Not affected). The diagram below illustrates this case.

	Control	Treatment	Total
Affected	n00	n01	n0.
Not Affected	n10	n11	n1.
Total	n.0	n.1	n..

Fisher's exact test is based on the probability of observing n_{01} affected subjects in the treatment group, if all marginal totals are considered fixed. This probability is given by the hypergeometric distribution: The stated probability of observing $n_{01.}$ affected subjects in the treatment group, given that $n_{0.}$ subjects were affected overall and a total of n.. subjects were in both groups combined is

$$P(x = n_{01}) = \frac{\binom{n_{.0}}{n_{0.} - n_{01}}\binom{n_{.1}}{n_{01}}}{\binom{n_{..}}{n_{0.}}}.$$

From this, of course, the probability of observing at least n_{01}subjects in the treatment group is

$$P(x \geq n_{01}) = \sum_{i=n_{01}}^{n_{0.}} P(X = i).$$

This gives the significance level of the observed response for a 1-sided test of an increase in the treatment group. Fisher's exact test can be applied to compare each treatment group to the control, independently of all other treatment to control comparisons. When this is done, a Bonferroni-Holm adjustment for the number of comparisons being made can be applied to control the over-all false positive rate.
Power: The power of Fisher's exact test is available in several software packages, including StatXact 4, Pass, and Study Size. The second of these can be found at http://www.Dataxion.com. The third is available at CreoStat@StudySize.com. A simple search of the internet will locate several such sites, some of which have free down loads. It is important to understand that the power of Fisher's exact test depends on the background (i.e., control) incidence rate, with power decreasing as background rate increases (up to 50%), when all other factors are fixed.

Assumptions: Subjects are independent within and among groups and subgroups (if present).
Confidence Intervals: Since no dose response relationship is assumed in using Fisher's exact test, only a crude confidence interval based on the binomial distribution or its normal approximation can be constructed for the incidence rate at a tested concentration.

Example: The same data will be analysed as for the Cochran-Armitage example. In this instance, Fisher's exact test reports the same NOEC as the tests given above.

FISHER EXACT TEST vs CONTROL FOR DEADLARV OBSERVED
TESTING FOR AN INCREASING ALTERNATIVE HYPOTHESIS

Dose	Number at risk	Number Responding	P-value (Right)	Significance
1	62	1	0.33155	
2	62	1	0.33155	
4	60	2	0.10400	
8	65	0	1.00000	
16	72	10	0.00003	**
32	65	29	0.00000	**

Poisson tests - Similar to the context of the Cochran-Armitage test, quantal (binary) data can be collected and categorized by explanatory factors (such as dosage or treatment level). An analysis of such data usually tries to indicate relationships between the response (binary) variable and factors. The counts within subgroups are assumed to follow a Poisson distribution whose mean is a function, usually linear, of some dose metric. The use of equally spaced dose scores (i.e., rank-order) makes this more a test for trend than raw doses do, when there are large differences in doses. An advantage of the Poisson model over the Jonckheere for quantal data is the use of a rate multiplier to adjust for unequal sample sizes.

Suppose the number of affected individuals is Y_{ij} within subgroup j of group i, a group of n_{ij} animals exposed to the i-th dose with dose score z_i. The proportion affected by this dosage is $p_i = Y_i / n_{ij}$. The model is determined by

$$p[Y_{ij} = y] = \frac{e^{-\mu_i}\mu_i^y}{y!},$$

where μ_i = is a function of dose.

For the trend version of the Poisson tes0t, it is typical to model

$\log(\mu_{ij}) = \log(r_{ij}) + \alpha + \beta z_i$,

where z_i is the dose-metric (e.g., log-dose or dose rank), r_{ij} is a rate multiplier (which can be the subgroup size) for the j-th subgroup of the i-th group, and α and β are parameters to be estimated. The test for linear trend is a test for $\beta=0$.

For the Poisson trend model, one fits the model with all dose groups present and tests the hypothesis $\beta=0$ against the alternative $\beta>0$. If this hypothesis is rejected at the 0.05 significance level, then the high dose group is omitted and the model is re-fit to the remaining dose groups. This process is continued until the hypothesis of positive slope is first not rejected. The high dose remaining at this stage is the NOEC.

For a non-trend version of this test, βz_i is replaced by $z_i'\beta$, where β is a parameter vector $<\beta_i>$and z_i is a column vector $< z_{iu} >$ whose u-th component is zero unless u=i. For i=0 (i.e., for the control group), all components of z_i are zero.

One then tests the hypotheses $\beta_i = 0$, for i=1, 2, . . . against the obvious 1- or 2-sided alternatives. Generally, a Bonferroni-Holm adjustment is made for the number of such comparisons made.

There are tests for lack of fit of the Poisson model, which will not be described here. References include McCullagh and Nelder (1989), Collett (1991), Aitkin et al. (1989), Morgan (1992), Mehta and Patel (1999), Hosmer and Lemeshow (1989), Thomas (1983). Software includes the GENMOD procedure in SAS version 8 and LogXact 4 for Windows. The availability in LogXact of exact permutation procedures

is especially appropriate for small samples or rare events. Robust versions of the Poisson tests are available. (Weller and Ryan (1998); Hirji and Tang (1998); Breslow (1990); Tarone and Gart (1980)).
Assumptions: Subjects are independent within and among groups and subgroups (if present). Within-group counts follow a Poisson distribution. For the trend version, a monotone trend in the dose-response is assumed. For the non-trend version, the dose metric is not used. Rather, an ANOVA-type model is used which assumes Poisson distribution rather than normal. Note: Robust versions of this test are available which allow for extra-binomial variation among subgroups.

5.2 Power of the Cochran-Armitage Test

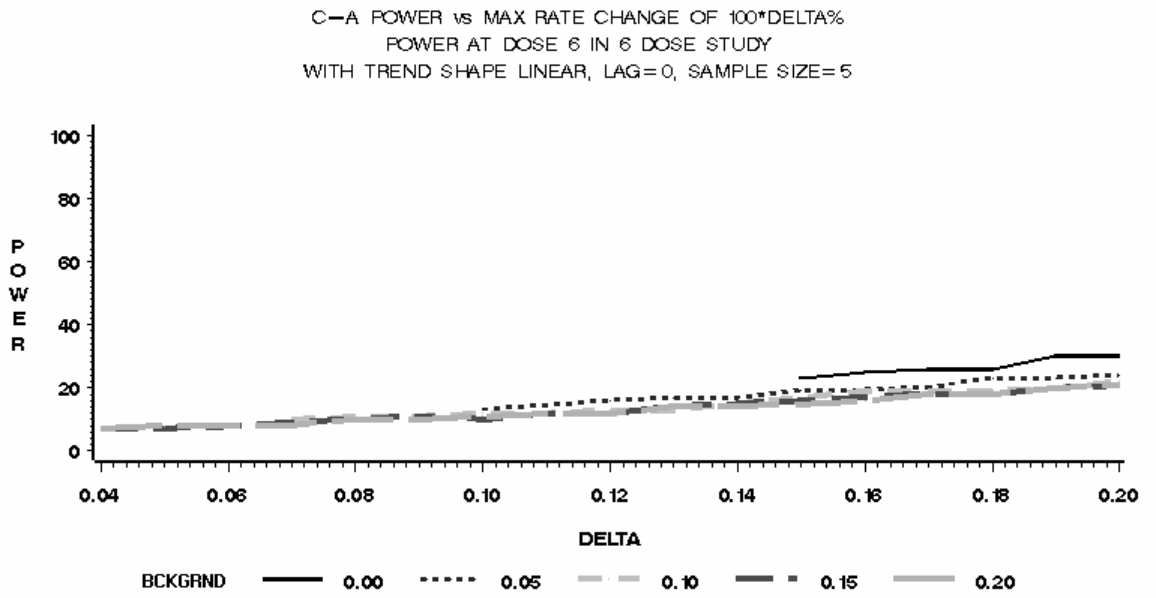
It is important to understand that the power of the Cochran-Armitage test depends on the background (i.e., control) incidence rate, with power decreasing as background rate increases (up to 50%), when all other factors are fixed.
The powers associated with the step-down application of the Cochran-Armitage test are not available, so far as we know, in any commercial or on-line source or even in published form. For that reason, a set of power plots has been included. The power of a statistical test also depends on the size effect to be found, the number of observations per treatment group and other factors. In comparing percent effect of treatments to a common control using a step-down trend test, additional factors affecting power are the shape of the concentration-response curve, the number of concentrations and the true threshold of toxicity (if such a thing exists). This last point is relevant only in that it affects the concentration-response shape by affecting the concentration at which the concentration-response begins to depart from horizontal.
A full treatment of the power of the step-down Cochran-Armitage test is being prepared for publication. As an aid in designing experiments, the following power curves can be used. These are for a response following a linear concentration-response shape, with no lag (i.e., threshold=0), for sample sizes 20, 40, 60 and 80 and number of concentrations ranging from 3 to 5.
A further consideration is whether the test is done in a 1-sided or 2-sided fashion. For quantal responses, it is unusual to be interested in anything but an increased incidence rate, so only 1-sided powers are presented. Powers for a 2-sided test will, of course, be lower for the same set of conditions.

Interpretation of the power curves

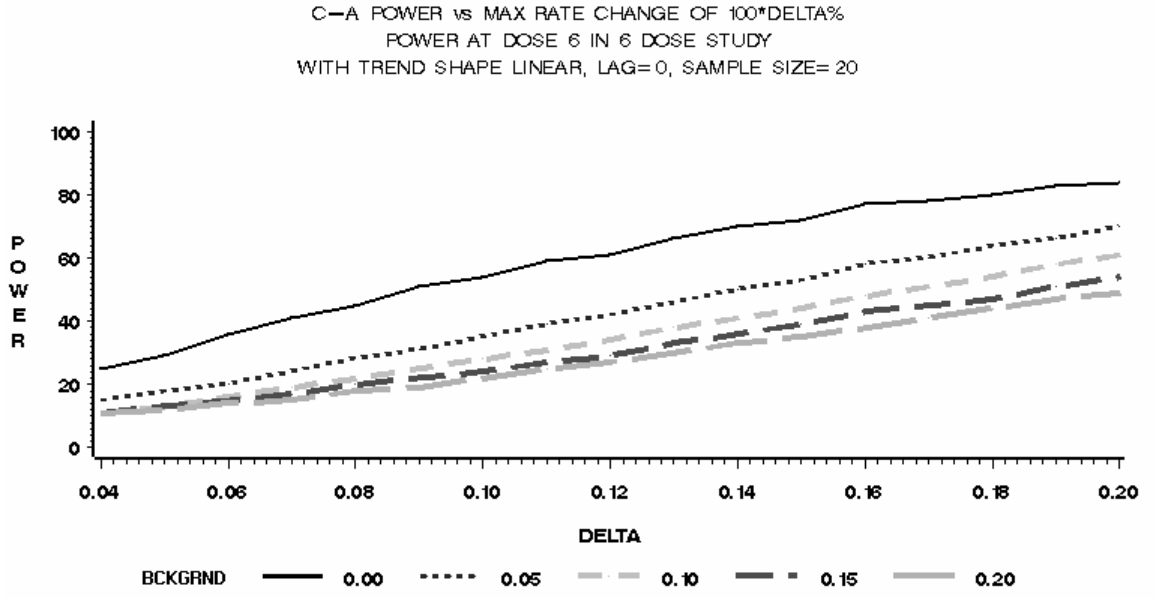
Three plots are presented for each of the sample sizes 20, 40, 60 and 80 subjects per concentration. The first of a set of three plots show the power of detecting an effect of a given size in the highest dose. The second of a set of three plots shows the power of detecting an effect of a given size at stage two of the step-down process; that is, in the second highest dose. The third of a set of three plots shows the power of detecting an effect of a given size in the third highest dose, that is, at stage three of the step-down process. For each plot, the vertical axis is the power or probability (expressed as a percent) of finding a significant effect if the true effect has the magnitude given on the horizontal axis. The horizontal axis shows the true change in percent effect from the control or background rate. Five power curves are drawn in each plot corresponding to background rates of 0, 5, 10, 15, and 20 percent.
To avoid ambiguity, it should be noted that these powers are expressed in terms of absolute, not relative, change. Thus, for example, they show the power of detecting an increase of 10 percentage points in the incidence rate as a function of background rate. For a background rate of 0, this is an increase from 0 to 10% incidence. For a background rate of 5%, this is an increase from 5% to 15%. Of course, these changes could be expressed in terms of a decrease in the rate of non-occurrence, e.g., as decreases in survival rate. Simple arithmetic will allow the use of these plots to compute relative rates of change as well.
In designing an experiment, no drastic change in number of replicates is generally needed to achieve at least 75% power of observing the size effect deemed important, bearing in mind that we do not know in advance with certainty at which concentration this size effect will be observed. Preliminary range-finding experiments usually provide some relevant information. It will be evident from the power curves that if we are to estimate small changes in percent effects, then large sample sizes are needed.
In all cases shown, the test used is a 0.05-level test, as described elsewhere in this document.
An example of power computations using these plots might be helpful. Suppose one wants to determine the NOEC for mortality in an experiment with daphnia magna, where past experience with this species and suggests the background mortality rate is near zero. The goal is to be able to detect a 20% mortality rate. Suppose that based on preliminary range finding experiments, a decision was made to do an experiment with five concentrations at 50, 100, 200, 400, and 800 ppm of the test compound plus a single (non-solvent) control. Furthermore, there is no anticipation of extra-binomial variance or within-tank correlations, so a standard Cochran-Armitage test can be done treating all subjects within a concentration

equally (i.e., ignoring any tank or replicate effect). The question is then, how many subjects per concentration should be planned?

First, consider designs with the same number, *n*, of subjects in each concentration as in the control. The power of the Cochran-Armitage test depends on the shape of the dose-response curve, which we do not know. Powers have been simulated for numerous shapes. Based on an examination of the various power plots, a reasonable choice for design purposes is the linear dose-response shape. In addition, the power depends on the threshold of toxicity. For design purposes, we will assume that is zero. The following plots will help.



This shows that 5 subjects per concentration would give very low power (about 25%) to detect a 20% change in the high concentration. There is little point in conducting the experiment for the purpose. Consider a design with 20 subjects per concentration.



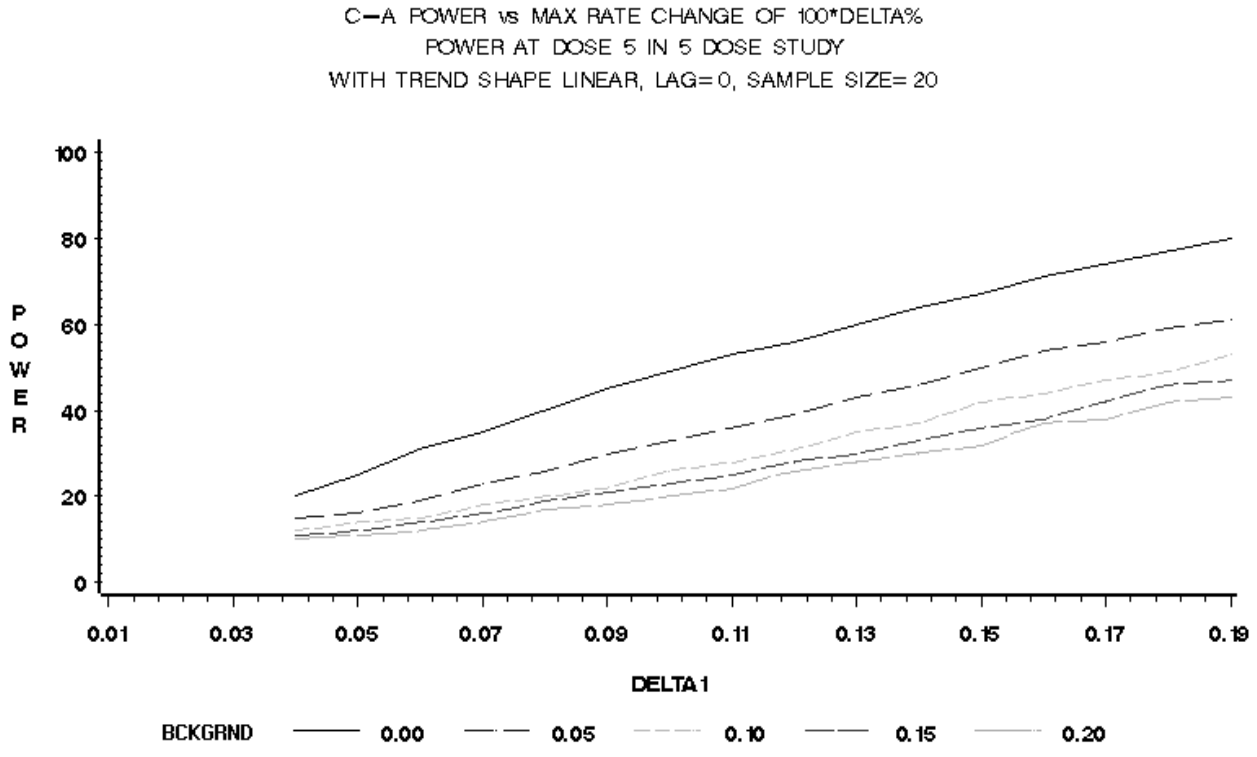
This sample size gives a power of 82% to detect a 20% mortality in the 800 ppm concentration. This may well be adequate. Next, consider the power to detect a 20% in lower concentrations. Fortunately, not

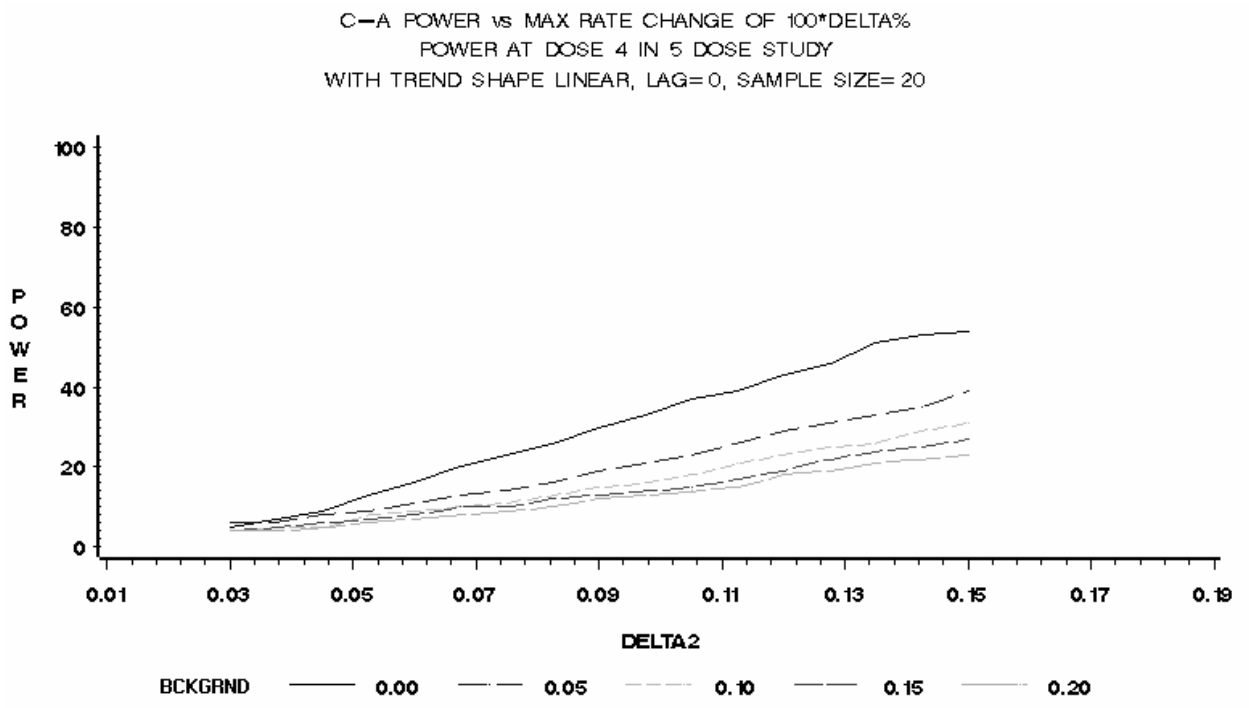
much power is lost in the step-down procedure. The power to detect a 20% mortality rate at 400 ppm is 80%, at 200 ppm is 78% and at 100 ppm is 76%. However, if the background incidence rate were 10% instead of zero, then the power to detect an increase in mortality rate of 20% drops to around 40%, which would be inadequate for most purposes.

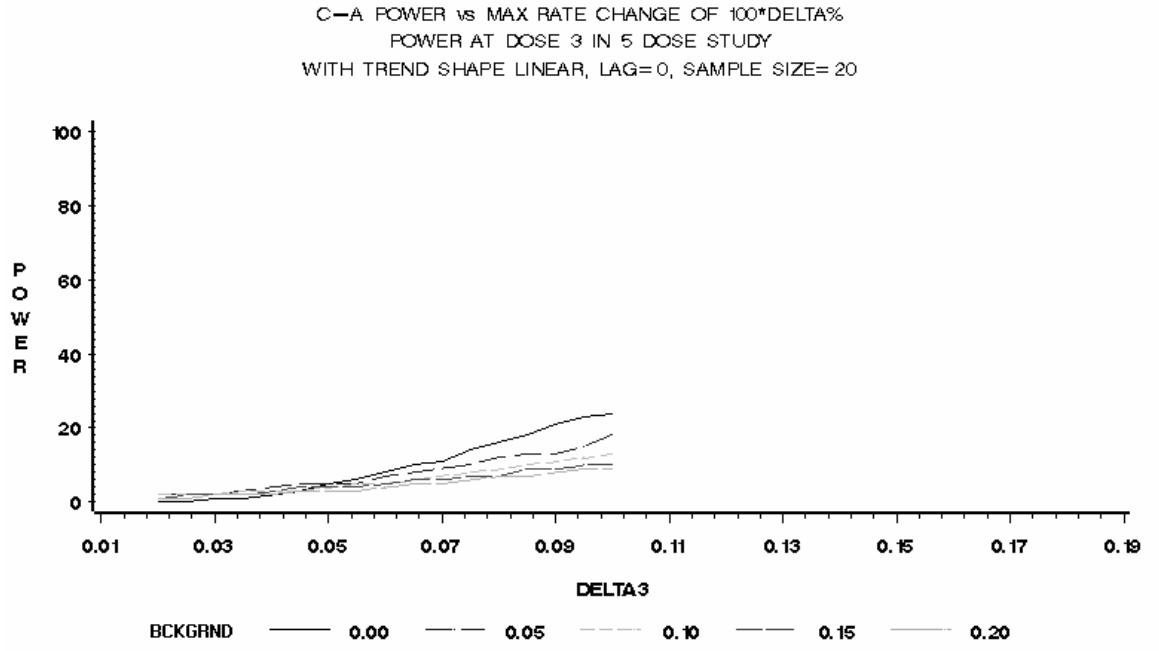
Replicates

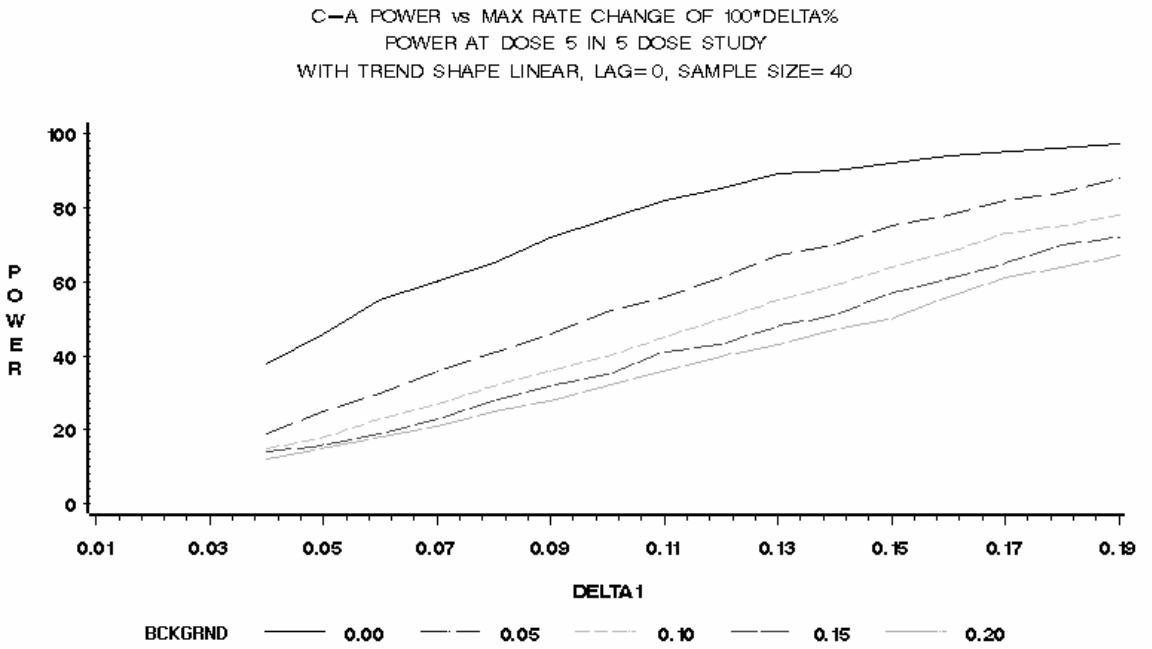
Decisions on the number of subjects per subgroup and number of subgroups per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-subgroup variation and correlation. If there are no subgroups, then there is no way to distinguish 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. 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 modeling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modeling 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.)

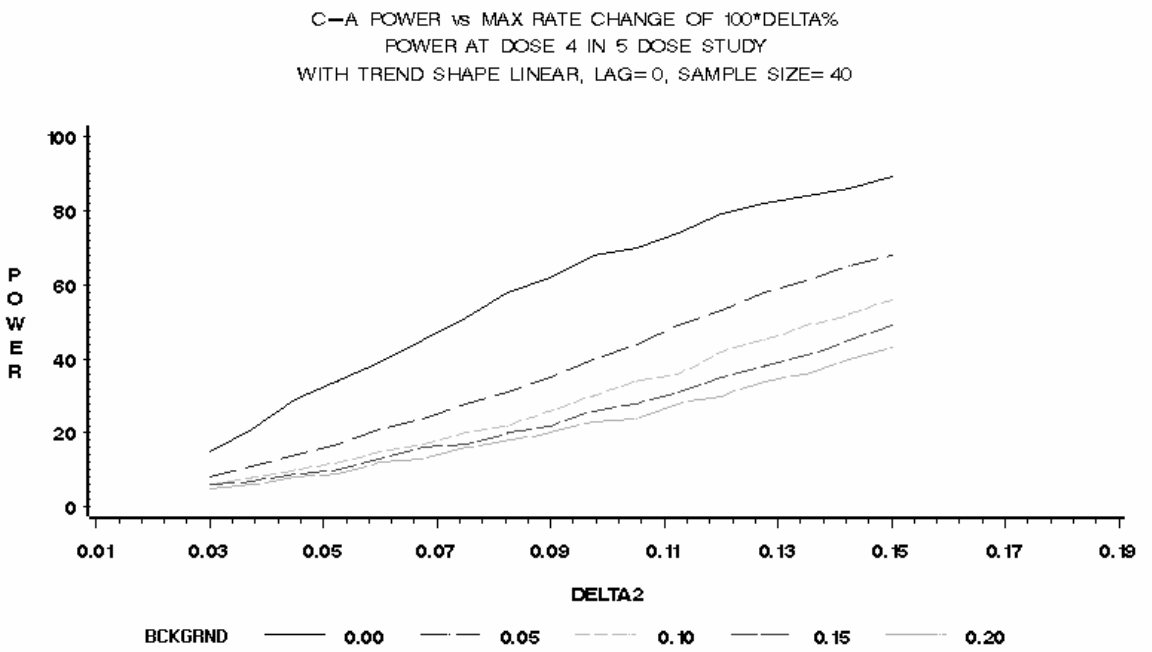
The number of subgroups per concentration and subjects per subgroup should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.

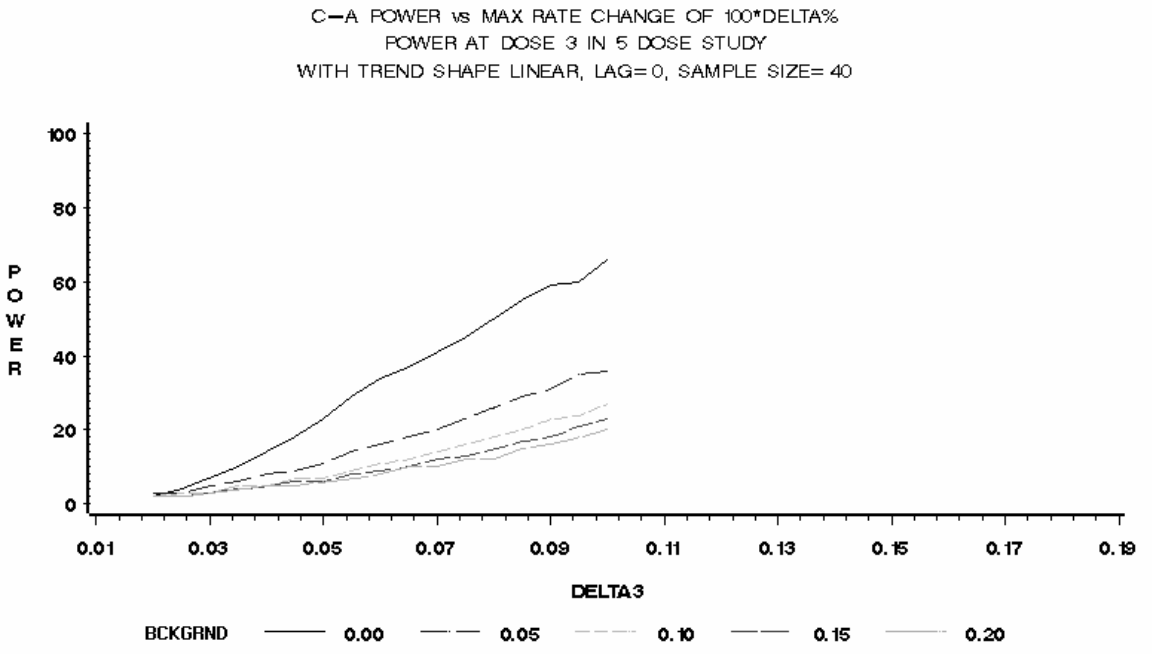


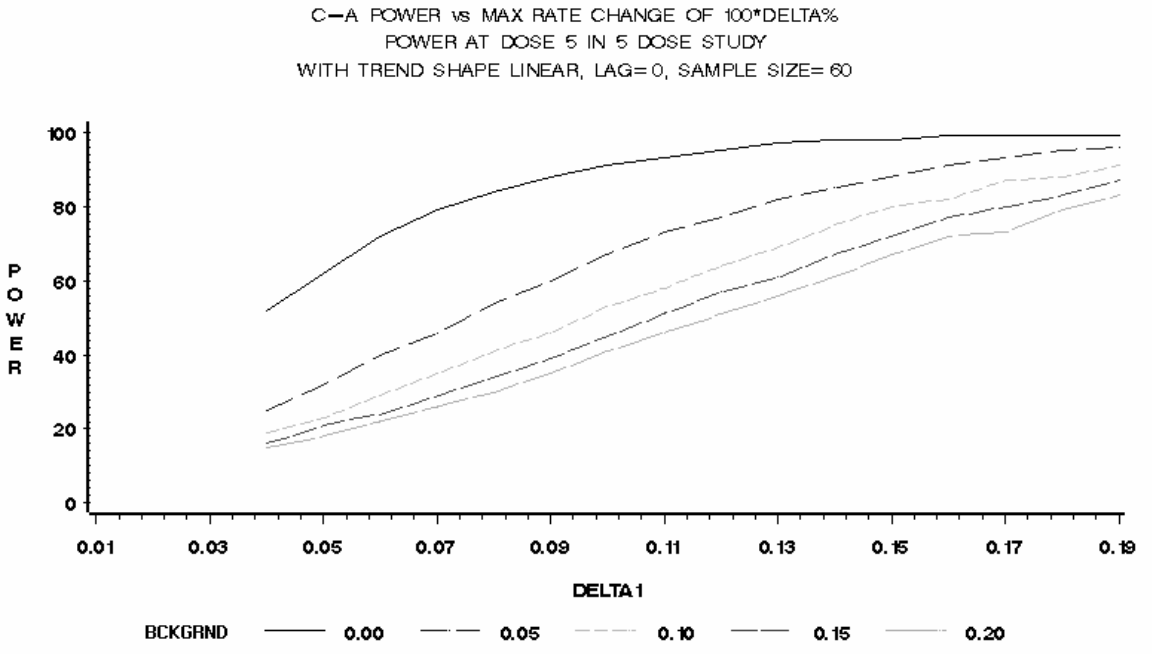

C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 4 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 20
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA2
BCKGRND
0.00
0.05
0.10
0.15
0.20

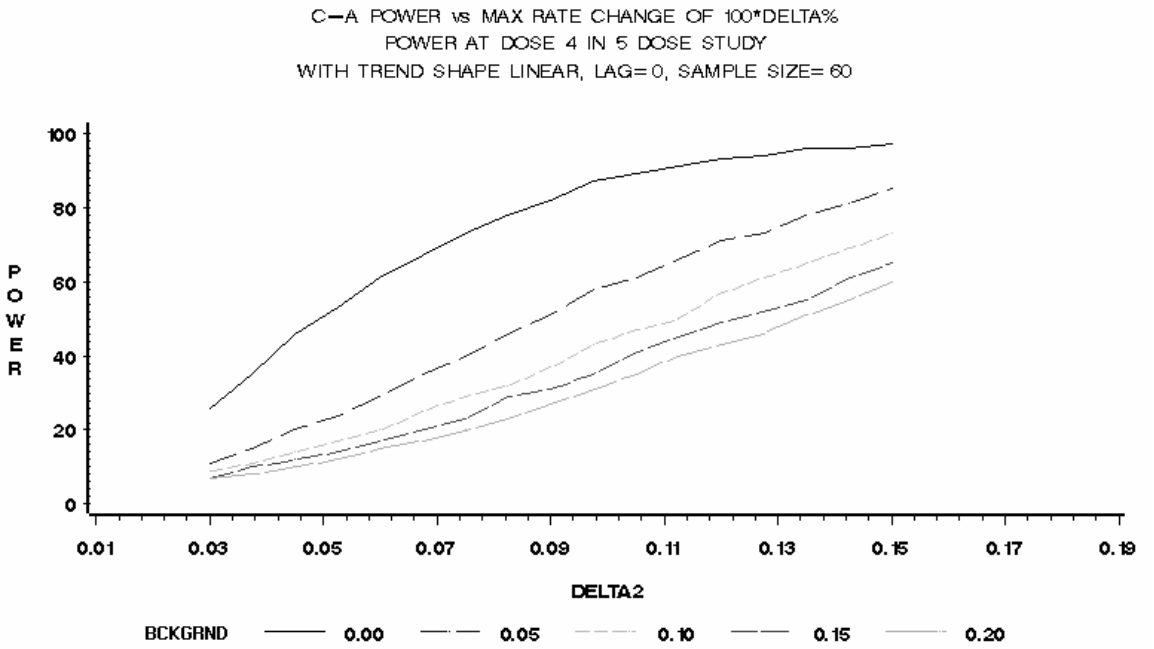

C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 3 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 20
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA3
BCKGRND
0.00
0.05
0.10
0.15
0.20

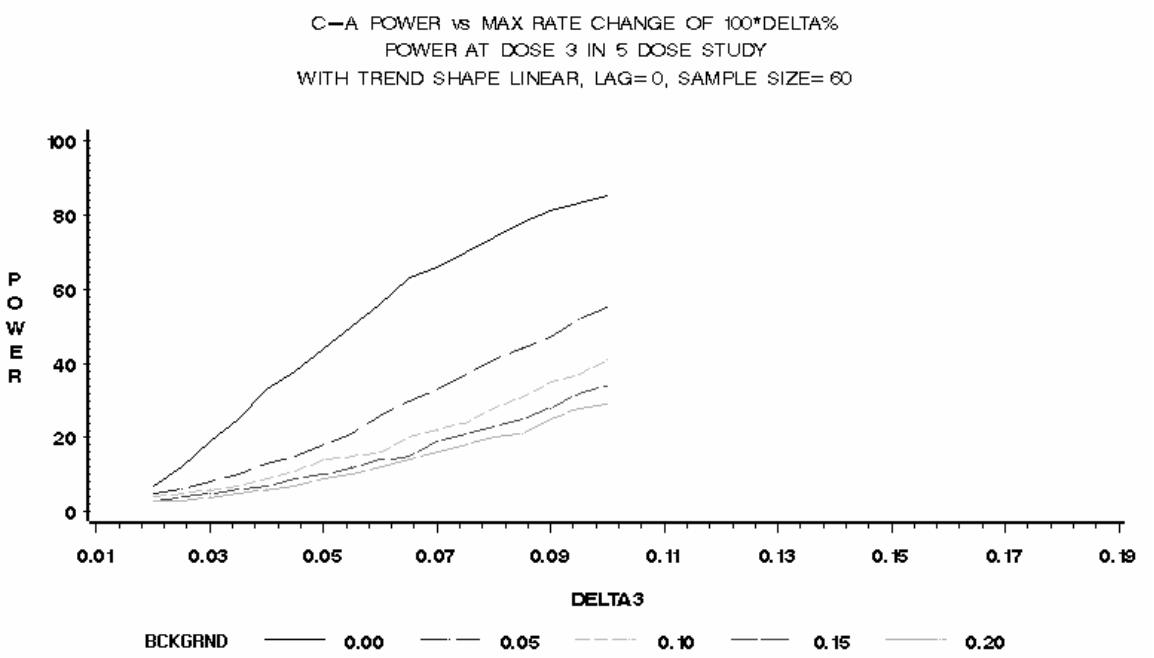

C—A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 5 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 40
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA1
BCKGRND
0.00
0.05
0.10
0.15
0.20

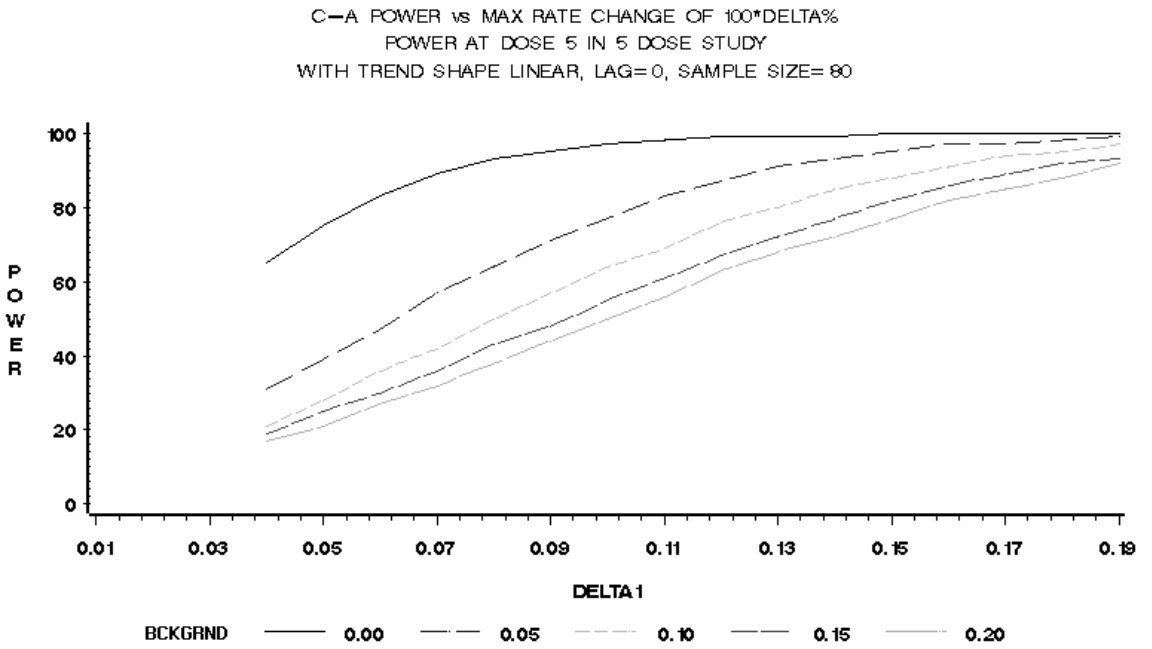

C—A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 4 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 40
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA2
BCKGRND
0.00
0.05
0.10
0.15
0.20

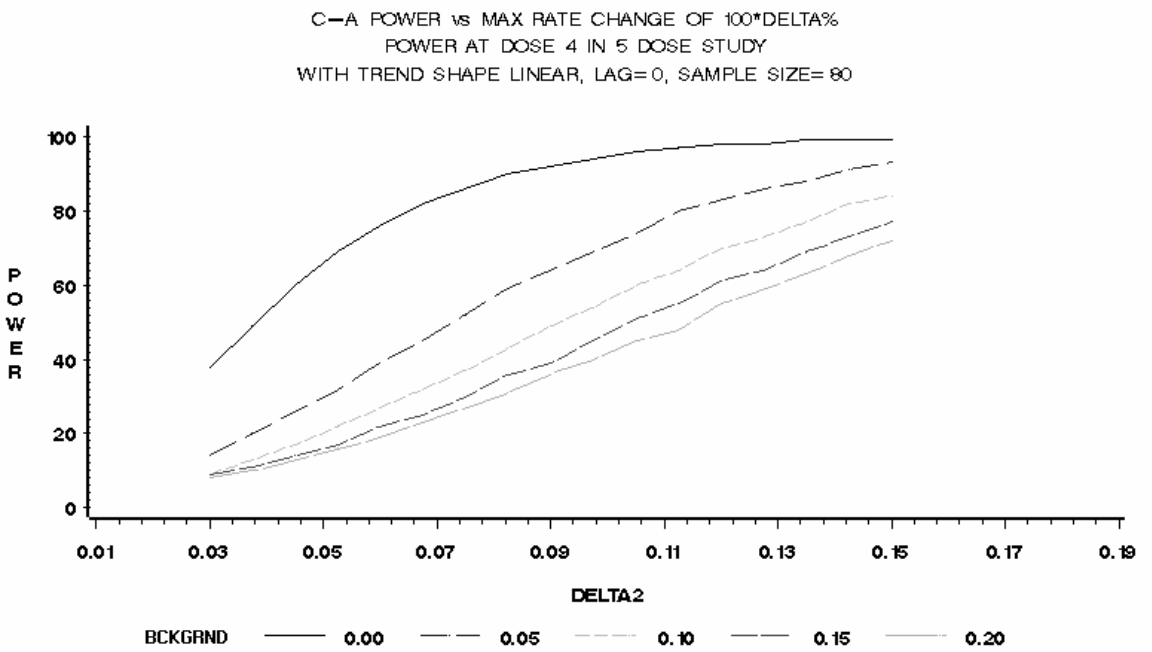

C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 3 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 40
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA3
BCKGRND
0.00
0.05
0.10
0.15
0.20

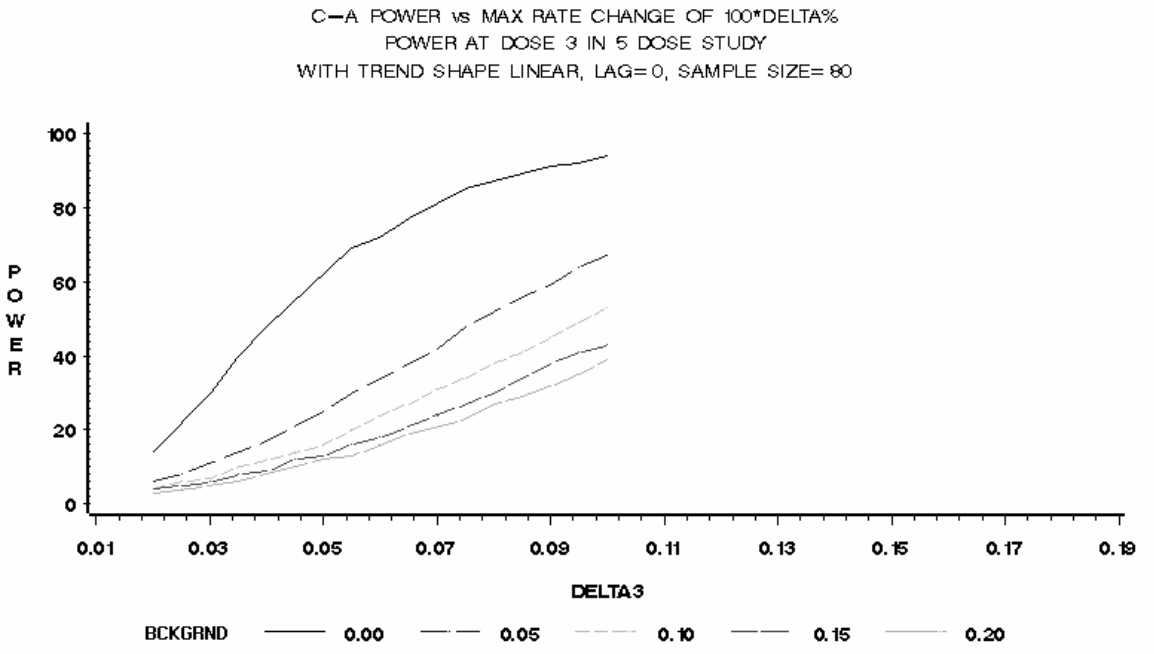

C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 5 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 60
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA1
BCKGRND
0.00
0.05
0.10
0.15
0.20


C—A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 4 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 60
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA2
BCKGRND
0.00
0.05
0.10
0.15
0.20


C—A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 3 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 60
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA3
BCKGRND
0.00
0.05
0.10
0.15
0.20


C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 5 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 80
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA1
BCKGRND
0.00
0.05
0.10
0.15
0.20


C-A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 4 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 80
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA2
BCKGRND
0.00
0.05
0.10
0.15
0.20


C—A POWER vs MAX RATE CHANGE OF 100*DELTA%
POWER AT DOSE 3 IN 5 DOSE STUDY
WITH TREND SHAPE LINEAR, LAG= 0, SAMPLE SIZE= 80
POWER
100
80
60
40
20
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
DELTA3
BCKGRND
0.00
0.05
0.10
0.15
0.20

5.3 Description of Selected Tests for Use With Continuous Data

General references on trend tests are Barlow *et al* (1972) and Robertson, *et al* (1988). Hochberg and Tamhane (1987) discuss step-down tests in general, Selwyn (1988) discusses the use of the Cochran-Armitage test in this way to establish the NOEC. USEPA (1995) recommends the use of step-down trend tests to establish NOECs for both quantal (incidence) and continuous responses. (USEPA (1995) recommends the Mantel-Haenszel test instead of the almost equivalent Cochran-Armitage test for incidence data and Jonckheere-Terpstra test for continuous data to establish the NOEC.) Dunnett and Tamhane (1991, 1992, 1995) discuss step-down trend tests to determine the equivalent of NOEC in medical tests. Tamhane and Dunnett (1996) discuss them in toxicology experiments, as do Tamhane, *et al*. (2001), Capizzi *et al*. (1984), Tukey *et al.* (1985). These authors criticize single step procedures as having low power of detecting real effects and offer step-down procedures as an improvement.

Both 1- and 2- tailed step-down procedures belong to the class of "fixed sequence" tests in the terminology of Westfall (1999) and Westfall et al (1999). Such tests also belong to the more general class of *closed systems of hypothesis tests* Peritz (1970) and Marcus, Peritz and Gabriel (1976). Budde and Bauer (1989) discuss a step-down procedure based on the Jonckheere-Terpstra test that differs from that discussed here. Kodell and Chen (1991) apply this same idea to quantal data, including the Cochran-Armitage test. These authors followed the more general but also more cumbersome closed system of Peritz (1970) and Marcus, Peritz and Gabriel (1976) rather than the fixed sequence approach.

There are several methods used to perform the tests of "fixed sequence" hypotheses for continuous responses, including Williams' test, the Jonckheere-Terpstra test, Bartholomew's test, Welch, and Brown-Forsythe tests, and sequences of linear contrasts, among others. These are well established as tests of the stated hypothesis in the statistics literature.

Williams' test is also a step-down trend test commonly used to establish NOECs in toxicological experiments. Tamhane and Dunnett (1996) discuss Williams' test and two new step-down procedures for use in toxicology and drug development. They note the low power of multiple comparison approaches to this work in toxicology in the context of discussing the benchmark dose approach of Gaylor (1983) and Crump (1984). They claim Williams' test loses power under some non-monotone alternatives, while their methods do not.

Salsburg (1986) discusses the low power of ANOVA methods in analysing dose-response experiments. He recommends Bartholomew's test as the most general test against ordered alternatives. He discusses the use of linear contrasts, but notes that they may not be powerful in experiments where the lower doses have no effect and all the effect is found only at the highest dose. Contrast tests make no direct use of the supposed monotone dose-response relationship, and, hence, are lower in power than alternative procedures that do. Also, linear contrasts test specifically for a linear relationship, not monotone relationships. If the dose-response is not linear with respect to the particular dose metric used, it loses power. Bartholomew's test is difficult to implement and Puri (1965) has shown that Jonckheere-Terpstra test is of very similar power. Robertson et al. (1988) has shown that under some conditions, Jonckheere-Terpstra test is more powerful, does not assume any specific shape (such as linearity) in the dose-response, and only requires monotonicity. It is also easily incorporated into step down procedures. Robertson et al. (1988) discuss in great depth various tests that are appropriate in dose-response experiments, or more generally, when the explanatory variable has an order restriction. They found that under certain conditions, namely the mean responses are approximately uniformly related to dose order (not magnitude); Jonckheere-Terpstra is more powerful than other alternatives considered.

Bartholomew (1961) compares Jonckheere-Terpstra test to his own. In the case of 3 or 4 treatment groups, Bartholomew sees little difference between the power of the two tests, for either equally spaced means or all but one mean equal. He in fact found Jonckheere-Terpstra test to be preferable, given its distribution-free nature. For larger k, however, Bartholomew's test is superior for the case of all means but one equal, while Jonckheere-Terpstra is still preferable for the equally spaced means case.

Williams' Test - Williams' test is step-down or fixed-sequence test procedure that can be used in the same situations as the Jonckheere-Terpstra test. Unlike the latter, Williams' is based on normally distributed, homogeneous responses and formally incorporates the presumed monotone dose-response in the estimated mean effects at each dose. These means are called isotonic estimates and are based on maximum likelihood theory, given the dose-response is monotone. Isotonic estimators were developed by Ayer *et al* (1955), who called their method Pool-the-Adjacent-Violators (PAVA) algorithm. Isotonic regression was introduced by Barlow *et al* (1972).

Assumptions: Independent random samples of normally distributed, homogeneous variables with monotone means (for example, $\mu_0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$.). The maximum likelihood estimate of μ_i under the monotone assumption is given by

$$\tilde{\mu}_i = Max_{1\leq u\leq i} Min_{i\leq v\leq k} \frac{\sum_{j=u}^{v} n_j \bar{Y}_j}{\sum_{j=u}^{v} n_j},$$

where $\bar{Y}_j$ is the arithmetic mean of dose group i and n_j is the sample size.

The roles of Min and Max are reversed for a non-increasing trend. It is easier to compute isotonic estimators than to describe them. Given that we expect a non-decreasing trend in the means, we look for a violation of this expected result. If $\bar{Y}_j > \bar{Y}_{j+1}$, then we pool (or amalgamate) these two means using a simple weighted average. We then re-examine the reduced set of means for violations of the expected order and do additional pooling as needed. It makes no difference in the final result which adjacent means we amalgamate first. We continue this amalgamation procedure until the means are in the expected non-decreasing order. The control is never amalgamated with positive dose groups.

It is evident from this description that, given unequal sample sizes, it can happen that doses i and i+1 are amalgamated and $\bar{t}_{i+1}$ is significantly different from the control but $\bar{t}_i$ is not. There are several ways to avoid this situation. Williams' suggests re-computing the isotonic means at each stage of the step-down procedure, using only those means remaining at each stage. In the balanced case, this is equivalent to the procedure already described. Another alternative is to use only the reduced set of amalgamated means and declaring all dose groups involved in the amalgamated mean to have significant effects if the amalgamated mean is significantly different from the control. Williams also suggests other modifications.

Advantages of Williams' test: This test makes direct use of the assumed monotone dose-response both in terms of the estimated mean effect at each dose and in the step-down conduct of the test. Under one of the modifications indicated above, it cannot happen that a low dose shows an effect and a higher dose does not. The test can be modified to take multiple sources of variation into account. The same method used in the Tamhane-Dunnett test below can be used for this purpose.

Disadvantages of Williams' test: This test loses power when additional higher dose groups are added to the design unless the effects at the new doses are substantially greater than at the lower dose levels. There can be a loss in power if there is a change in direction at the high dose, such as might occur if there is substantial mortality in that group. It is affected in an unknown way if the data are not normally distributed or heterogeneous. (However, Shirley's (1979) non-parametric alternative to Williams' is available for non-normal or heterogeneous data.) According to Bretz (1999) and Bretz and Hothorn (2000), the null distribution of Williams' test is known only for the balanced case. For the unbalanced case, the actual p-value when the nominal is 0.05 can be as much as 20% larger than the nominal. Williams (1972) claims that the equal sample size error probabilities are approximately correct if the difference in sample sizes is

not great. Bretz and Hothorn (2000) give reasons why this test should not be used for highly unbalanced data and they provide an alternative test, similar to Williams but overcoming several difficulties.

Power of Williams' Test: Definitive power properties of Williams' test are not readily available, though limited simulations have been published (Marcus, R. (1976); Poon, A.H. (1980); Shirley, E. A. (1979); Williams, D.A. (1971, 1972)) that suggest power characteristics similar to those for the Jonckheere-Terpstra test, so long as the data meet the requirements for Williams' test and there is no change in direction at the high dose. Large sample theoretical power properties have been published by Puri (1965), Bretz (1999) and Bretz and Hothorn (2000).

Confidence Intervals for NOEC from Williams' Test: It is sometimes of interest to have confidence bands for a dose-response curve. These bands can be used to construct simultaneous confidence intervals for mean or individual responses at different doses, form confidence intervals for doses at a given response, and to compare different dose-response curves. In the parametric regression context, it is quite straight forward to construct such confidence bands. When no such model is fit to the data, other methods must be employed.

Genz and Bretz (1999) have shown how simultaneous confidence intervals can be computed for the mean effect at each concentration that is applicable to Williams' and Dunnett's tests, as well as to various others. These procedures may suffer from the need to compute confidence intervals for means of no interest, namely at all concentrations rather than just at the NOEC. Hence, these intervals will be wider than what might be obtained by focusing on just the NOEC. The methods of Korn, Williams, and Schoenfeld described below are appropriate in the context of Williams' test.

According to Korn (1982), for a normally distributed response, one could construct simultaneous *1-*α confidence intervals for the means μ_i from the following.

$$\overline{Y}_j - m_{k,N-k}s/\sqrt{n_j} < \mu_j < \overline{Y}_j + m_{k,N-k}s/\sqrt{n_j} \quad (1),$$

where $m_{k,N-k}$ is the studentized maximum modulus distribution (Hochberg and Tamhane, 1987) for k treatment groups and N-k degrees of freedom, s is square-root of the pooled within-group variance estimate, N is the total sample size, and $\overline{Y}_j$ is the arithmetic mean of the j-th group. However, as Korn and others observe, these confidence intervals are wider than necessary for a monotone dose-response, in part because they make no use of the monotonicity.

Three methods for constructing more appropriate simultaneous confidence bounds are presented, all of which assume normality of the response within groups. If an alternative distribution is found for the response, it may be possible to modify these methods, but little such work appears to have been done. In addition, we may not want simultaneous confidence bounds at all tested concentrations, but only at the NOEC. Two methods will be presented for that.

The basis for all methods will be isotonic regression, which yields maximum likelihood estimates of the treatment means based on a monotone dose-response model. Isotonic estimators of the treatment means come from the Pool-the-Adjacent-Violators Algorithm, or PAVA, discussed in the context of Williams' test. Korn suggests the following modification of (1).

$$\max_{j\le i}(\overline{Y}_i - m_{k,N-k}s/\sqrt{n_i}) < \mu_j < \min_{i\le j}(\overline{Y}_i + m_{k,N-k}s/\sqrt{n_i}) .$$

These intervals are not derived from the isotonic estimates of means but from the sample means and the monotone (i.e., isotonic) assumption on the means. Korn suggests doing linear interpolation to obtain approximate confidence bounds for the response between tested doses.

Williams (1977) makes direct use of the isotonic estimators of the mean responses in his approach. An upper 1-α confidence bound on the true mean for the j-th treatment is given by $\tilde{\mu}_j + A_{\nu,\alpha}s$, where s is the square-root of the usual pooled variance estimate, $\tilde{\mu}_j$ is the isotonic estimate of the j-th treatment mean, and $A_{\nu,\alpha}$ is the critical value from the distribution of $(\tilde{\mu}_j - \mu)/s$, under the assumption $\mu_{j-1} = \infty$ and

$\mu_{j+1} = \mu_{j+2} = ... = \mu_k$. Williams (1977) contains tabulated critical values for this distribution. He also gives a modification of this to provide confidence bounds or intervals for the difference $\tilde{\mu}_0 - \tilde{\mu}_j$ and simultaneous confidence intervals for arbitrary contrasts of the means under the monotone assumption. These bounds are only available for the equal sample size case, though if sample size differences are small, they should be reasonable approximations.

Rather than use Williams' bounds, Schoenfeld (1986) developed isotonic estimators of the bounds based on likelihood ratio methods. His bounds are sometimes much tighter than those proposed by Williams (1977) and he specifically recommends them for toxicity experiments, especially when the dose-response curve is shallow and we desire upper confidence bounds on the effects at concentrations at or below the NOEC. The unfortunate aspect of Schoenfeld's otherwise attractive procedure is that for unequal sample sizes, the critical values used in constructing the upper bounds must be estimated by simulation. This makes the procedure more computer intensive than desirable for routine use. For this reason, the unequal sample procedure is not reproduced here. If the differences in sample sizes are not large, then approximate confidence bounds can be obtained by treating them as equal and using tables he provides. The method for equal sample sizes for an increasing trend follows.

Let $\bar{y}_j$ denote the arithmetic mean of the j-th treatment group, $\tilde{y}_j$ the isotonic mean of that group, and $\hat{y}_j$ the isotonic mean based on just groups 1 through j. Suppose that for some σ^2, the within-group variance of group j is σ^2/a_j, n_j is the sample size of that group and $N = n_0 + n^1 + n^2 + \ldots + n_k$. Let

$$\hat{\sigma}^2 = \sum_{j=1}^{k} a_j \sum_{m=1}^{n_j} (y_{jm} - \bar{y}_j)^2 / \sum_{j=1}^{k} (n_j - 1),$$

Finally, let $w_j = n_j * a_j$. Then the upper 1-α isotonic confidence bound on $\tilde{y}_i$ is the greatest solution x to the equation $T(x)=C_\alpha$, where C_α comes from a table of critical values given in the cited paper and

$$T(x) = \left\{ \sum_1 w_j (x - \tilde{y}_j)^2 - \sum_2 w_j (x - \hat{y}_j)^2 \right\},$$

where $\sum_1$ is summed over all j such that $\tilde{y}_i \le \tilde{y}_j \le x$ and $\sum_2$ is summed over all $j < i$ such that $\tilde{y}_i \le \hat{y}_j \le x$. The solution of this equation is simple, given that this is just a quadratic in x. It should be noted that for unequal sample sizes, the procedure is the same other than the manner of obtaining the critical value C_α.

It should be noted that this is a confidence bound for a single isotonic mean. Schoenfeld shows how this can be modified to yield simultaneous confidence bounds for estimating confidence bounds for several isotonic means simultaneously. He claims these simultaneous bounds, modified further to be two-sided, can be used to check whether a parametric regression model fits the data, since the predicted responses at all concentrations should fall within these simultaneous confidence intervals. Alternatively, though Schoenfeld does not suggest it, the same idea can be used to check for violations of the monotonicity assumption. The utility of either idea has not been fully explored.

Example 1. Williams' Test Consider an experiment with five doses of 10, 25, 60, 150 and 1000 ppm, respectively, plus a zero-dose control, with the following means and sample sizes. The response is trout weight in mg.

Dose	0	1	2	3	4	5
$\bar{Y}_i$	53.1	52.7	45.2	47.1	44.8	46.6
n_i	18	10	9	10	10	8

The null hypothesis is H_0: $\mu_0 = \mu_1 = \mu_2 = \ldots = \mu_k$ and the alternative is H_1: $\mu_0 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k$, a non-increasing trend. The first violation of the expected trend is from dose 2 to dose 3. We amalgamate these to get $\tilde{Y}_2 = \tilde{Y}_3$=(9*45.2 + 10*47.1)/(9+10)=46.2. The means are now 53.1, 52.7, 46.2, 44.8, 46.6 with sample sizes 18, 10, 19, 10, 8. The next violation is from (un-amalgamated) doses 4 to 5. We amalgamate these to obtain $\tilde{Y}_4 = \tilde{Y}_5$=(10*44.8 + 8*46.6)/(10+8)=45.6. The means are now 53.1, 52.7, 46.2, 45.6 with sample sizes 18, 10, 19, 18. Now that amalgamation is complete, we can write the dose group means as 53.1, 52.7, 46.2, 46.2, 45.6, 45.6 to retain the original number of groups. Williams' test statistic to compare dose i to the control is

$$\bar{t}_i = \frac{\tilde{Y}_i - Y_0}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}},$$

where s^2 is the usual ANOVA pooled within-group variance estimator. The statistic $\bar{t}_i$ is compared to a Williams' critical value $\bar{t}_{i,\alpha}$ which decreases as i increases. The amalgamation and tests can be carried out in some software packages, such as Probmc in SAS.

To complete the example, we need the pooled within-group standard deviation, s. Suppose that is 7.251. The resulting calculations are as follows, where WILL is the value of Williams' statistic, P2 and CRIT2 are the associated p-value and critical values of Williams' test, respectively, YBAR is the arithmetic mean of the indicated dose group, YTILDE is the isotonic mean, n is the sample size, YBAR0 is the mean of the control group, n0 is the control sample size, DFE is the degrees of freedom for Williams' test.

```
DOSE  YBAR  YTILDE  YBAR0   N  N0   DFE   WILL     P2       CRIT2

1     52.7   52.7   53.1   10  18   59   -0.1399  0.4446   -1.6711
2     45.2   46.2   53.1    9  18   59   -2.3309  0.0032   -1.3082
3     47.1   46.2   53.1   10  18   59   -2.4127  0.0013   -1.1721
4     44.8   45.6   53.1   10  18   59   -2.6225  0.0004   -1.1002
5     46.6   45.6   53.1    8  18   59   -2.4342  0.0006   -1.0557
```

Thus, for these data, the NOEC is dose 1, or 10 ppm. The observed percent effects at the NOEC and LOEC are 0.75% and 14.9%, respectively.

Suppose the same data was collected, as above, but in each concentration, there were two replicate subgroups. The breakdown of between- and within-replicate variance could affect the test results. In general, for a given total variance, the greater between-replicate variance is, the less sensitive the test will be. For example, the data above is consistent with VarBtween=9.076 and VarWithin=39.064 (as can be computed from the data for example 1 below using repa) and Williams' yields a NOEC of dose 2, or 25 ppm, when analysed to take both sources of variation into account. These data are also consistent with VarBtween=24.732 and VarWithin=21.844 (using repb below) and in that case, Williams' yields a NOEC exceeding 1000 ppm. These results arise because large between-replicate variance suggests the data is less

repeatable. Greater uncertainty in the results makes definitive conclusions more difficult. Note: The analyses just discussed used maximum likelihood estimates of the variance components. Other variance component estimators, such as REML estimators, can lead to different conclusions.
For reference, these same data were analysed by a 1-sided Dunnett's test and Jonckheere-Terpstra test. These analyses are described below.

The Jonckheere-Terpstra Trend Test The calculation of the Jonckheere-Terpstra test statistic is based on Mann-Whitney counts. These Mann-Whitney counts can be thought of as a numerical expression of the differences between observations in two groups in terms of ranks. The idea of the Jonckheere-Terpstra test is very simple. Order the observations from all groups combined, from smallest to largest. Decide, on biological or physical grounds, what direction (increasing or decreasing) the dose-response has.
For each two groups *i* and *j*, with $i < j$ and $d_i < d_j$, examine each pair (x_i, x_j) of observations that can be made, with x_i from group i and x_j from group j. Count the number, O_{ij}, of these pairs which follow the expected order, $x_i < x_j$ (for increasing trend; order reversed for decreasing trend). Add all of the O_{ij} and compare that sum to what would be expected if the dose-response were flat. A large positive difference is evidence of a significant increasing dose-response.
A step-by-step description of the calculation of the Jonckheere-Terpstra statistic follows. This description is largely from Rossini (1995, 1997).

1. Compute the $k(k\text{-}1)/2$ Mann-Whitney counts O_{ij} , comparing group *i* with group *j* , for $i < j$., as indicated above.

2. Define a test statistic *J* by

$$J = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} O_{ij}$$

i.e. *J* is the sum of the Mann-Whitney counts.

3. For an *exact* test, at the α level of significance, reject H_0 if

$$J \geq j(\alpha, k, (n_1, \ldots, n_k)),$$

and accept otherwise. The constant

$$j(\alpha, k, (n_1, \ldots, n_k))$$

is found in tabulated in Hollander and Wolfe 1973 and numerous other sources.

4. For an *approximate* test, at the level of significance, compute J^* by:

$$J^* = \frac{J - \mathrm{E}_{H_0}[J]}{\sqrt{\mathrm{Var}_{H_0}(J)}} = \frac{J - \left\{ \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \right\}}{\sqrt{\frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72}}}$$

J^* can be compared to the appropriate critical values of the normal distribution (i.e. reject H_0 if $J^* > z_\alpha$). Recall that n_i is the number of observations in the *i-th* group.

Advantages of Jonckheere-Terpstra Test. As a rank-based procedure, this test is robust against both mild and major violations of normality and homoscedasticity. There is an exact permutation version of the test available in commercial software (e.g., SAS and StatXact) to handle situations of small sample sizes or massive ties in the response values. It is based on a presumed monotone dose-response and is powerful against ordered alternatives for a wide variety of dose-response patterns and distributions. There is no problem with unequal sample sizes if individual subjects are the experimental units for analysis.

Disadvantages of Jonckheere-Terpstra Test. There is no way to take into account multiple sources of variance, such as within- and between-subgroups. A consequence of this is that if the experimental unit for analysis is a subgroup mean or median and these subgroups are based on unequal numbers of subjects, then no adjustment can be made for this inequality. There is no built-in procedure for deciding whether the data are consistent with a monotone dose-response model. Published power characteristics are limited, though extensive power simulations have been done (and are being prepared for publication) and an excerpt is included with this document.

Power of Jonckheere-Terpstra Test Asymptotic power properties of this test have been published (for example, Poon (1980); Bartholomew (1961); Odeh (1971, 1972); Puri (1965);), but they overestimate the power, sometimes grossly, for samples sizes less than 30 and are not computed for the step-down application described in this document. Limited simulations studies have also been published (e.g., Potter and Sturm (1981); Poon (1980); Shirley (1979)). A full treatment of the power of the step-down Jonckheere-Terpstra test is being prepared for publication. As an aid in designing experiments, the power curves in Annex 5.4 can be used. These are for a 2-sided test (as described in chapter 5 and in the next paragraph, a 2-sided test is used when one is confident that the dose-response is monotone, but not confident in advance whether the trend will be increasing or decreasing. That direction is set by the initial tests for increasing and decreasing trend with all data present) on a normally distributed response following a linear concentration-response shape, with no lag (i.e., response threshold=0), for sample sizes 5, 10, 20, 40 and 80 per concentration and number of concentrations ranging from 3 to 6. The full treatment considers other distributions, other shapes for the dose-response and other values of the response threshold. The sample size refers to the experimental unit used in analysis, whether it is an individual subject or a subgroup mean or median. In the latter case, it is assumed that the means or medians are all based on the same number of subjects.

A further consideration is whether the test is done in a 1-sided or 2-sided fashion. It is often true that in advance of the experiment, it is not known whether the response in question is likely to increase with increasing concentration or decrease. In this event, a 2-sided step-down test can be used. The power curves shown are for a 2-sided test. Powers for a 1-sided test will, of course, be higher. The choice to present only the linear dose-response shape was made to be intermediate between the powers for convex and concave shapes. The choice of presenting powers for 2-sided tests and the case of 0 lag will serve to make these curves conservative for experimental design purposes. That is, if one achieves P% power to detect a given effect size from these choices, one can be reasonably confident of one's design achieving at least P% power if some of the details of the dose-response differ from those presented.

In designing an experiment, the power curves can be used to design experiments of whatever power characteristics desired. In our experience, it should be possible with commonly used sample sizes, to achieve at least 75% power of observing the size effect usually deemed important, bearing mind that we do not know in advance with certainty at which concentration this size effect will be observed. Since we do not in advance at which concentration the started effect size will occur, at the design stage, we must consider the power of finding this size effect at any stage of the step-down procedure. Fortunately, as the power plots indicate, except for low power situations, the power is not much affected by the stage at which the effect is found. To be clear, the 75% power goal is somewhat arbitrary, but is a level achievable (and often exceeded) in experiments of the sizes typically used. Preliminary range-finding experiments usually provide some relevant information. In all cases shown, the test used is a 0.05-level test, as described elsewhere in this document.

Confidence Intervals for Effect Size at NOEC from Jonckheere-Terpstra Test

It is not clear that any meaningful method exists for constructing confidence intervals for non-normally distributed responses where non-parametric methods are used to evaluate the data. There is no true confidence interval associated with the Jonckheere-Terpstra test. If a gross estimate of a confidence interval of the mean response at the NOEC is desired, despite the fact that the test is non-parametric, then

one must resort to normal theory even though the NOEC determined by the Jonckheere-Terpstra test has very little association with normal theory. To do this, one could construct a traditional confidence interval of the form

$$\bar{x} \pm t^* s/\sqrt{n}$$

where n is the number of observations for the mean, s is the within-group ANOVA standard deviation, t is the $100(1-\alpha/2)$ quantile of the central t-distribution with ANOVA degrees of freedom. This formulation ignores the possible non-normality or heterogeneous nature of the data, the effect of outliers on the variance estimate and the fact that the NOEC was obtained without reference to this ANOVA. A more relevant measure of the adequacy of the NOEC is given by the power characteristics of the test used to determine it. Such power considerations are discussed in this document.

It is tempting to apply the confidence bounds described for Williams' test to means when another trend test is used, such as Jonckheere-Terpstra or Cochran-Armitage (where group proportions, suitable transformed, are used). However, these tests are not based on isotonic estimates of group means, so the applicability of these methods is not clear. Nevertheless, where there is no serious violation of the normality or homoscedasticity requirements or multiple variance components to take into account, these methods seem better than *ad hoc* methods discussed here.

Example 2 Jonckheere-Terpstra Test This uses the same data as Example 1 for Williams' test. However, it is reproduced in full here with more information to illustrate additional points. The monotonicity test presented is that described in section 5.1.3 in terms of linear contrasts.

First, consider the power characteristics of the test. For illustrative purposes, consider the case that at design, an experiment of 10 daphnid per concentration (in reps of size 1) is to be evaluated, with a total of six concentrations (five plus control). Based on historical control data, the expected within-group standard deviation σ=0.1. At design time, a fairly conservative approach would be to consider a dose-response linear with respect to log concentration. From the power curves for the Jonckheere-Terpstra test with 6 doses and linear dose-response shape, the power to detect a shift in the high dose group of 1.5 σ is about 93%. The power to detect the same size shift in the second highest concentration is about 90% and the power to detect this size shift at the third highest concentration is about 90%. The curves provided do not show the power to detect such a shift in the fourth highest concentration, but it is about 88%. Given a control mean of 4.1 (historical control mean), we have high power of detecting a shift of 1.5 σ or 100*1.5/4.1 =3.7% of the control mean in any of the top four concentrations. For other dose-response shapes, the power will be somewhat higher or lower, according to the shape.

In the present example, the actual sample standard deviation was slightly higher, at 0.0107 and a few daphnid died before the end of the experiment (final samples sizes 9, 10, 8, 7, 7, 10), so the achieved power will be a bit less, but it should still be satisfactory.

Raw Data for Example 1
REPA is for the first subgrouping, REPC is for the second.

CONC	DOSE	REPA	REPC	Y	CONC	DOSE	REPA	REPC	Y
0	1	2	1	39.5419	25	3	1	2	47.2060
0	1	1	1	42.1589	25	3	2	2	49.0197
0	1	2	1	45.1310	25	3	2	2	49.0803
0	1	1	1	45.7584	25	3	2	2	58.0969
0	1	1	1	46.3502	60	4	1	1	41.4250
0	1	1	1	48.4031	60	4	1	1	41.4392
0	1	1	1	49.1750	60	4	2	1	42.7406
0	1	1	1	51.2812	60	4	1	1	43.4315
0	1	1	1	51.8452	60	4	1	1	44.1543
0	1	2	1	52.9215	60	4	1	2	45.7393
0	1	1	2	54.8646	60	4	2	2	49.1442
0	1	1	2	55.5785	60	4	2	2	51.0632
0	1	1	2	56.4552	60	4	2	2	51.6004
0	1	2	2	56.9496	60	4	2	2	60.2623
0	1	2	2	60.4886	150	5	2	1	32.6964
0	1	2	2	62.0290	150	5	2	1	39.6888
0	1	2	2	64.3035	150	5	1	1	41.6733
0	1	2	2	72.5645	150	5	2	1	42.0448
10	2	2	1	43.2570	150	5	1	1	42.4995
10	2	2	1	45.3761	150	5	2	2	47.3381
10	2	2	1	50.7812	150	5	1	2	47.8829
10	2	1	1	53.8215	150	5	1	2	48.1557
10	2	2	1	53.8444	150	5	1	2	52.6260
10	2	1	2	54.4215	150	5	2	2	53.3945
10	2	1	2	54.4328	1000	6	1	1	31.2249
10	2	2	2	54.8774	1000	6	1	1	40.1562
10	2	1	2	56.6354	1000	6	1	1	43.2021
10	2	1	2	59.5528	1000	6	1	1	44.2384
25	3	1	1	33.7865	1000	6	2	2	44.9417
25	3	1	1	36.9014	1000	6	2	2	52.7730
25	3	1	1	42.2580	1000	6	2	2	56.0988
25	3	1	1	44.5783	1000	6	2	2	60.1649
25	3	2	1	45.8729					

Test for Monotonicity

SOURCE	NDF	DDF	F	P_F
LINEAR TREND	1	59	8.59	0.0048
QUADR TREND	1	59	2.38	0.1285

Test for monotonicity indicates trend test is appropriate.
The step-down Jonckheere trend test will be done.

KEY
ZC is Jonckheere statistic computed with tie correction
ZCCF is ZC with continuity correction factor
P1UPCF is p-value for upward trend
P1DNCF is p-value for downward trend
p-values are for tie-corrected test with continuity correction factor

Hi Dose	JONC	ZC	ZCCF	P1UPCF	P1DNCF	SIGNIF
1000	589	-3.17698	-3.1712	0.9992	0.0007	**
150	401	-3.32851	-3.3214	0.9996	0.0004	**
60	265	-2.61104	-2.6014	0.9954	0.0044	**
25	147	-1.96502	-1.9508	0.9745	0.0239	**
10	90	0	0.0240	0.4904	0.4904	

No significant trend DOWN from control to conc=10. No further testing below 10 is required. It will be observed that the NOEC is conc 10 and all higher concentrations indicate a significant effect. The above was computed with the individual subject as unit of analysis.

Suppose the same data was collected, as above, but in each concentration, there were two replicate subgroups. Then there would be two variance components to consider, VarBtween and VarWithin, the variance between replicates and the variance within replicates. The breakdown of between- and within-replicate variance could affect the test results. In general, for a given total variance, the greater between-replicate variance is, the less sensitive the test will be. For example, the data above is consistent with VarBtween=9.076 and VarWithin=39.064 and Williams' yields a NOEC of 25 ppm, when analysed to take both sources of variation into account. This decomposition of the data into replicates is identified as REPA. These data are also consistent with VarBtween=24.732 and VarWithin=21.844 and in that case, Williams' yields a NOEC exceeding 1000 ppm. This decomposition of the data into replicates is identified as REPC. These results arise because large between-replicate variance suggests the data is less repeatable. Greater uncertainty in the results makes definitive conclusions more difficult. Note: The analyses just discussed used maximum likelihood estimates of the variance components. Other variance component estimators, such as REML estimators, can lead to different conclusions.

If replicate means are analysed instead of individual observations as above, then all of the procedures (Williams', Jonckheere-Terpstra, and Dunnett) and exact permutation analysis yield NOECs greater than 1000 for either scenario of between- vs. within-rep variance. This is in contrast to the previous conclusion that the NOEC is 10 and all higher concentrations are effects levels. This is because of loss of power from the much reduced sample sizes, the sample sizes now being the number of reps per copncentration.. While there is a reduction in variability from using the means, that reduced variability is not enough to overcome the reduced sample sizes and a lower power test results. Thus, a decision to analyse replicate means should be made only if there is sufficient evidence of between-subject correlation as to render the indicated analyses invalid. This issue will be less important if there are more replicate subgroups, because the impact on sample size (for degree of freedom calculations) will be less. This consideration has obvious implications for experimental design.

The test was repeated using replicates, as defined by REPA and repeated again with reps as defined by REPC. In each case, the NOEC was found to exceed 1000 ppm. As would be expected, if we reduce the number of analysis units to 2 per concentration, there is substantial loss of power. It should also be recognized that unweighted an analysis of subgroup means ignores the unequal sample sizes. While that will not be a major consideration when sample size differences are very small, it can be quite significant when, as is the case here, sample size differ by 50% or more. The selection of weighting scheme to use is not trivial and is not discussed here.

It should be emphasized that these data are used to illustrate several ideas, such as indicating the potential effect of subgroups on an analysis. The fact that the data are analyzed in several ways to illustrate these ideas should not be taken to imply that all methods are appropriate for the same data set. Where there are subgroups, one must give careful consideration to how the subgroup information is used in the analysis.

Example 3 Jonckheere-Terpstra The length of daphnid in individual beakers was measured at the end of a 21-day study. A large percentage of the measurements were identical, as shown by the following frequency table.

FREQUENCY TABLE OF LENGTH BY CONC

LENGTH (MM) / CONC (mg/L)	0	1	2	4	8	16	Total
3.5	0	0	0	0	0	1	1
3.6	0	0	0	0	0	1	1
3.7	0	0	1	0	1	2	4
3.8	0	0	0	0	1	4	5
3.9	0	2	0	1	2	2	7
4.0	1	4	1	3	2	0	11
4.1	5	4	6	3	1	0	19
4	3	0	0	0	0	0	3
Total	9	10	8	7	7	10	51

Given the large percentage of ties in these measurements of daphnid length, the data cannot be regarded as normally distributed, nor will any transformation of the data make it so. If we nonetheless analyse this by Dunnett's or Williams' test, ignoring the normality issue, the result is that under Dunnett's and Williams' tests (either as is or after a log-transformation), the NOEC is less than the lowest concentration, whereas under Jonckheere-Terpstra (large sample), the NOEC is 2. If, on the other hand, we do an exact permutation analysis because of the tie issue, the exact Jonckheere-Terpstra NOEC is 8. It would be hard to justify 2 as the NOEC, given the higher percent change at 1, but calling a 2.5% change from control an effects level seems overly restrictive. The exact result seems cautious without being overly restrictive, and is the recommended procedure. Interestingly, Dunn's test, a large sample non-parametric procedure, sets the NOEC at 4, which has some intuitive appeal for these data. However, like Dunnett's test, Dunn's makes no use of the presumed monotone dose-response nature of the data and a single example should not persuade one to drop the step-down approach to analysis of toxicity data.
This example underscores the dependence of the statistical conclusion on the method of analysis. There is no escape from this under the regression approach, since the model used to estimate an EC20, EC10, EC5, etc, can have a large impact on the estimate and its confidence bounds.

TRTMNT	MEAN	MEDIAN	STD_DEV	COUNT	Pct Change from Control
Control	4.12222	4.1	0.06667	9	0
1 mg/L	4.02000	4.0	0.07888	10	2.5
2 mg/L	4.03750	4.1	0.14079	8	2.1
4 mg/L	4.02857	4.0	0.07559	7	2.3
8 mg/L	3.91429	3.9	0.13452	7	5.0
16 mg/L	3.75000	3.8	0.12693	10	9.0

An additional example using the Jonckheere-Terpstra is given in annex 5.1, where proportions are so analysed.

Tamhane-Dunnett Test - This test assumes normally distributed data but is optimized for variance heterogeneity. There is little power loss for homoscedastic data. The basic statistic is quite simple:

$$T = \frac{\bar{x}_i - \bar{x}_0}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_0^2}{n_0}}}$$

For a 2-sided test, T is compared to the maximum modulus distribution for k comparisons and $df=n_0+n_i-2$.
For a 1-sided test, T is compared to the Studentized maximum distribution for same k and df.
If there are subgroups in the treatment groups, the sample variances in the above formula are replaced by Satterthwaite-type expressions.

$$Let\ C_g = \sum_j \frac{n_{gj}^2}{n_g^2}$$

where n_{gj} is the number of subjects in the j^{th} subgroup of treatment group g, $g=0$ or i, and n_g is the total number of subjects in group g. Let

$$U_g = \frac{VWR_g}{n_g} + C_g * VBR_g$$

where VWR_g and VBR_g are the within- and between-subgroup variances, respectively, of treatment group g.
In the expression for T above, substitute U_i, U_0 for s_i^2, s_0^2.
The approximate degrees of freedom for this variance estimate is

$$df_i = \frac{(U_0 + U_i)^2}{\frac{U_0^2}{n_0 - 1} + \frac{U_i^2}{n_i - 1}}.$$

Tables for the Studentized maximum and maximum modulus distributions can be found in Hochberg and Tamhane (1987), among other sources. Alternatively, tables for these distributions can be computed using the following formulae.
Maximum Modulus Distribution

$$\Pr\{\max|T_i| \le t\} = \int_0^\infty \left[2\Phi(tx) - 1\right]^k dF_\nu(x)$$

Studentized Maximum Distribution

$$\Pr\{\max T_i \le t\} = \int_0^\infty \left[\Phi(tx)\right]^k dF_\nu(x)$$

In both formulae, k is the number of treatment groups in addition to the control and ν is the df.

Advantages of Tamhane-Dunnett Test. This test allows heterogeneous variances, but loses little power, compared to Dunnett's test, when the variances are homogeneous. As shown above, it can be adapted to handle multiple sources of variance (e.g., within- and between-subgroups).

Disadvantages of Tamhane-Dunnett Test. This test assumes within-group (or subgroup) normality. The main disadvantage is that it makes no use of a presumed monotone dose-response and, thus, loses power in comparison to tests that do.

Power of Tamhane-Dunnett Test According to Dunnett (1980) and Hochberg and Tamhane (1987), the power of this test is very similar to that of Dunnett's test when the assumptions underlying that test are met. The power for the unequal variance case will depend on the particular pattern of variances.

Confidence Intervals for Effect Size at NOEC for Tamhane-Dunnett Test. A general method is discussed above for simultaneous confidence intervals for effect size at all concentrations included in the experiment. A simple confidence interval for the difference between the mean effect at the NOEC and at the control comes immediately from the test statistic. A 95% confidence interval for that difference is given by

$$\mu_i - \mu_0 = \bar{x}_i - \bar{x}_0 \pm T * \sqrt{\frac{s_i^2}{n_i} + \frac{s_0^2}{n_0}},$$

where T is the maximum modulus distribution for k comparisons and $df=n_0+n_i-2$. Simple arithmetic can convert this into a confidence interval for the percent effect. Where appropriate, this can be given in 1-sided form based on the Studentized maximum distribution for same k and df.

Example 4 Tamhane-Dunnett Test. Daphnid were housed in individual beakers. The response is total live young after 21 days exposure to the chemical. Since the observation is a count, a square-root transform might be used prior to grouping the data and doing an analysis. The result of such a transform is not reported here, but the transformed response was not normally distributed so a non-parametric procedure would be appropriate. In this example, the control is labelled Dose 1.

TOTLY MEASUREMENTS FROM DATASET TAMHANE
GROUP STATISTICS FOR VALUE BY DOSE

DOSE	COUNT	MEAN	MEDIAN	STD	STD_ERR
1	9	17.2222	18.0	2.48886	0.82962
2	10	26.9000	26.5	1.28668	0.40689
3	10	21.4000	22.5	6.36309	2.01219
4	9	18.2222	17.0	8.48201	2.82734

SHAPIRO-WILK TEST OF NORMALITY OF TOTLY

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
38	5.21021	-0.12654	2.51771	0.94106	0.060989

LEVENE TEST FOR TOTLY

DF	LEVENE	P_VALUE	SIGNIF
3	4.17542	0.012769	**

The data was found to be normally distributed but group variances were unequal. A Tamhane-Dunnett analysis is appropriate.

Tamhane-Dunnett 2-sided test for difference in means in TOTLY Using maximum likelihood estimates of variance.

DOSE	MEAN	STDERR	DF	CONTROL	DIFF	CRIT	SIGNIF
2	26.90	0.40689	11.71	17.222	9.68	2.5482	*
3	21.40	2.01219	11.93	17.222	4.18	5.9843	
4	18.22	2.82734	9.37	17.222	1.00	8.4401	

Only the low dose group mean response was found significantly different from the control mean response.

MONOTONICITY CHECK OF TOTYOUNG
DOSES 0, 60, 75, 100 PPM

PARM	P_T	SIGNIF
DOSE TREND	0.8228	
DOSE QUAD	0.0038	**

Based on the formal test for monotonicity, a trend test is not indicated. The monotonicity test presented is that described in section 5.1.3 in terms of linear contrasts of normalized ranks. Of course, the decision of whether to apply the Tamhane-Dunnett test rather than a trend-based test will depend on a prior

determination of the appropriateness of a trend-based test. The results are presented in opposite order here, since the focus is on illustrating the Tamhane-Dunnett test.

Dunnett Test - This test assumes normally distributed responses with homogeneous within-group variances. The basic statistic is that for the Student T-test and is similar to that for the Tamhane-Dunnett :

$$T = \frac{\bar{x}_i - \bar{x}_0}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}},$$

where *s* is the square-root of the usual pooled within-group sample variance. *T* is compared to the *1*- or *2*-sided upper α equicoordinate point of a *k*-variate *t*-distribution with correlations defined below, and $df=n_0+n_i+n_i \ldots +n_{k-1}-k$.
The correlation of the *i-th* and *j-th* comparisons is given by

$$\rho_{ij} = \left[\frac{n_i n_j}{(n_i + n_0)(n_j + n_0)}\right]^{1/2}.$$

If all treatment sample sizes are equal, but the control sample size is possibly different, these all reduce to *1/(1 + r)*, where $r=n_0 / n$ and *n* is the common treatment sample size. Note that if $n_0 = n$, this is just 0.5. Tables for various values of *r* are available in many places and the calculations based on the general correlation structure are computed automatically by some commercial software packages, such as SAS.
Variations on this and computational formulas are given in Hsu (1992) and in Hochberg and Tamhane (1987). If there are subgroups in the treatment groups, the sample variances in the above formula can be replaced by Satterthwaite-type expressions as for the Tamhane-Dunnett procedure.

Advantages of Dunnett test. It is readily available in many books and statistical software packages, is familiar to most toxicologists as well as statisticians, is simple to calculate.

Disadvantages of Dunnett's test. It assumes normality and variance homogeneity (within group and within subgroup). More importantly, it ignores the monotonicity in the dose-response if present and, thus, loses power in relationship to step-down trend methods. The test was developed for qualitatively different treatments, such as different formulations of a drug or different drugs. As with all pairwise tests, this can lead to a low dose being found statistically significant and a higher dose not statistically different from the control.

Power of Dunnett's Test A key question to address here is the precise alternative hypothesis for which power is desired. One relevant power of Dunnett's test is the probability of a significant result for a particular concentration, given the true mean response for that concentration differs from the control by a specified amount, Δ. Such a power can be approximated using the distribution functions for Dunnett's test and the *t*-test. In SAS, such an approximate power can be computed as follows.

$$Let\ \lambda = \frac{\Delta}{\sigma\sqrt{\frac{2}{n}}}.$$

POWER1=1-PROBT(PROBMC('DUNNETT1',.,1-α,DF,GRPS),DF,λ),
POWER2=1-PROBT(PROBMC('DUNNETT2',.,1-α,DF,GRPS),DF,λ),
where α is the size of the test, Δ is the difference in means to be detected at the specified concentration, *n* is the number of replicates in each concentration (assumed equal here), *DF* is the ANOVA degrees of freedom for error, *GRPS* is the number of concentrations to be compared to the control, σ is the estimated within-group standard deviation. *POWER1* and *POWER2* are the approximate powers of 1-sided and 2-sided tests, respectively.

This and related powers can also be obtained using the %IndividualPower macro given in (Westfall *et al*, 1999). Additional references to power computations to Dunnett's test are Dunnett and Tamhane (1992) and Dunnett (2000).

Confidence interval for Effect at NOEC from Dunnett test. A general method is discussed above for simultaneous confidence intervals for effect size at all concentrations included in the experiment. A simple confidence interval for the difference between the mean effect at the NOEC and at the control comes immediately from the test statistic. A 95% confidence interval for that difference is given by

$$\mu_i - \mu_0 = \bar{x}_i - \bar{x}_0 \pm T * s\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}$$

where T is the *2*-sided upper α equicoordinate point of a k-variate t-distribution with correlations defined above, and $df=n_0+n_i+n_i \ldots +n_{k-1}-k$. Simple arithmetic can convert this into a confidence interval for the percent effect. This can also be given in 1-sided form, where appropriate.

Example 5 Dunnett's Test The data from example 1 were re-analysed using Dunnett's test. As shown below, the NOEC is still dose 2 when the analysis unit is the individual subject.

```
Shapiro-Wilk Test for Normality
  SW_STAT      SW_P     SIGNIF
  0.98735    0.92077
```

Data are normally distributed, so a parametric analysis can be done. Levene's test will be done to determine whether to use standard or robust ANOVA.

```
LEVENE TEST FOR  Y
DF     LEVENE    P_VALUE
 5    0.86180    0.51220
```

By Levene's test, the within-group variances are equal. A standard ANOVA will be done.

```
DOSE LSMEAN SERR     DIFF SE_DIFF DF  T    ADP_P  SIGNIF
   0 53.10 1.70910   0.00  .      .     .
  10 52.70 2.29299  -0.40 2.8599 59  -0.14 1.0000
  25 45.20 2.41703  -7.90 2.9602 59  -2.67 0.0443 **
  60 47.10 2.29299  -6.00 2.8599 59  -2.10 0.1653
 150 44.80 2.29299  -8.30 2.8599 59  -2.90 0.0241 **
1000 46.60 2.56364  -6.50 3.0811 59  -2.11 0.1613
```

The above analysis used individual daphnid as the experimental unit for analysis. If the analysis is repeated assuming the subgroup structure as shown in replicate A, the result is quite different, as discussed under the Jonckheere-Terpstra test example.

Dunn's Test - While Dunn's test is often described in texts as a way of comparing all possible pairs of treatment groups, her original paper (Dunn, 1964) provided a means of estimating any number of general contrasts and adjusting for the number of contrasts estimated. For present purposes, we shall describe only comparisons of treatments to a common control. The procedure is to rank the combined treatment groups, using the mean rank for tied responses. Compute the mean rank, R_i, for each treatment group. Compute the combined sample size, N, and the individual sample sizes n_i. Finally, compute the variance of $R_i - R_0$ as follows.

$$V_{i0} = \left[\frac{N(N+1)}{12} - \frac{\sum(t^3 - t)}{12(N-1)}\right]\left[\frac{1}{n_i} + \frac{1}{n_0}\right]$$

where the sum is over all distinct responses and t is the number of observations tied at that response.
The test statistic

$$Z = \frac{R_i - R_0}{\sqrt{V_{i0}}}$$

is compared to a standard normal distribution at $p=1-\alpha/k$, where k is the number of comparisons to control and α is 0.05 or 0.025, according as a 1- or 2-sided test is used.

Advantages of Dunn's test. This test is based on ranks, and thus is robust against a wide variety of distributions and heteroscedasticity. It is also flexible, so that arbitrary contrasts can be tested, not just comparisons of treatments to control. However, this latter is no advantage in a standard dose-response experiment.

Disadvantages of Dunn's test. This test does not permit modelling multiple sources of variances (e.g., within- and between-subgroups). The Bonferroni-Holm adjustment for multiple comparisons is statistically conservative. There is no exact permutation counterpart to this test, so it is not useful for very small samples or experiments with massive ties among the response values.

Power of Dunn's Test. Dunn (1964) describes an extreme situation, where all but one of k population means equal the control, but that every observation in that treatment group exceeds every observation in the control. The power of her 2-sided test to detect this difference is approximately

$$1 - \Phi\left[\frac{z_{1-\alpha/2k}\sqrt{\frac{k(kn+1)}{6}} - kn/2}{\sqrt{\frac{[(k-1)n+1][k-2]}{12}}}\right],$$

where n is the common within-group sample size. For a 1-sided test, the divisor $2k$ in the subscript of z is replaced by k.
The power of Dunn's test depends on the specific alternative, as is the case for other procedures. Nevertheless, the above provides an approximate power that can be used for design assessment purposes.

Mann-Whitney Test -The Mann-Whitney rank sum test compares the ranks of measurements in two independent random samples of n_1 and n_2 observations and aims to detect whether the distribution of values from one group is shifted with respect to the distribution of values from the other. It is equivalent to the Wilcoxon test.
To use the Mann-Whitney rank sum test, we first rank all $(n_1 + n_2)$ observations, assigning a rank of *1* to the smallest, *2* to the second smallest, and so on. Tied observations (if they occur) are assigned ranks equal to the average of the ranks of the tied observations. Then the ranks of the observations in each group are summed and designated as T_1 and T_2. If the distributions in the two groups are identical then T_1 and T_2 would be identical. If the two distributions differ, then the difference between T_1 and T_2 will be dissimilar, with the rank sums indicating the degree of overlap between the groups. There are one tailed and two tailed versions of the test, as well as small sample and large sample (asymptotic) versions.
For the one tailed test, the test statistic is T_1 if $n_1 < n_2$ or T_2 if $n_2 \geq n_1$. The rejection region is $T_1 \geq T_U$, if T_1 is test statistic; or $T_2 \leq T_L$, if T_1 is test statistic. The values of T_U and T_L can be obtained from tables available in many statistics texts.
If $n_1 \geq 10$ and $n_2 \geq 10$, then the large sample approximation can be used. In this case the test statistic is

$$z = \frac{T_1 - \left[\frac{n_1 n_2 + n_1(n_1+1)}{2}\right]}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

The rejection region is $z < -z_\alpha$ *(or* $z > z_\alpha$*)* for the one tailed tests and $z < -z_{\alpha/2}$ *or* $z > z_{\alpha/2}$ for the two tailed test. If more than one comparison is to be made, then an adjustment in the critical values is made, such as the Bonferroni-Holm correction discussed in 5.1.4. Tables of the probability of the *z* statistic are available in most statistics texts.
The assumptions of the Mann-Whitney rank-sum test are only that samples must be independent, and distributions of values must be symmetrical. The Mann-Whitney is nearly as efficient as the *t*-test when data are normally distributed with equal variances, and is generally more efficient when the test data fail to meet the assumptions for the *t*- test.

Advantages of Mann-Whitney test. This test is based on ranks, and thus is robust against a wide variety of distributions and heteroscedasticity. There is an exact permutation version of this test, so it can be used for very small samples or where there are massive ties in the response values.

Disadvantages of Mann-Whitney test. This test does not permit modelling multiple sources of variances (e.g., within- and between-subgroups). The Bonferroni-Holm adjustment for multiple comparisons is statistically conservative, whereas an unadjusted multiple comparison procedure has an unacceptable type I error. The usual large sample version of this test has low power for small samples.

Power of Mann-Whitney Test An approximate power of the Mann-Whitney test for comparing two normal populations can be derived from the power of the t-test from the fact that the asymptotic relative efficiency (A.R.E) of the Mann-Whitney relative to the *t-test* is about 0.95. The meaning of this is that the sample size, Nu, required to give an equal power using the Mann-Whitney as using the t-test is Nu/A.R.E. It should be cautioned that this is an asymptotic result and it obviously does not apply when the distribution is not normally distributed. There are a number of computer programs that will compute power of this test. These include StatXact, Pass, and Study Size. The second of these can be found at http://www.Dataxion.com. The third is available at CreoStat@StudySize.com. A simple search of the internet will locate several such sites, some of which have free down loads.

General Admonition: An additional practical caution for practitioners concerns the difficulty of explaining certain technical details to non-statisticians. For example, means are easier to explain than medians and untransformed responses are much easier to communicate than transformed ones. Plots can sometimes convey more than many paragraphs. In general, technical details should be presented sparingly. For example, explaining amalgamated means to non-statisticians has on some occasions led to serious and unhelpful resistance to acceptance of conclusions. It is probably best to avoid such details in reports and presentations.
As these examples show, we should not conclude that one method invariably gives lower NOECs than the others, nor should this be the criterion by which to judge a method. The selection of a method should be driven by biological and statistical considerations and then, a consistent analysis protocol should be used to avoid the appearance of selecting a method to give the desired answer. What this example suggests is that care must be taken in selecting statistical methods. In general, preference is given to step-down methods that actively use the expected monotone dose-response nature of the data.

5.4 Power of Step-Down Jonckheere-Terpstra Test

Interpretation of the power curves The vertical axis is the power or probability (expressed as a percent) of finding a significant effect if the true effect is of the magnitude given on the horizontal axis. For a given sample size, the first power curve is for the first step in a 6-concentration experiment and is labelled PCT61. That is, it shows the power of finding a significant effect at the high concentration of an experiment with six concentrations plus control, when the true effect at that concentration is as shown. The horizontal axis is in terms of standard deviation units. Thus, delta=1 means the true effect is 1 standard deviation away from the mean. The horizontal axis is labelled delta6, delta5, etc, but the numerical suffix is software generated for reasons of no importance here and should be ignored.

The second plot for a given sample size shows two curves. One is for the second step of a six-concentration experiment and is labelled PCT62. That is, it gives the power of finding a significant effect at the second highest concentration of an experiment with six concentrations plus control, when the true effect at that concentration is as shown. The other is for the first step of a 5-concentration experiment and is labelled PCT51. That is, it gives the power of finding a significant effect at the highest concentration of an experiment with five concentrations plus control, when the true effect at that concentration is as shown.

The third plot of a given sample size is for the third step in a six-concentration experiment and is labelled PCT63, the second step in a 5-concentration experiment (PCT52) and the first step in a 4-concentration experiment (PCT41).

To use these plots to determine the percent change from control it should be possible to detect with a given experiment, we need additional information. Suppose we want the power of detecting a 100p% change from the mean. Thus, if μ_0 is the control mean andμ_1 is a treatment mean, we want the power of detecting a shift of

$$\mu_1 - \mu_0 = p\mu_0,$$

or

$$\mu_1 - \mu_0 = p\sigma\frac{\mu_0}{\sigma} = \frac{p}{CV}\sigma,$$

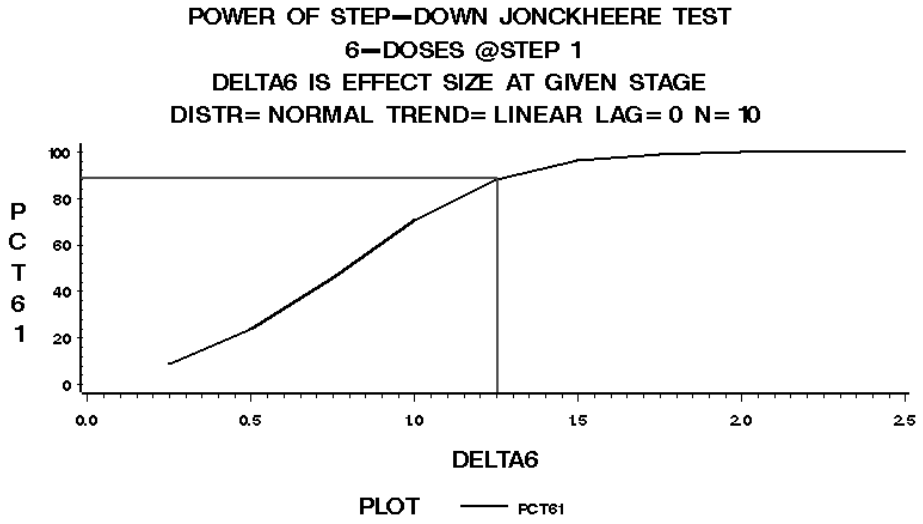
$$where\ CV = \frac{\sigma}{\mu}.$$

Here CV is the coefficient of variation. (Note: the coefficient of variation is often expressed as a percent, which is CV (as written) multiplied by 100%.) Example: Determine the power of detecting a 25% change from the control when the CV=.2 (or 20%). The shift in standard deviation units is given by

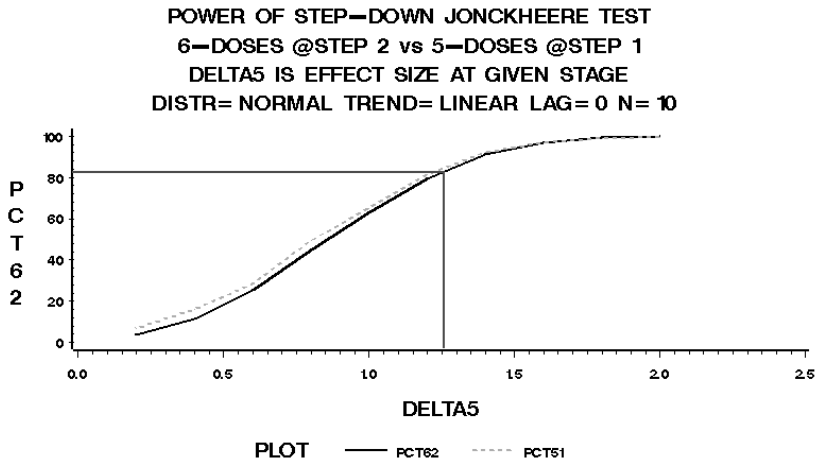
$$shift = \frac{p}{CV} = \frac{.25}{.2} = 1.25.$$

The power can then be read from the plot using DELTA=1.25 on the horizontal axis.

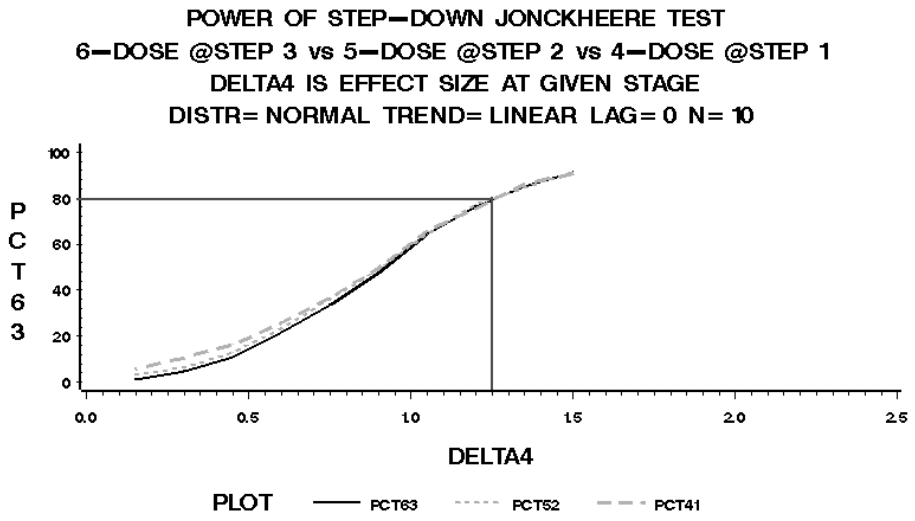
While it might have been helpful to have presented the power curves in terms of percent shift from the control mean directly, it would have been necessary to present power curves for a range of CVs, which would require much more space. Consider a 6-dose experiment (5 positive concentrations plus copntrol) with 10 experimental units per concentration. From the first of the plots below (reproduced from below), one sees that the power of declaring significant an effect at the highest concentration is approximately 88%.



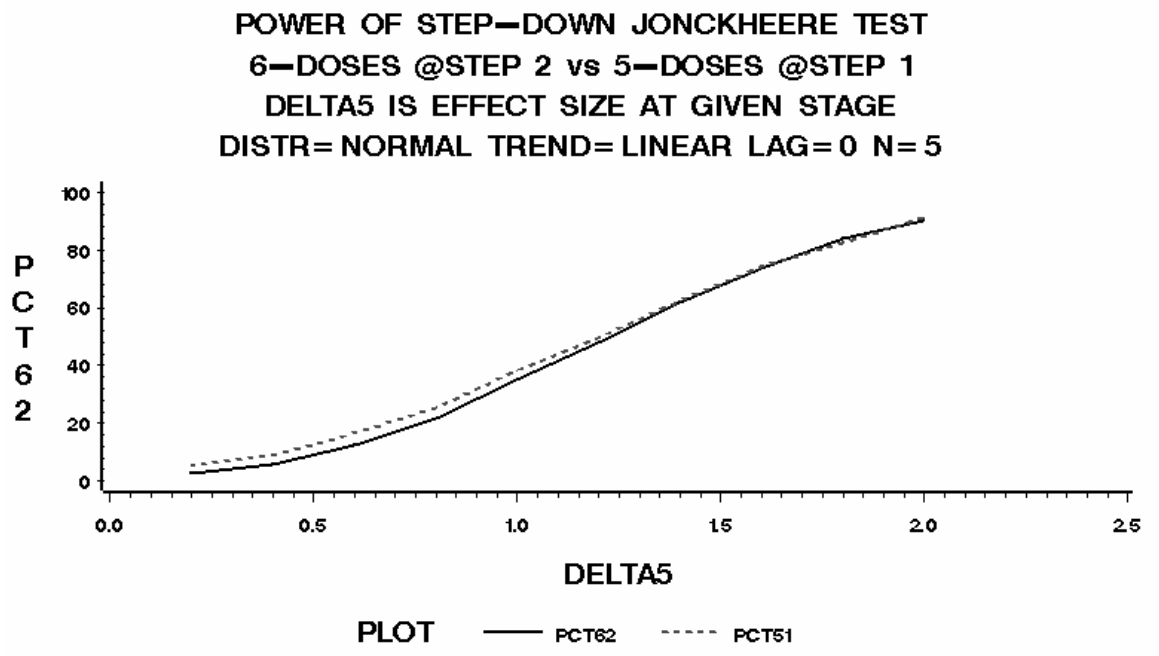
The power of declaring significant this size effect in the second highest concentration is obtained from the next plot and is seen to be about 84%.

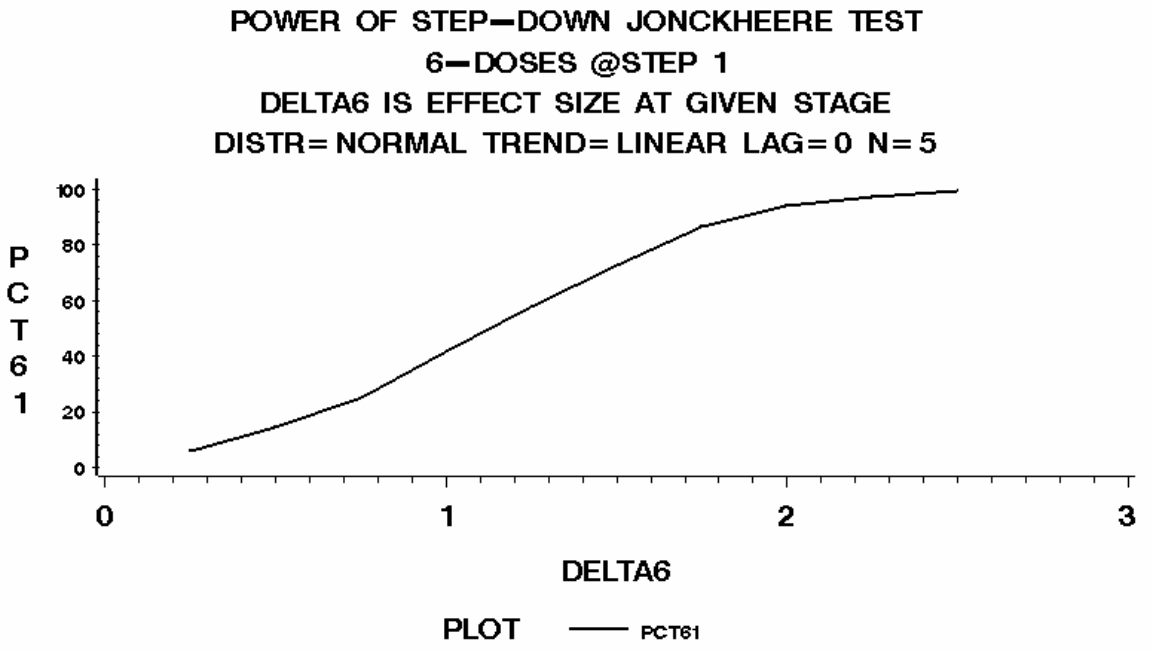


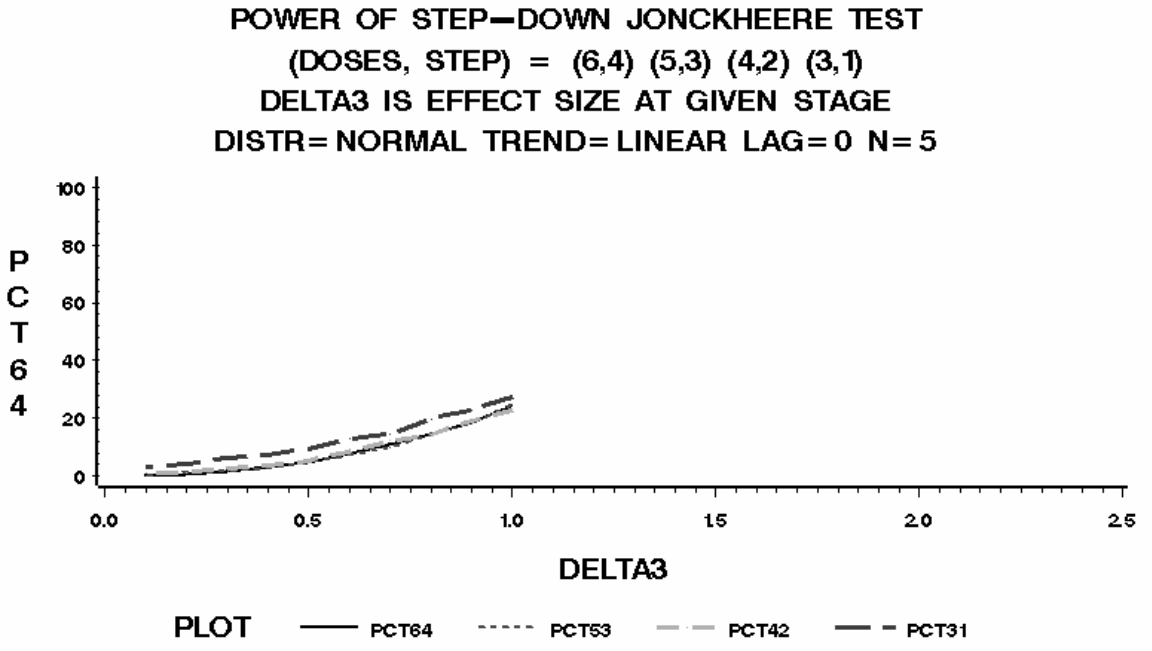
The power of declaring significant this sized effect at the third highest concentration is about 80%, as seen from the next plot. So, there is small drom in power as one steps down through the concentrations. This is to be expected, as there is less information in the reduced number of copncentrations under investigation. This plot also shows that it is not the step-down process itself that accounts for the small decrease in power., since theis plot also shows that the power to detect this size effect in the second highest concentration in a 5-dose experiment and the power to detect this sized effect in at the highest concentration in a 4-dose experiment are very nearly the same. Rather, it is the number of concentrations remaining at this stage.

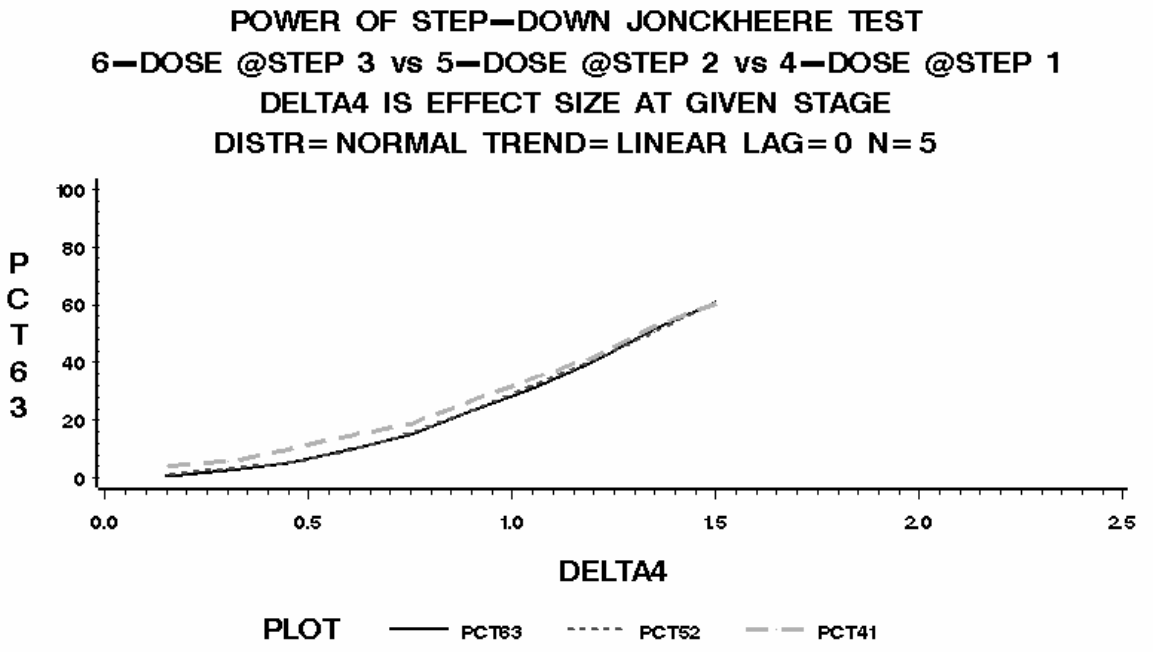


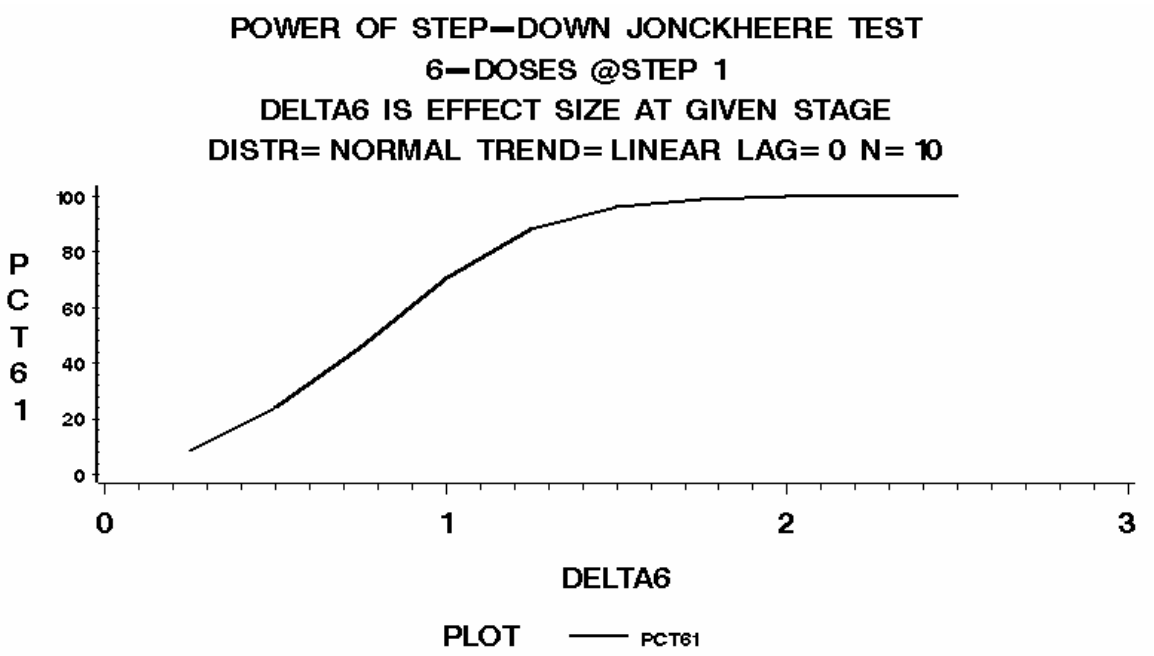
Additional information, and another example, on the use of these power curves is given in Annex 5.3 in the discussion of the Jonckheere-Terpstra test.

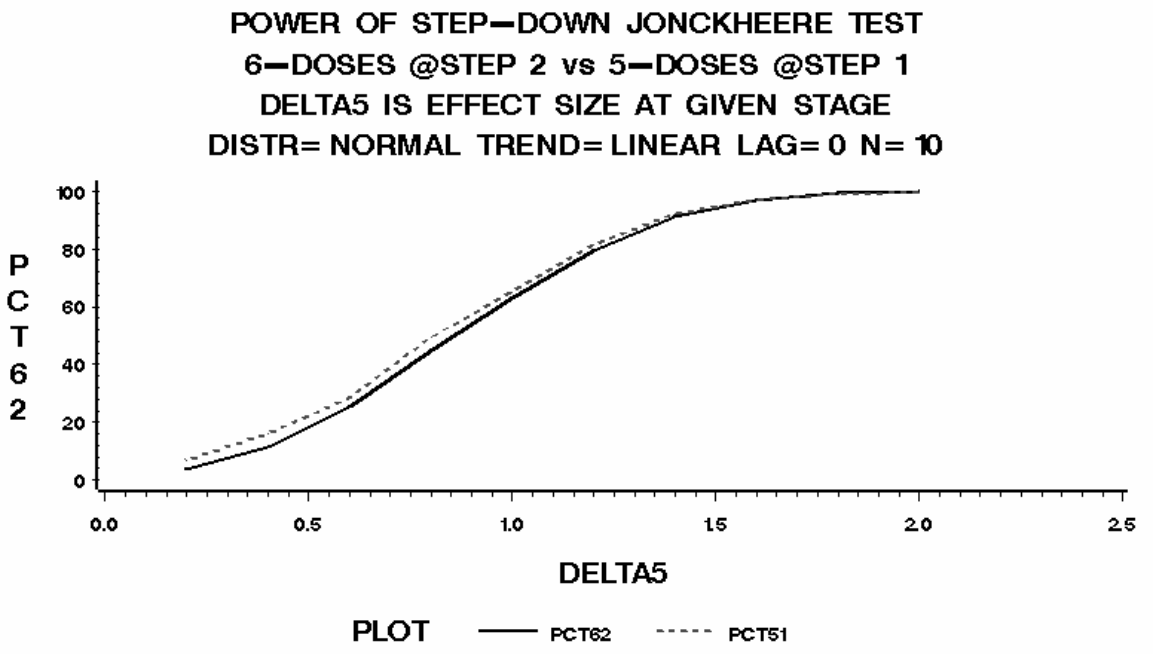


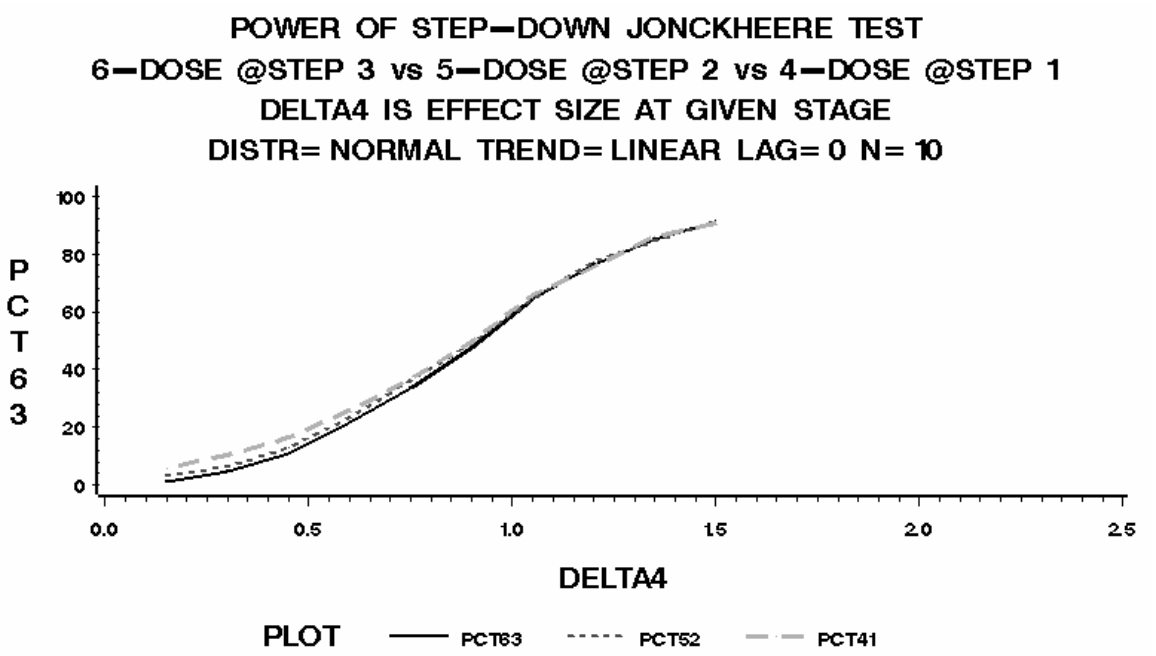

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 1
DELTA6 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 5
PCT61
100
80
60
40
20
0
0
1
2
3
DELTA6
PLOT
PCT61

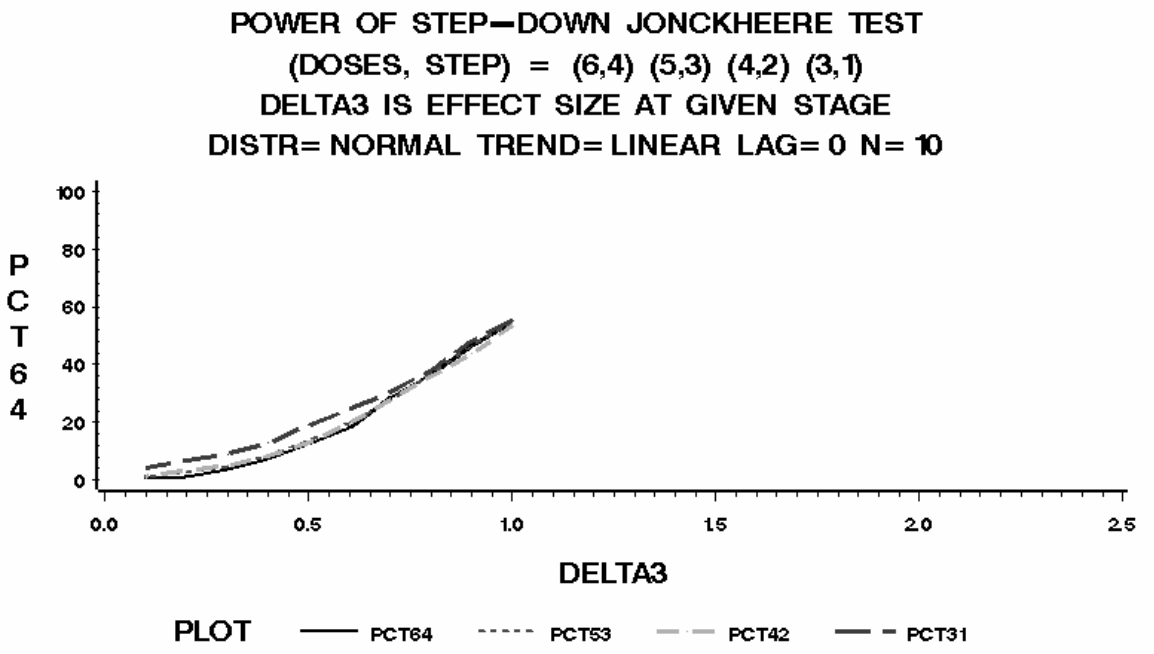

POWER OF STEP—DOWN JONCKHEERE TEST
(DOSES, STEP) = (6,4) (5,3) (4,2) (3,1)
DELTA3 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 5
PCT64
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA3
PLOT
PCT64
PCT53
PCT42
PCT31

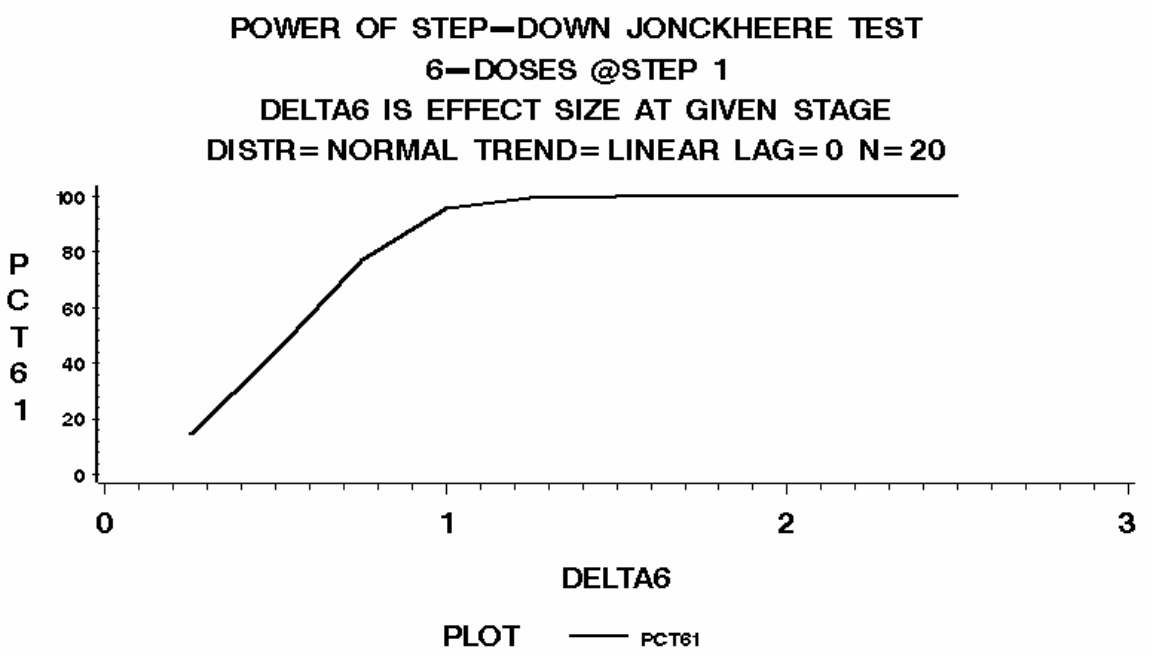

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSE @STEP 3 vs 5—DOSE @STEP 2 vs 4—DOSE @STEP 1
DELTA4 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 5
PCT63
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA4
PLOT
PCT63
PCT52
PCT41

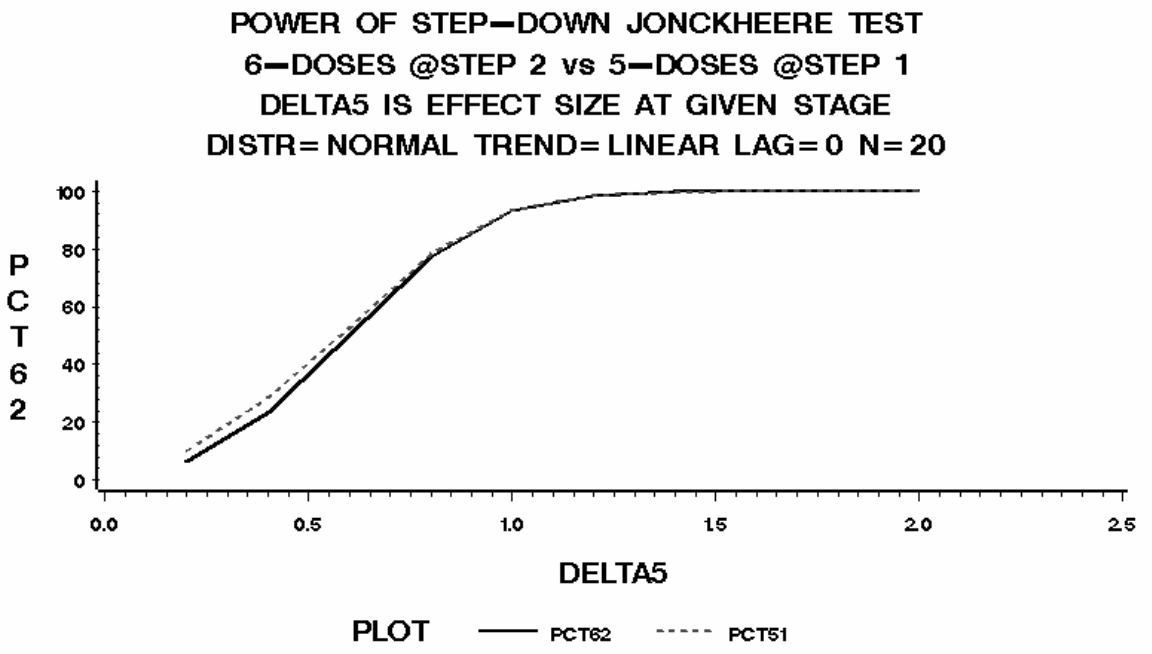

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 1
DELTA6 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 10
PCT61
100
80
60
40
20
0
0
1
2
3
DELTA6
PLOT
PCT61

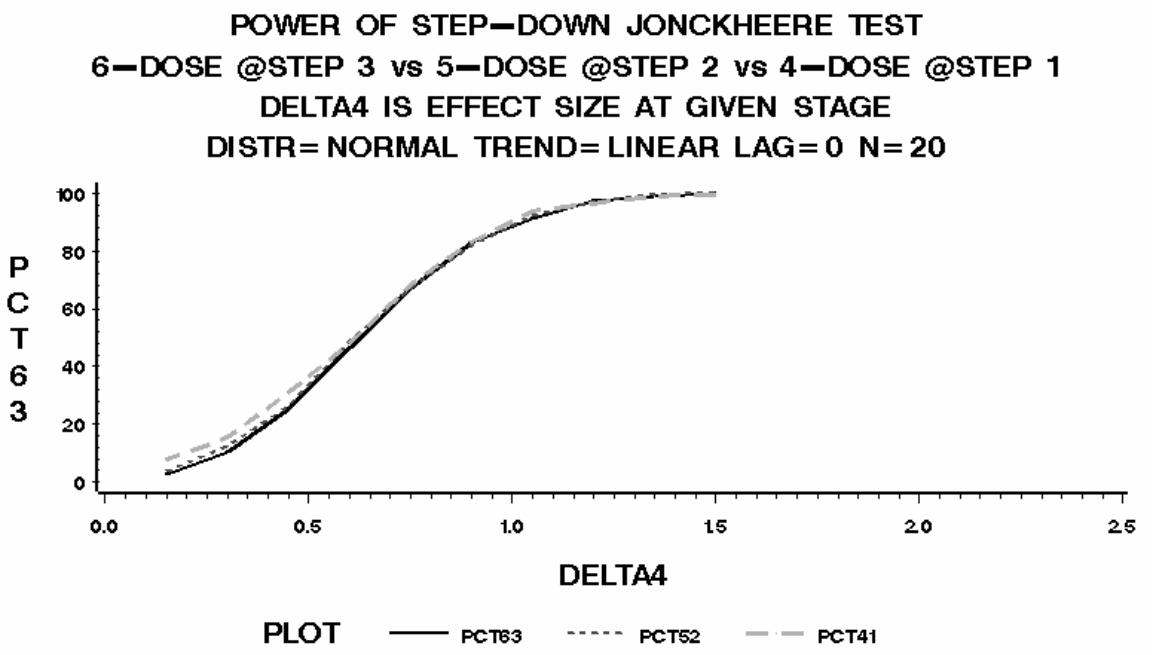

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 2 vs 5—DOSES @STEP 1
DELTA5 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 10
PCT62
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA5
PLOT
PCT62
PCT51

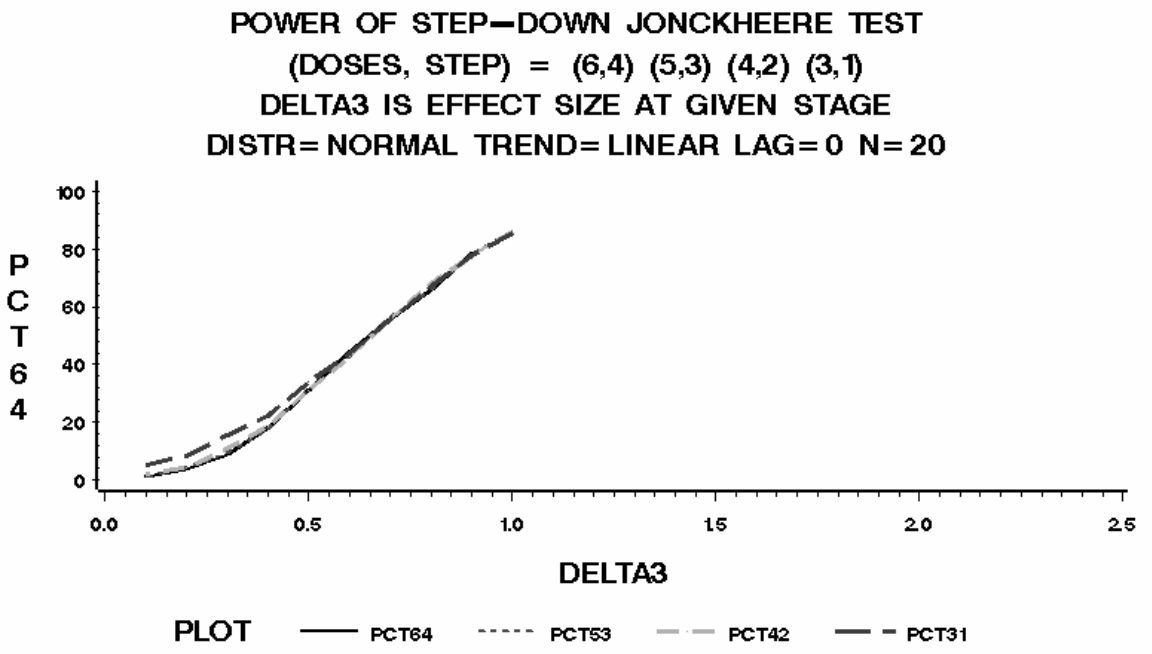

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSE @STEP 3 vs 5—DOSE @STEP 2 vs 4—DOSE @STEP 1
DELTA4 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 10
PCT63
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA4
PLOT
PCT63
PCT52
PCT41

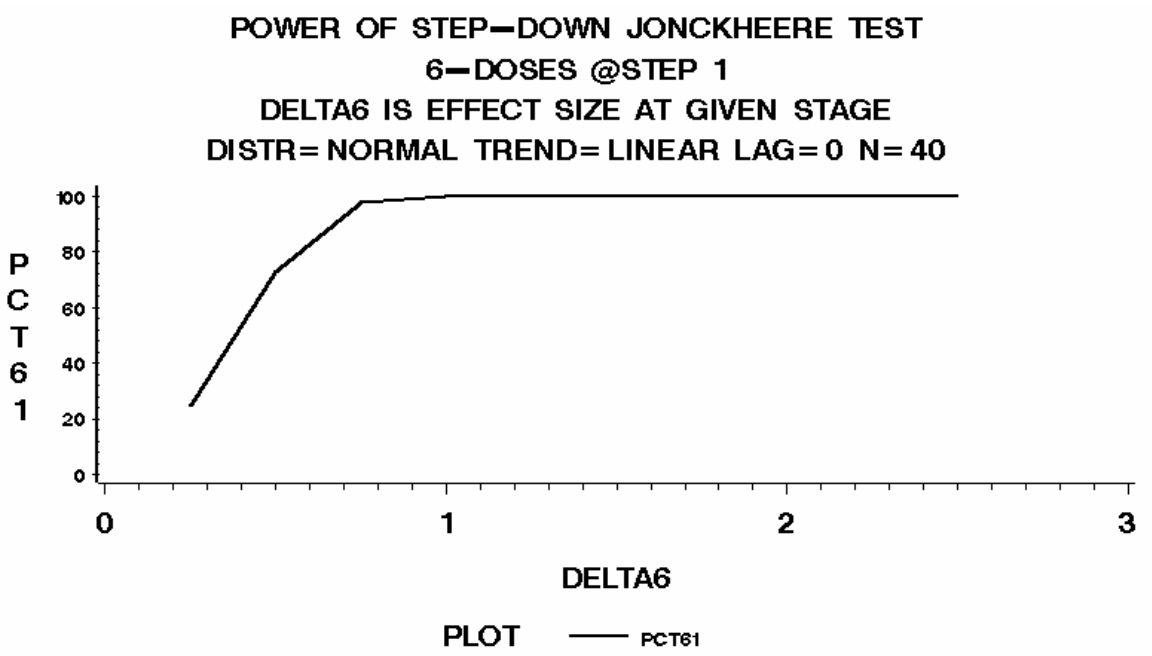

POWER OF STEP–DOWN JONCKHEERE TEST
(DOSES, STEP) = (6,4) (5,3) (4,2) (3,1)
DELTA3 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 10
PCT64
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA3
PLOT
PCT64
PCT53
PCT42
PCT31

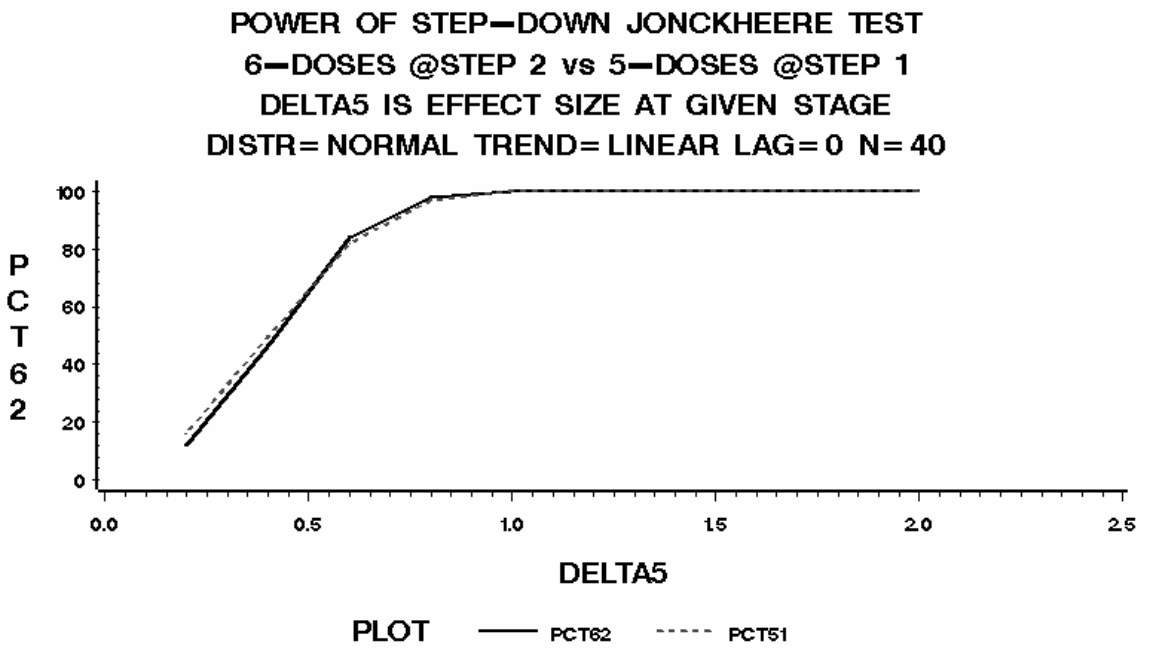

POWER OF STEP–DOWN JONCKHEERE TEST
6–DOSES @STEP 1
DELTA6 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 20
PCT61
100
80
60
40
20
0
0
1
2
3
DELTA6
PLOT
PCT61

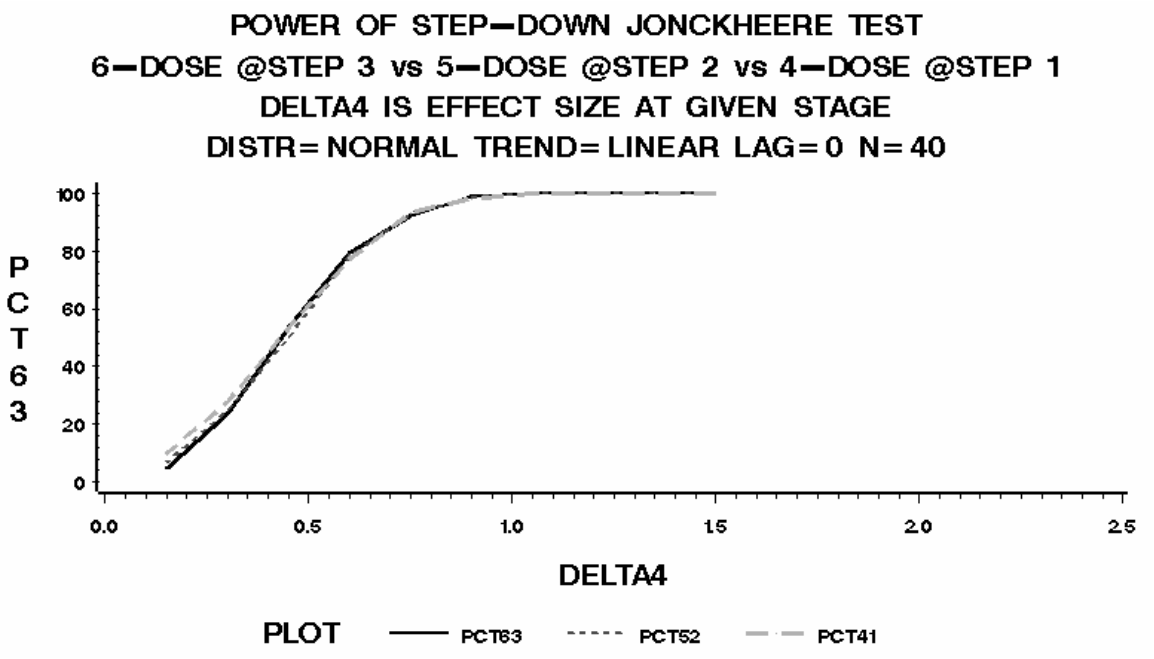

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 2 vs 5—DOSES @STEP 1
DELTA5 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=20
PCT62
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA5
PLOT
PCT62
PCT51

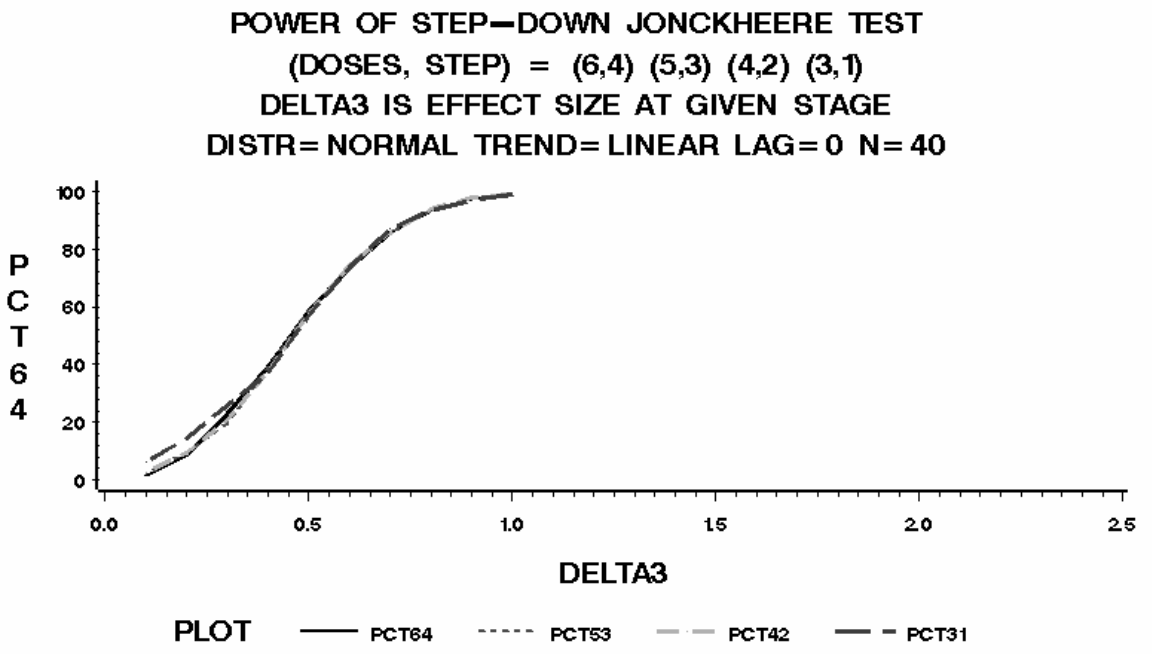

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSE @STEP 3 vs 5—DOSE @STEP 2 vs 4—DOSE @STEP 1
DELTA4 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=20
PCT63
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA4
PLOT
PCT63
PCT52
PCT41

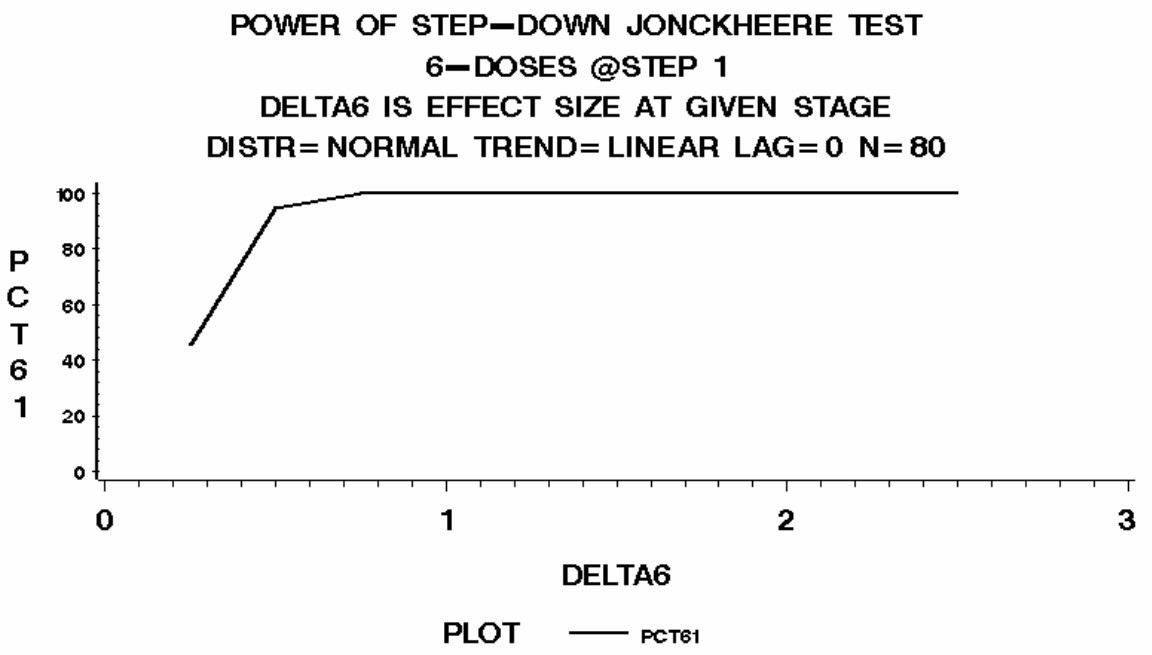

POWER OF STEP–DOWN JONCKHEERE TEST
(DOSES, STEP) = (6,4) (5,3) (4,2) (3,1)
DELTA3 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=20
PCT64
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA3
PLOT
PCT64
PCT53
PCT42
PCT31

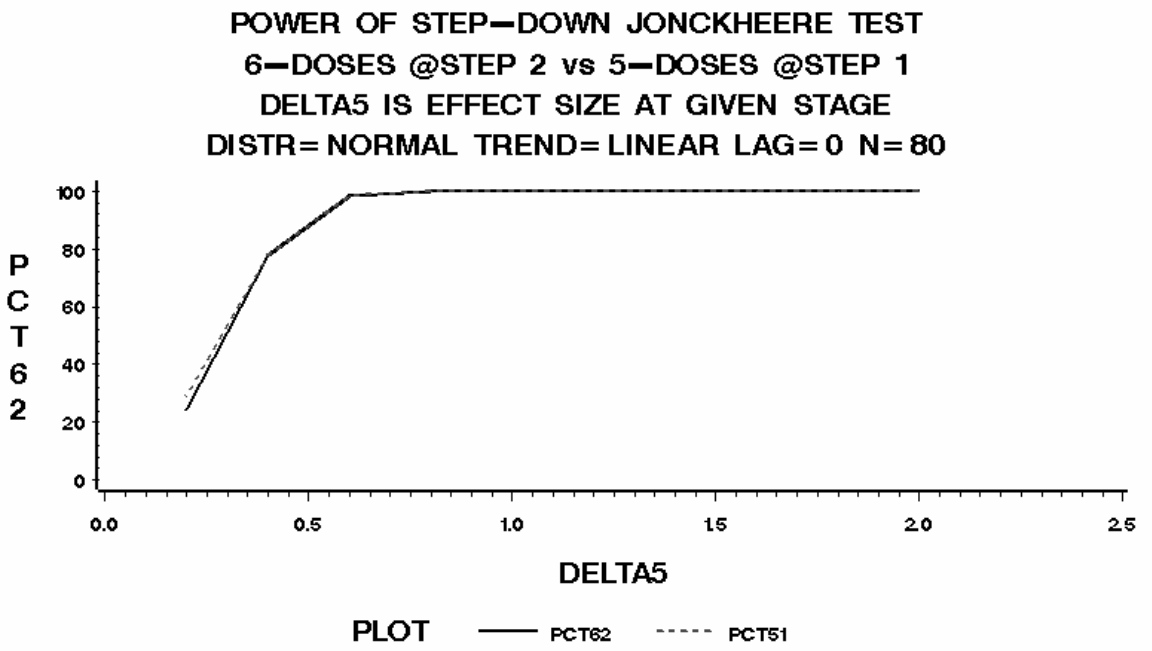

POWER OF STEP–DOWN JONCKHEERE TEST
6–DOSES @STEP 1
DELTA6 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=40
PCT61
100
80
60
40
20
0
0
1
2
3
DELTA6
PLOT
PCT61

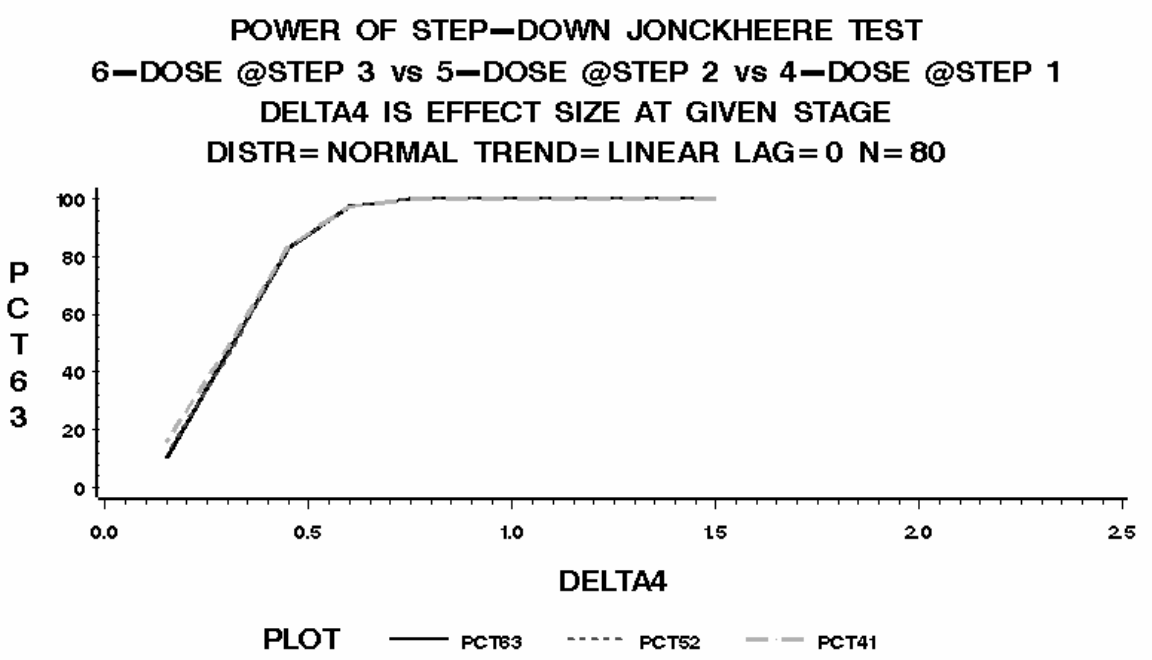

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 2 vs 5—DOSES @STEP 1
DELTA5 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 40
PCT62
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA5
PLOT
PCT62
PCT51

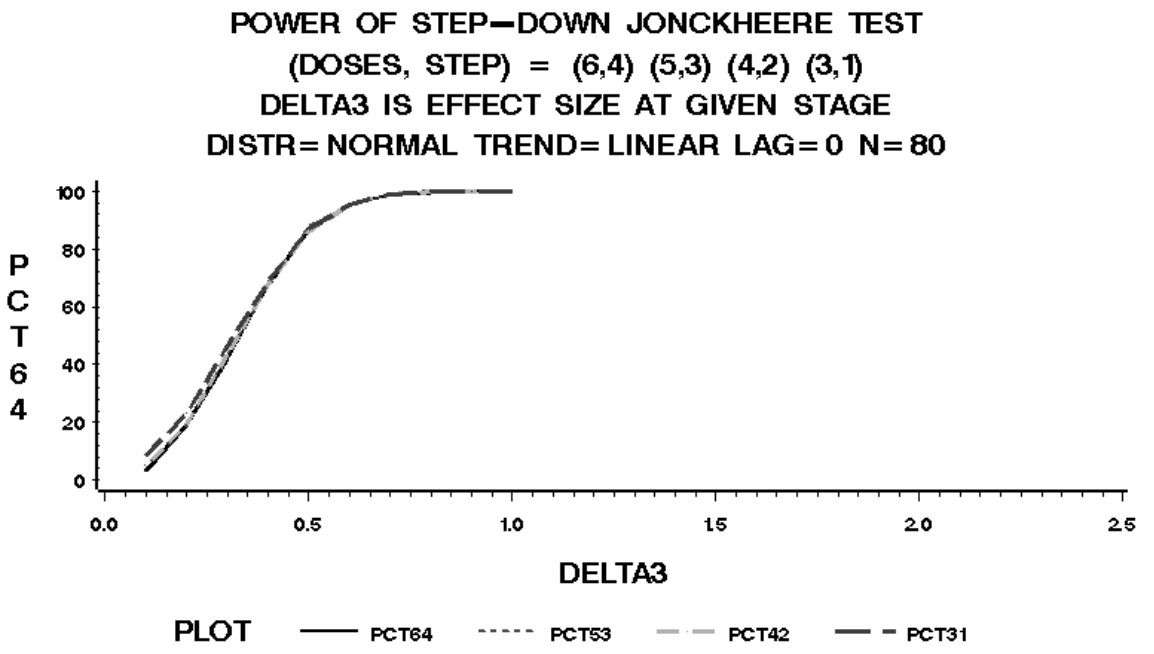

POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSE @STEP 3 vs 5—DOSE @STEP 2 vs 4—DOSE @STEP 1
DELTA4 IS EFFECT SIZE AT GIVEN STAGE
DISTR= NORMAL TREND= LINEAR LAG= 0 N= 40
PCT63
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA4
PLOT
PCT63
PCT52
PCT41


POWER OF STEP—DOWN JONCKHEERE TEST
(DOSES, STEP) = (6,4) (5,3) (4,2) (3,1)
DELTA3 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=40
PCT64
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA3
PLOT
PCT64
PCT53
PCT42
PCT31


POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 1
DELTA6 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=80
PCT61
100
80
60
40
20
0
0
1
2
3
DELTA6
PLOT
PCT61


POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSES @STEP 2 vs 5—DOSES @STEP 1
DELTA5 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=80
PCT62
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA5
PLOT PCT62 PCT51


POWER OF STEP—DOWN JONCKHEERE TEST
6—DOSE @STEP 3 vs 5—DOSE @STEP 2 vs 4—DOSE @STEP 1
DELTA4 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=80
PCT63
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA4
PLOT PCT63 PCT52 PCT41


POWER OF STEP–DOWN JONCKHEERE TEST
(DOSES, STEP) = (6,4) (5,3) (4,2) (3,1)
DELTA3 IS EFFECT SIZE AT GIVEN STAGE
DISTR=NORMAL TREND=LINEAR LAG=0 N=80
PCT64
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
2.5
DELTA3
PLOT
PCT64
PCT53
PCT42
PCT31

ANNEX 6: ANNEXE TO CHAPTER 7 "BIOLOGY-BASED METHODS"

This appendix specifies the models for bioassays on survival/immobilisation, body growth, reproduction and population growth. Biology-based methods put emphasis on the story behind the model, rather than the model itself; the derivation from underlying mechanistic assumptions is not given here, however. The assumptions themselves are given in the main text.

The dimensions and interpretations of all variables and parameters are given in tables for each type of bioassay. The dimensions are indicated with symbols that have the following interpretation

symbol	*interpretation*
-	dimensionless
t	time
mol	mole
l	length
#	number

Effects on survival

The target parameter is the hazard rate. At time $t_0 = -k_e^{-1} \ln\{1 - c_0 / c\}$ the survival probability starts to deviate from the control for $c > c_0$. The survival probability is given by

$q(t,c) = \exp\{-h_0 t + c(\exp\{-k_e t_0\} - \exp\{-k_e t\})b_k / k_e - b_k (c - c_0)(t - t_0)\}$ if $c > c_0$ and $t > t_0$

$q(t,c) = \exp\{-h_0 t\}$ if $c < c_0$ or $t < t_0$

DEBtox estimates up to four parameters from bioassay data. The variables and parameters are

variables	*dimension*	*interpretation*
t	t	exposure time
c	$mol\ l^{-3}$	external concentration
q	-	survival probability
Parameters		
h_0	t^{-1}	control mortality rate
c_0	$mol\ l^{-3}$	NEC
b_k	$mol^{-1}\ l^3\ t^{-1}$	killing rate
k_e	t^{-1}	elimination rate

Effects on body growth

Growth depends on the concentration of the compound in the tissue. This concentration is treated as a hidden variable and scaled to remove a parameter (the BioConcentration Factor BCF). The scaled tissue-concentration c_q relates to the tissue-concentration C_q as $c_q = C_q$ / BCF; the scaled tissue-concentration has the dimension of an external concentration, but is proportional to the tissue-concentration. The change in scaled tissue-concentration is

$$\tfrac{d}{dt} c_q = k_e \left(c - c_q - \frac{3c_q}{k_e L_m} \tfrac{d}{dt} L \right) \frac{L_m}{L} \text{ with } c_q(0) = 0 .$$

The third term in the second factor accounts for the dilution by growth; the change in body length depends on the mode of action of the compound and is specified below. The three modes of action are expressed in terms of the dimensionless "stress" function

$$s(c_q) = c_*^{-1} \max\{0, c_q - c_0\}$$

The modes of action are

- Direct effects on body growth: target parameter is the conversion efficiency from reserve to structure

$$\tfrac{d}{dt} L = r_B (L_m - L) \frac{1+g}{1+g(1+s(c_q))} \text{ with } L(0) = L_0 .$$

- Effects on maintenance: target parameter is the specific maintenance costs.

$$\tfrac{d}{dt} L = r_B \left(L_m - L(1+s(c_q)) \right) \text{ with } L(0) = L_0 .$$

- Effects on assimilation: target parameter is the maximum specific assimilation rate

$$\tfrac{d}{dt} L = r_B \left(L_m (1 - s(c_q)) - L \right) \text{ with } L(0) = L_0 .$$

DEBtox fixes three parameters at default values, and estimates up to four parameters from bioassay data. The variables and parameters are

variables	*dimension*	*fix*	*Interpretation*
t	t		exposure time
c	$mol\ l^{-3}$		external concentration
c_q	$mol\ l^{-3}$		scaled internal concentration
L	l		body length
$s(c_q)$	-		stress function
Parameters			
L_0	l	+	initial body length
L_m	l	-	maximum body length
g	-	+	energy investment ratio
r_B	t^{-1}	+	von Bertalannfy growth rate
c_0	$mol\ l^{-3}$	-	NEC
c_*	$mol\ l^{-3}$	-	tolerance conconcentration
k_e	t^{-1}	-	Elimination rate

Effects on reproduction

Body length is treated as a hidden variable and scaled to remove a parameter (maximum length L_m); scaled length relates to length as $l = L / L_m$. The reproduction rates are given as a function of scaled length, and external concentration. The scaled length and scaled internal concentration are given as differential equations. Their solutions are functions of time and external concentration. The endpoint in the *Daphnia* reproduction bioassay is the cumulated number of offspring, rather than the reproduction rate. This number *N* relates to the reproduction rate *R* as

$$N(t,c) = \int_0^t R\big(l(s),c\big)ds \text{ or } \tfrac{d}{dt}N = R\big(l(t),c\big) \text{ with } N(0,c) = 0.$$

The reproduction rate in the control amounts to

$$R(l,0) = \frac{R_m}{1-l_p^3}\left(\frac{g+l}{g+1}l^2 - l_p^3\right) \text{ for } l > l_p;\ R(l,0) = 0 \text{ for } l < l_p, \text{ where } l_p = L_p / L_m$$

Growth and reproduction depend on the concentration of the compound in the tissue. The change in scaled tissue-concentration is

$$\tfrac{d}{dt}c_q = k_e\big(c - c_q - 3c_q k_e^{-1}\tfrac{d}{dt}l\big)/l \text{ with } c_q(0) = 0.$$

The third term in the second factor accounts for the dilution by growth. The reproduction and growth rates depend on the mode of action, and are specified below. The effects are expressed in terms of the dimensionless "stress" function

$$s(c_q) = c_*^{-1}\max\{0, c_q - c_0\}$$

For indirect effects on reproduction (namely via effects on assimilation, maintenance or growth), the change in scaled body length and the reproduction rate $R(l,c)$ at scaled body length l and external concentration c are

- for effects on assimilation (target parameter is the maximum specific assimilation rate)

  $$\tfrac{d}{dt}l = r_B(1 - s(c_q) - l) \text{ with } l(0) = l_0$$

  $$R(l,c) = (1 - s(c_q))^3 R(l,0) \text{ for } l > l_p$$

- for effects on maintenance (target parameter is the specific maintenance costs)

  $$\tfrac{d}{dt}l = r_B(1 - l(1 + s(c_q))) \text{ with } l(0) = l_0$$

  $$R(l,c) = (1 + s(c_q))^{-2} R(l,0) \text{ for } l > l_p$$

- for effects on growth (target parameter is the conversion efficiency from reserve to structure)

  $$\tfrac{d}{dt}l = r_B(1-l)\frac{1+g}{1+g(s(c_q))} \text{ with } l(0) = l_0$$

  $$R(l) = \frac{R_m}{1-l_p^3}\left(\frac{(1+s(c_p))g + l}{(1+s(c_p))g + 1}l^2 - l_p^3\right) \text{ for } l > l_p$$

For direct effects on reproduction, the body growth is not affected and reduces to

$$\tfrac{d}{dt}L = r_B(L_m - L) \text{ with } L(0) = L_0, \text{ or } L(t) = L_m - (L_m - L_0)\exp\{-r_B t\}.$$

In scaled body length have

$\frac{d}{dt} l = r_B (1 - l)$ with $l(0) = l_0$, or $l(t) = 1 - (1 - l_0) \exp\{-r_B t\}$

Two types of direct effects on reproduction are delineated:

- for effects on the survival of (early) offspring (target parameter is the hazard rate of offspring):

  $R(l, c) = R(l, 0) \exp\{-s(c_q)\}$ for $l > l_p$

- for effects on the costs for reproduction (target parameter is the conversion efficiency of reserve from mother to offspring):

  $R(l, c) = R(l, 0)\left(1 + s(c_q)\right)^{-1}$ for $l > l_p$

DEBtox fixes four parameters at default values, and estimates up to four parameters from bioassay data. The variables and parameters are

variables	*dimension*	*fix*	*interpretation*
t	t		exposure time
c	$mol\ l^{-3}$		external concentration
c_q	$mol\ l^{-3}$		scaled internal concentration
l	-		scaled body length
$s(c_q)$	-		stress function
Parameters			
l_0	-	+	initial scaled body length
l_p	-	+	scaled body length at onset reproduction
g	-	+	energy investment ratio
r_B	t^{-1}	+	von Bertalannfy growth rate
R_m	$\#t^{-1}$	-	maximum reproduction rate
c_0	$mol\ l^{-3}$	-	NEC
c_*	$mol\ l^{-3}$	-	tolerance concentration
k_e	t^{-1}	-	elimination rate

Effects on population growth

The number of individuals in a population is partitioned into living and dead ones; the total number is counted or measured. The internal concentration is taken to be proportional to the external one, so the stress function can be written as

$$s(c) = c_*^{-1} \max\{0, c - c_0\}$$

Three modes of action are delineated:

- effects on growth costs

$$N(t,c) = N(0,c)\exp\{r(c)t\} \text{ with } r(c) = r_0\left(1 + s(c)\right)^{-1}$$

- effects on survival (during growth)

$$N(t,c) = N(0,c)\left(\frac{r(0)}{r(c)}\exp\{r(c)t\} + 1 - \frac{r(0)}{r(c)}\right) \text{ with } r(c) = r_0\left(1 + s(c)\right)^{-1}$$

- effects on adaptation (i.e. on survival at the start only)

$$N(t,c) = N(0,c)\left(\exp\{r_0 t - s(c)\} + 1 - \exp\{-s(c)\}\right)$$

DEBtox estimates up to four parameters from bioassay data. The variables and parameters ard

variables	*dimension*	*Interpretation*
t	t	exposure time
c	$mol\ l^{-3}$	external concentration
$s(c)$	-	stress function
Parameters		
N(0,c)	$\#\ l^{-3}$	inoculum size at concentration c
r_0	t^{-1}	control specific pop. growth rate
c_0	$mol\ l^{-3}$	NEC
c_*	$mol\ l^{-3}$	tolerance concentration