
Water quality — Guidance on statistical interpretation of ecotoxicity data

Qualité de l'eau — Lignes directrices relatives à l'interprétation statistique de données écotoxicologiques



PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

© ISO 2006

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

ii

Contents

Page

Foreword.....	xii
Introduction	xiii
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 General statistical principles	8
4.1 Different statistical approaches	8
4.1.1 General.....	8
4.1.2 Hypothesis-testing methods	8
4.1.3 Concentration-response modelling methods	10
4.1.4 Biology-based methods	11
4.2 Experimental design issues	11
4.2.1 General.....	11
4.2.2 NOEC or EC _x : Implications for design.....	12
4.2.3 Randomization	12
4.2.4 Replication.....	13
4.2.5 Multiple controls included in the experimental design.....	13
4.3 Process of data analysis.....	14
4.3.1 General.....	14
4.3.2 Data inspection and outliers.....	14
4.3.3 Data inspection and assumptions	15
4.3.3.1 Scatter	15
4.3.3.2 Heterogeneous variances and distribution	15
4.3.3.3 Heterogeneous variances and true variation in response.....	16
4.3.3.4 Consequences for the analysis	16
4.3.4 Transformation of data.....	16
4.3.5 Parametric and non-parametric methods	17
4.3.5.1 General	17
4.3.5.2 Parametric methods.....	17
4.3.5.3 Generalized linear models (GLMs)	18
4.3.5.4 Non-parametric methods.....	18
4.3.5.5 How to choose?.....	18
4.3.6 Pre-treatment of data.....	19
4.3.7 Model fitting.....	19
4.3.8 Model checking	20
4.3.8.1 Analysis of residuals	20
4.3.8.2 Validation of fitted dose-response model	21
4.3.9 Reporting the results.....	21
5 Hypothesis testing.....	21
5.1 Introduction	21
5.1.1 General.....	21
5.1.2 NOEC: What it is, and what it is not.....	25
5.1.3 Hypothesis used to determine NOEC	25
5.1.3.1 Understanding the question to be answered	25
5.1.3.2 One-sided hypothesis.....	26
5.1.3.3 Two-sided trend test	26
5.1.3.4 Trend or pair-wise test.....	26
5.1.4 Comparisons of single-step (pair-wise comparisons) or step-down trend tests to determine the NOEC	28

5.1.4.1	General discussion	28
5.1.4.2	Single-step procedures.....	28
5.1.4.3	Step-down procedures.....	29
5.1.4.4	Deciding between the two approaches	30
5.1.5	Dose metric in trend tests	31
5.1.6	Role of power in toxicity experiments	31
5.1.7	Experimental design	32
5.1.8	Treatment of covariates and other adjustments to analysis	33
5.2	Quantal data (e.g. mortality, survival)	34
5.2.1	Hypothesis testing with quantal data to determine NOEC values	34
5.2.2	Parametric versus non-parametric tests	35
5.2.2.1	Basis	35
5.2.2.2	Single-step procedures.....	36
5.2.2.3	Step-down procedures.....	36
5.2.2.3.1	Choice of step-down procedure.....	36
5.2.2.3.2	Test for monotone dose response	36
5.2.2.3.3	Analysing the monotonic response for quantal data — Step-down procedure	37
5.2.2.3.4	Possible modifications of the step-down procedure.....	37
5.2.2.4	Alternative procedures	37
5.2.2.4.1	Parametric and non-parametric procedures.....	37
5.2.2.4.2	Pair-wise ANOVA-based methods	38
5.2.2.4.3	Jonckheere-Terpstra trend test.....	38
5.2.2.4.4	Poisson tests	38
5.2.2.5	Assumptions of methods for determining NOEC values	38
5.2.3	Additional information	39
5.2.4	Statistical items to be included in the study report.....	40
5.3	Hypothesis testing with continuous data (e.g. mass, length, growth rate) to determine NOEC	40
5.3.1	General	40
5.3.2	Parametric versus non-parametric tests	41
5.3.3	Single-step (pair-wise) procedures	42
5.3.3.1	General	42
5.3.3.2	Dunnett's test.....	42
5.3.3.3	Tamhane-Dunnett test.....	42
5.3.3.4	Dunn's test	42
5.3.3.5	Mann-Whitney test.....	43
5.3.4	Step-down trend procedures	43
5.3.5	Determining the NOEC using a step-down procedure based on a trend test	43
5.3.5.1	General	43
5.3.5.2	Preliminaries	43
5.3.5.3	Step-down procedure.....	43
5.3.5.3.1	Preferred approach	43
5.3.5.3.2	Williams' test.....	44
5.3.5.3.3	Jonckheere-Terpstra test.....	44
5.3.6	Assumptions for methods for determining NOEC values	44
5.3.6.1	Small samples — Massive ties.....	44
5.3.6.2	Normality	45
5.3.6.3	Variance homogeneity	45
5.3.7	Operational considerations for statistical analyses.....	46
5.3.7.1	Treatment of experimental units.....	46
5.3.7.2	Identification and meaning of outliers	46
5.3.7.3	Multiple controls	46
5.3.7.4	General	47
5.4	Statistical items to be included in the study report.....	47
6	Dose-response modelling	48
6.1	Introduction	48
6.2	Modelling quantal dose-response data (for a single exposure duration)	49
6.2.1	General	49
6.2.2	Choice of model	50

6.2.2.1	General	50
6.2.2.2	Probit model	51
6.2.2.3	Logit model	53
6.2.2.4	Weibull model	54
6.2.2.5	Multi-stage models	55
6.2.2.6	Definitions of EC_{50} and EC_x	55
6.2.3	Model fitting and estimation of parameters	56
6.2.3.1	Software and assumptions	56
6.2.3.2	Response in controls	56
6.2.3.3	Analysis of data with various observed fractions at each dose group	57
6.2.3.4	Analysis of data with one observed fraction at each dose group	58
6.2.3.5	Extrapolation and EC_x	58
6.2.3.6	Confidence intervals	58
6.2.4	Assumptions	59
6.2.4.1	General	59
6.2.4.2	Statistical assumptions	59
6.2.4.3	Evaluation of assumptions	59
6.2.4.3.1	Evaluation of basic assumptions	59
6.2.4.3.2	Evaluation of the additional assumption	59
6.2.4.4	Consequences of violating the assumptions	60
6.2.4.4.1	Consequences of violating basic assumptions	60
6.2.4.4.2	Consequences of violating the additional assumption	60
6.3	Dose-response modelling of continuous data (for a single exposure duration)	60
6.3.1	Purpose	60
6.3.2	Terms and notation	60
6.3.3	Choice of model	61
6.3.3.1	First distinctions	61
6.3.3.2	Linear models	62
6.3.3.3	Threshold models	62
6.3.3.4	Additive versus multiplicative models	63
6.3.3.5	Models based on “quantal” models	63
6.3.3.6	Nested non-linear models	64
6.3.3.7	Hill model	67
6.3.3.8	Non-monotone models	67
6.3.4	Model fitting and estimation of parameters	68
6.3.4.1	Software and assumptions	68
6.3.4.2	Response in controls	68
6.3.4.3	Fitting the model assuming normal variation	68
6.3.4.4	Fitting the model assuming normal variation after log-transformation	68
6.3.4.5	Fitting the model assuming normal variation after other transformations	69
6.3.4.6	No individual data available	69
6.3.4.7	Fitting the model using GLM	69
6.3.4.8	Covariates	70
6.3.4.9	Heterogeneity and weighted analysis	71
6.3.4.10	Confidence intervals	73
6.3.4.11	Extrapolation	73
6.3.4.12	Analysis of data with replicated dose group	73
6.3.5	Assumptions	74
6.3.5.1	General	74
6.3.5.2	Statistical assumptions	74
6.3.5.3	Additional assumption	74
6.3.6	Evaluation of assumptions	75
6.3.7	Consequences of violating the assumptions	75
6.3.7.1	Basic assumptions	75
6.3.7.2	Additional assumption	76
6.4	To accept or not accept the fitted model?	77
6.4.1	Can the fitted model be accepted and used for its intended purpose?	77
6.4.2	Is the model in agreement with the data?	77
6.4.3	Do the data provide sufficient information for fixing the model?	77

6.5	Design issues	81
6.5.1	General	81
6.5.2	Location of dose groups	81
6.5.3	Number of replicates	81
6.5.4	Balanced versus unbalanced designs	82
6.6	Exposure duration and time.....	82
6.6.1	General	82
6.6.2	Quantal data.....	82
6.6.3	Continuous data.....	83
6.6.3.1	General	83
6.6.3.2	Independent observations in time	83
6.6.3.3	Dependent observations in time	85
6.7	Search algorithms and non-linear regression	85
6.8	Reporting statistics.....	86
6.8.1	Quantal data.....	86
6.8.2	Continuous data.....	87
7	Biology-based methods	87
7.1	Introduction	87
7.1.1	Effects as functions of concentration and exposure time.....	87
7.1.2	Parameter estimation.....	89
7.1.3	Outlook.....	89
7.2	Modules of effect-models.....	90
7.2.1	General	90
7.2.2	Toxico-kinetic model	91
7.2.3	Physiological targets of toxicants.....	91
7.2.4	Change in target parameter	92
7.2.5	Change in endpoint.....	93
7.3	Survival	93
7.3.1	Relationship between hazard rate and survival probability	93
7.3.2	Assumptions of survival probability at any concentration of test compound	94
7.3.3	Summary	94
7.4	Body growth	97
7.4.1	Routes for affecting body growth.....	97
7.4.2	Assumptions.....	97
7.4.3	Von Bertalanffy growth curve	98
7.5	Reproduction	99
7.5.1	Routes that affect reproduction.....	99
7.5.2	Assumptions.....	100
7.5.3	Implication	100
7.6	Population growth.....	101
7.6.1	General	101
7.6.2	Assumptions.....	101
7.7	Parameters of effect models	103
7.7.1	General	103
7.7.2	Effect parameters.....	103
7.7.2.1	Toxicity and dynamic parameters	103
7.7.2.2	Killing rate, b_k	104
7.7.3	Discussion	105
7.7.4	Eco-physiological parameters	107
7.8	Recommendations	109
7.8.1	Goodness of fit.....	109
7.8.2	Choice of modes of action	110
7.8.3	Experimental design	110
7.8.4	Building a database for raw data.....	110
7.9	Software support.....	110
7.9.1	General	110
7.9.2	DEBtox	111
7.9.3	DEBtool	111

8	List of existing guidelines with references to the subclauses of this Technical Specification.....	112
Annex A	(informative) Analysis of an “acute immobilization of <i>Daphnia magna</i> ” data set (OECD GL 202 — ISO 6341) using the three presented approaches.....	115
A.1	Data set (see Table A.1)	115
A.2	Examples of data analysis using hypothesis testing (NOEC determination)	115
A.3	Example of data analysis by dose-response modelling	120
A.4	Example of data analysis using DEBtox (biological methods).....	125
Annex B	(informative) Analysis of an “algae growth inhibition” data set using the three presented approaches.....	127
B.1	General.....	127
B.2	Examples of data analysis using hypothesis testing (NOEC determination)	128
B.3	Example of data analysis by dose-response modelling	135
B.4	Examples of data analysis using DEBtox (biological methods).....	139
Annex C	(informative) Analysis of an “ <i>Daphnia magna</i> reproduction” data set (OECD GL 211 – ISO 10706) using the three presented approaches	142
C.1	Examples of data analysis using hypothesis testing (NOEC determination)	143
C.2	Example of data analysis by dose-response modelling	148
C.3	Examples of data analysis using DEBtox (biological methods).....	155
Annex D	(informative) Analysis of a “fish growth” data set (OECD GL 204/215 – ISO 10229) using the three presented approaches	160
D.1	Data set.....	160
D.2	Examples of data analysis using hypothesis testing (NOEC determination)	162
D.3	Example of data analysis by dose-response modelling	172
D.4	Examples of data analysis using DEBtox (biological methods).....	177
Annex E	(informative) Description and power of selected tests and methods.....	180
E.1	Description of selected methods for use with quantal data	180
E.2	Power of the Cochran-Armitage test	189
E.3	Description of selected tests for use with continuous data	198
E.4	Power of step-down Jonckheere-Terpstra test	218
Annex F	(informative) Annex to Clause 7 “Biology-based methods”	231
F.1	General.....	231
F.2	Effects on survival	231
Bibliography		237
Figure 1	— Conceptual illustration of accuracy and precision.....	2
Figure 2	— Illustration of a concentration-response relationship and of the estimates of the EC_x and NOEC/LOEC	5
Figure 3	— Analysis of quantal data: Methods for determining the NOEC	23
Figure 4	— Analysis of continuous data: Methods for determining the NOEC.....	24
Figure 5	— Analysis of continuous data: Methods for determining the NOEC (<i>continued</i>)	24
Figure 6	— Flow-chart for dose-response modelling.....	50

Figure 7 — Probit model fitted to observed mortality frequencies (triangles) as a function of log-dose	52
Figure 8 — Logit model fitted to mortality dose-response data (triangles)	53
Figure 9 — Weibull model fitted to mortality dose-response data (triangles)	54
Figure 10 — Logit model fitted to mortality dose-response data (triangles), with background mortality	57
Figure 11 — Two members from a nested family of models fitted to the same data set.....	66
Figure 12 — Cholinesterase inhibition as a function of dose at three exposure durations.....	71
Figure 13 — Relative liver masses against dose, plotted on log-scale	72
Figure 14 — Dose-response model fitted to the data of Figure 13, showing that the heterogeneous variance was caused by males (triangles) and females (circles) responding differently to the chemical ..	73
Figure 15 — Model fitted to dose-response data with and without an outlier in the top dose	76
Figure 16 — Two different models (both with four parameters) fitted to the same data set resulting in similar dose-response relationships.....	79
Figure 17 — Two data sets illustrating that passing a goodness-of-fit test is not sufficient for accepting the model.....	80
Figure 18 — Observed biomasses (marks) as a function of time, for nine different concentrations of Atrazine	84
Figure 19 — Growth rates as derived from biomasses observed in time (see Figure 18) at nine different concentrations (including zero), with the Hill model fitted to them.....	84
Figure 20 — Estimated growth rates as a function of (log-)concentration Atrazine	85
Figure 21 — Fluxes of material and energy through an animal, as specified in the DEB model.....	92
Figure 22 — Time and concentration profiles of the hazard model, together with the data of Figure 27	95
Figure 23 — Time and concentration profiles for effects on growth of <i>Pimephalus promelas</i> via an increase of specific maintenance costs by sodium pentachlorophenate (data by Ria Hoofman, TNO-Delft).....	98
Figure 24 — Time and concentration profiles for effects on growth of <i>Lumbricus rubellus</i> via a decrease of assimilation by copper chloride (data from Klok and de Roos 1996)	99
Figure 25 — Effects of cadmium on the reproduction of <i>Daphnia magna</i> through an increase of the costs per offspring — Data from the OECD ring-test	101
Figure 26 — Example of application of the DEBtox method.....	102
Figure 27 — The effect of a mixture of C,N,S-compounds on the growth of <i>Skeletonema costatum</i> via an increase of the costs for growth — Data from the OECD ring test	103
Figure 28 — A typical table of data that serves as input for the survival model, as can be used in the software package DEBtox (Kooijman and Bedaux 1996).....	108
Figure 29 — This profile likelihood function of the NEC (right panel) for the data in Figure 28 results from the software package DEBtox (Kooijman and Bedaux 1996)	108
Figure A.1 — Probit model fitted to mortality response at day 2 — $CED = EC_{10}$	122
Figure A.2 — The Weibull (left panel) and the two-stage LMS model fitted to the mortality data at day 2	123
Figure A.3 — Probit model fitted to mortality data on day 1 (left panel), and fitted to both day 1 and day 2 simultaneously	124
Figure A.4 — DEBtox example: Parameters and asymptotic standard deviations (ASD).....	125
Figure A.5 — Graphical test of model predictions against data.....	125
Figure B.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)	136
Figure B.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)	136

Figure B.3 — Estimated growth rates (from individual flasks, see Figure B.2) as a function of the concentration of Atrazine	137
Figure B.4 — Regression residuals from analysis on log-scale (upper panels) and from analysis without transformation (lower panels)	138
Figure B.5 — Growth rate plots	139
Figure B.6 — DEBtox example: Data for effects of Atrazine in micrograms per litre on the growth of <i>Selenastrum capricornutum</i> in cells per millilitre	139
Figure B.7 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD).....	139
Figure B.8 — DEBtox example: Time profile (Population growth, growth model, CAS 1912-24-9).....	140
Figure B.9 — DEBtox example: Concentration profile (Population growth, growth model, CAS 1912-24-9) .	140
Figure B.10 — DEBtox example: Profile likelihood for NEC estimate (Population growth, growth model, CAS 1912-24-9).....	141
Figure C.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).....	149
Figure C.2 — Total live young (TLY) in third week plotted against TLY in first two weeks, showing the correlations between these counts.....	149
Figure C.3 — Exponential model fitted to the number of life young counted over week three	151
Figure C.4 — Plots of regression residuals	152
Figure C.5 — Means of number of young, plotted cumulatively against time	153
Figure C.6 — Estimated ET_{50} values from Figure C.5, plotted against the concentration, with a fitted dose-response model.....	154
Figure C.7 — Number of young, plotted cumulatively against time	154
Figure C.8 — ET_{50} s estimated per replicate (see Figure C.7) as a function of the concentration with a fitted dose-response model.....	155
Figure C.9 — ET_{50} s estimated per replicate (see Figure C.7) as a function of the concentration, with two outliers removed	155
Figure C.10 — DEBtox example: Data for the cumulative number of offspring per female as affected by an unknown compound	156
Figure C.11 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD).....	156
Figure C.12 — DEBtox example: Time profile (reproduction, maintenance model, ISO repro set)	157
Figure C.13 — DEBtox example: Concentration profile (reproduction, maintenance model, ISO repro set) .	157
Figure C.14 — DEBtox example: Profile likelihood for NEC estimate (reproduction, maintenance model, ISO repro set).....	158
Figure C.15 — DEBtox example: Body length at 21 days.....	158
Figure D.1 — Exponential model, $y = a \exp(bx)$, fitted to masses at 28 days	174
Figure D.2 — Plots of regression residuals for the analysis of Figure D.1	175
Figure D.3 — Plots of regression residuals	175
Figure D.4 — Dose-response analysis of the fish masses, but without log-transformation	176
Figure D.5 — Exponential model fitted to the body lengths	177
Figure D.6 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD).....	177
Figure D.7 — DEBtox example: Concentration profile	178
Figure D.8 — DEBtox example: Profile likelihood for NEC estimate.....	179
Figure E.1 — Cochran-Armitage test: Plot showing that 5 subjects per concentration would give very low power.....	190

Figure E.2 — Cochran-Armitage test: Design with 20 subjects per concentration.....	191
Figure E.3 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study, with trend shape linear, lag = 0, sample size = 20.....	192
Figure E.4 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study, with trend shape linear, lag = 0, sample size = 20.....	192
Figure E.5 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study, with trend shape linear, lag = 0, sample size = 20.....	193
Figure E.6 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 40.....	193
Figure E.7 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 40.....	194
Figure E.8 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 40.....	194
Figure E.9 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 60.....	195
Figure E.10 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 60.....	195
Figure E.11 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 60.....	196
Figure E.12 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 80.....	196
Figure E.13 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 80.....	197
Figure E.14 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 80.....	197
Figure E.15 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 10.....	219
Figure E.16 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step 1, N = 10....	220
Figure E.17 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 10.....	220
Figure E.18 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step 1, N = 5.....	221
Figure E.19 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 5.....	221
Figure E.20 — Power of step-down Jonckheere test: (Doses,Step) = (6,4) (5,3)(4,2) (3,1), N = 5.....	222
Figure E.21 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 5.....	222
Figure E.22 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 10.....	223
Figure E.23 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step 1, N = 10....	223
Figure E.24 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 10.....	224
Figure E.25 — Power of step-down Jonckheere test: (Doses, Step) = (6,4) (5,3) (4,2) (3,1), N = 10.....	224
Figure E.26 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 20.....	225
Figure E.27 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step 1, N = 20... 225	225
Figure E.28 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 20.....	226
Figure E.29 — Power of step-down Jonckheere test: (Doses, Step) = (6,4) (5,3) (4,2) (3,1), N = 20.....	226
Figure E.30 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 40.....	227

Figure E.31 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step 1, N = 40 ..	227
Figure E.32 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 40	228
Figure E.33 — Power of step-down Jonckheere test: (Doses, Step) = (6,4) (5,3) (4,2) (3,1), N = 40	228
Figure E.34 — Power of step-down Jonckheere test: 6 doses at Step 1, N = 80	229
Figure E.35 — Power of step-down Jonckheere test: 6 doses at Step 2 versus 5 doses at Step, N = 80	229
Figure E.36 — Power of step-down Jonckheere test: 6 doses at Step 3 versus 5 doses at Step 2 versus 4 doses at Step 1, N = 80	230
Figure E.37 — Power of step-down Jonckheere test: (Doses, Step) = (6,4) (5,3) (4,2) (3,1), N = 80	230

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of normative document:

- an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;
- an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TS 20281 was prepared by Technical Committee ISO/TC 147, *Water quality*, Subcommittee SC 5, *Biological methods*.

Introduction

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

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

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

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

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

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

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

There was an ISO resolution (ISO TC 147/SC 5/WG 10 Antalya 3), as well as an OECD workshop recommendation (OECD 1998), that the NOEC should be phased out from International Standards.

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

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

- acute toxicity on *Daphnia magna*;
- inhibition of algae growth;
- reproduction of *Daphnia magna*; and
- fish growth.

Water quality — Guidance on statistical interpretation of ecotoxicity data

1 Scope

This Technical Specification offers guidance on statistical methods used for the analysis of data of standardized ecotoxicity tests. It focuses on statistical methods for obtaining statistical estimates of parameters in current and future use, e.g. EC_x (LC_x), NOEC, NEC.

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

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

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

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

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1: —¹⁾, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1

accuracy

measure of how close the estimate is to the “true value” of the parameter (this true value is unknown)

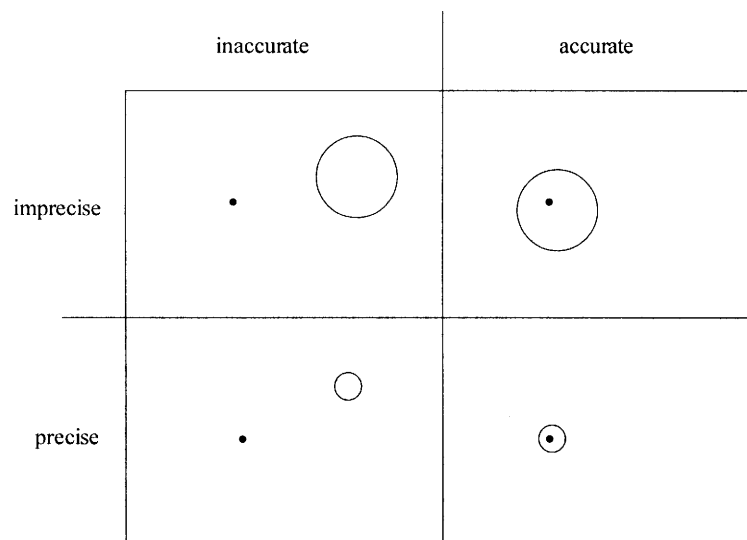
3.2

precision

measure of the amount of variability in the estimate (quantified by the standard error or the confidence interval of the estimate)

1) To be published. (Revision of ISO 3534-1:1993)

NOTE Precision may be increased by using larger sample sizes or by reducing the experimental variation. However, as Figure 1 illustrates, although an estimate may be precise, this does not imply that it is also accurate.



NOTE The dot represents the true parameter value. The circle represents the confidence interval of the estimate. Small circles indicate high precision. Large circles indicate low precision.

Figure 1 — Conceptual illustration of accuracy and precision

3.3 concentration

amount of test material in the testing environment

NOTE 1 It is expressed, for example, as milligrams per litre in water, and as milligrams per kilogram in soil and in food.

NOTE 2 Concentration and dose both refer to the amount of test material to which the test organism is subjected. Statistical methods for both types of studies are identical; however, interpretations are different. Although “concentration” is used throughout this Technical Specification, all the statistical methods presented here also apply to studies in which a dose is used.

3.4 dose

amount of test material administered to a subject

NOTE 1 It is expressed, e.g. in milligrams/kilogram-body mass in an avian bolus study.

NOTE 2 Concentration and dose both refer to the amount of test material to which the test organism is subjected. Statistical methods for both types of studies are identical; however, interpretations are different. Although “concentration” is used throughout this Technical Specification, all the statistical methods presented here also apply to studies in which a dose is used.

3.5 confidence interval

x % confidence interval for a parameter is an interval of values that theoretically covers the true value of the estimated parameter with x % confidence

NOTE 1 Standard confidence intervals are based on the assumption that the underlying mathematical model is correct. It does not capture model uncertainty.

NOTE 2 A more precise definition is the following: interval estimator (T_0, T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that $P[T_0 < \theta < T_1] \geq 1 - \alpha$. [ISO 3534-1,—].

NOTE 3 Associated with this confidence interval is the confidence level $100(1 - \alpha) \%$ where α is generally a small number. The confidence level is typically 90 % or 95 %. The inequality $P[T_0 < \theta < T_1] \geq 1 - \alpha$ holds for any specific but unknown population value of θ .

NOTE 4 A confidence interval does not reflect the probability that the observed interval contains the true value of the parameter (it either does or does not contain it). The confidence reflects the proportion of cases that the confidence interval would contain the true parameter value in a long series of repeated random samples under identical conditions.

3.6 Data types

3.6.1

quantal/binary data

data that arise when a particular property is recorded to be present or absent in each individual

NOTE 1 An individual shows an effect or it does not show an effect. Therefore, these data can exhibit only two states.

NOTE 2 Typically, quantal data are presented as the number of individuals showing the property (e.g. mortality) out of a total number of individuals observed in each experimental unit. Although this can be expressed as a fraction, it should be noted that the total number of individuals cannot be omitted.

3.6.2

continuous data

data that can (theoretically) take any value in an open interval, for instance any positive number

EXAMPLES Measurements of length or body mass.

NOTE 1 Due to practical reasons, the measured resolution depends on the quality of the measurement device. For example, if test units are observed once per day, then "time to hatch" can only be recorded in whole days; however, the underlying distribution of "time to hatch" is continuous.

NOTE 2 Typically, continuous data have a dimension (e.g. grams, moles/litre).

3.6.3

discrete data

data that have a finite or countable number of values

NOTE There are three classes of discrete data: nominal, ordinal and interval.

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

3.7

effect

change in the response variable under consideration compared to a control

NOTE 1 For quantal endpoints, an effect is usually described in terms of a change in the percentage of individuals affected.

NOTE 2 For continuous endpoints, an effect is typically described in terms of a percent change in the mean values of the endpoint, but it can also be described in terms of absolute change.

3.8 Effect concentrations

3.8.1

quantal effective concentration

quantal EC_x

concentration of test material in water, soil, or sediment that causes x % change in response (e.g. immobility) during a specified time interval.

EXAMPLE An example of a concentration-response relationship and its associated estimates of EC₁₀ and EC₅₀, are illustrated in Figure 2.

NOTE 1 It is expressed, e.g. in milligrams per litre or milligrams per kilogram.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.2

quantal effective dose

quantal ED_x

dose of test material that causes x % change in response (e.g. immobility) during a specified time interval

NOTE 1 It is expressed, e.g. in milligrams per kilogram body mass in an avian bolus study.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.3

quantal lethal concentration

quantal LC_x

concentration of test material in water, soil, or sediment that causes x % change in response (i.e. mortality) during a specified time interval.

EXAMPLE An example of a concentration-response relationship and its associated estimates of EC₁₀ and EC₅₀, are illustrated in Figure 2.

NOTE 1 It is expressed, e.g. in milligrams per litre or milligrams per kilogram.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.4

quantal lethal dose

quantal LD_x

dose of test material that causes x % change in response (i.e. mortality) during a specified time interval

NOTE 1 It is expressed, for example, in milligrams per kilogram body mass in an avian bolus study.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.5

continuous effective concentration

continuous EC_x

concentration of test material in water, soil, or sediment that causes x % in the size of the endpoint during a specified time interval

NOTE 1 It is expressed, for example, in milligrams per litre or milligrams per kilogram.

NOTE 2 This parameter is also estimated by dose-response modelling.

3.8.6**continuous effective dose****continuous ED_x**

dose of test material (e.g. mg/kg-body mass in an avian bolus study) that causes x % in the size of the endpoint during a specified time interval

NOTE 1 It is expressed, for example, in milligrams per kilogram body mass in an avian bolus study.

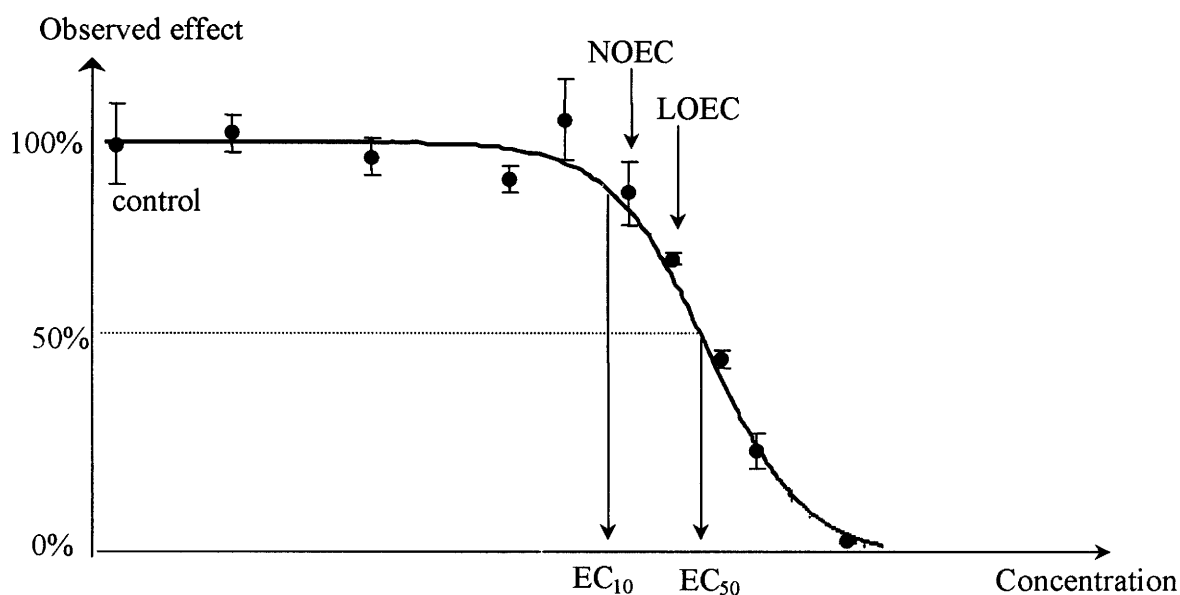
NOTE 2 This parameter is also estimated by dose-response modelling.

3.9**endpoint****response variable**

biological parameter observed

EXAMPLE Survival, number of eggs, size or growth, enzyme level.

NOTE An ecotoxicological study can have one or many endpoints.



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

Figure 2 — Illustration of a concentration-response relationship and of the estimates of the EC_x and NOEC/LOEC

3.10**effective time** **ET_x**

time at which an effect of x % is expected at a specified test concentration when the test organisms are exposed to a given concentration of material in water or sediment or soil

NOTE ET_x is estimated by modelling a time-response relationship.

3.11**lethal time** **LT_x**

time at which an effect of x % is expected at a specified test concentration when the test organisms are exposed to a given concentration of material in water, sediment or soil, and when the response of interest is mortality

3.12

experimental unit

replicate unit

smallest unit of experimental material to which a treatment can be allocated independently of all other units

EXAMPLES Aquariums, beakers, or plant pots.

NOTE By definition, experimental units shall be able to receive different treatments. Each experimental unit may contain multiple sampling units on which measurements are taken. Within each experimental unit, sampling units may not be independent. However, in some special case situations, individual organisms (housed in common units) can be treated as the experimental units: these special cases require some proof or strong argument of independence of organisms.

3.13

sampling unit

unit for which the measurement is taken

EXAMPLES Fish, daphnia or plants.

NOTE The sampling unit is not always identical with the statistical unit.

3.14

extrapolation

prediction of the value of variates outside the range of observations

NOTE 1 Extrapolation may not lead to a reliable estimate (see e.g. 6.4).

NOTE 2 When an EC_x estimated from a fitted concentration-response function is lower than the lowest nonzero concentration tested in the study, or higher than the highest concentration tested in the study, it is obtained by extrapolation.

3.15

interpolation

prediction of the value of variates within the range of observations

NOTE When the EC_x is between two consecutive nonzero test concentrations, it is said to be obtained by interpolation.

3.16

hormesis

effect where the tested substance is a stimulant in small concentrations, but it is inhibitory in large concentrations, resulting in a biphasic (or U-shaped) concentration-response relationship

NOTE 1 This observed stimulatory effect may be due to the tested substance, but it could also be due to an experimental artefact (e.g. solvent effect, non-random allocation of treatments to experimental units, experimental error).

NOTE 2 Models incorporating hormesis are not detailed in this Technical Specification. Two issues of *Critical Reviews in Toxicology* [2001, **31**(4) and **31**(5), pp. 351-694] and other journal articles discuss the issues concerning hormesis. Some discussion can be found in *Environment Canada* (2003).

3.17

lowest observed effect concentration

LOEC

lowest concentration out of the tested concentrations at which a statistically-significant difference from the control group is observed

EXAMPLE An example of LOEC is illustrated in Figure 2.

NOTE The LOEC is obtained by hypothesis testing.

3.18**no observed effect concentration****NOEC**

tested concentration just below the LOEC

EXAMPLE An example of NOEC is illustrated in Figure 2.

NOTE The NOEC is obtained by hypothesis testing.

3.19**monotonic concentration-response**

relationship in which the true, underlying concentration-response relationship exhibits an increase or a decrease over the range of concentrations in the study

NOTE If the concentration-response is monotone and non-increasing, the location parameters (mean or median) would exhibit the following relationship: $\gamma_0 \geq \gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \dots \geq \gamma_k$, where γ is the location parameter and 0, 1, 2, ..., k are the concentration groups. If the monotone relationship is non-decreasing, the inequalities are reversed.

3.20**non-monotonic concentration-response**

relationship in which the inequalities are not consistent across the concentrations

3.21**no effect concentration****NEC**

parameter, which, when used in concentration-response models, is interpreted as being the highest concentration in which the compound does not affect the endpoint, even after very long exposure to the compound

NOTE The NEC equals the EC_0 at infinite time.

3.22**response**

response corresponds to an observed value of any endpoint

3.23**parametric method**

method which assumes that all the properties of the model are specified, except for the values of the parameters

EXAMPLE 1 In classical analysis of variance (ANOVA), the residuals are assumed to follow a normal distribution with a mean of zero and some unknown variance.

EXAMPLE 2 In Poisson regression, the response variable is assumed to follow a Poisson distribution (parameters to be estimated in the fitting process).

3.24**non-parametric method**

method which contains weaker assumptions²⁾ than the **parametric method** (3.23) about the shape of the distribution of the residuals, and the analysis is often based on ranks of the observations

NOTE The non-parametric analogue to a two-sample t -test is the Mann-Whitney test. Non-parametric regression is often conducted using a variety of smoothing techniques.

2) With fewer constraints.

3.25

parsimony principle

principle that data should be described with as few parameters as possible

NOTE A common decision criterion of including more parameters in the model is the observation that such leads to a significantly better description of these data.

3.26

systematic error

situation that a single concentration (dose) group differs from the others not only with respect to the intended treatment (i.e. the concentration or dose) but also with respect to some unintended experimental factors

EXAMPLE Containers housing the animals may differ by themselves, and in a design with few or only one container per dose group a deviating container may lead to a systematic error in that group.

NOTE The factor of time may underlie systematic errors in various ways, e.g. time of feeding, time of observation. The problem of systematic errors is that they may be wrongly interpreted as an effect of the intended treatment.

3.27

statistical significance

observed deviation from what was expected under the null hypothesis, it is unlikely to be attributable to chance variation

EXAMPLE In hypothesis testing, a result is statistically significant at the chosen level α if the test statistic falls in the rejection region.

NOTE In this Technical Specification, the α -level is 0,05 unless otherwise stated.

4 General statistical principles

4.1 Different statistical approaches

4.1.1 General

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

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

- a brief description of the use of each method in ecotoxicity tests;
- a brief outline of specific analysis methods presented in the later clauses of this Technical Specification;
- a listing of some major assumptions and limitations for each approach.

4.1.2 Hypothesis-testing methods

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

Methods discussed in Clause 5 include analyses for quantal data and continuous data. For both types of data, parametric approaches (when an underlying distribution e.g.: normal, log-normal is characterised) and non-parametric approaches (when weaker assumptions are made regarding the distribution) are presented. In Clause 5, assessment is limited to conducting data analysis separately at each time point, though this is not a limitation of the method. Three terms often used when discussing hypothesis tests are: Type I errors, Type II errors, and power (Table 1).

- *Type I errors* (false positives) occur when the null hypothesis is the truth but the hypothesis test results in a rejection of the null hypothesis in favour of the alternative hypothesis. The probability of making a Type I error is often referred to as α and is usually specified by the data analyst – often at 0,05, or 5 %.
- *Type II errors* (false negatives) occur when the alternative hypothesis is true but the test fails to reject the null hypothesis (i.e. there is insufficient evidence to support the alternative hypothesis). The probability of making a Type II error is often referred to as β ($1 - \text{power}$).
- *Power* is the probability of rejecting the null hypothesis (H_0) in favour of the alternative hypothesis (H_A), given that the alternative hypothesis is true. Power of a test varies with sample size, variance of the measured response, the size of an effect that it is of interest to detect, and the choice of statistical test. Power to detect differences can be increased by increasing the sample size and/or reducing variation in the measured responses. Thus, if a test has low power to detect an effect of a given size, this is equivalent to saying that the test has a low probability of detecting an effect of that size.

Table 1 — Probabilities of finding a significant or non-significant test outcome, given that the null hypothesis is true or not

Result of hypothesis test	State of the world	
	H_0 true	H_A true
not significant	$1 - \alpha$	Type II error β
significant	Type I error α	$1 - \beta = \text{power}$

Several assumptions made when conducting hypothesis tests to determine the NOEC are given below.

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

Several limitations of using hypothesis testing to determine the NOEC are given below.

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

4.1.3 Concentration-response modelling methods

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

Methods discussed in Clause 6 include analyses for quantal data and continuous data. Parametric approaches (when a specific underlying distribution is assumed) are presented.

Although non-parametric methods have been developed for fitting concentration-response curves and estimating an EC_x , they are not presented in this Technical Specification.

— Sources on non-parametric regression include:

Green and Silverman (1994), Easton and Peto, Fan and Gibjels (1996), Hardle, W. (1991), Azzalini and Bowman (1997), Silverman B. (1985), Akritas and Van Keilegom (2001), Carroll *et al.* (1999) and Smith-Warner *et al.* (1998).

— Software for non-parametric regression can be found.

The effect of exposure time is considered in Clause 6.

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

These specific properties of the experimental design, the number and spacing of doses and the number of replicates, are related to the value of x of interest in the particular experiment. Different designs may be employed to estimate an EC_{50} , as opposed to an EC_{05} . Further guidance for the design of experiments of this type is discussed in Clause 6.

Several assumptions of concentration-response modelling are:

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

Several limitations of concentration-response modelling are:

- estimation of EC_x values outside the concentration range introduces a great deal of uncertainty (i.e. extrapolation outside the range of the data);

- once the experiment has been performed, the resulting concentration-response data may not be suitable for the estimation of parameters of a concentration-response model. In particular, when the gaps between consecutive response levels are so large that many different concentration-response models would fit equally well to the observed data, interpolation would not be warranted.

4.1.4 Biology-based methods

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

Methods discussed in Clause 7 include analyses for quantal data and continuous data for several aquatic toxicity tests (acute and chronic tests on survival/immobility for daphnia and fish, fish growth test, daphnia reproduction test, and alga growth inhibition test). The models presented in this Technical Specification utilize dynamic energy budget theory (see Clause 7 for details and associated references). This theory provides a quantitative description for the processes of feeding, digestion, storage, maintenance, growth, development, reproduction and ageing and their interrelationships. As with concentration-response modelling, the level of confidence in the parameter estimates (as evidenced by the width of the confidence interval) is a function of sample size and inherent variation in the response.

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

Several general assumptions made when using biology-based methods are the following.

- The models discussed in this Technical Specification assume the response has a monotonic concentration-response relationship.
- This analysis method incorporates mechanistic models for toxico-kinetics and physiology.

Several limitations of biology-based methods are the following.

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

4.2 Experimental design issues

4.2.1 General

The usual factors (independent variables) studied in ecotoxicity tests are the concentration of the tested substance and the duration of exposure. For the estimation of the effect at a given condition, it is necessary to replicate these conditions, to control experimental variation (see 4.2.3).

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

4.2.2 NOEC or EC_x : Implications for design

The estimation of an EC_x puts different demands on the study design than does the assessment of an NOEC.

- When the aim is to assess an NOEC, an important demand is that the study warrants sufficient statistical power. To that end, the concentration (dose) groups need a sufficient number of replicates (possibly at the expense of the number of dose groups). Many test guidelines are based on this principle.
- When the aim is, however, to provide an estimate of an EC_x , the primary demand on the study design is to have a sufficient number of concentration (dose) groups. This may be at the expense of the number of replicates per group (e.g. keeping the total size of the experiment the same), since the precision of the estimated EC_x depends more on the number and spacing of concentrations than on the sample size per concentration or dose group.
- The demands for study designs aimed at estimating an NOEC or an EC_x are further discussed in 5.1.7 and 6.5, respectively.

Therefore, the choice between assessing an NOEC and estimating an EC_x should actually be made before designing the study. If one wishes (or is required) to assess both, a compromise between the two opposing demands shall be made, i.e. a design with both a sufficient number of dose groups and a sufficient number of replicates in each group.

- The number of replications per group needed for assessing an NOEC depends on the desired power of the statistical test involved (see 4.1.1).
- For assessing an EC_x , three concentration groups, next to the control group, is an absolute (theoretical) minimum. However, when just one dose group appeared to have been unluckily chosen, the assessment of an EC_x would probably fail, and more concentration groups are therefore required in practice.

Design recommendations for experiments using a biology-based model include those for EC_x . Additional recommendations are discussed in 7.8.3.

4.2.3 Randomization

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

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

Randomization processes are used in designed experiments to eliminate bias in estimates of treatment effects, and to ensure independence of error terms in statistical models. Ideally, randomization should be used at every stage of the experimental process, from selection of experimental material and application of treatments, to measurement of responses.

To minimize the effects of the first two factors, animals need to be randomly distributed into concentration groups. To minimize the effects of the third factor (both intended and unintended, such as location in the room), application of treatments should be randomized as much as possible. To minimize the effects of the fourth factor, the measurement of responses should be randomized in time (e.g. although all responses are recorded at 24 h, the order in which the experimental units are measured should be randomized). With good scientific methods, measurement errors can be minimized.

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

4.2.4 Replication

As discussed above, noise cannot be avoided, and therefore it is necessary to assign a certain number of replicates (experimental units) to each treatment group and control group. The number of replicates influences the power in hypothesis testing and the confidence limits of parameter estimates. A standard assumption of all methods is that replicates are independent. Treating observations as independent replicates, whereas in fact they are not, represents an error called pseudo-replication (Hurlbert, 1984). This issue becomes important when animals are housed together, as in a tank or beaker.

There are two types of housing effects:

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

Both effects result in non-independence (or pseudoreplication) of the individual organism's responses.

The first effect may result in different mean responses between containers (at a given concentration). This type of non-independence can be addressed by taking the variation between containers into account in the statistical model. For instance, with continuous data this may be done using a nested ANOVA, where the individual observations are nested within the container.

The second effect might distort the distribution of the observations related to the individual organisms. For instance, with quantal data, the assumption of binomial distribution may not hold. In an example with continuous data, when there is competition among individuals in the same container, the responses of the individual organisms may appear bimodal. See Clause 5 and Clause 6 for more detailed discussions.

4.2.5 Multiple controls included in the experimental design

It is common in aquatic and certain other types of experiments that the chemical under investigation cannot be administered successfully without the addition of a solvent or vehicle. In such experiments, it is customary to include two control groups. One of these control groups receives only what is in the natural laboratory environment (e.g. dilution water), while the other group receives the dilution water with added solvent but no test chemical. In ecotoxicity experiments, these are often termed negative or dilution water (non-solvent) and solvent controls.

OECD recommends limiting the use of solvents (OECD 2000); however, appropriate use of solvents should be evaluated on a case-by-case basis. Details regarding the use of solvents (e.g. recommended chemicals, maximum concentrations) are discussed in the relevant guideline documents for a specific ecotoxicity test. In addition, regulatory guidelines shall be followed by both controls with regard to the range of acceptable values (e.g. minimum acceptable percent survival or mean oyster shell deposition rate). Multiple control groups can be utilized regardless of whether the experiment was intended for hypothesis testing (i.e. determination of an NOEC), regression analysis (i.e. determination of an EC_{50}), or biology-based methods. The focus of this subclause is to present a data analysis methodology for experiments in which a solvent is used.

Data from the two control groups are analysed to determine if the solvent had a statistically significant effect on the measured response. If there is a statistically significant difference between the negative and solvent

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

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

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

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

4.3 Process of data analysis

4.3.1 General

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

4.3.2 Data inspection and outliers

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

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

Obviously, treatment group outliers are highly undesirable, since they directly interfere with the effect that one wishes to measure, thereby increasing the probability of both false positive and false negative results. For example, an NOEC may be assessed at a level where substantial effects do occur, or an LOEC may be assessed at a level without real effects (i.e. from the chemical). The only way to prevent outliers at the group level is a design that is perfectly randomized with respect to all experimental actions that may potentially influence the (observed response of) the biological system. In practical biological studies, however, perfect randomization is hard to realise, and it is not feasible to reduce the probability of getting group outliers to nil.

Therefore, it is paramount to make the study design relatively insensitive to potential outliers, i.e. by randomized replicated dose groups, and/or by increasing the number of different doses (followed by dose-response modelling, see Clause 6).

Outliers at the individual level can only be detected in continuous data. When a particular distribution is assumed for the scatter in the data, the judgement of outliers may be based on a specific, small probability that any single data point could occur. This implies that the judgement of outliers can depend quite strongly on the assumed distribution. For example, values that appear to be extremely high when assuming a normal distribution may be judged as non-extreme when assuming a log-normal distribution. Vice versa, low values may be judged as extremes when assuming a log-normal distribution, but not so when a normal distribution is assumed.

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

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

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

4.3.3 Data inspection and assumptions

4.3.3.1 Scatter

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

4.3.3.2 Heterogeneous variances and distribution

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

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

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

Finally, in quantal data with replicated dose groups, it can be also be expected *a priori* that the scatter between the replicates depends on the mean response (this follows directly from the properties of

frequencies). Here, one may plot the frequencies after the transformation $\arcsin(\sqrt{p})$, where p is the observed frequency (fraction). This transformation is able to remove the (theoretical) relationship between the variance and the mean frequency (assuming a binomial distribution).

4.3.3.3 Heterogeneous variances and true variation in response

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

4.3.3.4 Consequences for the analysis

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

However, when the heterogeneity of variances appears to be due to real biological heterogeneity in responses among individual organisms, one should carefully consider if further analysis is meaningful. For example, when the organisms (or experimental units) consist of two distinct subpopulations, one responding, the other not, any estimated change in mean response has no useful meaning. When such two subpopulations can be distinguished from observable features (e.g. sex), the appropriate way to proceed is to analyse both subpopulations separately, or by using the observable feature as a covariate (see, e.g. 6.3.3, and Figure 14).

4.3.4 Transformation of data

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

For a variance-stabilizing transformation to exist, there shall be a relationship between the population means and variances. In many cases, the theoretical distribution of the response variable can guide the choice of a transformation. Examples are given below.

- If the underlying distribution is assumed to be Poisson, the square root transformation, $y'_i = (y_i)^{1/2}$ or $y'_i = [(y_i + 1)^{1/2}]$, is used.
- If the underlying data are log-normal, the log-transformation, $y'_i = \log(y_i)$, is often used.
- For proportions with binomial distributions, the arc-sin square-root [$y'_i = \arcsin(y_i^{1/2})$] and Freeman-Tukey [$y'_i = (y_i + 1)^{1/2} + (y_i)^{1/2}$] transformations are often used.

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

- If the underlying theoretical distribution is unknown, a data-based procedure (Box-Cox transformation) can be used (Box and Cox, 1964).

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

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

When a transformation is not used in the data analysis, the difference in the means is a logical measure for the size of an effect. This difference is interpreted as an absolute change in the original units (e.g. a decrease of 1,2 g). The back-transformed difference in means (of the transformed data) however has another, usually more difficult, interpretation. In the special case of a log-transformation, the difference between the back-transformed means does allow a simple interpretation: it estimates the ratio (or percent change) of the median responses.

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

4.3.5 Parametric and non-parametric methods

4.3.5.1 General

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

4.3.5.2 Parametric methods

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

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

Whatever distribution is assumed, parametric methods are based on the general principle of fitting the data to the model. In hypothesis testing, this may be the ANOVA model; in dose-response modelling, this may be a particular dose-response model.

- In applying parametric hypothesis tests, one shall examine the data for outliers, deviations from normality and homogeneity, assessment of monotonicity of the dose response (for some approaches), and do a general assessment of whether the proposed model adequately describes the data. These points are discussed in depth in 5.1, 5.1.4, 5.1.6, 5.2.2.3, 5.2.2.5, 5.3.1, 5.3.1.3, 5.3.1.4, 5.3.1.6, and 5.3.1.7.

- In dose-response modelling, the process of model fitting is eminent and indeed the focus of the analysis. Therefore, in any dose-response analysis (as discussed in Clauses 5, 6 and 7) the user should understand the general principles of model fitting. These are discussed in 4.3.6, 5.1.4, 5.2.2, 5.3.1, and 6.7.

4.3.5.3 Generalized linear models (GLMs)

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

4.3.5.4 Non-parametric methods

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

4.3.5.5 How to choose?

Parametric analyses have various advantages over non-parametric methods.

- They are typically simpler to conduct (wide availability of software).
- The methods have been developed for a wider array of designs (e.g. designs with replicated dose groups).
- The confidence intervals are more easily computed.
- The methods are more universally used.
- Interpretation of results is often easier.

Non-parametric methods have the following advantages.

- They are based on very weak assumptions.
- Further, since non-parametric analyses are based on the rank order of the data, they are less sensitive to outliers than parametric analyses.

When the data appear to comply with the assumptions (after a visual inspection) of a particular parametric analysis, parametric is the obvious method to choose. The assumptions can be further checked as part of the analysis (e.g. by examining the residuals, see below). It may be noted that parametric analysis based on normal assumptions is reasonably robust to mild violations against the assumptions. When a data transformation results in a (better) compliance with the normality assumptions, one should be reminded that transformations other than the log-transformation may impair the interpretation of the results. This is because the log-transformation is naturally linked to the intuitive notion that biological effects are proportional (or multiplicative) rather than additive (compare definition of EC_{50}). Thus, when omitting or applying a log-transformation does not make a large difference for complying with the assumptions, one might choose to apply it for reasons of interpretation.

In situations where specific distributions are natural candidates for the data type at hand, one may consider the use of GLMs.

When no regular distributions can be assumed, as e.g. in the case of tied observations⁴⁾, one may resort to non-parametric analysis.

4.3.6 Pre-treatment of data

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

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

A current habit in analysing continuous data is to divide the observed responses by the (mean) observed response in the controls. These corrected observations then reflect the percent change compared to the controls, which is usually the entity of interest. However, such a pre-treatment of the data is improper. Among other problems, it assumes that the (mean) response in the controls is known without error, which is not the case. Therefore, this should be avoided, and instead the background response should be estimated from the data by fitting the model to the untreated data. Thus, the estimation error in the controls is treated in the same way as the estimation errors in the other concentration groups. (see e.g. 6.2.3 and 6.3.3).

4.3.7 Model fitting

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

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

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

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

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

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

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

A general method of finding the best fit is by trial and error, i.e. in an iterative search one tries to improve the likelihood by changing the parameter values, until an improvement can no longer be found. General algorithms exist that can perform such an iterative search in an efficient way. In particular models ("linear" models), the maximum likelihood estimates can be derived from explicit formulae and search algorithms are not required (for that reason linear models used to be popular before the availability of computers). In non-linear models, search algorithms can hardly be avoided. Although the user need not worry about the calculations underlying these algorithms, fitting non-linear models does require some basic understanding of the general principles of search algorithms (see 6.7).

4.3.8 Model checking

4.3.8.1 Analysis of residuals

After a model has been fitted to the data, a final check for the appropriateness of the fitted model may be performed. Do the data indeed comply with the model assumptions?

For instance,

- Do the data comply with the assumed distribution (in parametric analyses)?
- Are the variances homogenous (e.g. in ANOVA)?
- Is the dose-response model suitable for the dose-response data at hand (in dose-response analysis)?

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

To check the distribution, the residuals can be taken together and be plotted in a single histogram, or in a (distribution-specific) QQ-plot⁵). Visual inspection of such plots may reveal deviations from the assumed distribution, in particular when inspecting a QQ-plot, which should be linear if the data comply with the assumed distribution. Formal tests exist as well (see Clause 5), but it should be noted that a mild violation of the assumptions is no reason for concern, and tests do not measure the degree of violation.

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

- against the predicted value (i.e. the group means, usually), to check if the variances are homogenous (if such were assumed);
- against the model prediction, to check for systematic deviations from the fitted model;
- against other experimental factors, if relevant, for instance the order in time by which the observations were made. Such a plot may show if the pertinent factor influenced the response systematically.

Finally, one may perform a formal goodness-of-fit test. This test is sensitive for all the assumptions simultaneously. A significant overall test of goodness-of-fit may indicate that one of the assumptions is not met, but this does not necessarily imply that the model is not useful for the particular purpose of the analysis. Here, one should judge the nature of the violated assumption and its potential impact on the results one is interested in. On the other hand, a nonsignificant goodness-of-fit test does not imply that the model used may be regarded as reliable. The test is more easily passed when the data contain relatively little information, and, as

5) The QQ-plot corresponds to a plot of observed quantiles versus expected quantiles.

a result, various models may pass the goodness-of-fit test, but lead to different conclusions. In other words, not only the model, but also the data should be “validated”, by asking the question of whether or not they contain the information needed for answering the question of interest. The evaluation of a dose-response model for describing a particular dataset is more fully discussed in 6.4.

4.3.8.2 Validation of fitted dose-response model

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

4.3.9 Reporting the results

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

The results of the analysis typically consist of summarizing statistics. In current practice, these are usually the means and standard deviations (or standard errors) per dose group. This may not generally be the best way of reporting results, however.

When a parametric analysis assuming homogenous variances is applied, it is more informative to report the estimate of the common variance (residual Mean Square), together with a justification of the homogeneity assumption (e.g. a plot of the individual data or of the residuals against dose).

When a log-transformation is applied before the analysis, it is more adequate to report the geometric means, and the (possible common) geometric standard deviation (GSD) or Coefficient of Variation (CV).

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

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

More specific guidance for reporting results is given at the end of Clauses 4, 5 and 6.

5 Hypothesis testing

5.1 Introduction

5.1.1 General

This clause provides an overview of both hypothesis testing and methodological issues specific to determining NOECs under various experimental scenarios. It is divided into three major parts.

The first part includes flow-charts summarizing possible schemes for analysing quantal (Figure 3) and continuous data (Figures 4 and 5), along with some basic concepts that are important to the understanding of hypothesis testing and its use in the determination of NOECs. Special attention is given to the choice of the hypothesis to be tested, as this choice may vary depending on whether or not a simple dose-response trend is expected, and on whether increases, or decreases (or both) in response are of concern.

The remainder of the clause is divided into two major subclauses that discuss statistical issues related to the determination of NOECs for quantal and continuous data (4.2 and 4.3 respectively) and provide further details on the methods listed in Figures 3 and 4.

This division reflects the fact that different statistical methods are required for each type of data, and that problems arise that are unique to the analysis of each type of data. An attempt has been made to mention the

most widely used statistical methods, but to focus on a set of methods that combine desirable statistical properties with reasonable simplicity. For a given set of circumstances, more than one statistical approach may be acceptable. In such cases, the methods are described, the limitations and advantages of each are given, and the choice is left to the reader.

The flow-charts in Figures 3 and 4 indicate a possible choice of methods. Examples of the application of many of these methods, mathematical details and properties of the methods are presented in E.1.

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

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

NOECs, as defined and discussed in this Technical Specification, are based on a concept sometimes called “proof of hazard”. In essence, the test substance is presumed to be non-toxic unless the data present sufficient evidence to conclude toxicity.

Alternative approaches to assessing toxicity through hypothesis testing exist. For example, Tamhane *et al.* (2001) and Hothorn and Hauschke (2000) developed an approach based on proof of non-hazard. Specifically, if an acceptable threshold of effect is specified, such as a 20 % decrease in mean, then the maximum safe dose (MAXSD in Tamhane *et al.*, 2001) is the highest concentration for which there is significant evidence that the mean effect is less than 20 %. These are relatively new approaches that have not been thoroughly tested in a practical setting, and for few endpoints is there agreement on what level of effect is biologically important to detect. All current guidelines regarding NOEC are based on the proof of hazard concept. For these reasons, this alternative approach is not presented in this clause, though it does hold some promise for the future. The only common exception to this is in regard to limit tests, where in addition to determining whether there is a statistically-significant effect in the single test concentration, one also tests for whether the effect in the test concentration is less than 50 %. A simple *t*-test can be used for that purpose.

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

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

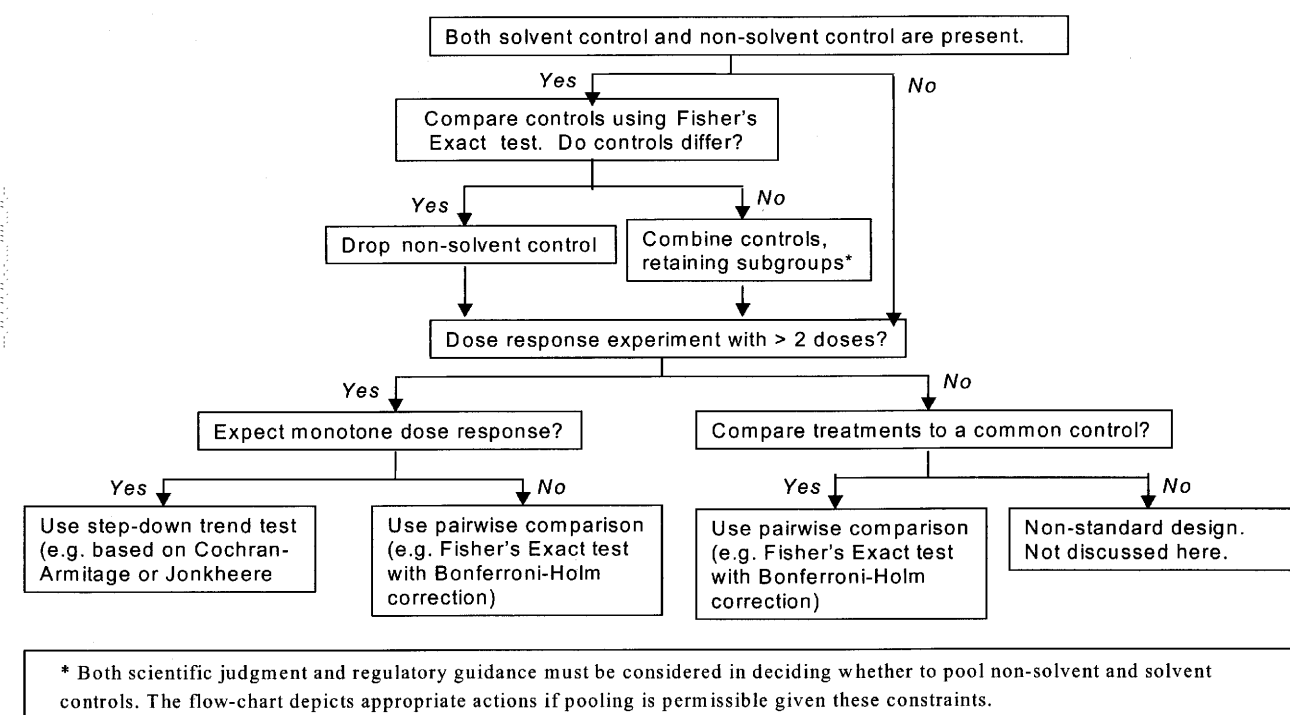
The flow-charts (Figures 3 and 4) are intended to include experiments which contain only two concentrations (control and one test concentration). Such experiments are generally referred to as limit tests and the methods described are applicable to these tests.

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

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

The tests discussed in this clause, with the exception of the Tamhane-Dunnett and Dunn tests, are all available in commercial software. For example, they are available in SAS version 8 and higher⁶⁾. The two-sided Tamhane-Dunnett test (though not called as such) is available in SAS through the studentized maximum modulus distribution provided by the probmc function. Where these tests are discussed, alternatives are provided, so that the reader can follow the general guidance of this clause without being forced to develop special programs.

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



NOTE Note that the dose count in '> 2' includes the control.

Figure 3 — Analysis of quantal data: Methods for determining the NOEC

6) SAS version 8 is an example of a suitable product available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this product.

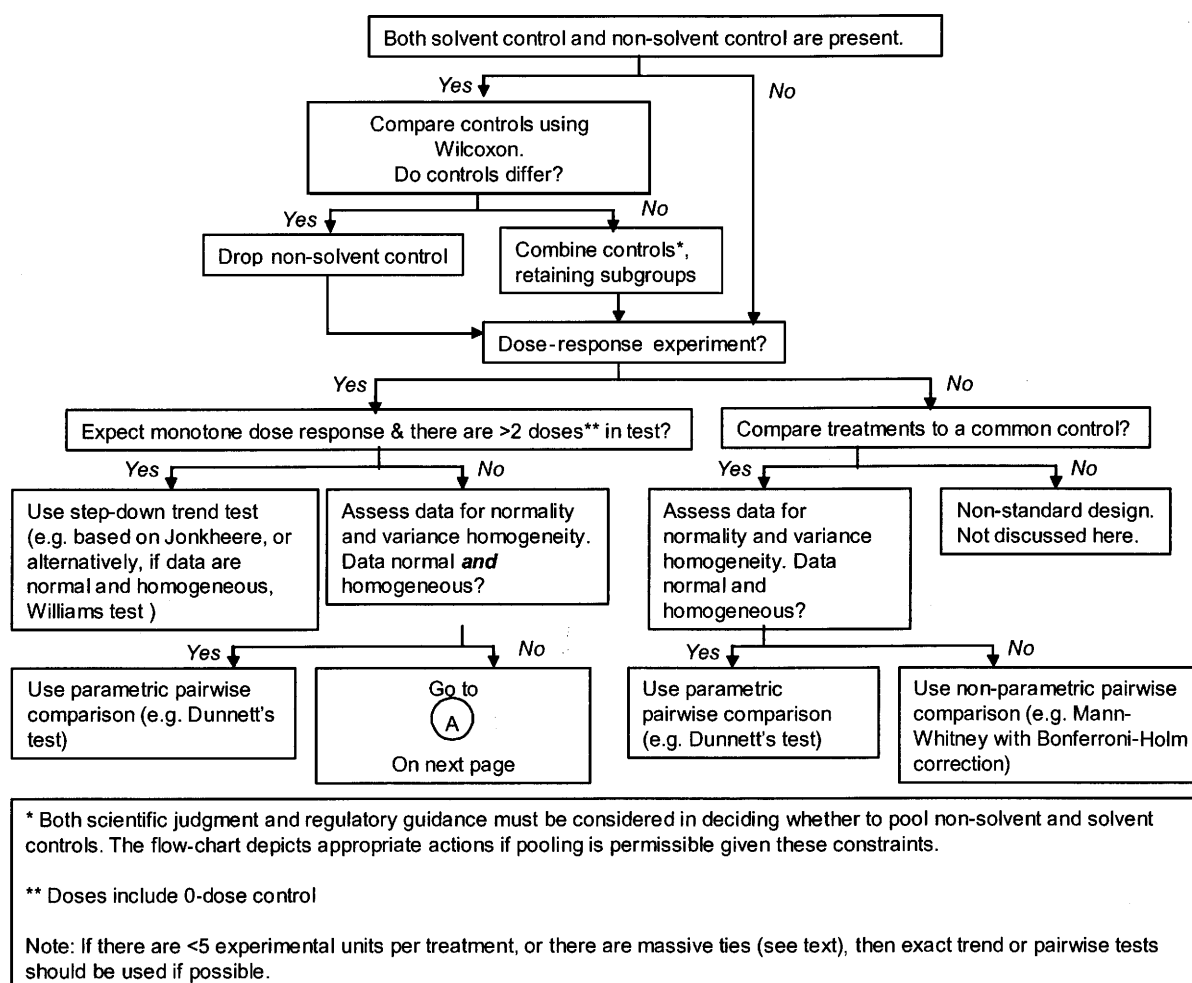


Figure 4 — Analysis of continuous data: Methods for determining the NOEC

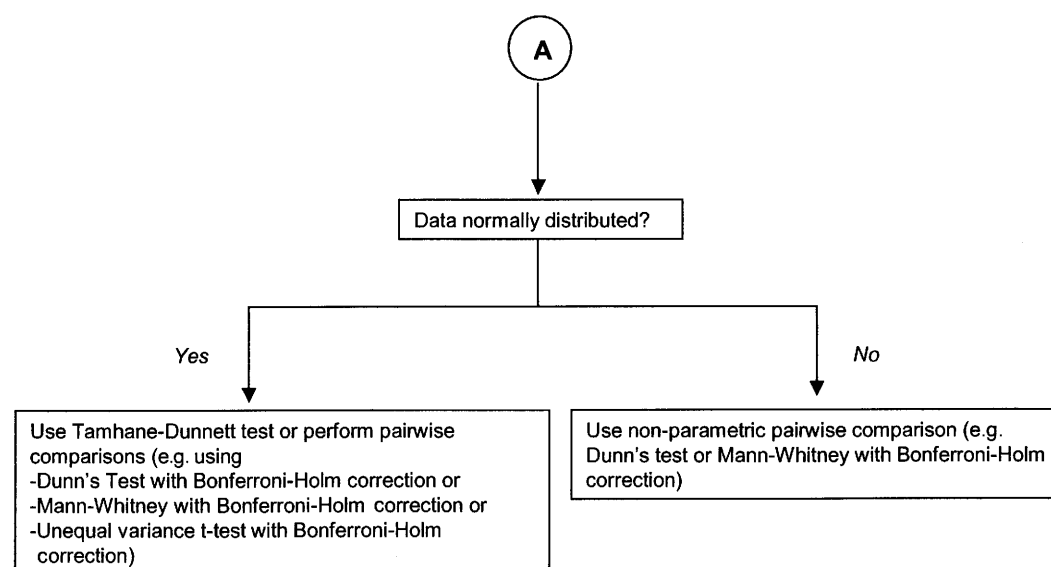


Figure 5 — Analysis of continuous data: Methods for determining the NOEC (continued)

5.1.2 NOEC: What it is, and what it is not

The NOEC (3.18) is defined as the test concentration below the lowest concentration that resulted in a significant effect in the specific experiment, i.e. the NOEC is the tested concentration just below the LOEC.

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

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

Meaningful confidence intervals for the effect size at a given concentration are sometimes possible. An application of this is discussed in 4.1.4 and methods for doing this are developed in E.3. For some techniques, obtaining meaningful confidence intervals is very difficult and this is discussed in greater detail in that annex.

5.1.3 Hypothesis used to determine NOEC

5.1.3.1 Understanding the question to be answered

The hypothesis that is tested in determining the NOEC for a toxicological experiment reflects the risk assessment question and the assumptions that are made concerning the underlying characteristics, or statistical model, of the responses being analysed (e.g. does the response increase in an orderly, i.e. monotone way with increasing toxicant concentration?).

The statistical test that is used depends on

- the hypothesis tested (e.g. are responses in all the groups equal?),
- the associated statistical model, and
- the distribution of the values (e.g. are data normally distributed?).

Thus, it is necessary to understand the question to be answered and to translate this question into appropriate null and alternative hypotheses before selecting the test procedure.

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

The simplest statistical model generally used in hypothesis testing assumes only that the distributions of responses within these populations are identical except for a location parameter (e.g. the mean or median of the distribution of values from each group).

Another statistical model that is often used assumes that there is a trend in the response that is associated with increasing exposure.

Each of these models suggests a set of hypotheses that can be tested to determine whether the model is consistent with the data. These two types of hypotheses can further be expressed as one-sided or two-sided. The discussion below is developed in terms of population means, but applies equally to hypotheses concerning population medians.

5.1.3.2 One-sided hypothesis

The most basic hypothesis (in one-sided form) can be stated as follows:

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \dots = \mu_k \text{ versus } H_1 : \mu_0 > \mu_i \text{ for at least one } i, \quad (\text{Model 1})$$

where

μ_i , $i = 0, 1, 2, 3, \dots, k$ denote the means of the control and test *populations*, respectively.

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

5.1.3.3 Two-sided trend test

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

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

5.1.3.4 Trend or pair-wise test

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

In toxicology, the treatment groups generally differ only in the exposure concentration (or dose) of a single chemical. It is further often true that biology suggests that if the chemical is toxic, then as the level of exposure is increased, the magnitude effect tends to increase. Depending on what response is measured, the effect of increasing exposure may show up as an increase or as a decrease in the measured response, but not both.

The statistical model underlying this biological expectation is what is called a trend model or a model assuming monotonicity of the population means:

$$\mu_0 \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq \dots \geq \mu_k \text{ (or with inequalities reversed)} \quad (\text{Model 2})$$

The null and alternative hypotheses can then be stated as

$$H_{02} : \mu_0 = \mu_1 = \mu_2 = \dots = \mu_k \quad \text{versus} \quad H_{12} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq \dots \geq \mu_k, \text{ with } \mu_0 > \mu_k.$$

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

Under the trend model and tests designed for that model, if tests of hypotheses H_{02} vs. H_{12} reveal that μ_3 is different from μ_0 , but μ_2 is not, the NOEC has been determined (i.e. it is the test concentration associated with μ_2), and there is no need to test whether μ_1 differs from μ_0 . Also, finding that μ_3 differs from μ_0 implies that a significant trend exists across the span of doses including μ_0 and μ_3 , the span including μ_0 and μ_4 , and so on.

For the majority of toxicological studies, a test of the trend hypothesis based on Model 2 is consistent with the basic expectations for a model for dose response. In addition, statistical tests for trend tend to be more powerful than alternative non-trend tests, and should be the preferred tests if they are applicable. Thus, a necessary early step in the analysis of results from a study is to consider each endpoint, decide whether a trend model is appropriate, and then choose the initial statistical test based on that decision. Only after it is concluded that trend is not appropriate, do specific pair-wise comparisons make sense to illuminate sources of variability.

Toxicologists sometimes do not know whether a compound causes measurements of continuous variables such as growth or mass to increase or decrease, but they are confident it acts in only one direction. For such endpoints, the two-sided trend test is appropriate, described in 4.1.7. One difference between implementing step-down procedures for quantal data and continuous data is that two-sided tests are much more likely to be of interest for continuous variables. Such a model is rarely appropriate for quantal data, as only increased incidence rate above background (control) incidence are of interest in toxicology.

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

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

or

$$\mu_0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \dots \leq \mu_k.$$

Under this model, in testing the hypothesis that all population means are equal against the alternative (that at least one inequality is strict), one first tests separately each one-sided alternative at the 0,025-level of significance with all doses present.

- If neither of these tests is significant, the NOEC is higher than the highest concentration. If both of these tests are significant, a trend-based procedure should not be used, as the direction of the trend is unclear.
- If exactly one of these tests with all the data is significant, then the direction of all further tests is in the direction of the significant test with all groups.

Thereafter, the procedure is as in the one-sided test, except all tests are at the 0,025 significance level to maintain the overall 0,05 false positive rate.

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

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

5.1.4 Comparisons of single-step (pair-wise comparisons) or step-down trend tests to determine the NOEC

5.1.4.1 General discussion

In general, determining the NOEC for a study involves multiple tests of hypotheses (i.e. a family of hypotheses is tested), and either pair-wise comparisons of treatment groups, or a sequence of tests of the significance of trend. For these reasons, statisticians have developed tests to control the family-wise error rate, FWE (the probability that one or more of the null hypotheses in the family are rejected incorrectly), in the multiple comparisons performed to identify the NOEC.

EXAMPLE For example, suppose one compares each of ten treatments to a common control using a simple *t*-test with a false positive error rate of 5 % for each comparison. Suppose further that none of the treatments has an effect, i.e. all of the treatment and control population means are equal. For each comparison, there is a 5 % chance of finding a significant difference between that sample treatment mean and the control. The chance that at least one of the ten comparisons is wrongly declared significant is much higher, possibly as high as $1 - 0,95^{10} = 0,4$ or 40 %.

The method of controlling the family-wise error rate has important implications for the power of the test. There are two approaches that are discussed: single-step procedures and step-down procedures. There are numerous variations within each of these two classes of procedures that are suited for specific data types, experimental designs and data distributions.

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

5.1.4.2 Single-step procedures

Single-step procedures amount to performing all possible comparisons of treatment groups to the control. Multiple comparisons to the control may be made, but there is no ordered set of hypotheses to test, and no use of the sequence of outcomes in deciding which comparisons to make. Examples of the single-step approach include the use of the Fisher's Exact test, the Mann-Whitney, Dunnett and Dunn tests. Since many comparisons to the control are made, some adjustment shall be made for the number of such comparisons to keep the family-wise error (FWE) rate at a fixed level, generally 0,05. With tests that are inherently single comparison tests, such as Fisher's exact and Mann-Whitney, a Bonferroni adjustment can be made: a study with k treatment levels would be analysed by performing the pair-wise comparisons of each of the treatment groups to the control group, each performed at a significance level of α/k instead of α . (This is the Bonferroni adjustment.) Equivalently, the calculated p -value ignoring multiplicities is multiplied by k . That is, $p_i^b = k \cdot p_i$. The Bonferroni adjustment is generally overly conservative, especially for large k . Modifications reduce the conservatism while preserving the FWE at 0,05 or less.

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

Table 2 — Comparison of adjusted and unadjusted p -values

Comparison	Unadjusted p value	Bonferroni-Holm Adjusted p value $p^*(i)$	Bonferroni Adjusted p -values p^b_i
Control – Treatment 4	$p_{(1)} = 0,002$	$0,002^*4 = 0,008$	$0,002^*4 = 0,008$
Control – Treatment 2	$p_{(2)} = 0,013$	$0,013^*3 = 0,039$	$0,013^*4 = 0,052$
Control – Treatment 3	$p_{(3)} = 0,020$	$0,020^*2 = 0,040$	$0,02^*4 = 0,08$
Control – Treatment 1	$p_{(4)} = 0,310$	$0,310^*1 = 0,310$	$0,310^*4 = 1$

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

5.1.4.3 Step-down procedures

Step-down procedures are generally preferred where they are applicable. All step-down procedures discussed are based on a sequential process consisting of testing an ordered set of hypotheses concerning means, ranks, or trend. A step-down procedure based on trend (for example) works as follows.

- First, the hypothesis that there is no trend in response with increasing dose is tested when the control and all dose groups are included in the test.
- Then, if the test for trend is significant, the high dose group is dropped from the data set, and the hypothesis that there is no trend in the reduced data set is tested.

This process of dropping treatment groups and testing is continued until the first time the trend test is non-significant. The highest dose in the reduced data set at that stage is then declared to be the NOEC. Distinguishing features of step-down procedures are that the tests of hypothesis shall be performed in a given order, and that the outcome of each hypothesis test is evaluated before deciding whether to test the next hypothesis in the ordered sequence of hypotheses. It is these two aspects of these procedures that account for controlling the family-wise error (FWE) rate.

A step-down method typically uses a critical level larger than that used in single-step procedures, and seeks to limit the number of comparisons that need to be made. Indeed, the special class of "fixed-sequence" tests described below fix the critical level at 0,05 for each comparison but bound the FWE rate at 0,05. Thus, step-down methods are generally preferable to the single-step methods as long as the response means are monotonic.

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

Procedures that employ step-down trend tests have more power than procedures that rely on multiple pair-wise comparisons when there is a monotone dose response because they make more use of the biology and experimental design being analysed. When there is a monotone dose response, procedures that compare single treatment means or medians against the control, independent of the results in other treatments (i.e. single-step procedures), ignore important and relevant information, and suffer power loss as a result.

The trend models used in the step-down procedures do not assume a particular precise mathematical relationship between dose and response, but rather use only monotonicity of the dose-response relationship.

The underlying statistical model assumes a monotone dose response in the *population* means, not the *observed* means.

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

5.1.4.4 Deciding between the two approaches

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

A second philosophy recommends formal tests to determine if there is significant monotonicity or significant departure from monotonicity. With continuous data, one can use either

- a) a positive test for monotonicity (such as Bartholomew's test) and proceed only if there is evidence of monotonicity, or
- b) use a "negative" test for departure from monotonicity (such as sets of orthogonal contrasts for continuous responses and a decomposition of the Chi-squared test of independence for quantal responses) and proceed unless there is evidence of non-monotonicity.

Details on these procedures are given in E.1 and E.3. Either philosophy is acceptable. The second approach is grounded in the idea that monotonicity is the rule, and that it should take strong evidence to depart from this rule. Both approaches reduce the likelihood of having to explain a significant effect at a low or intermediate concentration when higher concentrations show no such effect. The "negative" testing approach is more consistent with the way tests for normality and variance homogeneity are used and is more likely to result in a trend test than a method that requires a significant trend test to proceed. This is what is shown in the flow diagrams presented below.

Formal tests for monotonicity are especially desirable in a highly automated test environment.

One simple procedure that can be used in this situation for continuous responses is to construct linear and quadratic contrasts of normalized rank statistics (to avoid the complications that can arise from non-normal or heterogeneous data). If the linear contrast is not significant and the quadratic contrast is significant, there is evidence of possible non-monotonicity that calls for closer examination of the data or pair-wise comparison methods. Otherwise, a trend-based analysis is used.

A less simple, but more elegant procedure would be to construct simultaneous confidence intervals for the mean responses assuming monotonicity (i.e. isotonic estimators based on maximum likelihood criteria; see E.3) and use a trend approach unless one or more sample (i.e. non-isotonic) means fall outside the associated confidence interval. For quantal data using the Cochran-Armitage test, there is a built-in test for lack of monotonicity.

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

5.1.5 Dose metric in trend tests

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

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

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

5.1.6 Role of power in toxicity experiments

The adequacy of an experimental design and the statistical test used to analyse study results are often evaluated in terms of the power of the statistical test. Power is defined as the probability that a false null hypothesis is rejected by the statistical test in favour of a true alternative. That power depends on the alternative hypothesis. In the context of toxicology, the larger the effect, the higher the power to detect that effect. So, if a toxicant has had some effect on the organisms in a toxicity test, power is the probability that a difference between treatment groups and the control is detected. The power of a test can be calculated if we know the size of the effect to be detected, the variability of the endpoint measured, the number of treatment groups, and the number of replicates in each treatment group. (Detailed discussions are given in 5.2 and 5.3 and E.1 and E.3).

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

The primary use of power analysis in toxicity studies is in the design stage. By demonstrating that a study design and test method have adequate power to detect effects that are large enough to be deemed important, if we then find that, at a given dose, there is no statistically significant effect, we can have some confidence that there is no effect of concern at that dose. However, power does not quantify this confidence. Failure to adequately design or control an experiment so that statistical tests have adequate power can result in large effects being found to be statistically insignificant. On the other hand, it is also true that a test can be so powerful that it finds statistically significant effects of little importance.

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

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

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

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

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

5.1.7 Experimental design

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

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

The trade-off between number of subjects per subgroup and number of subgroups per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-subgroup variation and correlation. If there are no subgroups, then there is no way to distinguish housing effects from concentration effects and neither between- and within-group variances nor correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to subgroup means other than the Jonckheere-Terpstra test. Thus, a minimum of two subgroups per concentration is recommended; three subgroups are much better than two; four subgroups are better than three. The improvement in modelling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modelling is improved if we get better estimates of

both among- and within-subgroup variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance has, 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 verifies the statements made.) The precise needs for a given experiment depend on factors such as the relative and absolute size of the between- and within-replicate variances. Examples 1 and 2 in Annex E.3 illustrate the trade-offs between replicates per concentration and subjects per replicate.

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

Since the control group is used in every comparison of treatment to control, consideration should be given to allocating more subjects to the control group than to the treatment groups in order to optimize power for a given total number of subjects. The optimum allocation depends on the statistical test to be used. A widely used allocation rule was given by Dunnett (1955), which states that for a total of N subjects and k treatments to be compared to a common control, if the same number, n , of subjects are allocated to every treatment group, then the number, n_0 , to allocate to the control to optimize power is determined by the so-called square-root rule. By this rule, the value of n is (the integer part of) the solution of the equation $N = kn + n\sqrt{k}$, and $n_0 = -kn$. (It is almost equivalent to say $n_0 = n\sqrt{k}$.) This has been shown to optimize power for Dunnett's test. It is used, often without formal justification, for other pair-wise tests, such as the Mann-Whitney and Fisher exact tests. Williams (1972) showed that the square-root rule may be somewhat sub-optimal for his test and optimum power is achieved when $\sqrt{k}$ in the above equation is replaced by something between $1,1\sqrt{k}$ and $1,4\sqrt{k}$.

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

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

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

5.1.8 Treatment of covariates and other adjustments to analysis

It is sometimes necessary to adjust the analysis of toxicity data by taking into account some restriction on randomization, compartmentalization (housing) or by taking into account one or more covariates that might affect the conclusions.

EXAMPLE Examples of potential covariates include: initial body masses, initial plant heights and age at start of test.

While a thorough treatment of this topic is not presented, some attention to this topic is in order.

For continuous, normally distributed responses with homogeneous variances, analysis of covariance (ANCOVA) is well developed. Hocking (1985) and Milliken and Johnson (1984) are among the many

references on this topic. For continuous responses that do not meet the normality or homogeneity requirements, non-parametric ANCOVA is available.

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

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

Adjustments shall be made to statistical methods when there are restrictions on randomization of subjects such as housing of subjects together. This is discussed for both quantal and continuous data in 5.2.2.5, 5.2.3, and 5.3.7.1, where the possibility of correlations among subjects housed together is considered, as are strategies for handling this problem. In the simple dose-response designs being discussed in this clause, other types of restrictions on randomization are less common. However, there is a large body of literature on the treatment of blocking and other issues that can be consulted. Hocking (1985) and Milliken and Johnson (1984) contain discussions and additional references.

Transformation of the doses (i.e. *not* response measures) in hypothesis testing is restricted, in this clause, to the use of rank order of the doses. For many tests, the way that dose values (actual or rank order) are expressed has no effect on the results of analysis. An exception is the Cochran-Armitage test. (See E.1.)

5.2 Quantal data (e.g. mortality, survival)

5.2.1 Hypothesis testing with quantal data to determine NOEC values

In this clause, the selection of methods and experimental designs for determining NOEC values focuses on identifying the tests most appropriate for detecting effects. The appropriateness of a given method hinges on the design of the experiment and the pattern of responses of the experimental units. Figure 3 illustrates an appropriate scheme for method selection, and identifies several statistical methods that are described in detail below. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

If there are two negative controls (i.e. solvent and non-solvent), Fisher's Exact test, applied just to the two controls, is used to determine whether the two groups differ wherever it is appropriate to analyse individual sampling units. Where replicate means or medians are the unit for analysis, the Mann-Whitney rank sum test can be used. Further discussion of when each approach is appropriate is given in 5.2.2 and 5.2.2.4. 4.2.5 contains discussions of issues regarding multiple controls in an ecotoxicity study.

Figure 3 identifies a number of powerful methods for the analysis of quantal data. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

The methods used for determining NOEC values on quantal data can be categorized according to whether the tests involved are parametric or non-parametric and whether the methods are single-step or step-down. Table 3 lists methods that can be used to determine NOEC values. Some of these methods are applicable only under certain circumstances, and some methods are preferred over the others.

Except for the two Poisson tests, those tests listed in the column “Parametric” can be performed only when the study design allows the proportion of organisms responding in replicated experimental units to be

calculated (i.e. there are multiple organisms within each of multiple test vessels within each treatment group). Such a situation yields multiple responses, namely proportions, for each concentration, and these proportions can often be analysed as continuous. For very small samples, such a practice is inappropriate.

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

Table 3 — Methods used for determining NOEC values with quantal data

Method	Parametric	Non-Parametric
Single-step ^a (pair-wise)	Dunnett Poisson comparisons	Mann-Whitney ^c with Bonferroni-Holm correction Chi-squared with Bonferroni-Holm correction Steel's Many-to-One Fisher's Exact test with Bonferroni-Holm correction
Step-down ^b (trend-based)	Poisson trend Williams Bartholomew Welsch Brown-Forsythe Sequences of linear contrasts	Cochran-Armitage Jonckheere-Terpstra test Mantel-Haenszel
NOTE The tests listed in this table are well established as tests of the stated hypothesis in the statistics literature.		
^a All listed single-step methods are based on pair-wise comparisons.		
^b All step-down methods are based on trend tests.		
^c The Mann-Whitney test is identical to the Wilcoxon rank-sum test.		

5.2.2 Parametric versus non-parametric tests

5.2.2.1 Basis

Parametric tests are based on assumptions that the responses being analysed follow some given theoretical distribution. Except for the Poisson methods, the tests listed in Table 3 as parametric all require that the data be approximately normally distributed (possibly after a transformation). The normality assumption can be met for quantal data only if the experimental design includes treatment groups that are divided into subgroups, the quantal responses are used to calculate proportions responding in each of the subgroups, and these proportions are the observations analysed. These proportions are usually subjected to a normalizing transformation (see 4.3.2, 4.3.3 and 4.3.4), and a weighted ANOVA is performed, perhaps with weights proportional to subgroup sizes (Cochran 1943). (It is noteworthy that some statistical packages, such as SAS version 6⁷⁾, do not always perform multiple comparisons within a weighted ANOVA correctly.) This approach limits the possibilities of doing trend tests to those based on contrasts, including Welsch and Brown-Forsythe tests (Roth 1983, Brown and Forsythe 1974). Non-trend tests include versions of Dunnett's test for pair-wise comparisons allowing for unequal variances (Dunnett 1980, Tamhane 1979). These methods may not perform satisfactorily for quantal data, partly due to a loss of power in analysing subgroup proportions. An example is given in E.1.

The Cochran-Armitage test is listed as non-parametric, even though it makes explicit use of a presumed binomial distribution of incidence within treatment groups. Some reasons for this are given in E.1. Fisher's Exact test is likewise listed as non-parametric, even though it is based on the geometric distribution. The Jonckheere-Terpstra test applied to subgroup proportions is certainly non-parametric. An advantage of Jonckheere-Terpstra over the cited parametric tests is that the presence of many zeros poses no problem for

7) SAS version 6 is an example of a suitable product available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this product.

the analysis and it provides a powerful step-down procedure in both large- and small-sample problems, provided the number of subgroups per concentration is not too small. An example in E.3 illustrates this concern.

5.2.2.2 Single-step procedures

Suitable single-step approaches for quantal data are Fisher's Exact test and the Mann-Whitney test to compare each treatment group to the control, independently of other treatment groups, with a Bonferroni-Holm adjustment. Details of these tests are given in E.1.

5.2.2.3 Step-down procedures

5.2.2.3.1 Choice of step-down procedure

Suitable step-down procedures for quantal data are based on the Cochran-Armitage and Poisson trend tests. First, a biological determination is made whether or not to expect a monotone dose response.

- If that judgement is to expect monotonicity, then the step-down procedure described below is followed, unless the data strongly indicate non-monotonicity.
- If the judgement is not to expect monotonicity, then Fisher's Exact test is used.

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

5.2.2.3.2 Test for monotone dose response

If one believes on biological grounds that there is a monotone dose response, then the expected course of action is to use a trend test. However, statistical procedures should not be followed mindlessly. Rather, one should examine the data to determine whether it is consistent with the plan of action. There is a simple and natural way to check whether the dose response is monotone.

The $k - 1$ df Pearson Chi-squared statistic decomposes into a test for linear trend in the dose response and a measure of lack of fit or lack of trend,

$$\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$$

where

$\chi^2_{(1)}$ is the 1 df calculated Cochran-Armitage linear trend statistic; and

$\chi^2_{(k-2)}$ is the $k - 2$ df Chi-squared test statistic for lack of fit.

The details of the computations are provided in E.1.

If the trend test is significant when all doses are included in the test, then proceed with a trend-based step-down procedure.

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