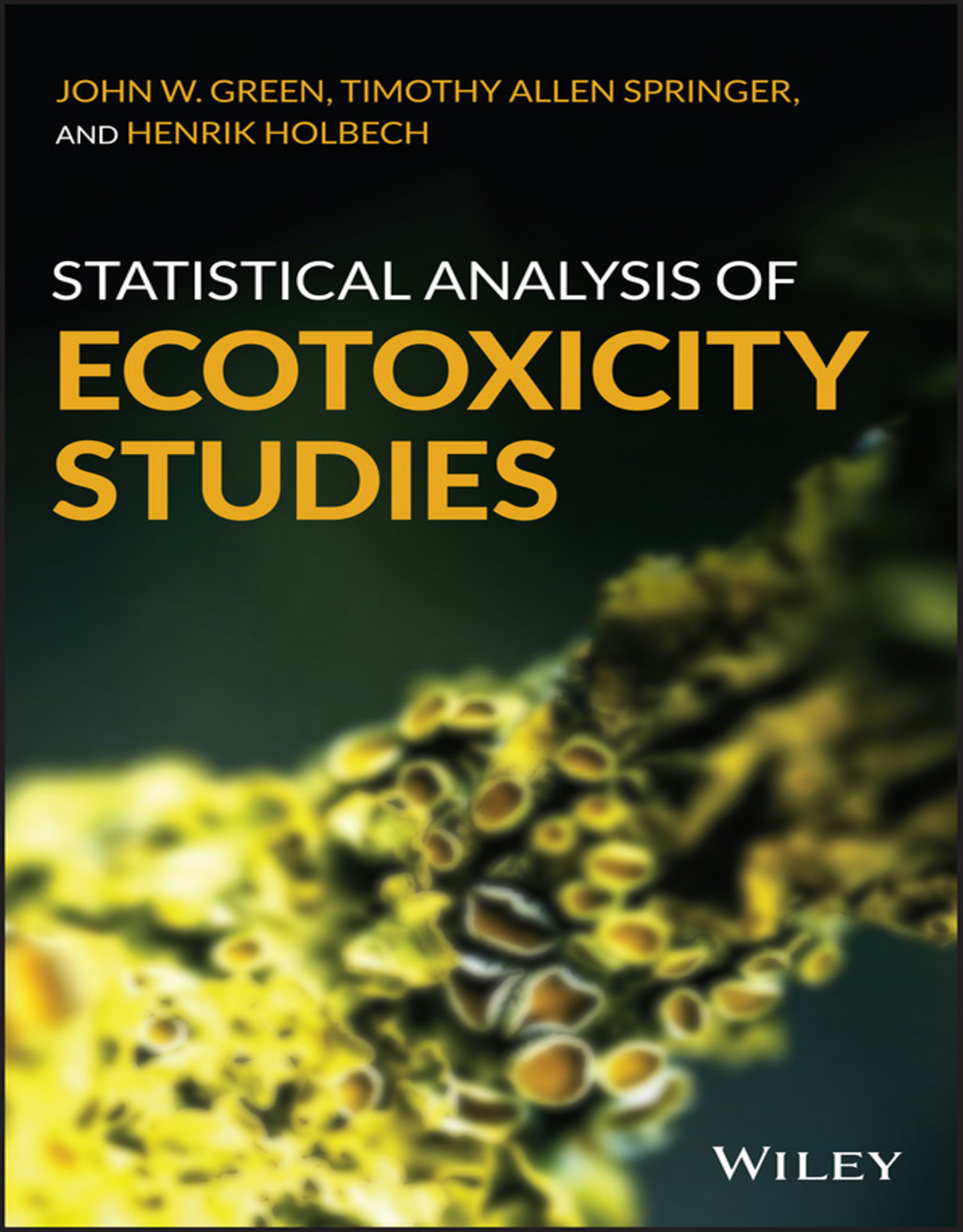
JOHN W. GREEN, TIMOTHY ALLEN SPRINGER,
AND HENRIK HOLBECH

STATISTICAL ANALYSIS OF ECOTOXICITY STUDIES



WILEY

Statistical Analysis of Ecotoxicity Studies

Statistical Analysis of Ecotoxicity Studies

John W. Green, Ph.D. (Mathematics), Ph.D. (Statistics)
Principal Consultant: Biostatstics
DuPont Data Science and Informatics

Timothy A. Springer, Ph.D. (Wildlife and Fisheries Science)
Director of Special Projects and IT Operations
EAG/Wildlife International

Henrik Holbech, Ph.D. (Ecotoxicology)
Associate Professor
Institute of Biology, University of Southern Denmark

WILEY

This edition first published 2018

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Green, John William, 1943– author. | Springer, Timothy A., 1948 March 3– author. | Holbech, Henrik, 1969– author.
Title: Statistical analysis of ecotoxicity studies / by John W. Green, Ph.D., Timothy A. Springer, Ph.D., Henrik Holbech, Ph.D.
Description: First edition. | Hoboken, NJ : John Wiley & Sons, 2018. | Identifiers: LCCN 2018008775 (print) | LCCN 2018014558 (ebook) | ISBN 9781119488828 (pdf) | ISBN 9781119488811 (epub) | ISBN 9781119088349 (cloth)
Subjects: LCSH: Environmental toxicology–Statistical methods. | Toxicity testing–Statistical methods.
Classification: LCC QH541.15.T68 (ebook) | LCC QH541.15.T68 G74 2018 (print) | DDC 615.9/07–dc23
LC record available at https://lccn.loc.gov/2018008775

Cover design by Wiley
Cover image: © Ross Collier/Alamy Stock Photo

Set in 10.25/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

John Green and Tim Springer developed a one-day training course, Design and Analysis of Ecotox Experiments, for the Society for Environmental Toxicology and Chemistry (SETAC) and delivered it for the first time at the SETAC Europe 13th Annual Meeting in Hamburg, Germany, in 2003. Since then, in many years we have taught this course at the annual SETAC conferences in Europe and North America, updating it each time to stay abreast of the evolving regulatory requirements. In 2011, Henrik Holbech joined us and has made valuable contributions ever since. In 2014, Michael Leventhal of Wiley approached us with the idea of turning the training course into a textbook. The result is the current book, and we appreciate the opportunity to reach a wider audience.

This book covers the statistical methods in all current OECD test guidelines related to ecotoxicity. Most of these have counterparts in the United States Environmental Protection Agency (USEPA) guidelines. In addition, statistical methods in several WHO and UN guidelines are also covered, as are guidelines in development or that have been proposed. Chapter 11 provides a good coverage of all the test guidelines covered in this book with reference to the chapters in which guideline-specific statistical methods are developed. With very few exceptions, the data used in the examples and exercises are from studies done for product submissions or in developing some regulatory test guideline. The authors have been members for a combined total of more than 30 years of the OECD validation management group for ecotoxicity (VMG-eco) responsible for development and update of significant portions of numerous current test guidelines including OECD TG 210, 229, 230, 234, 236, 240, 241, 242, and 243. We have also been actively involved in designing and analyzing ecotoxicity studies for more than a combined total of 60 years. One or more of us were also members of the expert groups that developed (i) the European Framework for Probabilistic Risk Assessment (Chapman et al., 2007), (ii) OECD *Fish Toxicity Testing Framework* (OECD, 2014c), (iii) *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application* (OECD, 2014a, 2006a), (iv) OECD test guideline 223 that describes a sequential test designed to measure mortality in avian acute tests, (v) OECD *Guidance Document on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption* (OECD, 2012a) and (vi) OECD test guideline 305 for assessing bioaccumulation in fish.

Our intent is to provide an understanding of the statistical methods used in the regulatory context of ecotoxicity. However, the coverage and treatment of the topics should appeal to a much wider audience. A mathematical appendix is included to provide technical issues, but the focus is on the practical aspects of model fitting and hypothesis tests. There are numerous exercises based on real studies to help the reader enhance his or her understanding of the topics. Ample references are provided to allow the interested reader to pursue topics in greater depth. We have not shied away from controversies in the field. We think it important that the reader understand that statistics is not free of controversy and should be well-informed on these issues. Nonetheless, while we have points of view on these topics and express them, we have tried to take an even-handed approach in describing the different points of view and provide references to allow the reader to more fully appreciate the arguments on these issues.

A frequent question from participants in the training course was where one could find software to carry out the methods of analysis we taught and were required or at least recommended in regulatory test guidelines. While we have developed in-house proprietary SAS-based software for this purpose, it has not been possible to share it. One of the benefits of this textbook is the availability of a website created by Wiley where we are providing SAS and R programs for almost all methods presented. In some instances, rather than present programs, we provide a link to free online

software that has been developed for specific guidelines or for a more general use. In some cases, we have been unable to find R programs to carry out the recommended methods. For those cases especially, we invite the readers of this book to develop and send such programs to us. In a few cases, no SAS program is provided. In all cases, a program or link is provided for all analyses discussed. After we test programs supplied by readers, we will put them on the website with appropriate acknowledgments. Also, if any shortcomings are found in the initially provided programs, we encourage the readers to bring them to our attention and we will post corrections or improvements. As regulatory requirements change or methods improve, we will update the website.

We have had support from numerous people over the years in developing the training material and the material for this book. Colleagues too numerous to name from DuPont, Wildlife International/EAG, USEPA, OECD, and other companies, universities, and CROs have contributed ideas and data that have been very helpful in improving our understanding of ecotoxicology. Two instructors joined us, Michael Newman of Virginia Institute of Marine Science, School of Marine Science, The College of William and Mary, and Chen Teel of DuPont, each for one offering of the course and both added value. In addition, we have SAS expertise, but more limited experience with R. As a consequence, while we developed some R programs ourselves, several very capable people were engaged to develop most R programs for the website. Several deserve special acknowledgment. We have modified their programs in minor ways to fit the needs of the website and accept responsibility for any errors.

Joe Swintek is a statistician working with the Duluth office of the USEPA. He was a contributor to one of our publications (Green et al., 2014) and turned the SAS version of the StatCHARRMS software John and Amy Saulnier developed under contract for the USEPA into an R package. The SAS version is provided in Appendix 1 (the website) and the R version is now in the CRAN library. A link is provided in the references (Swintek, 2016). In addition to the RSCABS program for histopathology severity scores (Chapter 9), StatCHARRMS contains the Dunnett and Dunn tests, the step-down trend tests Jonckheere–Terpstra (Chapter 3), Cochran–Armitage and Fisher's exact tests (Chapter 6), Shapiro–Wilk and Levene tests for normality and variance homogeneity (Chapter 3), and repeated measures ANOVA for multi-generation medaka reproduction studies (Chapter 5). Several of these tests are provided in Appendix 1 in stand-alone versions, as well as in the full CRAN version. In addition, Joe developed a versatile R program for the important Williams' test, and that is in Appendix 1 and has been added to the StatCHARRMS package. We were surprised to find that this test had not previously been released in an R package, so far as we are aware. There is an R package, multcomp, that refers to Williams' type contrasts within the function mcp, but the results deviate substantially from Williams' test. We have verified with the developer, Ludwig Hothorn, that package mcp does not provide Williams' test. More discussion on this is provided in Chapter 3. Joe also provided numerous other R programs for several chapters as well as pointing out a simple R function based on the package sas7bdat for reading a SAS dataset into R without the need to have SAS installed or converting the dataset to excel or text first. We are very grateful for his contributions.

Chapter 13 leans heavily on discussions of the expert group that developed guidance on implementation of OECD test guideline 305 on bioaccumulation in fish. In particular, Tom Aldenberg of RIVM has provided invaluable communications to us concerning the R program, bcmfR, that he has provided to OECD for analysis of bioconcentration and biomagnification studies.

Georgette Asherman also deserves special mention, primarily for her R programming work for Chapter 5. Among her notable contributions were versatile and robust versions of the Shapiro–Wilk and Levene tests, the Shirley nonparametric ANCOVA program, two parametric ANCOVA programs, programs to add confidence bounds to the graphic output for nonlinear regression, and zero-inflated binomial and beta-binomial models.

Erand Smakaj provided training in the use of R-Studio and contributed programs for survival analysis and for several topics in Chapter 13 and was very accommodating throughout the text and code development.

Xiaopei Jin made important contributions to the R programs for Chapter 8 and demonstrated useful capabilities of R that can be applied to programs in all chapters.

Finally, we would be remiss not to acknowledge the many contributions Amy Saulnier has made to SAS programming used in this book and elsewhere. John has worked with Amy over the entire 29+ years of his DuPont career. In addition to turning his SAS programs into the user-friendly StatCHARRMS program, she has done the same for two other heavily used SAS-based in-house software packages routinely used for our toxicology and ecotoxicology analyses for regulatory submissions. She has maintained these programs, updated them as needed to stay current with regulatory requirements and changes in the computing environment, and has been an essential contributor to DuPont's work for over three decades.

The term GLMM is used for generalized linear models regardless of whether there is a random term. This encompasses both generalized linear mixed models and fixed effects models. The term GLM is reserved to the classic general linear model with normal errors.

Acknowledgments

In addition to the people mentioned above for programming and other professional support, John would like to acknowledge his wife Marianne, without whose unwavering support and understanding, this book would not have been possible. He would also like to acknowledge the support he received from his daughters and step-daughter, M'Lissa, Janel, and Lauren, who encouraged him throughout. Finally, he would like to thank his companions Sam, Max, Ben, and of course Jack for their warmth and comfort through the countless hours devoted to this work. Henrik would like to acknowledge his wife Bente, always supporting the work on the book.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/Green/StatAnalysEcotoxicStudy

The companion website contains programs in SAS and R to carry out the analyses that are described in the text. These programs will be updated as improvements are identified or regulations change. Readers are invited to send corrections or improvements to the authors through Wiley. Once these are verified and judged appropriate, they will be added to the website with appropriate acknowledgment. Also on the website are datasets referenced in the text but too large to include there. These are in the form of excel files or SAS datasets. An R program is provided to convert SAS datasets to R without the need to have access to SAS. In a few instances noted in the text, links are given to specialized programs developed specifically for some regulatory test guideline when there seemed no purpose in creating a new program.

Chapter 1

An Introduction to Toxicity Experiments

This chapter introduces some basic concepts that apply to all chapters. It begins with a discussion of the nature of toxicology or ecotoxicology studies that distinguish them from experiments more generally. Then some basic experimental design issues are discussed, such as types of control groups, replicates and pseudo-replicates, and units of analysis. The various types of responses that occur are introduced, with pointers to chapters in which methods of statistical analysis of the various types of response are developed. An introduction is given to the use of historical controls and how these studies relate to regulatory risk assessment of chemicals in the environment. Then a hierarchy of statistical models is provided that, in broad terms, defines the statistics used in this field of study and, specifically, in this text. Finally, a topic is introduced that is the cause of considerable tension in ecotoxicology and biological analysis of data in general, namely the difference between biological and statistical significance.

1.1 NATURE AND PURPOSE OF TOXICITY EXPERIMENTS

The purpose of a toxicity experiment is to obtain a quantifiable measure of how toxic a given substance is to a group of organisms or community of organisms. The primary purpose of this book is to describe the design and statistical analysis of laboratory experiments on groups of organisms of a single species exposed to controlled levels of a substance thought to have the potential to produce an adverse effect on the test organisms. Such experiments have the goal of quantifying the level of exposure to the substance that has an adverse effect of biological concern. Some consideration is also given to how information from multiple toxicity experiments on different species can be combined to assess the adverse effect of the test substance on an ecological community. This chapter is intended to provide a general overview of toxicity studies and an introduction to the topics covered in this book.

1.1.1 Designed Experiments Compared to Observational Studies

Historically, the toxicity of chemicals has been studied using experiments performed under carefully controlled conditions in the laboratory and by observation of responses in uncontrolled settings such as the environment. Observational studies that gather information by survey or monitoring have the advantage of providing insight into toxicological responses under real-world conditions. Such studies are valuable in alerting researchers to potential problems resulting from chemical exposure. However, in surveys and monitoring studies, many uncontrolled factors can affect responses, and exposure of organisms to a chemical of interest (e.g. dose and concentration) usually cannot be estimated accurately. As a result, conclusions concerning the relationship between possible toxicological responses and exposure to the chemical are difficult to establish with certainty.

On the other hand, designed experiments typically control most of the factors that affect response, and dose or exposure concentration can be accurately measured. Designed experiments performed in a laboratory are usually performed at constant temperature with constant exposure to a test substance. Control of test substance exposure and other experimental factors allow the relationship between exposure and response to be modeled.

Exposure to the test substance in these experiments may be: via food or water ingested, air breathed, from

Statistical Analysis of Ecotoxicity Studies, First Edition. John W. Green, Timothy A. Springer, and Henrik Holbech.

Companion website: www.wiley.com/go/Green/StatAnalysEcotoxicStudy

contact with the soil or sediment or contact with spray application or spray drift on plants, through gavage or intravenous injection, or by direct application to the skin or eyes. The measure of exposure can be the concentration in the food or water or air, the quantity of chemical per unit of body weight, the quantity of chemical per unit of land area, or the concentration of the chemical in the blood.

Toxicity experiments are generally classified as acute, if the exposure is of short duration relative to the life span of the organism; or subchronic, if the exposure is of medium duration relative to a full life time; or chronic, if the exposure is for approximately a normal life span of the test substance.

Toxicity is measured in many ways. In its simplest form, it refers to the exposure level that kills the whole organism (e.g. laboratory rat or fish or tomato plant). Many sublethal responses are measured and the types of measurements are varied. The types of response encountered in toxicology fall broadly into one of the following categories: Continuous, quantal, count, and ordinal. Below is an introduction to each of these types of responses together with an indication of some of the challenges and methods associated with each type. Later chapters will discuss in detail all the points mentioned here.

1.1.1.1 Continuous Response

This class includes measurements such as plant yield, growth rate, weight and length of a plant or animal, the amount of some hormone in the blood, egg shell thickness, and bioconcentration of some chemical in the flesh, blood, or feathers. Typical continuous response data are shown in Tables 7.6 and 7.7 and Figures 7.2 and 7.3.

Continuous responses also include responses that exist in theory on a continuous scale, but are measured very crudely, such as days to first or last hatch or swim-up or reproduction, or time to tumor development or death, which are observed (i.e. "measured" only once per day). Hypothesis testing methods of analyzing continuous data are presented in Chapter 3 and regression models are presented in Chapters 4 and 5.

Example 1.1 Daphnia magna reproduction

The experimental design is seven daphnid individually housed in beakers in each of six test concentrations and a water control. Once each day, it is recorded whether or not each daphnid has reproduced. Ties in first day of reproduction are very common. In this typical dataset, there were a total of six distinct values across the study. While in theory, time to reproduction is continuous, the measurement is very crude and, as will be seen in Chapters 3 and 4, analysis will be different from that for responses measured on a continuous scale.

See Figure 1.1. The solid curve connects the mean responses in the treatment groups with line segments. Recall that there are seven beakers per treatment, but many beakers have the same first day of reproduction, so each diamond can represent from 1 to 6 observations. See Table 1.1 for the actual data.

1.1.1.2 Quantal Response

Quantal measures are binary (0–1 or yes/no) measurements. A subject is classified as having or not having some characteristic. For each subject, the possible values of the response can be recorded as 0 (does not have the characteristic of interest) or 1 (has the characteristic of interest). The quintessential example is mortality. Outside Hollywood films about zombies and vampires, each subject at a given point in time is either alive (value 0) or dead (value 1). Other quantal responses include immobility, the presence

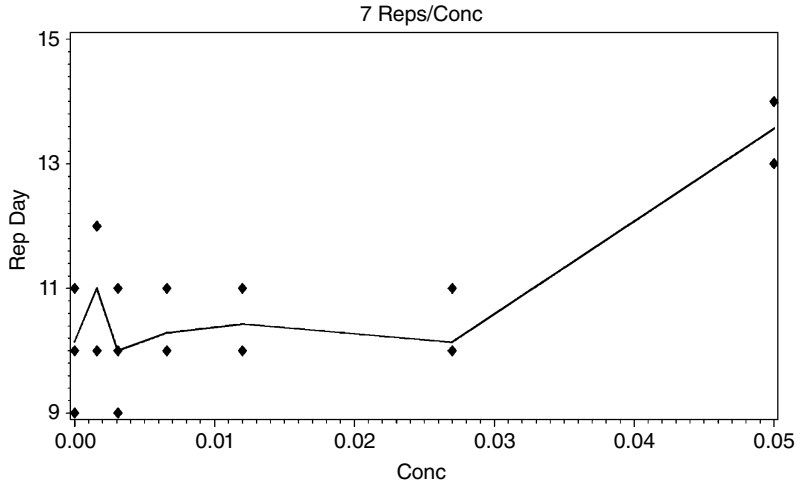


Figure 1.1 First day of daphnid reproduction. Diamonds, replicate means; solid line, joins treatment means.

Table 1.1 Daphnid First Day of Reproduction Data for Example 1.1

Conc	Rep	RepDay	Conc	Rep	RepDay	Conc	Rep	RepDay
−1	1	9	0.001 6	6	10	0.012	4	10
−1	2	9	0.001 6	7	12	0.012	5	11
−1	3	11	0.003 1	1	10	0.012	6	10
−1	4	9	0.003 1	2	11	0.012	7	10
−1	5	10	0.003 1	3	10	0.027	1	10
−1	6	9	0.003 1	4	10	0.027	2	10
−1	7	9	0.003 1	5	10	0.027	3	10
0	1	9	0.003 1	6	9	0.027	4	10
0	2	11	0.003 1	7	10	0.027	5	11
0	3	10	0.006 6	1	10	0.027	6	10
0	4	11	0.006 6	2	10	0.027	7	10
0	5	10	0.006 6	3	11	0.05	1	14
0	6	10	0.006 6	4	11	0.05	2	13
0	7	10	0.006 6	5	10	0.05	3	13
0.001 6	1	10	0.006 6	6	10	0.05	4	13
0.001 6	2	12	0.006 6	7	10	0.05	5	14
0.001 6	3	10	0.012	1	10	0.05	6	14
0.001 6	4	12	0.012	2	11	0.05	7	14
0.001 6	5		0.012	3	11			

Conc = −1 is water control. Conc = 0 is solvent control. Controls should be combined (with Rep numbers altered to distinguish replicates in the two controls) prior to further analysis, or else one control should be discarded (see Sections 1.3.1 and 1.3.2). RepDay, first day of reproduction of daphnid in the beaker.

Table 1.2 Mite Survival Data

Conc	Unit	Risk	Alive	Conc	Unit	Risk	Alive
0	1	5	3	75	1	5	4
0	2	5	5	75	2	6	2
0	3	5	5	75	3	5	3
0	4	5	5	75	4	5	2
18.75	1	5	5	150	1	5	1
18.75	2	5	5	150	2	5	1
18.75	3	5	5	150	3	5	0
18.75	4	5	3	150	4	5	0
37.5	1	5	5	300	1	5	0
37.5	2	6	6	300	2	5	0
37.5	3	5	5	300	3	5	0
37.5	4	5	2	300	4	5	0

Unit, replicate vessel; Risk, number of mites placed in vessel at study start; Alive, number of mites alive at the end of the study period.

of matted fur, pregnant, lethargic, and the presence of liver tumor. Hypothesis testing methods of analyzing quantal data are presented in Chapter 6 and regression models are presented in Chapter 7. See Table 1.2 for an example of survival data for mites.

The data in Table 1.2 are from an experiment on mites. Mites were exposed to varying levels of a pesticide as part of a risk assessment for product registration. Each housing unit consists of a frame with a glass plate at the top and bottom of the frame. Pesticide residue is sprayed on the inner side of each glass plate. For the control, water is sprayed on the inner plate surface. After the plates dry, mite protonymphs are placed between the plates. Fresh air is circulated within the frame by an air pump. The mites are examined 7 days after exposure begins. Risk is the number of mites in each housing unit. Alive is the number alive at the end of the experimental period. The concentrations were in ppm. There were nominally five mites per unit, including control. Due to initial counting problems two units actually included six mites. Chapters 6 and 7 will discuss how to analyze such data.

1.1.1.3 Count Response

While quantal responses involve counts of the number of animals with the characteristic of interest, as we use the term, counts are the number of occurrences in a single subject or housing unit of some property. These include the number of eggs laid or hatched, the number of cracked eggs, the number of fetuses in a litter, the number of kidney adenomas, and the number of micronucleated cells. See Table 1.3 for an example dataset from Hackett et al. (1987) showing variable litter sizes and sex ratios.

Methods for analyzing count data will be presented in Chapter 8. As will be discussed there, count data can sometimes be analyzed as though it were continuous (usually after a transformation). Count data can also be analyzed through specialized distributions, such as Poisson

Table 1.3 Mouse Litter Size and Sex Ratio

Conc	Dam	Males	Litter	Conc	Dam	Males	Litter	Conc	Dam	Males	Litter
0	1	6	9	40	27	8	16	200	53	7	12
0	2	6	10	40	28	6	11	200	54	7	15
0	3	2	3	40	29	6	12	200	55	0	2
0	4	5	13	40	30	3	6	200	56	8	16
0	5	7	12	40	31	7	13	200	57	7	15
0	6	5	13	40	32	6	11	200	58	9	11
0	7	8	15	40	33	4	10	1000	59	4	9
0	8	7	13	40	34	8	15	1000	60	5	14
0	9	4	13	40	35	7	14	1000	61	5	11
0	10	6	13	40	36	3	14	1000	62	7	12
0	11	11	12	40	37	7	13	1000	63	6	14
0	12	6	13	200	38	7	13	1000	64	9	14
0	13	3	9	200	39	8	12	1000	65	7	15
0	14	8	13	200	40	7	12	1000	66	7	12
0	15	6	9	200	41	4	13	1000	67	9	16
0	16	3	13	200	42	6	14	1000	68	10	14
0	17	7	14	200	43	9	15	1000	69	9	14
0	18	6	14	200	44	5	11	1000	70	7	11
40	19	6	11	200	45	7	14	1000	71	4	10
40	20	3	11	200	46	6	11	1000	72	7	9
40	21	3	13	200	47	6	12	1000	73	5	12
40	22	8	14	200	48	7	11	1000	74	3	8
40	23	8	13	200	49	8	11	1000	75	6	10
40	23	11	15	200	50	5	13	1000	76	9	13
40	25	6	13	200	51	6	14	1000	77	2	11
40	26	6	12	200	52	8	12	1000	78	6	15

Dam is an ID for the pregnant female mouse. Litter is the number of fetuses for that dam. Males = number of males in the litter and conc is the exposure parentage dosage of 1,3-butadiene in ppm. Questions of interest include whether the chemical affects the litter size or sex ratio and whether there is an association between litter size and sex ratio. Fetal, placenta, and dam body weights were also included in the original dataset and other questions were also addressed.

or zero-inflated Poisson, in the context of what are called generalized linear models (GLMM). We will present and compare these methods in Chapter 8.

1.1.1.4 Ordinal Response

Ordinal responses indicate relative severity or level but not magnitude. Examples include amphibian developmental stage and histopathology severity scores. Amphibian developmental stages are represented by numbers 1–66 (as derived from Nieuwkoop and Faber, 1994), but the difference between stage 55 and 56 is not comparable to the difference between 56 and 62. The larger number indicates a more advanced development, but this development is defined by the presence or absence of specific physical characteristics, not otherwise quantifiable. Consider the following stages as examples:

1. Stage 56 typically occurs on day 38 post hatch. Forelimbs of stage 56 animals are visible beneath the skin of the tadpoles. The tadpoles are filter-feeding.
2. Stage 57 typically occurs on day 41 post hatch. Stage 57 animals lack emerged forelimbs, and metamorphosis in the alimentary canal is just beginning.
3. Stage 58 typically occurs on day 44 post hatch. Stage 58 animals have emerged forelimbs and there is significant histolysis of the duodenum (animals can no longer digest food).
4. Stage 59 typically occurs on day 45 post hatch. Stage 59 animal forelimbs now reach to the base of the hindlimb and there is now histolysis of the non-pyloric part of the stomach (animals still can no longer digest food).
5. Stage 60 typically occurs on day 46 post hatch.

In terms of development rates, a stage 57 animal is 3 days behind a stage 58 animal, whereas a stage 58 animal is only 1 day behind a stage 59 animal. Also, in terms of development rate, a stage 56 animal is 6 days behind a stage 58 animal, whereas a stage 58 animal is only 2 days behind a stage 60 animal.

The biological significance of moving between two stages might vary greatly depending on which stages are being considered. For example, a stage 56 animal can filter-feed. None of the animals in the other stages listed above can.

Developmental stage is a key endpoint in the OECD TG 231 Amphibian Metamorphosis Assay (AMA). The experimental design in the test guideline is for four tanks per test concentration, 20 tadpoles per tank, and three test concentrations plus a water control. In developing the test guideline, other designs were explored, including designs with five test concentrations plus control, two tanks per concentration, and 20 tadpoles per tank. See Table 1.4 for an example with this latter design.

In Table 1.4, there was an apparent shift right in group 5 and perhaps in group 4, but groups 2 and 3 have increased frequencies of smaller stages. It is not clear what a 10% effects concentration would mean for this response. Averaging stages in a group is meaningless (i.e. Stage 57.2 is meaningless), as stage is an ordinal, not a quantitative, variable. The response measure should not be based on simply considering the proportion of tadpoles above some stage (e.g. >stage 58), since calculation of the concentration causing a 10% increase in the percent of tadpoles with stage greater than 58 ignores the effects on the distribution of stages above and below 58. Analysis based on median stages in tanks ignores too much within-tank information. Chapter 9 will describe the analysis of such data.

Clearly, the analysis of the stage data requires care, and it is important not to think of the stages as representing equal increments of development. It should be clear that a shift in the stage of metamorphosis of a single stage *might* be, but need not be, biologically meaningful. The analyses of developmental stage data will be discussed in detail in Chapter 9.

Histopathology severity scores are similar to developmental stages in terms of being ordinal, not numeric, but differ in another way that requires a different type of analysis. Here, pathologist-grade organ slides on a scale 0–4, with score 0 meaning no abnormality was observed, score 1 meaning only a minimal abnormality, score 2 meaning mild abnormality, and scores 4 and 5 meaning moderate and severe abnormalities, respectively. It would be more accurate to describe score 0 as meaning there was nothing remarkable, rather than no abnormality. A severity score is assigned to a tissue sample by a trained pathologist. These scores depend on the type of tissue damage found and an assessment of its importance to the health of the animal. See Figure 1.2 for an example tissue slide. Assigning severity scores to such slides is not a simple exercise. More discussion of this and a more detailed example are provided in Chapter 10.

With most toxicology severity scores, there is no uniform change in severity between scores, that is, the difference between minimal and mild is not the same as the difference between mild and moderate or between moderate and severe. See Figure 1.3 for a simple illustration that may help keep these scores in mind.

Stated this way, the nature of severity scores is straightforward. Few people would suggest that if half of the tissue samples have a minimal finding and half have a moderate finding, then on average, the finding is mild.

Confusion arises from the common practice of labeling a finding of none as 0, minimal as 1, mild as 2, moderate as 3,

Table 1.4 Example Developmental Stage Data from AMA Study

Stage		56	57	58	59	60	61	62
Group	Tank							
1	1	2[a]	3	7	4	2	2	
	2			9	8	2	1	
2	1			6	9	3		2
	2		7	8	4	1		
3	1		2	9	6		3	
	2		1	6	7	2	2	2
4	1			1	13	3		3
	2			1	3	8	3	5
5	1			5	6	6	1	2
	2				5	7	3	5

Stage, developmental stage reached by some tadpole in the indicated tank; Group, treatment group, with control = group 1; Tank, replicate vessel.

[a] Number in cell is the number of tadpoles in the tank at the indicated developmental stage.

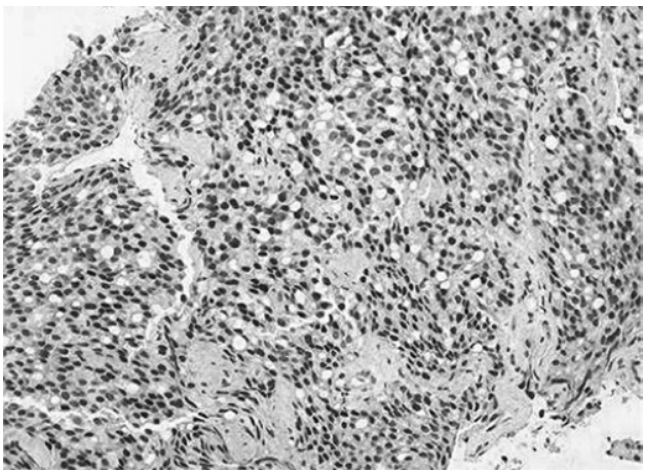


Figure 1.2 Example tissue slide for histopathology grading. Expert judgment is used to score tissue slides such as these. Image from Google images altered to black and white and cropped using Photoshop. https://image.slidesharecdn.com/cpc-4-4-2-ren-bph-pathlec-view-091013211114-phpapp02/95/pathology-of-prostate-53-728.jpg?cb=1255468480.

and severe as 4. These labels are numbers and a simple-minded statistical approach is to treat them as though these labels behave as numbers do rather than recognizing them merely as labels: that is, one can average them, compute the standard deviation, and employ all the simple statistical tools one learned in an introductory course, such as the T-test. However, moving from a score of 1 to 2 does not indicate a doubling of severity, and moving from 3 to 4 may not indicate a change in severity equal to that in moving from 1 to 2.

It should be emphasized that these scores are just labels. To average scores 1 and 2 is the same as averaging minimal and mild. What is the average of minimal and mild or of mild and severe? These scores are arbitrary except for order. We could just as well use the numbers 1, 2, 5, 7, and 12 as scores (see Figure 1.3) to emphasize that the difference between "adjacent" scores is not the same as a subject progresses from no effect to severe effect. So the average of minimal and severe could be $(1+4)/2=2.5$ or $(2+12)/2=7$. Neither average makes sense.

Such numerical approach is nonsensical, but it does highlight a real concern. If the tank in an aquatic experiment is the unit of analysis, what value do we give to the tank? Leaving aside for now how to analyze severity scores, if there are five fish in a tank with severity scores 0, 0, 3, 3, and 4, what single value do we assign to the tank for statistical analysis? Note that the arithmetic mean of these numerical labels is 2. Does 2 capture something meaningful about this set of scores?

While the mean score is objectionable, what about the median score? The median inherently treats the labels as equally spaced across the spectrum of severities. Think about where in the wide range of moderate (score 3) tissue damage assessments in Figure 1.3 the moderately damaged slide lies. As shall be discussed in Chapters 3 and 5, rank ordering is a basic idea in most nonparametric testing, and the set of all values from treatment and control are ranked as a whole and then the sum of the ranks in the treatment and control are compared. Such nonparametric tests take the spread of severity scores into account, not just the median.

One of the two approaches is typically taken in rodent histopathology analysis. (i) Apply a nonparametric test such as the Mann–Whitney (Chapter 3), which compares the median scores in treatment tanks to those in the control. But the tank median ignores the spread of the data. The tank with scores 0, 0, 3, 3, and 4 has the same median as a tank with scores 3, 3, 3, 3, and 3, but the first is much more dispersed than the second and this may signal a difference of biological importance missed by the comparison of medians. The need for a summary measure for each tank limits the appropriateness of traditional nonparametric procedures for severity score analysis. (ii) Some scientists simply do not perform a statistical analysis, either because they recognize the shortcomings of the above approach or because they have little value for statistics altogether.

Given the restricted number of possible severity scores and the small sample sizes typical in histopathology, at least in ecotoxicology studies, analysis methods for severity scores are different from those for developmental stage. See Table 1.5 for an example from a medaka multi-generation test.

In the dataset in Table 1.5, there were no score 0 fish. The empty tanks (A in treatment 1 or control and C in treatment 2) do not represent mortality. Rather, medaka could not be sexed at the initiation of the study and by chance, these tanks contained no females. This inability to know the sex at study initiation leads to highly imbalanced experimental designs. The tank is the unit of analysis, not the individual fish, it is thus important to retain tank identification and not lose the distribution of scores within the tank. Also, because fish cannot be sexed at the beginning of the study and must be analyzed by sex at the end of the

Table 1.5 Severity Scores for Liver Basophilia in Female F2 Medaka at 8 Weeks

		Frequency of score per tank						
Trt	Score	A[a]	B	C	D	E	F	Total
1	1		1[b]	1	2	2	3	9
	2		3		1			4
	3					1		1
2	1				2	1	1	4
	2	4	2		1	1		8
3	1		1			2	4	7
	2	2	2	1	1			6
4	1	1	1			1		3
	2	1		1	3	3	2	10
	3						1	1
5	1	1	2	1	1	1		6
	2			1	1	2		4
	3	1			2		1	4

Trt, treatment group, with 1=control; Total, number of fish in all tanks in the indicated treatment group with the indicated score.

[a] Tanks are labeled A, B, and F.

[b] Numbers in cells indicate the number of fish (ignoring tanks) in the treatment group with that score.

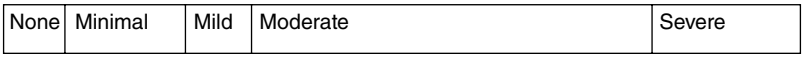


Figure 1.3 Example severity scale. Varying widths for different scores indicate possible differences in the range of severities given the same score.

study, tank sizes are highly variable and this complicates the analysis. For that reason and others, analysis of tank medians, for example, would discard important information.

Appropriate methods for analysis of ordinal data are discussed in detail in Chapter 9.

1.1.2 Analysis of Laboratory Toxicity Experiments

The variety of sublethal endpoints measured suggests the need for multiple statistical tools by which to analyze toxicity data. It is the objective of this book to discuss many of the statistical methods that have been used for this purpose and to indicate what additional tools could be brought to bear. Science is not static and advances in statistical methods and computer power and software have made available techniques that were impossible only a few years ago. It is fully expected that additional advances will be made in the time to come that cannot be foreseen today. The authors will attempt to present the main statistical methods in use now, and to the extent possible, those likely to be included in the near future.

In its simplest form, a toxicity experiment is conducted on a single species for a fixed amount of time. Different groups of subjects are exposed to difference levels of the test substance. More complex experiments include other factors, such as measurements of lethal and sublethal effects over time, differences among the sexes of the subjects, different ambient conditions, and mixtures of chemicals. The object of the statistical analysis is to identify the level of exposure that causes a biologically meaningful adverse effect under each set of conditions in the experiment. Ideally, subject matter experts (e.g. toxicologists or biologists) will determine what level of effect is biologically meaningful. Criteria for making that determination can be on the basis of the health of the individual animal or on the ability of the population as a whole to thrive. For example, it may be the scientific judgment of biologists that a 10% change in body weight of a Sprague-Dawley rat, a 3% change in the length of a Daphnia magna, and only a 300% or greater increase in vitellogenin (VTG) are of biological importance. This is not a statistical question but it is very important to the statistician in designing or interpreting a toxicity study to know what size effect it is important to find. Without the information on what size effect it is important to detect, the statistician or data analyst can only determine what is statistically significant or estimate an arbitrary percent effect that may have no inherent value. The result is unsatisfying to the statistician, biologists, and risk assessor.

Ethical concerns about the use of animals in toxicity experiments are increasingly important and the authors share this concern. There is a very active worldwide effort underway to reduce or eliminate the number of animals for various species (mice, fish, birds, etc.) used in toxicity experiments. We will not pursue the question of the desirability of animal testing. Our purpose is to provide scientifically sound methods for analyzing the range of responses that arise from toxicity experiments. Most of these methods apply whether the test subject is a fathead minnow, tomato plant, cell, or bacterium. In all cases, experiments should be designed to use the minimum number of test subjects needed to provide scientifically sound conclusions. This is an instance where ethical and cost considerations coincide.

1.2 REGULATORY CONTEXT FOR TOXICITY EXPERIMENTS

Many toxicity studies are done to meet a regulatory requirement needed to obtain permission to use a chemical that may lead to an environmental exposure. Such toxicity experiments are used by regulatory authorities, such as the United States Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS), United States Environmental Protection Agency (USEPA), Office of Pesticide Programs (OPP), European Food Safety Association (EFSA), the European Chemicals Bureau (ECHA), The Institute for Health and Consumer Protection (IHCP), or one of the European country environmental agencies, including the Danish Environmental Protection Agency (DK-EPA) and Umweltbundesamt (UBA) following standardized test guidelines issued by the Organization for Economic Co-operation and Development (OECD) or the USEPA to assess the likelihood of adverse impacts on populations and communities of organisms in the environment.

To minimize data requirements and avoid unnecessary tests, regulatory risk assessments in the US have a tiered structure. Tier I studies estimate hazard and exposure under "worst-case" conditions. If no adverse effects are found under these conditions, there may be no need for further data. In its simplest form, a so-called limit test may be done with a single very high concentration of the test chemical and a control. In other instances, there may be several exposure levels. In either case, except for determining lethal exposure levels, the emphasis is on testing hypotheses regarding whether an adverse effect exists, but there is no need for a precise quantification of the size effect at each exposure level. If a higher tier test is needed, the focus of such tests is usually on sublethal effects, so it is important for the tier I tests to establish exposure levels that are lethal to a substantial portion of the exposed subjects. Early tier tests tend to be simple in design and may indicate that there is no need for the more detailed information that can come from higher tiered tests. Higher tier tests are designed

either to assess risk under more realistic conditions or to obtain more precise quantification of the exposure–effect relationship.

In the European Union (EU) chemicals expected to enter the environment are mainly regulated by three regulations: (i) REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) covering industrial chemicals, (2) PPPR (Plant Protection Products Regulation) covering pesticides, and (3) BPR (Biocidal Products Regulation) covering biocides. The test information requirements in REACH are driven by tonnage, i.e. the yearly volume produced in or imported to the EU. Test requirements start when more than 1 ton of a chemical is produced or imported yearly. The most test requirements are applied to chemicals exceeding 1000 ton year^{-1}.

Chapters 2–10 and 13 will develop methods appropriate for all levels of this tiered process. Much more information on the regulatory process will be provided in Chapter 11. Chapter 12 will develop an important tool for combining the information from individual studies into a single summary distribution useful for risk assessment. References that can be explored now and returned to throughout a course based on this text include http://www.epa.gov/pesticides/biopesticides/pips/non-target-arthropods.pdf, http://www.epa.gov/oppefed1/ecorisk_ders/toera_analysis_eco.htm, http://www.epa.gov/pesticides/health/reducing.htm, and http://www.eea.europa.eu/publications/GH-07-97-595-EN-C2/riskindex.html.

1.3 EXPERIMENTAL DESIGN BASICS

While observational studies of animals or plants captured in the wild are valuable to environmental impact studies, such studies can be quite frustrating in that routes and conditions of exposure are often unknown, sample sizes are often inadequate, and measurements are all too often non-standardized, so that comparisons among studies are very difficult. This book is not concerned with observational studies, even though one of the authors has been very actively involved in several such studies, including one major study lasting for more than 12 years. We will restrict ourselves to designed experiments.

Considerations of study objectives should include what and how measurements will be taken to address the objectives. For a study of fish, for example, how is death to be determined? It may be difficult to know with certainty whether a fish floating upside down at the top of the tank is dead or just immobile. How long should it be allowed to float before deciding it is dead or near death and should be euthanized to prevent suffering? If a fish or plant is weighed, is it weighed wet or first blotted dry or desiccated? Specific protocols should be provided to address such questions.

Experiments intended for regulatory submissions of new pharmaceuticals or crop protection products or food stuffs will receive special attention in this book. In studies done to meet regulatory requirements, objectives are generally very detailed in test guidelines that must be followed. What is often unclear in test guidelines is the size of effect it is important to detect or estimate. Guidelines, especially older guidelines, simply refer to effects that are statistically significant. As a result, it has often been argued, with some merit, that such guidelines reward poor experimentation, since the more variable the data, the less likely an observed effect will be found statistically significant. A good study should state explicitly what size effect is important to detect or estimate for each measured response and the power to detect that size effect or the maximum acceptable uncertainty for that estimate in the proposed study. Detailed discussion of statistical power is introduced in Chapter 2 and discussed in detail in Chapters 3, 5, 6, 8, and 9 in the context of specific tests. There has been increasing interest in the last 15 years or so in replacing the use of hypothesis tests to determine a NOEC by regression models to estimate a specific percent effects concentration, ECx. One goal of the regression approach is to replace the ill-defined connection between biological and statistical significance with an estimate of the exposure level that produces an effect of a specific size. Such methods are introduced in Chapter 2 and explored in depth in Chapters 4, 6, 7, and 8. A hypothesis testing method with the same goal is discussed in Chapter 13.

The basic toxicity experiment has a negative control, where subjects are not exposed to the test substance, and one or more treatment groups. Treatment groups differ only in the amount of the test substance to which the subjects are exposed, with all other conditions as nearly equal as possible. For example, treatment groups might be tanks of fish exposed to different concentrations of the test substance, or pots or rows of plants exposed to different application rates of the test chemical, or cages of mice with different amounts of the test substance administered by gavage. Apart from the amount of chemical exposure, the same species, strain, age, sex, ambient conditions, and diets should be the same in all treatment groups and control.

1.3.1 Multiple Controls

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 in an aquatic experiment, a water spray in a

pesticide application experiment, and unadulterated food in a feeding study), while the other group receives the dilution water with added solvent but no test chemical, a spray with surfactant but no test chemical, or an oral gavage with corn oil but no test substance. 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 and maximum concentrations) are discussed in the relevant guideline documents for a specific ecotoxicity test. In addition, regulatory guidelines must be followed by both controls with regard to the range of acceptable values (e.g. minimum acceptable percent survival or mean oyster shell deposition rate). Multiple control groups can be utilized regardless of whether the experiment was intended for hypothesis testing or regression analysis.

In rodent studies where the chemical is administered by oral gavage using a corn oil vehicle (or some other vehicle), one control group should be given just the corn oil by gavage. The intention is to rule out a gavage effect or separate it from any effect from the test chemical. Not all such rodent experiments include a control group that is simply fed a standard diet with no gavage administered. The statistical treatment of multiple controls will be addressed in Chapter 2 and in specific types of analyses in later chapters.

In some experiments, a positive control group is also used. Here a different compound known to have an effect is given to one group of subjects. The purpose is to demonstrate that the experimental design and statistical test method are adequate to find an effect if one is present. If the positive control is not found to be significantly different from the control, the experiment will generally have to be repeated. More information on how to analyze experiments with a positive control group will be given in subsequent chapters. There are other ways to demonstrate the sensitivity of the design and analysis method, including power analysis and computer modeling. These topics will also be addressed later.

1.3.2 Replication

In almost all toxicity experiments, each treatment group and control is replicated, so that there are multiple subjects exposed to each treatment. The need for replication arises from the inherent variability in measurements on living creatures. Two animals or plants exposed to the same chemical need not have the same sensitivity to that chemical, so replication is needed to separate the inherent variability among subjects from the effects, if any, of the test substance. The number of replicates and the number of subjects per replicate influence the power in hypothesis testing and the confidence limits of parameter estimates and other model evaluation measures in regression models and will be discussed in depth in later chapters.

It is important to understand what constitutes a replicate and the requirements of statistical methods that will be used to analyze the data from an experiment. A replicate, or experimental unit, is the basic unit of organization of test subjects that have the same ambient conditions and exposure to the test substance. To paraphrase Hurlbert (1984), different replicates are capable of receiving different treatments and the assignment of treatments to replicates can be randomized. The ideal is that each replicate should capture all the sources of variability in the experiment other than the level of chemical exposure. Two plants in the same pot will not be considered replicates, since they will receive the same application of the test chemical and water and sunlight and other ambient conditions at the same time and in the same manner. Different pots of plants in different locations in the greenhouse will generally be considered replicates if they receive water, test compound, and the like through different means, for example, by moving the applicator and water hose. If 25 fish are housed together in a single tank and the chemical exposure is through the concentration in the water in that tank and the ambient conditions and chemical exposure in that tank are set up uniquely for that tank, then the tank constitutes one replicate, not 25. Furthermore, if two tanks sit in the same bath and receive chemical from a simple splitter attached to a single reservoir of the test substance so that the chemical exposure levels in the two tanks are the same and do not capture all the sources of variability in setting up an exposure scenario, then the two tanks are not true replicates.

Hurlbert (1984) describes at some length the notion of pseudoreplication, "defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent. In ANOVA terminology, it is the testing for treatment effects with an error term inappropriate to the hypothesis being considered." Hurlbert defines the rather colorful term nondemonic intrusion as "the impingement of chance events on an experiment in progress" and considers interspersion of treatments as an essential ingredient in good experimental design. Oksanen (2004) extends the idea of spatial interspersion to interspersion along all potentially relevant environmental axes so that nondemonic intrusions cannot contribute to the apparent treatment effects. The primary requirements of good experimental design, according to Hurlbert, are replication, randomization, interspersion of treatments, and concomitant observations. Many designed experiments fail to meet these ideals to some degree. For example, in an aquatic experiment, tanks of subjects in the same nominal treatment group may receive their chemical concentrations from a common source through a physical

splitter arrangement. Rodents may be housed throughout a chronic study in the same rack. The latter is usually compensated for by the rack frame that rotates positions of the racks to equalize air flow, light, room temperature variations, and other ambient conditions across the experiment as a whole. Furthermore, it is sometimes impossible to make concomitant measurements on all subjects in a large experiment, so that a staggered experimental design may be necessary in which subjects are measured at equivalent times relative to their exposure. For Oksanen (2004), "the proper interpretation of an experiment of a demonstrated contrast between two statistical populations hinges on the opinion of scientists concerning the plausibility of different putative causes." Oksanen (2001, 2004) would accept the results of an experiment if the scientific judgment was that the observed treatment effects could not be plausibly explained by the shortcomings of the experimental design, even if it was possible to *imagine* some form of nondemonic intrusion (Hurlbert, 2004) that could account for the observed effect. However, it must be stated that true replication, randomization, concomitant observation, and interspersion of treatments is the goal.

In some toxicity experiments, subjects are individually housed, such as one bird per cage, one daphnid per beaker, or one plant per pot. In these experiments, the replicate is usually the test vessel, which is the same as the subject, unless there are larger restrictions on clusters of vessels, such as the position in the lab. In other experiments, multiple subjects are housed together in the same cage or vessel and there are also multiple vessels per treatment. In these latter experiments, the replicate or experimental unit is the test vessel, not the individual subject.

In a well-designed study, one should investigate the trade-off between the number of replicates per treatment and the number of subjects per replicate. Decisions on the number of subjects per subgroup and number of subgroups per group should be based on power calculations, or in the case of regression modeling, sensitivity analyses, using historical control data to estimate the relative magnitude of within- and among-subgroup variation and correlation. If there are no subgroups (i.e. replicates), 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 to be described to subgroup means. Thus, a minimum of two subgroups per concentration is recommended; three subgroups are much better than two; and four subgroups are better than three. The improvement in modeling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modeling is improved if we get better estimates of both among- and within-subgroup variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance will have, at least approximately, a chi-squared distribution. The quality of a variance estimate can be measured by the width of its confidence interval and a look at a chi-squared table will verify the statements made.)

The number of subgroups per concentration and subjects per subgroup should be chosen to provide adequate power to detect an effect of magnitude judged important to detect or to yield a slope or ECx estimate with acceptably tight confidence bounds. These determinations should be based on historical control data for the species and endpoint being studied. There are two areas of general guidance. If the variance between subjects greatly exceeds the variance between replicates, then greater power or sensitivity is usually gained by increasing the number of subjects per replicate, even at the expense of reducing the number of replicates, but almost never less than two per treatment. Otherwise, greater power or sensitivity generally comes from increasing the number of replicates and reducing the number of subjects per replicate. This claim will be developed more fully in the context of specific types of data in Chapter 3. The second generality is that for hypothesis testing (NOEC determination), generally there need to be more replicates per treatment and fewer treatments, whereas with regression analysis, it is generally better to have more treatments, and there is less need for replicates. As will be illustrated in Chapter 4, the quality of regression estimates is affected by the number of replicates unless there are a large number of treatments.

Since the control group is used in every comparison of treatment to control, it is advisable to consider allocating more subjects to the control group than to the treatment groups in order to optimize power for a given total number of subjects and thoroughly base the control against which all estimates or comparisons are to be made. The optimum allocation depends on the statistical method to be used. A widely used allocation rule for hypothesis testing 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 positive 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 $n_0 = N - kn$. (It is almost equivalent to say $n_0 = n\sqrt{k}$.) Dunnett showed this to optimize power of his test. It is used, often without formal justification, for other pairwise tests, such as the Mann–Whitney and Fisher exact test. Williams (1972) showed that the square-root rule may be somewhat suboptimal 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 square-root allocation rule will be explored in more detail in Chapter 2 and in subsequent chapters in the context of specific tests or regression models.

1.3.3 Choice and Spacing of Test Concentrations/Doses

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

The choice of test substance concentrations or doses or rates is one aspect of experimental design that must be evaluated for each individual study. The goal is to bracket the concentration/dose/rate[1] at which biologically important effects appear and to space the levels of the test compound as closely as practical. If limited information on the toxicity of a test material is available, exposure levels 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. Effects are usually expected to increase approximately in proportion to the log of concentration, so concentrations are generally 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 limited.

Of course, the idea of bracketing the concentration/dose/rate at which biologically important effects appear is much simpler to state than to execute, for if we knew what that concentration was, there would no longer be a need to conduct an experiment to determine what it is. To that end, it is common to do experiments in stages. Conceptually, a small range-finding study is done to give an idea of the exposure levels likely to produce effects of interest. Based on that, a larger definitive study is done. Experience indicates that this process is not fail proof, so exposure levels generally start well below the expected level and extend well beyond. There are practical issues as well. If concentration levels are too small, analytical chemistry methods may not be sufficiently sensitive to measure these levels and it sometimes happens that there is an inversion, where some mean measured concentrations are in reverse order to the planned nominal concentrations. This complicates the interpretation of results and brings into question the entire experiment. Another issue is involved in the intended use of the test substance. For example, a pesticide or pharmaceutical product has to be administered at a high-enough level to be effective. Testing much below the effective level may only make sense if the concern is of environmental exposure that might arise from dilution in a stream or from rainfall.

At the other extreme, in aquatic experiments, chemicals have a solubility limit that cannot be exceeded and this obviously restricts the range of exposure levels that can be included. In all types of studies aimed at determining sublethal effects, the exposure levels must be below the level that produces high mortality. Generally, separate studies are done to determine lethality and that information is used in both the range finder and definitive tests for sublethal effects.

1 To avoid repeated awkward phrases such as concentration/dose/rate, the text will frequently use only one of these terms, usually concentration when the context clearly requires an aquatic environment, but commonly dose regardless of context. The terms will be used interchangeably in this text except in rare instances that are clear from context.

1.3.4 Randomization

Variability (often called noise) is inherent in any biological dataset. The following factors affect the level of noise in an experiment:

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

Randomization is used in designing 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 will be 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.

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 and modifications are made to the model to account for this restriction on randomization.

1.3.5 Species Used for Experiments

Many species are used in ecotoxicity experiments. In aquatic toxicology, rainbow trout, fathead minnow, zebrafish, Japanese medaka, sheepshead minnow, and silverside are common fish species tested. In addition, Daphnia magna, various species of algae, macrophytes, lemna, and sediment dwelling chironomid and endobenthic species round out the common aquatic species used in laboratory experiments. To the aquatic species must be added earthworms, honeybees, mites, numerous avian species, and many non-target crop plants and wild plant species. Mammalian toxicity studies are most often done on some rat and mouse species, plus rabbits and guinea pigs. Other species are also used. With rare exceptions, humans are not test subjects for toxicity experiments. Humans are, of course, used to test pharmaceuticals in clinical trials and sublethal toxic effects may be observed in these trials. Clinical trials are not considered in this text.

Some statistical methods apply across species. There is no widely accepted specific fish statistical test or model used in toxicity studies. There is, on the other hand, much variety in the types of responses that arise and while these are sometimes linked to specific species, it is the nature of the response that determines the statistical method to be used, not the species per se.

1.3.6 Extrapolation to Human Toxicity

Given that humans are not subjects for experiments, but are exposed to various chemicals in the course of work or in food consumption, wearing apparel, and use of home products, the risk assessor needs some mechanism for extrapolating from animal studies to human exposure. It is not the purpose of this text to explore the ways in which such extrapolations are done, other than to indicate that generally this involves some uncertainty factor to apply to the animal studies. For example, the lowest level found to have a toxic effect on a rodent may be divided by 100 or 1000 in assessing the safe level of human exposure. Further discussion of such extrapolations and human risk assessment can be found in Brock et al. (2014), Vose (2000), Warren-Hicks and Moore (1998), and Hubert (1996, pp. 401ff).

1.4 HIERARCHY OF MODELS FOR SIMPLE TOXICITY EXPERIMENTS

There is a model underlying every statistical test used to derive a NOEC or estimate an ECx. A basic experimental design in ecotoxicity is one in which independent groups of subjects of common species, age, and sex are exposed to varying concentrations of a single test chemical for the same length of time, so that the only non-random source of difference among these subjects is the level of chemical exposure. It is expected for most species, chemicals, and responses to be analyzed that if there is an effect of the chemical it will tend to increase as the chemical concentration increases. The basic statistical model for this simple toxicity experiment is given by

$$Y_{ij} = \mu_i + e_{ij,} \tag{1.1}$$

where μ_i is the expected mean response in the ith concentration, and e_{ij} are independent identically distributed random errors, often assumed to be normally distributed with homogeneous variances, though that is not by any means an absolute requirement. What distinguishes one model from another is what additional restrictions or assumptions are placed on the treatment means, μ_i.

The simplest model that is used for hypothesis testing is usually stated in terms of null and alternative hypotheses as

$$\mathrm{H}_{02}: \mu_0 = \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs. } \mathrm{H}_{a2}: \mu_0 \neq \mu_i \text{ for some } i, \tag{1.2}$$

where μ_0 is the control mean.

This model implies no relationship among the treatment means and one merely tests each treatment against the control. In the contest of toxicology, Model (1.2) ignores the expected relationship between the exposure concentration and the response. A more appropriate model is given by

$$\mathrm{H}_{03}: \mu_0 \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_k \text{ vs. } \mathrm{H}_{a3}: \mu_0 > \mu_i \text{ for some } i. \tag{1.3}$$

This model assumes a non-increasing concentration–response, which is what is expected biologically for most responses from ecotoxicity experiments. It should also be evident that if $\mu_0 > \mu_i$ for some i, then $\mu_0 > \mu_j$ for all $j > i$. Of course, the inequalities can be reversed where an increased response is expected with increasing concentrations. A two-sided version of this model is also possible for those situations where the researcher is confident that the subjects will respond to the chemical insult in a monotone dose–response fashion, but they are unsure for the compound and endpoint in question whether the direction of effect will be an increase or a decrease. For either Model (1.2) or (1.3), one would estimate the μ_i from the data.

Do we expect a monotone dose–response? If not, then tests based on monotonicity should not be used and, of course, regression models are likewise inappropriate. An exception exists for the case of hormesis where a specialized model can be employed to capture that phenomenon. Hormesis, or more generally, low-dose stimulation, is the presence of an apparently beneficial effect at low exposure

levels followed by an adverse effect at higher exposure levels. Such effects are observed in many testing laboratories throughout the world and can arise from several causes. Pharmaceuticals are used because they are beneficial at low dosages but they can be toxic at high dosages. Duke et al. (2014) reported that allellochemicals, which are phytotoxins released from plants, are known to induce hormesis and some chemicals stimulate the production of allellochemicals. They speculate that some mechanisms producing hormetic responses could represent physiological attempts to compensate for chemical stress. For example, plants could produce more seeds, thereby increasing the chance of the next generation to germinate under more favorable conditions. In rodent studies, animals are often fed *ad libitum*. Since there is little opportunity for exercise or other physical stimulation, control animals sometimes get grossly overweight, which has health implications for those species as it does for humans. At low doses, the chemical can make the animals slightly ill so their appetite is reduced and they maintain a more healthy body weight and do not suffer the adverse weight-related health effects of the control animals. At higher dose levels, the adverse effects of the test compound overwhelm any benefit from reduced weight. In an aquatic experiment, the low dose of the test compound may stimulate the growth of algae because of a mild increase in nutrients in the chemical, but higher concentrations inhibit growth because the toxic effect overwhelms the nutrient effect. In a pesticide spray application, a low application rate may inhibit pests without damaging the plants, but at higher applications, plant damage may occur.

Calabrese and Baldwin (2002) have a very interesting discussion of hormesis in which they interpret it an "adaptive response with distinguishing dose-response characteristics that is induced by either direct acting or overcompensation-induced stimulatory processes at low doses. In biological terms, hormesis represents an organismal strategy for optimal resource allocation that ensures homeostasis is maintained." van der Woude et al. (2005) and Carelli and Iavicoli (2002) contain further discussion of hormesis. Lloyd (1987) discusses this in connection with mixtures of chemicals. Dixon and Sprague (1981) and Stebbing (1982) discuss this for a variety of phyla.

Even if we expect a monotone dose–response, it is important to assess the data for consistency with that judgment. There are several reasons for this. 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. While our experience does not substantiate the idea, Davis and Svendsgaard (1990) and others have suggested that departures from monotonicity may be more common than previously thought. These concerns suggest that a need for caution exists. We advocate both formal tests and visual inspection to determine whether there is significant monotonicity or significant departure from monotonicity. Further discussion of this issue will be presented in Chapter 3.

Regression models assume a specific mathematical form for the relationship between treatment mean and concentration, for example, when modeling length or weight of fish from an aquatic experiment or shoot height from a non-target plant study, one might hypothesize

$$\mu_i = ae^{bx_i} \tag{1.4}$$

where x_i is the concentration in the ith treatment, and a and b are positive parameters to be estimated from the data. The essential difference between Model (1.3) and models of the type illustrated by Model (1.4) is the specific mathematical relationship assumed between concentration and response. It is rare in ecotoxicology to have a priori models based on biological principles, so the regression approach becomes an exercise in curve fitting and evaluation. An advantage of Model (1.3) is that there is no need to specify an exact form. An advantage of Model (1.4) is the ability to predict the mean response at a range of chemical concentrations by interpolating between the x_i. The typical goal in a regulatory setting is to estimate the concentration or rate that produces a specific percent change (increase or decrease, as appropriate) in the measured response compared to the control mean response. The label ECx is used for the concentration that produces an x% change from the control and is referred to as the x-percent effects concentration. There are other differences that will be addressed later.

Biologically-based models are mathematical models having the form of regression models, such as Model (1.4), but derived from, or based on, concepts observed in biology that encompass much more than just the concentration of test chemical. The Dynamic Energy Budget (DEB) theory developed by S.A.L.M. Kooijman and colleagues (e.g. Kooijman and Bedaux, 1996) is built on the idea that the hazard rate and the parameters that quantify the energy budget of the individual are proportional to the concentration of the test substance in the animal. DEB theory "specifies the rules that organisms use for the energy uptake of resources (food) and the ensuing allocations to maintenance, growth, development and propagation."

1.5 BIOLOGICAL VS. STATISTICAL SIGNIFICANCE

Statistical analysis of toxicity data, especially the hypothesis testing or NOEC approach, is often criticized for finding treatment means statistically significantly different from the control mean that are not biologically important. On the other hand, in some experiments, differences are found that are thought to be biologically important but are not found

statistically significant. The difference between biological and statistical significance is real and should be appreciated. Moreover, an ideal study should be designed to have high power to find biologically important treatment effects and low likelihood of finding significant a treatment "effect" that is not biologically important. In terms of estimating an *x*% effects concentration, EC*x*, from a regression model, the percent effect, *x*, to be used should be biologically determined, not an arbitrary value selected purely to satisfy some legislation or convenience.

There are several factors that make it difficult or impossible to design such an ideal experiment. First, it is unusual for an experiment to have a single endpoint or response of interest. This means that while it may be possible to design the experiment to be optimal in some sense for one response, the experiment may be decidedly suboptimal for another response. Second, it has proved very difficult to find agreement within the ecotoxicity scientific community to reach agreement on the size effect for a given response that is biologically important to detect, or even the basis for making such a determination. Is it more important to judge biological importance on the health of the individual subject, as is usually the case in assessing the carcinogenic risk in rodent experiments (and by inference, to humans), or on the ability of population as a whole to thrive, as suggested by Iwasaki and Hanson (2014) and Mebane (2012)? Third, when there is, say, 90% power to detect, i.e. found to be significant, a 10% change in some response, there may also be a 50% chance of detecting a 5% change and 25% power to detect a 3% change. In plain language, then, sometimes very small effects will be found statistically significant. If a regression model can be fit to the data to try to eliminate this problem, then it might turn out that the 95% confidence interval for EC10 spans the entire range of tested concentrations including the control. (Examples will be given in Chapter 4.) This hardly seems informative. Data may be highly variable, given practical limitations on the size experiment than can be run and the inherent variability of the subjects and sensitivity of the measuring equipment, and will mean a 50% effect cannot be detected or no model can be fit or the aforementioned wide confidence bounds on EC10 estimates encountered.

One should also question the idea that 10% is some universally applicable size effect. The measurement of VTG in fish to evaluate possible endocrine effects is extremely variable and effects of 1000% or higher increases are observed. There also frequently very high inter-lab and even intra-lab variability in this measurement. The data are continuous but by no means normally distributed or homogeneous in variance. For hypothesis testing purposes, a log-transform or even a rank-order transform might be used to deal with the huge spread in the data. Regression models, even on log-transformed responses, are often very poor and generate very wide confidence bounds. It is totally pointless to estimate EC10 with such data, so what size effect should be estimated? Furthermore, do we estimate an *x*% effect based on the untransformed control mean or on the log-transform? If the former, the model-fitting algorithm will usually not converge. If the latter, the meaning of EC*x* will vary from experiment to experiment in a much bigger way than with more well-behaved data. For example, if the mean control response is 10, 100, 1000, or 10 000, then a 10% increase in the logarithm corresponds to a 26, 58, 100, or 151% effect in the untransformed values.

Isnard et al. (2001) note the problem of determining an appropriate choice of *x*, noting "Biological arguments are scarce to help in defining a negligible level of effect x for the ECx." An insightful presentation by C. Mebane of the US Geological Survey/NOAA Fisheries Liaison at the SETAC conference in Long Beach, California in 2012 (Mebane, 2012) reported on a review of early-life stage toxicity testing with aquatic organisms (reduced growth, fecundity, and survival) in the context of responses of wild fish populations to disturbances associated with changes in mortality or growth rates. The review suggested that different EC*x* values would be appropriate for different endpoints and for species with different life histories. Under some conditions and for some species, differences in length of as little as 5% can disproportionately determine survival. Growth reductions of the magnitude of 20% could predict extremely high indirect mortalities for juvenile fish. For some other populations, where juveniles have to compete for limited shelter to survive their first winter, much greater than 20% loss of the young-of-year is routinely absorbed. These observations suggested to Mebane that an EC20 for fecundity or first-year survival in density-dependent fish populations would conceptually be sustainable, yet for reduced growth (as length of juvenile fish) an EC5 would be a more appropriate endpoint.

These findings and others suggest a real need to identify the size effects of biological importance across a wide spectrum of measured endpoints. In any event, to design properly an experiment, it is critical to know what size effects are of interest for each response to be analyzed. This must be tempered by the understanding mentioned above that an experiment well designed for one response may be over- or under-sensitive for other responses. One practical workaround is to design around the most important responses. It is also important to maintain a working historical control database within the testing laboratory for each type of study, species, and response, and then interpret the results of each new study in light of the distribution of control responses. This can be done informally or through Bayesian methodology applied to incorporate such information formally in the analysis.

Before leaving this topic, it is appropriate to discuss how statistical significance is decided and to appreciate a