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

09-May-2006

English - Or. English

ENV/JM/MONO(2006)18
Unclassified

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

OECD SERIES ON TESTING AND ASSESSMENT
Number 54

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

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

JT03208537

English - Or. English

OECD Environment Health and Safety Publications

Series on Testing and Assessment

No. 54

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

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2006

PBN1446

TABLE OF CONTENTS

5. HYPOTHESIS TESTING

5.1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

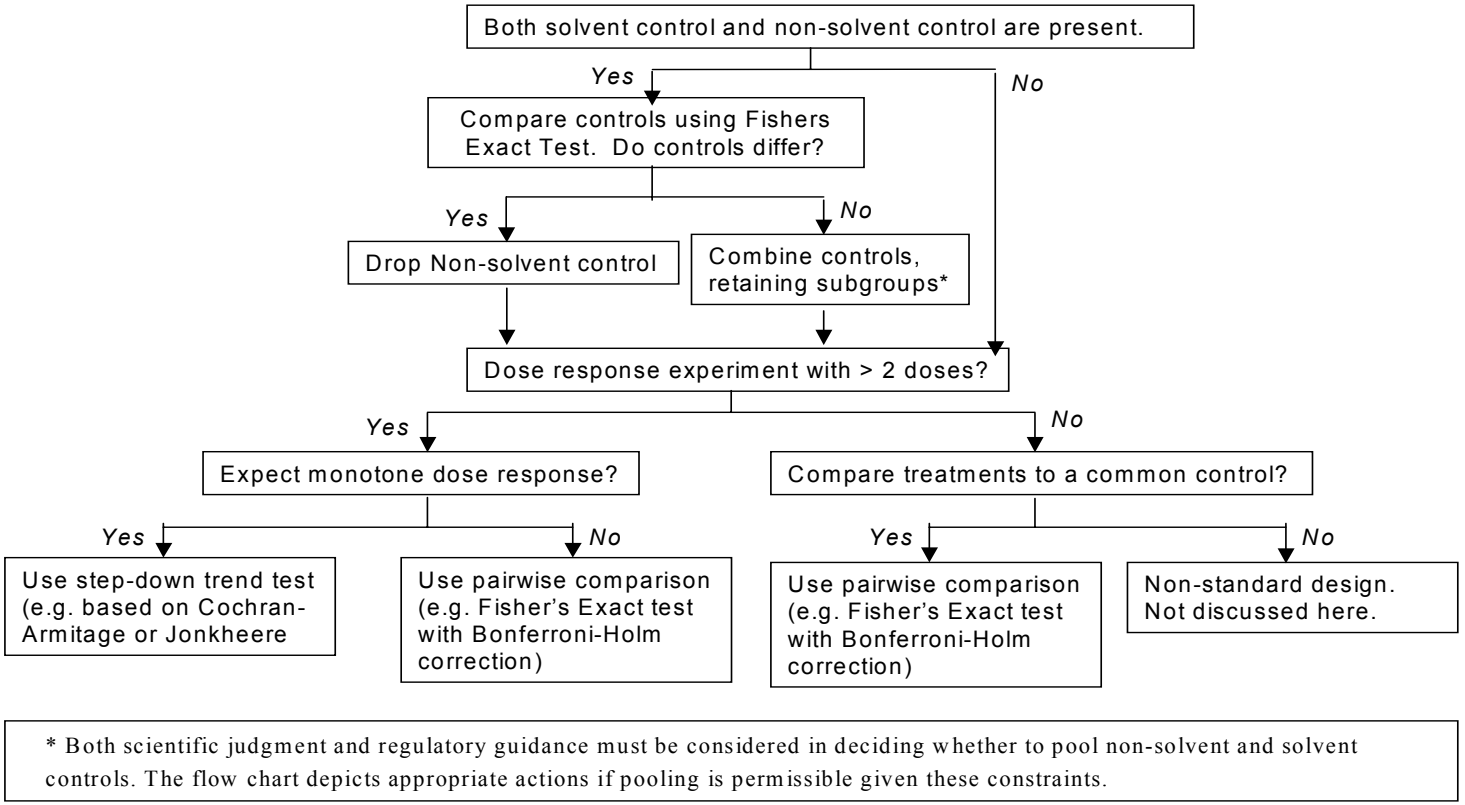

Figure 5.1. Analysis of Quantal Data: Methods for determining the NOEC. Note that the dose count in '>2' includes the control.

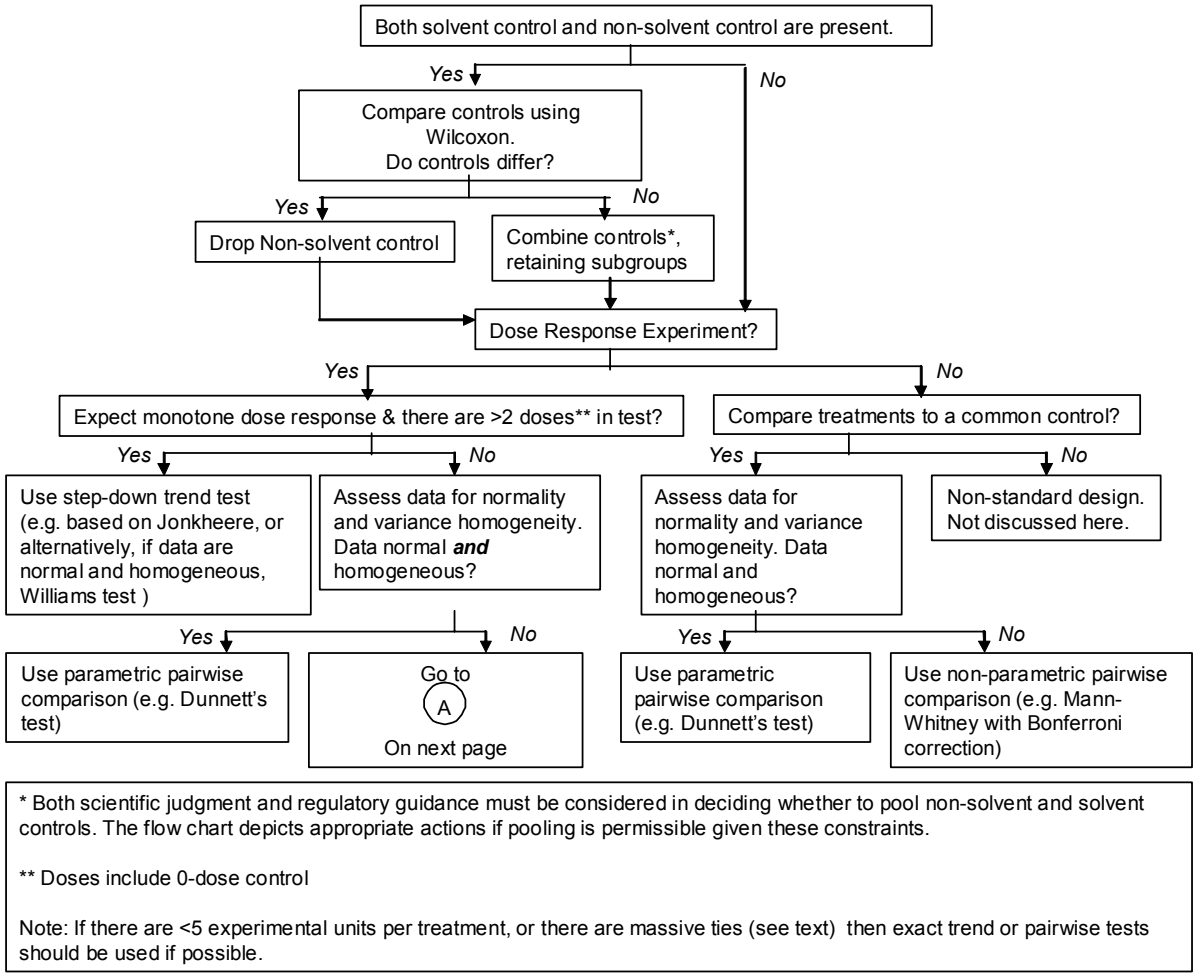

Figure 5.2. Analysis of Continuous Data: Methods for determining the NOEC

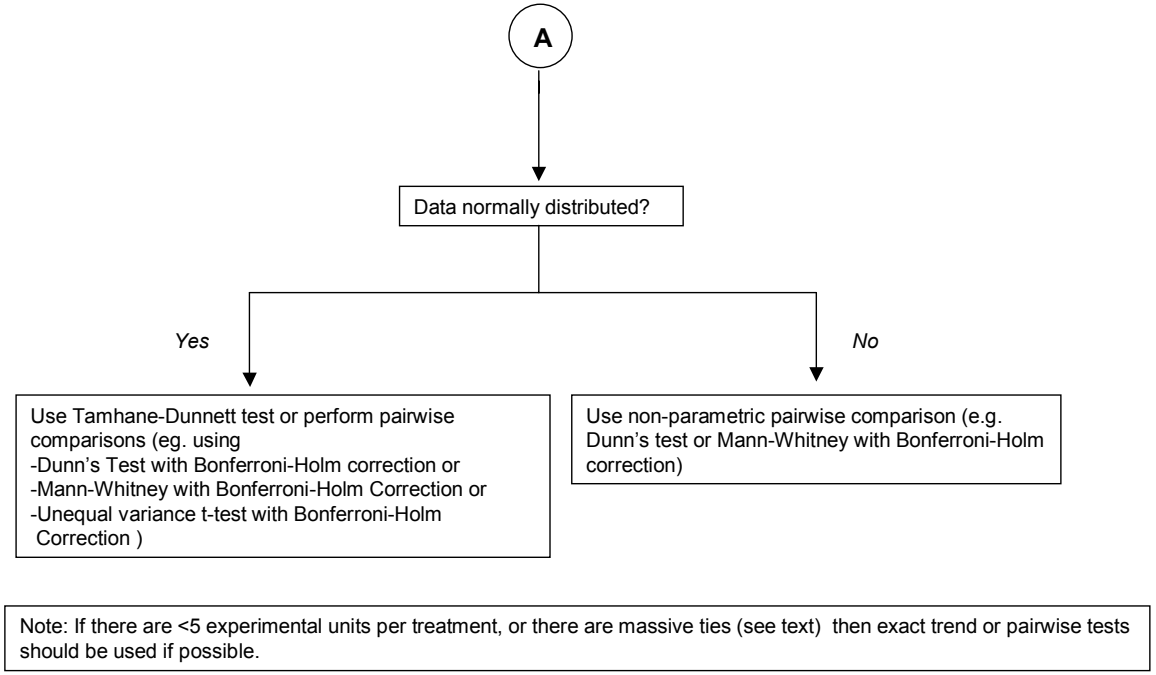

Figure 5.3. Analysis of Continuous Data: Methods for determining the NOEC.

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

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

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

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

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

5.1.2. Hypothesis Used to determine NOEC

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

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

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

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

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

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

Two-sided Trend Test

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

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

Trend or Pairwise test

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

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

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

The null and alternative hypotheses can then be stated as

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

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

123. Under the trend model and tests designed for that model, if tests of hypotheses H_{02} vs. H_{12} reveal that μ_3 is different from μ_0, but μ_2 is not, the NOEC has been determined (i.e. it is the test concentration associated with 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 trend is not appropriate do specific pairwise comparisons make sense to illuminate sources of variability.

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

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

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

or

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

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

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

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

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

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

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

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

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

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

Table 5.1 Comparison of Adjusted and Unadjusted P-Values

133. Alternatives based on the Sidak inequality (each comparison at level $1-(1-\alpha)^k$) 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.

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

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

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

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

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

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

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

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

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

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

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

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

5.1.4. Dose metric in trend tests

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

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

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

5.1.5. The Role of Power in Toxicity Experiments

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

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

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

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

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

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

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

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

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

5.1.6. Experimental design

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

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

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

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

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

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

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

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

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

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

5.1.7. Treatment of Covariates and Other Adjustments to Analysis

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

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

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

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

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

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

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

5.2.1. Hypothesis testing with quantal data to determining NOEC values

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

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

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

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

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

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

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

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

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

5.2.2. Parametric versus non-parametric tests

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

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

5.2.2.1. Single-step procedures

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

5.2.2.2. Step-Down Procedures

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

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

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

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

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

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

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

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

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

5.2.2.3. Alternative Procedures

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

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

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

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

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

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

5.2.2.4. Assumptions of methods for determining NOEC values

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

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

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

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

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

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

Treatment of multiple controls

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

5.2.3. Additional Information

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

5.2.4. Statistical Items to be Included in the Study Report

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

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

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

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

5.3.1. Hypothesis testing with continuous data to determine NOEC

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

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

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

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

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

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

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

5.3.1.1. Parametric versus non-parametric tests

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

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

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

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

5.3.1.2. Single-step (pairwise) procedures

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

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

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

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

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

5.3.1.3. Step-down trend procedures

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

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

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

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

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

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

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

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

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

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

5.3.1.5. Assumptions for methods for determining NOEC values

Small Samples / Massive Ties

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

Normality

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

Variance Homogeneity

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

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

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

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

5.3.1.6. Operational considerations for statistical analyses

Treatment of Experimental Units

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

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

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

Identification and Meaning of Outliers

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

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

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

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

Multiple Controls

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

General

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

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

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

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

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

6. DOSE-RESPONSE MODELLING

6.1. Introduction

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

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

$$y = f(x)$$

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

$$y = a + b\,x$$

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