3. **Statistical considerations**

3.1 **Outline**

The statistical methods used to analyse results of regulatory ecotoxicology studies must be consistent with regulatory frameworks, they must be statistically robust and maximize efficiency in terms of animal use, time and costs. Different national and regional risk assessment schemes have often been developed to balance these factors in a variety of ways. For example, in Europe, the environmental hazard assessments of industrial chemicals or biocides focus on the calculation of the predicted no effect concentration (PNEC) values. These are typically based on either acute effects concentration (EC₅) type studies or chronic no observed effect concentration (NOEC) type studies, using safety factors as appropriate (EU 2003). For pesticides, in Europe either acute ECₐ type studies or chronic NOEC studies are used to derive toxicity exposure ratios that are employed for risk assessment (Directive 91/414; EU 1991). In the United States, the regulatory terminology implies the use of chronic NOEC data as a basis for the calculation of hazard quotients used in the risk assessment of pesticides (US EPA 2004). In contrast, chronic exposure in aquatic algae (e.g. OECD TG 201) studies are typically analysed by calculation of an EC₁₀ and EC₂₀, which are often used alongside aquatic fish studies in the assessment of risk. Also, some probabilistic risk assessment schemes might require information on the slope and confidence limits of the dose-response curve.

Regulatory needs, test designs and statistical methods cannot be considered independently, and the impact of change in one of these factors on the others must be considered carefully. For example, in Chapter 4 it is stated that for endocrine screens (OECD TG 229 and OECD TG 230) “the definitive test exposes fish to a suitable range of concentrations maximizing the likelihood of observing the effect. The important distinction being that achieving a NOEC is not the purpose of the test”. This statement accurately reflects the original basis for the design of the test. Yet, as the basis for the original test design fades in memory, there appears to be a tendency to expect the calculation of both EC₅ and NOEC values based on the results of the screens. It cannot be assumed that the design of this test can support adequate estimates of EC₁₀ and EC₂₀ values without adequate studies of the accuracy and precision of estimates. Alternative testing methods may have new or novel strengths/limitations in terms of statistical power compared to standard guidelines (e.g. testing based on upper threshold concept).

Statistical methods must be capable of detecting, or modelling, the smallest effects that are biologically meaningful (for discussion, see below). A key issue in the interpretation of fish toxicity tests, as they grow in complexity, is to distinguish between biologically important effects caused by the test chemical *versus* statistically detectable differences. This aspect of ecotoxicology test guideline data interpretation is identical to the principals developed for many mammalian test guidelines in recent years (Länge et al. 2002 Williams et al. 2007). Against this background, a key element of this chapter is to illustrate key principles of data interpretation (e.g. importance of historical control values for endpoints of interest, adequate replication, etc.) that can be used as required for different fish test guidelines.
Toxicological endpoints should not be interpreted in isolation from other information relevant to the test. For example, it is usually assumed that the responses will follow an underlying monotone concentration response pattern (i.e. there is a general tendency for the effect to increase as concentration increases) in the absence of compelling evidence to assume otherwise. Use of the knowledge that responses are likely to follow such a pattern can lead to better statistical tests and allow variations not related to treatment to be identified. For example, this assumption makes the Jonckheere-Terpstra trend test a powerful tool for the calculation of NOEC values and forms the entire basis for calculation of ECx values.

There is some controversy on the question of whether hypothesis testing (NOEC/LOEC [no/lowest observed effect concentration]) or regression (ECx) is the better way to evaluate toxicity data (e.g. Chapman et al. 1996, Dhaliwal et al. 1997). It is not the intention to replay that debate here. The intention of this chapter is to indicate how best to do each type of analysis and to indicate the types of data and experimental designs under which each type of analysis can be done with reasonable expectation of useful results. Therefore, requirements for the different approaches are considered.

The OECD guidance document no. 54 (OECD 2006) describes current approaches to the statistical analysis of ecotoxicity data and should be consulted. However, the recent development of new fish test guidelines (e.g. draft Fish Sexual Development Test) has raised additional specific issues worthy of discussion in addition to some general considerations.

### 3.1 Biological versus statistical significance

The question of what magnitude of effect is biologically important to detect or what effects concentration (ECx) to determine is not a statistical issue. This issue is not unique to fish tests, but is valid for other ecotoxicity test species such as *Daphnia* and algae. Scientific judgment, grounded in repeated observation of the same response in the same species under the same conditions, is required to specify this (i.e. the understanding of historical control data). Statistics provides a means to determine the magnitude of effect that a given experimental design can quantify. To put this another way, once an effect size of biological importance has been determined by subject matter scientists, it is possible to design an experiment that has a high likelihood of producing the desired information (i.e. whether an effect of the indicated size occurs at some test concentration or what concentration produces the specified effect).

The relationship between biological significance and statistical significance can be understood in terms of the magnitude of effect that can be detected statistically. For a continuous response, such as growth or fecundity, this in turn depends on the relative magnitude of the between-replicate and within-replicate variances. The standard error of the sample control mean response is given by:

\[
SE = \text{SQRT}[\text{Var}(\text{Rep})/r + \text{Var}(\text{ERR})/rn] = \sigma \cdot \text{SQRT}[R/r+1/rn],
\]
where $\sigma$ is the within-replicate or error standard deviation, $R$ is the ratio of the between-replicate variance to the within-replicate variance, $r$ is the number of replicates in the control group and $n$ is the number of fish per replicate, $\text{Var(Rep)}$ is the between-replicate variance and $\text{Var(ERR)}$ is the within-replicate or fish-to-fish variance.

The 95% confidence interval for the mean is, approximately, $\text{Mean} \pm 2 \cdot \text{SE}$. It is often convenient to express $2 \cdot \text{SE}$ as a percent of the mean, so the 95% confidence interval for the mean can be expressed as $\text{Mean} \pm P \%$, where $P = 200 \cdot \text{SE}/\text{Mean}$. The true mean is statistically indistinguishable from any value in this confidence interval for the sample mean.

This means that the smallest treatment effect that can be distinguished statistically is $P \%$ of the control mean. This holds for both the NOEC and EC$_x$ approaches.

It is then incumbent on the study director to determine the magnitude of effect, $Q$, which is judged biologically important. For a given experimental design and endpoint, $Q$ is compared to $P$. If $Q > P$, then the experiment is suited for the purpose, otherwise not.

For example, in a recent fish full life-cycle (FFLC) test, the control mean for percent male offspring was 69%, with standard error of the mean = 19%. Thus, the smallest effect that can be found statistically significant is 19% and EC$_x$ for $x < 19$ cannot be reliably estimated. Another way of considering this is to observe that the lower bound of the 95% confidence interval for EC$_{10}$ and EC$_{20}$ is 0. Also, the NOEC is a concentration at which a > 19% effect was observed. (NOTE: The confidence interval for the difference between the control mean and a treatment mean is actually greater than 19% by a factor of $\sqrt{2}$). Vitellogenin (VTG) is another highly variable response and only large effects, around 40%, can be expected to be statistically significant in a practical experiment. Equivalently, EC$_{40}$ might reasonably be estimated, but not EC$_{25}$.

At the other extreme, the standard error of a growth measurement for *Daphnia* is often 2-3% of the control mean, so very small effects can be found statistically significant. For this response, it is quite feasible to estimate EC$_5$. It is a matter of scientific (non-statistical) judgment whether such small effects are biologically meaningful. Similar findings hold for avian eggshell thickness measurements.

From the formula for standard error (SE), it will be evident that there is a trade-off between the number of fish per replicate and the number of replicates per control or test concentration. For example, from the formula, it is evident that if the number of replicates is doubled and the number of fish per replicate is reduced by 50%, then the second term in brackets is unchanged, but the first term is reduced by 50%. This might indicate a preference for more replicates of fewer subjects per replicate. However, if $\text{Var(REP)}$ is already relatively small, not much is gained by such an approach. Instead, if we cut the number of replicates by 50%, but increase the number of subjects per replicate by a factor of 4, then the first term remains small but the second term is reduced by 50%. Thus, whether it is better to have a few replicates with many subjects in each, or many replicates with few subjects in each, depends on...
the relative magnitude of the two variances. A general rule is that if the ratio $R$ of variances exceeds 0.5, then more emphasis is given to the number of replicates, otherwise, more emphasis is given to the number of subjects within each replicate. For example, shoot height of some emergent crops (e.g. oat, tomato, rape) tend to have variance ratios exceeding 0.5 (John Green, pers. comm.).

As further illustration, recent experiments conducted for the OECD found VTG measurements to be very highly variable and the within-replicate variance ranged between two and 10 times that of the between-replicate variance. Thus, good experimental design would call for relatively few replicates with numerous fish in each. In the case of medaka (*Oryzias latipes*), it was found that a control and three test concentrations with two replicates per control and test concentration and five fish per replicate were adequate to give 80% power to detect a 60% effect. For fathead minnow (*Pimephales promelas*), four replicates of four fish each in the control and each test concentration were required to give 80% power to detect a 94% effect. While test effect sizes may seem large and the number of fish small, there were constraints on the number of replicates and fish that could be accommodated from practical considerations. Furthermore, such size effects were observed in high test concentrations in these experiments during validation.

A complication is when there are multiple endpoints to be analysed in a single experiment. If the experimental design is optimal for one response, it may be sub-optimal for another response. This may mean that only very large effects can be estimated or detected statistically for one endpoint and perhaps very small, biologically unimportant effects will be found statistically significant for another response. It is important to understand this in interpreting the data. For a biologically important effect may be missed in the first instance, which should not be interpreted to mean the chemical in question has no effect on that response; while a sound study might be rejected because of a tiny effect found statistically significant. Any statistical result should be interpreted in light of the biologically important effects determined before the experiment was conducted.

To design an experiment with high likelihood of detecting a P% effect (or estimating a meaningful EC$_P$), it suffices to find $r$ and $n$ so that $P = 200 \cdot \text{SE}/\text{Mean}$. It is a simple matter to construct a table showing 200·SE/Mean for various values of $r$ and $n$ and seeking the most practical combination to define the design based on historical control estimates of the two variances.

### 3.2 NOEC/LOEC

For the purpose of determining an NOEC or an LOEC, it is important to design the experiment so as to be able to have a reasonable chance of finding a biologically relevant effect statistically significant and minimize the chance of finding biologically irrelevant effects statistically significant. These two objectives are somewhat incompatible, and judgment will be useful in reaching appropriate regulatory conclusions. A fish study will have a water control group (dilution water control), and if a solvent is used, a solvent control, and at least one test
concentration. Unless the design is for a limit test, there will usually be three or more test concentrations approximately equally spaced on a log scale. With very few exceptions, the control(s) and test concentrations should be replicated. Replicate here refers to the test vessel, not to individual animals unless they are housed individually in a test vessel. The trade-off between number of fish per replicate and number of replicates per test concentration and control will vary according to the response and species, and a power calculation may be needed to determine the best design. Since multiple responses are usually tested from the same experiment, it will often not be possible to design an experiment that is optimal for all responses. Judgment is needed to decide on the most important response(s) and experiments should be designed to provide adequate power (75 - 80 %) to detect biologically relevant changes in those responses. Power simply refers to the probability of finding statistically significantly an effect of a given true magnitude, taking into account variability in the response of interest and variability arising from sampling. It is also important to quantify the size effect likely to be found significant in all other responses. This may indicate that there is a need to rethink the objectives of the experiment. There should be no surprises at the end of the study about what can be analysed, by what test, and with what ability to detect effects.

Most responses are analysed using 2-sided tests, unless there is a clear biological reason to expect or be concerned only with changes in one direction (e.g. an increase). Furthermore, for most responses, there is an expectation that the concentration-response is approximately monotone. The effect may be measured as an increase or a decrease in some measurement (e.g., weight might decrease, mortality might increase). That being true, a test that is designed for a monotone trend is more powerful than one that simply compares each test group to the control independent of effects in other test groups. Thus, there is a preference for a step-down Williams or Jonckheere-Terpstra test over a Dunnett, Dunn, Mann-Whitney or other pairwise test, provided the data are consistent with a monotone concentration-response. All tests referenced in this chapter are discussed in detail in OECD (2006).

All statistical procedures are based on some data requirements. In addition to the monotonicity requirement for Williams and Jonckheere-Terpstra for continuous responses and Cochran-Armitage for quantal responses, there are additional requirements. The Williams and Dunnett tests require normally distributed data with homogeneous variances. While these tests are robust against mild violations of these requirements, they are not impervious and some checking of these requirements is appropriate. A visual check from a scatter plot may be sufficient to assess monotonicity and variance homogeneity, and even normality. There are also formal tests for all three and OECD (2006) provides details.

Where normality or variance homogeneity are violated, a transformation of the data to achieve these requirements can be sought or non-parametric methods employed, which have fewer requirements or are much less sensitive to violations. Be mindful that different agencies may have different jurisdictions on how, or whether, data will be transformed. Contrary to widely held opinions, non-parametric tests are not always inferior to parametric tests. For example, the power properties of the step-down Jonckheere-Terpstra test are very similar to those of the step-down Williams test, when the data are normally distributed with homogeneous variances, and are superior to Williams, when those conditions are violated. On the other
hand, for datasets with few replicates, the power properties of the Mann-Whitney and Dunn tests are worse, sometimes much worse, that those of Dunnett’s test. Fig. 3.1 indicates a typical comparison of these tests.

Fig. 1 shows the power of seven tests for an experiment with three positive test concentrations and a single control. The horizontal axis shows the percent change from the control, and the vertical axis shows the probability that size effect will be found statistically significant. The red curve shown with diamonds is the step-down Jonckheere-Terpstra (JT) test (standard asymptotic version), orange with triangles for the exact permutation version of JT, the black dark curve with asterisks is William’s test, blue with circles is for Dennett’s, cyan with asterisks and green with squares are for the standard (asymptotic) and exact permutation versions of Mann-Whitney (also known as the Wilcoxon) test, and grey dots for Dunn’s test. In the top row, the design called for 2 reps of 8 fish in the control and each test concentration, while the bottom row is for two replicates of two fish each. The left column is for a design following the square-root allocation rule (see below), and the right column is for a design with equal replication in control and all treatment groups. On the left, the gray Dunn power curve is hidden by the green and cyan Mann-Whitney curves. The data generated were normally distributed with homogeneous variances and with variability that was observed for VTG in some OECD validation experiments for fish endocrine screening studies. It is clear that the power of the Jonckheere test is greater than that of Williams test on the left, whereas William’s test sometimes has slightly greater power on the right and both tests exceed in power that of all the pair-wise tests (Dunnett, Dunn, Mann-Whitney). A striking feature of these results is that Mann-Whitney has zero power to detect effects regardless of magnitude in either design, whereas Dunn’s has zero power under the square-root rule and low power under equal allocation. This knowledge is clearly important in deciding on design and test selection.

3.3 ECx

Standard regression models also depend on data meeting certain requirements. Among these requirements, a key prerequisite is that the observations are mutually independent. This requirement is violated, for example, if all responses are divided by the control mean in an attempt to “normalize” the data or reduce variability. While it is not impossible to model correlated responses, specialized models are required to do so. For a continuous response (i.e. proportion of males is analysed as a continuous response), the data are assumed to be normally distributed with homogeneous variances. There are modifications to accommodate heterogeneous variances, such as alternative variance-covariance structures or weighting. It is also possible to accommodate some types of non-normality.
Fig. 3.1: Power of various tests to detect effects

**POWER OF TESTS TO DETECT A VTG EFFECT AT DOSE 3**
From Fish Experiment with Control and 3 Positive Doses
CVTYPE=MAX REPS=2 SAMPLSIZE=8 SQRTRULE=YES

**POWER OF TESTS TO DETECT A VTG EFFECT AT DOSE 3**
From Fish Experiment with Control and 3 Positive Doses
CVTYPE=MAX REPS=2 SAMPLSIZE=2 SQRTRULE=NO
It is recognized that regression is robust against mild deviations from normality and variance homogeneity, but it can be adversely affected by serious violations of these requirements. Thus, some checking of the distribution of responses is in order, either visually from a scatter plot or through formal testing. For quantal data, normality is not required, but typically the data are assumed to follow a binary distribution within a given treatment group. Quantal data should be checked for extra-binomial variance (more variation than can be accounted for by the simple binomial distribution), the quantal analogue of the homogeneous variance requirement for continuous responses. If extra-binomial variance is observed, there are statistical test methods which take this into account (e.g. Rao-Scott, as described in OECD 2006). Finally, attention should be paid to goodness-of-fit of the model to the data. There are several ways to assess goodness-of-fit. Among the simplest are (1) visual comparisons of the responses predicted by the model to the observed responses, and (2), where replicates are available, comparing the residual mean square from the model against the pure error mean square. If these residual mean squares are significantly larger than the pure error mean square, then the model does not fit the data well. With small datasets typical in this field, this may not be a powerful test. (3) Confidence bounds on the model predictions are very important to show whether the model predictions have any meaning. If no confidence bounds can be produced or they are very wide, then predictions from the model are scientifically unreliable. It should also be understood that typical confidence bounds do not capture model uncertainty, which is one reason for conducting other goodness-of-fit assessments such as items (1) and (2). Model confidence bounds are constructed based on the assumption that the model is correct. If no confidence bounds can be computed or they are very wide, then the model is not internally consistent regardless of how well the model appears to fit the data from a visual inspection. It is also possible to compute an R-square value to judge the proportion of the total sum of squares that is accounted for by the model. While R-square is a useful measure for linear models, it is an unreliable guide for the non-linear models which are most often used to model ecotoxicity responses. For comparing two models for the same data, Akaike’s AIC criteria can be useful.

In more general terms, a search of OECD TGs 204, 210, 212, 215, 229 and 230 was made to determine procedures specified for ECx calculation. Only OECD TGs 212 and 215 describe ECx procedures. OECD TG 212 describes a normalization procedure, but does not specify fitting a regression curve to the normalized percentages from which to estimate ECx. In contrast, OECD TG 215 describes two test designs: one for ECx and one for NOEC determination. Acknowledging the necessary differences: “that a design which is optimal (makes best use of resources) for use with one method of statistical analysis is not necessarily optimal for another. The recommended design for the estimation of a LOEC/NOEC would not therefore be the same as that recommended for analysis by regression.”

If ECx designs are to be more widely used, consideration of the following will be necessary prerequisites:
• General guidance should be given on how the need to estimate EC$_x$ affects the optimum spacing of concentrations, number of treatments, and number of replicates to be used. This guidance will probably suggest test designs quite different for the minimum designs currently described in the OECD TGs optimized for NOEC determination.

• Different endpoints may elicit responses at very different concentration levels. Therefore, a strategy is required to handle this within one test.

• What constitutes a meaningful EC$_x$ estimate, and what is the implication, if a meaningful estimate cannot be obtained for one or all endpoints?

Validation of the draft Fish Sexual Development Test raised particular concerns about the use of regression analysis for determination of effects on sex ratio. These issues are explored in detail in Appendix 1. From these discussions, it is clear that the EC$_x$ approach is inappropriate for the sex ratio endpoint and a NOEC design should be carried forward.

3.4 NOEC versus EC$_x$ designs

It is stated in OECD (2008) that: “[92.] In summary, ANOVA designs for fish testing appear inferior to regression designs, and the latter are considered to show more promise for fish life-cycle tests given the generally large inherent variability in egg production (fecundity) between individuals, which inevitably reduces the power of the ANOVA approach. Final decisions on which design strategy to use should be made on a case-by-case basis, taking into account factors such as the known variability in reproductive output of the species in question.”

The example datasets and analyses discussed in Chapter 3.x (Appendix) should serve as a caution against overstating the advantages of regression over what is referred to as the ANOVA approach. While regression has always been an important tool for statisticians, it is not appropriate for some datasets and can suggest a level of precision that is not supported by the data. Experiments for regression analysis call for different experimental designs than those for which ANOVA methods are intended. Just as ANOVA methods call for designs with adequate power to detect biologically relevant adverse effects, regression methods call for designs that are capable of providing reliable or meaningful estimates of an x % effects concentration and this requires designing around the specific x or percent effect to be estimated. Basic requirements include the following: (1) There should be test concentrations on both sides of EC$_x$. The zero concentration control does not figure into this requirement. However, see the following paragraph. (2) If the 95 % confidence interval for the control response is of the form Mean ± P %, then estimates of EC$_x$ are meaningful only for x > P. For example, if the control mean is estimated only with 20 % error, then it is meaningless to estimate EC$_{10}$. (3) If the confidence interval for EC$_x$ is very wide, perhaps spanning several test concentrations, then there can be little or no confidence in the EC$_x$ estimate. These basic requirements are important to keep in mind, because once a re-
gression model is fit to the data, it is a simple mathematical exercise to use the resulting equation
to estimate \( EC_x \) for any percent \( x \), and yet not all such values of \( x \) lead to plausible or meaningful
estimates. A mathematical equation is not a substitute for valid interpretation of data. This is akin
to the requirement in the NOEC approach of adequate power to detect an effect of magnitude
deemed biologically important.

It can be appropriate to determine both an NOEC (provided there is sufficient power to detect
biologically relevant effects) and an \( EC_x \) (provided \( x \) is beyond the range of control variability).
Ideally, \( x \) should be between two tested concentrations. However, it is important to recognise
extrapolation beyond the range of data adds significant uncertainty and needs to be justified. It is
permissible to extrapolate modestly beyond the range of tested concentrations, provided this does
not violate restriction (2) in the previous paragraph. Such extrapolation necessarily comes with
increased uncertainty and assumes the model fit is valid beyond the range of tested concentra-
tions, something that is untestable from the data. The uncertainty increases the further from the
experimental range one extrapolates.

Although it is not recommended to combine NOEC and ECx approaches in the same study, there
may be some compelling reasons to do so. For certain existing regulatory frameworks, it might
be appropriate to focus on NOEC test designs for fish chronic endpoints (e.g. FIFRA: US EPA
1996). However, for future regulatory frameworks it could be required to have both \( EC_x \) and
NOEC determinations in fish chronic studies (e.g. draft Sanco 2010 review document). However,
the latter has serious implications for experimental design, time and costs, as well as ethical and
statistical interpretations. It might not be practical to design tests with multiple endpoints to de-
terminate both the NOEC and ECx values for the endpoints of interest.

3.5 Alternate designs (e.g. square root allocation rule)

There are several factors that affect the power of a given test. These are experimental design (e.g.
number of replicates per control and treatment group, number of fish per replicate, number of
treatment groups), shape of the concentration-response, and inherent variability of the response
of interest. One simple, but important decision is whether the control and treatment groups
should be equally replicated or whether more replicates should be allocated to the control. The
argument for the latter is two-fold: First, it gives a better measure of the undisturbed population
against which all treatment groups are compared, and, second, it tends to increase the power of
the test, in part by increasing the degrees of freedom for the test statistics.

Dunnett (1955) showed that the power of his test is optimized using what is called the square-
root allocation rule, which provides a specific formula for the number of replicates in the control
and all treatment groups. Details are given in OECD (2006). Further theoretical published work
and extensive power simulation studies have shown that this same rule (or a minor modification)
also maximizes the power of the Williams and Jonckheere tests and usually increases the power of the Mann-Whitney and Dunn tests.

### 3.6 Solvent/carrier control

One of the issues brought up in the first version of the OECD Fish Sexual Development Test Review, and which is a consideration on all of the test guidelines, is how the two controls (dilution water and solvent controls) should be used when there is a solvent used in the treatment groups. There are advantages to pooling the two controls to test for treatment effects: (1) By doubling the number of control replicates, the power of the tests for treatment effects is increased, achieving at least part of the advantages of the square-root allocation rule described above. (2) All the data are used and the pooled control provides the best estimate of the background population from the experiment. Permissible solvents are those which have been well-established in fish experiments and have been found to have no practical effect on fish at the concentrations used. A preferable alternative to always pooling the controls is to compare them statistically and pool them, if no significant difference is found, and otherwise use only the solvent control to test for treatment effects. The justification for the latter is that solvent is in all the treatment groups at approximately the same concentration as in the solvent control, so that one compares solvent plus treatment to solvent, the difference being the treatment effect. This is a plausible hypothesis based on the apparent additivity of effects in most aquatic chemical mixtures that is supported by concentration addition. References on this include Belden et al. (2007), Backhaus et al. (2010), as well as Kortenkamp et al. (2007). The last communication addresses endocrine disrupting chemicals specifically as well as other classes of chemicals.

Currently, there is a lack of harmonisation amongst different regulatory authorities on what is the control for statistical analysis (dilution water control or solvent control and if they should be pooled or not). A definitive answer to this question cannot be provided at present, but it has been recommended that a working group to progress this issue. A topic that might be addressed by such a working group is the reduction in the number of animals that could arise by eliminating one of the controls.

### 3.7 Power

In the design stage, the primary use of power analysis in toxicity studies is to demonstrate adequate power to detect effects that are large enough to be deemed important. If our methods have sufficient power, and we find that, at a given concentration, there is no statistically significant effect, we can have some confidence that there is no effect of concern at that concentration. Failure to achieve adequate power can result in large effects being found to be statistically insignifi-
significant. 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.

Deciding on what effect size is large enough to be important is difficult. In some cases, the effect
size may be selected by regulatory agencies or may be specified in guidelines.

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 mov-
ing frame of reference should be long enough to average out noise without being so long that un-
detected 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 inci-
dence rate in the same moving frame of reference already mentioned.

At the design stage, power must, of necessity, be based on historical control data for initial vari-
ance estimates. It may also be worthwhile to do a post-hoc power analysis to determine whether
the actual experiment is consistent with the criteria used at the design stage. If there is signifi-
cantly 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. 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 histori-
cal control data for the species and endpoint being studied should be reported.

Power Example

Suppose we want to determine the NOEC for mortality in an experiment with rainbow trout (*On-
corhynchus mykiss*), where past experience with this species suggests that background mortality
rate at the relevant age and test duration is near zero. We want to be able to detect a 20 % mortal-
ity rate and, based on preliminary range-finding experiments, we have decided on an experiment
with five test substance concentrations at 50, 100, 200, 400, and 800 ppm, plus a single (non-
solvent) control. Furthermore, suppose previous experience suggests that extra-binomial variance
and within-tank correlations of responses are unlikely, so a standard Cochran-Armitage test can
be done treating all fish within a concentration equally (i.e. ignoring any tank or replicate effect).
How many fish per concentration should we plan?
First, consider designs with the same number, \( n \), of fish in each concentration as in the control. The power of the Cochran-Armitage test depends on the shape of the concentration-response curve, which we do not know. Powers have been simulated for numerous shapes. Based on an examination of the various power plots, a reasonable choice for design purposes is the linear concentration-response shape. In addition, the power depends on the threshold of toxicity. For design purposes, we will assume that is zero. The following plots will help (Figs. 3.2 and 3.3).

Fig. 3.2: Power versus maximum change of \( 100 \cdot \text{Delta} \% \) for \( n = 5 \) fish per concentration.

Fig. 3.2 shows that 5 fish per concentration would give very low power (about 25 \%) to detect a 20 \% change in the high concentration. There is little point in conducting the experiment for the purpose.

Consider a design with 20 fish per concentration: This sample size gives a power of 82 \% to detect a 20 \% mortality in the 800 ppm concentration (Fig. 3.3). This may well be adequate. What is the power to detect a 20 \% in lower concentrations? Fortunately, we do not lose much power as we step down. The power to detect a 20 \% mortality rate at 400 ppm is 80 \%, at 200 ppm it is 78 \%, and at 100 ppm it is 76 \%. Notice, however, that if the background incidence rate were 10 \%, then the power to detect an increase in mortality rate of 20 \% drops to around 40 \%, which would be inadequate for most purposes.
Fig. 3.3: Power versus maximum change of 100·Delta % for \( n = 20 \) fish per concentration.

### 3.8 Replicates

Decisions on the number of fish per tank and number of tanks per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-tank variations and correlations. If there is only one tank per test concentration, then there is no way to distinguish housing effects from concentration effects and neither between- and within-group variances or correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to tank means. Thus, a minimum of two tanks per concentration is recommended; three tanks are much better than two; four tanks are better than three. Some non-parametric tests (e.g. not Mann Whitney) require a minimum of four replicates. The improvement in modelling falls off substantially as the number of tanks 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-tank 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). For further discussion and other tests (parametric and non-parametric), see OECD (2006).

The number of tanks per concentration and fish per tank should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.
Since the control group is used in every comparison of treatment to control, consider allocating more fish to the control group than to the treatment groups in order to optimize power for a given total number of fish. 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 \( n \) fish and \( k \) treatments to be compared to a common control, if the same number, \( n \), of fish are allocated to every treatment group, then the number, \( n_0 \), to allocate to the control to optimize power is determined by the so-called square-root rule. By this rule, the value of \( n \) is (the integer part of) the solution of the equation \( N = kn + n\sqrt{k} \), and \( n_0 = N - kn \). [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 \( \sqrt{k} \).

Computer simulations show that for the step-down Jonckheere-Terpstra and Cochran-Armitage tests, power gains of up to 25% can be realized under the square-root rule compared to results from equal sample sizes.

**Power example, continued**

What if we used the square-root rule in the above power example? Based on the above, we will examine the case where a total of 120 fish are used (20 per concentration and control in the above design). Under the square-root rule, we solve \( 120 = 5n + n\sqrt{5} \) for \( n \) to get \( n = 16 \). Then \( n_0 = 120 - 5 \cdot 16 = 40 \). The following power plot is based on this allocation (Fig. 3.4). Note that the power to detect a 20% increase in mortality rate in the 800 ppm group is now 92%. So, with the same number and spacing of concentrations and the same total number of fish, the power to detect a 20% increase in mortality rate has increased from 82% to 92% by using the square-root allocation rule instead of equal sample sizes. An alternative way to use the square-root rule would be to reduce the total number of fish required without loss of power. Indeed, power curves for nominal sample size \( N = 15 \) under the square-root rule show the power to detect a 20% increase in mortality is 86%. Thus, with a smaller total number of fish allocated optimally, the power to detect a 20% increase is actually increased. This result underscores the importance of good experimental design and test selection.
In experiments where two controls (dilution water and solvent controls) are used and 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 fish is not necessarily immediately clear. This of course would not be a consideration, if a practice of pooling of controls is not followed.

The reported power increases from the square root rule do not consider the effect of any increase in variance as concentration increases. One alternative is to add additional fish 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.

The square-root allocation rule holds little, if any, advantage for regression analysis. The reason for this is that the curve fitting activity only happens once, with all data. There is no special consideration or use of the controls.
3.9 Appendix 1: Detailed consideration of regression analysis for sex ratio endpoints

3.9.1 Comparison of alternative models

When using regression models based on treating the proportion male or female as a continuous response, there will be a need to select one model from among several candidate models fit to the data. There are both formal and informal selection criteria appropriate for the class of models that will be used for sex ratio data. The simplest approach to comparing models should be used in all cases, even when formal tools will also be used. This is visually inspecting the fits to the data. It is important to identify regions of concentrations where each model provides a poor fit. Next, the widths of the confidence bounds about the fitted curves should be compared. Generally, a model that gives narrower confidence bounds is preferred, but this does not outweigh the fit of the model to the data. There are situations where narrow confidence bounds are obtained for a model that clearly does not fit the data. Next, confidence bounds for all estimated parameters should be examined. If the confidence interval for a model parameter contains zero, then the model is suspect, as that parameter is evidently not required. Beyond that, preference is given to the model where the confidence intervals for the parameters are smallest. Where replicate data are available, the residual mean square from the model should be compared to the pure error mean square, which can be obtained from an ANOVA. Finally, Akaiki’s or Schwartz’s information criterion can be used. The preferred form for Akaiki’s information criterion is given by the formula below. A good discussion of AICc is given in Motulsky and Christopoulos (2004).

\[
AICc = \ln(RSS/n) + \frac{n+k}{n-k-2},
\]

where RSS is the residual sum of squares from the model, n is the total number of observations, and k is the number of parameters estimated for the model.

In general, the model with the smaller AICc is preferred. If the values of AICc are close for two models, it is helpful to compute the probability that model A is better than model B using the following formula:

\[
Prob = \frac{e^{-D/2}}{1+e^{-D/2}},
\]

where

\[
D = AICc(B) - AICc(A)
\]

Here AICc(B) denotes the value of AICc for model B. If the probability is high, then model A is favoured. The following plot of these probabilities may be helpful (Fig. 3.5).
Fig. 3.5: Probability of correct model selection

It will be observed that if $\text{AIC}_c(B) - \text{AIC}_c(A) \geq 10$, the model $A$ is almost certainly better than model $B$. This criterion is limited, however, where weighted fits are used, as two models can be compared using this criterion only if they use the same weights. So, in comparing two un-weighted model fits (i.e., weight=1), the criterion is sound. For weighted models where the weights depend on the function being estimated, the criterion is not appropriate and comparing an unweighted to a weighted model fit is certainly not appropriate. Some discussion of weights in using $\text{AIC}_c$ is given in http://www.boomer.org/Manual/ch05.html and http://www.micromath.com/products.php?p=scientist&m=statistical_analysis.

3.9.2 The meaning of an $x\%$ effect

The third issue concerns the meaning of an $x\%$ effect. For incidence data (such as percent males), there are three distinct concepts that are sometimes confused:

- Absolute risk is when $x\%$ of the population is affected.
- Additional risk is when $x\%$ above the “background” is affected, so that if the background incidence rate is $c\%$, then the total risk is $(x+c)\%$.
- Relative risk is when $x\%$ of the population that would “normally” not be affected is affected, or $c\% + (1 - c/100) \cdot x\%$. 
To illustrate the difference, consider the meaning of EC$_{50}$ when the background incidence is 20 %, i.e., C = 0.2 (background incidence rate).

**Fig. 3.6**: Absolute risk: 50 % of the population is affected (only 30 % above background)

**Fig. 3.7**: Added risk: 70 % is affected (total = 20 % + 50 %)
Fig. 3.8: Extra (or relative) risk: 60% is affected (total = 20% + 50 \cdot (1 - 0.2)\% = 20\% + 40\% = 60\%)

Probit analysis of incidence data is based on the concept of relative risk and care must be taken to arrive at the correct EC₅₀ estimate, if there is background incidence. For the sex ratio, if there is no background proportion male (i.e. no males in the control) and there are only males and females, then EC₅₀ for males is the same as EC₅₀ for females. However, if there is a background incidence of males, the two approaches are not equivalent. This is because it is important in probit analysis to analyse an increasing proportion, when there is background incidence. Probit analysis can handle background incidence for an increasing function, but gets thoroughly confused accounting for background in a decreasing function, for what does a “background” incidence rate of 70\% mean when the incidence rate in a treatment group is 40\%?
**Glossary**

1. **NOEC - Definition**
2. **ECx – Definition**
3. **Chi-square - Definition and reference to the statistical test**
4. **Cochran-Armitage – Definition and reference to the statistical test**
5. **Dunn – Definition and reference to the statistical test**
6. **Dunnett – Definition and reference to the statistical test**
7. **Fisher exact test - Definition and reference to the statistical test**
8. **Jonckheere-Terpstra – Definition and reference to the statistical test**
9. **Mann-Whitney – Definition and reference to the statistical test**
10. **Rao-Scott - Definition and reference to the statistical test**
11. **Williams - Definition and reference to the statistical test**

**References**


risk assessment for existing substances and Directive 98/8/EC of the European Parliament and of
the Council concerning the placing of biocidal products on the market.

Kortenkamp, A., Faust, M., Scholze, M., Backhaus, T. (2007). Low-level exposure to multiple

Länge, R., Hutchison, T.H., Croudance, C., Siegmund, F., Schweinfurth, H., Hampe, P., Panter,


OECD (2008) Detailed review paper on fish life-cycle tests, OECD Series on Testing and As-
essment No. 95, ENV/JM/MONO(2008)22. Organisation for Economic Cooperation and De-
velopment, Paris.


Programs, U.S. Environmental Protection Agency. Endangered and threatened species effects
determinations. Office of Prevention, Pesticides and Toxic Substances, Office of Pesticide Pro-


Evaluation of the reproductive effects of tamoxifen citrate in partial and full life-cycle studies