STATISTICAL ISSUES IN FISH LIFE-CYCLE TESTS WITH MANY ENDPOINTS

ECOΣTAT

N. van der Hoeven[1] and D.R. Dietrich[2]

[1]ECOSTAT, Vondellaan 23, 2332 AA Leiden, The Netherlands, NvdH@ecostat.nl, [2]Univ. of Konstanz, Germany

Abstract

Methods are described for the statistical analysis of a large set of variables as will be observed in a fish life-cycle test with a presumed endocrine disrupting compound.

- The significance level should be corrected for the number of comparison.
- Histopathological parameters can be tested using a non-parametric test for ordinal data.
- For the statistical tests observed cell numbers should be transformed to logits of relative cell frequencies.
- The gain in MSD of counting cells in more gonadal slides is calcluated.
- Power analysis should be used to calculate the number of replicates needed in new experiments.

Experimental design: gonadal cell type counting

Digital images can be taken from gonadal slides. In these images cell types can be counted when cells are either on or touching the intersect of a grid. Per slide one or more digital images can be analysed. In testis slides four stages of the reproductive cells can be distinguished: Spermatozoa, Spermatids, Spermatocytes and Spermatogonia. To estimate the gain in accuracy of analysing several slides per testis, the variance between slides can be estimated from a small set of gonads for which several slides are analysed.

Cell type frequency

The frequencies of four testis cell types can be analysed to answer the question whether a presumed endocrine disrupting compound (EDC) induces shifts in these frequencies. Absolute frequencies of cell types are correlated, since these frequencies always sum to 1. Therefore, instead of absolute frequencies, relative frequencies should be used. These relative frequencies can be calculated by first putting the cell types in their natural inverse order of development, i.e. from most to least mature. Then, the frequency of the most mature cell type is calculated within the group of all cell types (i.e. the absolute frequency of the most mature cell type). Then, the group of all but the most mature cell type is analysed, and in that group the frequency of the most mature cell type is calculated. This procedure is repeated up until the group of the most juvenile but with the exception of the most juvenile cell type. Four cell types thus lead to three relative cell frequencies. The relative frequencies could all have values between 0 and 1, independent of the observed values of the other relative frequencies.

The difference in relative frequency between control and treatment is not a very good measure to express the size of the treatment effect. Small changes in relative frequency are much more important if the relative frequency is very small or very high than if the relative frequency is intermediate. For instance, a decrease in relative frequency of 10% from 55% to 45% is of much less consequence than a decrease from 10% to 0%. Therefore, the analysis should be performed on the logit of the relative frequency which can be calculated per gonad. The logit of the cell frequency per fish is the mean of its two gonadal logits (logit of left and right gonad).

Statistical analysis of cell type frequencies

The Dunnett test [1] can be used to analyse whether the cell frequency in any of the concentration differs from the cell frequency in the control. In this calculation the vessel mean should be used as observation. The variable-wise significance level can be calculated using the Dunn-Šidák Bonferroni equation.

It is assumed that if an EDC affects the relative frequency of the more mature cell types it will decrease that specific frequency. For each frequency, the minimum significant difference (MSD) can be calculated. Experimental studies [5] showed that the MSD in mean vessel logit varies between -1.2 and -0.6. Figure 1 shows how the MSD for the logit can be translated to the minimum observable decrease in relative cell frequency in EDC experiments. Note that the MSD in cell frequency is less for small or large control frequencies than for intermediate cell frequencies.

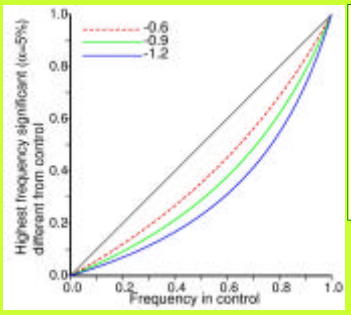


Figure 1 Relationship between the control frequency and the MSD in frequency given three different MSD values for the logits.

Introduction

The statistics for fish life-cycle tests to investigate the effects of a presumed endocrine disrupting compound (EDC) are discussed. For example, an experimental set-up is used in which a male and a female fish are housed together in one compartment, and each test vessel is divided into four compartments. Many variables are measured for each fish, such as length, survival, VTG level, gonadal cell type distribution and histopathological lesions. For each of these variables it has to be tested whether it is affected by the EDC. The statistical test to be used depends on the type of variable.

Determining the size of a relevant effect

In fish life-cycle experiments, just as in other ecotoxicological experiments, the statistical endpoint of the test is often an NOEC. It should be kept in mind that a NOEC does not guarantee that the effect of the chemical is acceptably small, but only that the experiment was not sufficiently accurate to show an effect at the chosen significance level.

Before performing the experiment, it should be decided which effect size is biologically relevant, and should therefore be observed with reasonable high probability. Determining which shift in the observed variable (e.g. length, gonadal cell frequency, ..) is relevant for the organism, for the population or for the ecosystem is outside the field of statistics. If an effect is deemed relevant, it is reasonable to require that the power to observe such an effect is high. In concordance with the usually chosen significance level (5%), a power of 95% would be the logical choice. This means accepting missing a biological relevant effect in 5% of the experiments. A common choice for the required power is 80%.

Designing an experiment with sufficient power

Only when knowing the required power, *P*, to observe a certain relevant effect size, δ, the experiment can be designed. For that design it is necessary to have a reasonable estimate of the variance, σ^2, in the variables to be observed. If, for instance, the one-sided Dunnett test at significance level α is used to determine the NOEC, and *k* treatments are compared with the same control, the necessary number of replicates per treatment and in the control, *n*, should be at least

$$2\left(U_{\alpha,n,k}+\Phi^{-1}(P)\right)^2\left(\frac{s}{d}\right)^2$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution and $U_{\alpha,\nu,k}$ the appropriate one-sided critical value for a test with ν degrees of freedom and *k* comparisons between a treatment and the control at significance level α [4].

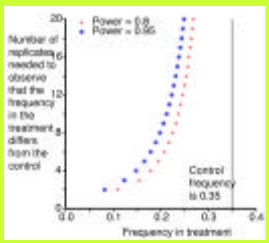


Figure 2 If four treatments are compared with one control and the frequency in the control is 0.35 with a standard deviation of the logit of 0.4 the number of replicates (vessels) needed is shown.

More accuracy by counting more slides per testis?

To determine whether counting more than one slide per testis increases the accuracy of the measurement of the actual cell type distribution sufficiently to warrant the extra cost and effort, cell types should be counted in several slides of the left and of the right testis in several exposure groups. The cells counted in each slide should be considered as a sample of the cell type distribution of a small location within the gonads (local cell type distribution). Variance between slides is due to the combination of the variance between locations and the variance due to taking, per location, a sample of the cells for counting (one or more grids per slide). In figure 3 the reduction in MSD obtained by analysing more slides per testis is shown.

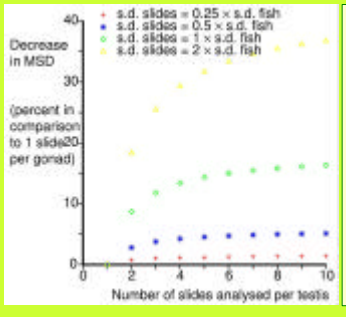
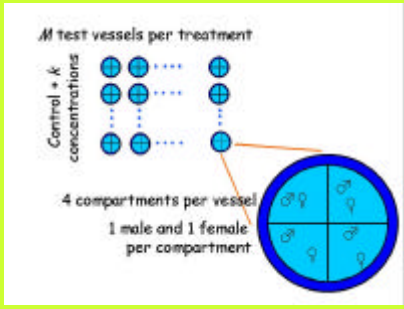


Figure 3 The reduction in MSD obtained by analysing 2 to 10 slides from each gonad instead of 1. The reduction depends on the ratio between the s.d. between slides of the same gonad and the s.d. between gonads of different fish.

Experimental design

The fish are housed in flow-through test vessels which are divided in four compartments. At the start of the experiment, a male and female adult fish is put into each compartment. At the end of the experiment, a large set of variables will be measured.



The experimental unit

Fish housed together in one vessel, although in different compartments, cannot be considered as independent. The test vessel is the experimental unit. The observed variance between fish in different vessels can be split in the variance due to the fish and the variance due to the vessel effect.

Significance level in case of many comparisons

In a fish life-cycle tests a large number of different variables are tested. To keep the simultaneous probability of wrongly concluding that a certain concentration has induced an effect below a chosen (simultaneous) significance level α (α=5%), the simultaneous significance level has to be split over the tested variables. This can be done using the Dunn-Šidák Bonferroni method which splits the simultaneous significance level α over the *k* comparisons to a per comparison significance level of $\alpha' = 1-(1-\alpha)^{1/k}$.

It is necessary to split the significance level in some way over the comparisons. If the number of variables is e.g. 70 and a significance level of 5% is used for each variable separately this would lead to a probability of 97% of observing a significant effect in at least one of these 70 variables.

Sometimes the variables can be classified into several groups, e.g. "histopathological lesions", "gonadal cell type distribution", and "general body measures". In this case, the significance level can first be split over these groups, and then within these groups over the variables.

Histopathological lesion scores

The tissues of several organs such as liver, kidney or gonads can be screened for all kinds of lesions. For each lesion the tissue is scored on an ordinal scale running from 0 (no lesion) to *U* (most severe lesion). The observations can be tested with an ordinal non-parametric trend test, such as the Jonckheere-Terpstra test [2]. For this test it is assumed that the frequency and severity of the lesions increases with increasing EDC exposure. This test allows to use the severity score of the lesion and not just the absence or presence of lesions.

The number of lesions scored is often very large. Therefore, the per-variable significance level is low, leading to a small power to observe effects. Nevertheless, using this methods significant effects can be observed in some experiments with EDC, e.g. Bisphenol A [5].

Redundancy analysis of lesion scores

The number of lesions observed can be very large leading to a small power if the significance level is distributed over the lesions using a Bonferroni equation. Instead, a weighted sum score of all lesions can be calculated. This sum score is calculated to maximise the difference between the sum score of different treatment levels. Then, it can be calculated whether this maximum difference between treatments is significantly larger than the difference would be if the observations are distributed randomly over the treatments. This kind of analysis is described in [3]

References

[1] Dunnett, C.W., 1964. New tables for multiple comparison with a control. *Biometrics* **20**, 482-491

[2] Hollander, M. & D.A. Wolfe, 1973. Nonparametric statistical methods. John Wiley & Sons, New York, USA

[3] Ter Braak, C.J.F. & P. Šmilauer, 1998. CANOCO Reference manual and User's Guide to CANOCO for Windows. Centre for Biometry Wageningen

[4] Hoeven, N. van der, 1998. Power analysis for the NOEC: What is the probability to detect small toxic effects on three different species using the appropriate standardised test protocols? *Ecotoxicology* **7**, 355-361

[5] Friederich, U. et al., 2005. Effects of bisphenol A on adult fathead minnw (P. promelas) somatic organ and gonadal histopathology: a 4s2-days exposure study. Poster on SETACEurope Lille 2005.