Observations on the nature of statistical distributions around sea lice count data from Norwegian salmon farms and their implications for interpretation of accuracies and uncertainties around estimates

Including summaries based on empirical field data and Monte Carlo simulations of potential scenarios used to estimate sea lice abundance levels and likely measures of treatment efficacy.

A report by Dr. Jaewoon Jeong and Prof. Crawford W. Revie

June 2018

Atlantic Veterinary College, UPEI, Charlottetown, PEI, Canada

This short report outlines the results of some research carried out at the Atlantic Veterinary College over the period April to June 2018, under the FHF Project 901411 coordinated by NINA. The report is divided into two sections which deal with (A) the analysis of some empirical field data collected as part of the project activities and (B) the use of a Monte Carlo based simulation model to consider the likely outcome of sampling sea lice under a variety of hypothetical scenarios.

The details below are provided in relatively short-hand form as there has been limited time to coordinate inputs with colleagues from the group at NVI who are exploring similar questions. It is expected that in due course the work of researchers at AVC and NVI can be integrated to provide the basis of a scientific publication around these topics.

The report begins by looking at the results of analyses from sea lice data observed on a number of Norwegian farms. The main purpose here was to explore overall distributional characteristics of these sea lice counts and to gain insights into the range of parameters relating to dispersion, abundance, etc., that should be used in the hypothetical modelling exercises that followed on from this initial exploration. The second section reports on some of the main outcomes of relevance seen in the Monte Carlo simulation models (developed using the R package).

Overall Summary

While the details are given in the pages that follow, some of the material is somewhat technical, involving discussion of statistical distributional assumptions, etc. While it is important to report these details to make explicit the extent, and limitations, of any broader statements, it is also useful to summarise the key implications of this initial exploration.

Key findings from field-based base sets

- the distributions explored confirmed that the best option to adopt when summarising these types of count data in the ***negative binomial*** (NB) distribution. As expected, an

increase in the variance-to-mean ratio within these populations will increasingly render the Poisson distribution ineffective for these types of data;

- perhaps surprisingly, there did not appear to be a need to model these data using a *zero-inflated negative binomial* (ZINB) distribution;
- the value of ***k*** (the dispersion parameter) did not increase with mean abundance as significantly as had been proposed in previous (hypothetical) studies. This means that even in situations with increased abundance (here up to 10 sea lice) assuming a Normal distribution may lead to inaccurate interpretations;
- a reasonable set of ***k*** parameters should therefore be explored in any set of simulation experiments. In addition the level of cage-to-cage clustering, measured for example by the ICC estimate (Revie *et al*., 2007) will have to be considered in tandem with this dispersion parameter, and it is likely that the two will vary in a systematic manner.

Key outcomes from the Monte Carlo simulations

- the probability that a sample size of 20 fish would generate an estimate that was ±0.5 given a true abundance level of 1 sea louse was around 90% when limited dispersion was present. However, in the present of modest dispersion (***k***=0.5) this fell to around 70%;
- if an accuracy of ±0.25 (on a true abundance of 1 sea louse) was being sought, then even a sample size of 50 fish would only result in a probability of around 60% that this could be achieved;
- when looking to accurately estimate the effect of a sea lice treatment, the Monte Carlo simulations illustrated that when either the expected efficacy was low (e.g. less than 50% as may be typical in the case of modest resistance levels in the parasite population) or the level of dispersion is modest (***k***=0.5), the statistical power to detect treatment effects is severely limited even with sample sizes of as many as 50 fish;
- while an analytical formula to estimate the statistical power to detect a meaningful treatment effect exists, it was found that it significantly over-estimated the likely power (i.e. tended to under-estimate the impact of random variation in the field as demonstrated by the Monte Carlo approach).

Analyses of empirical data

Sea lice count data were obtained from eight sites and their key parameters from the perspective of sampling are summarised in Figure 1. These data were not categorised by cage or sampling date; rather the data are categorised by sites and developmental stage.

One of the datasets used in Figure 1 was then more thoroughly investigated (Figure 2). Sea lice counts were obtained from two sites consisting of eight cages at 28 sampling dates. Each site, cage, and every sampling date were analysed as separate data points. The data consist of three developmental stages (chalimus, mobile, ovigerous) though as can be seen from the graphs there was little evidence of significant differences among these developmental stage.

Fit to negative binomial distribution

Many ΔAIC values are around zero (Figure 2), indicating that both the negative binomial distribution and Poisson distribution have a reasonable fit to the sea lice counts. However, as the value for VMR rises so the ΔAIC tends to reduce, indicating that the negative binomial distribution is a more appropriate distribution with which to represent sea lice counts under those situations involving high variance to mean ratios for abundance.

Relationship between mean abundance and *k*

It was expected that increased abundance would lead to an increasing value for ***k*** (Heuch *et al*, 2011); however, this did not appear to be the case in these analyses. VMR and abundance indicated a broadly positive relation (Figure 3), whereas it had been hypothesised that VMR would be constant regardless of abundance (Heuch *et al*, 2011). In other words, over-dispersion (high VMR) tends to be more marked as sea lice levels increase.

Table 1. Description of abbreviations

k	Dispersion parameter of the negative binomial distribution
VMR	Variance-to-Mean Ratio
ΔAIC	*Delta* AIC value. Here used to compare model fit as ([Negative binomial distribution] – [Poisson distribution])

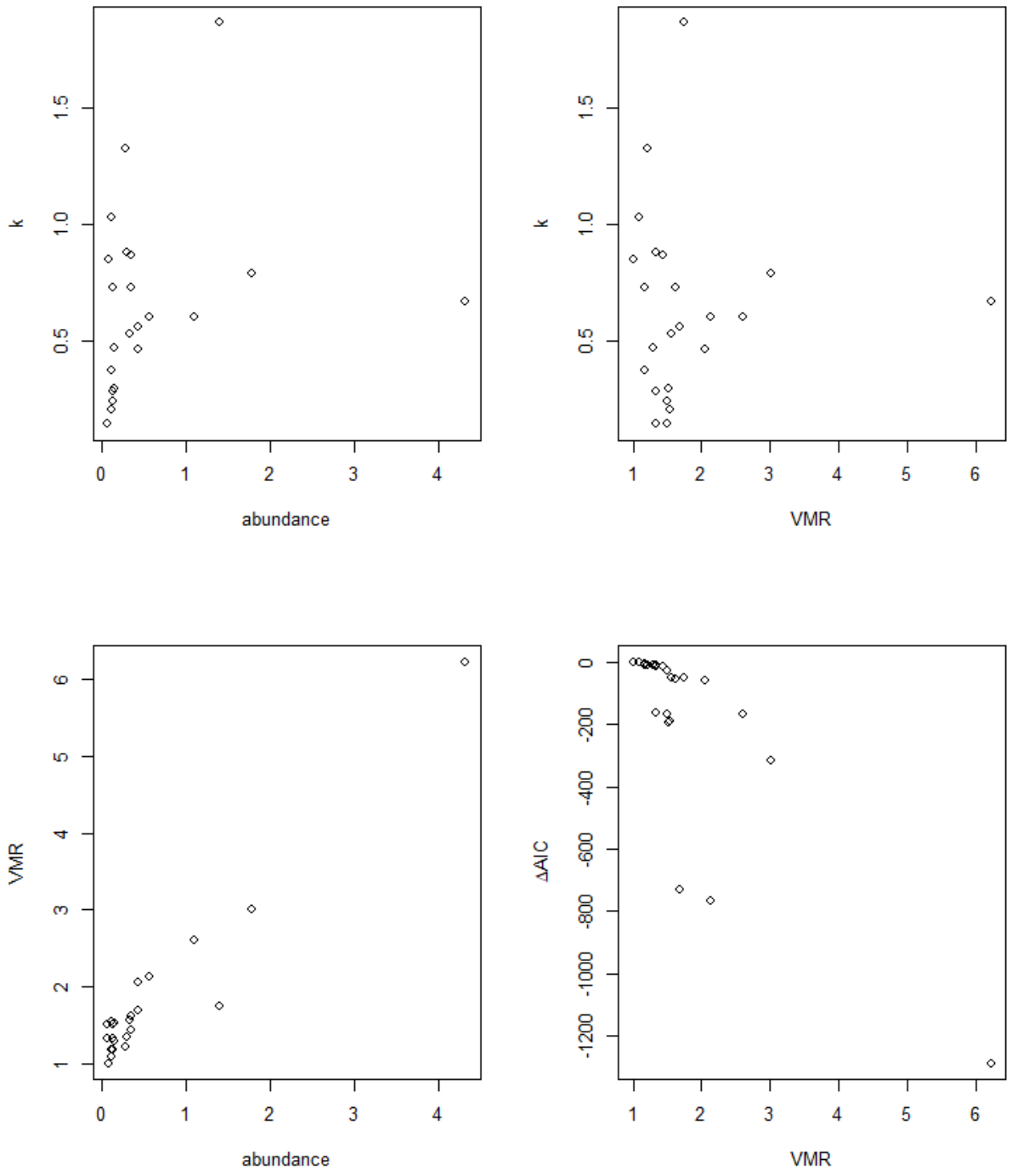

Figure 1. Sea lice count data from multiple locations in Norway. Sites and developmental stages were analysed separately, but cages and sampling dates were not differentiated. Eight sites with two or three different developmental stages were analysed.

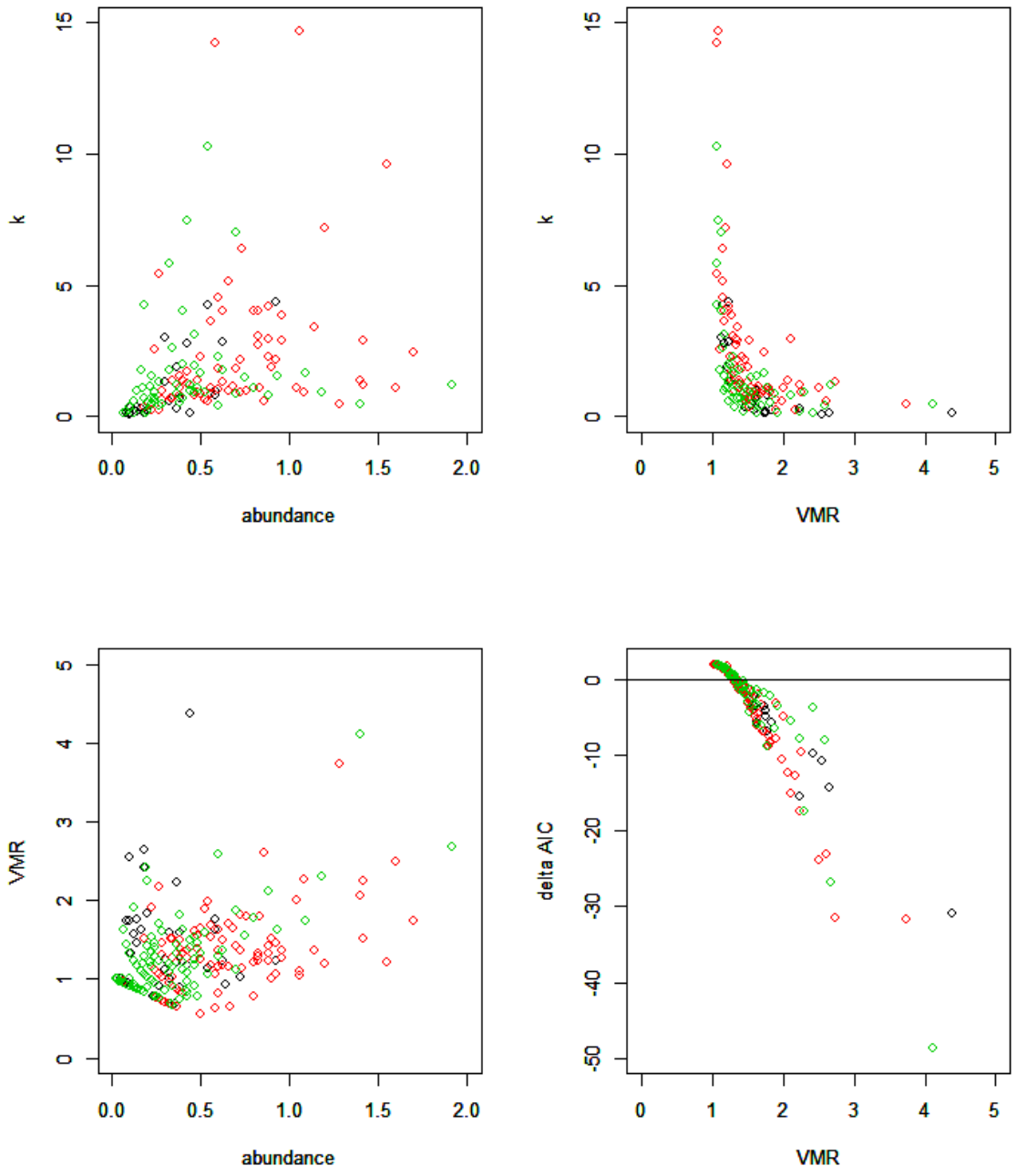

Figure 2. A dataset of a location in Norway was thoroughly investigated. Sites, cages, sampling dates were separately analysed. This analysis involved 26 datasets from five sources; each source had a slightly different way to categorize sea lice. As such total number of sea lice from each source for the analyses are given. Black, red, and green dots represent chalimus, mobile, and ovigerous stages, respectively.

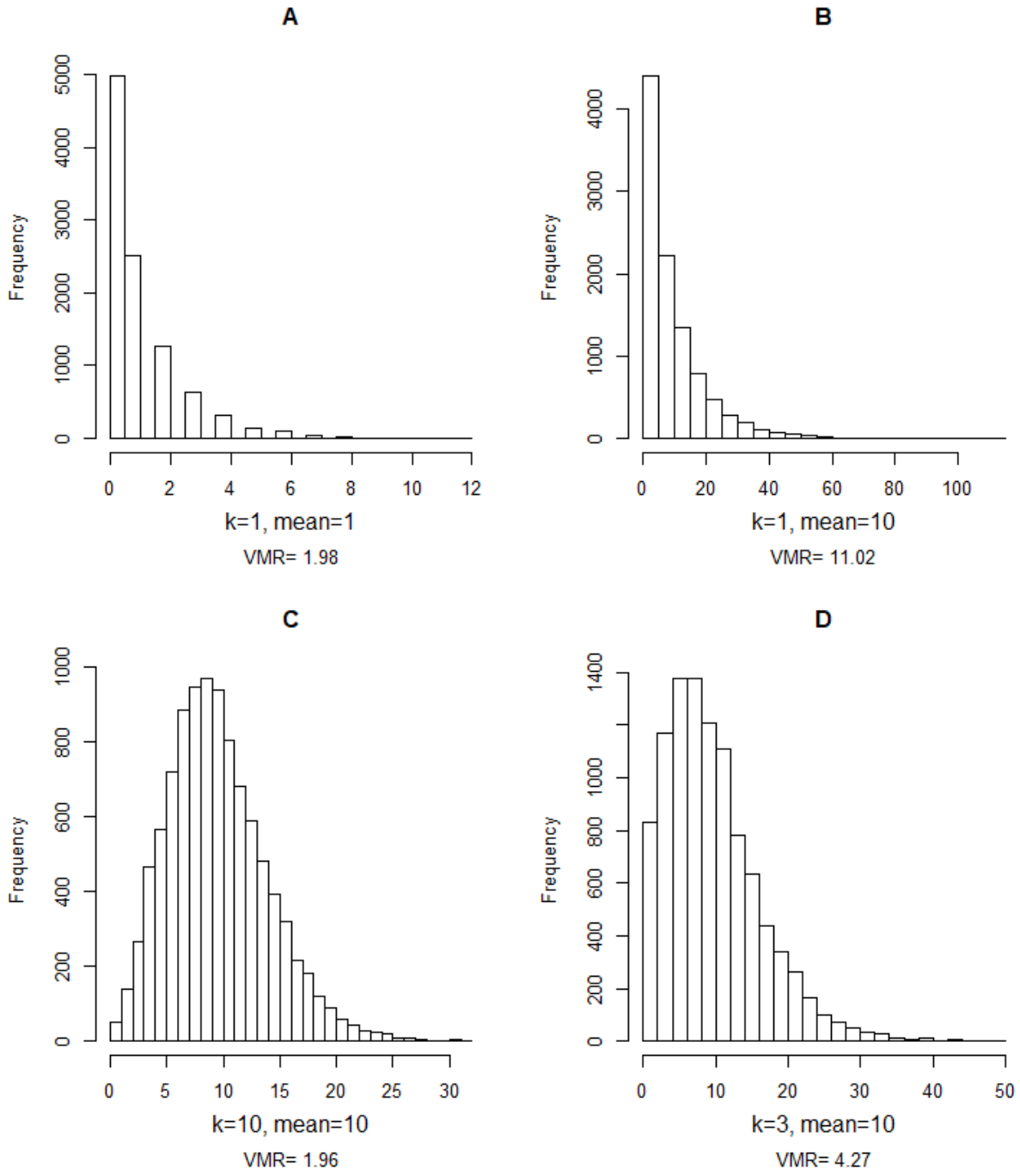

Figure 3. A simulation (N=10,000) to demonstrate the potential relationships between ***k*** and mean abundance. When the situation in [A] (***k***=1, mean abundance=1) shows an increase in mean abundance, the dispersion parameter (***k***) may remain constant [B] or may also increase [C]. Heuch *et al* (2011) suggested that [A] becomes [C] as mean abundance increases. However, our analysis of empirical datasets suggested that [A] becomes [D], which is somewhere between [B] and [C].

Simulations based on a hypothetical model

Monte Carlo simulation was used to explore how different sample sizes related to different outcomes under differing levels of mean abundance and dispersion (based in the value of the ***k*** parameter). This was explored under the assumption that the sampling was designed to achieve one of two broad goals: 1) to estimate sea lice abundance; and 2) to test for differences in sea lice abundance following a treatment. In all cases shown in the graphs below the mean values that have been reported are based on 10,000 runs of the simulation code.

1) To estimate sea lice abundances

The implications for getting the correct estimate of mean abundance under different assumptions of dispersion (the lower the value of ***k*** the more dispersed) using different sample sizes is illustrated on Figure 4. In both panels the true abundance is 1.0 sea lice per fish and not surprisingly the probability of making an estimate ±0.5 of this value (left panel in Figure 4) is higher than the probability of being within 0.25 of the true value (right panel in Figure 4). In both cases it can be seen that the larger the sample size, the more likely that our estimate will be within these accuracy bounds. In the case where we are willing to be only within ±0.5 of the true value of 1 louse and have limited dispersion (***k***=1), the benefit of moving from a sample size of 20 to 50 is fairly marginal (85% to 95% likelihood of being within the correct range). However, on the case where is wish to be within ±0.25 of the true value of 1 louse and have more marked dispersion (***k***=0.3; black line in right panel of Figure 4) a sample size of 20 will only give a 45% likelihood of this level of accuracy, rising to just over 60% with a sample size of 50 fish. The estimation of abundance was more accurate with high ***k*** than with low ***k***, which is due to the fact that it is easier to obtain estimates from a large number of fish with fewer sea lice (Figure 3C) than from only a few fish with high levels of sea lice (Figure 3B). As expected we gain better estimates when more fish are sampled but we tend to require larger sample sizes when ***k*** is lower. A fuller assessment of these relationships using a slight more 'academic' approach is given in Figure 5. Here it can be seen that using a sample size of 5 (red lines),will tend to consistently under-estimate the true mean abundance, because too many fish with no sea lice are likely to be sampled; this was particularly obvious in the cases with lower values of ***k***.

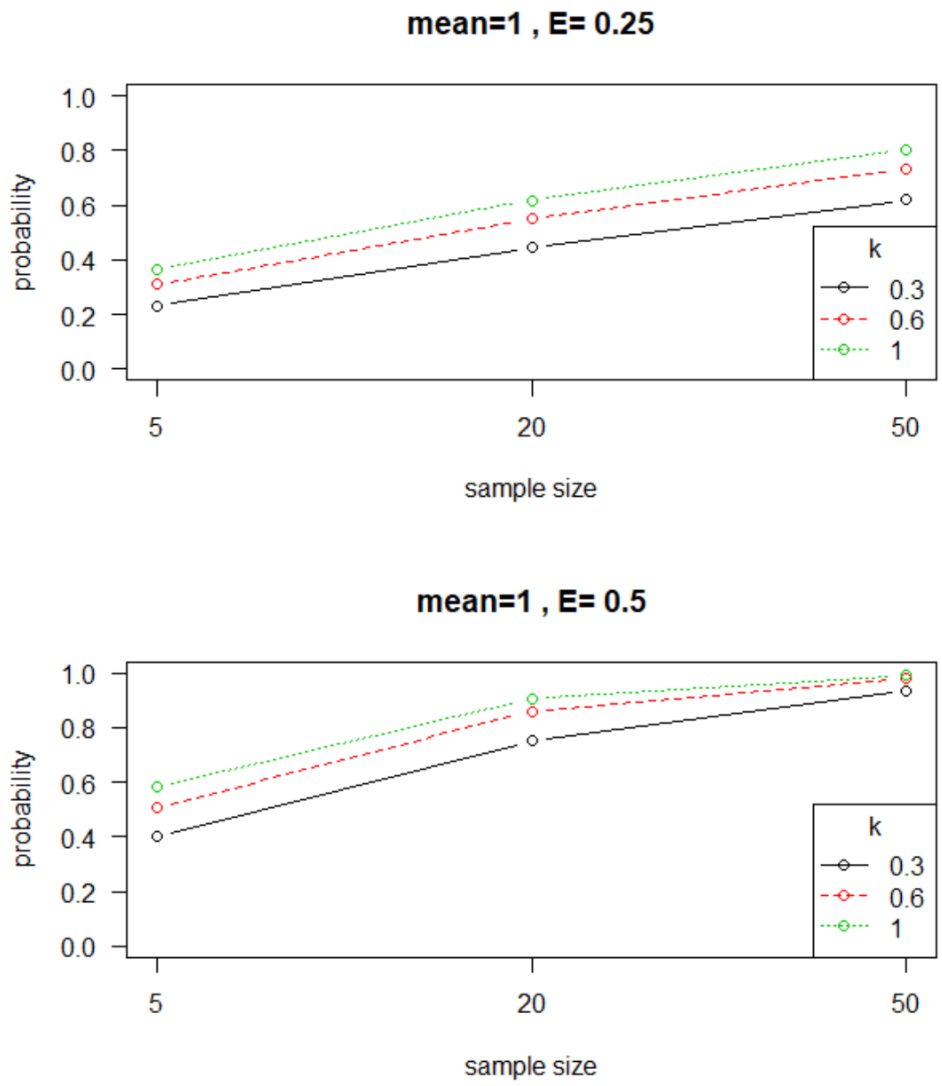

Figure 4. Accuracy of estimated mean abundance levels. The simulations illustrate the probability of obtaining two different errors (±E in absolute terms) in estimated abundance given a mean abundance of one louse per fish (mean=1). These errors were (E=0.25) and (E=0.5) in the top and bottom panels, respectively.

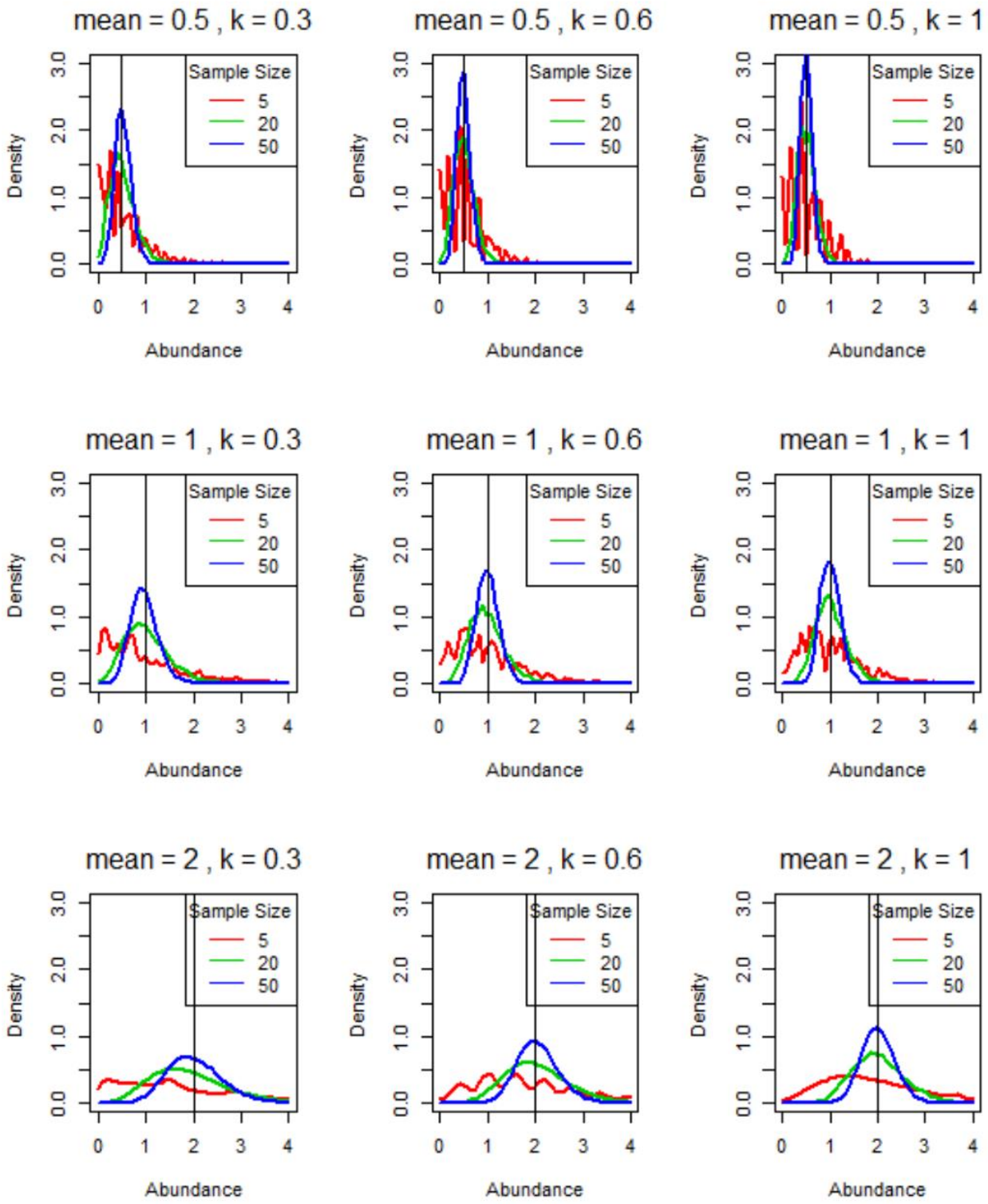

Figure 5. Estimation of sea lice abundances through Monte Carlo simulation. In plots, mean abundance differs by row, while the dispersion parameter (***k***) differs by columns. Density of probabilities around estimated abundances are shown on the y-axis.

2) To test for differences in sea lice abundance following a treatment

A common practical setting in which the accurate estimation of sea lice levels is of great importance is when assessing the impact (efficacy) of a treatment. Clear as the mean abundance before treatment and/or the expected efficacy increase, the power to detect a difference in the mean abundance following treatment will also increase (Figure 6). Clearly where a mean efficacy of 70% is expected when treating on fish with a mean abundance of 1 sea louse (green lines in bottom row of Figure 6) a sample size of around 30 fish gives a power of around 0.85 when there is limited dispersion (***k***=2). However, in a setting with modest dispersion (***k***=0.5) even sampling 50 fish would only give rise to a power of around 0.55 to pick up this effect.

The outputs in Figure 6 are based on a Monte Carlo simulation of treatment outcomes and sampling but for this reasonably restricted case it is also possible to generate an analytical solution based on an equation in Cundill and Alexander (2015):

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}(\frac{1}{\mu_1} + \frac{1}{k_1}) + \frac{1}{Q_0}(\frac{1}{\mu_0} + \frac{1}{k_0})}}{\log(\mu_0) + \log(\mu_1)}$$

Overall the trends in terms of statistical power to detect differences are largely similar to those seen when using the Monte Carlo simulation-based approach. However, there were some differences, particularly when assuming small sample sizes (Figure 7). The analytical results were less sensitive to different values of mean abundance, ***k***, and sample size than was the case in the Monte Carlo simulations, but this may well be due to the lack of stochastic variation in the analytical approach as well as the fact that certain simplifying assumptions around the parameters which describe the distributions are made in the analytical solution.

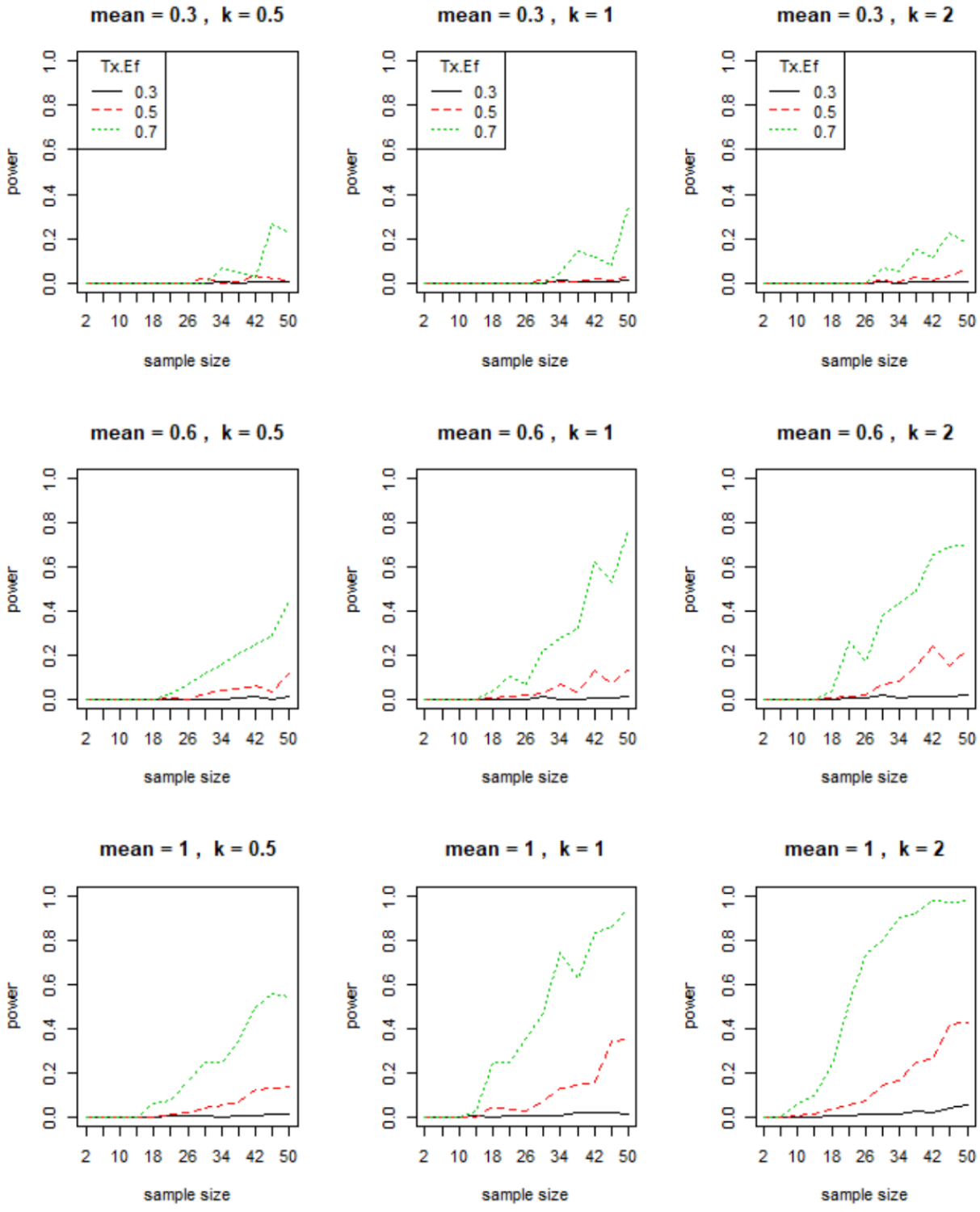

Figure 6. Testing for differences (Monte Carlo simulation). In plots, mean abundance (before treatment) differ by row, while the dispersion parameter (***k***) differs by column. The statistical power (1-β) is shown on the y-axis, based on a Confidence level of 0.95. ('Tx.Ef' indicates the treatment effect – e.g. if Tx.Ef =0.3, then a mean abundance of 1 louse would reduce to 0.7 after treatment).

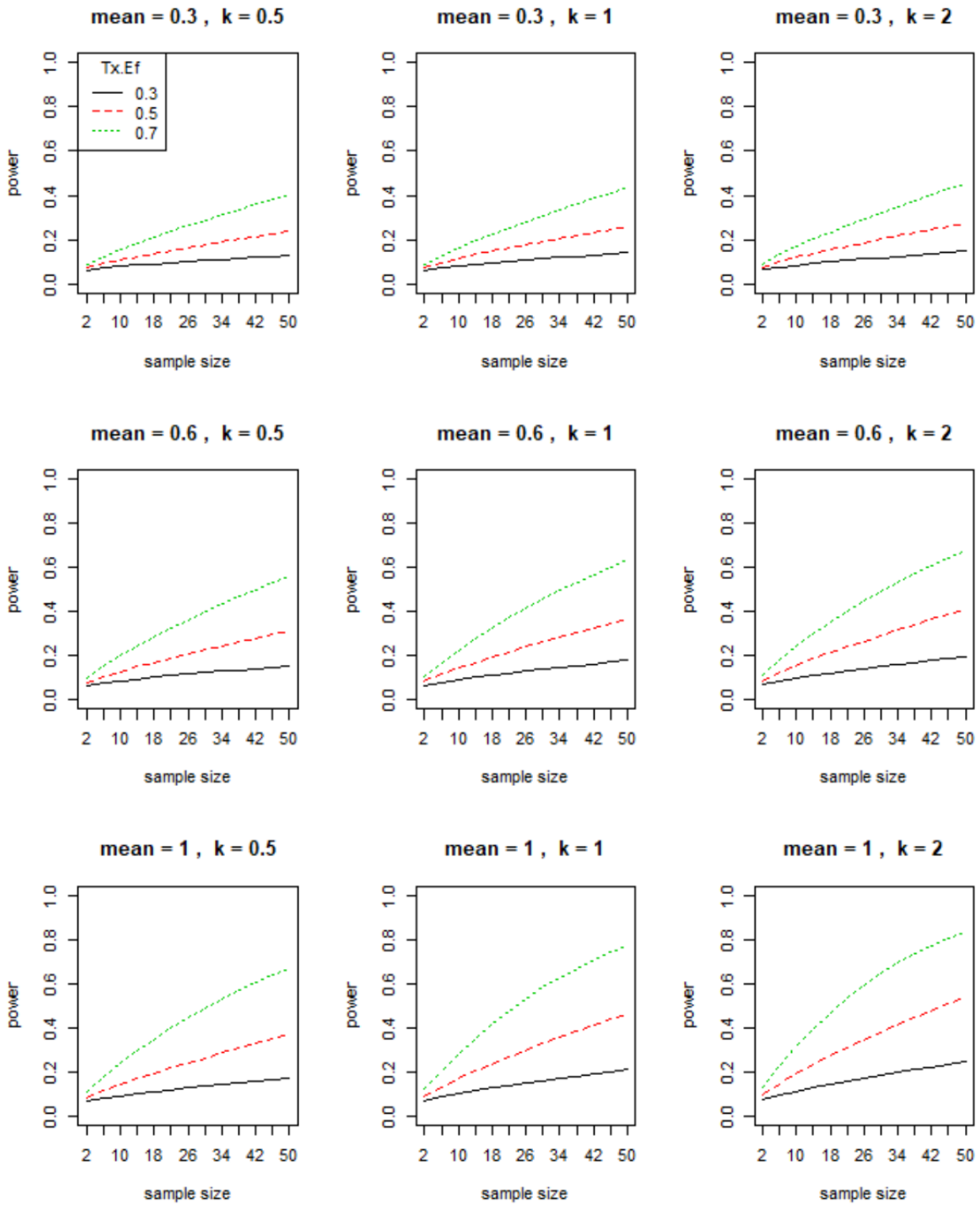

Figure 7. Testing for differences in abundance following a treatment using a fixed analytical equation (Cundill and Alexander, 2015). The layout of plots is the same as in Figure 6.

References

Cundill, B., Alexander, N.D., 2015. Sample size calculations for skewed distributions. *BMC Medical Research Methodology* 15, 28.

Heuch, P.A., Gettinby, G., Revie, C.W., 2011. Counting sea lice on Atlantic salmon farms—empirical and theoretical observations. *Aquaculture* 320, 149-153.

Revie, C.W., Hollinger, E., Gettinby, G., Lees, F., Heuch, P.A., 2007. Clustering of parasites within cages on Scottish and Norwegian salmon farms: Alternative sampling strategies illustrated using simulation. *Preventive Veterinary Medicine* 81, 135-147.