# Modelling Zero-inflated Fish Counts in Estuaries – A Comparison of Alternate Statistical Distributions

**[1]Mayer, D., [2]D. Roy, [2]J. Robins, [2]I. Halliday and [2]M. Sellin**

**[1]Animal Research Institute, and [2]Southern Fisheries Centre, Department of Primary Industries and Fisheries. E-Mail:** david.mayer@dpi.qld.gov.au

***Keywords:*** *environmental flows; assemblages; delta distribution; generalized linear model; GenStat.*

**EXTENDED ABSTRACT**

Freshwater flows have an important influence on the balance of fish species in estuaries. As well as supporting general ecosystem health, these flows are also necessary to comply with Australia-wide legislation aimed at the sustainable management of water resources. The Coastal Zone CRC and Fisheries Research and Development Corporation (FRDC) instigated the 'environmental flows for estuaries' project in three dry-tropics estuaries in central Queensland, namely the Fitzroy, Calliope and Boyne Rivers. The effects of the independent variables, including freshwater flows, on catch rates now need to be quantified. Typically, for each species the count data from the two-minute spatial and temporal trawl samples include many zeros, and are significantly skewed.

This inflated zero-class violates the statistical assumptions of many standard analytical techniques. A more appropriate method is to use two-part conditional distributions – firstly, a Binomial to represent the proportion of zeros (simple presence or absence of each species in each trawl), and secondly, if present, a truncated distribution modelling the catch numbers (>0). For this second part there is a range of available distributions – researchers in ecology and entomology have typically used the Poisson or Negative Binomial for their discrete counts, whereas in meteorology, health statistics, and fisheries and air-pollution research, continuous distributions such as the Gamma or log-Normal have been profitably employed. Whilst catches are necessarily integers, the addition of modelling terms (such as effort) effectively converts these onto a continuous basis, so statistically the use of continuous distributions is quite acceptable.

To compare statistical methods, we consider a subset of the species in our extensive data set – the most common (*Thryssa hamiltoni*, an anchovy, with 44% of samples being zero), somewhat rarer (*Favongobius exquisitus*, a goby, 72% zeros) and rare (*Valamugil* sp., mullet, 82% zeros), as well as a commercial fish (*Pomadasys kaakan*, banded grunter, 63% zeros). The zero-truncated distributions of catch were all positively and significantly skewed. For each species, the range of available conditional distributions was fitted, with the final model including site and month as factors, flowmeter reading (i.e., effective effort) as a linear covariate, and salinity, pH and turbidity as quadratic covariates. Residual plots were used to compare the observed values with the expected distributions. Back-transformed adjusted means were estimated for the model terms, and compared.

Whilst the approximate Poisson and Negative Binomial models occasionally performed well, this does not hold across all situations, especially when the mean catch rates are low. Hence the correctly zero-truncated versions of these distributions must be used. However, these are yet to be incorporated into the major statistical packages, so complex coded optimisation procedures are required. The Negative Binomial, in particular, suffered from poor convergence and computational problems.

Across our four example data sets, there was no consistent pattern regarding how the different distributions performed. The zero-truncated Poisson was shown to be unsuitable, as it could not accommodate the degree of overdispersion in the data. The Negative Binomial generally performed well, but for the species with the lowest degree of model fit produced unacceptably low fitted means. The residual plots for the Gamma models were quite acceptable, however some question remains over the fitted means, as these tended to follow the Poisson models.

The truncated log-Normal distribution gave the best overall results – significance of the independent factors and variates, good distributions of the residuals, and generally responsive fitted means which were usually 'amongst' the other models. In particular, they never went into 'unbelievable' regions, unlike the other distributions.

One remaining problem, however, is the current lack of an acceptable method to estimate the standard errors of this log-Normal by Binomial combination. This is an area of ongoing investigation.

## 1. INTRODUCTION

The balance of fish species in estuaries is affected by many factors. These include freshwater flows, which must be allowed to pass down into estuaries to support ecosystem health, and to comply with new Australia-wide legislation aimed at the sustainable management of water resources. The 'environmental flows for estuaries' project, funded by FRDC and conducted by DPI&F and the Coastal Zone CRC, has been running in three dry-tropics estuaries in central Queensland (the Fitzroy, Calliope and Boyne Rivers), for a number of years.

The relationship between estuarine demersal assemblages and freshwater flows was investigated. Samples were collected using a beam trawl (1 m wide by 0.5 m high, with a mesh size of 6 mm). The beam trawl was towed in water 0.5 to 1 m deep for two minutes (~50m in distance), with the volume of water flowing through the mouth of the net measured using a flowmeter. Sites were selected for sampling on the following attributes: (i) presence of mangroves; (ii) presence of a runoff channel; (iii) depth; and (iv) ability to be trawled (i.e., no snags). Samples were collected on the falling tide, within two hours of high water, around the new moon between October and May in 2002 and 2003. Data for salinity, pH and temperature were collected using a Yellow Springs Instruments water quality meter, and turbidity was measured as secchi depth (which is inversely related to the actual level of turbidity). Demersal assemblages were characterized by a larger number of species (~160), with 10 species dominating the overall catch (95% of all animals sampled).

For the comparison of statistical methods, we consider a subset of the species in our extensive data set – the most common (*Thryssa hamiltoni*, an anchovy), somewhat rarer (*Favongobius exquisitus*, a goby) and rare (*Valamugil* sp., mullet), as well as a commercial fish (*Pomadasys kaakan*, banded grunter). Descriptive statistics for these species are listed in Table 1, and the distribution for the first is shown in Figure 1. Histograms for the latter three species appeared similar, but somewhat less smooth.

**Table 1.** Descriptive statistics by species.

| | *T-ham.* | *P-kaak.* | *F-exq.* | *Val.* |
|---|---|---|---|---|
| Ppn. zeros | 0.44 | 0.63 | 0.72 | 0.82 |
| Mean | 8.7 | 1.3 | 1.0 | 3.8 |
| Mean (>0) | 16 | 3.7 | 3.5 | 21 |
| Median (>0) | 6 | 2 | 2 | 2 |
| Maximum | 196 | 56 | 26 | 876 |
| Skewness | 3.2 | 4.9 | 2.5 | 7.7 |

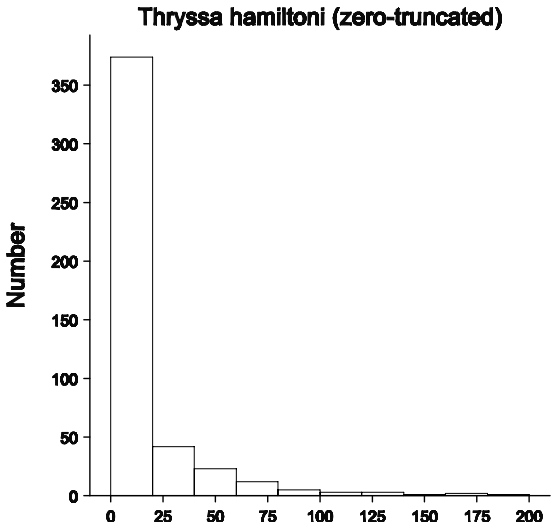


**Figure 1.** Histogram of non-zero catches per trawl.

## 2. STATISTICAL MODELLING OF POPULATION DATA

Obviously, multivariate analyses of species assemblages are appropriate for this data set, and these analyses are being undertaken. However, one key aim of the study was to investigate the effects of environmental flows, and the other known important covariates, on each individual species. This requires sequential univariate analyses of the fish counts for each.

If the observed data skewness and heteroscedastic variance can be assumed to be mainly attributable to the population strata and other independent variates, analyses of the raw data (assuming the usual Normal distribution) may be appropriate (Haddon 2001). Heteroscedastic regression models (Welsh et al. 2000) extend this by allowing the variance to be a function of the mean. Unfortunately, residual plots showed these simpler methods were not appropriate for our skewed data.

Transformation models, such as the square root or log, have been mentioned as potentially useful for abundance data (Welsh et al. 1996, Ye et al. 2001). In instances where sampling time is 'lengthy', these are appropriate – for example, the boat catch per day data in the northern trawl fishery (Bishop et al. 2004) has no zeros, and is approximately log-Normally distributed. However, in cases with smaller sampling times which result in even a mildly-inflated zero class, these are unlikely to be appropriate. The simulation study of Welsh et al. (2000) showed that these transformation methods produce markedly biased results, even when using bias-corrected back-transformations.

A range of more flexible and skewed distributions has been used for modelling animal abundances, including the Poisson (Gardner et al. 1995, Welsh et al. 1996), the overdispersed Poisson or Negative

Binomial (Gardner et al. 1995, Welsh et al. 2000), and Poisson mixtures or Neyman type A (Dobbie and Welsh 2001, who concluded that conditional models were superior). Again, these distributions tend to become less valid as the proportion of zeros in the data increases.

On the other hand, conditional models show good potential to accommodate all data sets. This approach combines a model for the binary (presence/absence) nature of the zeros; and then a second distribution modelling these numbers (conditional upon their presence). As Welsh et al. (1996) demonstrate, the overall log-likelihood is the sum of the two independent components, and the fitted parameters are orthogonal. Hence, it is valid and usual to fit these distributions separately (Feuerverger 1979, Pennington 1983, Tu 2002, O'Neill and Faddy 2003).

For the binary component of the conditional model, it is most common to use a logistic model assuming the Binomial distribution (Feuerverger 1979, Welsh et al. 1996, 2000, Ye et al. 2001, O'Neill and Faddy 2003). The influence of the various factors and covariates on this probability of capture is usually investigated via generalized linear models (McCullagh and Nelder 1989).

There is a range of distributions available for the second component of the conditional model, i.e., for modelling the numbers present. Distributions which allow zero values need to have these truncated, before being useful as a conditional distribution with the Binomial. With larger mean values this truncation may not be necessary as the untruncated version can provide an acceptable approximate solution (Welsh et al. 1996), but this will not apply generally. Available distributions, and their use in conditional models, include -

Truncated Poisson – this discrete distribution was used to model industrial processes (Lambert 1992), and possum counts (Welsh et al. 1996).

Truncated Negative Binomial – this distribution has a more flexible shape than the Poisson, as it allows for extra variation (Tu 2002). It has been used to model seabird nesting counts (Welsh et al. 1996) and recreational fishery catch rates (O'Neill and Faddy 2003), with both these data sets displaying overdispersion.

Extended Poisson process models – based on a Markov birth process representation of discrete distributions (Faddy 1997), this allows a range of distributional shapes, with the Poisson and Negative Binomial being particular cases. Podlich et al. (2002) applied it to ecological data, where results were similar to the truncated Poisson. O'Neill and Faddy (2002) adopted it for the analysis of recreational fishery catches, and found close agreement with their truncated Negative Binomial models (O'Neill and Faddy 2003).

Log-Normal - this combination (of the Binomial with the log-Normal) was initially termed the delta distribution (Aitchison 1955), and has been used to model mackerel egg counts (Pennington 1983), health statistics (Zhou and Tu 2000), and air contaminant levels (Tu 2002). In a review of 78 zero-truncated fisheries and ecological samples, Myers and Pepin (1990) found that only 7 showed significant ($P$<0.05) departure from the log-Normal distribution.

Gamma – this has been used successfully as a conditional distribution in meteorological models (Feuerverger 1979), and fishery catch rates in Kuwait (Ye et al. 2001), where zero catches occurred in 50% or more of the data. Myers and Pepin (1990) showed their Gamma models to be generally similar, but more often slightly superior, to the log-Normal fits.

## 3. METHODS

The design factors in this survey were year, river, location (side-creek or main branch) and month. For each species, catch number per trawl is taken as the independent variable, with 837 spatial and temporal observations. Generalized linear models (McCullagh and Nelder 1989), in GenStat (2005), were used to fit the independent terms to the observed counts. Continuous covariates, uniformly fitted as quadratics to accommodate curvature, included temperature, salinity, pH and turbidity. Flowmeter reading is dependent on both trawl speed and time, and is a direct measure of water entering the trawl net. Flow is thus an effective measure of effort, and was included as a linear covariate. Correlations between these variates are listed in Table 2. Whilst a correlation coefficient of 0.36 is statistically significant, this relationship only explains 13% of the variation, so there is ample scope for each to contribute approximately independently to the catch models.

**Table 2.** Correlation matrix for the independent variables.

| | Temperature | Salinity | pH | Turbidity |
|---|---|---|---|---|
| Salinity | -0.10 | | | |
| pH | 0.02 | -0.07 | | |
| Turbidity | -0.15 | 0.36 | 0.22 | |
| Flow | -0.01 | -0.11 | 0.22 | -0.13 |

Across all species and models, the main effects were generally pronounced and significant ($P$<0.05). The exceptions were year, where the observed differences appeared to be adequately explained by the covariates; and the temperature covariate when in combination with the month factor, as these also tended to explain the same effect. Hence, year and temperature were dropped. Except for the pronounced 'river by location' interaction (the levels of these factors were then combined into a new 'site' factor), interactions tended to be of a lower order of magnitude, so were omitted. Hence, the final model included site and month as factors, flow as a linear covariate, and salinity, pH and turbidity as quadratics.

Firstly, a 'basic' general linear model was fitted to all data (including zeros), using the transformation ln(x+1). The more complex conditional models were fitted as follows.

It was assumed that the binary presence/absence of each species was adequately modeled using the Binomial distribution with a logit link. As each data point is a single trawl sample the dispersion parameter could not be estimated and, as usual, was taken as one.

Distributions used for the conditional (zero-truncated) data included the Gamma, log-Normal, Poisson and Negative Binomial. The latter two were fitted as both approximate (under standard GenStat models, which allow for the missing zero class), and correctly-truncated (via complex procedures). The suitability of each model can be judged by half-normal (or quantile) plots, which show how the model residuals align with the expected distribution. On this basis almost all the Poisson models were judged to be inappropriate.

Figure 2 shows two such plots. The residuals for the ln(x+1) analysis of all data at first appear quite reasonable. However, they fall outside the 95% simulated intervals for most of the low and high ends of the range, and it may be expected that this pattern will only get worse as the proportion of zeros in the data increases. Contrast this with the residuals from the log-Normal model of the truncated data – these match expectation throughout, and of course this pattern is unaffected by the proportion of zeros.

## 4. RESULTS

Despite the significance of the independent terms, the amount of variation explained by these models is relatively low (Table 3). Unfortunately, this is somewhat common with fisheries data, particularly when shorter sample times tend to result in 'hit or miss' counts with high variability.

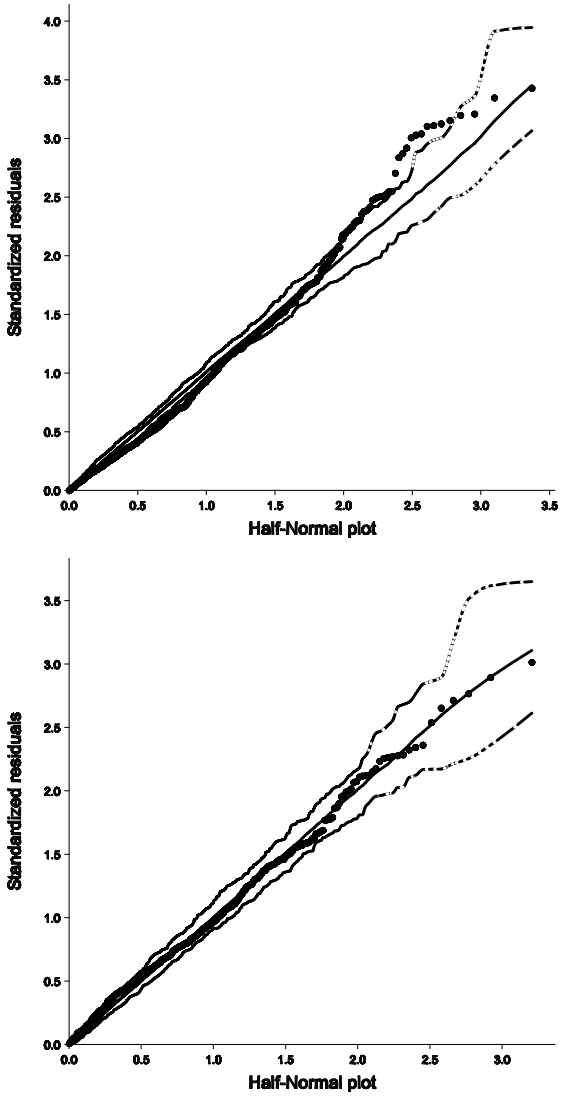


**Figure 2.** Half-normal plots, with 95% simulated envelopes, of residuals of *Thryssa hamiltoni* for a). ln(x+1) model on all data, and b). log-Normal model on truncated data.

**Table 3.** Degree of fit for Binomial model of binary data (percentage of deviance explained), and log-Normal model of the zero-truncated data (adjusted $R^2$, percent).

| | *T-ham.* | *P-kaak.* | *F-exq.* | *Val.* |
|---|---|---|---|---|
| Binomial | 27.4 | 7.5 | 27.3 | 15.8 |
| log-Normal | 33.0 | 8.6 | 11.9 | 21.4 |

Adjusted means and their standard errors were estimated for each model. For the ln(x+1) and zero-truncated log-Normal models, the bias-corrected back-transformation (Zhou and Gao 1997, Tu 2002) was used. We also used the adjustment for sample size of Kendall *et al.* (1983), as it is simpler than the adjustment listed in Pennington (1983). For the larger sample size of this study, these two produced similar results.

For all conditional models, the overall means for each level of each independent term were calculated as the Binomial proportion (of presence) multiplied by the conditional (on presence) mean. Standard errors for each mean were estimated, using the standard formula for the variance of the product of two random variables (Goodman 1960).

Within species, the fitted (or adjusted) means tended to form the same patterns for each independent term. Figure 3 shows two such terms for the most common species, *Thryssa hamiltoni.*

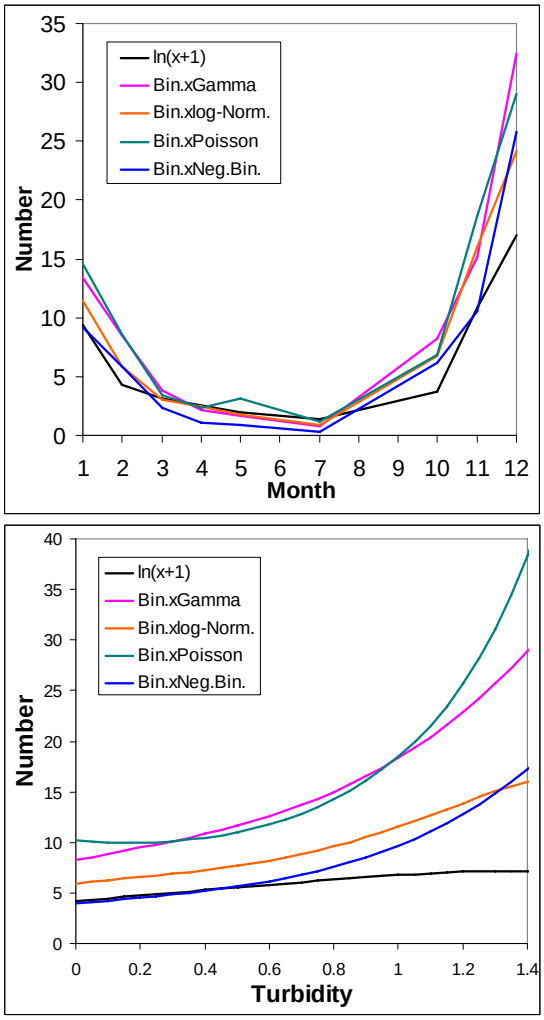


**Figure 3.** Adjusted means for *Thryssa hamiltoni*, for a). Month main effect, and b). Polynomial for observed turbidity levels.

Only the correctly zero-truncated versions of the Poisson and Negative Binomial are shown here. For the approximate models, the Negative Binomial tended to give quite acceptable results – and the parameter estimates proved to be most useful as initial values for the optimization procedure for the zero-truncated version (this always experienced convergence problems with anything other than good initial estimates). The back-transformed means of the approximate and the exact versions were quite similar for the Negative Binomial model.

In contrast, the approximate Poisson model performed poorly. The means were often 'well-removed' from the exact Poisson model, as well as the other models, and went into 'unbelievable' regions (given the observed mean catch rates). As indicated from the half-Normal plots (not presented here), the Poisson distribution appears to have insufficient skewness to accommodate short-sample fisheries data.

For the five models in Figure 3, the patterns across months appear quite similar, with the exception that the ln(x+1) model was less responsive and lower overall. This proved to be a common feature for this model, across all analyses. Across turbidity levels, there was good agreement between the Gamma and Poisson models, and also between the log-Normal and Negative Binomial models.

Figure 4 shows the adjusted means across pH levels, for *Favongobius exquisitus*. This data set displayed the least skewness (Table 1). Again, the 'flawed' ln(x+1) model showed the least responsiveness to increased pH levels. In practical terms, all of the conditional models displayed a similar response.

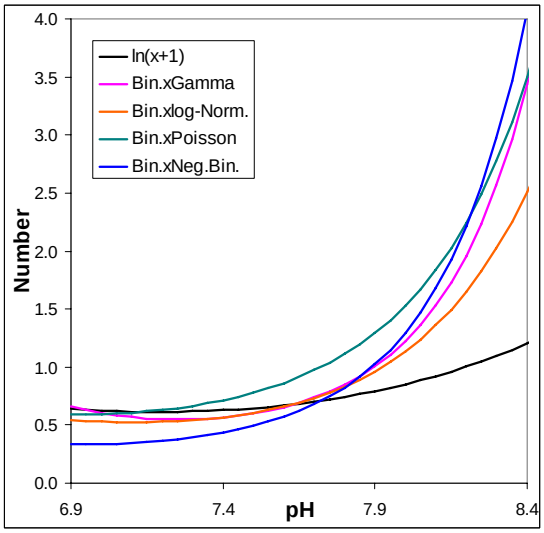


**Figure 4.** Adjusted means for *Favongobius exquisitus*, across observed pH levels.

These results, however, are in contrast to the *Pomadasys kaakan* means against salinity levels (Figure 5). Here, whilst a general increase with salinity is indicated, the shapes of the (possibly over-fitted) polynomials differ. The models tend to predict different ranges of catch numbers - in particular the Negative Binomial means are even lower than the ln(x+1) model. This was common across all the independent terms for this species, and may have been caused by its higher skewness (Table 1), or the lower degree of fit (Table 3).

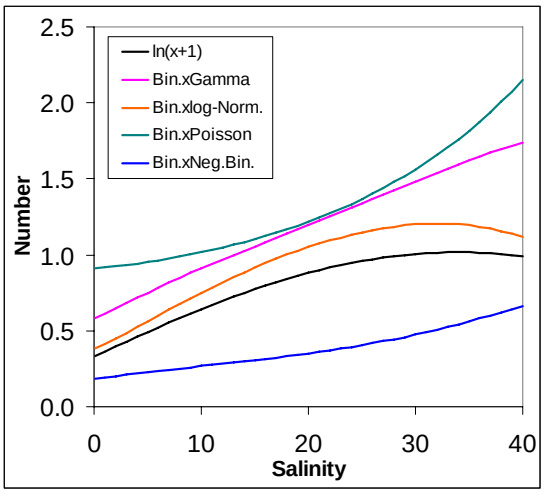


**Figure 5.** Adjusted means for *Pomadasys kaakan*, across observed salinity levels.

The adjusted means for *Valamugil* sp. produced yet another set of patterns (Figure 6). This species had the highest number of captured fish in a single sample (876), and the highest skewness, so the models had to accommodate this extreme tail. In Figure 6, the ln(x+1) is again unresponsive and low, the Poisson and the Gamma conditional models provide the more extreme fitted values, and the log-Normal and the Negative Binomial models indicate close agreement.

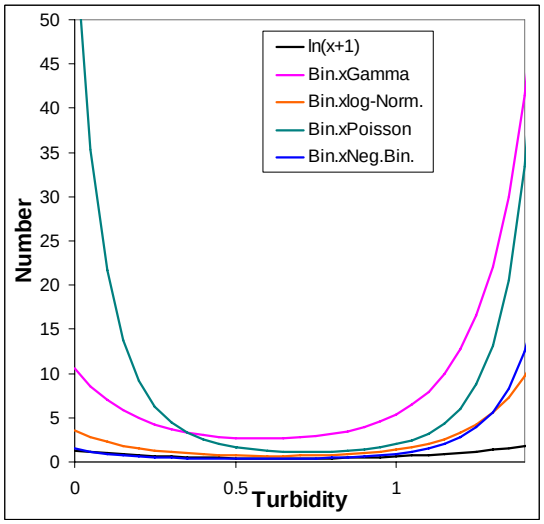


**Figure 6.** Adjusted means for *Valamugil* sp., across observed turbidity levels.

The issue of appropriate standard errors for these means is yet to be adequately addressed. Standardising these as a percentage of their respective means, some common patterns were observed. For an example, we take the site main effect for the Boyne River (this had the median number of observations per site, of 65). Across species, the standard errors for the 'basic' ln(x+1) model ranged between 10 and 36%, which given the low degree of model fits appears reasonable. For the gamma conditional model this range was 19 to 58%, and for the Negative Binomial, 19 to 60%. From our implementation of the formula in Pennington (1983) for the variance of the conditional log-Normal model, the range across species was 5 to 11%. These appear over-precise, especially when compared with the asymmetrical intervals for the log-Normal distribution recommended in Zhou and Gao (1997) - these ranged from 13 to 31% on the lower side, with upper values of 15 to 45%. These only apply to the zero-truncated portion of the data, i.e., cannot be incorporated with the Binomial component. We are thus unable to recommend an acceptable theoretical method for estimating the standard errors for the Binomial by log-Normal model, and are currently investigating bootstrap methods.

## 5. CONCLUSIONS

Fisheries data are frequently right-skewed, with a sizable proportion of zero values. No single distribution appears sufficiently versatile to accommodate this wide range of distributional shapes. Although the ln(x+1) transformation may appear adequate, for our data the residual plots and the general lack of responsiveness of the fitted means indicate this model performs poorly.

This leaves conditional distributions as the recommended statistical method. The Binomial appears an adequate representation of the binary (zero proportion) component, and there is a range of distributions available for the conditional (>0) catch data. Their suitability can be judged by the distribution of residuals, which needs to be considered separately for each species. However, this can only be determined after the final model has been decided, and the significance of the various competing model terms may well depend on the adequacy of the selected distribution. Hence this process will tend to be iterative, between model selection and distribution selection.

Based on residual plots and the fitted means, the truncated Poisson was shown to be generally inappropriate for our data. The truncated Gamma gave similar means – so whilst its residual plots appear acceptable, the means may be questionable. In three of our four species, the Negative Binomial and the log-Normal gave similar means, but for *Pomadasys kaakan* the Negative Binomial was notably below all other models.

Overall, the conditional (zero-truncated) log-Normal performed best. It gave good residual plots, and the resultant fitted means were generally responsive without ever going into 'unbelievable' regions. This is in agreement with Myers and Pepin (1990) that most zero-truncated fisheries

data sets will approximately conform to the log-Normal. However, one remaining problem is that we currently have no recommended method to estimate the standard errors for the overall means of this Binomial by log-Normal combination.

## 6. ACKNOWLEDGMENTS

This research is funded by the Fisheries Research and Development Corporation. We also thank our collaborators in the Coastal Zone CRC, and Michael O'Neill, Tony Swain and Christine Donnelly for their statistical codes and assistance.

## 7. REFERENCES

Aitchison, J. (1955), On the distribution of a positive random variable having a discrete probability mass at the origin, *Journal of the American Statistical Association*, 50, 901-908.

Bishop, J., W. N. Venables and Y-G. Wang (2004), Analysing commercial catch and effort data from a Penaeid trawl fishery. A comparison of linear models, mixed models, and generalised estimating equations approaches, *Fisheries Research*, 70, 179-193.

Dobbie, M. J. and A. H. Welsh (2001), Models for zero-inflated count data using the Neyman type A distribution, *Statistical Modelling*, 1, 65-80.

Faddy, M. J. (1997), Extended Poisson process modeling and analysis of count data, *Biometrical Journal*, 39, 431-440.

Feuerverger, A. (1979), On some methods of analysis for weather experiments, *Biometrika*, 66, 655-658.

Gardner, W., E. P. Mulvey and E. C. Shaw (1995), Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models, *Psychological Bulletin*, 118, 392-404.

GenStat (2005), GenStat for Windows, Release 8.1, Eighth Edition, VSN International Ltd., Oxford.

Goodman, L. A. (1960), On the exact variance of products, *Journal of the American Statistical Association*, 55, 708-713.

Haddon, M. (2001), Modelling and quantitative methods in fisheries, Chapman and Hall/CRC, 406 pp., London.

Kendall, M., A. Stuart and J. K. Ord (1983), The Advanced Theory of Statistics (Volume 3, 4th edition), Griffin, 780 pp., London.

Lambert, D. (1992), Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34, 1-14.

McCullagh, P. and J. A. Nelder (1989), Generalized Linear Models (2nd ed.), Chapman and Hall, 511 pp., London.

Myers, R. A. and P. Pepin (1990), The robustness of lognormal-based estimators of abundance, *Biometrics*, 46, 1185-1192.

O'Neill, M. F. and M. J. Faddy (2002), Analysis of recreational fish catches – Dealing with highly skewed distributions with many zeros, 3rd World Recreational Fishing Conference, 21-24 May 2002, Northern Territory, Australia, 67-69.

O'Neill, M. F. and M. J. Faddy (2003), Use of binary and truncated negative binomial modelling in the analysis of recreational catch data, *Fisheries Research*, 60, 471-477.

Pennington, M. (1983), Efficient estimators of abundance, for fish and plankton surveys, *Biometrics*, 39, 281-286.

Podlich, H. M., M. J. Faddy and G. K. Smyth (2002), A general approach to modeling and analysis of species abundance data with extra zeros, *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 324-334.

Tu, W. (2002), Zero-inflated data, in Encyclopedia of Environmetrics Vol. 4 (eds. A. H. El-Shaarawi and W. W. Piegorsch), Wiley, Chichester, 2387-2391.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly and D. B. Lindenmayer (1996), Modelling the abundance of rare species: statistical models for counts with extra zeros, *Ecological Modelling*, 88, 297-308.

Welsh, A. H., R. B. Cunningham and R. L. Chambers (2000), Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay, *Biometrics*, 56, 22-30.

Ye, Y., M. Al-Husaini and A. Al-Baz (2001), Use of generalized linear models to analyse catch rates having zero values: the Kuwait driftnet fishery, *Fisheries Research*, 53, 151-168.

Zhou, X-H. and S. Gao (1997), Confidence intervals for the log-Normal mean, *Statistics in Medicine* **16**, 783-790.

Zhuo, X-H. and W. Tu (2000), Confidence intervals for the mean of diagnostic test charge data containing zeros, *Biometrics*, 56, 1118-1125.