Gregory Walsh - ID 29785685 - gw2g17@soton.ac.uk - COMP6235 - Statistics Coursework

Analysis of Hypothetical Fishing Data

Abstract

This report presents an analysis of a set of fishing data. It describes the distributions of weights and catch times, analyses the relationship between the two variables, and investigates the best time to go fishing for a hypothetical scenario.

Methods and Results

Characterisation of the Data Set

Fig. 1 shows that the frequency density of catch weights is broadly distributed and possibly bimodal. Because of the non-normal distribution, the Freedman-Diaconis rule was used to calculate the bin width (0.62) since it scales with the IQR. A range of other bin widths were also tested, but none gave a significantly clearer picture of the distribution. It seems plausible that the water source could contain several species of fish, so the bandwidth of the kernel density estimate (KDE) was determined using a "solve-the-equation" method since it handles multimodal data well (Botev, Grotowski & Kroese - 1991).

Both the KDE and histogram for weight hint at a bimodal distribution, suggesting that the water source may contain two distinct groups of fish (perhaps different species) with different distributions.

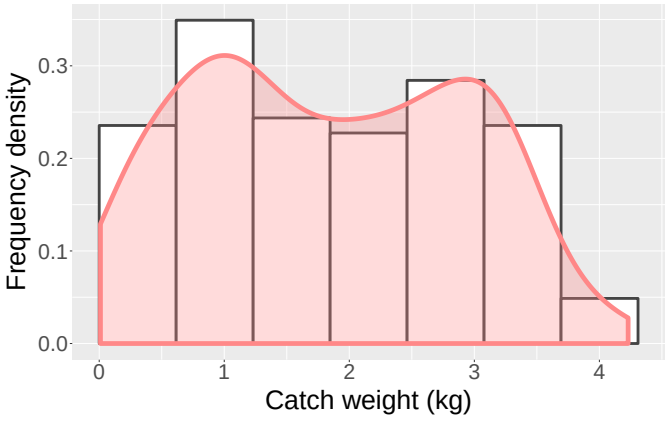

Figure 1: Histogram and KDE for weights.

The mean, median, and skewness values for the weights are 1.84 kg, 1.83 kg, and 0.09 respectively. With such similar mean and median values, and small skewness, we would expect the weights of new observations to fall above the mean as frequently as they do below. The sample variance is quite large (1.16 kg^2) due to the broad distribution of data.

A similar set of analyses were performed on the catch times (24hrs). Fig. 2, created using the same methodology as Fig. 1, shows the frequency of catch times are non-normally distributed.

The mean, median and skewness for the times are 9.39 hours, 8.95 hours, and 0.25 respectively. Since the mean is greater than the median, and there is a slight positive skewness, we expect more fish to be caught earlier in the day than later. As a consequence of the broad distribution, the variance of the catch times is quite large (32.0 hours^2).

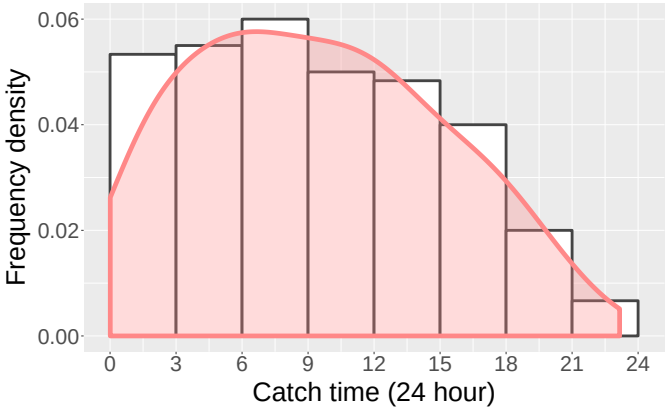

Figure 2: Histogram and KDE for catch times.

The sample based estimates of the population means for weight and catch times, with 95% certainty (critical value of 1.96), are 1.69 to 1.99 kg and 8.60 to 10.17 hours respectively. The relative precision of these estimates is almost identical.

Time and Weight Dependencies

From visual inspection of the points in Fig. 3, there is no obvious correlation between weight and catch time, which is supported by a small value for the Pearson's correlation coefficient ($r = -0.13$). Using r we can calculate r^2, and determine that just 1.6% of the variance in weight is explained by the catch time.

A two tailed test of significance for r was

carried out to determine the reliability of the value for r, with $\alpha = 0.05$ and null hypothesis $H_0 : r = 0$. A p-value of 0.07 was found, so H_0 is accepted and we conclude that weight is not linearly dependent on catch time. Equivalently, we see in Fig. 3 that the 95% confidence band for the regression line includes a set of lines with slope 0, i.e. cases with no linear dependence.

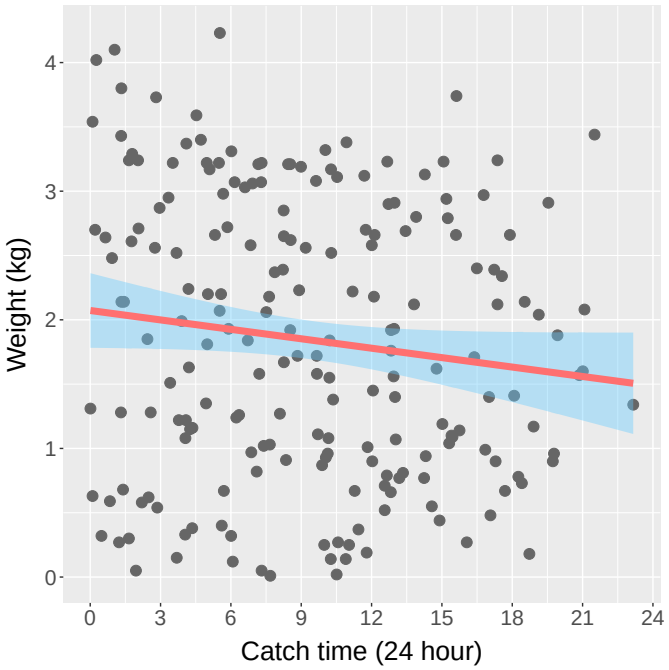

Figure 3: Catch time vs. weight with regression line and 95% confidence band.

What Time is Best to go Fishing?

There are many ways to interpret this question. For the purposes of this analysis, a scenario involving a commercial fisherman working an eight hour shift, who wants to maximise the total weight of fish he catches, was investigated.

In the analysis of this scenario, two key assumptions were made. First, that maximising the number of fish caught will also maximise the total weight of fish caught, since it has been shown that catch time and weight are linearly independent. Secondly, that the frequency density ought to be smooth over midnight, since there is no reason to believe fish make step-wise changes in their behaviour at this time, and so catch time should be treated as a circular quantity.

The analysis comprised several steps. First, a circular von-Mises KDE for catch times was calculated using the "circular" R package. The bandwidth was determined using a procedure which maximises the cross validation likelihood with respect to the bandwidth. The expected fraction of the total daily catch was then calculated for a set of eight hour periods, each beginning at one of 512 evenly spaced points around the clock, by computing the area beneath the KDE for each period. Fig. 4 shows a curve fitted through these values. We expect to catch the maximum amount of fish at the global maximum of this curve, which occurs just before 4am.

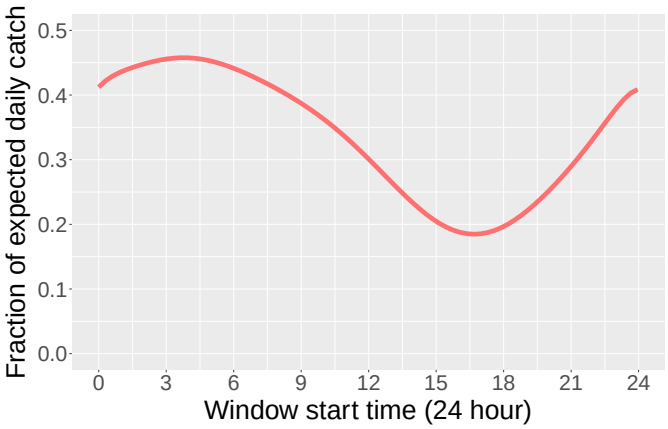

Figure 4: Estimated fraction of expected daily catch for all eight hour windows.

It should be noted that the value for the predicted best time will be influenced by the bandwidth, since the analysis is dependent on the position of the KDE's peak, which is directly dependent on the choice of bandwidth. For scenarios involving shorter periods, which are more sensitive to the locations of peaks, the bandwidth would be even more influential.

Discussion

There are a couple of areas where the analysis of the report could be extended. The confidence interval for the estimated mean catch time was calculated using the normal distribution to determine the critical values, however, since the distribution is not perfectly symmetric, a bootstrapping method may yield a more accurate result.

Also, the estimate of the best time to go fishing would be more useful if combined with a confidence interval. E.g. if the confidence interval is large, it would have commercial implications, and therefore it may be prudent to spend more money tracking catch data to get a better estimate.