Statistics Coursework

Alun Meredith

November 5, 2015

1 Reading data

Downloaded fish.txt, data about the catch of a hypothetical fishing fleet from:
"http://www.edshare.soton.ac.uk/view/courses/COMP6235/2015.html" on Thu Nov 05 10:13:37 2015.

2 Visualising Data

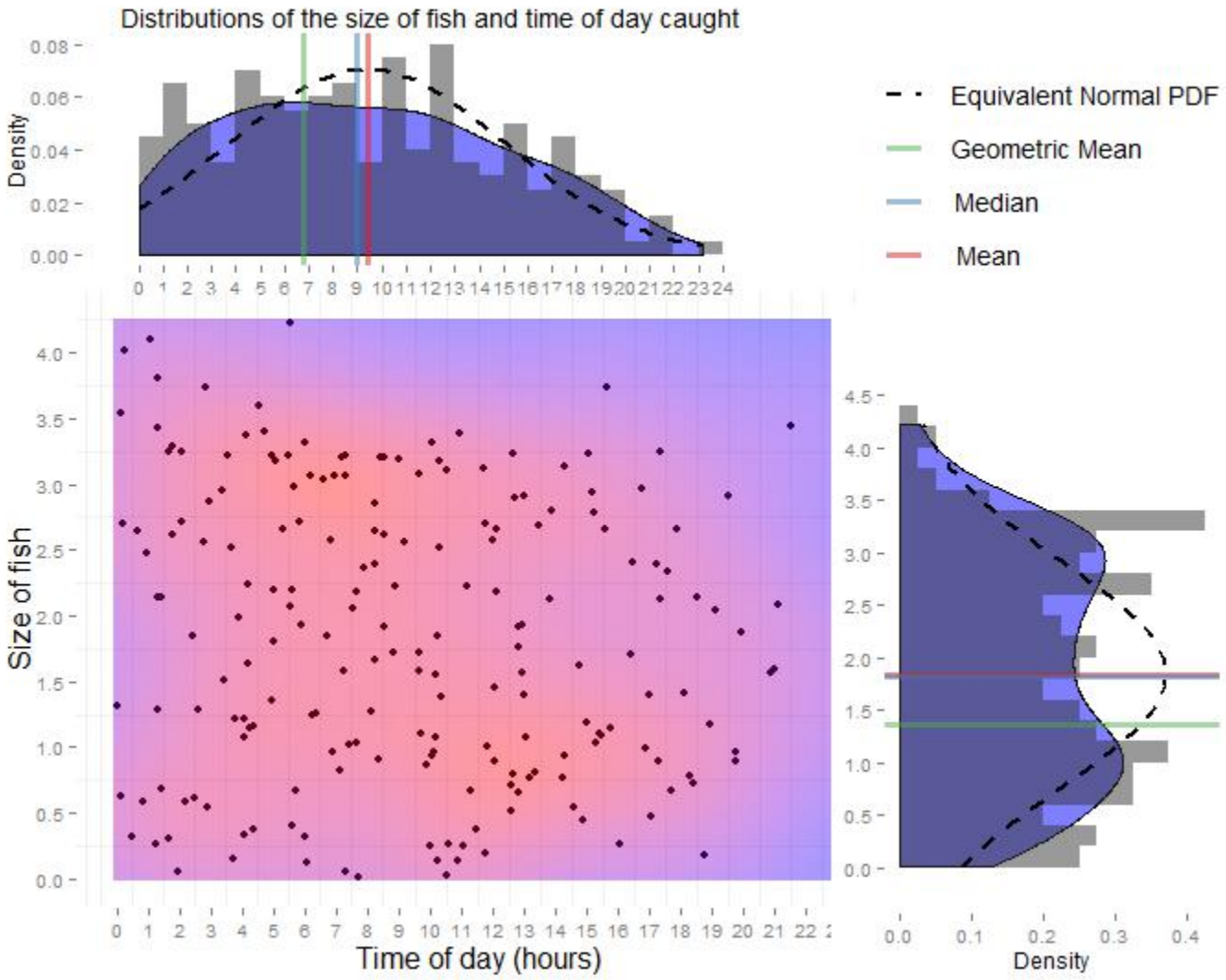

Figure 1: ***Bottom Left****: Scatter graph of size of fish caught vs time of day (hours), coloured background demonstrating gaussian kernel 2d density* ***Right****: Histogram showing distribution of size of fish caught and gaussian kernel density* ***Top****: Histogram showing time of day fish caught overlayed by gaussian kernel density*

Table 1: Numeric summary statistics for Time and Size Distributions

	min	Q1	median	Q3	max	mean	s.dev	var	skewness	kurtosis	geometric mean
sizeSummary	0.01	0.94	1.83	2.74	4.23	1.84	1.08	1.16	0.09	-1.17	1.37
timesSummary	0.01	4.88	8.95	13.22	23.16	9.39	5.66	0.25	-0.89	6.79	0.01

3 Analysing Distributions

By looking at the above summary statistics and the density distributions in figure 1, we can see some interesting features of the distributions. The time fish were caught was very broard with high standard deviation. There is some skewness which can be seen clearly in the plot. This is fairly surprising as you would expect the density at 12:01 to be approximately equal to the density at 11:59 in a cyclical manner. This suggests that an effect such as the data being recorded on a Saturday where no fishing is done on a Sunday is occuring, but requires more domain knowledge to analyse fully.

The distribution of fish sizes (right hand density plot in figure 1) shows a distribution which is visibly bimodal. The standard deviation covers 51.18% of the data range, which is greater than the wide flat times data of 48.9, this is probably due to the bimodal distributions. Both the mean and median sits centrally between the two peaks with low skewness because the peaks are approximately symmetric and equal in size.

3.1 Confidence Intervals

Both of the distributions are not normally distributed, the normal distributions in fig 1 doesn't accurately describe the observations. This introduces some complexity in producing confidence intervals. For the time caught confidence interval it is possible to closely approximate the normal distribution by transforming the data, e.g. applying an exponent of approximately 0.79 removes the bias.

In order to evaluate a confidence interval from these non-standard distributions a Bootstrapping method is used. As shown in fig.3 by resampling our data and plotting the means we get a distribution that closely approximates the normal distribution via the central limit theorem. The mean of a sample distribution approximates the mean of the population it was sampled from. By calculating a confidence interval for this mean we can estimate a 95% confidence interval for the population of:[1]

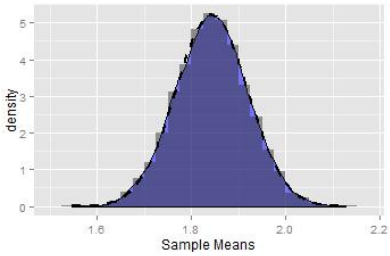

Figure 2: 10000 sample means taken from fish size data, blue region is density, dashed line shows normal distribution with equal mean and variance.

SizeConfidence	1.69	1.99
TimeConfidence	8.61	10.17

Table 2: Bootstrapped confidence intervals for the mean (95 percent)

3.2 Codependency and Time intervals

Visually Figure 1 (bottom left), showing a scatter plot of time of fish vs size of day demonstrates no obvious correlation between time of catch and size of catch. The covariance between the variables is -0.78, this is difficult to interpret so we use a Pearson product-moment correlation yielding, -0.13 (0.95 % CI: -0.26 to 0.01).

The correlation is very weak (0.1-0.3), and the 95% interval overlaps 0 so there is not enough evidence to claim a relationship but one explanation of weak correlation could be that you would expect the bigger fish to be easier to catch, therefore the population of bigger fish in the water shrinks over the course of the day. However considering the size of the space fishing boats typically fish in any effect like this is likely neglibible.

We can also use a linear regression model yields $y = -0.02x + 2.07$, with a p value of 0.704 which is typically too high to reject the null hypothesis. Similarly the 95% confidence intervals overlap 0.

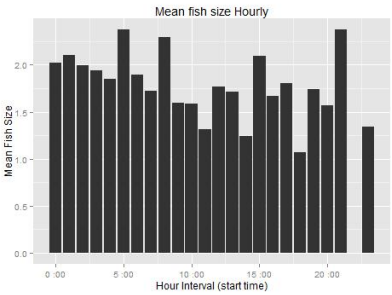

Figure 3: Mean size of fish caught Hourly

Splitting the times into 24 intevals corresponding to each hour of the day we can show the interval of time with the highest rate of catch is 12:00 to 13:00. The interval with the largest average size of fish caught is 21:00 to 22:00.

[1] At 3 significant figures the students t test yields the same confidence intervals, for applications where accuracy is only required at this level the students t test or normal approximation is sufficient.