Statistics Coursework

Alun Meredith

October 30, 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 Fri Oct 30 23:31:39 2015.

```
> readLines("fish.txt", 5)
```

```
[1] "15.25 2.79 " "13.45 2.69 " "5.58 2.20 "  "7.17 3.21 "  "13.81 2.12 "
```

From reading the first 5 lines of the file we can see that the data is in 2 numeric columns. Each variable is seperated by a single whitespace, there is no header and no apparent NA characters or comment/escape characters. As such we can use the default values of read.table, including the col.names argument to name the variables.

2 Visualising Data

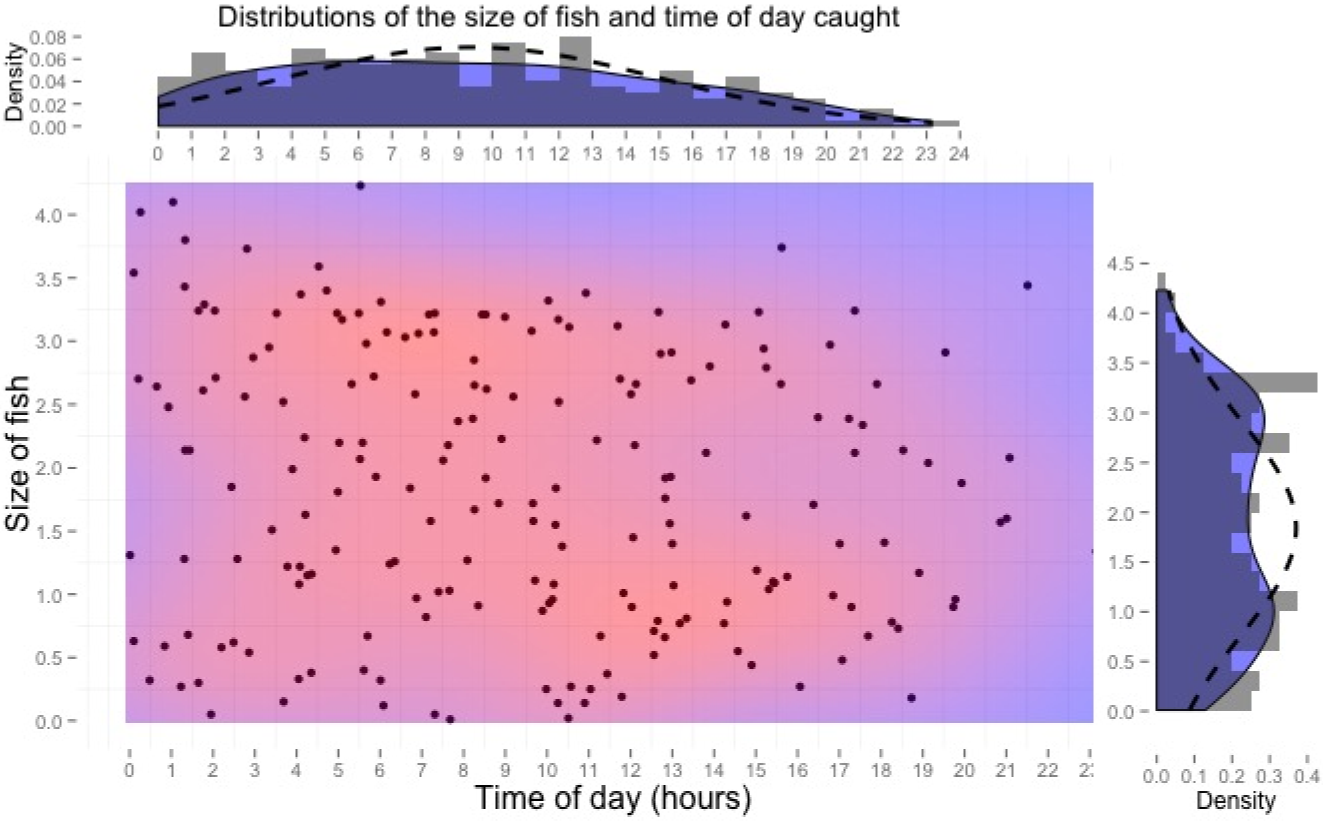

*Figure 1: **Bottom Left**: Scatter graph of size of fish caught vs time of day (hours), coloured background demonstrating gaussian kernel 2d density **Right**: Histogram showing distribution of size of fish caught with gaussian kernel density plot overlayed **Top**: Histogram showing time of day fish caught overlayed by density plot, dashed line shows normal distributions for equal mean and standard deviation*

Table 1: Numeric summary statistics for Time and Size Distributions

	min	Q1	median	Q3	max	mean	s.dev	skewness	kurtosis	geometric mean
sizeSummary	0.01	0.94	1.83	2.74	4.23	1.84	1.08	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

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 between the two peaks and there is negligible skewness because the peaks are approximately symmetric and equal in size.

3.1 Confidence Intervals

Both of the distributions are clearly not normally distributed as shown by the dashed normal distributions (with same mean/variance) shown in the figure. 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.

In order to evaluate a confidence interval from these non-standard distributions a method called Bootstrapping is used. As shown in fig.2 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:

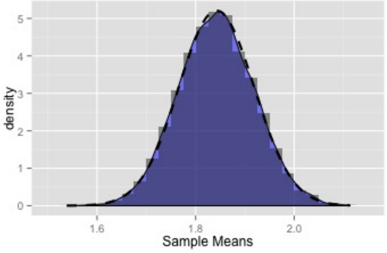

Figure 2: 10000 sample means taken from fish size data, blue region is density, dashed line shows normal distribution with equal mean and variance.

SizeConfidence	1.69	1.99
TimeConfidence	8.61	10.17

Table 2: Bootstrapped 95 percent confidence intervals

4 Codependency

The central figure above shows a scatter of size of fish and time of day. It is hard to identify any dependencies in the data visually but it looks as though it may be slightly weakly correlated. After numerical analysis we can see a covariance of -0.78 and a Pearson product-moment correlation of -0.13.

The correlation is very weak but one explanation 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. We would need more domain knowledge to conclude an effect like this. For instance if the fishing is taking place over a larger area (larger population of fish) then the fish caught early in the day will have quite a small effect on the ammount left at the end of the day.

5 Intervals

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.

Although this period has the largest average size of fish caught, it is not what would be initially expected considering the negative correlation calcualted. If each interval is inspected individually this apears to be an outlier. The general trend is that until 8.00 the average is around 2 and after 12 is almsot entirely below 2 (with just 2 exceptions).