University of Southampton

Msc Cyber Security

Foundation of Data Science

Statistics with R

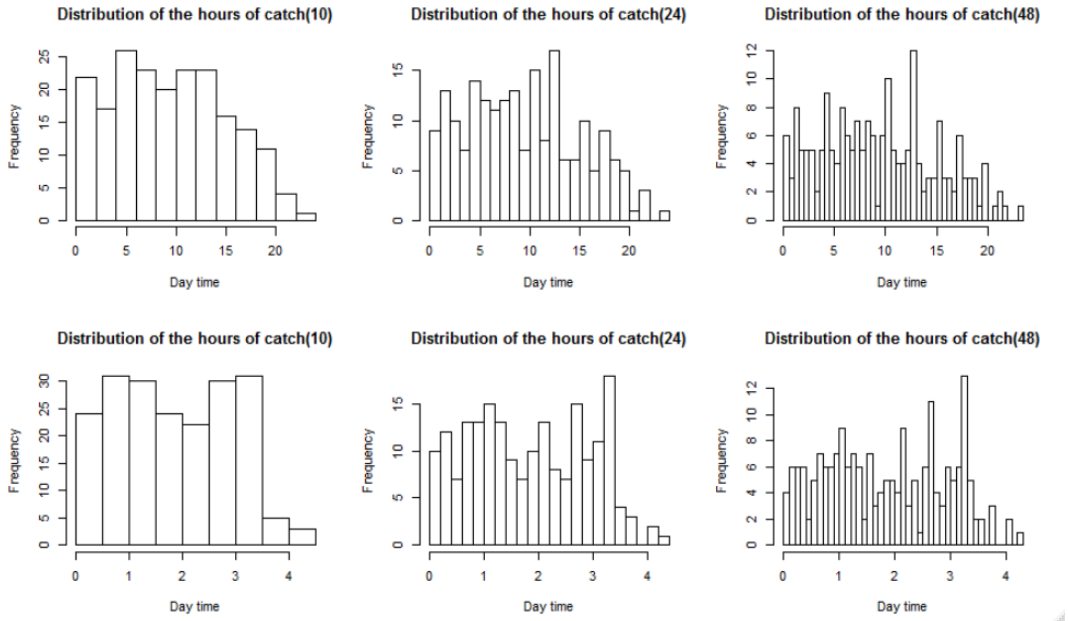
Author:
Gerard Tio Nogueras

Supervisors:
Pr. Elena Simperl
Dr. Chris Phethean
Dr. Ramine Tinati
Dr. Markus Brede

December 10, 2016

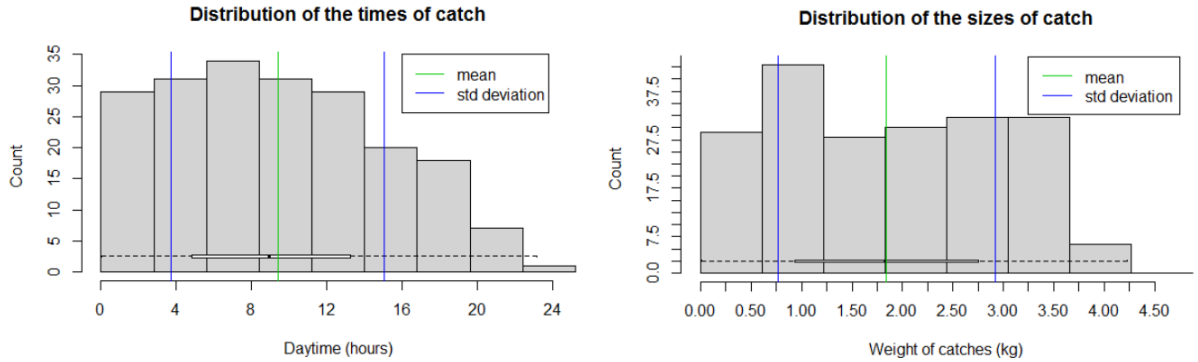
1 First step: Distributions

Different bin sizes:



1.1 Typical Scores

Freedman-Diaconis rule for bin sizes which gives us 2.8h for the times and 0.61kg for the weights.



Times		**Weights**	
mean	9.388	mean	1.8431
median	8.950	median	1.8250
mode	1.65	mode	3.22
geometrical mean	6.786465	geometrical mean	1.36948

We can see for both distributions that the means and the medians are almost equal this shows that each side of the median has the same total counts. The mode and geometrical mean are given as general information but I considered them irrelevant for this distributions because there are almost no repetitions in the dataset and the geometrical mean is mostly used to normalize different scales.[Mathsteacher, 2000][Geometric mean, 2016]

1.2 Range of Scores

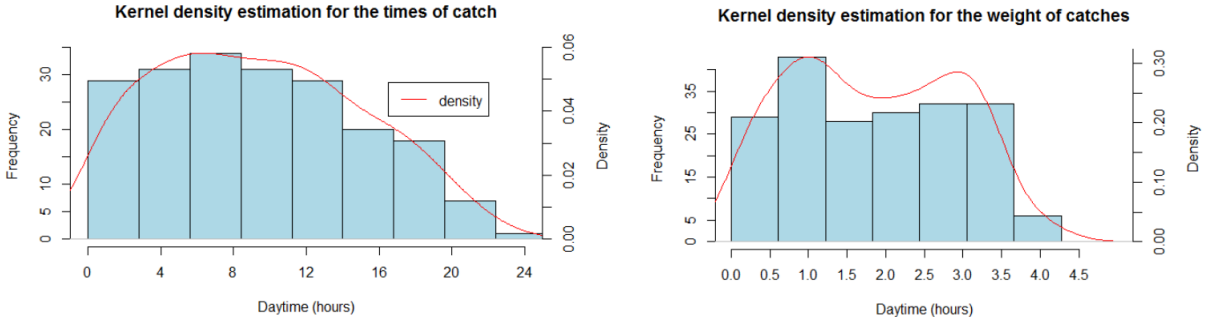
Times		**Weights**	
minimum	0.010	minimum	0.0100
maximum	23.160	maximum	4.2300
variance	31.99831	variance	1.162248
standard deviation	5.656705	standard deviation	1.078076
interquartile	range 8.335	interquartile	range 1.8
skewness	0.2477339	skewness	0.08966
kurtosis	-0.8870124	kurtosis	-1.169334

The IQR is smaller then the standard deviation range, this shows that the values between the first and the third quartile deviate less from the mean than the rest of the values.
For the time distribution, combining this with the fact that the mean and median are clearly on the left side of the center of 24h scope we expect similar counts between the standard deviation range and low counts for the right edge of the distribution. For the weight distribution, combining the previous information with the mean and median values we expect a more uniformed distribution, which is the case with the exception of higher counts between 0.6kg and

1.25kg. This is balanced by the extreme right edge of the histogram which is really low.[Standard deviation, 2016]

1.3 Kernel density estimation

KDE allows for a smoother estimate of the distribution then the histogram. Normally this is used to extrapolate the PDF for a larger population.[Kernel density estimation, 2016] [Probability density function, 2016]

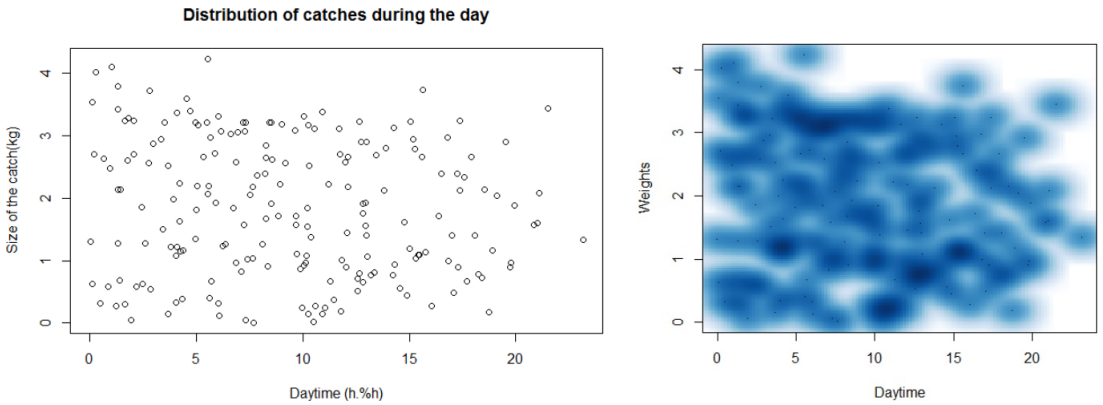


1.4 Mean values with 95% confidence intervals for both distributions

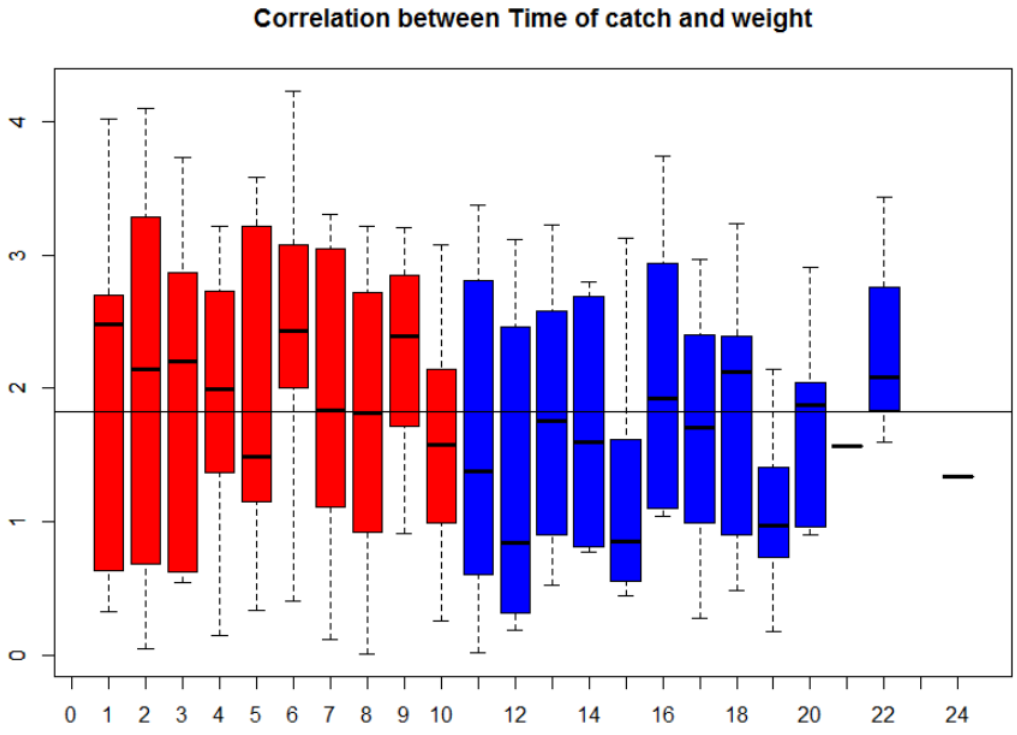
The formula that will give us the confidence interval for population mean $= \bar{x} \pm z * s/\sqrt{n}$ with z being the z value for 95% confidence interval, s the standard deviation of the sample, n the sample size and $\bar{x}$ the sample mean. The confidence interval of the mean for the "times" is $[8.604371, 10.17233]$ and for the weights it is $[1.693636, 1.992464]$ [M. Lane, no date][Confidence intervals for means, 2016]

2 Second step: Correlation

There isn't any apparent correlation, even when smooth scattering the distribution. After this approach we end up with a static snow distribution(no correlation).



By regrouping the data by data frames of hours we can start retrieving interesting information. The median values before 10am are generally higher then the median of weights compared to the values after 10am which tend to have medians under weights median. In conclusion, the fishermen should go fish in the morning before 10am if they wish to obtain big catches.



The intervals with the lowest average rate of catching is d and with the highest average is d.

References

[Geometric mean, 2016] Geometric mean (2016) in Wikipedia. Available at: https://en.wikipedia.org/wiki/Geometric_mean (Accessed: 10 December 2016).

[Mathsteacher, 2000] Mathsteacher (2000) Mean, median and mode. Available at: http://www.mathsteacher.com.au/year8/ch17_stat/02_mean/mean.htm (Accessed: 10 December 2016).

[Standard deviation, 2016] Standard deviation (2016) in Wikipedia. Available at: https://en.wikipedia.org/wiki/Standard_deviation (Accessed: 10 December 2016).

[Kernel density estimation, 2016] Kernel density estimation (2016) in Wikipedia. Available at: https://en.wikipedia.org/wiki/Kernel_density_estimation (Accessed: 10 December 2016).

[Probability density function, 2016] Probability density function (2016) in Wikipedia. Available at: https://en.wikipedia.org/wiki/Probability_density_function (Accessed: 10 December 2016).

[M. Lane, no date] M. Lane, D. (no date) Confidence interval for the mean. Available at: http://onlinestatbook.com/2/estimation/mean.html (Accessed: 10 December 2016).

[Confidence intervals for means, 2016] 7.5 - confidence intervals for means (2016) Available at: https://onlinecourses.science.psu.edu/stat200/node/49 (Accessed: 10 December 2016).