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 4, 2016

1 First step

We start by plotting the distribution of the times of catch and the distribution of the weights of the fishes caught. Let us look at the information we can retrieve from these distributions following the lectures recommendations. One of them is to find the good bin sizes using the Freedman-Diaconis rule which gives us 2.8h for the times and 0.61kg for the weights.

1.1 Typical Scores

We start with the mean, the median, the mode and finally the geometrical mean. **Times**

- mean 9.388
- median 8.950
- mode 1.65
- geometrical mean 6.786465

Weights

- mean 1.8431
- median 1.8250
- mode 3.22
- geometrical mean 1.36948

1.2 Range of Scores

Here we will look at the range of the distributions, their spread and how they spread. **Times**

- minimum 0.010
- maximum 23.160
- variance 31.99831
- standard deviation 5.656705
- interquartile range 8.335
- skewness 0.2477339
    1. old
    2. new
- kurtosis -0.8870124

Weights

- minimum 0.0100
- maximum 4.2300
- variance 1.162248
- standard deviation 1.078076
- interquartile range 1.8
- skewness 0.08966
    1. old
    2. new
- kurtosis -1.169334

1.3 Kernel density estimation

Here we plot a continuous estimation of the PDF (probability density function) for our variables. This allows for a smoother estimate of the distribution then the histogram. Normally this is used to extrapolate the PDF for a larger population. Strangely for the weights distribution the sum of the KDE values exceeds 1 by far which seems a bit odd since it is supposed to estimate a PDF. On the other hand the "times" KDE seems perfectly normal

2 Second step