Reports of coursework 1
Statistics with R

YULIN LI - 30526515 - yl8n18@soton.ac.uk

Abstract

This report focuses on using some static method to analyze the data of fishing, which is consist of three variables (weight, time, baits). In this reports, it analyzes the weight, time distribution and gives mean values with 95% confidence intervals for both distributions. Moreover, it analyzes the relationship between weight and time. Also, it solves the question of the best time for fishing, the effectiveness of each bait, the best baits to use at 3 pm. All the plots in this report are plotted by R.

Analysis

According to the dataset, we can get the histogram of x (weight) distribution and y (time) distribution, as figure 2.1:

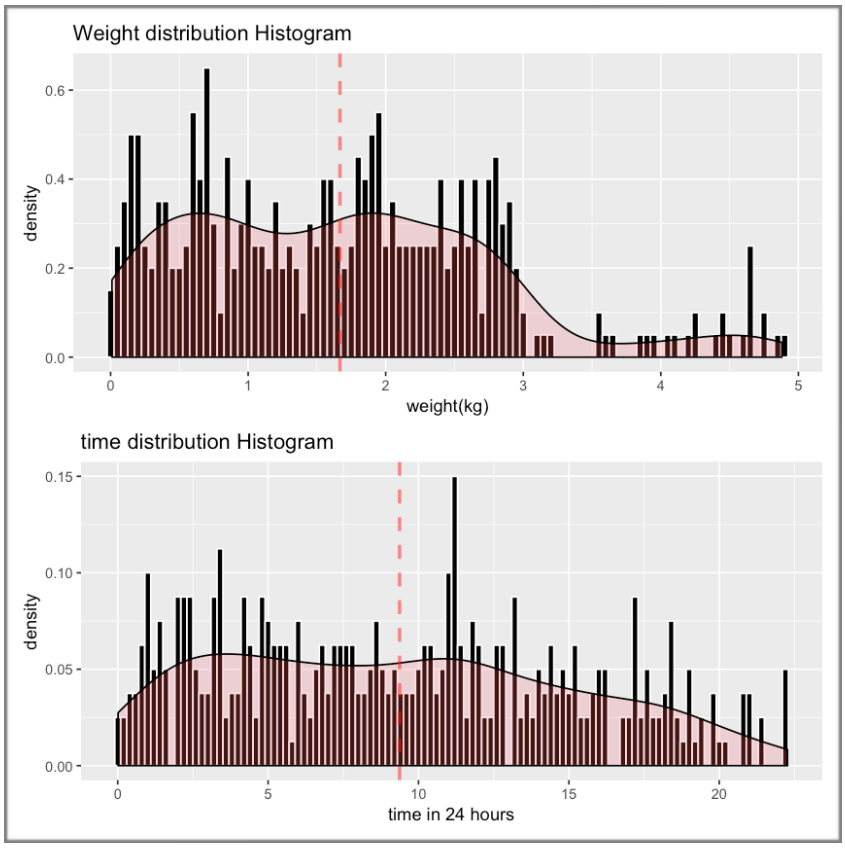

Figure 2.1 histogram of weight and time distribution

In figure 2.1, the red line represents the mean value for both distributions. The mean time to catch a fish is 9.371, and the average weight of fish is 1.667. The weight ranges from 0.01 to 4.88 with a standard deviation of 1.1081 which means that the weight fluctuates steadily. Moreover, the mean of the time is 9.371, the standard deviation of time is 5.796, which can be explained that most of the time can the fish be caught. The Assuming that if the data are a sample from a large population, we can get the means values with 95%

confidence intervals for x, y distribution. For the time distribution, it ranges from 8.8 to 9.94. For the weight distribution, it ranges from 1.558 to 1.776.

According to the distributions of X values (times of catch), Y values (size of catch), we plot a scatter plot for three different baits, as the figure 2.2:

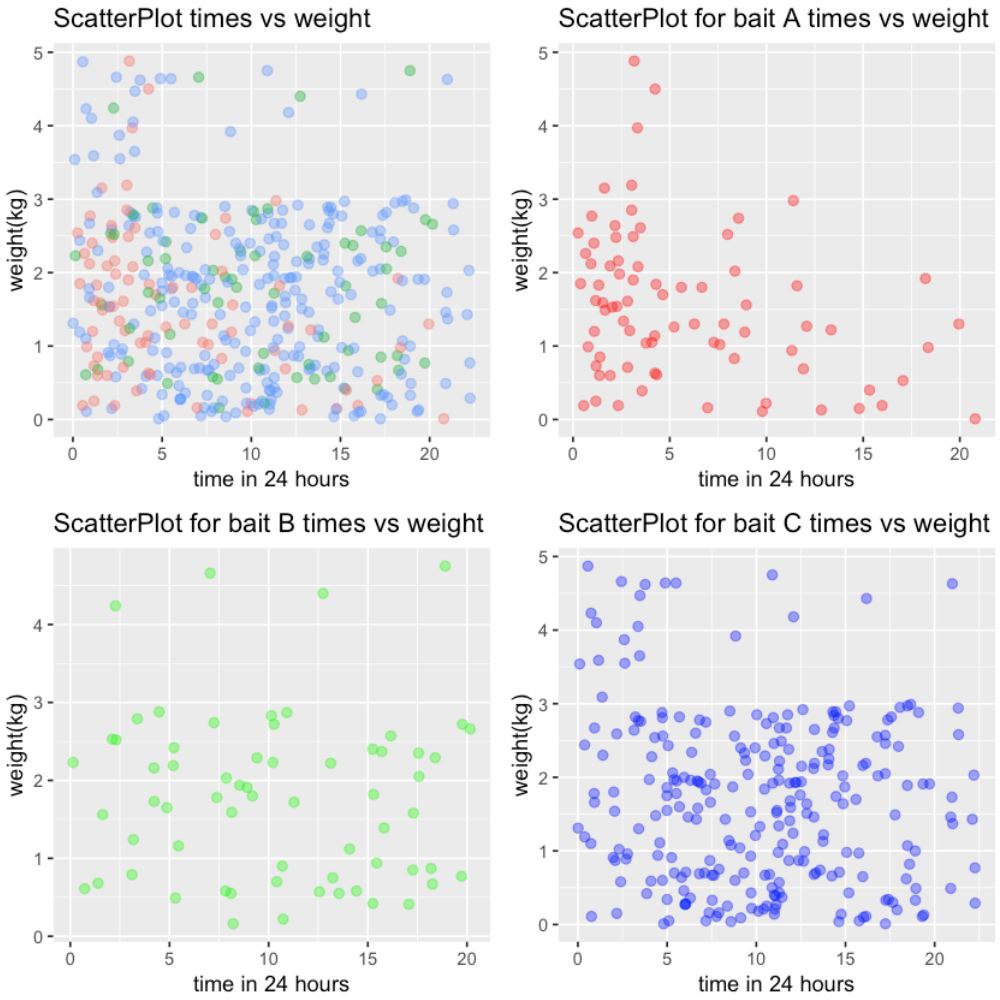

Figure 2.2 the scatterplot of time and weight distributions for 3 baits

Figure 2.2 applies that, the weight frequency of three different baits which can show the efficiency of three baits. The bait C gets the most number of fishes which is higher than bait A and bait B. However; the scatter plot shows that the data is non-linear. We use Spearman's rho method to try to find out the correlation between rank of x, y values. We find that the correlation coefficient is -0.108. Also, I use Pearson's method; the correlation coefficient is -0.121. So we can infer that the x value is weak relative to y value.

3. Question

- **What is the best time to go fishing at this lake?**

  To find out the best time to go fishing at this lake, it is necessary to decide a way to judge that what means the best time. I define the hour the fisherman catch the heaviest amount of weight of fish means the best time for fishing. Then we can get figure 3.1:

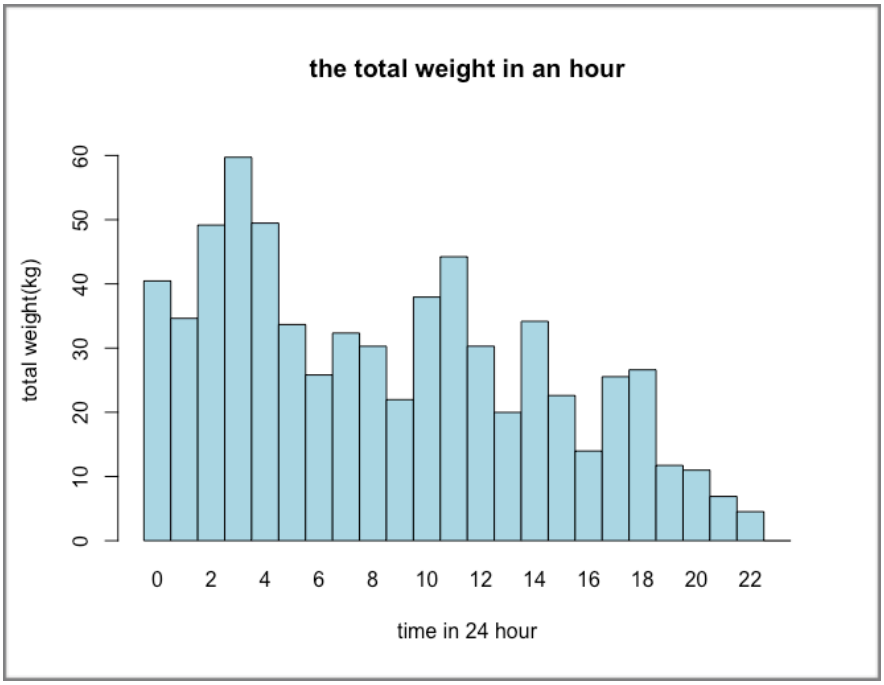

Figure 3.1 the total weight distribution

Figure 3.1 shows that the total weight distribution of the fish in an hour. It is easily found that the fisherman catches nearly 60 kilograms of fishes from 3 to 4 which is the highest number during the total weights in 24 hours. So we can infer that 3-4 is the best time to go fishing in this lake.

- **Which bait is the most effective?**

To find out which bait is the most effective, it is necessary to define what is effective. We think that the bait can catch the most number of fishes and the sum of weight which is also the highest mean the most effective. According to figure 3.2 and figure 2.2, it is easily found that bait C is the most effective. Because The bait C catches 257 fishes totally and the total weight of C is 432.017 kilograms, the bait B catches 64 fishes totally, and the total weight is 119.528. The bait A catches 79 fishes, and the total weight is 120.791. It can infer that bait C is the most effective one.

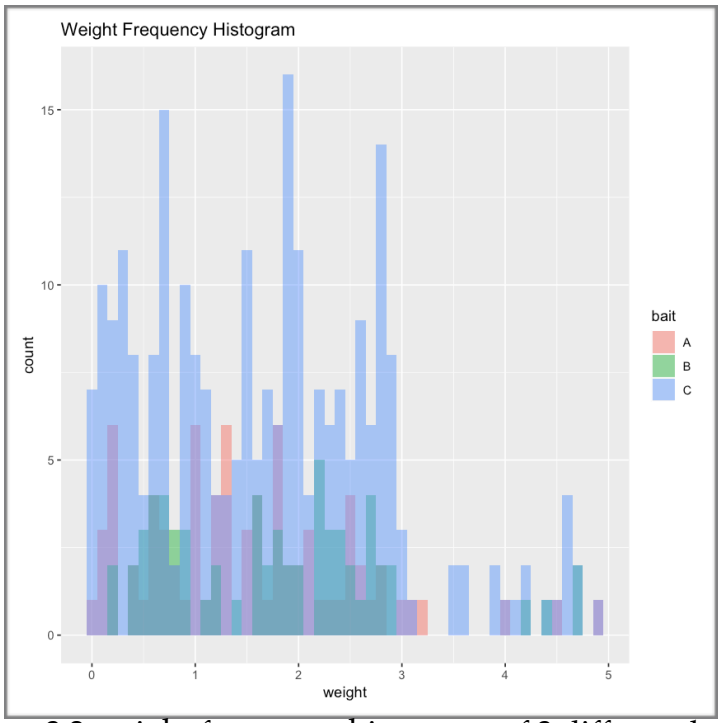

Figure 3.2 weight frequency histogram of 3 different baits

- **What is the best type of bait to use at 3pm in the afternoon?**

To find out the best type of bait to use at 3 p.m., we define a period from 15:00 to 15:20. If we catch the fish in this period, we think that it is caught by the bait the fisherman used at 3 pm. With this definition, we convert 15:20 into y value to get 15.33. So we need to find out the distribution of weight with 3 different baits, and y value is limited between 15 and 15.33. According to the data, we get the figure 3.3 as follow:

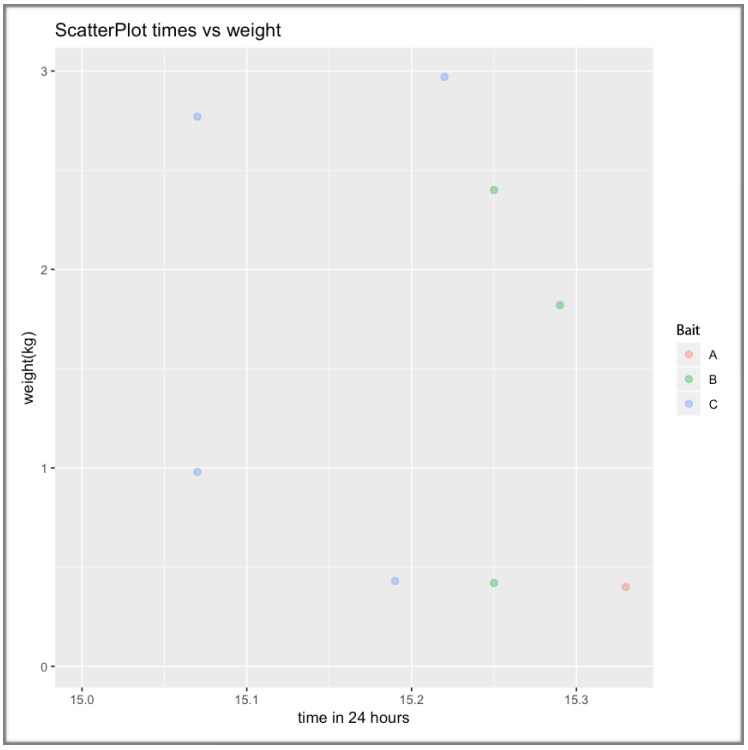

Figure 3.3 scatter plot of 3 baits from 15:00 to 15.20

Figure 3.3 applies that the time and weight for 3 different baits from 15:00 to 15:20. There are only 8 data points in this period. Moreover, it shows that there are 4 fishes are caught by bait c which is more than any other two baits. Also, the average weight of the fishes caught by bait C is 1.895; bait B is 1.547, bait A is 0.4. So we can infer that bait C is the best type to use at 3 pm.

4. Conclusion

In this report, we can analyze the performance of three different baits. According to the distribution of time and weight of three baits, it can infer that bait C has the highest efficiency among the three baits. Moreover, at 3 p.m., the bait c is also the best bait to use for fishing. Also, 3 a.m. is the best time for fishing for the whole day. What's more, the time of catching and the weight of fishes have a weak correlation.