COMP6235(20/21): Foundations of Data Science (MSc) Coursework 1 Report

Yan Zhang (yz10u20@soton.ac.uk)

***Abstract*—A method of analyzing the efficiency of different bait and the best time to go fishing is developed based on the statistical knowledge on the skewed data. In this report, the data were sorted on the time of fishing, then a new column was constructed, which indicates the time per catch. After the stage of data preprocessing, this report used the total sum of weights in every hour and the median value of time per catch in each hour to measure the efficiency of baits during the 24 hours. In the last stage, several visualizations of data demonstrated the relationship between time of catch, size of catch and the type of bait. It is found that there is a poor linear relationship between time of catch and size of the catch, but both variables of time of catch and size of the catch are correlated with the type of bait.**

Introduction

Fishing is an excellent recreational activity for people to enjoy a peaceful day. Nevertheless, sometimes the wrong choice of types of baits or size of the hook may cause little harvest but take a long time to wait. Different kinds of baits play different roles; some may be more attractive to bigger fish, and some may be good for smaller fish. It is significant to find out how different types of bait affect the harvest. Besides, fishing at a different time of a day also leads to disparity in efficiency. For instance, fishing early in the morning is better than fishing in the afternoon, because fishes will be foraging in the morning and are easy to be hooked.

Fishing efficiency is an essential determinant of the usefulness of a rod using different baits that may have a marked impact on increasing the gain of a day's fishing. Using a low-efficiency bait is likely to waste much time waiting for the hook. The objectives of this report are to find the relation between time that people go fishing and the efficiency of fishing, and which kind of bait is the most efficient. The results are described in this report

Methodology

The data were recorded from a period of a day during which a fisherman has fished in the lake. Considering the fisherman used three different baits, the data were filtered into three groups by the type of bait, and then sorted on time of catch(X). It is hard to distinguish which bait is more efficient by observing the original data. The variable time of catch(X) was used to calculate the time per catch, the speed of catching a fish, by getting the difference between two adjacent X samples, storing it as variable T. After transforming the data, it is found that at some point, two fishes were caught at the same time, resulting in $T = 0$. This report adopted the method of replacing the value zero on the k^{th} row of the column T with the value on the $(k-1)^{th}$ row of the same column.

This report used two metrics to measure the efficiency of the bait. The first one is the sum of the weight of fishes caught in every hour,

$$EW_{i,j} = \sum_{k}^{n_{i,j}} w_{k,i,j} \tag{1}$$

where $i = 1, 2, ..., 24, j = A, B, C$, $n_{i,j}$ is the number of fishes caught in the i^{th} hour using bait j, $w_{i,j}$ is the weight of k^{th} fish caught in the i^{th} hour using bait j.

The second approach is using the median of time per catch in each hour, as the median is the best measure of centrality for an asymmetric distribution.

$$ET_{i,j} = Median(W_{i,j}) \tag{2}$$

where $i = 1, 2, ..., 24, j = A, B, C$, $W_{i,j}$ is a list of each time length of catching fishes in the i^{th} hour using bait j. In such a method, the bait with the highest $EW_{i,j}$ and the smallest $ET_{i,j}$ will be the most efficient.

Skewness is a measure of the direction and degree of lopsidedness in a specific distribution, which would be zero when the distribution is symmetric. Carrying out an analysis of skewness is essential because it is associated with how to measure the centrality and spread of the distribution. When the distribution is symmetrical, such as normal distribution, the mean, the middle of the data, is the best description for the centrality.

Since both distributions of X and Y are skewed, thus the mean is not appropriate for measuring the centrality. Therefore, a better measure of the centre for the asymmetric distribution would be the median of it. As for the measure of spread, the standard deviation is a widely used measure of spread. Nevertheless, the standard deviation is only suitable for symmetric distribution. To measure an asymmetric distribution like X or Y, the inter-quantile range(IQR) is the best choice, because it is unaffected by extreme values or outliers. IQR could be calculated by $IQR = Q3 - Q1$, the distance between the 25th and 75th percentiles of the data.

The Central Limit Theorem(CLT) implies that the distribution of the sample means can be regarded as normally distributed regardless of the distribution of the original data when the size of the sample is large enough(n>400). When the sample size is large enough, the mean of sample data can represent the mean of the overall population. Therefore, calculating the mean values with confidence intervals is essential to estimate the mean of the overall population. Having the assumption that the data are a sample from a larger population, and the sample is large enough, the mean and standard deviation could be calculated, and the 95% confidence interval could be constructed as:

$$\left(\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) \tag{3}$$

where $\overline{x}$ is the mean, s is the standard deviation, and $t_{\alpha/2,n-1}$ is the Student's t-distribution score(t-score) with the degrees of freedom = $n - 1$ and $\alpha = 1 - C(C = 0.95)$.

It is important to check if there is a linear relationship between X and Y, and also enjoyable to find whether the increase of

X results in an increase of Y. This report used the Pearson coefficient to measure the linear relationship between X and Y,

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{4}$$

If there is a linear relationship between X and Y, the absolute value of the outcome of (4) will be above 0. In detail, the closer it is to one, the more significant the linear relationship is. There is no linear relationship if the outcome of (4) is zero.

This report used a violin plot to compare the distribution of quantitative data across several levels of categorical variables. One good point about the violin plot is that it shapes the density function of the distribution, and it shows how the data of different groups varies concerning the categorical variable.

Results & Discussion

The data was divided into three data sets according to the type of bait(column Z). It was then used to calculate the speed of catching after the data were sorted by X, saved as variable T.

TABLE I: Measures of centrality, spread, and suitable additional measures

Measurements	X	Y
Count	400.000000	400.00000
Mean	9.370525	1.66740
Std	5.796400	1.10816
Min	0.010000	0.01000
25%	4.325000	0.70750
50%	9.020000	1.61500
75%	13.747500	2.40000
IQR	9.4225	1.69
Max	22.270000	4.88000
Skewness	0.266856	0.653793
Kurtosis	-0.946594	0.161891

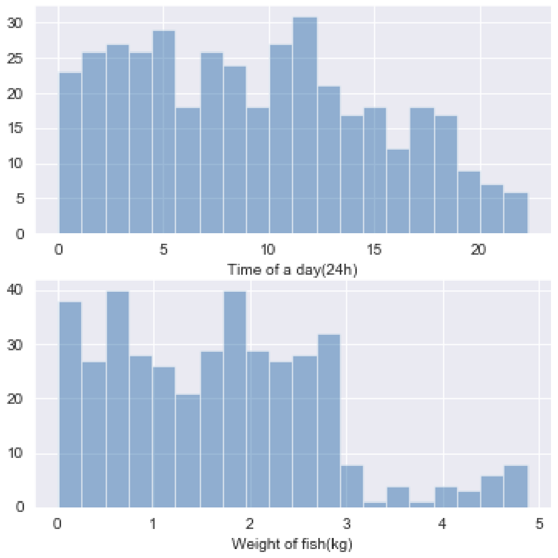

Fig. 1: Distributions of X and Y.

Histograms(Fig.1) were plotted to illustrates the distributions of variables X and Y. From the figure, it is evident that both distributions are positive skew, with long right tails and heavy left tails. As is shown in Fig.1, Y has a sharper peak than X, with its kurtosis larger than that of Y. The median is more suitable for measuring the centrality instead of mean in an asymmetric distribution. IQR is a useful metric for measuring the spread of the asymmetric data rather than the standard deviation. The results of all suitable measures about the distribution are in Fig.1. As shown in Fig.2, the distribution of variable T is heavily positive skew, and the median is only the way of measuring centrality. In Fig.3, there are two ways to compare the effectiveness of types of bait, the upper graph is the sum of weights of fishes caught in every hour, and bait C is more compelling all the time compared to bait A and B. The second graph in Fig.3 is the median value of time per catch in each hour, which compares different baits by the speed of catching a fish. With the assumption that the data are a sample from a larger population, the 95% confidence could be calculated,

$$CI_X = (8.807, 9.940), CI_Y = (1.558, 1.776) \tag{5}$$

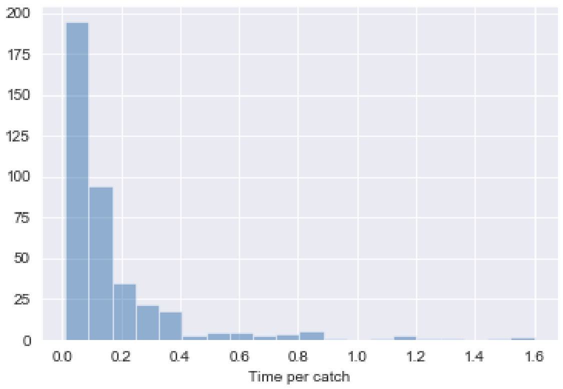

Fig. 2: Distribution of time per catch(T)

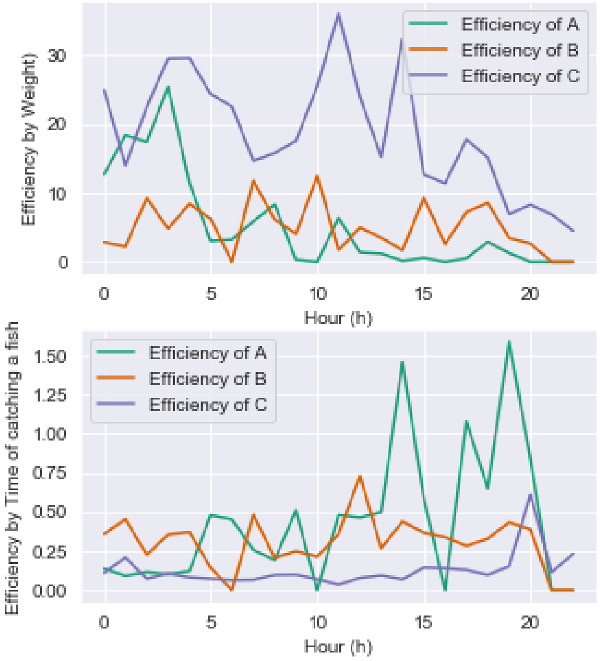

Fig. 3: Efficiency of each bait by weight and time

This report used Pearson correlation to measure the monotonic relationship between two variables. It carried out a statistical hypothesis test to determine whether or not the correlation coefficient is significantly different from zero. If the P-value is below 0.05, the null hypothesis(non-linear relation) is rejected [1]. The results($\rho_{X,Y} = -0.12, p{-}value = 0.0158$) show that there is a weak negative linear relationship between X and Y. and a p-value below 0.05, indicating that the results may be statistically significant and the linear relationship exists. Even though it is a weak relationship, the results could be significant with a large sample.

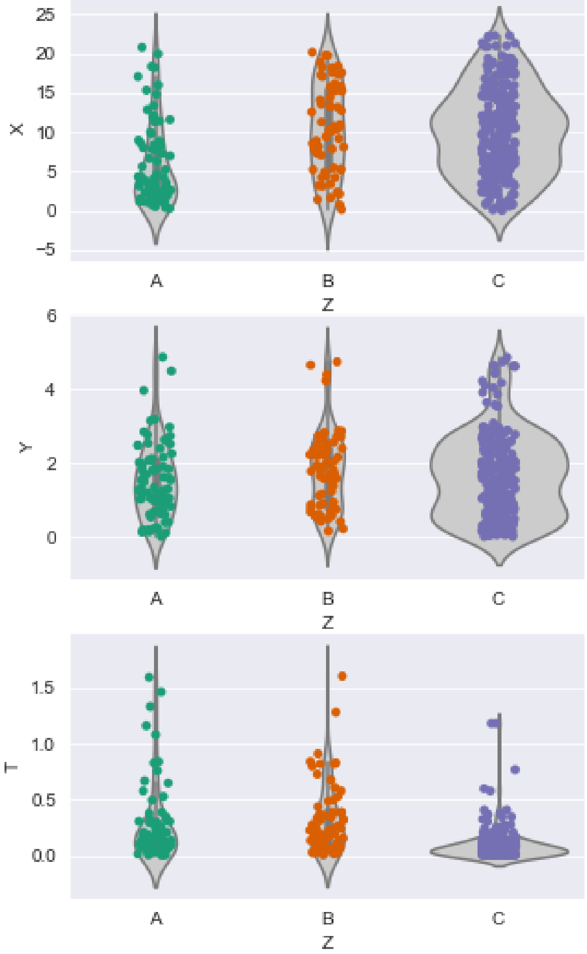

Fig. 4: Violin plot w.r.t. X, Y, T, Z

According to the line plot of efficiency of each bait, the best time to go fishing is 2 am to 5 am using both bait C or A, and 10 am to 12 am using bait C. Afternoon, the efficiency of fishing begins to decrease. Therefore, the best time for fishing is in the morning, as early as possible. In terms of the most efficient bait, bait C is the most efficient one because it could be used to catch more fishes every hour compared to other baits. Also, using bait C, the fisherman caught the most fishes in the time of 24h(in Fig. 3). At 3 pm, considering which bait is more useful, it is evident that bait C still has a higher efficiency than the other two baits.

According to the violin plot shown in Fig.4, the upper graph is how X varies w.r.t different types of bait. As shown in this graph, bait A caught more fishes before 5 am than other periods, bait B seems to have a stable performance during all day but did not catch too many fishes, and bait C could catch much more fishes than bait A and B, where the peak performance was at around noon. The graph at the middle shows the distribution of the size of catch across different types of bait. Bait C was able to catch more heavy fishes than bait A and B, while bait A and B perform similarly. The graph at the bottom of Fig.4 shows the distributions of time per catch according to different types of bait. Bait C has the fastest speed of fishing and is the most efficient.

Conclusion

The systematic analysis above generated practical results on what is the primary determinant of the efficiency of fishing, utilizing several statistical methods. A relationship between the time of fishing and the efficiency was found that fishing early always results in higher efficiency and fishing late after 3 pm leads to low efficiency. Therefore, a general decision of time of fishing should be made before going to fishing, which is that people should go fishing as early as possible, especially before 6 am or at the noon when the efficiency reaches the peak.

In term of the choice bait, this report carried out an exploratory data analysis on the efficiency of baits. It is found that bait C is the most efficient one among all other baits, as more massive fishes were caught using bait C during the same period compared with bait A and B. Moreover, bait C has the lowest value of median time per catch(T) in each hour, which means the time of waiting for a fish to be hooked is shorter using bait C.

Although the general efficiency of fishing is decreasing in the afternoon, the efficiency of bait C is still higher than the other two baits, as illustrated in the figure above. In conclusion, if people decide to go fishing, the best time for this recreational activity is before noon, and the recommended bait is bait C.

References

[1] K. H. Zou, K. Tuncali, and S. G. Silverman, “Correlation and simple linear regression,” *Radiology*, vol. 227, no. 3, pp. 617–628, 2003. PMID: 12773666.

Appendix

Code of the report stored in Github

https://github.com/zhangyan560/Data-science-.git.