Coursework1 for Foundation of Data Science

Han Zhang hz5g21@soton.ac.uk

Abstract

Knowing how to fish successfully is a common wish for fishermen. The research focuses on discovering how to fish more successfully and efficiently. By analyzing the data in one day, the relationship between X(time of catch), Y(size of catch), and Z(type of baits) has been identified. The histogram is employed to briefly describe the distribution of X and Y, followed by the Pearson correlation coefficient to analyze deeper correlation and Mahalanobis distance to detect outliers. The results show that the relationship between X and Y is not significant and the bait B is the most effictive in fishing.

1 Introduction

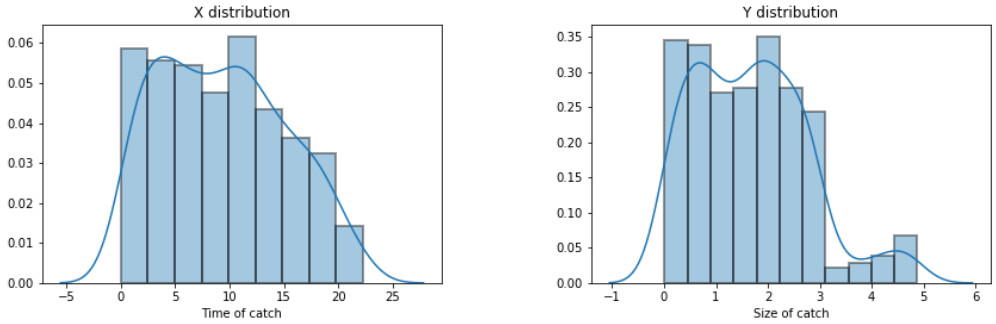
Do you know how to fish effectively? Data analysis is critical and can help us with this problem. To enhance fishing efficiency, this article examines the data about the Time of Catch, Size of Catch, and Type of baits. The following sections are arranged in the following order: Section 2 will include a literature review of several technologies about data analysis. The results are discussed in Chapter 3. Section 4 will provide the ultimate conclusion.

2 Literature Review

In 2001, The central limit theory helps to calculate the confidence interval[1].It is proposed a method in 2004 for detecting multivariate outliers,which identifying outliers is defined by the deviation measure between the empirical distribution function of Mahalanobis distance and the theoretical distribution function[2]. After that, the method is implemented in 2020 and detect outliers through python[4]. In 2009, the article proposed guidelines for interpreting the correlation coefficient, r and coefficient of determination, R^2 [3]. In 1989, the author improved ANOVA method by pairwise comparison methods which can analyze the relationship between numerical and catrgory numbers [5].

3 Discussion

3.1 Basic distribution



(a) X histogram (b) Y histogram

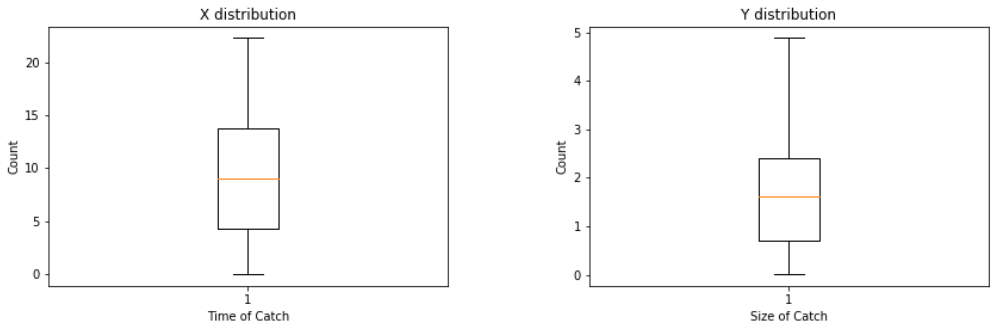
Figure 1: X and Y distribution

To analysis the data, it's important to observe their distribution. The histogram and the sketched curve are demonstrated in Figure 1. The Freedman-Diaconis rule is used to choose suitable the number of bars of X, Y distribution, which is 9 and 11 respectively.

Distribution	Mean	Geo-mean	Mode	SD	Skewness	Kurtosis
X	9.37	6.84	0.93	5.79	0.27	2.05
Y	1.67	1.18	0.11	1.10	0.65	3.16

Table 1: Basic Description

The statistical descriptions including mean, geometric mean, mode, standard deviation, skewness, and kurtosis are shown in Table 1. The skewness is bigger than 0 which means that the sketched curve is skewed to left, besides, the kurtosis is smaller than 3 which indicates that the sketched curve is flatter than the normal distribution. So the sketched curve of X is skewed to the left and flatter than the normal distribution, and the sketched curve of Y is skewed to the left and shaper than the normal distribution.



(a) X box plot (b) Y box plot

Figure 2: Outliers Detection

It is also to identify whether there are outliers in distribution. The box plot in Figure 2 shows that no outlier exists.

From a personal perspective, the effectiveness of bait should be the size of the catch divided by the number of catches. The results come out that the B bait is the most effective in catching fish as shown in Figure 3.

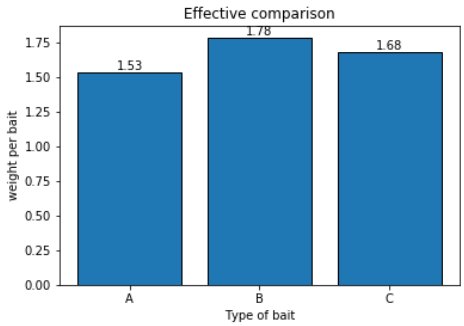
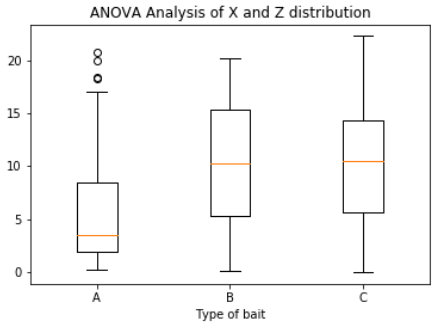


Figure 3: Effectiveness plot for bait A,B,C

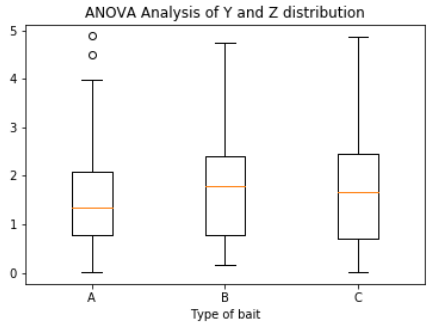
For the confidence interval, according to Central Limit theory[1], The mean of a sample selected from a large population should be a normal distribution if these samples are chosen from a very large population. The results are that the mean value of X is between 8.20 and 9.94, the mean value of Y is between 1.56 and 1.78 with 95% confidence.

3.2 Deeper correlation

For analyzing deeper correlation, the method is to compare two of them(X, Y, Z). Firstly, to analyze dependence between X and Y, the Pearson correlation coefficient is used and the coefficient between them is -0.12 which means that they are nearly independent with each other[3].



(a) ANOVA analysis between X and Z



(b) ANOVA analysis between Y and Z

Figure 4: ANOVA Analysis

For analyzing dependence between X and Z, Y and Z, the ANOVA was adopted to analyze their dependence[5]. The results show in Figure 4. Z are remarkably distributed differently on X and are not significantly distributed differently on Y as P-Value between X and Z is 5.75×10^{-9} which is smaller than 0.05 and P-Value between Y and Z is 0.37 which is better than 0.05. The means that X is dependent with Z while Y is independent with Z[5].

The results can be proofed by Figure 5. It can be observed that Z spread difference in X, to be more specific, A lies around 5 am while B, C lies around 10 am but nearly has the same spread in Y, A, B, C lies around 1.5 kg.

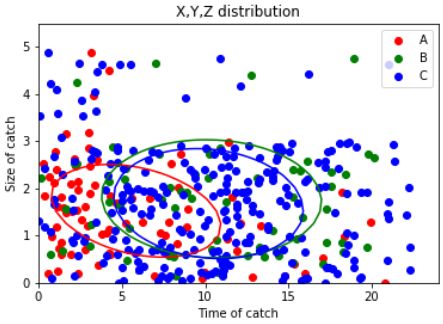
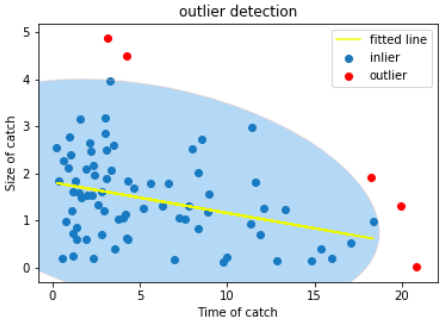
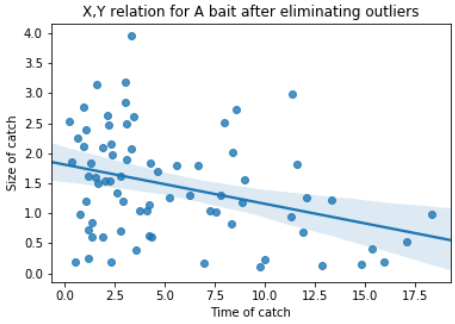


Figure 5: Distribution for bait A,B,C

To analysis the the correlation between X and Y, it is sensible to calculate the correlation coefficient. Besides, three baits are divided to observe the relationship between X and Y to eliminate the influence of different baits.



(a) Outliers elimination



(b) Linear Regression after eliminating outliers

Figure 6: Correlation between X and Y for bait A

In beginning, for bait A, the coefficient between X and Y is only -0.32. Figure 4 shows that there are some outliers. so Eliminating outliers could be a suitable choice. The Mahalanobis distance is applied to solve this problem and Chi-Square distribution provides the cutoff value since they have some correlation[4].In the end,the points outside the 0.95 (two-tailed) will be considered as outliers. There are five outliers and

the results are shown in Figure 6. After removing these outliers,the coefficient is -0.34. As for the bait of B,C, the correlation between X and Y is -0.05 and -0.12 respectively.

The amount of information about Y that is given by knowledge of X is R^2 in [3]. So X can explain 0.1156 of Y for bait A, 0.0144 of Y for bait B, and 0.0025 of Y for bait C.

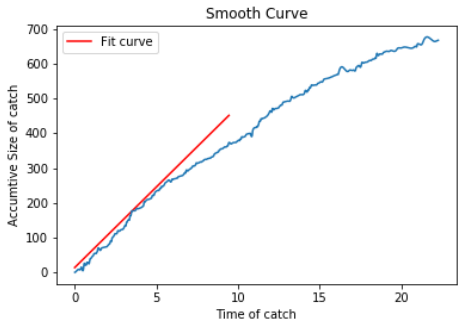
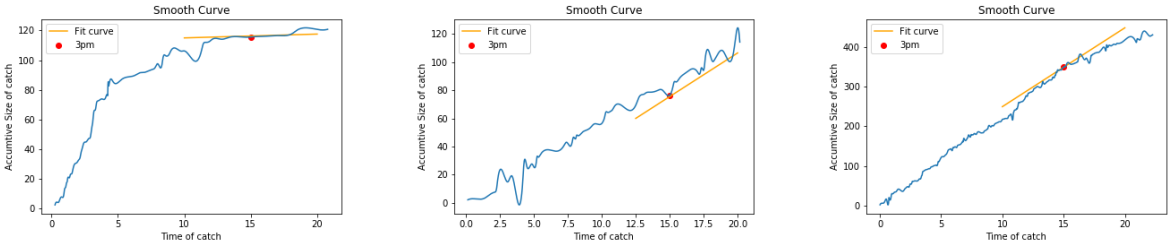


Figure 7: The slope of best time to go fishing

To find the best time to go fishing, the accumulative Y, that is the total size of catch in this time point, is a vertical line and X is a horizontal line. The slope of the graph is the speed of the size of the catch. So the steepest tangent line occurs at 3.15 with 52.18. The time slice is 1000 and δh is 0.001. So the best time to go fishing is 3 hours and 9 minutes in the morning.

As mentioned above, bait B is most effective with 1.78 wights per bait shown in figure 3.



(a) grown speed for A (b) grown speed for B (c) grown speed for C

Figure 8: The slope for A,B,C

Using the same method to calculate the slope of the bait A, B, C at 3 pm, the results are 0.82, 8.96, 44.42 for A,B,C individually. So bait C should be advised to be used at 3 pm. The slope graphs are shown in Figure 8.

4 Conclusion

This study examines one day's fishing data using a range of statistical analysis methodologies. Through different statistical tools, the report explained the distribution of data, the connection between the data, and the best time to go fishing at a certain moment, and at a certain moment, the best bait should be used. These all helped fishermen to fish effectively.

References

[1] István Berkes and Endre Csáki. A universal result in almost sure central limit theory. *Stochastic Processes and Their Applications*, 94(1):105–134, 2001.

[2] Peter Filzmoser, C Reimann, and RG Garrett. *A multivariate outlier detection method.* Citeseer, 2004.

[3] Bruce Ratner. The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, 17(2):139–142, 2009.

[4] Sergen. Multivariate outlier detection in python. https://towardsdatascience.com/multivariate-outlier-detection-in-python-e946cfc843b3, 2020.

[5] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.