COMP6235(2021/2022) Foundation of Data Science — Coursework 1 report

Name : Teeraphon Issaranuluk — Email : ti1n21@soton.ac.uk

Abstract

The report investigates sufficient time and bait type for fishing by using a sample of fishing data in a single day. Using the existing library, for example, Pandas, Matplot, NumPy, Seaborn, to manipulate or visualise the sample data to demonstrate the insight of sample data. Consequently, the statistic method and related graph are used to answer the question of the best time for fishing and bait, which is suitable for fishing in either general time or a specific time.

1 Introduction

Nowadays, data visualisation is the most powerful tool for revealing the hiding knowledge from raw data. The key factor which includes and take an essential role behind these tools is a statistic. In order to make a reasonable assumption and measure the accuracy of statistic value from sample data, several statistic methods are taken into account for data analysis. By use the provided dataset on fishing record in one day, this work tries to answer the question of how to identify the best time for fishing, the most effective bait for fishing, and which bait must be used in some specific time range. This report focuses on analysing the distribution and correlation between the existing variables. Precisely, the discussed question tries to find the best time to go fishing, the most effective bait and the suitable bait to use at 3.00 p.m.

2 Dataset exploration

The fish1 dataset represents a sample of fishing data of one day period. This dataset consists of three columns: time, size of fish in kilogram, and type of fishing rods used to catch the fish at a specific time labelled as A, B and C, respectively. In this report, each variable might use X, Y, and Z to represent the dataset column; X stands for time, Y is a fish weight that catches in a specific time record, and Z is rods with different baits types that use to catch a fish in each time.

2.1 Distribution of sample data

the dataset's characteristic is initially represented by the "describe" function, as shown in figure 1. In addition, Figures 2 and 3 illustrate the distribution of numeric variables, which included Time and Fish weight, and demonstrate the dataset's spread and centrality. The median of two features (9.02 and 1.61) is shown as a verticle line in bow space, whereas the 25 and 75 percentile are simulated as the box left and right limit. Moreover, the minimum and maximum values are present as the end tails on both sides in the plotted graph.As mentioned in the graph, both variables seem skewed to the right as graphical observation.

	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
max	22.270000	4.88000

Figure 1: Statistic values generated by describe() function

Significantly, the skewness value is calculated through the skew() function. The skewness value of the time sample is approximately 0.26. This number indicates that the sample was oversampled in the morning (around 0 to 5 a.m.) and undersampled in the evening (around 7 to 11 p.m.). In addition, the weight sample graph is also dramatically skewed to the right, which is highlighted by nearly 0.65 skewness.

This observation shows that both set of samples are non-normally distributed. Hence, in this report, non-parametric methods should be utilized to analyze the data [3].

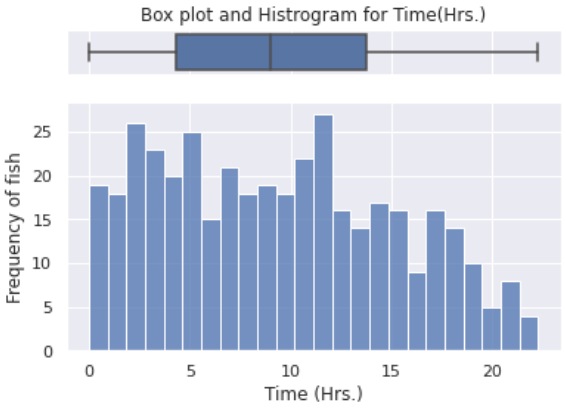

Figure 2: Box plot and Histogram of Time data

Figure 3: Box plot and Histogram of Weight data

Nonetheless, the distribution for the third column (Baits type) may use mode to measure the spread of data as the bait type is represented in qualitative data [4]. The mode value is calculated by mode() function. The result shows that bait type C has the highest frequency, as shown in figure 4.

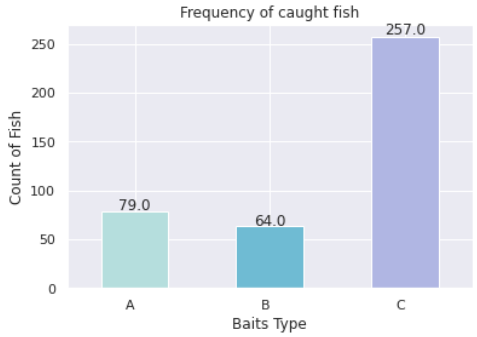

Figure 4: Bar chart for frequency of each bait

2.2 Correlation analysis

According to the previous section, a non-parametric method is more suitable for the dataset's non-normally distributed nature. In the beginning, there are two candidate methods for this analysis: Kendall's tau and Spearman's rank correlation coefficient [1].

Based on the fact that weight samples contain outliers which can be observed from the graph in Figure 3, and the fact that Spearman's rank correlation coefficient is much more susceptible to outliers, Kendall's tau seems to be an intuitive choice. Kendall's tau method will be used in this report = to analyze the correlation between the time a fish was caught and the weight of the fish since it is the most.

Kendall's tau method [2] calculates the correlation coefficient value between caught time and the weight of fish at each particular time which as shown here;

- Kendall value for all bait = -0.071983.
- Kendall value for bait A = -0.198113.
- Kendall value for bait B = -0.032787.
- Kendall value for bait C = -0.053322.

Overall, the correlation between time and weight tends to be weak. However, bait A offers the strongest correlation. Also, even though the correlation is considered weak, it is worth pointing out that time and weight has a negative monotonous relation, meaning the smaller one variable is, the greater the other variable becomes. In other words, bigger fish tends to be caught in the morning and vice versa. Furthermore, bait B seems to have the least correlation between time and weight, indicating that bait B is the most reliable bait since the weight of the fish caught was not affected by the time. However, it is crucial to note that bait B also got the least number of sample which could make the result bias.

2.3 Conditional Probability of catching a fish using each bait

Since this sample was selected from a larger population with 95 percent confidence intervals, the range of probability that represent the actual population can be computed by using the standard error of sample proportion which is $SE = P * Q/n$ [5] where P is the population proportion, Q is $1-P$ and n is number of samples. Since the population proportion is unknown in this research, sample proportion will be used to estimate the population proportion. Lastly, if the weight sample was split into two bins, the value that separates the upper and lower bin is computed as 2.445 kg. This value will be considered as a threshold deciding between a smaller and bigger fish.

These ranges of conditional probability were computed by following (please note that the actual computation is excluded due to the lack of space);

1. Compute $P(Big|X)$ which refers to the probability of catching a big fish (weight > 2.445) when using bait X.
2. The formula for sample proportion is $P = \frac{X}{N}$

where; X = count of successes, N = size of the population. Hence, $P(Big|X)$ can be used in the SE formula as $SE = P(Big|X) * (1 - P(Big|X))/n$.

3. Since SE of the sample is an estimate of the standard deviation of the population. It is acceptable to use SE as SD
4. Finally, with 95 percent confidence intervals, the range of the probability that the actual $P(Big|X)$ of the population can lie within can be computed as $P(Big|X) \pm 2SD$

In summary, with 95 percent certainty:

- 10.17% to 27.81% of the fish caught by bait A is heavier than 2.445 kg.
- 12.84% to 34.04% of the fish caught by bait B is heavier than 2.445 kg.
- 19.87% to 30.71% of the fish caught by bait C is heavier than 2.445 kg.

It is worth pointing out that since bait B got the least sample size (n), the estimated double standard deviation range is more significant than other baits. Even though bait B got the highest upper bound of the probability of catching a bigger fish, it also introduces a larger span of error.The most effective bait seems to be bait C, which gives the highest probability of catching a bigger fish and the narrowest error span.

Also, as can be recalled from the previous section regarding correlation, bait C is the second least dependent on the time variable, which means it is a reliable bait regardless of the time that it was being used.

3 Discussion

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

There are several ways to define the best time to go fishing; one might represent the time with the highest average of fish weight, the other might focus on the quantity of caught fish. In addition to the first definition of the best time for fishing, according to figure 5, it is noticed that the range of fish weight is lie in the highest range in hours of 3, which may be related to the average value in this range of time.However, to emphasise the highest average of fish weight in each hour, the mean value of fish weight on each hour was calculated and plotted as present in figure 6. Besides, the mean value of fish at 3 p.m. is approximately 2.6 Kg which is the highest value in the time range.

In another way, figure 7 shows the frequency of catching fish in each range of hours. The graph mentions that the highest total number is around 31 units at 11 a.m.; in other words, the fisherman can catch the most fish at 11.Notwithstanding, according to sample data, the best time to go fishing might depend on the different points of view. At 3 p.m. might refer to the best for fishing in terms of the highest average of fish weight, whereas 11 a.m. seems to be the proper hour for fishing if that number of fish is taken into account for consideration.

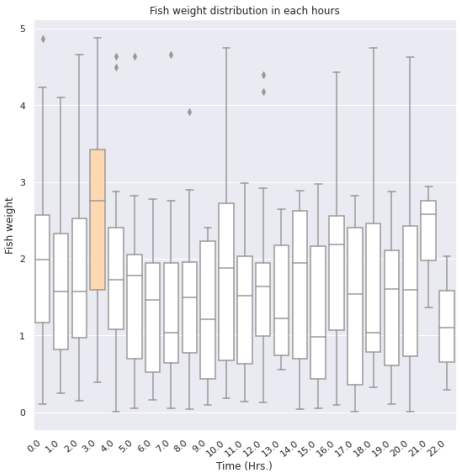

Figure 5: Box plot for weight of fish in each hours

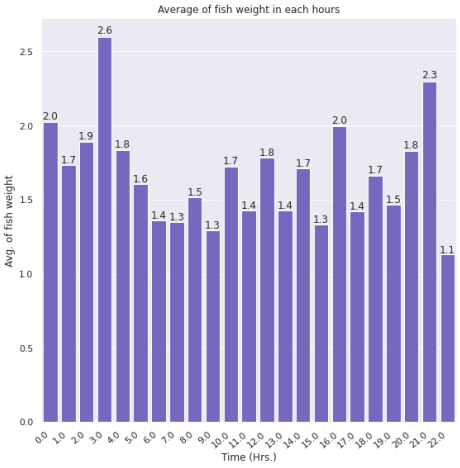

Figure 6: Bar graph to show average of fish weight in each hours

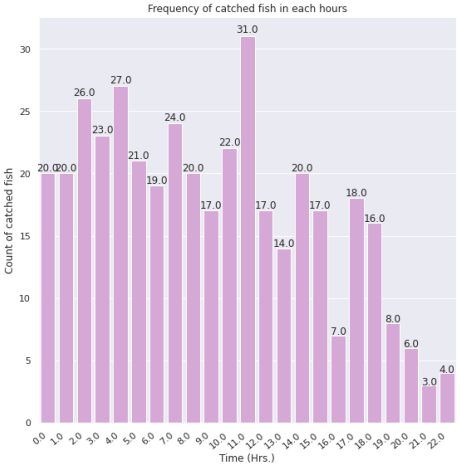

Figure 7: Bar graph to show frequency of fish weight in each hours

3.2 Which bait is most effective?

However, to concrete the assumption in correlation analysis which claim that bait type C is given the most effective for fishing. The specific time is chosen to demonstrate the distribution of fish weight in each bait type. Figure 8 and 9 shows the distribution of each bait type at 11 a.m. and 3 a.m., respectively. These two charts seem that bait type C has either the broadest range of data that may present the frequency of catching fish or lie in the high weight of caught fish.According to the graph, this might be concluded that bait type C noticeably has the highest efficient base from the provided sample dataset.

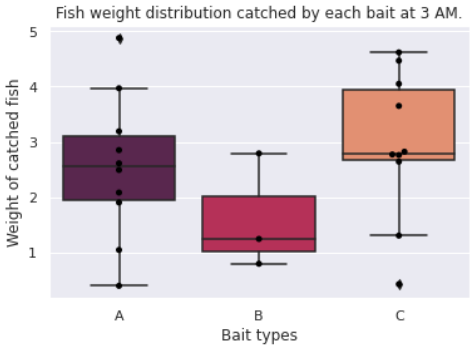

Figure 8: Box plot of weight of each type of bait at 3 a.m.

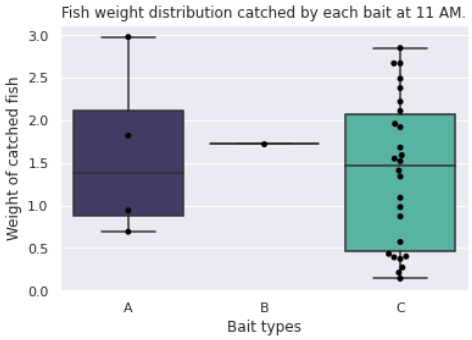

Figure 9: Box plot of weight of each type of bait at 11 a.m.

3.3 What is the best type of bait to use at 3 p.m.?

Subsequently, similar to the previous analysis, the distribution of weight by each bait type was plotted as represented in figure 10. Due to the graph seems that baits type B and C might give a slight difference of qualities result. However, to analyse the quality of bait, there appear that it has two main perspectives that may take into account; the average weight and the number of fish. Figure 11 demonstrate the average fish weight caught at 3.00 p.m. Although bait type C gives a slightly lower average than bait B, it gives a significantly higher fish number. Thus baits type C may be suitable for fishing at 3 p.m. compared with the other type.

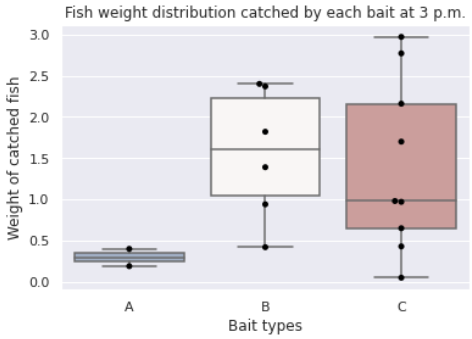

Figure 10: Box plot of weight of each type of bait at 3 p.m.

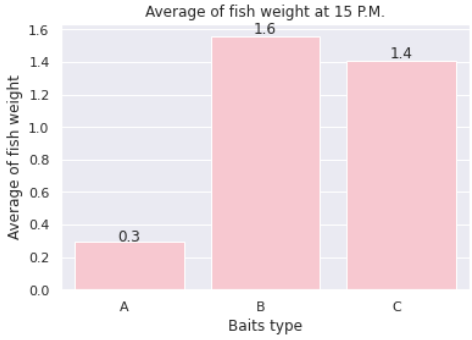

Figure 11: Bar chart of average of fish at 3 p.m.

4 Conclusion

In conclusion, the variables from sample data might represent a non-normal distribution as analysed in the data observation section. The distribution of sample data can affect the parametric method. Thus non-parametric method is suitable to apply for finding the relationship between variables. In addition, the analysis indicates that bait C, which has the weakest correlation between weight and time, seems to be the most effective bait if the word refers to (1) the ability to catch a fish regardless of time and (2) having high chance of catching a bigger fish. In addition, the best time for fishing discovered in this report may refer to 11 a.m. and 3 a.m. as the best time with the highest number of fish and the highest average fish weight, respectively.

References

[1] Daniel, W. [1990], *Applied Nonparametric Statistics*, Duxbury advanced series in statistics and decision sciences, PWS-KENT Pub.
URL: *https://books.google.co.jp/books?id=0hPvAAAAM*

[2] Kendall, M. and Gibbons, J. D. [1990], *Rank Correlation Methods*, 5 edn, A Charles Griffin Title.

[3] Mircioiu, C. and Atkinson, J. [2017], 'A comparison of parametric and non-parametric methods applied to a likert scale', *Pharmacy* **5**(2).
URL: *https://www.mdpi.com/2226-4787/5/2/26*

[4] Rust, R. T. and Cooil, B. [1994], 'Reliability measures for qualitative data: Theory and implications', *Journal of Marketing Research* **31**(1), 1–14.
URL: *https://doi.org/10.1177/002224379403100101*

[5] Zwillinger, D. [2018], *CRC standard mathematical tables and formulas*, chapman and hall/CRC.