Foundations of Data Science

Abstract

Different statistical tools and measures help to analyze and interpret the true nature of the data. In order to identify the best bait among the three categories, the best time of fishing, and the effectiveness of baits at a specific time, different statistical techniques and mathematical computation have been utilized. It is also observed that to understand the best fishing time or best bait type, there can be two different scenarios depending on the fisherman's requirement: either weight of the fish or the number of fish. As the fisherman's point of view is unknown, both the scenarios are examined, and a conclusion is drawn. Different visualization plots have also contributed a lot to shed some light on the data's true pattern and easier interpretation.

Introduction

Fishing is one of the exciting hobbies or occupations for different people. However, a little bit of analytics can improve the overall efficiency and outcome of fishing. The dataset that has been considered here has information about 'Time of catch', 'weight of fish', and 'bait type'. Statistical analysis has deciphered how the fishing outcome can be changed by using different bait types or different fishing times. In the report, the best time is determined by a few plots and mathematical computations and comparisons. How the change in time impacts the weight of fish has been clearly shown in the analysis with statistical measures. In order to understand which bait type has the maximum effectiveness, a comparative study has been shown, considering different scenarios. Also, after critically analyzing the sample data, an estimation has been made on the population data.

Background and Literature Review

[1] The 95% confidence interval provides a range (with the help of sample mean and standard deviation) with 95% certainty that the population mean will lie within that range. More precisely, it implies if multiple samples are collected from the population dataset, 95% of the samples will provide the same range of values within which the population mean will lie. Still, there can be 5% (significance level – α) [7] unsuccessful cases. α for 95% confidence interval is 0.05, and the Z-score is 1.96 (from Z-table). The interval range increase if % confidence interval increases.

[2] IQR is the range obtained by the difference between 75(Q_3) percentile and 25(Q_1) percentile value. The outlier range can be calculated using IQR. The formulas for the same are as follows:

$IQR = (Q_3 - Q_1)$

$\text{Outlier Range} = [Q_1 - (IQR*1.5), Q_3 + (IQR*1.5)]$ (1)

Methodology

This is a sample dataset of 400 data points collected from a larger fishing dataset. The data set contains 3 columns: X, Y, Z. X represents the time of catching fish in 24-hour format, Y represents the weight(kg) of fish, and the Z column represents different bait types. Also, it has been observed from the data that at a single timestamp, multiple fishes have been caught by the same bait, which indicates that the fisherman has more than one bait of the same type. For example, at time 12:38 A.M., bait 'C' has caught two fishes of different weights 1.19 and 2.44 kg.

EDA Performed:

- **Feature Engineering:** The 'Time' variable is modified because the minute part of the time data is presented as a percentage of a minute. For example, if the time is 22.95, it means it is 22.57 (as 95% of 60 is 57 min), and the total time is shown in 24-hour format.
- **Outlier Detection:** For outlier detection, the IQR formula is used to estimate the range. If data falls outside the range, then it is considered to be an outlier. Based on EDA, it is found that the dataset contains no outliers.
- **Duplicate and Missing value check**: It is essential to remove any duplicate record from the data and replace the missing values with a suitable statistical measure to avoid data discrepancies. After performing EDA, no duplicate or missing values were found in the given dataset.
- **QQ-plot for distribution check:** To understand if values of both Weight and Time column are normally distributed, QQ plot [4] has been drawn to interpret the result visually.

A 95% confidence interval for the population mean for both 'Time' and 'Weight of fish' is calculated with the help of sample mean, sample standard deviation, and Z-score.

Covariance is used to show the direction of the relationship between Time and Weight of Fish (kg) of the fishing data, and correlation coefficient is used to determine the strength of the relationship.

To identify the efficient bait, two scenarios have been considered.

a) **Weight of the fish is the primary concern:** In this case, avg. weight of fish captured by each type of bait would be the best parameter to identify the best bait type. To visually interpret the effectiveness as per avg. weight a box plot of weight distribution with highlighted mean for different bait types is shown.

b) **The number of fish is the primary concern:** In this case, counting the number of fishes captured by each type of bait will provide the best bait in terms of the number of fish caught. A bar plot of the count of fishes for different bait types is shown in the report for visual interpretation.

To determine the best time for fishing, the time scale is divided into 24-time bins of 59 min duration, and the two scenarios: avg weight of the fish per time bin and number of fish caught per time bin are validated, and the decision is made.

Discussion

From the statistical distribution of the weight of fish (Kg) (Fig: 1), it can be observed that the skewness is positive and mean > median, which implies that the data is slightly right-skewed. Also, the positive kurtosis indicates the same conclusion of right-skewed data. The density distribution graph shows that most of the fish's weight lies between 0.01 kg to 3 kg.

Here, IQR (Q_3-Q_1) is 1.6925, and the black dotted shows outlier range (using formula 1) between -1.83 to 4.94, outside of which no data point is present implies no outlier in the data. The descriptive statistics table (Fig: 2) shows other statistical measures [5] like IQR, standard deviation, variance, etc. The QQ plot (Fig: 3) indicates that the weight of the fish is normally distributed as it is almost accurately following the theoretical quantile line.

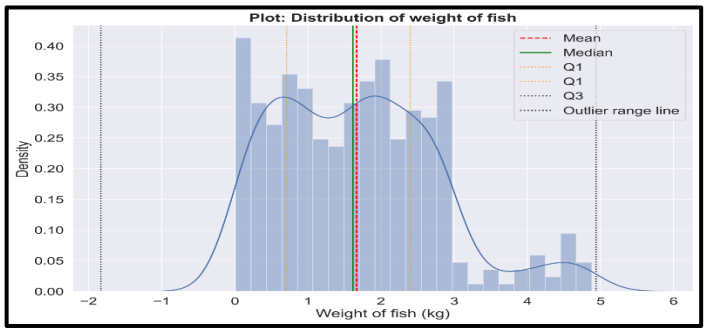

Fig:1 Weight distribution of fish (Kg)

Quantile statistics		Descriptive statistics	
Minimum	0.01	Standard deviation	1.108159639
5-th percentile	0.15	Coefficient of variation (CV)	0.6646033576
Q1	0.7075	Kurtosis	0.1618912657
median	1.615	Mean	1.6674
Q3	2.4	Median Absolute Deviation (MAD)	0.855
95-th percentile	3.9225	Skewness	0.6537928749
Maximum	4.88	Sum	666.96
Range	4.87	Variance	1.228017784
Interquartile range (IQR)	1.6925	Monotonicity	Not monotonic

Fig:2 Descriptive statistics for the weight of fish (kg)

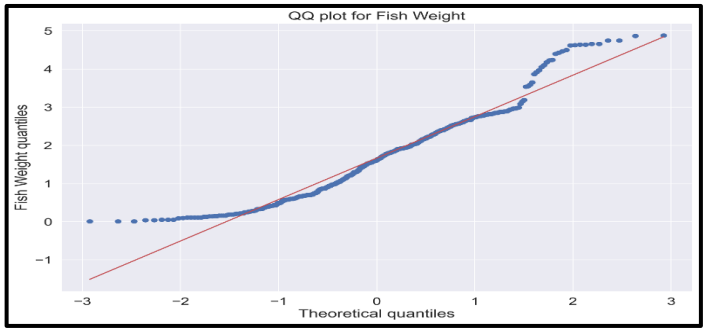

Fig:3 QQ plot for the weight of fish (kg)

1) A similar observation to the weight plot was evident while observing the time distribution plot (Fig: 4). Even though the distribution says that mean is greater than median with positive skewness, i.e., the distribution is right-skewed, the entire data falls within the range -9.26 to 27.29 (outlier detection range calculated using formula 1), indicates no outlier.
   From the descriptive statistics table (Fig. 5) of Time, the interquartile range (IQR) can be seen as 9.26, which is much larger than the weight signifies to a larger spread. A greater standard deviation as compared to weight columns also points towards the greater spread. It can also be observed that most of the fishing is done within the time interval of 11 A.M. to 12 P.M. and with time, the count of catching fishes has reduced. The QQ plot (Fig: 6) shows the data is normally distributed.

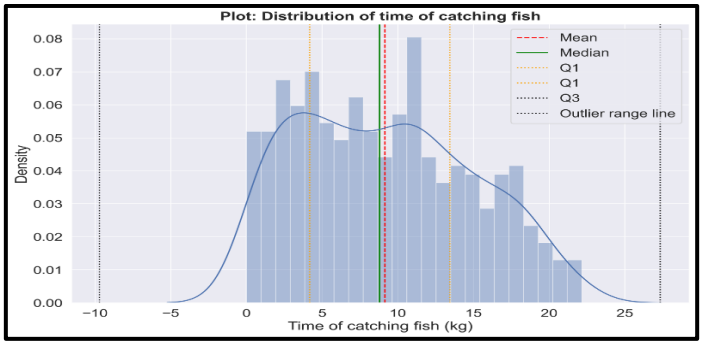

Fig:4 Time distribution of catching fish

Quantile statistics		Descriptive statistics	
Minimum	0.01	Standard deviation	5.78779223
5-th percentile	0.9885	Coefficient of variation (CV)	0.6320465239
Q1	4.185	Kurtosis	-0.9529125172
median	8.8	Mean	9.157225
Q3	13.445	Median Absolute Deviation (MAD)	4.635
95-th percentile	19.0105	Skewness	0.2673819442
Maximum	22.16	Sum	3662.89
Range	22.15	Variance	33.4985389
Interquartile range (IQR)	9.26	Monotonicity	Not monotonic

Fig:5 Descriptive statistics for the time of catching fish

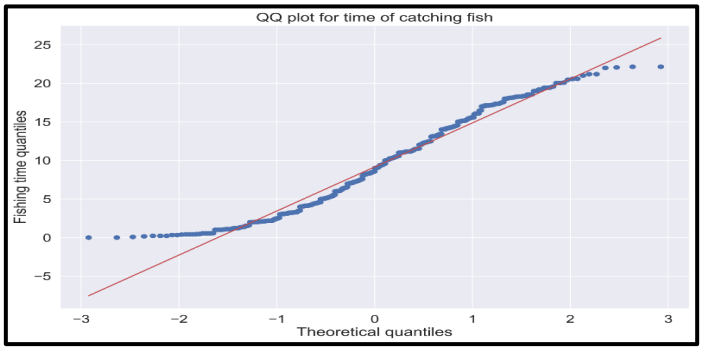

Fig:6 Descriptive statistics for the time of catching fish

Estimation of population mean from the sample data:

As the dataset that has been analyzed is a subset of a larger population, it is not certain that the sample mean will be equal to the population mean [6]. However, for the estimation of population, a 95% confidence interval is calculated using the formula:

$$\text{Population mean } (\mu) = \text{Sample mean } (\bar{X}) \pm \left(\frac{Sample\ SD\ (\sigma)}{\sqrt{n}} * 1.96\right) \qquad (2)$$

Where 1.96 is the z score for 95% confidence interval obtained from Z-table

Using the formula 2, the population estimations are:

$8.5898 \leq$ Time of catching fish (population) ≤ 9.7241

$1.5584 \leq$ Weight of fish (Kg) (population) ≤ 1.7755

Relation between Time of catch and Weight of fish:

The Relationship graph between 'Time' and 'Weight of fish' (fig. 7) shows a negative covariance [3] calculated using:

$$Cov[X,Y] = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \quad (3)$$

$$\text{Correlation coefficient (r)} = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})\text{^}2}\ \sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})\text{^}2}} \quad (4)$$

Where X and Y represent the Time and Weight of Fish (kg), and N represents the number of data points, and N-1 is used to handle the biasness because the analysis is made on sample data where $\bar{X}$ *and* $\bar{Y}$ denotes the sample mean of X and Y in formula 3.

A negative covariance between the variables shows that the two variables are inversely related and correlation coefficient (r) equals to -0.12 (using formula 4) indicates a weak negative correlation b/w Time (X) and Weight of fish (Y) which implies that the two variables are not dependent on each other. Evidently, from Fig:7 the relationship is negative and weak between the two variables and with time the fishing events are also reducing.

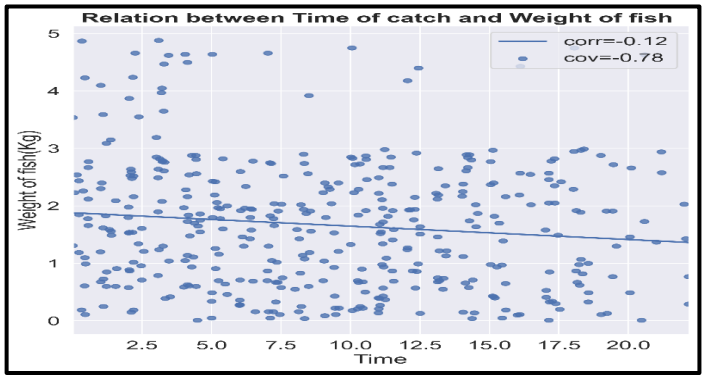

Fig:7 Relationship between time of catching fish (X) and Weight of fish (Kg) (Y)

Effectiveness of bait:

The effectiveness of bait can be analyzed under two conditions:

a) Effectiveness in terms of the average weight of the fish caught by each bait:

When the weight of the fish is the most important factor for the fisherman to measure the effectiveness, then the effectiveness can be computed as:

$$\text{Effectiveness of a bait} = \frac{\text{Total weight captured by the bait type}}{\text{Total number of times that bait has been used}}$$

Here' Effectiveness of a bait' represents the average weight of fish captured by a bait type. The table below (Fig: 8) shows bait 'B' has captured fish with a maximum avg weight of approximately 1.78 kg. The box plot [8] of the weight of the fish caught for different bait with a highlighting mean shown in Fig: 9 indicates bait 'B' is most effective when weight matters.

Note: *From fig-9 it can be observed that there are outliers in the case of bait "A". However, even with the influence of outliers, it is unable to outperform bait "B" and "C".*

bait_type	count	Sum of weight	Effectiveness by weight
A	79	120.79	1.528987
B	64	114.16	1.783750
C	257	432.01	1.680973

Fig:8 Tabular Representation of effectiveness of bait by weight

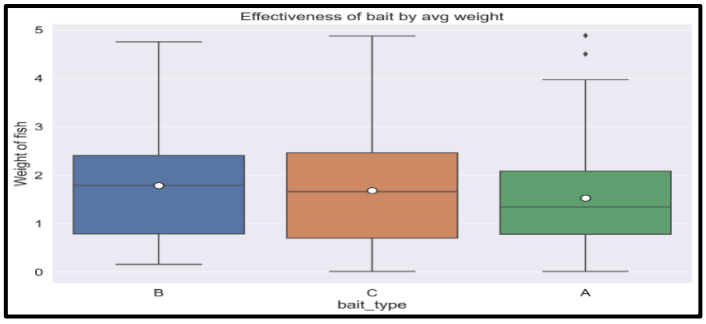

Fig:9 Effectiveness of bait in terms of avg weight of fish caught

b) Effectiveness in terms of the number of fish caught:

If the number of fish caught by a bait type is the measure of effectiveness for the fisherman, the effectiveness can be calculated by the total number of fish caught per bait type. The graph below (Fig: 10) shows bait 'C' has caught maximum fish, which is 257, supported by the above table (Fig: 8). So, it can be concluded that if the number of fish is the important factor for the fisherman, then bait 'C' is the best bait.

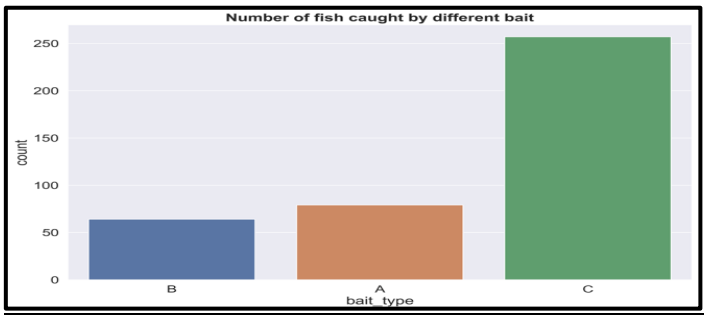

Fig:10 Effectiveness of bait by number of fish caught

Best Bait at 3 P.M.:

For the analysis of best bait at 3 P.M., a time interval must be considered because there is no fishing event at 3 P.M.

For this analysis, the 14:30 to 15:30 time range has been considered, and the baits' performance was measured using the same two approaches; by avg weight of fish captured and by the number of fish caught by each bait.

a) *The best bait for 3 P.M. in terms of the average weight of the fish caught:* The average weight of fish captured by bait 'B' is again the maximum within the time interval. So, we can say 'B' is the best bait when the weight of the fish matters the most for the fisherman. The bar plot (Fig: 11) of bait's performance w.r.t avg weight and the table (Fig:12) is shown to substantiate the statement

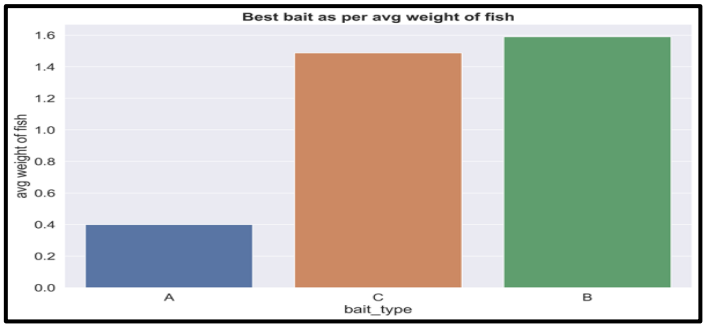

Fig:11 Best bait as per avg weight of fish

bait_type	Number_of_fish_caught	Total_weight	avg weight of fish	Time_range
A	1	0.40	0.400000	14:30-15:30
B	5	7.95	1.590000	14:30-15:30
C	11	16.37	1.488182	14:30-15:30

Fig:12 Best bait as per avg weight of the fish caught

b) *The best bait in terms of number of fish caught:* If the number of fish caught is the main deciding factor for the fisherman, then at 3 P.M. (considering the time range from 14:30 to 15:30), bait 'C' is the best bait supported by the table below (Fig: 13)

bait_type	Number_of_fish_caught	Total_weight	avg weight of fish	Time_range
A	1	0.40	0.400000	14:30-15:30
B	5	7.95	1.590000	14:30-15:30
C	11	16.37	1.488182	14:30-15:30

Fig:13 Best bait as per the number of fish caught

Best Time for fishing:

If avg weight of the fish is important to the fisherman than the number of fish caught, then from the plot (Fig:14), it can be seen that the maximum average weight of fish in time interval of 3 to 4 A.M. is 2.59 Kg. So, 3 to 4 A.M. is the best time for fishing if the fisherman wants to catch fish with huge weight as the highest bar (length of the bar representing avg weight of the fish) corresponds to the time interval 3–4 A.M.

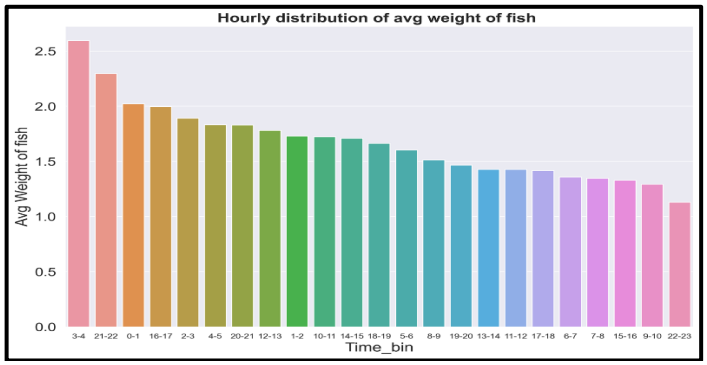

Fig:14 Best fishing time as per average weight

In the graph mentioned below, the length of the bar represents the number of fish caught, and the x-axis represents hourly time bins which shows the maximum number of fish to be 31, caught at 11AM-12 PM bin. Hence, if the fisherman considers the quantity of fish as the best measure for fishing, then he should target 11AM-12 PM time interval.

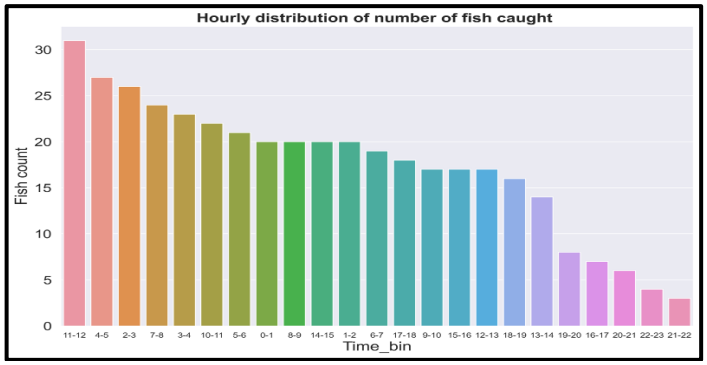

Fig:15 Best fishing time as per the number of fish caught

Conclusion

Using various statistical measures and mathematical computation, the best bait type is determined in two different ways.

Approach1: Weight of the fish is more important than the number of fish to the fisherman
Approach2: The number of fish is more important than the weight of fish to the fisherman.

Using both approaches, the best bait is determined, and a conclusion is made that bait 'B' is the best if the weight of fish is concerned, and bait 'C' is best if the number of fish is concerned. To determine fishing efficiency, both approaches were taken into consideration, and the best fishing time was found out. Morning 3-4 A.M. is the best time in terms of weight, and 11-12 A.M. is the best time in terms of the number of fish. As there are no direct fishing events at 3 P.M. in the afternoon, the fishing event is observed between 14:30 to 15:30, and the best bait type is determined using both approaches. For 3 P.M. in the afternoon, bait 'B' is the one that is catching heavy-weighted fish, and bait 'C' is catching the maximum number of fish.

References

[1] Brookmeyer, R., & Crowley, J. (1982). "A Confidence Interval for the Median Survival Time", *Biometrics*, *38*(1), 29–41. https://doi.org/10.2307/2530286

[2] Jiawei Yang, Susanto Rahardja, and Pasi Fränti. (2019). "Outlier detection: how to threshold outlier scores" In Proceedings of the International Conference on Artificial Intelligence, New York, Article 37, https://doi.org/10.1145/3371425.3371427

[3] Jöreskog, K.G. "Structural analysis of covariance and correlation matrices", *Psychometrika* **43,** 443–477 (1978). https://doi.org/10.1007/BF02293808

[4] Z. Djurovic, B. Kovacevic and V. Barroso, "QQplot-based probability density function estimation," Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing 2000, pp. 243-247, https://doi.org/10.1109/SSAP.2000.870120

[5] Manikandan, S. "Measures of central tendency: Median and mode", Journal of Pharmacology and Pharmacotherapeutics; Mumbai Vol. 2, (Jul 2011): 214-215. https://doi.org/10.4103/0976-500X.83300

[6] I. J. Good, "The Population Frequencies of Species and The Estimation of Population Parameters", *Biometrika*, Volume 40, Issue 3-4, December 1953, Pages 237-264, https://Doi.Org/10.1093/Biomet/40.3-4.237

[7] Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). "Statistical Inference for Coefficient Alpha". *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

[8] Neil C Schwertman, Margaret Ann Owens, Robiah Adnan, "A simple more general boxplot method for identifying outliers", Computational Statistics & Data Analysis, Volume 47, 2004, Pages 165-174, https://doi.org/10.1016/j.csda.2003.10.012