# The Statistical Report of Fishing Data Set

ID(29299675)

**Abstract**

The report analyses the data set which records the fishing data for one fisherman has fished in a lake during a time period of one day by the statistical methods. It provides the distribution of fishing time and fish size, mainly analyses the dependence of each variable and correlation between fishing time and fish size to find the inner relationship among them.

## Introduction

1. The motivation of this analysis is to analyse the fishing data set through the statistical techniques intending to discover whether there are correlations between variables.
2. The available resources and tools are common statistical techniques.

## 1 Methodology

1. This analysis utilised standard statistical techniques, mean value, median value, standard variance, correlation analysis, and ANCOVA analysis.
2. The data set consists of three columns with X values giving the times where the fisherman has made a catch, the Y values indicating the size of catch (i.e. its weight in kg), and the Z values giving a letter A, B, or C showing which fishing rod was used to make that catch.
3. The analysis is based on the 95% confidence level.

## 2 Results

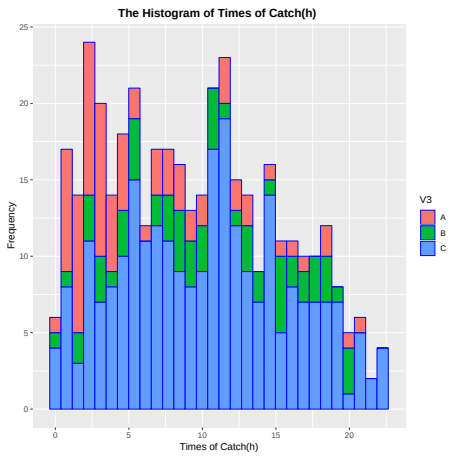


Figure 1: The Histogram of Time of Catch

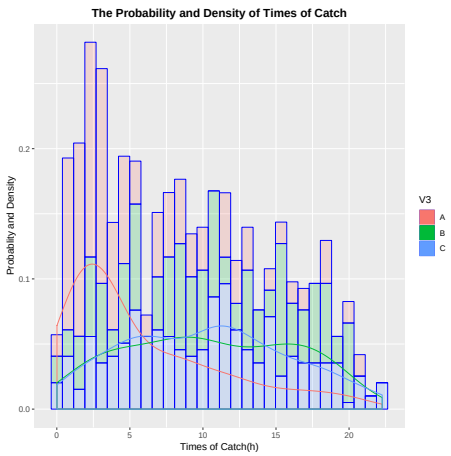


Figure 2: The Probability and Density of Time of Catch

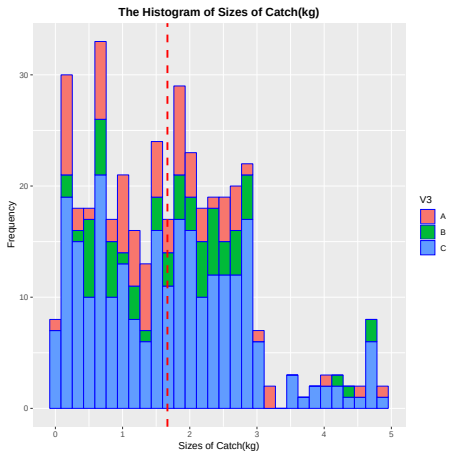


Figure 3: The Histogram of Sizes of Catch(The red- dashed line indicates the mean value)

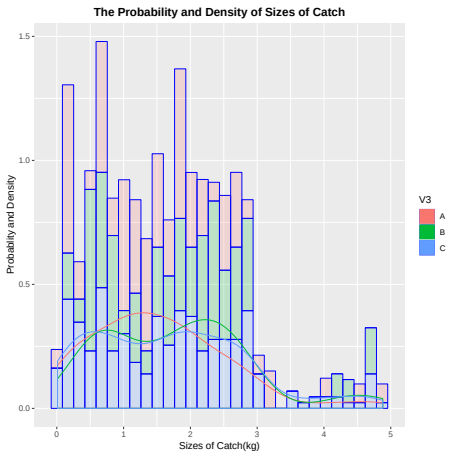


Figure 4: The Probability and Density of Sizes of Catch

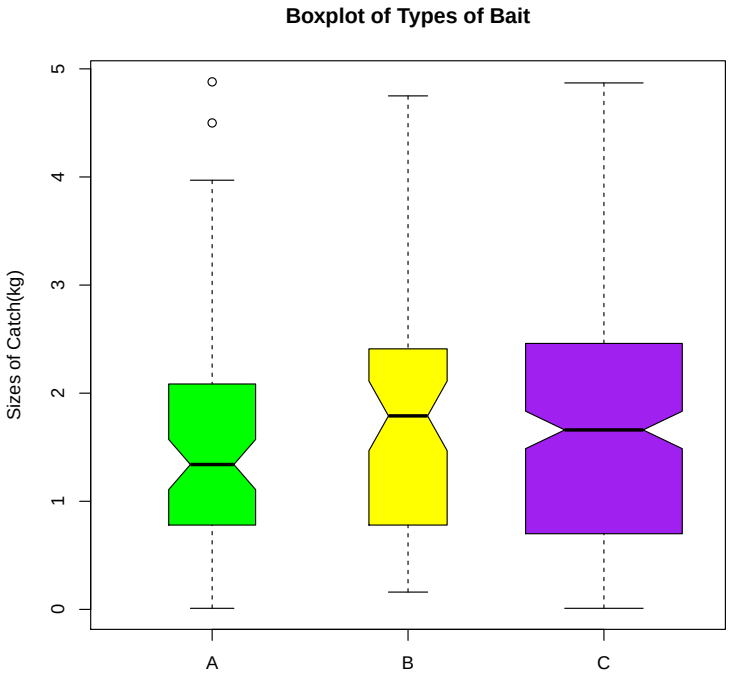


Figure 5: The Boxplot of types of bait for the entire time period of fishing.

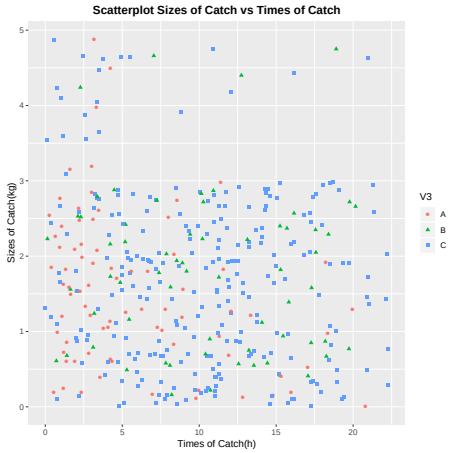


Figure 6: The Scatterplot Sizes of Catch vs Time of Catch.

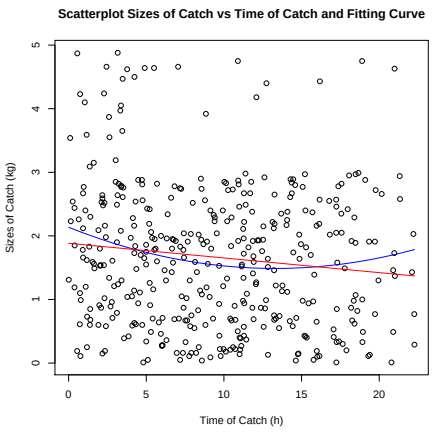


Figure 7: Fitting Curve for Sizes of Catch vs Time of Catch.

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max | Sd. |
|---|---|---|---|---|---|---|---|
| X(time of catch) | 0.010 | 4.325 | 9.020 | 9.371 | 13.748 | 22.270 | 5.7964 |
| Y(sizes of catch) | 0.0100 | 0.7075 | 1.6150 | 1.6674 | 2.4000 | 4.8800 | 1.1081 |

Table 1: Measures of X values and Y values

| | Type A bait | Type B bait | Type C bait |
|---|---|---|---|
| Frequency | 79 | 64 | 257 |
| Probability | 0.1975 | 0.1600 | 0.6245 |

Table 2: The statistical probability of Z(type of bait) values

| | X (time of catch) | Y (sizes of catch) |
|---|---|---|
| Population Mean | [8.80076, 9.94029] | [1.558472,1.776328] |

Table 3: Mean values with 95% confidence intervals for X and Y distributions

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| A type bait | 0.010 | 0.780 | 1.340 | 1.529 | 2.085 | 4.880 |
| B type bait | 0.160 | 0.785 | 1.790 | 1.784 | 2.405 | 4.750 |
| C type bait | 0.010 | 0.700 | 1.660 | 1.681 | 2.460 | 4.870 |

Table 4: The summary of fishing effectiveness for different types of bait

| | Adjusted R-squared | p-value |
|---|---|---|
| Linear model | 0.01207 | 0.01582 |
| Quadratic polynomial model | 0.02174 | 0.004695 |

Table 5: The correlation analysis between X and Y

| | Adjusted R-squared | p-value |
|---|---|---|
| Linear model for categorical data | -5.911e-05 | 0.3732 |

Table 6: The dependence analysis between Y and Z

| | X | Z | X:Z |
|---|---|---|---|
| p-value | 0.0154 | 0.0902 | 0.2633 |

Table 7: The interaction analysis of Z and X to Y

| Index | X | Y | Z |
|---|---|---|---|
| 290 | 3.16 | 4.88 | A |

Table 8: The data item which the Y(size of catch) value is the largest

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|---|
| Type A bait | 0.1500 | 0.2125 | 0.2750 | 0.2750 | 0.3375 | 0.4000 |
| Type B bait | 0.420 | 0.810 | 1.380 | 1.395 | 1.965 | 2.400 |
| Type C bait | 0.040 | 0.570 | 1.640 | 1.488 | 2.395 | 2.970 |

Table 9: The summary for different types of bait from 2.30pm to 3.30pm

| | Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|---|
| Type A bait | 0.150 | 0.180 | 0.295 | 0.490 | 0.605 | 1.220 |
| Type B bait | 0.4200 | 0.7075 | 1.2550 | 1.4275 | 2.2575 | 2.5700 |
| Type C bait | 0.0400 | 0.7175 | 1.9450 | 1.7057 | 2.5075 | 4.4300 |

Table 10: The summary for different types of bait from 1pm to 5pm

# 3 Discussion

1. **The correlation between X (time of catch) and Y (sizes of catch)**
   It is discovered from the scatter plot (Figure 6) between X and Y that there is no direct correlation between X and Y.
   - The result of fitting a linear model to Y and X
     From Table 5, "Adjusted R-squared" shows around just 1.207% of the variation in Y accounted for by X. The p-value is 0.01582 that is larger than 0.01, which is showed that the hypothesis the linear model to Y and X is incorrect.
   - The result of fitting a quadratic polynomial model to Y and X
     From Table 5, although p-value is 0.004695 which is less than 0.01, Adjusted R-squared is not impressive. According to the quadratic polynomial model curving fitting(Figure 7), the quadratic polynomial model does not fit well for Y and X.
2. **The dependence between Z (the type of bait used) and Y**
   The Table 6 indicates "Adjusted R-squared" is not impressive, and the p-value is 0.3732 that is larger than 0.05, which shows the model is not significant as well. The dependence on Z (the type of bait used) from Y is not significant. Z has no significant effect on Y.
3. **The interaction analysis of Z and X to Y**
   Consider X as the predictor and Y as the response variable, considering the interaction between X and Z. From Table 7, it is indicated X and Z do not effect Y significantly and there is no significant effect between X and Z, because the p-values are all larger than 0.01 or 0.05.
4. **Question1:What is the best time to go fishing at this lake?**
   From Figure 6, it is showed that the interval from 0am to 5am is more effective for fishing, because there are more points which sizes of catch are larger than other intervals. The time which size of catch is the largest (4.88kg) is 3.16am (Table 8) which is also in the interval from 0am to 5am. It can be summarized that 3.16am is the best time for fishing.
5. **Question2:Which bait is most effective?**
   From Figure 5 and Table 4, it is indicated that the type B bait has the largest median and mean value for sizes of catches. From Figure 5, it also can be showed how the median values of different baits match each other. Although type B bait do not have the largest maximum, the distribution for sizes of catch is more uniform and concentrated. It is summarized that type B bait is most effective.
6. **Question3:What is the best type of bait to use at 3pm in the afternoon?**
   The key point of this question is how size of the time interval to select around 3pm to analyse. The subset of the whole sample can also be regarded as a sample. If selecting 95% confidence level for the analysis, this means the probability that the standard error between the mean value of the sub sample and the mean value of population is less than certain value is 95%. For this question, Y variable (sizes of catch) is selected to consider. Considering the mean value of Y for the sample is 1.6674 kg (Table 1), the standard error is selected from 0.1 to 0.5, based on the formula:

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq error$$

   Among this, z is 1.96 corresponding to 95% confidence level, $\sigma$ is 1.1081 (Table 1). The required sample size corresponding to different standard error for analysis is as follows:

| Standard error(kg) | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |
|---|---|---|---|---|---|
| The required sample size | 19 | 30 | 53 | 118 | 472 |

Table 11: The required sample size corresponding to different standard error

   The sample size in different time intervals is:

| Time interval | 2.30pm to 3.30pm | 2pm to 4pm | 1.30pm to 4.30pm | 1pm to 5pm |
|---|---|---|---|---|
| Sample size | 17 | 37 | 47 | 58 |

Table 12: The sample size corresponding to different time intervals

   From Table 9 10, for different time intervals although corresponding to different standard error, it can be inferred the same result, the size of catch for type C bait has the largest mean value and median value, which indicates type C bait is the best type to use at 3pm.

# 4 Conclusion

Overall, it can be inferred from the above analysis there is no correlation between X and Y, there is no direct dependence between Z and Y as well.