# A Report for Foundations of Data Science Statistics Coursework

Ziyi Guo

zg2u21@soton.ac.uk

**Abstract**

Mathematical statistics plays an important role in data analysis and tendency prediction and is widely applied in various practical problems. This paper firstly summarizes the related methodologies in statistics, including the descriptions of data distributions, the correlation evaluations between variables and the prediction of population based on random samples. Secondly, the paper expounds the experiment based on a fishing data set, introducing the raw data set and the initialize operation, showing the visualization plots generated to describe the data distributions and correlations of the variables and giving mathematical analysis based on experiment outputs. To conclude, the paper sums up the experiment results and purposes reasonable suggestions on future fishing.

## 1 Introduction

Mathematical statistics are widely applied in the daily life. It can reveal the main feature of the data from the perspectives of numerical analysis as well as graphic inference. With the increase of data complexity nowadays, it is becoming more and more important to carry out appropriate statistics methods in different cases. Based on the known statistics of the samples from a large object group or the current state of a phenomenon, the data distribution of the large group and the developing tendency can be inferred through prediction methodologies.

Fishing data is a specific problem for data analysis, which often contains information of fishing time, the size of catch, the bait used, etc. Reasonably analyzing the current situation of a fishing can be helpful to know about the data distributions and propose sensible suggestions for future fishing. Statistical analysis therefore has practical significance to fishery.

## 2 Methodology and Related Work

### 2.1 Descriptions of the Data Distributions

Descriptions of data distributions gives the features about the current data set, which is crucial for the following data analysis. Researchers generally describe a data distribution with the Central Tendency, the Dispersion Degree, and the Distribution Shape[1].

Central Tendency is that the data of a large number of individuals often have the distribution characteristics of fluctuating around a centre in a certain range. Determining the Central Tendency index, it is able to observe the trend of centralizing and the general level of the data.To measure the Central Tendency index, researchers often use two kinds of indicators: the numerical mean, including arithmetic mean, harmonic average and geometric average, and the location characteristic, including mode and median[2, 3, 4].

Dispersion Degree represents the difference extent of data. It can describe the stability and equilibrium of data and measure the representativeness of the mean value. The more dispersed the data distribution and the bigger the degree of dispersion is, the smaller the representativeness of the mean value becomes. Researchers often use range, quartile deviation, variance, standard deviation and confidence intervals for interval data and variation ratio for categorical data[4].

To characterize data distribution, researchers also use Skewness and Kurtosis to describe the distribution shape. Skewness is measured by mean, median and mode and describes the asymmetric and skew extent of the data distribution. Data distributions with skewness value of positive, 0 and negative correspond to right-skewed distribution, normal distribution and left-skewed distribution. Usually in the case of a left-skewed distribution distribution, the value of mode is larger than median larger than arithmetic mean and vice versa. Kurtosis describes the central extent of variables and the steepness of distribution curve. Positive Kurtosis indicates the distribution to be more pointed and centralized while negative value indicates higher level of discreteness[5].

### 2.2 Evaluations of the Correlations between Variables

Correlation is used to denote the association between two quantitative variables. The correlation coefficient $r$ is represented by a value that varies from +1 through 0 to 1. The absolute value of $r$ indicates the weak or strong association between variables. Complete correlation is expressed by +1 or 1 while complete absence of correlation is represented by 0[4].

### 2.3 Estimating Population with Samples

Population is the aggregate of all the subjects in the statistic. Samples are the aggregate of the random individuals in the population. Based on specific rules, the data distribution of the population can be estimated through statistics of the samples. Unbiased estimation is an unbiased inference when using sample statistics to estimate population parameters. If the mathematical expectation of the estimator is equal to the true value of the estimated parameter, the estimator is called the unbiased estimation of the estimated parameter. For unbiased estimation, the mean of population equals the samples' mean and the variance of population equals the sum of the sample-to-mean distances' square divided by the amount of samples minus 1[6].

# 3 Experiment

The experiment focuses on a specific question about multi-variate data analysis. A data set recording the fishing data is used to plot the distributions of the statistics, analyze the correlations between the variables and propose theoretical guidance to the recommendations for fishing based on the experiment results. The experiment environment for code development is on Jupyter Notebook with bottom layer of Python 3. The operating system is Windows 10.

## 3.1 Experiment Data Set

The experiment uses an existing data set that describes a hypothetical situation of the catch of a fishing fleet on a single day. The data set consists of three columns with $X$ values showing the times of catches, $Y$ values giving the sizes of each catch and $Z$ values in $A, B, C$ representing the categories of the fishing rods. There were totally 400 records of catches, corresponding to the 400 rows of data in the data set. In order to make the raw data analytical, tools from *panda* were used to initialize the file to a basic data frame of 400 rows and 3 columns. Considering the effect of each bait, the basic data frame was divided into 3 independent data frames based on the $Z$ value, which were of 79 rows, 64 rows and 257 rows, and 3 columns all, corresponding to 79 catches with bait $A$, 64 with bait $B$, and 257 with bait $C$.

## 3.2 Distributions of the Variables

In the first part of the experiment, it was aimed to fully describe the distributions of the data and give confidence intervals for the distribution mean. The figures generated are as below:

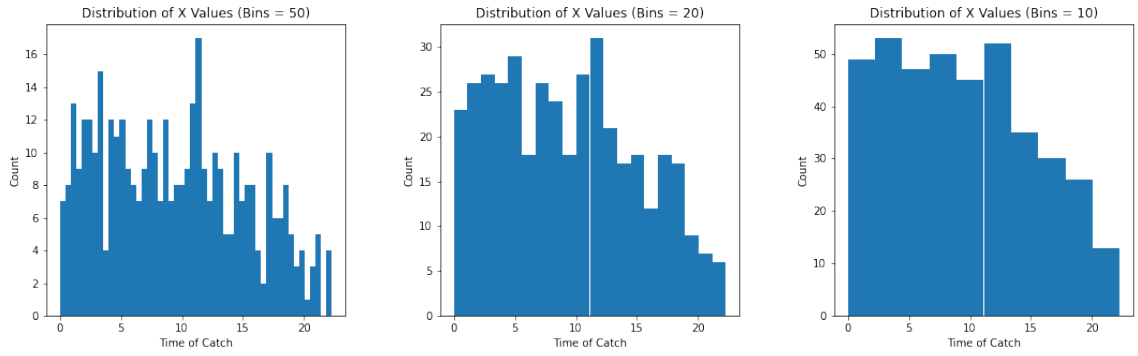



Figure 1: Histograms of the Catching Time Distribution

Here a series of histograms with distinct *bins* values of 50, 20 and 10 were drawn to show the distribution of $X$ values. Considering the noises and the necessary data tendency, the histogram with *bins* value of 20 is generally the best choice on this problem. For the $X$ values of the experiment data, the range is $[0.01, 22.27]$, the value of mean is 9.37, the median is 9.02, the standard deviation is 5.80 and the Interquatile Range is $[4.33, 13.75]$. Given the complexity of the real problem and the possible representativeness of the data point, outlier elimination is not discussed under this circumstance. Because of the high level of dispersion degree reflected by the value of the standard deviation, the mean value is indicating the central tendency of the data to be centering around an $X$ value intervals of $[9, 9.4]$ but is representing the general degree limitedly. And the histogram is showing approximate intervals of $[1, 5]$ and $[10, 14]$ with the most data falling in them.

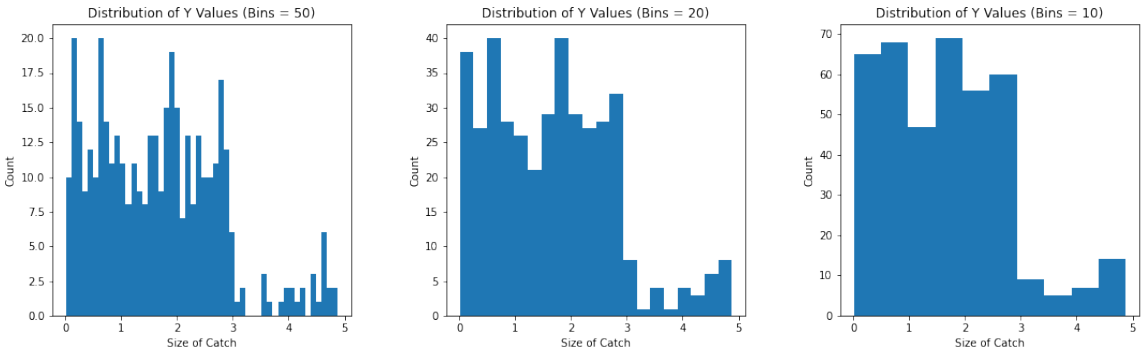



Figure 2: Histograms of the Catching Size Distribution

Similarly, the histogram with *bins* value of 20 is the best choice to describe the distribution of $Y$ values as well. For the $Y$ values of the experiment data, the range is $[0.01, 4.88]$, the value of mean is 1.67, the median is 1.62, the standard deviation is 1.11 and the Interquatile Range is $[0.71, 2.40]$. The outliers here are mainly catches of bigger sizes that are important information for the problem and therefore need to be taken into consideration. Affected by the extreme data and the dispersion degree though, the mean of $Y$ values show the general level of data distribution to some extent. However, the intervals covering the most data are still need to be discussed to better describe the central tendency of data distribution.

Further, plots of Kernel Density Estimation and the Fitting Curving of normal distribution were generated to describe the shape of the distributions as below[7]. According to numerical analysis and the comparison between the curves, the skewness values of the two distributions are 0.27 and 0.65, indicating both of the two distributions to be right-skewed distributions and the mode are likely to locate on the left of mean.

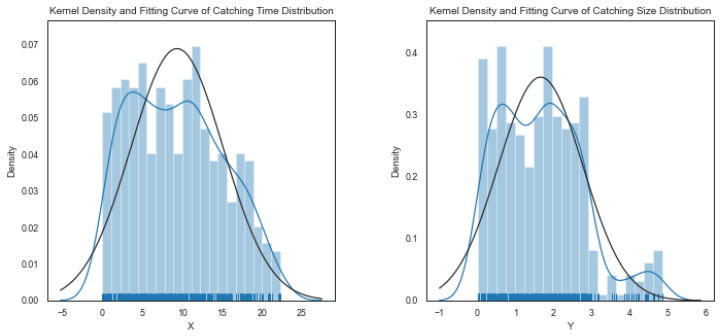



Figure 3: Kernel Density and Fitting Curve of Catching Data Distributions

The kurtosis values of the two distributions are -0.95 and 0.16, indicating that the distribution of $X$ is flatter than the normal distribution with fat tails on the right while the distribution of $Y$ is steeper than the normal distribution with thin tails on the right. From the perspective of Kernel Density Estimation, the data is approximately gathering at the intervals of $[1, 5]$ and $[10, 13]$ for variable $X$ and the intervals of $[0, 0.8]$ and $[1.4, 1.8]$ for variable $Y$, which actually reflects the concentration of data.
Assuming that the data are a sample of a large population and the data of the population are of normal distribution, with the known standard deviation 5.796 of distribution $X$ and 1.108 of distribution $Y$, mathematical calculation were carried out to acquire the unbiased estimate of the standard deviation of the whole population and the results are 5.804 and 1.110 corresponding to the standard deviation of $X$ and $Y$ in the large population. Illustrating that the unbiased estimate of the mean of the whole population equals the mean of the samples, here the 95% confidence intervals for the mean of the two distributions are figured out with the known information. The 95% confidence intervals of the distribution of $X$ for the mean values is $[-2.00, 20.75]$ and the confidence intervals of $Y$ is $[-0.51, 3.84]$. Apparently, both of the two mean values are positive and thus there is a probability of 95% for the mean value of $X$ to fall in the intervals of $(0, 20.75]$ and for the mean value of $Y$ in $(0, 3.84]$.

Finally, to analyze the effectiveness of each type of bait, scatter plots describing the data distributions were generated as below:

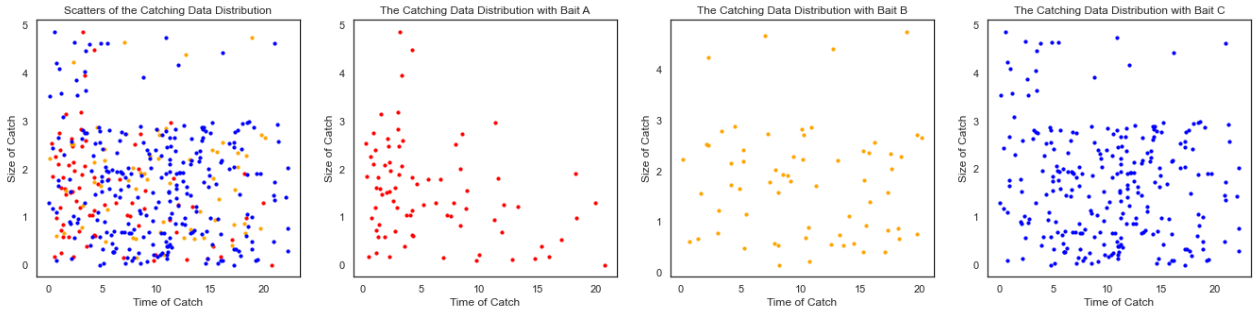
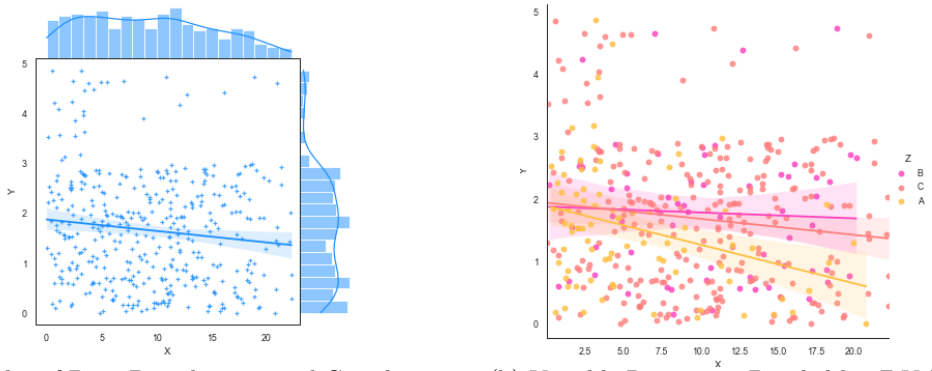



Figure 4: Scatters of the Catching Data Distribution

Illustrating that histograms about $X$ and $Y$ values separated by bait type could absolutely show the distribution details of each variable clearly though, the scatter plots here are actually giving the overall distribution of all data containing both the variables and thus a better choice. According to the observation of the scatter plots and mathematical statistics, there are 79 data points of bait $A$ in red, 64 of bait $B$ in orange, and 257 of bait $C$ in blue. Estimating the probability of the population based on the sample frequency, it is indicated that the bait $C$ has the largest possibility of about 0.65 for a successful fishing in the whole day. And the mean values of $Y$ in the three distributions are 1.52, 1.78 and 1.68 which are approximate, but it is more likely to make a catch with bait $C$.

## 3.3 Correlations between the Variables

In the second part of the experiment, it was aimed to work out the correlation between the variables and analyze the dependence of the time, the size of the catches and the bait used. The figures are as below:




(a) Joinplot of Data Distributions and Correlation (b) Variable Regression Divided by $Z$ Value

Figure 5: Correlation and Regression of the Variables

The first part of Figure 5 gives a joint plot of the variable distribution histograms and the linear regression fitting model. It is observed that the data points are distributed irregularly and the absolute value of the slope curve of the linear regression is low, showing that there is no obvious correlation between $X$ and $Y$. More specifically, using numerical calculation with Pearson function, it is found that the value of the correlation coefficient between $X$ and $Y$ and its p-value is $(-0.12, 0.015)$, indicating the fine statistical significance of the correlation but the very weak relationships between the variables in this case[8]. However, the degree of data sparseness and dispersion are changing with the $X$ value. Sorting the data points into four parts based on dividing the $X$ values intervals $[0.01, 22.75]$ into four continuous equidistant intervals, the main features of data in each intervals are as below:

| $X$ Intervals | Count | Mean($Y$) | Variance($Y$) | Range($Y$) |
|---|---|---|---|---|
| [0.01, 5.69] | 132 | 1.97 | 1.44 | [0.01, 4.88] |
| (5.69, 11.38] | 124 | 1.41 | 1.00 | [0.04, 4.75] |
| (11.38, 17.07] | 90 | 1.63 | 1.01 | [0.04, 4.43] |
| (17.07, 22.75] | 54 | 1.59 | 1.28 | [0.01, 4.75] |

Table 1: Main Features of $Y$ Distributions Divided by $X$ Intervals

It can be inferred from the statistics that generally the amount of data points, the mean, the variance and the range of $Y$ values of the data are decreasing with the increase of the $X$ value, indicating the decrease of the distribution density and the data dispersion degree. Through reasonable inference based on the developing tendency of the growing data centralization, the amount of $Y$ value is actually decreasing with the increase of $X$.

The second part of Figure 5 aims to find out the influence of bait types on data distribution. Although the linear regression curves are showing differences between data with distinct $Z$ values, the correlation coefficients and p-values of the three groups of data with $Z$ values of $A$, $B$ and $C$ are $(-0.32, 0.004), (-0.05, 0.697)$ and $(-0.13, 0.044)$, all indicating weak relationships between the variables or low statistical significance of correlation . The linear regression model is not reliable in all the cases above. Based on the observation of the data distributions, the data points with $Z$ value of $A$ are likely to fall in the $X$ value intervals of $(0, 5)$ and are falling in $(5, 8)$ and $(12, 14)$ with $Z$ value of $C$. Also, the data with $Z$ value of $A$ are distributed less discetely and has an fuzzy central tendency around the point $(3, 1.5)$ while the others are comparatively dispersedly distributed with no obvious central tendency.

# 4 Conclusion

In this paper, specifically focusing on the fishing data set, the experiment analysed the distributions and correlations of the variables. To integrate the experiment results and give available suggestions on future fishing, general conclusions are made as follows.

Firstly, according to the histogram with *bins* value of 20 of the distribution and Kernel Density of $X$, the data distribution generally has two peaks located in the intervals of $[2.23, 5.57]$ and $[10.03, 13.37]$. Based on the sample frequency, it can be predicted that the mode of the population is falling into the two intervals above of the highest probability. For further consideration, from the perspective of catching size, it can be observed from the scatter plot of the data distribution that there are more data points having bigger $Y$ values with $X$ values falling in the intervals of $[2.23, 5.57]$ than in $[10.03, 13.37]$; numerically analyzing the data, the mean values of $Y$ of the data points in the two intervals are 2.04 and 1.60, indicating that the data points with $X$ value in the intervals $[2.23, 5.57]$ have bigger expectation of catching size. And this can be used to infer the overall situation of the population. Combining the two points above, the best time to go fishing at this lake can be from 2 to 5:30 in the morning.

Secondly, according to the statistics and the scatter plot of sample distribution, there is a most amount of 257 catches of 400 with bait $C$ in the whole day. Using sample frequency to estimate the population, bait $C$ has the biggest probability for a successful catch. More specifically considering the size of the catch, the mathematical expectation of the catches for each bait equals the probability multiplied by the mean value. With the known mean values of $Y$ with each bait and the probability estimated by sample frequency, it is figured out that the expectations of fishing with bait $A$, $B$ and $C$ correspond to the values of 0.30, 0.28 and 1.08, which actually reflect the effectiveness of bait. In general, bait $C$ is the most effective for fishing at this lake the whole day.

Finally, to discuss the general situation at 3pm in the afternoon, since the linear regression model is not reliable in the case, data with $X$ value within the intervals of $[14, 16]$ are specifically sorted out as the study objects and there are 37 data points with 3 of bait $A$, 8 of bait $B$ and 26 of bait $C$. Similarly figuring out the fishing expectation, and the results are correspondingly 0.02, 0.30 and 1.22 for bait $A$, $B$, $C$. In another way, taking 20 data points with $X$ values nearest to 15, there are 2, 4 and 14 data points of bait $A$, $B$, $C$ and the result of expectation comparison is similar. Considering the reliability of prediction, taking less data points is infeasible while taking more data points will take in noises. To conclude, Bait $C$ is the best choice to use at 3pm in the afternoon.

# References


[1] Longyu Chen. Analysis on the Measurement of Data Distribution Characteristics [J]. Northern economy and trade, 2014(08):192

[2] Xiaosu Sun. Analysis of Several Problems in Central Tendency Measurement [J]. Journal of Lanzhou Business School, 2009,25 (04): 107-111

[3] Haijian Wu. My Opinion on Mode Calculation: Discussion on the Central Tendency Measurement of Numerical Data Sets [J]. Statistical research, 2002 (08): 58-60. Doi: 10.19343/j.cnki.11-1302/c.2002.08.015

[4] Ali Z, Bhaskar S B, Sudheesh K. Descriptive statistics: Measures of Central Tendency, Dispersion, Correlation and Regression. Airway 2019;2:120-5

[5] Doric D, NikolicDoric E, Jevremovic V, Malisic J. On Measuring Skewness and Kurtosis. QUALITY and QUANTITY, 2009(06);3:481-493. Doi: 10.1007/s11135-007-9128-9

[6] Estimating Population with Samples [J]. New century intelligence, 2021 (ZC): 78-80

[7] Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. Scandinavian Journal of Statistics. 9 (2): 65–78. JSTOR 4615859

[8] Xiaotian Zhang. The Meaning of Significance Level [J]. Sociological research, 1997 (2): 6