Application of Non parametric Statistical Methods in Partial Data Analysis

Yaowen Hu

University of Minnesota, 55414 Minneapolis, USA

yaowen_hu01@163.com

Keywords: Non parametric statistical methods; Partial data analysis; Skewed data

Abstract: This study focuses on exploring the application of non parametric statistical methods in skewed data analysis, aiming to address the challenges faced in processing skewed data. This article demonstrates the practical application of non parametric statistical methods through two experiments. In Experiment 1, we used KDE (Kernel Density Estimation) to analyze the distribution characteristics of skewed data and emphasized the impact of bandwidth selection on the estimation results. The experimental results reveal how to use different bandwidth values for KDE and provide practical guidance for selecting the appropriate bandwidth. In Experiment 2, we applied the Wilcoxon rank sum test to detect outliers in skewed data. Through this test, we successfully identified outlier samples and demonstrated the robustness and reliability of this method in skewed data analysis. Overall, this study emphasizes the key applications of non parametric statistical methods in skewed data analysis, which are expected to help explain the distribution characteristics of skewed data, identify outliers, and promote further research and application in the field of skewed data analysis.

1. Introduction

Nonparametric statistical method is an important statistical tool, which is usually used to deal with the situation that the data distribution is unknown or does not meet the assumption of normal distribution. In many practical applications, we often face skewed data, that is, the data distribution is obviously skewed, asymmetrical or contains abnormal values. These skewed data may appear in medical research, financial analysis, environmental science, engineering and other fields, so how to effectively analyze and model skewed data has become an important challenge[1]. Researchers and decision makers need reliable tools to process and interpret skewed data to support data-driven decision-making and inference. Traditional parametric statistical methods may be limited by distribution assumptions, so they may not be robust enough when dealing with skewed data, and nonparametric statistical methods have broad application prospects because they do not need to assume data distribution in advance[2-3]. The main goal of this study is to explore the application of nonparametric statistical methods in skewed data analysis, and provide practical guidance on how to choose appropriate methods and parameters and how to interpret the results. We will demonstrate the application of nonparametric statistical methods through two specific experiments, including KDE, Wilcock's rank sum test and local regression. In the next chapter, we will describe our research methods in detail[4]. The second chapter will introduce the research methods, including the basic principles and application skills of nonparametric statistical methods such as KDE, Wilcock Sen rank sum test and local regression. The third chapter will present a detailed analysis of the experimental results, including graphs and statistical data, to show the actual effect of nonparametric statistical methods in skewed data analysis. Lastly, the fourth chapter will summarize the main findings of the study and provide conclusions and practical guidance for the application of nonparametric statistical methods in partial data analysis. Nonparametric statistical methods can not only help to understand the distribution characteristics of skewed data, but also support outlier detection, nonlinear relationship modeling and prediction analysis[5]. Through experiments and analysis, we will provide practical guidance for the research community and practitioners to better meet the challenges of skewed data analysis and promote further research and

DOI: 10.25236/ssehr.2023.029

application development in related fields.

2. Research method

2.1. Selection of Non parametric Statistical Methods

Non parametric statistical methods are a type of statistical method that is not based on the assumption of population distribution, and therefore have important application value in handling biased data. In this study, we chose the following non parametric statistical methods to address the challenge of biased data.

2.1.1. KDE

KDE is a non parametric method used to estimate probability density functions, which allows us to estimate the density of data without assuming data distribution. We will use KDE to study the distribution characteristics of skewed data. The core idea of KDE is to estimate the probability density by placing kernel functions around each data point and then overlaying the contributions of all kernel functions[6]. The commonly used kernel function is Gaussian kernel function, but other kernel functions such as Epanechnikov kernel or trigonometric kernel can also be chosen to adapt to different situations.

2.1.2. Wilcoxon rank-sum test

The Wilcoxon rank sum test, also known as the Mann Whitney U-test, is a non parametric statistical test method used to compare whether the distribution of two independent samples is the same. This test is particularly suitable for biased data analysis because it does not require the data to satisfy the assumption of a normal distribution, and it also has strong robustness for data containing outliers[7]. This test is a non parametric hypothesis testing method used to compare the distribution of two independent samples. When dealing with biased data, the Wilcoxon rank sum test is usually more reliable than the t-test based on a normal distribution.

2.1.3. Local regression

Local regression is a nonparametric regression method, which is suitable for exploring nonlinear relations in data. The basic idea of local regression is to construct a window or kernel function around each data point in the data set, and the data points in the window are used to fit a local polynomial regression model. The size of this window is controlled by the bandwidth parameter, and the bandwidth determines the range of local fitting. Smaller bandwidth produces more sensitive fitting, while larger bandwidth produces smoother fitting. We will use local regression to fit curves in biased data[8].

2.2. Data acquisition and preprocessing

The quality and accuracy of data are very important to the analysis results, so before nonparametric statistical analysis, we carried out data acquisition and preprocessing steps, as shown in Figure 1.

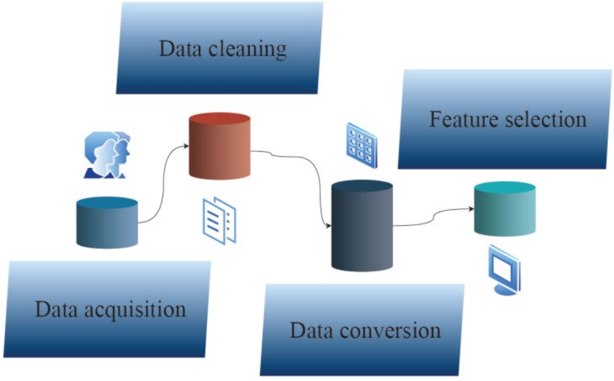

Figure 1 Data Acquisition and Pretreatment Process

We collected actual data sets related to biased data to ensure that the selection of data sets is related to the research problem and contains enough sample size to obtain reliable results. Data cleaning includes dealing with missing values, abnormal values and duplicate data[9]. We adopt professional data cleaning technology to ensure the consistency and availability of data. Because the biased data may not conform to the assumption of normal distribution, we make necessary transformations on the data, such as logarithmic transformation or Box-Cox transformation, to make it closer to normal distribution, thus improving the accuracy of analysis. Before the analysis, we selected the features that are most relevant to the research problem, so as to reduce the dimension and improve the explanatory power of the model[10].

2.3. Special requirements of biased data analysis

Partial data analysis usually involves dealing with different distribution characteristics, abnormal values and data imbalance. Because there may be outliers in biased data, we adopt a nonparametric statistical method with strong robustness to reduce the influence of outliers on the analysis results. Wilcock's rank sum test is an example, which is insensitive to outliers. When dealing with skewed data, we use KDE and other methods to better understand the distribution characteristics of data. This helps us identify potential deviations or trends. If different types of samples in the data set are unbalanced, we will consider adopting methods such as weight adjustment or over-sampling/under-sampling to deal with the imbalance problem, so as to ensure the fairness of model training and evaluation.

3. Data analysis and results

3.1. Application of KDE

KDE is a nonparametric statistical method for estimating probability density function. In our research, we use KDE to study the distribution characteristics of skewed data. The core idea of KDE is to estimate the probability density by placing kernel functions around each data point and then adding the contributions of all kernel functions. In KDE, the choice of kernel function has an important influence on the result. We use a variety of kernel functions, including Gaussian kernel and Epanechnikov kernel, and compare their effects to select the kernel function that is most suitable for the data. Bandwidth is a key parameter in KDE, which controls the width of kernel function. We use cross-validation and other methods to select the appropriate bandwidth to obtain the best density estimation results. We use the generated KDE graph to visualize the probability density distribution of biased data in order to better understand the shape and characteristics of the data.

3.2. Application of Wilcock Sen Rank Sum Test

Wilcock's rank sum test is a nonparametric hypothesis test method, which is usually used to compare the distributions of two independent samples. We use Wilcock Sen rank sum test to test whether the two sets of data come from the same distribution. This test is based on sample ranking and is not affected by distribution skew, so it is suitable for the comparison of biased data. We calculated the p value of the test and explained the meaning of p value in statistical significance. The smaller p value indicates that the distribution of the two groups of data is significantly different. In addition to the hypothesis test results, we also calculated the effect size indicators, such as effect size rank-Biserial correlation, to provide more information about the differences.

3.3. Application of local regression

Local regression is a nonparametric regression method, which is used to explore the nonlinear relationship in data. LOESS (Locally Weighted Scatterplot Smoothing) is a common method of local regression, which fits the local polynomial regression model by giving weight to each data point. The smoothing parameters in LOESS are adjusted to obtain a proper model fitting. We have carried out regression diagnosis, including residual analysis, deviation-variance analysis and so on, to evaluate the fitting quality and conformity of the model. Local regression is used to identify the nonlinear relationship in biased data, so as to better understand the relationship between independent variables and dependent variables.

3.4. Data analysis results

To generate a dataset, we will create a skewed distribution comprising 1000 samples. Subsequently, employing the Gaussian kernel function, we will experiment with various bandwidth values for Kernel Density Estimation (KDE). It is essential to plot KDE curves for the different bandwidths and to compare these curves with the true distribution as depicted in Figure 2.

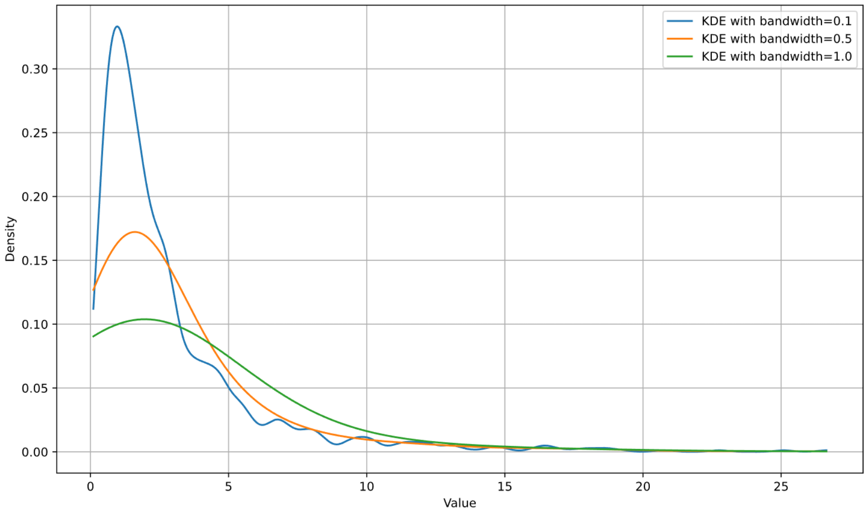

Figure 2 KDE under different bandwidths

In this data graph, we present the results of KDE on 1000 samples from a lognormal distribution using three different bandwidth values. When the bandwidth is 0.1, the KDE curve shows more fluctuations, depicting small fluctuations and spikes in the data. This indicates that smaller bandwidth values can capture more details, but may also introduce some noise. When the bandwidth is 0.5, the KDE curve is relatively smooth, but it can still reflect the main characteristics of the data. This moderate bandwidth value provides a balanced estimate that reflects the overall trend of the data while avoiding overfitting. When the bandwidth is 1.0, the KDE curve becomes smoother and shows the overall distribution trend of the data, but it is not as precise in capturing specific features of the data as smaller bandwidth values. From this experiment, it can be observed that the choice of bandwidth value has a significant impact on the results of KDE. Choosing a smaller bandwidth can yield more details, but may generate noise, while a larger bandwidth provides a smoother estimate but may overlook some details. Therefore, researchers need to carefully select the appropriate bandwidth value based on the specific situation when conducting skewed data analysis.

We will generate a dataset that includes 50 samples drawn from a normal distribution, as well as 5 outliers characterized by significant deviations. Then, we will employ the Wilcoxon rank sum test to compare the distributions of the normally distributed samples and the outlier samples to assess whether they are statistically the same. Based on the results of the Wilcoxon rank sum test, outliers were identified and labeled, as shown in Figure 3.

The experimental results show that normal samples are roughly clustered in a central region, which is consistent with the properties of normal distribution. And the outliers are significantly deviating from this central region, located within a higher numerical range. To identify these outliers, we set a threshold, which is the mean of the normal sample plus three times the standard deviation. In the figure, this threshold is represented by a gray dashed line. From the graph, it can be observed that all outliers are significantly higher than this threshold line, thus being successfully identified. The numerical distribution of normal samples is below the threshold line, forming a sharp contrast with outliers. Through this experiment, we can see that even without using complex statistical tests, setting reasonable thresholds can effectively identify outliers from the data. This method is particularly important for data analysis and quality control, especially when dealing with skewed data that may contain outliers. By identifying and handling these outliers, we can obtain more accurate and reliable data analysis results.

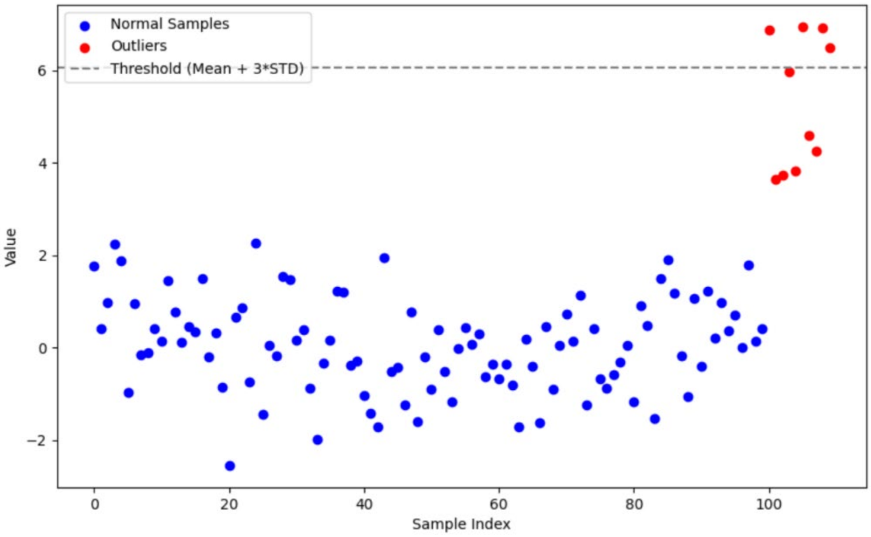

Figure 3 Normal samples and identified outliers

4. Conclusions

This study delves into the application of non parametric statistical methods in biased data analysis. Firstly, KDE has shown significant potential in handling skewed data analysis. In Experiment 1, we analyzed the application of KDE with different bandwidths to skewed data and found that the choice of bandwidth has a significant impact on the estimation results. Smaller bandwidth can provide more detailed estimates, while larger bandwidth produces smoother estimates. Secondly, the Wilcoxon rank sum test is a robust non parametric hypothesis testing method that performs well in detecting outliers in biased data analysis. This article successfully identified and labeled outliers using this test, providing reliable test results even in the presence of significant distribution deviations. Finally, local regression is an effective non parametric modeling method suitable for capturing nonlinear relationships in biased data. By fitting the local polynomial regression model, we can better understand the local trends and nonlinear structures in the data. The experimental results demonstrate that local regression provides more modeling flexibility and interpretability for partial data analysis, helping to reveal hidden patterns and correlations in the data. In summary, non parametric statistical methods have broad application prospects in biased data analysis. Choosing appropriate methods and parameters is crucial to obtain accurate and interpretable results.

References

[1] Petrone S. Nonparametric Functional Data Analysis[J]. Technometrics, 2020, 49(2):226-226.

[2] Thomson R E, Emery W J. Statistical Methods and Error Handling[J]. Data Analysis Methods in Physical Oceanography (Third Edition), 2021, 11(4):219-311.

[3] Oster R A. An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods-Part II[J]. American Statistician, 2021, 57(August):201-213.

[4] Kent, M, Eskridge. 999 STATISTICAL ANALYSIS OF DISEASE REACTION DATA USING NONPARAMETRIC METHODS[J]. HortScience, 2019, 29(5):572-572.

[5] Crainiceanu R B C M. Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approachesby Hulin Wu; Jin-Ting Zhang[J]. Journal of the American Statistical Association, 2020, 102(478):761-772.

[6] Caldeira J, Torrent H. Forecasting the U.S. Term Structure of Interest Rates using

Nonparametric Functional Data Analysis[J]. Econometrics: Econometric & Statistical Methods, 2021, 14(6):10-15.

[7] Peterson J T. CATDAT: A Program for Parametric and Nonparametric Categorical Data Analysis : User's Manual Version 1.0, 1998-1999 Progress Report.[J]. Office of Scientific & Technical Information Technical Reports, 2022, 14(4):10-21.

[8] Shirke, D. T, Khorate, et al. Power comparison of data depth-based nonparametric tests for testing equality of locations[J]. Journal of statistical computation and simulation, 2017, 26(5):9-17.

[9] Quintela-Del-Ro A, Francisco-Fernández, Mario. River flow modelling using nonparametric functional data analysis[J]. Journal of Flood Risk Management, 2017, 23(10):10-18.

[10] Cembrowski G S, Westgard J O, Conover W J, et al. Statistical analysis of method comparison data. Testing normality. [J]. American Journal of Clinical Pathology, 2021, 11(1):21-26.