Dean&Francis
ISSN 2959-6157

Predict True Distribution from a Part of Gaussian Distribution

Chengtao Zhang

International department from Beijing National Day School, Beijing, China

Corresponding author: chengtao_zhang@sina.cn

Abstract:

This paper tackles the challenge of predicting dataset properties from a Gaussian-distributed subset, a key issue in statistical analysis and data science. The objective is to estimate the expected value and standard deviation of a comprehensive dataset with solely a subset. The methodology includes hypothesis testing, confidence interval estimation, regression analysis, Bayesian inference, and simulation methods. The study employs statistical models like linear and non-linear regression and Bayesian updating for prediction refinement. Resampling techniques such as bootstrapping and Monte Carlo simulations are used to ensure prediction reliability. By giving the smallest 50 GPA data in the 200 GPA data, the results show the methods' effectiveness in predicting dataset parameters, with detailed calculations indicating the expected value likely falls between 4.073 and 4.173, and variance between 0.098 and 0.540. The precision and dependability of these predictions are emphasized in the study's results, providing a strong basis for additional statistical investigation. Because it offers a thorough methodology for estimating population parameters from small samples, this research is significant for statistics and data analysis. Its findings are valuable in educational research, finance, and other fields dealing with incomplete or skewed data. The paper also highlights the importance of understanding statistical prediction limitations and uncertainties, offering a robust framework for future research and applications.

Keywords: Gaussian distribution; Linear regression; Confidence interval.

1. Introduction

In the realm of statistical analysis, the ability to predict the properties of an entire dataset using a subset is a pivotal skill accurately, particularly when faced with incomplete or partially obscured data. This problem is not only an intellectual one; it has significant ramifications for many academic fields, such as engineering, economics, and social sciences. For instance, in educational research, understanding the

ISSN 2959-6157

overall academic performance of a student body from a sample of available grades can inform educational policy and restrict allocation. Similarly, in finance, predicting market trends from a limited set of data points can guide investment decisions. The importance of this research is further magnified by the increasing reliance on data-driven decision-making in both public and private sectors. This research is important because Gaussian-distributed data are normal in various fields [1]. For example, a registrar may have a record of student grades with a portion of the data be damaged by ink blots or something else. In these situations, the capacity to extrapolate the general expected value and standard deviation from the remaining data is critical for assessing student performance and obtaining accurate statistics. Similarly, in quality control within manufacturing, people can test some products for defects and use this aphorism to get the total data. Then, the results can be used to predict the quality of the entire batch. Accurate predictions in these scenarios can prevent the release of substandard products and maintain consumer trust [2].

This paper provides a result of how to predict the properties of a complete dataset using a Gaussian-distributed subset. In the first section, the paper delves into the advanced statistical theory and methodological framework necessary for understanding the complexities of Gaussian distributions and their application in hypothesis testing and confidence interval estimation, which includes a detailed explanation of the Gaussian distribution, the basics of hypothesis testing, and a thorough description of the Gaussian distribution. The paper then addresses regression analysis in the second section to model the relationship between the parameters of the Gaussian distribution and the subset data. Then the author discusses the Bayesian inference. It offers a probabilistic approach to updating beliefs about population parameters as more data becomes available. In the third section, the paper introduces simulation and resampling methods such as bootstrapping and Monte Carlo simulations. Since those techniques are particularly helpful when dealing with complex data or when the underlying distribution is unclear, they are important for evaluating the accuracy of predictions. The fourth section outlines the algorithm development process for predicting the properties of the entire dataset and details each step from data preprocessing to uncertainty quantification, particularly the detailed algorithm for calculating confidence intervals. Finally, the paper offers an illustration of how the algorithm is used. In the example, only using the lowest 50 scores as the given data, the school predict the standard deviation and expected value of a complete dataset of student GPA scores. This section provides a real-world example of the application of the methodologies discussed, demonstrating the effectiveness of the approach in a tangible scenario. In the end, the paper provides the conclusion. The paper makes a valuable contribution to the field of statistical analysis by providing a solid framework that facilitates problem solving while dealing with partial datasets.

2. Methods and Theory

2.1 Data analyze

With a typical parabolic shape, the expected value (μ) and standard deviation (σ) are the two essential parameters which define the Gaussian distribution. The standard deviation quantifies the dispersion of the data around the expected value, while the mean shows the tendency of the data. The Gaussian distribution can be fully described by its probability density function (PDF) [3]:

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

This function reaches its maximum at $x = \mu$ and declines symmetrically towards zero as x moves away from μ, and the area under the curve representing the total probability, which has a total probability equals to 1.

In this study, the author will use hypothesis testing to determine if the properties of the full dataset can be inferred from the sample data. Specifically, A null hypothesis, and an alternative hypothesis will be established.

Null Hypothesis (H_0) represents postulates that the sample data and the presumptive population parameters do not significantly differ from one another. For this study, H_0 is that the expected value of the entire dataset is equal to the expected value of the sample dataset; then, the alternative Hypothesis (H_1) suggests the population parameters and the sample differ significantly, which expected values there's a different between the expected value of entire dataset and the expected value of the sample dataset [4].

For the test statistic, the author will choose z-statistic. Then, the p-value will receiving test results that are at least as extreme as the observed results if H_0 is correct. In other words, if the p-value is less than 0.05, it can prove strong evidence against H_0. The confidence intervals will provide a range which the true is likely to fall inside with a given degree of confidence, In this case, 95%, is suitable.

For the expected value, the confidence interval can be calculated as

$$\mathrm{CI} = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}} \quad (2)$$

In this equation, z is the z-score which shows the confidence level.

2.2 Regression Analysis

The relationship between subset data and the parameters of the Gaussian distribution can be modeled using regression analysis, and the author can predict the expected value and variance of the entire dataset by fitting a regression model and based on the subset data. First, using the linear regression, the author can model the link between a dependent variable and independent variables. In this case, the sample mean or variance are the dependent variable, and other characteristics of the dataset are the independent variables. Then, if multiple variables are considered, multiple regression can be used to predict the outcome based on several predictors. Lastly, non-linear regression models can be utilized to capture intricate interactions when the relationship is non-linear.

As more data becomes available, Bayesian inference increases the confidence about the population parameters. The author needs to use some useful methods of bayesian Inference the author used in the calculation. Prior Distribution shows the revised belief about the parameters after taking the data and the prior into account is known as the posterior distribution, and the likelihood function measures how well the data fits the model parameters. The posterior distribution is the belief about the parameters after considering the data and the prior; the posterior distribution can be sampled from using Markov Chain Monte Carlo methods, particularly in cases when the posterior is complicated or analytically unmanageable [5].

Simulation and resampling methods can be used to assess the accuracy and reliability of the predictions. Therefore, those methods help make sure the predictions calculated by this algorithm are accurate. Therefore, the author uses some simulation and resampling method. First, resampling the data with replacement to create many simulated datasets is important, and the properties of these simulated datasets can be used to estimate the sampling distribution of the sample statistics. And in order to understand the variability and uncertainty in the predictions, the author use Monte Carlo simulations to model the entire dataset and generate synthetic datasets based on the assumed Gaussian distribution, and cross-validation techniques can be used to evaluate the predictive performance of the model by dividing the data into training and validation data sets.

In the paper, the author is going to development an algorithm to predict properties of the entire dataset. The author uses several steps predict the data set. First, the author preprocesses the data, which means cleaning the data, finding missing values, and standardizing the data. The second step is visualizing the data, calculating some statistics, and finding patterns and outliers, which is called Exploratory Data Analysis (EDA). Then based on the EDA, an appropriate statistical model is chosen for the prediction. Then the model is trained on the subset data to estimate the parameters, and then model's performance can be proved using cross-validation, and all of the dataset's properties are predicted using the trained model. Finally, the predicted values are post-processed to ensure they are within the plausible range and consistent with the observed data, and the author quantify the uncertainty in the predictions.

2.3 Calculation

The author needs to get the expected value and variance at the beginning

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (3)$$

Here, $\bar{x}$ represents the sample mean of the dataset, and $\sum_{i=1}^{n} X_i$ aggregates all individual data points X_i, and calculating the average value by dividing by the total number of observations. The average value is a central measure of the dataset's central tendency

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \quad (4)$$

The sample variance s^2 is calculated by taking each data point X_i, subtracting the sample expected value $\bar{x}$, squaring the result to eliminate negative values and emphasize larger deviations, and then averaging these squared differences. The $n-1$ in the denominator corrects for bias (Bessel's correction), providing an unbiased estimate of the population variance. Then, the author should determine the Confidence Level, choosing the confidence level 95% for the confidence interval.

To calculate the Standard Error for the expected value, it is found that standard error is $s/\sqrt{n}$, which is divided by the sample size's square root, the sample mean's standard deviation is the standard error. It gauges the accuracy of the sample mean as an approximation of the population mean by measuring the standard deviation of the expected value in feasible samples of a specific size in the population. To find the critical values, the author should determine the critical values z or t for the chosen confidence level based on the standard normal distribution or the t -distribution, respectively.

Then, one needs to calculate the confidence interval for

ISSN 2959-6157

the expected value $\mathrm{CI}_{\mu} = \bar{x} \pm z \times \frac{s}{\sqrt{n}}$ with a given degree of confidence, the range that the population mean is expected to lie inside is provided by the confidence interval for the expected value. The $\pm z \times \frac{s}{\sqrt{n}}$ represents the margin of error, where z is the z-score corresponding to the desired confidence level, indicating how far from the sample expected value the author should extend to capture the population expected value with the given probability.

To calculate the confidence interval for the variance [6]

$$\mathrm{CI}_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right) \tag{5}$$

This formula calculates the confidence interval for the population variance σ^2. It uses the chi-square distribution because variances are inherently positive and can vary widely, particularly in small samples. From the chi-square distribution, the $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are critical values which define bounds of the interval within which the true population variance fall into a given confidence level. In the end, the author should interpret the results: The confidence intervals provide a range where the genuine population parameters are predicted to lie inside.

3. Results and Application

3.1 Preliminary Analysis

In this study, the author presents with a dataset comprising the GPA scores of 200 students from a school. Unfortunately, three-quarters of the data have been lost, leaving people with only the lowest 50 scores. This task is to predict the expected value and standard deviation of the entire dataset using the remaining subset. This scenario provides a practical application of the Gaussian distribution and the algorithms developed in the "Methods and Theory" section.

The subset of data available is as follows: $[4.05, 4.56, 4.24, 4.19, 3.66, 4.34, \ldots, 4.43, 4.02]$. This array consists of the lowest 50 GPA scores. This goal is to predict the expected value (μ) and standard deviation (σ) of the entire dataset of 200 students.

First, the author calculates the sample mean ($\bar{x}$) and sample variance (s^2) of the available subset: (Tex translation failed) and (Tex translation failed). One can calculate the sample mean as

$$\bar{x} = \frac{4.05 + 4.56 + 4.24 + \ldots + 4.02}{50}, \tag{6}$$

and calculate the sample variance as

$$s^2 = \frac{1}{49}\left((4.05 - \bar{x})^2 + (4.56 - \bar{x})^2 + \ldots + (4.02 - \bar{x})^2 \right). \tag{7}$$

To estimate the confidence interval for the expected value, the author uses the standard error and the z-score for a 95% confidence level. It is found that standard error is $\frac{s}{\sqrt{50}}$ and confidence interval for means is $\bar{x} \pm 1.96 \times \frac{s}{\sqrt{50}}$. For this variance, the author uses chi-square distribution to find the lower and upper bounds of the confidence interval is $\frac{49 \times s^2}{\chi^2_{0.975,49}}$ and $\frac{49 \times s^2}{\chi^2_{0.025,49}}$.

3.2 Algorithm Implementation and Outlook

First, input the 50 GPA scores into the algorithm, then calculate the sample expected value and variance. After that, compute the standard error of the expected value, and determine the critical chi-square values for the desired confidence level. Then calculate the confidence interval, which includes the confidence interval for the expected value and the confidence interval for the variance. Finally, output the predicted expected value and variance with their respective confidence intervals.

The Sample Expected Value Calculation is given by

$$\bar{x} = \frac{4.05 + 4.56 + 4.24 + \ldots + 4.02}{50} = 4.123 \tag{8}$$

The Sample Variance Calculation is given by

$$s^2 = \frac{1}{49}\left((4.05 - 4.123)^2 + (4.56 - 4.123)^2 + \ldots + (4.02 - 4.123)^2 \right) = 0.132 \tag{9}$$

The Standard Error is $\frac{0.364}{\sqrt{50}} = 0.025$ and the Confidence Interval for Expected value is $\mathrm{CI}_{\mu} = 4.123 \pm 1.96 \times 0.025 = (4.073, 4.173)$. The Chi-Square Values are $\chi^2_{0.975,49} = 68.024, \chi^2_{0.025,49} = 11.171$, and the lower and upper bounds of Confidence Interval for Variance are $\frac{49 \times 0.132}{68.024} \approx 0.098$ and $\frac{49 \times 0.132}{11.171} \approx 0.540$, respectively.

Prediction of Gaussian distribution has so many problems, and the experiment of this paper just make little strides in developing a robust framework for predicting dataset properties from a Gaussian-distributed subset. There are many problems and areas for potential enhancement and future research. One area of improvement of this study is the incorporation of more complex data scenarios. Extending the algorithms to handle such complexities would increase their applicability and robustness in real-world

scenarios and increased the universality of this algorithm. Additionally, the integration of machine learning techniques could offer a more dynamic approach to predicting dataset properties. Machine learning algorithms, particularly those that can handle missing data, could be trained on large datasets to improve the accuracy and efficiency of predictions.

Future research could also focus on the development of more sophisticated simulation and resampling methods. As computational power increases, more complex simulations can be run. Therefore, people can decrease the variability and uncertainty in the predictions. In conclusion, this paper provides a briefly analysis of the methods of predicting the properties of a complete dataset using a Gaussian-distributed subset. The paper shows some finding, but there's still many problems which can be discovered, solved, and benefit mankind.

4. Conclusion

This paper has presented a comprehensive study on the prediction of dataset properties using a Gaussian-distributed subset. The author began by outlining the advanced statistical theory and methodological framework necessary for understanding the complexities of Gaussian distributions and their application in hypothesis testing and confidence interval estimation. The detailed explanation of the Gaussian distribution, the principles of hypothesis testing, and the calculation of confidence intervals provided the foundational knowledge required for the subsequent analysis.

The paper delved into the application of regression analysis, both linear and non-linear, to model the relationship between the subset data and the parameters of the Gaussian distribution. The use of Bayesian inference was also explored, offering a probabilistic approach to updating beliefs about population parameters as more data becomes available. This section highlighted the importance of combining prior knowledge with observed data to form a posterior distribution, which is crucial for making informed predictions. Simulation and resampling methods, such as bootstrapping and Monte Carlo simulations, were introduced as important tools for assessing the reliability of predictions. These methods were particularly emphasized for their utility in situations where the data is complex or the underlying distribution is not fully known. The algorithm development process for predicting the properties of the entire dataset was outlined in detail, from data preprocessing to uncertainty quantification. The detailed algorithm for calculating confidence intervals was presented as a critical component of the prediction process, ensuring that the predictions are not only accurate but also robust against potential variations in the data. In the practical application section, the developed algorithms were applied to predict the expected value and standard deviation of a complete dataset of student GPA scores, using only the lowest 50 scores as the subset. This real-world example demonstrated the effectiveness of the methodologies discussed and provided a tangible scenario for the application of the study's findings.

References

[1] Dytso, A., Bustin, R., Poor, H. V., & Shamai, S. Analytical properties of generalized Gaussian distributions. Journal of Statistical Distributions and Applications, 2018, 5(6): 1-40.

[2] Montgomery, D. C., & Runger, G. C. Applied statistics and probability for engineers. John Wiley & Sons, 2010.

[3] DeGroot, M. H., & Schervish, M. J. Probability and statistics. Pearson Education, 2012.

[4] Altman, Naomi and Martin Krzywinski. Simple Linear Regression. Nature Methods, 2015, 12(11): 999–1000.

[5] Wiedermann, Wolfgang, et al. Heteroscedasticity as a Basis of Direction Dependence in Reversible Linear Regression Models. Multivariate Behavioral Research, 2017, 52(2): 222–41.

[6] Lili Ilma and Anxhela Kosta. Applying an Ordinary Least Squares (OLS) Regression Model On Processed Air Quality and Environment Data. British Journal of Environmental Sciences, 2024, 12(2): 49–58.