Application of Nonparametric Statistics on Stock Trading Volume Distribution

Yaqiang Fan[1,*], Ximing Cheng[2]

Corresponding author: 2022021058@bistu.edu.cn*, Chengxm@bistu.edu.cn

School of Science, Beijing Information Science and Technology University, Beijing 100096, China

Abstract. Stock trading volume, as a key indicator reflecting market trading activity, is widely used in market trend analysis and investment strategy formulation. In recent years, many scholars have conducted research on it, and the existing methods mainly include parameter statistics and machine learning algorithms. This article applies nonparametric statistical methods such as kernel density estimation and the single sample Kolmogorov- Smirnov (KS) test to analyze the daily hourly trading volume of A-shares. It is found that the Johnson-U distribution has the best fit to the data, which is of great significance for understanding and analyzing the trading behavior of the stock market.

Keywords: Stock trading volume; Nonparametric statistics; Kernel density estimation; Kolmogorov-Smirnov test

1 Introduction

The stock market, as a core component of the financial system, plays multiple roles such as resource allocation, financing, and investment. Many domestic and foreign scholars have conducted research on the distribution of various indicators of stocks. Mandelbrot and Fama [1,2,3] proposed that although normal or Gaussian distributions are the most familiar and easily computable stable distributions, they are often assumed to describe actual distributions. However, asset returns often exhibit thicker tails or higher kurtosis than Gaussian distributions, which naturally leads people to consider non-Gaussian stable distributions. The research of Zhao and Zeng [4] shows that the peak and thick tail characteristics of stock returns make the Laplace distribution more accurate in describing its distribution characteristics than the normal distribution. Similarly, Wang et al. [5] found through empirical research that stable distributions can more effectively handle the thick-tailed distribution characteristics of stock returns. The research of Wang and Song [6] revealed the peak and thick tail distribution characteristics of the Chinese stock market and pointed out that the investment risk in the Shenzhen market is relatively high. Qiu [7] proposed a nonlinear statistical model that describes the relationship between returns and trading volume changes in the stock market and pointed out that the return distribution can converge to a linearly stable distribution. Hu and Han [8] further developed this model and explored the dynamic relationship between stock returns, historical trading volume, and other stocks. Uchaikin and Zolotarev [9] theoretically discussed the hypothesis that financial asset returns may follow a stable distribution, emphasizing the application of the central limit theorem in financial models.

ICBBEM 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347465

In order to more accurately capture the peak and thick tail characteristics of stock returns and their risks, multiple studies have adopted nonparametric kernel density estimation methods. Ren and Zhang [10] characterized the asymmetric tail correlation between financial assets by constructing a Copula Kernel model and conducted an in-depth analysis of portfolio investment risks in the Chinese stock market. Researchers such as Liu and He [11], Wang and Wang [12] have revealed the distribution changes of stock returns under the influence of the limit up and limit down system through kernel density estimation, as well as the characteristics of the Shanghai and Shenzhen stock markets in terms of returns and risks. Gao [13] and Xia [14] conducted an accurate analysis of the volatility of stock returns in specific industries by applying nonparametric estimation methods. The study by Niu and Hu [15] demonstrated the effectiveness of nonparametric estimation methods in expressing the distribution of A-share returns in the Shenzhen and Shanghai stock markets. Huang's [16] empirical research further proves the advantage of non-parametric kernel density estimation over normal distribution in capturing market risk characteristics.

In recent years, multiple studies have explored the distribution characteristics of stock market returns and their changes in different economic cycles through statistical testing methods. Yan et al. [17] found that during the periods of 2004–2006 and 2007–2009, the daily logarithmic return of stock indices was more suitable to be fitted with a t-distribution rather than a normal distribution, especially during financial crises when the risk of stock indices significantly increased and the degree of freedom value was usually less than 2. Li et al. [18] verified the non-normality, homogeneity, and high correlation of the returns of the Chinese and American stock markets through statistical analysis and found that non-parametric kernel density estimation can more accurately reflect market risk characteristics. Wang's [19] research shows that the mixed Clayton Gumbel Frank copula function can effectively describe the asymmetric tail correlation between the Shanghai Composite Index and the CSI 300 stock index futures. The work of Cao and Chen [20] confirmed the existence of weekend effects in the Shanghai and Shenzhen stock markets and explored their causes. The studies of Han and Yan [21], as well as Yan et al. [22], further support the non-normality of the distribution characteristics of stock market returns from the perspectives of high dimensional dynamic Vine Copula modeling preparation and volatility similarity in the stock market. The research by Du and Zhang [23] supported the use of the t distribution to describe stock return data through the KS test.

With the development of the stock market, trading volume has gradually become one of the most important indicators for market participants and investors. [24,25] However, research on the distribution types of trading volume in the Chinese stock market is still relatively limited. Previous studies have mainly relied on parameter statistical methods, usually based on assumptions about specific distribution types. However, finding the distribution type that best fits the distribution of stock trading volume is very difficult, and this method has limitations when facing market complexity and volatility. This article aims to analyze the distribution types of daily hourly trading volume in A-shares through non-parametric statistical methods, especially kernel density estimation and the single-sample KS test. They do not rely on assumptions about specific distribution types and are more suitable for analyzing complex data. By combining multiple continuous distribution types, they can comprehensively explore the essential characteristics of stock trading volume data, thereby enhancing their understanding of the operating laws of the stock market.

2 Data analysis

As of March 2024, there are a total of 3054 stocks in the A-share market, and the collected data is the daily hourly trading volume of A-shares (all with over 4700 data), sourced from Dongfang Wealth Network. With randomly selected data on the trading volume of 601318 stocks on March 12, 2024, and calculated statistical indicators such as central trend and dispersion, the distribution of stock trading volume can be preliminarily understood. The example results are shown in Table 1:

Table 1. Statistical indicators of Stock trading volume for 6013181

Stock code	Average value	Median	Mode	Range	Skewness	Kurtosis
601318	143	79	30	15791	27	1305

From the above indicators, it can be seen that a skewness of 27 positive values indicates a longer right tail of the data distribution, with more extreme values on the right and fewer extreme values on the left. This means that most of the data is concentrated on smaller values. The kurtosis is 1305, indicating an extremely sharp distribution of data. According to the range of 15791, it is speculated that there are outliers in the data. With the help of the box plot, it can be visually seen as shown in Figure 1. Further draw a histogram to remove the largest outliers (which do not affect the overall distribution trend), as shown in Figure 2. According to the indicator results and histogram, it can be seen that the distribution of stock trading volume does not follow a normal distribution.[26]

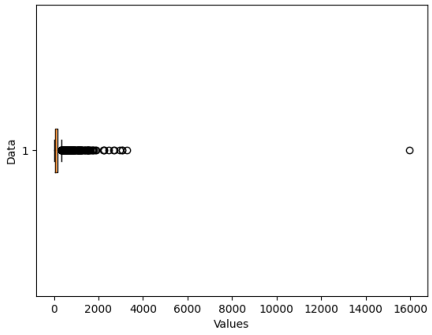


Fig. 1. 601318 Stock trading volume Box chart

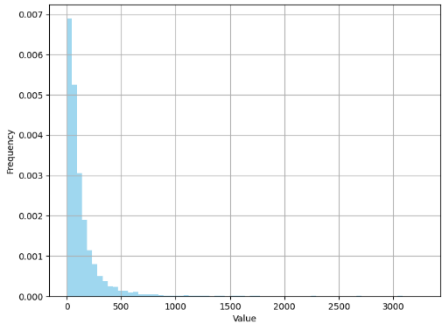


Fig. 2. 601318 Stock trading volume Histogram

3 Kernel density estimation

3.1 Kernel density estimation theory

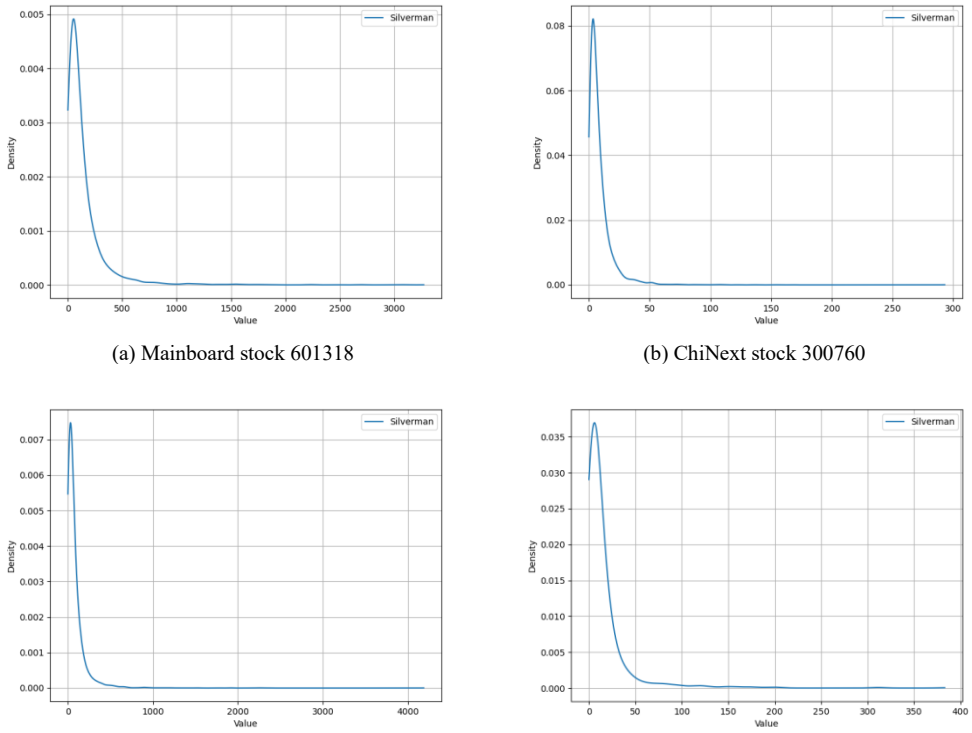
Kernel density estimation is a commonly used nonparametric estimation method that does not require any assumptions about the data. By smoothing the density function of the sample data, the overall density function can be obtained.[27] The kernel density estimation equation (1) is as follows:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{1}$$

where $K(\cdot)$ is the kernel function and h is the bandwidth.

3.2 Experimental results

Multiple datasets were selected for testing on March 12, 2024, including mainboard stocks, ChiNext stocks, small and medium-sized board stocks, and science and technology innovation board stocks. The specific experimental results are as follows:



(a) Mainboard stock 601318

(b) ChiNext stock 300760

(c) Small and medium-sized board stocks 002415

(d) Sci-Tech innovation board stock 688227

Fig. 3. Kernel density estimation images of trading volume for stocks in different sectors

From Figure 3, it can be seen that all four images have only one sharp peak located near 0, indicating a large number of data points concentrated around 0 values. The distribution is right skewed, gradually decreasing to the right (larger values), and the long tail extends to larger values, indicating the presence of some outliers or extreme values in the data. Based on the above characteristics, this distribution may resemble a lognormal distribution, a Pareto distribution, or other long tailed distributions.

4 Single sample KS test

4.1 Single sample KS test theory

Single sample KS test is a commonly used hypothesis testing method used to test whether a sample conforms to a theoretical distribution, using KS statistics to measure the difference between the sample and the selected distribution type. [28] The KS statistic is defined as equation (2):

$$D = \max \left| F_n(x) - F(x) \right|, \tag{2}$$

where $F_n(x)$ is the cumulative distribution function of sample experience, $F(x)$ is the cumulative distribution function of the selected distribution type.

4.2 Experimental results

This study first selected 2400 stocks as the training set, obtained trading volume data at a certain time on a certain day from Dongfang Wealth Network, conducted KS tests on 100 distribution types, and counted the proportion of stocks that met the conditions. The results are as follows:

Table 2. The proportion of stocks of different distribution types in the training set

Distribution type	The proportion of stocks that match the distribution type
Johnson-U distribution	88.75%
Inverse gamma distribution	87.25%
Normal inverse Gaussian distribution	86.88%
Non-central t distribution	86.42%
Johnson-B distribution	85.96%
Generalized inverse Gaussian distribution	83.33%
other	<80.00%

From the results in Table 2, it can be seen that for the distribution types of daily trading volume of stocks, the Johnson-U distribution, inverse gamma distribution, normal inverse Gaussian distribution, non-center t distribution, Johnson-B distribution, and generalized inverse Gaussian distribution have better fitting degrees.

Subsequently, more than 600 other stocks were used as test sets for the same processing, and the results are as follows:

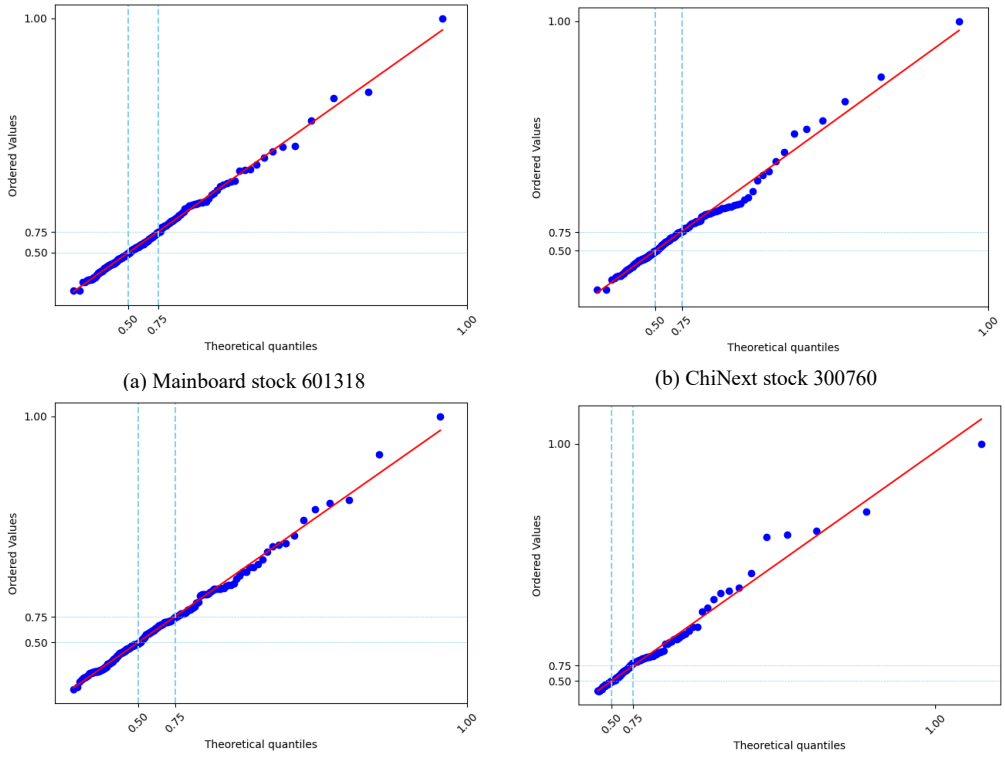
Table 3. The proportion of stocks with six distribution types in the test set

Distribution type	The proportion of stocks that match the distribution type
Johnson-U distribution	88.50%
Inverse gamma distribution	85.83%
Normal inverse Gaussian distribution	86.17%

Non-central t distribution	84.67%
Johnson-B distribution	84.67%
Generalized inverse Gaussian distribution	84.67%

According to the results of the test set in Table 3, the Johnson-U distribution with the highest probability is selected as the distribution type that best fits the daily hourly trading volume distribution of A-shares.

Using the QQ chart to further verify [29], it is still recommended to select stocks from the four sectors shown in Figure 3 above. From the results in Figure 4 below, most of the points of the four stocks are concentrated on a straight line, with some outliers at the tail that deviate from the theoretical distribution line. Overall, the data distribution fits the theoretical distribution well. By comparing the 0.5 and 0.75 quantiles of the data with the corresponding quantiles of the theoretical distribution, the correspondence is good and the data conforms to the selected theoretical distribution.[30]



(a) Mainboard stock 601318

(b) ChiNext stock 300760

(c) Small and medium-sized board stocks 002415

(d) Sci-Tech innovation board stock 688227

Fig. 4. QQ chart of trading volume for stocks in different sectors

The previous analysis was based on the hourly trading volume data of stocks on March 12, 2024. In order to further prove that the Johnson-U distribution is not only applicable to that day but also to data from other dates, considering that only the hourly trading volume data of

the last five trading days can be found on the website, 1000 stocks were randomly selected every day from March 8, 2024, to March 14, 2024 (with 9 and 10 days as rest days and no data). Calculate the proportion of stocks that fit the Johnson-U distribution type, as well as the proportion of stocks that perform well in the inverse gamma distribution on the training and testing sets. For comparison, the results are shown in Table 4.

From the results in Table 4, it can be seen that the proportion of 1000 randomly selected stocks that conform to the Johnson-U distribution from March 8, 2024, to March 14, 2024, remained stable at over 80%, while the proportion that conforms to the inverse gamma distribution was slightly lower than that of the Johnson-U distribution. This further confirms the effectiveness of selecting the Johnson-U distribution as the daily hourly trading volume distribution for stocks.

Table 4. Proportion of stocks that meet distribution types

Date	Johnson-U distribution	Inverse gamma distribution
March 8th, 2024	84.30%	80.80%
March 11, 2024	86.80%	86.10%
March 12, 2024	87.50%	85.80%
March 13, 2024	87.90%	86.00%
March 14, 2024	84.70%	80.90%

5 Conclusions

This article analyzes the hourly trading volume data of A-shares on a certain day. Based on kernel density estimation in nonparametric statistics and the single sample KS test, it is believed that the Johnson-U distribution fits the best, and the inverse gamma distribution and normal inverse Gaussian distribution fit relatively well. Further research was conducted on the trading volume data of multiple dates and different stocks, and it was found that the Johnson-U distribution has the strongest applicability to the data. This result is of great significance for understanding and analyzing the trading behavior of the stock market.

References

[1] Mandelbrot, B.: New methods in statistical economics. Journal of political economy, Vol. 71, No. 5, pp. 421-440 (1963)

[2] Mandelbrot, B.: The variation of certain speculative prices. Springer New York (1997)

[3] Fama, E.F.: Mandelbrot and the stable Paretian hypothesis. The journal of business, Vol. 36, No. 4, pp. 420-429 (1963)

[4] Guiqin, Z., Zhenyu, Z.: Non-normal distribution model of stock returns. Contemporary Finance & Economics, No. 10, pp. 40-43 (2002)

[5] Jianhua, W., Yuling, W., Kaiming, K.: Stable distribution fitting and testing of China's stock returns. Journal of Wuhan University of Technology, Vol. 25, No. 10, pp. 99-102 (2003)

[6] Xinyu, W., Xuefeng, S.: Statistical distribution fitting of Chinese stock market returns. Systems Engineering Theory and Practice, Vol. 26, No. 12, pp. 40-46 (2006)

[7] Wenjia, Q.: The Limit Distributions of Return in Stock Market. Chinese Journal of Applied Probability and Statistics, No. 2, pp. 130-140 (2005)
[8] Xijian, H., Dong, H.: Asymptotic Distributions of Return for Several Stocks with Trading Volume. Applied Probability and Statistics, Vol. 24, No. 1 (2008)
[9] Uchaikin, V.V., Zolotarev, V.M.: Chance and stability: stable distributions and their applications. Walter de Gruyter (2011)
[10] Xianling, R., Shiying, Z.: Investment portfolio risk analysis based on kernel density estimation and multivariate Archimedean Copula. Management Science, Vol. 20, No. 5, pp. 92-96 (2007)
[11] Hongzhong, L., Wenzhong, H.: Kernel density estimation and Monte Carlo simulation test of stock return distribution - A comparative study based on data before and after the introduction of the limit system. World Economic Literature, No. 2, pp. 46-55 (2010)
[12] Yong, W., Guangming, W.: Research on the distribution of stock returns based on non-parametric methods. China Foreign Investment, No. 4, pp. 116-116 (2010)
[13] Yunyue, G.: Application of Gamma Kernel in Positive Sequence Data and Probability Density Estimation of Stock Return Volatility. Business Economy, No. 19, pp. 79-81 (2010)
[14] Shi, X.: Empirical Study on VaR in Agricultural Sector Based on Non-parametric Kernel Density Estimation. Journal of Hubei University of Technology: Humanities and Social Sciences Edition, Vol. 30, No. 1, pp. 70-72 (2013)
[15] Yukun, N., Xiaohua, H.: Research on the Distribution of China's Stock Market Returns Based on Non-parametric Kernel Density Estimation Method. Journal of Hainan Normal University: Natural Science Edition, Vol. 26, No. 4, pp. 363-367 (2013)
[16] Wen, H.: Analysis of SSE A-share Returns Based on Kernel Density Estimation. Journal of Jiamusi University: Natural Science Edition, Vol. 33, No. 5, pp. 779-783 (2015)
[17] Yan, Y., Xiaowu, Z., Aimin, X., et al.: Financial Risk of the World's Major Stock Indexes Before and After the Financial Crisis: An Empirical Study Based on t Distribution. Systems Engineering Theory and Practice, Vol. 31, No. 5, pp. 841-847 (2011)
[18] Yizhou, L., Hongyu, Z., Yeli, W.: Comparative Analysis of Volatility Characteristics of Stock Returns in China and the United States - Based on Non-parametric Test and Its Distribution Kernel Density Estimation. China Foreign Investment, No. 20, pp. 1-4 (2013)
[19] Yan, W.: Research on the Optimal Mixed Copula Function in the Stock Market. Journal of Taiyuan Normal University: Natural Science Edition, Vol. 14, No. 1, pp. 38-43 (2015)
[20] Lingjuan, C., Yuan, C.: Empirical Study on China's Stock Market Weekend Effect - Based on the Analysis of Changes in Shanghai Composite Index and Shenzhen Component Index. Journal of Yichun University, Vol. 38, No. 8, pp. 31-36 (2016)
[21] Chao, H., Taihua, Y.: Research on Industry Portfolio Risk in China's Stock Market - Based on High-dimensional Dynamic C-Vine Copula Model. Journal of Chongqing University (Social Science Edition), Vol. 23, No. 2, pp. 40-50 (2017)
[22] Zehao, Y., Wen, L., Qun, W.: Empirical Study on the Similarity of Volatility Trends in China's Stock Market. Modern Business, No. 11, pp. 96-97 (2017)
[23] Jia, D., Jinping, Z.: Empirical Analysis of Stock Risk Based on Quantile Regression VaR Model. Statistics and Application, Vol. 7, pp. 407 (2018)
[24] Panjaitan, HP., Chandra, T.: The Influence of the Work Creation Law Draft on Abnormal Return and Trading Volume Activity in LQ45 Share[J]. Journal of Applied Business and Technology, Vol. 3, No. 1, pp. 17-25 (2022)
[25] Stevany, S., Wati, Y., Chandra, T., et al.: ANALYSIS OF THE INFLUENCE EVENTS ON THE INCREASE AND DECREASE OF WORLD OIL PRICES ON ABNORMAL RETURN AND

TRADING VOLUME ACTIVITY IN MINING SECTOR COMPANIES THAT REGISTERED IN INDONESIA STOCK EXCHANGE[C]//International Conference on Business Management and Accounting. Vol. 1, No. 1, pp. 181-192 (2022)
[26] Demir, S.: Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients[J]. International Journal of Assessment Tools in Education, Vol. 9, No. 2, pp. 397-409 (2022)
[27] Dong, W., Sun, H., Tan, J., et al.: Regional wind power probabilistic forecasting based on an improved kernel density estimation, regular vine copulas, and ensemble learning[J]. Energy, No. 238, 122045 (2022)
[28] Gao, F., Cheng, H.: Application of Kolmogorov-Smirnov Test in the Distribution of Saturn's Regular Satellites[J]. BULGARIAN ASTRONOMICAL JOURNAL Учредители: Institute of Astronomy and Rozhen NAO, pp. 37 (2022)
[29] Xu, X., Wang, Q., Li, C.: The impact of dependency burden on urban household health expenditure and its regional heterogeneity in China: Based on quantile regression method[J]. Frontiers in public health, No. 10, 876088 (2022)
[30] Nyakurukwa, K.: Revisiting the dynamic stock return–volume relationship in South Africa: a non-parametric causality in quantiles approach[J]. Macroeconomics and Finance in Emerging Market Economies, Vol. 17, No. 1, pp. 136-152 (2024)