

# ***Log-normal Distribution in Market Capitalization of Listed Companies in China***

**Diming Hong<sup>1\*</sup>, Kamang Fong<sup>2</sup>**

<sup>1</sup>*School of Insurance and Economics, University of International Business and Economics, Beijing, China*

<sup>2</sup>*College of Business and Public Management, Wenzhou-Kean University, Wenzhou, China*

*\*Corresponding Author. Email: hongdiming20001215@163.com*

**Abstract.** As the second largest economy in the world, the development of China's stock market has been of great interest to investors worldwide. By collecting the stock market capitalisation of the listed companies included in China's SSE Composite Index from 2003 to 2023, this paper finds that the market capitalisation of these companies shows a log-normal distribution. This work also conducts robustness tests using MSE, R2, PDF images, CDF images, and Q-Q plots, and the preliminary findings remain robust. Further research found that investor sentiment in the market affects the parameters of the log-normal distribution. By constructing market-level investor sentiment indicators using principal component analysis, the paper finds that when investor sentiment in the market is higher, the average market capitalisation of a company will be more prominent, and the gap in market capitalisation between different companies will be more significant. This paper provides a reference value for understanding the current development of China's stock market and improving the capital market regulation.

**Keywords:** Log-normal Distribution, Investor Sentiment, Capital Market

## **1. Introduction**

Distribution, the way objects and statistics spread or exist over a specific pattern, has been a popular topic of research across various fields. People often see statistics showing that a large number of variable values gather around the average while a small number shows a divergence trend. Commonly, human height, blood pressure, shoe size, IQ, weight, test scores, etc., obey certain distributions. It is important for us to understand the changing character of things.

The earliest history of log-normal distribution can be traced back to 1879 in "The Law of the Geometric Mean", published by Donald Mcalister. Based on Francis Galton's work on standard errors and Gaussian distribution, he found the logarithmic cumulative distribution function of continuous series to answer Galton's question of finding a better medium value than Arithmetic Mean in discordant measurements.

In the first chapter, the introduction, our paper briefly describes the need to study distribution. The second chapter is a literature review that briefly reviews the current state of research on the distribution and characteristics of global and Chinese stock markets. The third chapter is

Methodology, which is the core part of this paper. It includes introducing log-normal distribution, data source, data visualisation, and the data fitting process. The fourth chapter is the research conclusion.

This paper studies the market capitalisation of listed companies on the SSE in China from 2003 to 2023. It finds that the distribution of market capitalisation of listed companies is characterised by a log-normal distribution, and the findings are robust. High investor sentiment increases the mean firm market capitalisation and widens the standard deviation of the capitalisation.

## 2. Literature review

The log-normal distribution is extensively used in various fields. In the field of ecology, Crow and Shimizu found that both the abundance of marine organisms and terrestrial plant communities obeyed a log-normal distribution to some extent [1]. Stedinger also found that the two-parameter log-normal distribution has the ability to explain well the distribution of hydrosphere data. In sociology, the log-normal distribution also has many good applications [2]. In daily life, some scholars have found that the operation time in the hospital can be well explained and fitted with the two-parameter log-normal distribution [3]. Fonseca and Tayman also found that the adjusted log-normal distribution model can explain the income level distribution in a county below California [4].

In the past, many scholars have concentrated on the research of share prices in capital markets. Research on company share prices focuses on factors influencing stock pricing, stock price forecasting, portfolio optimisation and so on [5-9].

Apart from that, there are also many scholars focusing on the distribution of listed companies' share prices and the market value of listed companies. However, quantitatively describing the distribution of stock prices or company market capitalisation is a challenging task. Teichmoeller found that the distribution of stock prices is a complex mixture of normal distributions [10]. They may exhibit more complex distributions. Using maximum likelihood estimation, Linden found that stock returns on the Helsinki stock exchange market in the late 1980s were consistent with Laplace mixture distribution [11]. Besides, Kaizoji & Kaizoji quantify stock returns and volatility of the Nikkei 225 index and discover that they can be affected by asymmetric Laplace distribution and double exponential distribution, respectively [12]. Masoliver et al. used the S&P 500 cash index to fit a model to high-frequency financial data and found that the index is not Gaussian distribution. However, their centre is well revised by Lévy distributions [13].

Since China's capital market started relatively late, research on Chinese stock prices and stock returns is still in dire need of supplementation compared to other developed countries. Hussain and Li compared the fitting effects of three distributions, Generalized Extreme Value (GEV), Generalised Logarithmic (GL) and Generalised Pareto distributions (GP), by using the Power Weighted Method (PWM) approach for parameter estimation and found that the GL distribution is more suitable for the minimum series of Chinese stock market returns. It is found that the GL distribution is more suitable for the minimum series of Chinese stock market returns. The GEV distribution is more suitable for the maximum Chinese stock returns [14]. Xu et al. found that the alpha stability law can explain the asymmetric leptokurtic features distributed in the returns of the Shanghai Composite Index and the Shenzhen Composite Index [15].

Based on the above analyses, there has yet to be a consensus on the distribution of market capitalisation in China-listed companies. Therefore, this paper attempts to fit the market capitalisation distribution of listed companies in the Shanghai Stock Exchange by model fitting.

### 3. Methodology

#### 3.1. Definition

Before the discussion, this briefly introduces the "Log-normal Distribution" (LD).

The log-normal distribution is a continuous probability distribution that describes a random variable whose logarithmic values follow a normal distribution. Specifically, if  $Y = \log(X)$  obeys a normal distribution, then  $X$  obeys a log-normal distribution:

$$X \sim \text{Log-normal}(\mu, \sigma^2) \quad (1)$$

Whereas the  $\mu$  is the mean of the normal distribution and the  $\sigma^2$  is the variance of the normal distribution.

Let  $x$  represent the variables which are interested. The probability density function (PDF) of the log-normal distribution can be written as:

$$f(x, \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right] \quad (2)$$

Where  $x$  is the observed variable and  $x > 0$ .  $\mu$  is the mean of logarithmic values of  $x$  and  $\sigma^2$  is the variance of logarithmic of  $x$ .

Calculating the cumulative distribution function of the log-normal using the cumulative distribution function (CDF) of the normal distribution shows that:

$$F(x, \mu, \sigma) = \int_0^x \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(t)-\mu)^2}{2\sigma^2}\right] dt \quad (3)$$

Commonly used parameter estimation methods include Moment estimation (MM), least squares estimation (LSE), and maximum likelihood function estimation (MLE). MLE is a statistical inference method used to estimate model parameters in a way that maximises the likelihood of the observed sample data. Since MLE has good parameter estimation for large samples and applies to continuous distributions, this paper uses MLE to estimate the parameters when performing model fitting.

First, assume that for random variables:  $x_1, x_2, \dots, x_n$ , they follow a log-normal distribution, then the function is known to be:

$$L(\mu, \sigma) = \prod_{i=1}^n \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right] \quad (4)$$

Taking logarithms on both sides of the equation:

$$\log L(\mu, \sigma) = \log L(\mu, \sigma) \quad (5)$$

$$\log L(\mu, \sigma) = \log \left[ \prod_{i=1}^n \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right) \right] \quad (6)$$

$$\log L(\mu, \sigma) = \sum_{i=1}^n \left[ \log \left[ \frac{1}{x_i\sigma\sqrt{2\pi}} \right] - \frac{(\ln x_i - \mu)^2}{2\sigma^2} \right] \quad (7)$$

$$\left(\mu, \sigma\right) = -n \log(\sigma \sqrt{2\pi}) - \sum_{i=1}^n \frac{1}{2\sigma^2} \sum_{i=1}^n (\ln x_i - \mu)^2 \log \quad (8)$$

After that this work take the partial derivatives of the parameters  $\mu$ ,  $\sigma$  of the log-likelihood function  $(\mu, \sigma)$  :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \ln x_i \quad (9)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \hat{\mu})^2 \quad (10)$$

### 3.2. Data source

The research is mainly conducted using Scipy, a Python package that helps to fit the empirical data using MLE and calculates all the required statistical quantities.

This paper's data comes from the CSMAR dataset. It collects data regarding each company's market value on the last opening day of the month. To begin with, this work filters out all the months with fewer than 500 data points, that is, the periods when fewer than 500 companies are presented in the market. This act aims to avoid an extremely poor fit due to the small sample size, as it could lead to highly inaccurate parameter estimations and false conclusions.

This monthly dataset contains the market capitalisation of listed companies in the Shanghai Securities Composite Index, collected and used from 2003 to 2023.

### 3.3. Data visualisation

The paper visualises company market capitalisation data for the market capitalisation of companies included in the SSE Composite Index for each month from 2003 to 2023. Besides, the work selects samples at five-year intervals and plots histograms of the frequency distribution of formula market values. Meanwhile, since Chinese listed companies have already published their annual financial statements for the previous year in June each year, and the market capitalisation of the companies is relatively stable, this paper selects the distribution of the market capitalisation in the month of June at five-year intervals for data visualisation. The results are shown in Figure 1. We normalised the company's market capitalisation, where the horizontal axis represents the normalised company market capitalisation and the vertical axis represents the market capitalisation frequency.

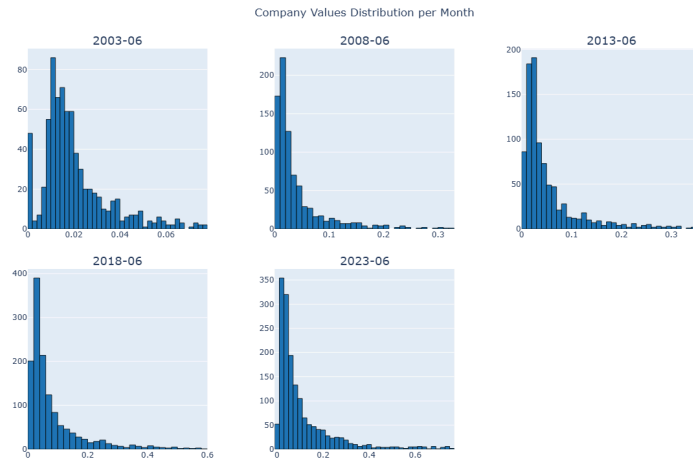


Figure 1. Histogram of frequency distribution of market capitalization of companies

Based on the histogram of the data frequency distribution, it is conjectured that the market capitalisation of listed companies may obey a log-normal distribution, a Weibull distribution, a Skewed Normal distribution, or a Powerlaw distribution. Table 1 shows the expressions for the probability density functions of these four distributions and the meaning of each parameter.

Table 1. Definitions of the statistical distributions

| Name                       | $f(x)$ for Continuous Data                                                                                                                       | Parameter Definitions                                                                                                                                                        |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Log-normal Distribution    | $f\left(x, \mu, \sigma^2\right) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right]$                            | $\mu$ : mean of the normal distribution<br>$\sigma^2$ : variance of the normal distribution                                                                                  |
| Weibull Distribution       | $f\left(x, \lambda, k\right) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^k\right]$        | $\lambda > 0$ : scale parameter<br>$k > 0$ : shape parameter                                                                                                                 |
| Skewed Normal Distribution | $f\left(x, \xi, \omega, \alpha\right) = \frac{2}{\omega} \phi\left(\frac{x - \xi}{\omega}\right) \Phi\left(\alpha \frac{x - \xi}{\omega}\right)$ | $\xi$ : location parameter<br>$\omega > 0$ : scale parameter<br>$\phi(\cdot)$ : PDF for standard normal distribution<br>$\Phi(\cdot)$ : CDF for standard normal distribution |
| Powerlaw Distribution      | $f(x) = \frac{\alpha - 1}{x_{\min}} x^{-\alpha}$                                                                                                 | $\alpha$ : the power-law exponent<br>$x_{\min}$ : the minimum value of $x$                                                                                                   |

### 3.4. Goodness of fit

It is crucial to perform a goodness of fit to evaluate and represent the effect the distribution fits to the data. Here, the paper considers the Mean-Squared-Error (MSE) and the R Square (  $R^2$  ) to determine whether the data comes from the particular distribution and whether the fit is appropriate. For  $R^2$ , if the value of  $R^2$  is larger, it proves that the fitted model explains the real data more strongly; for MSE, if the value of MSE is smaller, it proves that the difference between the fitted data and the real data is smaller, and the better the model is fitted. For evaluation, this paper chooses a significance level of 1%, which means if the resulting  $p$  -  $value$  is smaller than 0.01, it would be rejected the null hypothesis that the data comes from the particular distribution.

$$MSE = \frac{1}{N} \sum_{i=1}^n \left[ \hat{F}(X_i) - F(X_i) \right]^2 \quad (11)$$

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (13)$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (14)$$

After fitting log-normal, weibull, skewed normal, and power-law distribution, the paper gets the fitting results shown in Figure 2 and Figure 3. Figure 2 shows the fit under MSE and Figure 3 shows the fit under  $R^2$ . It is evident that log-normal has the least MSE in all fits, and the fits for power-law show the worst performance. This shows that the log-normal fit is the best of the four possible distributions. From Figure 3, our paper gets the same interpretation by looking at  $R^2$ , which means the market value data fits the log-normal distribution best compared to the three other distributions.

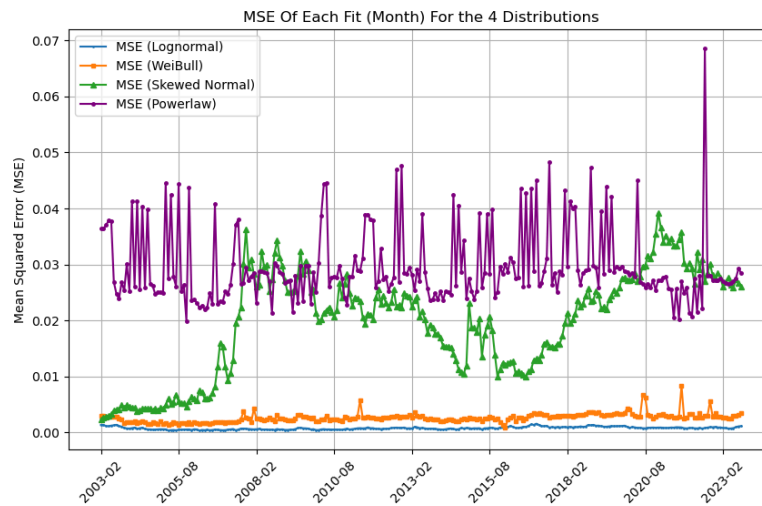


Figure 2. MSE under four methods of fitting

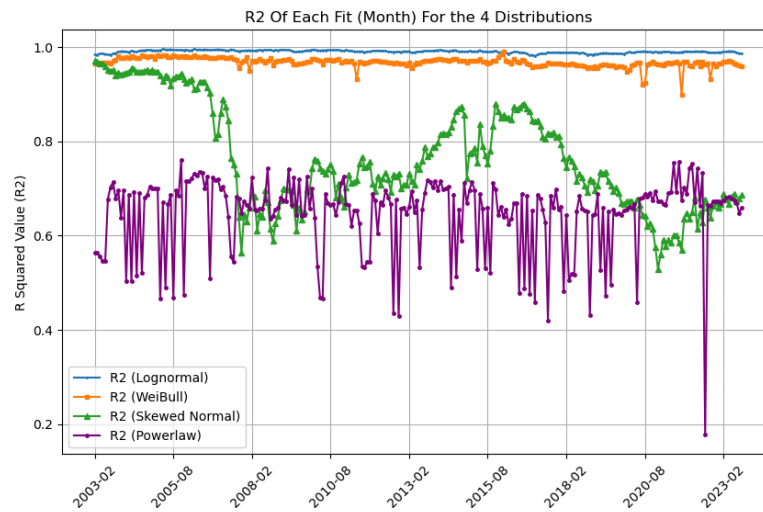


Figure 3.  $R^2$  under four methods of fitting

Based on comparing the four possible distributions mentioned above, this paper then verifies them by the CDF and PDF images of the LD, as shown in Figures 4 and 5. In Figure 4, the KDE is plotted and the fitted PDF of nine months of data, and in Figure 5, the corresponding empirical CDF is plotted and the fitted CDF of the same data. Figures 4 and 5 show that the LD explains the distribution of the company's market capitalisation data points well.

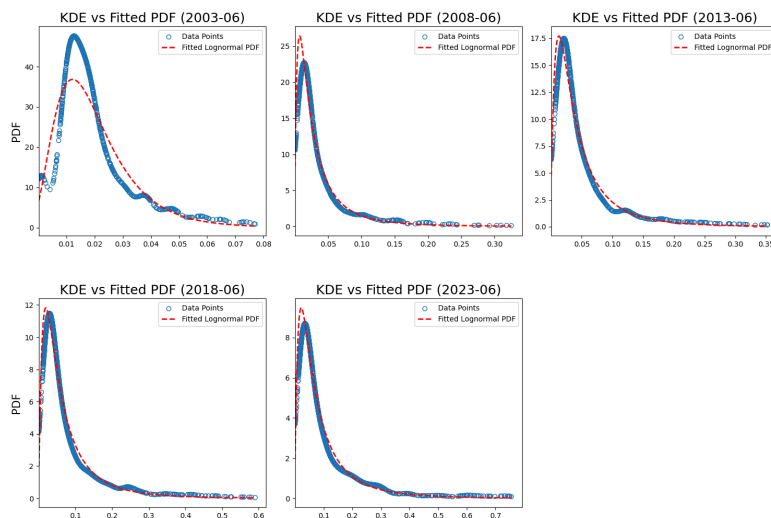


Figure 4. PDF under fitted log-normal

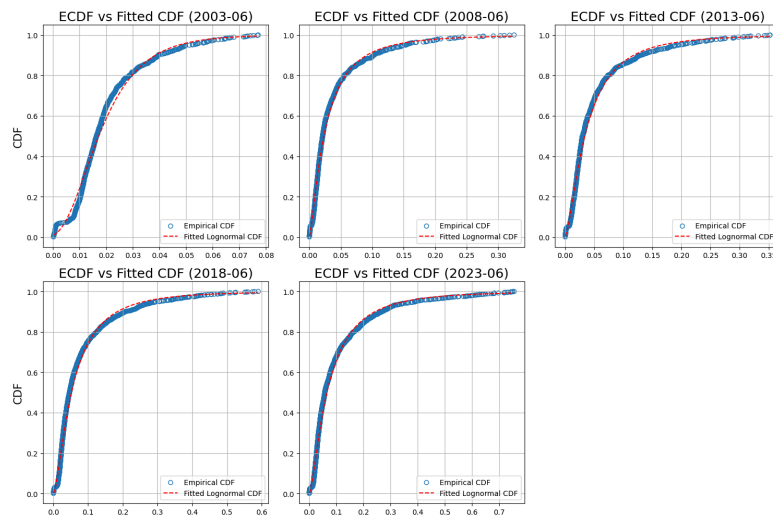


Figure 5. CDF under fitted log-normal

These graphs generally indicate decent fits of log-normal distribution to the data as all the lines follow similar trends and overlap each other. Nevertheless, this does not imply that the data obey the distribution in the whole range, especially for the tails where extreme values lie. Hence, in Figure 6, the quantile-quantile plot (Q-Q Plot) of those same data is shown to examine the deviations at these tails since the Q-Q plot is particularly sensitive to differences in the tails of distributions. Because the points deviate from the line at higher quantiles, the data exhibit a heavier tail compared with the fitted log-normal distribution, which reflects that the distribution does not do an excellent job of capturing the extremes of the data.

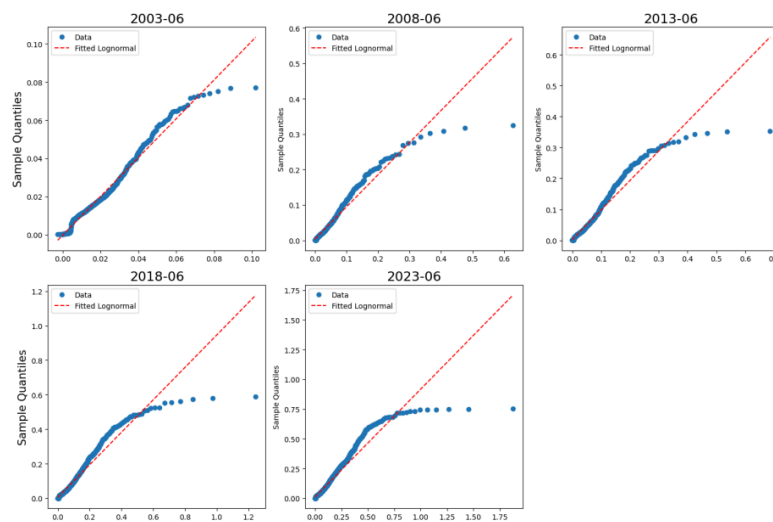


Figure 6. Q-Q plot under fitted log-normal

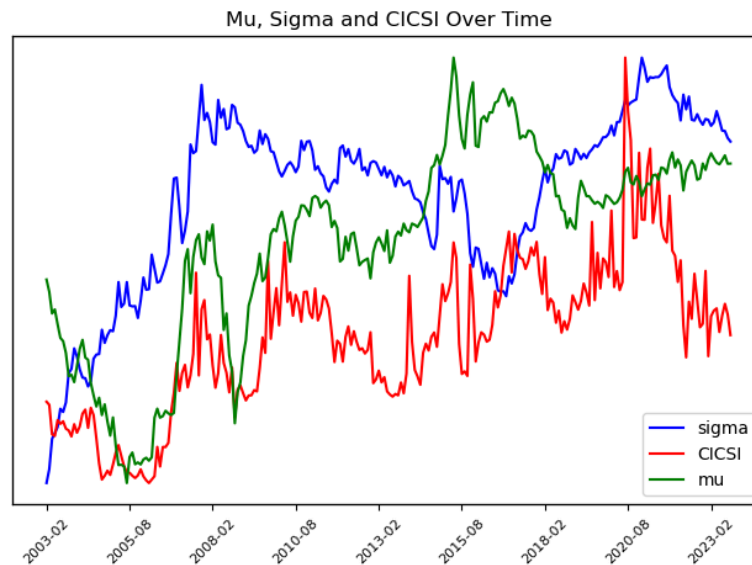


Figure 7. Mu, Sigma and CICI over time

Based on the above analysis, it is concluded that the market capitalisation of listed companies measured by China's SSE Composite Index has a log-normal distribution. According to the description in the Definition section, when  $\mu$  is larger, the market value will be larger, and the overall distribution will be shifted to the right; when  $\sigma$  is larger, the overall distribution will be more dispersed, and the difference in market capitalisation of the listed companies will be larger.

### 3.5. Further analysis

Investor sentiment refers to the psychological state and emotional tendencies investors exhibit during decision-making. In 1990, De Long et al. found that noise traders can profit from capital markets and persist for a long time due to the limited rationality of investors and sentiment [16]. At the same time, investor sentiment also distorts the stock pricing of listed companies, making them deviate from their true value [17]. In addition to this, past research has not only found that investor sentiment can affect the return and VaR of a company's stock but also found that investor sentiment has some synergies with stock returns [18-20].

Based on the above analyses, this paper seeks to investigate whether investor sentiment impacts the  $\mu$  and  $\sigma$  of the log-normal distribution of a firm's market capitalisation. Referring to Baker and Wurgler and Cao et al., this paper measures market-level investor sentiment using the China Investor Composite Sentiment Index (CICI) [21,22].

Our paper plots the investor sentiment index with the parameters  $\mu$  and  $\sigma$  of the log-normal distribution in Figure 7. In Figure 7, the paper plots trends in CICI,  $\mu$  and  $\sigma$  over a 20-year period from 2003 to 2023. It can be seen that the trend of the CICI and the parameters  $\mu$  and  $\sigma$  are the same for a more significant number of time periods, from which this work initially infers that investor sentiment is positively correlated with the mean and standard deviation.

Next, this paper makes a causal inference of the investor sentiment index with  $\mu$  and  $\sigma$  by OLS regression. The regression coefficient between investor sentiment and  $\mu$  is 0.7836, with a  $p$ -value less than 0.000; the regression between investor sentiment and  $\sigma$  is 0.6016, with a  $p$ -value less than 0.000.

It can be seen that when investor sentiment is high, the overall market capitalisation of the market increases, but the difference in market capitalisation between companies also increases; when investor sentiment is low, the overall market capitalisation of the market decreases, and the difference in market capitalisation between companies decreases.

#### 4. Conclusion

Capital markets have always interested investors and economists worldwide, who consider them macroeconomic "barometers". By collecting data on the market capitalisation of companies included in China's SSE Composite Index from 2003 to 2023, this paper discovers that the market capitalisation of listed companies shows a log-normal distribution in general. By comparing the MSE and  $R^2$  of several possible distributions of firms' market capitalisation and drawing the PDF, CDF and Q-Q plot of the market capitalisation distribution, we find that the findings remain robust.

This work also explores the link between investor sentiment and the parameters  $\mu$  and  $\sigma$  of the log-normal distribution using empirical analyses. It is found that when investor sentiment is higher, the values of  $\mu$  and  $\sigma$  are larger, the mean of the overall distribution increases and the distribution becomes more dispersed.

However, this study tested only four types of distributions. It did not investigate other potential distributions, such as the Poisson and Gamma distributions.

#### References

- [1] Crow, E. L., Shimizu, K. (1987). Lognormal Distributions. Marcel Dekker Publishing, New York.
- [2] Stedinger, J. R. (1980) Fitting log normal distributions to hydrologic data. *Water Resources Research*, 16(3): 481-490.
- [3] May, J. H., Strum, D. P., Vargas, L. G. (2000) Fitting the lognormal distribution to surgical procedure times. *Decision Sciences*, 31(1): 129-148.
- [4] Fonseca, L., Tayman, J. (1989) Postcensal estimates of household income distributions. *Demography*, 26(1): 149-159.
- [5] Allen, F., Gale, D. (1992) Stock-price manipulation. *The Review of Financial Studies*, 5(3): 503-529.
- [6] Jennings, R., Starks, L. (1985) Information content and the speed of stock price adjustment. *Journal of Accounting Research*: 336-350.
- [7] Lu, W., Li, J., Wang, J., Qin, L. (2021) A CNN-BiLSTM-AM method for stock price prediction. *Neural Computing and Applications*, 33(10): 4741-4753.
- [8] McDonald, J. G., Fisher, A. K. (1972) New-issue stock price behaviour. *The Journal of Finance*, 27(1): 97-102.
- [9] Zhang, Z., Zohren, S., & Roberts, S. (2020) Deep learning for portfolio optimisation. *Arxiv Preprint Arxiv*, 2005: 13665.
- [10] Teichmoeller, J. (1971) A note on the distribution of stock price changes. *Journal of the American Statistical Association*, 66(334): 282-284.
- [11] Linden, M. (2001) A model for stock return distribution. *International Journal of Finance & Economics*, 6(2): 159-169.
- [12] Kaizoji, T., Kaizoji, M. (2003) Empirical laws of a stock price index and a stochastic model. *Advances in Complex Systems*, 6(3): 303-312.
- [13] Masoliver, J., Montero, M., Porrà, J. M. (2000) A dynamical model describing stock market price distributions. *Physica A: Statistical Mechanics and its Applications*, 283(3-4): 559-567.
- [14] Hussain, S. I., Li, S. (2015) Modeling the distribution of extreme returns in the Chinese stock market. *Journal of international Financial Markets, Institutions and Money*, 34: 263-276.
- [15] Xu, W., Wu, C., Dong, Y., & Xiao, W. (2011) Modeling Chinese stock returns with stable distribution. *Mathematical and Computer Modelling*, 54(1-2): 610-617.
- [16] De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J. (1990) Noise trader risk in financial markets. *Journal of Political Economy*, 98(4): 703-738.

- [17] Antoniou, C., Doukas, J. A., & Subrahmanyam, A. (2016) Investor sentiment, beta, and the cost of equity capital. *Management Science*, 62(2): 347-367.
- [18] Li, X., Xing, Y. (2023) When stock return synchronicity meets investor sentiment. *Finance Research Letters*, 53: 103655.
- [19] Liu, Q., Lee, W. S., Huang, M., Wu, Q. (2023) Synergy between stock prices and investor sentiment in social media. *Borsa Istanbul Review*, 23(1): 76-92.
- [20] Shen, Y., Liu, C., Sun, X., Guo, K. (2023). Investor sentiment and the Chinese new energy stock market: A risk–return perspective. *International Review of Economics & Finance*, 84: 395-408.
- [21] Baker, M., Wurgler, J. (2006) Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4): 1645-1680.
- [22] Cao, S., Li, Z., Koedijk, K. G., Gao, X. (2022) The emotional cost-of-carry: Chinese investor sentiment and equity index futures basis. *China Finance Review International*, 12(3): 451-476.