Remarks on Normal Distribution and Central Limit Theorem

Zixi Cheng*

Department of Applied Math, University of California, Santa Barbara, California, USA

*Corresponding author: zixicheng@ucsb.edu

Abstract. The normal distribution holds significant importance in the fields of mathematics, science, and engineering. Statisticians often use the normal distribution to get a good idea of how likely different real-life events are to happen. For example, they make it easier to show how data is spread out during the measurement process. This research looks at the basic ideas behind the normal distribution. The analysis examines the shape of the curve, the average value, the measure of variability, and its applicability across several disciplines. It also explains the central limit theorem, which says that when adding up a lot of random variables, they often form a distribution that looks a lot like a normal distribution. The study also talks about other statistical ideas, like confidence intervals and Z-tests, and how they relate to the normal distribution. Formulas and methods for changing statistical measures are broken down in the book, which also briefly talks about related distributions like the χ-squared and t-distributions.

Keywords: Normal distribution; Central limit theorem; Confidence interval; Distribution function.

1. Introduction

The normal distribution, sometimes known as the Gaussian distribution, is a fundamental statistical model. Gauss made some important mathematics advances in 1795 that led to the Gaussian or standard normal distribution, which is linked to the bell-shaped curve [1]. The prime number distribution theorem and the least squares method are two of these innovations. In probability estimates, the Gaussian distribution is very important. Well-known groups like the International Bureau of Metrology, the International Organization for Standardization, the International Organization for Legal Metrology, and others work together to show that the principles of this system are the basis for important global standards [2]. These partnerships are in many areas, such as clinical chemistry, theoretical and applied chemistry, and theoretical and applied physics. Metrology, the scientific discipline concerned with accurate measurement in all fields with different degrees of confidence, depends on the mathematical foundations of the normal distribution to establish and calculate uncertainty [3].

The normal distribution is a symmetrical Gaussian curve with lower values at the extremes and greater values in the middle. The unique form of the normal distribution curve has resulted in its frequent designation as a bell curve. The notation $N(\mu, \sigma^2)$ represents a random variable x that conforms to a normal distribution, with an expected value of μ and a variance of σ^2. The expected value μ has an impact on the location of the normal distribution, whereas the amplitude is defined by the standard deviation [4]. The standard normal distribution is defined as a normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1. Nevertheless, not all general normal populations display absolute symmetry with respect to the y-axis. Thus, for a given normal population, people may determine the likelihood of its value being smaller than x or falling within a particular interval [5]. In order to streamline computations and practical use, it is usual to transform a generic normal variable into a standard normal distribution. This transformation allows people to directly obtain values from a table of the standard normal distribution. Therefore, this process of conversion is referred to as a standard transformation.

This paper is organized as following. Section 2 will give a brief description of normal distribution and central limit theorem. Section 3 will present the application of normal distribution in statistics and some relevant distributions. The last section is devoted for the conclusion.

2. Normal Distribution and Central Limit Theorem

2.1. Normal Distribution

The normal distribution is very important and it has a lot to do with statistics and chance. The normal distribution, commonly known as the Gaussian distribution, is a probability distribution for a random variable that can have real values. It is a continuous distribution. In a broad sense, the probability distribution function looks like this [6]

$$f(x) = \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

By virtue of the definition of expectation, it is found that

$$E(x) = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx + \mu \int_{-\infty}^{\infty} f(x)\, dx = \mu. \tag{2}$$

On the other hand, it is observed that

$$\mathrm{Var}(x) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{dx} = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2\, e^{-\frac{1}{2}t^2} dt = \sigma^2. \tag{3}$$

The above equality follows by $\lim_{t\to-\infty}\left(-te^{-\frac{1}{2}t^2}\right) = -\lim_{t\to-\infty}\left(-te^{-\frac{1}{2}t^2}\right) = 0$ and also the other counterpart $\lim_{t\to+\infty}\left(-te^{-\frac{1}{2}t^2}\right) == \lim_{t\to+\infty}\left(\frac{-t}{e^{\frac{t^2}{2}}}\right) = \lim_{t\to+\infty}\left(\frac{-1}{te^{\frac{1}{2}t^2}}\right) = 0$. The standard normal distribution, which is frequently emphasized, has a mean of 0 and a variance of 1. The standard normal distribution is denoted as $\mathcal{N}(0,1)$ and the cumulant distribution function of $\mathcal{N}(0,1)$ is $\Phi(z)$.

2.2. Central Limit Theorem

The central limit theorem (CLT) is a significant outcome in the realm of probability that pertains to the normal distribution. According to this theorem, when there is a specific number (n) of independent random variables with equal mean and variance, their average will increasingly resemble a normal distribution. The content of the message is as follows.

Theorem 1. Suppose $\{\mathrm{X_1}, \mathrm{X_2}, \cdots, \mathrm{X_n}\}$ is a sequence of i.i.d (independent and identically distributed) random variables with $\mathrm{E(X_i)} = \mu$ and $\mathrm{Var(X_i)} = \sigma^2 < \infty$. As n tends to ∞, the random variable $\sqrt{\mathrm{n}}(\bar{\mathrm{X}}_\mathrm{n} - \mu)$ converges in distribution to a normal distribution $\mathcal{N}(0, \sigma^2)$ [7].

When the expected value and variance of each random variable are distinct, Lyapunov has the following CLT.

Theorem 2. Suppose $\{\mathrm{X_1}, \mathrm{X_2}, \cdots, \mathrm{X_n}\}$ is a sequence of i.i.d random variables each with finite expected value μ_i and variance σ_i^2. Define $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$. If for some $\delta > 0$ and the Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left[X_i - \mu_i^{2+\delta}\right] = 0 \tag{4}$$

is satisfied. Then the sum of $\frac{\mathrm{X_i}-\mu_\mathrm{i}}{\mathrm{s_n}}$ converge in distribution to a standard normal variable as n goes to ∞ [8].

Here is an example. The paper considers a scenario where it has a coin and the likelihood of it landing on heads is $1/4$ and the likelihood of obtaining a tails outcome is $3/4$. Roll the coin 1000 times and denote that Y is the number of the coin landing heads. To estimate the probability that $P(200 \le X \le 210)$, one way is the directly calculation which is found that $P(200 \le X \le 210) =$

$\sum_{i=200}^{210} C_{1000}^{i} \left(\frac{1}{4}\right)^{i} \left(\frac{3}{4}\right)^{1000-i}$, while another way to estimate it is by Theorem 2.1. By directly calculation, it is found that $E(x) = 1000 * \frac{1}{4} = 250$ and $Var(X) = 1000 * \frac{1}{4} * \frac{3}{4} = 187.5$. Since 1000 is a large number, by Theorem 2.1, it is inferred that

$$\begin{aligned} P(200 \le X \le 210) &\approx P\left(\frac{199.5-250}{\sqrt{187.5}} < \frac{X-250}{\sqrt{187.5}} < \frac{211.5-250}{\sqrt{187.5}}\right) \\ &= P\left(\frac{199.5-250}{\sqrt{187.5}} < Z < \frac{211.5-250}{\sqrt{187.5}}\right) \quad Z \sim \mathcal{N}(0,1) \\ &= \Phi\left(\frac{211.5-250}{\sqrt{187.5}}\right) - \Phi\left(\frac{199.5-250}{\sqrt{187.5}}\right) \end{aligned} \tag{5}$$

It is found that Theorem 2 is much easier to calculate than Theorem 1.

3. Applications and Extensions of Normal Distribution

3.1. Normal Distribution in Statistics

The normal distribution is a crucial component in the field of statistics. One of the distributions is to calculate the Confidence interval. A confidence interval in mathematics is an interval that is predicted to usually include the estimated value. To be more precise, a confidence interval is a stochastic interval that encompasses the estimated parameter with a confidence level α, typically set at 95% or 99%. If there are n object, it is assumed that each object conforms to the same normal distribution $\mathcal{N}(\mu, \sigma^2)$. Given the variance σ^2 and the sample mean $\bar{X}$. Given Theorem 1, which states that when the number of samples is large, then $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$. Use this information to estimate the confidence interval of μ at the level α as follows $P\left(\left|(X-\mu)/\sqrt{\sigma^2/n}\right| \le z_{\alpha/2}\right) = \alpha$. The confidence interval of μ at the level α is $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right]$. Calculating the confidence interval of proportions p at the level α, it follows that $\left[p' - z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}, p' + z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}\right]$, where $p'$and $q'$the point estimate of p and q taken from the sample [9].

In mathematics, a test is referred to as a Z-test when the distribution of the test statistic may be approximately approximated by a normal distribution under the assumption that the null hypothesis is true. The one-sample location test compares the average value of a dataset to a fixed value, assuming that the variability of the dataset is already known. The Z-test is typically employed for this purpose. For example, if the observed data $X_1, X_2, \cdots, X_n$ are independent, have a common mean μ, and have common variance σ^2. Given the null hypothesis is that the mean value of $X = \frac{X_1+X_2+\cdots X_n}{n}$ is a given number μ_0, since the variance σ^2 is known and the statistic $Z = \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}}$ follows $\mathcal{N}(0,1)$, then the author can calculate that at each level β whether the author reject or accept the null assumption.

3.2. Other Distribution Related to Normal Distribution

The first is the χ-square distribution. If $Z_1, Z_2, \cdots, Z_n$ are i.i.d standard normal random variables then the sum of their squares $Q = \sum_{i=1}^{n} Z_i^2$ follows a χ-squared distribution with n being the degree of freedom, indicated as $\chi^2(n)$. The probability distribution function of $\chi^2(n)$ is following

$$f_n(x) = \begin{cases} \frac{x^{n/2}-1}{2^{n/2}\Gamma(n/2)} & x \ge 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

where Γ denotes the Gamma function. What the author cares is the expectation and variance of $\chi^2(n)$ which are following [10]:

$$E\left(\chi^2(n)\right) = n, Var\left(\chi^2(n)\right) = 2n. \tag{7}$$

The second is the t-distribution. In probability and statistics, a t-distribution with freedom ν is a distribution t which can be written as $T = \frac{Z}{\sqrt{U/\nu}}$ where $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(\nu)$. Denote the t-distribution with freedom ν as t_ν. Then the probability distribution function of t_ν is following:

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{8}$$

The expectation and variance of t_ν is

$$E(t_\nu) = \begin{cases} 0 & \nu > 1 \\ DNE & \nu = 1 \end{cases}, Var(t_\nu) = \begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ DNE & \nu = 1,2 \end{cases} \tag{9}$$

The next theorem is the base for people to use the t-distribution to do something. Let $X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d samples from a normal distribution with mean μ and variance σ^2 .The sample mean and unbiased sample variance are given by $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, then the t-statistic is given by

$$T = \frac{\bar{X}-\mu}{\sqrt{s^2/n}} \sim t_{n-1} \tag{10}$$

and is distributed according to a t-distribution with freedom n− 1. Hence if there are If there are n object. Assume that each object follows the same normal distribution $\mathcal{N}(\mu, \sigma^2)$. If the author just knows the sample mean $\bar{X}$ and the sample variance s^2, and when n is large then $\frac{\bar{X}-\mu}{\sqrt{s^2/n}} \sim t_{n-1}$, then the author can estimate the confidence interval of μ at the level α as following $P\left(\left|(X-\mu)/\sqrt{s^2/n}\right| < t_{n-1,\alpha/2}\right) = \alpha$. Hence, the confidence interval of μ at the level α is $\left[\bar{X} - \frac{s}{\sqrt{n}}t_{\frac{n-1,\alpha}{2}}, \bar{X} + \frac{s}{\sqrt{n}}t_{n-1,\frac{\alpha}{2}}\right]$. The t-test is another way that the t-distribution can be used. The t-test is a mathematical test that determines whether the difference between the answers of two groups is statistically significant.

4. Conclusion

The main ideas of normal distribution are studied, along with the importance of normal distribution in statistical analysis, with a focus on the central limit theorem. The textbook talks about how approximation likelihoods can be used in real life and how they relate to statistical ideas like confidence intervals and Z-tests. It also gives a full description of the central limit theorem, focusing on its importance in the field of probability and including real-life examples to help readers understand better. This essay also talked about how the normal distribution can be used in different statistical methods, such as hypothesis testing and regression analysis. For many statistics methods, the normal distribution is the basic idea behind them. When conducting hypothesis testing, analysis of variance, correlation analysis, or regression analysis, one needs to believe that the variables people are looking at have a normal distribution. In metrology, using normal distributions is very important for making statistical data more general and figuring out error that has real-world effects. Many statistical methods don't need to assume that the indicators being studied have a normal distribution, but when working with big samples, the statistics that go with them tend to be close to a normal

distribution. Because of this, these statistical methods for drawing conclusions depend on the idea that big samples follow a normal distribution.

References

[1] Yan Surong, Cui Hongxin et al. Basis of Probability and Statistics. National Defense Industry Press, 2011.

[2] Fu Hui-Min, Guo Jian-Chao, Fu Yue-Shuai, Li Zi-Ang. Reliability Assessment Methods with Type-II and Type-I Censored Data under Normal Distribution. Development and Innovation of Machinery and Electrical Products, 2023, 36(6): 1-4.

[3] Yao Jihai, Zhang Yuexin. The Misuse of Normal Distribution of Course Test Scores in Higher Education and Suggestions for Improvement. Journal of China Examinations. 2023, 7: 76-83.

[4] Iglesias P. M. C., Vidal-Puga J, Pino J. M. R. The role of self and peer assessment in higher education. Studies in Higher Education, 2022, 47(3): 683-692.

[5] Li Ling, Xu Zhangtao. The Instructional Design of Normal Distribution: Looking for the Growing Point of Students' Cognition from History. Journal of Mathematics Education. 2023, 32(2): 12-17.

[6] Stahl S. The evolution of the normal distribution. Mathematics Magazine, 2006, 79 (2): 96-113.

[7] Wei Tanrong, Zeng Zhenbing. A limit theorem related to unbiased estimates of normal distribution. College Mathematics, 2022, 38(4): 96-99.

[8] Wei Chijiang. An Inference on the Density Function of Normal Probability Distribution. Statistics and Information Forum. 2009, 24(5): 3-6.

[9] Fienberg S E, Hinkley D V, Fisher R A. An application lecture notes in statistics. New York: Springer - Verlag, 1980.

[10] Wang Ronghua, Gu Beiqing, Liu Jinmei, Xu Xiaoling. Approximate Confidence Interval for the Difference and Quotient of Variation Coefficients of Two Normal Distributions. Statistics and Decision, 2022, 589(1): 38-42.