Dean&Francis

ISSN 2959-6157

Research on Derivation and Application of Normal Distribution

Bowen Xiang

Pennon Education, Qingdao, China

*Corresponding author: xiangbowen@ldy.edu.rs

Abstract:

With the progress of the times, probability theory has been widely used. The normal distribution is an important concept in the Chapter of probability theory. The normal distribution is a probability distribution with a bell-shaped probability density function curve. The probability density function is a function that describes the probability distribution of continuous random variables in a certain range. Now the normal distribution is also widely used. For example, in the weather, the average temperature, average humidity and rainfall in July each year, and the water level in the hydrology are also approximately normal distribution. Based on the definition of normal distribution and the theoretical knowledge of normal distribution, this paper studies the derivation of normal distribution formula and the application of normal distribution in daily life. Finally, the application of normal distribution is diversity in daily life, it is expected that this paper will provide specific reference for the study of the academic content of normal distribution and its practical application in professional fields.

Keywords: Normal distribution; Gaussian distribution; Likelihood function; Asphalt scales.

1. Introduction

After the normal small sample theory was fully developed in the 20th century. The French mathematician Laplace soon learned of Gauss's work and immediately connected it to his discovery of the central limit theorem. He published an article in 1812 showing that the binomial distribution can be approximated by the normal distribution. Nowadays, Gaussian distribution is widely used in many fields such as life production, scientific and technological experiments, product production, biology and meteorology. Applications of the normal distribution in these areas mainly includes estimating frequency distributions, formulating ranges of reference values, and serving as the theoretical basis for many statistical methods. The application of normal distribution can also make sure that the quality of a product or service meets predetermined standards during the production process [1].

On the topic of normal distribution, there has been a deep level of research in the field of mathematics. This paper mainly writes the derivation of the formula of normal distribution, and also involves the application of normal distribution in some fields. However, it can not only be limited to the research in the field of mathematics, but also make the normal distribution more widely applied in other fields [2].

In the section 2, the paper starts with the background knowledge of normal distribution, then proves some lemmas and explains the application of the theorem from the derived formula, finally deduces the formula of normal distribution. In the section 3, this paper obtained examples of the application of normal distribution from the students' test scores and the error test of asphalt scale calibration.

2. Methods and Theory

Normal distribution is also called Gaussian distribution. It is one of the probability distributions of continuous random variables, which plays an important part in statistics. The standard normal distribution is often used in everyday mathematical calculations, and the standard normal distribution is a normal distribution of $\mu = 0, \sigma = 1$. The standard normal distribution has all the characteristics of the normal distribution [1]. The normal distribution was first found by Abraham de Moivre in the asymptotic formula. For the binomial distribution. Gauss derived the normal distribution from another aspect. Thanks to the discoveries of these mathematicians, the foundation was laid for subsequent related research.

2.1 Lemma and Theorem

Lemma 1. If the function $g(x)$ is an even function with a second derivative, then $g'(x)$ is an odd function; $g''(x)$ is an even function again [3].

Proof. Because $g(-x) = g(x)$ so the $g'(-x)(-1) = g'(x)$, namely: $g(-x) = -g'(x)$, with: $g''(-x) = g''(x)$.

Lemma 2. If the function g(x) satisfies the following conditions: (a) $g(0) = 0$, (b) $g(x)$ derivable and continuous function, (c) $g(x)$ is even function, (d) For any natural number m and real number $x : g(mx) = g'(x)$, then for any real number x the function $g(x)$ must have the form: $g(x) = cx$ (where c is a constant).

Proof. When $x = 0$, this is true. When $x \neq 0$, let $x = \frac{1}{m}$, then $g'(1) = g'\left(m \times \frac{1}{m}\right)$. When $x > 0$, let $x = \frac{n}{m}$, then $g'(x) = g'\left(\frac{n}{m}\right) = g'\left(\frac{1}{m}\right) = g'(1)$.

For any irrational number x, A series of rational numbers that can be constructed. such that the sequence tends to x, $\lim_{n\to\infty} xn = x$. Thus, $g'(x) = g'\left(\lim_{n\to\infty} Xn\right) = \lim_{n\to\infty} g'(Xn) = \lim_{n\to\infty} g'(1) = g'(1)$. Let $g'(1) = c$, integrating both sides can lead to $g(x) = cx$.

Lemma 3. The integral value of the function e^{-x^2} over the whole field of real numbers is $\sqrt{\pi}$, that is $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$. Assume D is the first quadrant of the rectangular coordinate system. So $D = [0, +\infty] \times [0, \infty)$.

Proof. let Dr be the circle with the origin as the center and radius r, the intersection with D, the intersection is the first quadrant of the circle (see Fig. 1 for illustration). According to the double integral theory, it can be obtained

$$\iint_D e^{-(x^2+y^2)} d\sigma = \lim_{R\to\infty} \iint_D e^{-(x^2+y^2)} d\sigma = \lim_{R\to\infty} \int_0^{\frac{\pi}{2}} d\theta \int_0^R e^{-r^2} r dr = \lim_{R\to\infty} \frac{\pi}{4}\left(1 - e^{-R^2}\right) = 4\pi \#(1)$$

Assume $Sa = [0, a] \times [0, a], a > 0$, then integrating S_a and one finds that

$$\iint_{Sa} e^{-(x^2+y^2)} d\sigma = \int_0^a e^{-x^2} dr \int_0^a e^{-y^2} dy = \left(\int_0^a e^{-x^2} dx\right)^2. \quad (2)$$

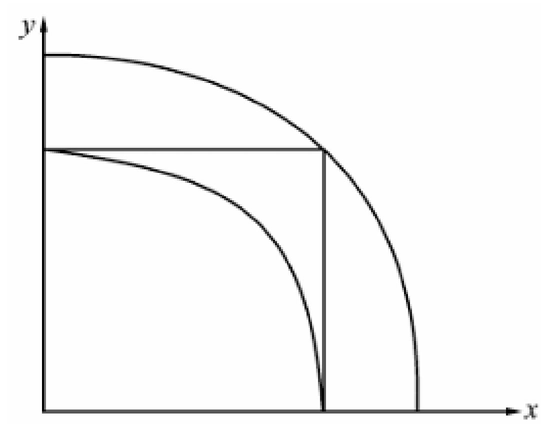

Fig. 1 The integral areas. Sa represents the area of the square in the graph, and Da represents the area of the larger $1/4$ circle in the graph.

Since $e^{-(x^2+y^2)}$ is a non-negative function. So:

$$\iint_{Da} e^{-(x^2+y^2)} d\sigma \leqslant \iint_{Sa} e^{-(x^2+y^2)} d\sigma. \quad (3)$$

Thus,

$$\left(\int_0^{\infty} e^{-x^2}\right)^2 = \left(\lim_{a\to\infty} \int_0^a e^{-x^2} dx\right)^2 = \lim_{a\to\infty}\left(\int_0^a e^{-x^2} dx\right)^2 = \iint_D e^{-(x^2+y^2)} d\sigma = 4\pi \quad (4)$$

ISSN 2959-6157

Therefore, the integral $\int_0^\infty e^{-x^2}dx=\frac{\sqrt{\pi}}{2}$, or equivalently, $\int_{-\infty}^{\infty} e^{-x^2}dx=\sqrt{\pi}$.

2.2 Derivation of Normal Distribution Form

Random error is a typical random variable. When measuring an object, the arithmetic mean of the results of multiple measurements as the overall average true value is better than the results of word measurements as an estimate. Then the distribution that the error random variable obeys must be such a "meticulous" function, which can always make the arithmetic mean of the sample become the best estimator among the overall true value estimators [4].

Assume that the true value of the population is θ, there is a simple random sample $x_1, x_2 \ldots, x_n$, regarding the measurement error, consider the estimator when the following likelihood function takes the maximum value:

$$L\left(\hat{\theta}\right)=\max_{\theta} L(\theta) \tag{5}$$

$$L\left(\hat{\theta}\right)=\max_{\theta} f(x_1-\theta)\,f(x_2-\theta)\ldots f(x_n-\theta) \tag{6}$$

Gauss first admitted that $\bar{x}$ was already a desirable estimate, and then went to the error density function to accommodate this, that is to find $f(x)$ such that $\hat{\theta}$,as determined by the above equation, is $\bar{x}$. Because of the problem is equivalent to

$\max_{\theta} ln[f(x_1-\theta)\,f(x_2-\theta)\ldots f(x_n-\theta)]$. That is, it is equal to $\max_{\theta}\sum_{I=1}^{n} ln\left(f(x_i-\theta)\right)$. When taking the maximum value,

$$\sum^{n}_{i=1}\ln\acute{\theta}\left(f(x_i-\theta)\right)=\sum_{i=1}^{n}\frac{f\acute{\theta}\left(x_i-\hat{\theta}\right)}{f\left(x_i-\hat{\theta}\right)}=0 \tag{7}$$

Use the auxiliary function $g(x)=\frac{f'(x)}{f(x)}$, then:

$$\sum_{i=1}^{n} g\left(x_i-\hat{\theta}\right)=0. \tag{8}$$

According to lemma 1 and $f(x)$ even function properties, one can know $g(x)$ is an odd function which satisfies $g(x)=-g(-x), g(0)=0$. Quote the natural number m, and let $n=m+1$, so $x_1=x_2=x_3=\ldots=x_m=-x$, $x_{m+1}=mx$, at this time $\hat{\theta}=\bar{x}=0$. Because of

$$\sum_{i=1}^{n} g\left(x_i-\hat{\theta}\right)=\sum_{i=1}^{n} g(x_i)=0, \tag{9}$$

and according to the previous formula, then $g(mx)=mg(x)$. This formula is consistent with all natural numbers m and real numbers x. Assume $g(x)$ is derivable and the derivations are continuous, take the derivative of x on both sides of this equation, then one can get

$$\frac{dg(mx)}{d(mx)}\times\frac{d(mx)}{dx}=m\frac{dg(x)}{dx}. \tag{11}$$

Because of $g'(x)=\frac{f''(x)\,f'(x)-\left[f'(x)\right]^2}{\left[f'(x)\right]^2}$, $f(x)$ is an even function. According to lemma 1, $f''(x)$ is even function, so $g'(x)$ is even function. Since $f(x)$ has second-order conductance, so $g(x)$ is continuous. Then $g(x)$ conforms to all conditions from Lemma 2, so

$$\frac{f'(x)}{f(x)}=cx. \tag{11}$$

Let $g(x)=cx$ and since $\frac{f'(x)}{f(x)}=\frac{d\left[ln(f(x))\right]}{dx}$, thus $ln\left(f(x)\right)=\int cxdx=\frac{1}{2}cx^2+c$. Therefore,

$$f(x)=e^{\frac{1}{2}cx^2+c'}=Me^{\frac{1}{x}cx^2}\left(M\triangleq e^{c'}\right) \tag{12}$$

As $f(x)$ always greater than 0, $f(x)$ is a density function when the integral of the whole real field is equal to 1. c must be a constant less than 0. $c=-\frac{1}{\sigma^2}(\sigma>0)$. Thus, $\int_{-\infty}^{\infty} Me^{\frac{1}{2\sigma^2}x^2}dx=1$. Through lemma 3, one finds that $\int_{-\infty}^{\infty} e^{-x^2}=\sqrt{\pi}$. Let $y=\frac{1}{\sqrt{2}\sigma}=x$, then $dx=\sqrt{2}\sigma dy$. So,

$$1=\int_{-\infty}^{\infty} Me^{-\frac{1}{2\sigma^2}x^2}dx=M\int_{-\infty}^{\infty} e^{-y^2}\left(\sqrt{2}\sigma dy\right)= M\left(\sqrt{2}\sigma\right)\int_{-\infty}^{\infty} e^{-y^2}dy=M\left(\sqrt{2}\sigma\right)\sqrt{\pi}. \tag{13}$$

Thus, the constant $M=\frac{1}{\sigma\sqrt{2\pi}}$. Finally, the formula for the normal distribution is

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}x^2}. \tag{14}$$

3. Results and Application

3.1 Application of normal distribution in student test score statistics

The data is a random selection of first-year students (200) in a certain English test, and the selected data is shown in Table 1 [5].

Table 1. The grade intervals and the corresponding student numbers.

Grade Interval	20.5-30	30-40	40-50	50-60	60-70	70-80	80-90	90-90.5
Students Number	5	15	30	51	60	23	10	6

In order to analyze these data, the first step is to draw a histogram of the data set. In this table 1, can know the maximum grade value 90.5, the minimum grade value 20.5, all the data is in the range [20.5-90.5]. Now take interval [20-100], it can cover the interval [20.5-99.5], The interval [20-100] is equally divided into 8 small intervals, see Table 2. Let the length between each small interval equal Δ, $\Delta = 10$. Δ means class interval. The endpoints between cells are called group limits. Then calculate the frequency (f_i) and frequency ($\frac{f_i}{n}$) in each small interval.

Table 2. The 8 small intervals and the corresponding information.

Class interval limit	Frequency (f_i)	Frequency ($\frac{f_i}{n}$)	Cumulative relative frequency
20-30	5	0.025	0.025
30-40	15	0.075	0.1
40-50	30.	0.15	0.25
50-60	51	0.255	0.505
60-70	60	0.3	0.805
70-80	23	0.115	0.92
80-90	10	0.05	0.97
90-100	6	0.03	1

Next, one can make the x^2 fit test. Assume the probability density of $H_0 : X$ is:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\times 15^2}}, -\infty < x < +\infty \tag{15}$$

Since the value of μ, σ^2 is not given in H_0, so one need to estimate μ, σ^2 first. Through maximum likelihood estimation can get $\hat{\mu} = 60, \hat{\sigma}^2 = 15^2$. If H_0 is true, the estimated probability density of X is:

$$f(X) = \frac{1}{\sqrt{2\pi}\times 15} e^{-\frac{(x-60)^2}{2\times 15^2}}, -\infty < x < +\infty \tag{16}$$

On the basis of the above formula, the distribution function table of the standard normal distribution is queried and one can get the estimation of probability $P(A_i)$. Namely,

$$\hat{P}_1 = \hat{P}(A_1) = \hat{P}(20 < x < 30) = \Phi\left(\frac{30-60}{15}\right) - \Phi\left(\frac{20-60}{15}\right) = 0.019. \tag{17}$$

The calculation results are shown in Table 3.

Table 3. The data set and the corresponding probability.

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2 / n\hat{p}_i$
$A_1 : 20 < x < 30$	5	0.019	3.8	6.6
$A_2 : 30 < x < 40$	15	0.069	13.8	16.3
$A_3 : 40 < x < 50$	30	0.1596	31.96	28.2

ISSN 2959-6157

$A_4: 50 < x < 60$	51	0.2486	49.72	52.3
$A_5: 60 < x < 70$	60	0.2486	49.72	72.4
$A_6: 70 < x < 80$	23	0.1596	31.91	16.6
$A_7: 80 < x < 90$	10	0.069	13.8	7.2
$A_8: 90 < x < 100$	6	0.019	3.8	9.5

In Table 3, one can get $\sum_{i=0}^{8}\frac{f_i^2}{n\hat{p}_i} = 209.1$. Now $x^2 = 209.1 - 200 = 9.1$, since $x_{0,1}^2(k-r-1) = x_{0,1}^2(8-1) = x_{0,1}^2(7) = 12.017 > 9.1$, one should accept H_0 because it is below the level 0.1. Also, one can show that the data comes from the normally distributed population. Through the above analysis, it can be concluded that students' exam grades fit the normal distribution.

3.2 Normal Distribution in Calibration and Error Test of Reduced Asphalt Scales

According to the error sources and influencing factors in the calibration of the reduced asphalt scales, it can be considered that the test data of buoyancy coefficient conform to the normal distribution, so the normal distribution method is used to analyze the calibration and inspection of the reduced asphalt scales [6].

Assume population X is subject to a normal distribution, that is to say $X\ N(\mu,\sigma^2)$, $(X_1, X_{2,\dots} X_n)$ is the sample of volume n taken from the population X,and $(x_1, x_{2,\dots} x_n)$ is the measured value of $(X_1, X_{2,\dots} X_n)$, now the author will establish confidence probability $1-a$. The confidence interval is (Z_1, Z_2); If $P(Z_1 < Z < Z_2) = 1-a$ is true, then $(Z_1 and Z_2)$ are the confidence intervals of the estimated parameter Z in the confidence probability $1-a$. The interval length of Z_1 and Z_2 is the interval estimation accuracy.

Since the population sample follows a normal distribution, then $X\ N(\mu,\sigma^2)$, the values taken from the population also follow a normal distribution, can get $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} N(0, 1)$, but the value of σ^2 is not given, so should consider using the sample variance s^2 to estimate the population variance σ^2, the test statistic is $T = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} t(n-1)$, the probability that T is between $t_{1-\frac{a}{2}}(n-1)$ and $t_{\frac{a}{2}}(n-1)$ is $1-a$, to wait $P\left\{t_{1-\frac{a}{2}}(n-1) < T < t_{\frac{a}{2}}(n-1)\right\} = 1-a$. Since $t_{1-\frac{a}{2}}(n-1) = -t_{\frac{a}{2}}(n-1)$, can get a confidence interval for the population mean:

$$\left[\bar{X} - \frac{s}{\sqrt{n}} t_{1-\frac{a}{2}}(n-1), \bar{X} + \frac{s}{\sqrt{n}} t_{1-\frac{a}{2}}(n-1)\right] \#(18)$$

The sample error calculation formula can be obtained as

$$e = \frac{s}{\sqrt{n}} t_{1-\frac{a}{2}}(n-1) \#(19)$$

Then the sample size expression can be obtained as

$$n = \left[\frac{t_{1-\frac{a}{2}}(n-1) C_v}{|e|}\right]^2, C_v = \frac{s}{\bar{X}}, |e| = \frac{e}{\bar{X}} \#(20)$$

Here, C_v is sample deviation coefficient, $|e|$ is the relative value of the sample error. The expression of interval estimation accuracy is:

$$\frac{e}{s} = \frac{t_{1-\frac{a}{2}}(n-1)}{\sqrt{n}} \#(21)$$

where e is sample error, s is sample standard deviation, n is sample size, a take 0.5. The conclusion is that the accuracy of interval estimation under different sample sizes can be obtained from the standard normal distribution table.

4. Conclusion

This paper provides a comprehensive overview of normal distribution, beginning with the essential background knowledge necessary to understand its significance. It explores the fundamental theorem behind normal distribution, detailing the derivation of its formula to illustrate how it is mathematically formulated. Additionally, the paper highlights the practical applications of normal distribution in everyday life, demonstrating its relevance in diverse fields such as education, and manufacture. The introduction demonstrates normal distribution is not only a mathematical concept, but also is an important tool in each area of real life, the analysis and understanding of everyday data and probabilities has also been enhanced. It is hoped that in future research, the application of normal distribution in other fields besides mathematics can be further explored. For example, in the medical field, normal distribution can be used to analyze clinical trial data and promote the development of new drugs to optimize treatment programs. In finance, normal distributions are applied in risk assessment and portfolio management to help investors better understand market volatility and risk. Generally speaking, the normal distribution will continue to play a fundamental role in scientific research and practical applications. Its significance spans across multiple areas, including psychology, finance, and healthcare, promoting the development of various fields by providing essential tools for data analysis, risk assessment, and informed decision-making in real-world scenarios.

References

[1] JIA Gui-cong, Liu Jian-hui and SHI Cui-fang. Investigation and analysis of clinical blood records in medical institutions of Hebei Province. Chinese Health Quality Management 2014, 21(6): 97.

[2] WEI Chi-jiang. Derivation of Gaussian distribution function form. Statistics &Information Forum 2019, 24(5): 3-5.

[3] Xu Xiao-li. Student test Score Statistics from a study of Normal Distribution. China Academic Journal Electronic Publishing House. 2012, 24: 160-161.

[4] ZHU Jia-Jian, ZHENG Zhi-Xuan, GUO Xin, LI Fu-le. Research on calibration and on-line material inspection technology of reduced asphalt scale based on normal distribution. China Academic Journal Electronic Publishing House. 2023, 11: 115-118.

[5] Agu, Friday, I. And E. Francis, Runyi. Comparison of Goodness of Fit Tests for Normal Distribution. Asian Journal of Probability and Statistics, 2018, 6: 1-32.

[6] Meeker W Q, Escobar L A, Pascual F G. Statistical methods for reliability data. John Wiley & Sons, 2022.