Dean&Francis
ISSN 2959-6157

Unveiling Different Probability Distribution Functions and Their Applications

Chengman Li

Wuhan Britain-China School, Wuhan, China
lixiaol.zhsh@sinopec.com

Abstract:

Nowadays, data is everywhere from social economic activities to the observation of natural phenomena, the complexity and scale of data requires researchers to master more advanced data analysis tools. As a bridge between random events and mathematical models, probability distribution function not only provides a theoretical basis for describing data characteristics, but also provides an important means for quantitative analysis of data and modeling prediction. This paper focuses on several typical probability distributions which have wide application background and deep theoretical foundation in their respective fields. This paper mainly focuses on two core issues. The first is the introduction of various probability distribution functions and their applications, while the second is how to use these distribution functions for effective data analysis and prediction. Through a series of empirical studies, this paper verifies the applicability and validity of different probability distribution functions in different fields. The results show that the correct selection and application of probability distribution function can significantly promote the accuracy and efficiency of data analysis, and provide strong support for decision making. Finally, the research in this paper not only makes the theoretical system of probability distribution function more abundant, but also offers some valuable references to practical applications. By revealing the key role of different distribution functions in data analysis and prediction, this paper provides new ideas and new methods for interdisciplinary research.

Keywords: Probability distribution function; Expectancy value; Variance.

ISSN 2959-6157

1. Introduction

Nowadays, data collection, processing and analysis have become an indispensable part in many fields such as natural science, social science, engineering technology, economy and finance. The inherent regularity and uncertainty of data are often described by probability distribution functions, which not only reveal the mathematical characteristics of random phenomena, but also provide a solid theoretical basis for prediction, decision optimization and risk management. Therefore, it is of great significance to deeply understand and master various probability distribution functions and their applicability in different application scenarios to reinforce the depth and range of data science research [1]. And nowadays the big data technology is developing rapidly, the application of probability distribution function has made remarkable progress. How to select and fit probability distribution models more effectively and how to apply these models to practical problem solving has been widely discussed in academia and industry. On the one hand, researchers constantly propose new probability distribution models to better adapt to the complex and changeable data characteristics; On the other hand, the expansion and application of traditional probability distribution functions continue to deepen, such as mixed distribution model, truncated distribution model. In the processing of non-standard data show unique advantages. In addition, the introduction of machine learning and deep learning technology can support some new ideas and methods to parameter estimation and model selection of probability distribution function [2].

As the core concept of statistics and probability theory, probability distribution function describes the probability law of random variable values. From simple the discrete distributions such as binomial distribution and Poisson distribution to the continuous distributions such as normal distribution, exponential distribution and Weibull distribution, each distribution corresponds to a specific random phenomenon and data generation mechanism. These distributions not only occupy an important position in theoretical research, but also are powerful tools for dealing with randomness and uncertainty in practical applications. In finance, for example, normal distribution is used to describe the volatility of asset returns; In biomedical research, Poisson distribution is often used to analyze the number of appears per unit time [3].

In view of the above background, this paper aims to comprehensively review and discuss the basic characteristics, theoretical basis, and application examples of various probability distribution functions in many fields. Specifically, this paper will first introduce several common probability distribution functions (including but not limited to the above-mentioned ones), and analyze their mathematical expressions, properties, and applicable conditions; Then, through case analysis, it shows the specific application of these distributions in financial risk management, biomedical statistics, engineering technology optimization.

2. Discrete Probability Distribution Function

2.1 Background Knowledge

The discrete probability distribution function (PDF) is that describes probability corresponding to their possible value of a discrete random variable. It has two fundamental properties, the first one is non-negativity, which means that all values x_i is this random variable X have probabilities greater than 0 $p(x_i > 0)$. The second one is regularity, which means that the sum of the probabilities of all possible values is 1 $\Sigma_i p(x_i) = 1$. For the expectancy value, it means the average of all possible values of a random variable [4]

$$E(X) = \sum_i [x_i \cdot p(x_i)] \tag{1}$$

In addition, it has two properties, the first is linear and the second is related to variance

$$E(aX + bY) = aE(X) + bE(Y), Var(X) = E[(X - E(X))^2] \tag{2}$$

2.2 Different Types

Bernoulli distribution. The probability distribution of only two possible outcomes in a randomized experiment. The PDF is

$$P(X = k) = \begin{cases} p, if k = 1 \\ 1 - p, if k = 0 \end{cases} \tag{3}$$

where p means the probability of success. For this distribution, the expectancy value is $E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$, and the variance:

$Var(X) = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p)$.

Binomial distribution. It is probability distribution consisting of several independent and repeated Bernoulli tests. The PDF is [5]

$$P(X = k) = C(n,k) \times p^k \times (1-p)^{(n-k)} \tag{4}$$

where n means the number of tests, k means the number of successful tests then p means that the probability of the number of successful tests. The expectancy value is $E(X) = np$, and the variance is $Var(X) = np(1-p)$.

Poisson distribution. A probability distribution that describes number of times an event appears. The PDF is

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!} \tag{5}$$

where λ means the average occurrence of an event and λ>0. It satisfies the relation $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}= \quad e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=$ $e^{-\lambda}e^{\lambda}=1$. The expectancy value is $E(X)=\lambda$, and the variance is $Var(X)=\lambda$.

Uniform distribution. It is a symmetric probability distribution in which each result is equal and they occur at any points in this distribution. The PDF is

$$f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0, others\end{cases} \tag{6}$$

Regarding the cumulative distribution function, $F(x)$ shows the random variable X that is less than the probability of x, i.e.,

$$F(x)=\begin{cases}0, x<a\\ \frac{x-a}{b-a}, a\leqslant x<b\\ 1, x\geqslant b\end{cases} \tag{7}$$

The expectancy value is $E(X)=\frac{a+b}{2}$ and the variance is $Var(X)=\frac{(b-a)^2}{12}$.

Exponent distribution. It describes the probability of an event appearing within a constant time interval. The PDF is

$$f(x)=\begin{cases}\frac{1}{\theta}e^{-x/\theta}, x>0\\ 0, others\end{cases} \tag{8}$$

where $1/\theta$ means rate parameter and $1/\theta>0$. For the cumulative distribution function, F(x) shows the random variable X that is less than the probability of x, i.e.,

$$F(x)=\begin{cases}1-e^{-x/\theta}, x>0\\ 0, others\end{cases} \tag{9}$$

The expectancy value is $E(X)=1/\lambda$ and the variance is $Var(X)=1/\lambda^2$.

2.3 Application

The first example is about the Poisson distribution. If the probability of k customers coming to the store in a day follows the Poisson distribution, and each customer arriving at the store buys goods independently, the probability is P.

Supposed that A means the mall had k customers a day, and B means the mall had r customers buy products a day. Therefore, $P(A_k)=\frac{\lambda^k e^{-\lambda}}{k!}(k=0,1,2,...,r,...)$ and $P(A_k\mid B)=C_k^r p^r(1-p)^{k-r}(k=r,...)$ [6]. Thus,

$$P(B)=\sum_{k=0}^{\infty}P(A_k)P(B\mid A_k)=\sum_{k=r}^{\infty}\frac{\lambda^k e^{-\lambda}}{k!}C_k^r p^r(1-p)^{k-r} \tag{10}$$

After some derivation, it is found that

$$P(B)=\frac{(\lambda p)^r e^{-\lambda}}{r!}e^{\lambda(1-p)}=\frac{(\lambda p)^r e^{-\lambda p}}{r!}. \tag{11}$$

It is also found that the expectancy value is $P(X=r)=\frac{(\lambda p)^r e^{-\lambda p}}{r!}(r=0,1,...)$. Therefore, $X\,P(\lambda p)$, $E(X)=\lambda p$.

The second example is about the Binomial distribution. A workshop has 10 machines of the same type, each machine equipped with a motor which has power of 10 kilowatts, and when each machine works, average hours are started for 12 hours, and whether it is started is independent of each other. Due to local electricity is in short supply, for the power supply department only provides 50 kilowatts of power, and what is the probability that all 10 machines will appear at the same time and they cannot work due to lack of power? In an 8-hour shift, what is the approximate amount of time that is not working normally? Suggested that the actual operation of 10 machines is a random variable ξ , and the probability of each machine starting P, therefore $P=\frac{12}{60}=\frac{1}{5}$. Thus,

$$P(\xi=k)=C_{10}^k\left(\frac{1}{5}\right)^k\left(\frac{4}{5}\right)^{10-k}\left(k=0,1,2\ldots,10\right) \tag{12}$$

As a result, it is found that

$$P(\xi\leqslant 5)=C_{10}^0\left(\frac{4}{5}\right)^{10}+C_{10}^1\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^9+\ldots+C_{10}^5\left(\frac{1}{5}\right)^5\left(\frac{4}{5}\right)^5$$
$$\approx 0.994\#(13)$$

Therefore, under the condition of 50 kW power supply, the probability of the machine tool not working normally is only 0.006, so that the time of not working normally within 8 hours of a working day is about 2.88 minutes (cf. $8\times60\times0.006=2.88$).

The third example is about the uniform distribution. According to the statistics of a certain product, the demand in the international market is a uniform distribution of a random variable between 2000 and 4000. If 2800 tons are organized according to this law this year, what is the probability that the demand cannot be met?

ISSN 2959-6157

Supposed the probability density function is X, then

$$f(x)=\begin{cases}\frac{1}{2000} & 2000<x<4000\\ 0 & other\end{cases} \quad (14)$$

Thus, it is calculated that the probability is

$$P(2800<x<4000)=\int_{2800}^{4000}\frac{1}{2000}dx=\frac{1200}{2000}=0.6 \quad (15)$$

Therefore, the random event $(2800<X<4000)$ indicates that the amount of preparation cannot meet the demand, and its probability is that the probability of failing to meet the international market demand is 60%.

Finally, combined with current research hot-spots and future development trends, the challenges and opportunities that probability distribution function may face in the background of big data are discussed, and corresponding research directions and suggestions are put forward. At the same time, the results of this article are of great importance to promote the development of data science and other related fields, and help to promote the deep integration of theoretical innovation and practical application in these fields.

3. Continuous Probability Distribution Function

3.1 Background Knowledge

The Continuous Probability Distribution Function is a mathematical function that describes the probability distribution of continuous random variables taking any value within its defined interval. It has two properties, the first is that probability density function f(x) in the domain is 1, and the second one is that the probability of continuous random variable taking any single value is 0. Therefore, $\int_a^b f(x)dx=1$ and $P\{X=a\}=0$. Moreover, the expectancy value

$$E(X)=\int_{-\infty}^{\infty}xf(x)dx \quad (16)$$

3.2 Different Types

To begin with, the normal distribution is given by [7]

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (17)$$

where μ means central location of distribution and σ^2 means variance. The expectancy value $E(X)=\mu$ and the variance $Var(X)=\sigma^2$.

This distribution has two features. The function is symmetric about the mean, which has the same value as its mode and median. The inflection point of the density function is the position a standard deviation away from the mean.

The second distribution is Cauchy distribution, whose behavior is

$$f(x;x_0,\gamma)=\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}=\frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right] \quad (18)$$

Here, x_0 means location of distribution, γ means the Parameter sets half of the width at half of the maximum value. The major feature is that no mean, variance, or moment is defined. Its mode and median are equal x_0.

The Gamma distribution is defined as

$$f(x;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \quad (19)$$

The expectancy value is $E(x)=\int_0^\infty x\bullet f(x;k,\theta)dx=k\theta$, and the variance is $Var(X)=k(k+1)\theta^2-(k\theta)^2=k\theta^2$.

For the Log-normal distribution, its probability density is given by

$$\phi(y)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad (20)$$

and $-\infty<y<+\infty$. The expectancy value is $EX=e^{\mu+\frac{\sigma^2}{2}}$ and variance $DX=e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$. There are several features available. Firstly, it is found that $InY=aX+b\ N(a\mu+b,a^2\sigma^2)$. Secondly, supposed that X and Y are independent events, and $X\ N(\mu_1,\sigma_2^2)$, $Y\ N(\mu_2,\sigma_2^2)$, then $Z=e^{aX+bY}$. Thirdly, supposed that X and Y are independent events, and $InX\ N(\mu_1,\sigma_1^2)$, $Iny\ N(\mu_2,\sigma_2^2)$, then $InZ=Ln\left(X^aY^b\right)=aInX+\ bInY\ N(a\mu_1+\ b\mu_2,\ a^2\sigma_1^2+b^2\sigma_2^2)$. Fourthly, $\frac{1}{n}\sum_{i=1}^{n}InX_i\ N(\mu,\frac{1}{n}\sigma^2)$ and $\frac{1}{\sigma^2}\sum_{i=1}^{n}(InX_i-\mu)^2\ X^2(n)$. Finally, supposed that $W^2=\frac{1}{n-1}\sum_{i=1}^{n}(InX_i-\frac{1}{n}\sum_{i=1}^{n}InX_i)^2$, then $\frac{n-1}{\sigma^2}W^2\ X^2(n-1),\frac{\frac{1}{n}\sum_{i=1}^{n}InX_i-\mu}{w/\sqrt{n}}t(n-1)$.

3.3 Application

The first example is about the Normal distribution. A major recruit 20 graduate students, of which the top 10 are

free, the number of applicants is 1000 people, the full score of the exam is 500 points, after the exam, only to know that the average score of this professional exam is $\mu = 300$ points, if the score line for recruiting graduate students is determined to be 350 points, may I ask if someone is now taking 360 points, he is likely to be admitted as a free student?

The author assumes that graduate test scores conform to a normal distribution. According to the content, it is said that $P(X \geqslant 350) = \frac{20}{1000} = 0.02$ and $?\left(\frac{350-300}{\sigma}\right) = 0.98$.

Therefore, $?(2.05) \approx 0.98$ and $\sigma = \frac{50}{2.05} = 24.4$ [8]. Thus,

$$P(X \geqslant 360) = 1 - P(X < 360) = 1 - ?(\frac{360-300}{\sigma}) = 0.007 \tag{21}$$

Therefore, the candidate probably ranked seventh ($1000 \times 0.007 = 7$) in the exam, so he can be admitted as a free student.

The second example is about the Cauchy distribution. In physics, in signal processing, the Cauchy distribution is generally used to simulate pulse noise in signals. In economics, the Cauchy distribution can be used to model stock price fluctuations in financial markets, to assess the risk of financial products, and to quantify the design and validation of investment strategies. For example, the volatility and risk level of financial markets, with smaller scaling parameters, the distribution curve becomes steeper, indicating a lower degree of dispersion of the distribution.

The third example is about Log-normal distribution. A log-normal distribution is a continuous probability distribution characterized by the fact that when logarithmic values of random variables are taken, the values of these pairs follow a normal distribution. In other words, if a random variable X follows a log-normal distribution, then $In(X)$ follows a normal distribution, $X\ Log-Normal(\mu,\sigma^2)$ Log-normal distribution is widely used in finance, economics, biology, and engineering, especially when dealing with variables that are naturally log-normal, such as stock prices, the size of certain populations of organisms, and urban populations. In these cases, using a log-normal distribution can more accurately describe the distribution characteristics of the data.

The last example is about Brownian movement. If random progress $\{X(t), t \geqslant 0\}$ and $(i)X(0) = 0$; $(ii)\{X(t), t \geqslant 0\}$; $(iii)t > 0, X(t)N(0, \sigma^2 t)$, and $\sigma > 0, \{X(t), t \geqslant 0\}$. Suppose that a person owns a European call option of a certain stock whose delivery time is T and delivery price is P_1, that is, he has the purchasing right of such stocks at a fixed price P_1 at time T, so as to obtain benefits, otherwise he will give up the option. Suppose that today the price of such a stock is P_2, and that the price of the stocks changes according to the geometric Brownian motion of the parameter $\mu = 0, \sigma = 1$, then the average value of this option.

The author supposes that $P_2(T)$ means the prices of certain time T, and $P_2(0) = P_2, P_2(T) = P_2 e^{B(T)}$, if $P_2 > P_1$ and the option will be exercised. The expectancy value can be calculated as

$$\begin{aligned} &E\left[max\{P_2(T) - P_1, 0\}\right] = \\ &\int_0^\infty p\{P_2(T) - P_1 > u\}du = \\ &\int_0^\infty p\left\{B(T) > In\frac{P_1+u}{P_2}\right\}du \end{aligned} \tag{22}$$

Because $B(T)N(0,T)$, and probability density $f_{B(T)}(x) = \frac{1}{\sqrt{2\pi T}} e^{-\frac{x^2}{2T}}$ $(-\infty < x < +\infty)$,

so $P\left\{B(T) > In\frac{P_1+u}{P_2}\right\} = \frac{1}{\sqrt{2\pi T}} \int_{In\frac{P_1+u}{P_2}}^{+\infty} e^{-\frac{x^2}{2T}} dx$.

4. Conclusion

The author discusses the various probability distribution functions and their applications, discrete and continuous, which are not only fundamental tools in probability theory and statistics, but also play a crucial role in other subject areas. Such as in the design of bridges and dams, the probability distribution function should be used to estimate the probability of the highest water level of a river. In medical research, probability distributions can be used to analyze the incidence and mortality of diseases; In economics, it can be used to predict market trends and formulate investment strategies. However, with the increasing amount of data and the improvement of computing power, the application of probability distribution function is becoming more and more extensive. Thus, although the probability distribution function and its related theories have made remarkable achievements, with scientific and technological of the continuous development and faced new problems, there are still many topics worthy of further study. For example, how to better apply probability distribution function to complex system modeling, how to increase the efficiency and accuracy of probability distribution estimation, and how to expand the application of probability distribution function in new fields are all important directions

ISSN 2959-6157

of future research. In a word, as one of the cornerstones of probability theory, probability distribution function not only has a deep theoretical basis, but also has a wide application prospect. In the future research and practice, people should continue to deepen the understanding and application of probability distribution function, to promote the development of science and technology and social progress to make greater contributions.

References

[1] Yu Y. Application of lognormal Distribution in Stock price model. Journal of Langfang Normal University (Natural Science Edition),2012, 12(05): 69-72.

[2] Yu Yang. Some properties of lognormal Distribution and its Parameter Estimation. Journal of Langfang Normal University (Natural Science Edition),2011, 11(05):8-11.

[3] Dong Hongling. The application of gamma distribution in the carbon market research. Zhejiang university of science and technology, 2022.

[4] Zhang C N. Statistical inference based on autoregressive model of non-normal distribution network. Jilin University, 2024.

[5] Rowling, Xiao Chengying, Wu Yannan. Differential method in the application of the probability density function with examples. Computer knowledge and technology, 2020, (4): 221-223.

[6] Wu Shuchen, Hou Chaoqun, Sun Zhibin. Research on stability calculation model of infinite slope with non-uniform initial water content distribution. Journal of Hefei University of Technology (Natural Science Edition), 2024, 47(05): 690-695.

[7] WU Yiqi, Xiao Xiang, Gu Xi. Objective Bayesian Analysis based on 0-1 dilatation binomial Distribution. Computer Applications and Software, 2019, 41(04): 46-52+59.

[8] Hu Chunyan, Hu Liangping. Reasonable mean comparison: Poisson distribution regression model. Sichuan Mental Health, 2019, 36(S1): 13-17.