# Quantile estimates of entropy uncertainty for distributions supported on bounded interval

*Vitaly G* Polosin*

Medical Cybernetics and Computer Science Department, Medical Institute of Penza State University, 40 Krasnaya Street, Penza, 440026, Russian Federation

**Abstract.** Formulas for calculating the Shannon information entropy and the entropy uncertainty interval are obtained, which are based on calculating quantile estimates of random variable distributions and are set over a limited range of random variable values supported both on a semi-infinite interval and on the entire real line. In this paper, using the example of a generalized beta distribution of the first kind, the possibility of determining quantiles for the entire variety of possible shapes of a given distribution subfamily is illustrated. To assess the quality of the approximation construction, a study was conducted, the purpose of which was to compare estimates of the uncertainty of a complex system using analytically specified information entropy for the Kumaraswamy distribution and information entropy obtained on the basis of an approximating formula using quantile estimates of the Kumaraswamy distribution. Based on the study, it is shown that when choosing the sampling intervals specified by the percentiles of the distribution, the approximation error did not exceed 1% for the range of the most used parameters of the power and shape of the Kumaraswamy distribution.

## 1 Introduction

Modern methods of establishing control over complex stochastic systems are based on tracking the correspondence of the output distribution to the controlled parameters of the target output distribution model. Due to the fact that most of the controlled output parameters are characterized by a limited range of random values, the distributions should be distinguished among the entire variety of distributions those that are specified in a limited interval. The reasons for the limitations of the range of random values are often related to both the design features of technological processes and the physical properties of the measured quantity. For example, when the marginal energy is accumulated by the system that it is limits its controllable properties. In this case, statistical distributions are often used as a simulation model of the output parameters of complex systems if these are supported on a bounded interval.

In all areas of human activity complex systems with the bordered range of output values are applauded. As an example, it is possible to single out such research areas as particle size distributions limited by the technological process [1-3], distribution of the lengths of

* Corresponding author: polosin-vitalij@yandex.ru

molecules of petroleum products [4], population growth restrictions due to the limitation of limiting reserves in the study demographic and economic processes [5], limitations of properties for phase transitions of the material [6], limitation of the energy of microparticles during nuclear decay [7] and others.

A special place is occupied by studies of the characteristics of complex computer networks and computing systems that are used for control in queuing systems and the production of the Internet of Things [8-10]. The shapes variety of used distributions makes it possible to take into account many different facts such as the non-linearity of the characteristics of the system elements, a variety of external impacts, and the cyclical nature of internal processes during the development of various epochs in the system [11]. Since the properties of the model largely depend on the shape variety where are available for implementation and sufficient for solving the assigned control tasks, at the moment it is remained relevant to the research of methods for constructing criteria for controlling distribution shapes. The choice of a distribution model is often associated with providing the condition according to which the model retains the ability to implement the required shape set with the least complexity of its mathematical formalism. This is necessary to solve the problems of both selecting a model and estimating the residual variability of a set of its forms. In this paper, it is proposed to apply quantile estimates of entropy uncertainty as the characteristic for the distribution of data over the bordered interval.

## 2 Estimates for information uncertainty

The behavior of a complex object is often characterized by its uncertainty, which is usually estimated by the variability of the output observed values. The average square deviation is often used as a measure of uncertainty intervals in statistical physics. In information theory, the uncertainty of a system is estimated using information entropy, which characterizes the unpredictability of the appearance of a message or symbol.

To determine the differential information entropy on the set of continuous probability distribution functions $f(x)$ of a random variable $X$, the formula proposed by K. Shannon [12] is used, which is given as

$$H(X) = E\left(-\log\left(f(X)\right)\right) = -\int_{-\infty}^{\infty} f(x)\cdot\log f(x)\cdot dx\cdot \qquad (1)$$

Information entropy is a measure of the uncertainty of a system, which is characterized by its unpredictability about an experiment or events that are observed in the system. The greater the number of states is possible for a system, the greater the uncertainty value characterized this system. To assess the uncertainty of the system, entropy is used as a quantitative characteristic of the information system.

Since the value of the differential information entropy on a set of continuous functions of probability distributions $f(x)$ cannot always be obtained in the form of an analytical random variable, its value is more conveniently expressed if is using methods of approximate numerical integration. The main idea these methods is to replace the integral function with a simpler one that made it possible the integral is easily calculated analytically. The most general form of the quadrature formula for calculating the integral is given by [13]:

$$H(X) = \sum_{i=1}^{n} w_i \cdot f_i \quad . \qquad (2)$$

Where $n$ is the number of points where the value of the $f_i$ integrand function; is the weight of the nodal point

For quadrature calculations, rectangle methods with constant length, height, or area are often used. If the interval lengths are set based on the equal probability condition, then the

interval boundaries will be determined by the quantiles of the distribution, which is convenient for constructing computational processes.

Since the grouping interval is equal to the difference of the probability quantiles, the product of the grouping interval by the distribution density will determine the constant discrete probability of an observation in the *i*-th grouping interval.

If the integration is expressing by summation that it is obtain a formula for determining the differential entropy. It is given by

$$H(X) = -\sum_{i=1}^{n} p_i \cdot \log f(x_i) \,. \tag{3}$$

Where $p_i$ is the probability of the $X$ random variable for the $i$-th summation interval.

Since in the case of quadrature calculations is using the rectangle method with a constant area than for observing random variable $X$ the $p_i$ probabilities are remain constant and these are equal to the $\Delta p$ probability. Then the values of the boundaries of the $x_i$ observation intervals will be determined as the quantiles of the probability distribution $Q_i$, that equal to the product $(i \cdot \Delta p)$.

As for as the probability $\Delta p$ of observing a random variable $X$ in the $i$-th interval remains constant, the expression for determining entropy is possible to transformed as

$$H_Q(X) = \ln\left( n\prod_{i=1}^{n} \left(Q(p_i) - Q(p_{i-1})\right)^{(1/n)} \right). \tag{4}$$

Where $Q_i$ is the $i$-th probability quantiles that is equal to the $(i \cdot \Delta p)$. product.

## 3 Quantile entropy estimates for the generalized beta distribution

The paper investigates quantile entropy using the example of a well-known commonly used subfamily, the generalized beta distribution of the first kind. For this distributions the standardized cumulative function is given using a regularized incomplete beta function as the ratio of the incomplete beta function $B(x^a,u,v)$ of the $X$ random variable to complete the beta function $B(u,v)$

$$F(x,a,u,v) = \frac{B(x^a,u,v)}{B(u,v)} \,. \tag{5}$$

Where $u$, $v$ are the first and second shape parameters of the beta function $B(u,v)$, $a$ is the power parameter of the distribution shape.

The density function of the standardized form of the generalized beta distribution of the first kind is given by

$$f(x,a,u,v) = \frac{a}{B(u,v)} x^{au-1}\left(1-x^a\right)^{v-1} \,. \tag{6}$$

Generalized beta distributions of the first and second kind were introduced in [14]. A set of shapes corresponds to a standardized distribution, provided that the scale parameter is one. Due to the variety of possible forms included in the subfamilies of the generalized beta distribution, its functions are often used in modern research [15].

The distribution function (6) contains beta distributions of the first kind and Kumaraswamy distributions, the choice of standardized forms of which is achieved by setting the power parameter a or the first parameter of the form u equal to one, respectively. If these $Q^*(p,a,u,v)$ and $Q(p,u,v)$ are quantile functions of generalized and standardized beta distributions of the first kind, then the quantile functions are related by

$$Q^*(p,a,u,v) = \left(Q(p,u,v)\right)^{\frac{1}{a}} \,. \tag{7}$$

Tables of relations of the incomplete beta function for $u \geq v$ are given in [16]. Many packages of applied mathematical programs, such as MatLab, MathCad, SciLab, and others,

contain the inverse beta distribution function of the 1st kind, which makes it possible to calculate quantiles for all possible forms of the generalized beta distribution using the ratio (8). Various approximations are also used to calculate quantiles, which contain various scientific studies, for example [17-20].

It should be noted that the quantile function of separate subfamilies often has an algebraic representation form. For example, the quantile function for the simplified form of the Kumaraswamy distribution can be obtained from the distribution function (5) by taking the parameter *u* equal to 1. The formula for calculating the quantiles of the Kumaraswamy distribution is given by

$$Q_{Kw}(p,a,v)=\left(1-\left(1-p\right)^{\frac{1}{v}}\right)^{\frac{1}{a}}. \tag{8}$$

By substituting quantiles (8) in expression (4) made it possible to calculate information entropies for any forms of implementations of the generalized beta distribution of the first kind. In particular, for a quantile estimate of the entropy of the Kumaraswamy distribution, an expression of the form is valid

$$H_{Kw}(X)=\ln\left(n\prod_{i=1}^{n}\left(\left(1-\left(1-p_i\right)^{\frac{1}{v}}\right)^{\frac{1}{a}}-\left(1-\left(1-p_{i-1}\right)^{\frac{1}{v}}\right)^{\frac{1}{a}}\right)^{\left(1/n\right)}\right). \tag{9}$$

Since expression (5) is a quadrature approximation for calculating the integral expression (1) using rectangles with a constant area, an estimate of the quality of the approximation can be obtained by comparing the results with the information entropies of the generalized beta distribution obtained in a theoretical form. Expressions for the Information entropy of the generalized beta distribution can be found as:

$$\begin{aligned}H_{GB1}(\vartheta,a,u,v)=&\ln\left(\vartheta\cdot a^{-1}\cdot B\left(u,v\right)\right)+\left(a^{-1}-u\right)\cdot\\&\cdot\left(\psi(u)-\psi(u+v)\right)+\left(v-1\right)\cdot\left(\psi(u+v)-\psi(v)\right).\end{aligned} \tag{10}$$

Where $\vartheta$ is a parameter of the distribution scale equal to the range of the random variable *X*. For standardized distributions, it is equal to 1.

Accordingly, if the condition is accepted that the $\vartheta$ scale parameter and the *u* first shape parameter are equal to one then expression (10) is transformed to the information entropy of the Kumaraswamy distribution. For calculating this information entropy, the formula is given

$$\begin{aligned}H_{Kw}(a,v)=&\ln\left(a^{-1}\cdot v^{-1}\right)+\left(1-v\right)\psi(v)+\\&+\left(a^{-1}-1\right)\psi(1)+\left(v-a^{-1}\right)\psi(1+v).\end{aligned} \tag{11}$$

To display the uncertainty of the distribution, the entropic uncertainty interval is more convenient, which is a potentiation operation that is applied to the expression of information entropy. For determining the interval of entropic uncertainty, the formula is given as

$$\Delta_H=\exp(H). \tag{12}$$

Fig.1 shows the variability of the entropy uncertainty range for the Kumaraswamy distribution with the change in both the *a* power parameter and the v shape parameter. The maximum of the entropy uncertainty interval is $\vartheta$ and it is defined as the uniform distribution, which reduces the distribution when choosing *a* and *b* parameters are accordingly equal to zero and one.

Fig.2 illustrate the reduced error of entropy approximation when quantile functions (9) of the Kumaraswamy distribution are used depending on the distribution parameters. The given error was estimated using the formula

$$\delta_H=\frac{\exp(H_{Kw}(a,v))-\exp(H_Q(a,v))}{\vartheta}\cdot 100\%. \tag{13}$$

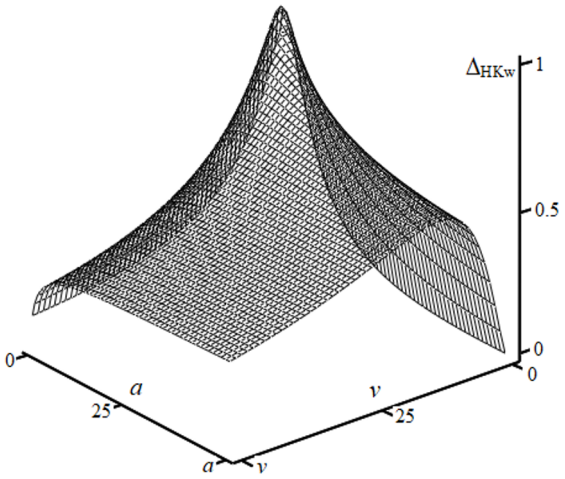


**Fig. 1**. Uncertainty of the Kumaraswamy distribution of the entropy uncertainty interval.

As can be seen from the examination of Fig.2, the approximation error in the most commonly used range of shape and power parameters did not exceed 1%. When constructing an approximation, the discreteness for a number of quantiles is chosen from the condition that the probability of values falling into the sampling intervals is 1%. At the same time, the number of n approximation intervals is equal to 100.

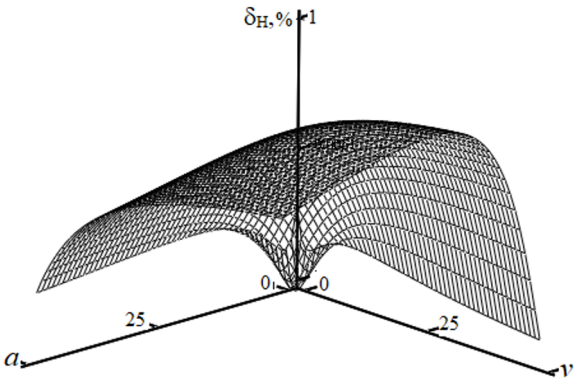


**Fig. 2**. The reduced approximation errors.

## 4 Conclusion

Thus, the proposed method for calculating Shannon's information entropy using quantile estimates of the distribution is a new effective tool for studying information uncertainty in complex systems, which makes it possible to calculate information entropies and information uncertainty intervals of random variables for which there are no explicit entropy formulas. An obvious area of applied research of the proposed method is the search and evaluation of distribution forms over arrays of experimental data and output random variables of complex systems.

## References

1. S. K. Niazi, *Handbook of Pharmaceutical Manufacturing Formulations: Compressed Solid Products*. (Informa Healthcare USA, New York, 2009)
2. S. A. Heyam, S. S. Rasha, M. A. E. Babiker, S. Raina, A Recent Progresses and Manufacturing Techniques in Pharmaceutical Powders and Granulation. International Journal of Pharmaceutical and Clinical Research, **11(1)**, 1-12 (2019)
3. E. I. Vedenin, S. V. Polovchenko, P. V. Charty, V. G. Shemanin, Particle Size Distribution Functions at Dust Separation Equipment's Various Operating Modes. Safety in Technosphere. **1**, 41-47 (2016). doi:10.12737/19022
4. A. A. Sherbakova, V. G. Polosin, B. V. *Chuvykin, Calculation of the Optical Path Optimal Length of the Flow Cell of Spektrometric System for the Research of Gasolin*es, in 20th International Conference of Young Specialists on Micro/Nano-technologies and Electron Devices, EDM, 407-412 July, 2019, Erlagol, Russia, (2019).
5. A. V. Korotaev, A. S. Malkov, D. A. Khalturina, A mathematical model of the growth of Earth's population, economics, technology, and education. The New in Synergetics: New Reality, New Problems, and New Generation. 148-186. (2007), Preprints of the Keldysh Institute of Applied Mathematics, **013**, 1-39 (2005)
6. D. D. Mishin, *Magnetic materials*. (Higher school, Moscow, 1991)
7. A. P. Trunev, Quantization of energy of electrons in a magnetic beta-spectrometer. Chaos *and Correlation*. **19** 1-22 (2010)
8. Yu. V. Knyazheva, Improving the efficiency of the mass service system of a trade enterprise through numerical statistical modeling. NSU Bulletin. Series: Social and Economic Sciences, **14(2),** 83-100 (2014)
9. T. L. Saati, *Elements of Queueing Theory*. (Soviet Radio, Moscow, 1965)
10. L. Ya. Saveliev, Distributions of the number of states in binary Markov stochastic models. Sib. J. Comput. Mathematics. RAS. Sib. Branch. **18(2),** 191–200 (2015)
11. A. Yu. Krasnov, *Statistical methods in engineering research* (ITMO University, St. Petersburg, 2022)
12. K. Shannon, *Works on information theory and cybernetics*. (Publishing House of Foreign Literature, Moscow, 1963)
13. I.P. Mysovskikh, *Interpolation cubature formulas*. (Main editorial office of physical and mathematical literature, Moscow, 1981)
14. J. B. McDonald, Y. J. Xu, A generalization of the beta distribution with applications. Journal of Econometrics, **66(1–2),** 133–152 (1995)
15. D. W. W. Ng, S. Z. Sim, M. C. Lee, The study of properties on generalized Beta distribution. IOP Conf. Series: Journal of Physics: Conf. Series, **1132,** 012080 (2019)
16. K. Pearson, *Tables of incomplete beta function*. (VTs FR USSR, Moscow, 1974)
17. V. Egorova, A. Gil, J. Segura, N. M. Temme, Computation of the regularized incomplete Beta function. Dolomites Research Notes on Approximation. **16**, 10–16 (2023)
18. A. Li, H. Qin, Some Transformation Properties of the Incomplete Beta Function and Its Partial Derivatives. IAENG International Journal of Applied Mathematics, **49(1)** (2019)
19. J. L. González-Santander, A Note on Some Reduction Formulas for the Incomplete Beta Function and the Lerch Transcendent. Mathematics, **9**, 1486 (**2021)**
20. M. A. Ozarslan, C. Ustaoglu, Extension of Incomplete Gamma, Beta and Hyper-geometric Functions. Progr. Fract. Differ. Appl. **5**(1), 21-35 (2019)