УДК 303.4 : 519.2

5.2.2. Математические, статистические и инструментальные методы экономики (физико-математические науки, экономические науки)

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЫБОРОК ИЗ БЕТА-РАСПРЕДЕЛЕНИЯ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор

РИНЦ SPIN-код: 4342-4994
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru

Эконометрика - важная составная часть математических, статистических и инструментальных методов экономики. Одна из основных задач эконометрики - оценивание параметров по выборочным данным. В настоящей статье рассматриваются методы статистического анализа выборок из бета-распределения. Семейство таких распределений - одно из параметрических семейств, которые обычно перечисляются в учебниках и справочниках по теории вероятностей. Бета-распределения используются при решении задач экономики и управления, в частности, в сетевом планировании. Однако методы оценивания параметров таких распределений по эмпирическим данным требуют дальнейшего развития. Точечные оценки параметров бета-распределения методом моментов приводятся в литературе (хотя и без вывода). С целью построения доверительных интервалов (т.е. интервальных оценок параметров) целесообразно уметь находить асимптотические распределения точечных оценок. Настоящая статья посвящена решению этой задачи. В ней впервые найдены распределения оценок параметров бета-распределения, полученные методом моментов. Они являются асимптотически нормальными. Предельные распределения являются асимптотически нормальными. Они получены методом линеаризации (путем выделения главного линейного члена при изучении приращения функции от выборочных моментов. Приведен численный пример точечного и интервального оценивания параметров бета-распределения. Рассмотрено применение бета-распределения в сетевом планировании при оценке продолжительности опытно-конструкторских работ. Исследования целесообразно продолжить. Целесообразно применить метод моментов для проверки статистических гипотез (в том числе для проверки согласия эмпирических данных с семейством бета-распределений), Необходимо проанализировать свойства оценок метода максимального правдоподобия и разработать алгоритм получения одношаговых оценок. Заслуживает внимания проблема практического

UDC 303.4 : 519.2

5.2.2. Mathematical, statistical and instrumental methods of economics (physical and mathematical sciences, economic sciences)

STATISTICAL ANALYSIS OF SAMPLES FROM THE BETA DISTRIBUTION

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
RSCI SPIN-code: 4342-4994
Bauman Moscow State Technical University, Moscow, Russia

Econometrics is an important part of the mathematical, statistical and instrumental methods of economics. One of the main tasks of econometrics is the estimation of parameters from sample data. This article discusses methods for statistical analysis of samples from the beta distribution. The family of such distributions is one of the parametric families that are usually listed in textbooks and reference books on probability theory. Beta distributions are used in solving problems of economics and management, in particular, in network planning. However, methods for estimating the parameters of such distributions from empirical data require further development. Point estimates of the parameters of the beta distribution by the method of moments are given in the literature (although without a derivation). In order to construct confidence intervals (that is, interval parameter estimates), it is advisable to be able to find asymptotic distributions of point estimates. This article is devoted to solving this problem. In it, for the first time, distributions of estimates of beta-distribution parameters obtained by the method of moments are found. They are asymptotically normal. The limit distributions are asymptotically normal. They are obtained by the linearization method (by extracting the main linear term when studying the increment of a function from sample moments. A numerical example of point and interval estimation of the parameters of the beta distribution is given. The use of beta distribution in network planning when estimating the duration of development work is considered. It is advisable to continue research. It is advisable to apply the method of moments to test statistical hypotheses (including checking the agreement of empirical data with a family of beta distributions), It is necessary to analyze the properties of the maximum likelihood method estimates and develop an algorithm for obtaining one-step estimates. when planning development work, more generally, when applying statistical methods of network planning and management, including the modern version of the PERT system and its varieties in the management of production military and research projects

использования бета-распределения, в том числе при планировании опытно-конструкторских работ, более общо, при применении статистических методов сетевого планирования и управления, в том числе современного варианта системы ПЕРТ и ее разновидностей при управлении производственными и научно-исследовательскими проектами

Ключевые слова: СТАТИСТИЧЕСКИЕ МЕТОДЫ, БЕТА-РАСПРЕДЕЛЕНИЕ, ОЦЕНИВАНИЕ ПАРАМЕТРОВ, МЕТОД МОМЕНТОВ, АСИМПТОТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ, ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ, СЕТЕВОЕ ПЛАНИРОВАНИЕ, ОПЫТНО-КОНСТРУКТОРСКИЕ РАБОТЫ

Keywords: STATISTICAL METHODS, BETA DISTRIBUTION, ESTIMATION OF PARAMETERS, METHODS OF MOMENTS, ASYMPTOTIC NORMAL DISTRIBUTIONS, CONFIDENCE INTERVALS, NETWORK PLANNING, DEVELOPMENT WORK

http://dx.doi.org/10.21515/1990-4665-187-017

Introduction

Econometrics is an important part of the mathematical, statistical and instrumental methods of economics. One of the main tasks of econometrics is the estimation of parameters from sample data. This article discusses methods for statistical analysis of samples from the beta distribution. The family of such distributions is one of the parametric families that are usually listed in textbooks and reference books on probability theory. Beta distributions are used in solving problems of economics and management, in particular, in network planning. However, methods for estimating the parameters of such distributions from empirical data require further development. Point estimates of the parameters of the beta distribution by the method of moments are given in the literature (although without a derivation). For the purpose of constructing confidence intervals (i.e. interval parameter estimates) it is expedient to be able to find asymptotic distributions of point estimates. This article is devoted to solving this problem.

Problems of Statistical Analysis of Samples from the Beta Distribution

Random value$X$has a beta distribution if it takes values between 0 and 1 and its probability density is:

$$f(x) = f(x; p, q) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, 0 < x < 1, \quad (1)$$

where the parameters p and q are positive and

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \Gamma(a) = \int_0^{+\infty} y^{a-1}e^{-y}\mathrm{dy}, \quad (2)$$

and f(x) = 0 for x outside the segment [0, 1]. Thus, the beta distribution is given by two parameters p and q and is determined using special mathematical functions - the gamma function $\Gamma(a)$ and beta features $B(p,q)$.

The beta distribution is one of the basic distributions of probability theory. Basic information about it is included in the reference books on probability theory and mathematical statistics [1 - 3]. However, methods for estimating parameters and testing hypotheses for samples from the beta distribution are not sufficiently represented in the literature.

Indeed, comparable with similar methods for the gamma distribution [4]. For it, point estimates by the method of moments and their asymptotic distributions were found, which made it possible to construct interval estimates for the parameters of the gamma distribution and methods for testing hypotheses about the values of the parameters. A similar research program has been carried out for asymptotically optimal one-step estimates of distribution parameters (which are used in modern applied mathematical statistics instead of maximum likelihood estimates). These results were included in our prepared GOST 11.011-83. Let us note the approach obtained later on the basis of statistics of interval data and the method for checking the agreement between empirical data and the gamma distribution.

Although studies on the statistical analysis of samples from the beta distribution are ongoing (see, for example, [5–7]), only point estimates of the method of moments have been obtained [2, 5]. This work is devoted to research in order to obtain scientific results that correspond to those previously proven for the gamma distribution.

Point estimates of the parameters of the beta distribution by the method of moments

Let*X*- a random variable with a beta distribution with density (1). The moments of the random variable X are expressed in terms of the parameters p and q as follows [1, p.145]:

$$M(X) = \frac{p}{p+q}, M(X^2) = \frac{p(p+1)}{(p+q)(p+q+1)}, \\ D(X) = \frac{pq}{(p+q)^2(p+q+1)}.(4)$$

Note that according to (4)

$$M(X^2) = \frac{p+1}{p+q+1} M(X).(5)$$

Using (4) - (5), we express the parameters of the random variable*X*through her moments. According to the formula for mathematical expectation

$$q = \frac{p(1-M(x))}{M(X)} \quad (6)$$

Substituting (6) into (5), we obtain

$$M(X^2) = \frac{p+1}{p+1+\frac{p(1-M(x))}{M(X)}} M(X).(7)$$

Because the

$$p + \frac{p(1-M(X))}{M(X)} = p + \frac{p}{M(X)} - p = \frac{p}{M(X)}. \quad (8)$$

then from (7) it follows that

$$M(X^2) = \frac{p+1}{1+\frac{p}{M(X)}} M(X). (9)$$

We multiply both sides of equality (9) by the denominator of the right side of (9), we get that

$$M(X^2)\left(1 + \frac{p}{M(X)}\right) = (p+1)M(X).(10)$$

Let's expand the brackets in (10):

$$M(X^2) + \frac{M(X^2)p}{M(X)} = \text{pM}(X) + M(X). (10)$$

Hence,

$$M(X^2) - M(X) = p\left(M(X) - \frac{M(X^2)}{M(X)}\right).(\text{eleven})$$

We get the expression of the parameter p through the initial moments:

$$p = \frac{M(X^2)-M(X)}{M(X)-\frac{M(X^2)}{M(X)}} = \frac{(M(X)-M(X^2))M(X)}{M(X^2)-(M(X))^2}.$$ (eleven)

According to (6)

$$q = \frac{(M(X)-M(X^2))(1-M(X))}{M(X^2)-(M(X))^2} = \frac{(M(X)-M(X^2))(1-M(X))}{D(X)}.(12)$$

Formulas (11) and (12) make it possible to construct estimates of the method of moments for the parameters of the beta distribution. To do this, it suffices to replace the initial moments on the right-hand sides of these formulas by their sample estimates.

Because the

$$M(X^2) = D(X) + (M(X)^2),(13)$$

then in formulas (11) - (12) it is possible to use dispersion instead of the second initial moment. We get:

$$p = M(X)\left\{\frac{M(X)(1-M(X))}{D(X)} - 1\right\},(14)$$

$$q = (1 - M(X))\left\{\frac{M(X)(1-M(X))}{D(X)} - 1\right\}.(15)$$

Let$X_1, X_2,..., X_n$-independent identically distributed random variables whose distribution function is a beta distribution with parameters p and q. As an estimate of the mathematical expectation M(X), we will use the sample arithmetic mean

$$\overline{X} = \frac{X_1+X_2+...+X_n}{n},$$ (16)

and as an estimate of the variance D(X) - the sample variance

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$ (17)

Substituting into (14) and (15) instead of the mathematical expectation and variance of their estimates (16) and (17), we obtain estimates for the method of parameter moments*p*and q:

$$p * \overline{X}\left\{\frac{\overline{X}(1-\overline{X})}{s^2} - 1\right\},(18)$$

$$q * (1 - \overline{X})\left\{\frac{\overline{X}(1-\overline{X})}{s^2} - 1\right\}.(19)$$

It is in this form that the estimates of the method of moments for the parameters of the beta distribution are given in [2, p.75]. Since we did not find the

derivation of formulas (18) - (19) in the available literature, we considered it useful to analyze in detail the derivation of these formulas.

The sample arithmetic mean and sample variance are consistent estimates of the mathematical expectation and variance (the proof uses the fact that the distribution of the considered random variable is concentrated on the segment [0; 1]). Since the right-hand sides of formulas (18) and (19) contain continuous functions of these consistent estimates, then*p** and q* are consistent estimates of the parameters p and q of the beta distribution, i.e.

$$\lim_{n\to\infty} p * p, \lim_{n\to\infty} q * q \quad (20)$$

(convergence in probability).

Linearization Method for Beta Distribution Parameter Estimations

We will find the asymptotic distributions of the estimates of the method of moments according to the algorithm given in [4, Section 6.1]. Let us describe the main idea of this method for the formulation under consideration. Consider the function

$$f(\overline{X}; s^2) - f(M(X); D(X)) \quad (21)$$

If the function f(x; y) is sufficiently smooth (for example, it has continuous second partial derivatives), then we can single out the main linear term

$$f(\overline{X}; s^2) - f(M(X); D(X)) = \frac{\partial f}{\partial x}\left(\overline{X} - M(X)\right) + \frac{\partial f}{\partial y}(s^2 - D(X)) + \varepsilon, \quad (22)$$

where the partial derivatives are taken at the point (x; y) = (M(X); D(X)), andε is the remainder term of a higher order of smallness (of order 1/n) compared to the first two terms in (22) (their order of decreasing is equal to$1/\sqrt{n}$).

It follows from (21) that the asymptotic distribution of the difference on the left side of this formula coincides with the mathematical expectation of the linear combination

$$Z = \frac{\partial f}{\partial x}\left(\overline{X} - M(X)\right) + \frac{\partial f}{\partial y}(s^2 - D(X)) \quad (23)$$

The distribution of the random variable is asymptotically normal (this follows from the multivariate central limit theorem) with mean 0 and variance

$$\sigma^2 = \left(\frac{\partial f}{\partial x}\right)^2 M\left(\overline{X} - M(X)\right)^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} M(\overline{X} - M(X))(s^2 - D(X)) + \\ + \left(\frac{\partial f}{\partial y}\right)^2 M(s^2 - D(X))^2 \quad (24)$$

In (24) the square of the partial derivative with respect to*x*is multiplied by D(X)/n. Let us find two other characteristics of the random variable X involved in (24).

Because the

$$M(\overline{X} - M(X))(s^2 - D(X)) = M(\overline{X} - M(X))s^2 - M(\overline{X} - M(X))D(X)$$

And

$$M\left(\overline{X} - M(X)\right) = 0,$$

That

$$M(\overline{X} - M(X))(s^2 - D(X)) = M(\overline{X} - M(X))s^2. \quad (25)$$

Based on (16) and (17), we conclude that

$$M(\overline{X} - M(X))s^2 = \frac{1}{n^2} M\left(\sum_{i=1}^{n}(X_i - M(X)) \sum_{j=1}^{n}(X_j - \overline{X})^2\right). \quad (26)$$

Because the

$$(X_j - \overline{X})^2 = \left((X_j - M(X)) + \left(M(X) - \overline{X}\right)\right)^2 = \\ (X_j - M(X))^2 + 2(X_j - M(X))\left(M(X) - \overline{X}\right) + \left(M(X) - \overline{X}\right)^2 \quad (27)$$

That

$$M(\overline{X} - M(X))s^2 = A + B + C, \quad (28)$$

where according to (27)

$$A = \frac{1}{n^2} M\left(\sum_{i=1}^{n}(X_i - M(X)) \sum_{j=1}^{n}(X_j - M(X))^2\right), \quad (29)$$

$$B = \frac{1}{n^2} M\left(\sum_{i=1}^{n}(X_i - M(X)) \sum_{j=1}^{n} 2(X_j - M(X))(M(X) - \overline{X})\right) \quad \text{(thirty)}$$

And

$$C = \frac{1}{n^2} M\left(\sum_{i=1}^{n}(X_i - M(X)) \sum_{j=1}^{n}(M(X) - \overline{X})^2\right) \quad (31)$$

Let's study each of the terms*A*, B and C separately. To simplify calculations, we introduce independent random variables Yi = Xi - M(X), i = 1, 2. ..., n. Then

$$X_i - M(X) = Y_i, M(Y_i) = 0, \overline{X} - M(X) = \overline{Y}. \quad (32)$$

Substituting equalities (32) into (29) - (31), we obtain that

$$A = \frac{1}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} Y_j^2\right), \quad (33)$$

$$B = \frac{2}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} Y_j(-\overline{Y})\right) \quad (34)$$

And

$$C = \frac{1}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} (-\overline{Y})^2\right) \quad (35)$$

The mathematical expectation of the sum is equal to the sum of the mathematical expectations, so

$$A = \frac{1}{n^2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} M(Y_i Y_j^2)\right). \quad (36)$$

At$i \neq j$ we have$M(Y_i Y_j^2) = M(Y_i) M(Y_j^2) = 0$ due to the independence of the random variables Yi, i = 1, 2. ..., n, and$M(Y_i Y_j^2) = M(Y_j^3)$ for i = j = 1, 2..., n.

Hence,

$$A = \frac{1}{n} M(Y^3). \quad (37)$$

For*IN*from (34) we have:

$$B = \frac{2}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} Y_j(-\overline{Y})\right) = \frac{2}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} Y_j \left(-\frac{1}{n}\sum_{k=1}^{n} Y_k\right)\right) = \frac{(-2)}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} M(Y_i Y_j Y_k). \quad (38)$$

If at least two lower summation indices are different, then by virtue of (32) and the independence of the random variables Yi, i = 1, 2. ..., n, the term on the right-hand side of (36) is equal to 0. Therefore, it is equal to the third initial moment Y only for n terms for which all three indices are equal to each other. Hence,

$$B = \frac{(-2)}{n^2} M(Y^3). \quad (39)$$

Let us proceed to the calculation of C from (35). We have:

$$C = \frac{1}{n^2} M\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} (-\overline{Y})^2\right) = \frac{1}{n^2} M\left(\sum_{i=1}^{n} Y_i n(-\overline{Y})^2\right) = \frac{1}{n} M \sum_{i=1}^{n} Y_i \left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)^2 = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{m=1}^{n} M\left(Y_i Y_k Y_m\right). \quad (40)$$

As in the analysis of the right side of (38), we state that in the case when at least two lower summation indices of the summand on the right side of (40) are different, then due to (32) and the independence of the random variables Yi, i =

1, 2. ..., n, this term is equal to 0. Therefore, the mathematical expectation is equal to the third initial moment Y only for n terms, for which all three indices are equal to each other. Hence,

$$C = \frac{1}{n^2} M(Y^3). \quad (41)$$

Let's summarize. From (25), (28), (37), (39) and (41) it follows that

$$M(\overline{X} - M(X))(s^2 - D(X)) = \frac{1}{n} M(Y)\left(1 - \frac{1}{n}\right). \quad (42)$$

In accordance with the remark after formula (22), the final factor in (42) can be replaced by 1.

The third characteristic of a random variable*X*participating in (24) can be found similarly. The result is known [8, p.419]:

$$M(s^2 - D(X))^2 = \frac{1}{n} D(X). \quad (43)$$

From (24), (42) and (43) it follows that

$$\sigma^2 = \frac{1}{n}\left(\left(\frac{\partial f}{\partial x}\right)^2 D(X) + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} M(X - M(X))^3 + \left(\frac{\partial f}{\partial y}\right)^2 D(X^2)\right). \quad (44)$$

Formula (44) is valid for any sufficiently smooth function of the sample mean and sample variance. We apply (44) to estimate the parameters of the beta distribution. According to (18), the estimate*p** parameter p has the form

$$p * f_1(\overline{X}, s^2), \quad (45)$$

Where

$$f_1(x, y) = x\left(\frac{x(1-x)}{y} - 1\right). \quad (46)$$

Let us find the partial derivatives of the function (46):

$$\frac{\partial f_1}{\partial x} = \frac{\partial}{\partial x}\left(\frac{x^2 - x^3}{y} - x\right) = \frac{2x - 3x^2}{y} - 1, \quad (47)$$

$$\frac{\partial f_1}{\partial y} = x^2(1-x)\frac{\partial}{\partial y}\left(\frac{1}{y}\right) = -\frac{x^2(1-x)}{y^2}. \quad (48)$$

Substituting (47) and (48) into (44), we obtain the asymptotic variance*D*(p*) estimates p* of parameter p of the beta distribution

$$\frac{1}{n}\left(\left(\frac{2M(X) - 3(M(X))^2}{D(X)} - 1\right)^2 D(X) - 2\left(\frac{2M(X) - 3(M(X))^2}{D(X)} - 1\right) \times \left| \right)\left(\left(\frac{(M(X))^2(1 - M(X))}{(D(X))^2}\right) M(X - M(X))^3 + \left| \right)\right.$$

(49)

Replacing the mathematical expectation and variance with their sample counterparts, we can use formula (49) to calculate the estimate D*(p*) of the variance D(p*) of the asymptotically normal estimate p* of the parameter p of the beta distribution.

According to (18), the estimate$q$* parameter q has the form

$$q * f_2(\overline{X}, s^2), \tag{50}$$

Where

$$f_2(x, y) = (1 - x)\left(\frac{x(1-x)}{y} - 1\right). \tag{51}$$

Let us find the partial derivatives of the function (51):

$$\frac{\partial f_2}{\partial x} = \frac{\partial}{\partial x}\left(\frac{x-2x^2+x^3}{y} - 1 + x\right) = \frac{1-4x+3x^2}{y} + 1, \tag{52}$$

$$\frac{\partial f_2}{\partial y} = x(1-x)^2 \frac{\partial}{\partial y}\left(\frac{1}{y}\right) = -\frac{x(1-x)^2}{y^2}. \tag{53}$$

Substituting (52) and (53) into (44), we obtain the asymptotic dispersion$D$(q*) estimates q* of the parameter q of the beta distribution

$$\frac{1}{n}\left(\left(\frac{1-4M(X)+3(M(X))^2}{D(X)} + 1\right)^2 D(X) - 2\left(\frac{1-4M(X)+3(M(X))^2}{D(X)} + 1\right) \times \left| \right)\left(\left(\frac{M(X)(1-M(X))^2}{(D(X))^2}\right) M(X - M(X))^3 + \left| \right)\right. \tag{54}$$

Replacing the mathematical expectation and variance with their sample analogs, we can use formula (54) to calculate the estimate D*(q*) of the variance D(q*) of the asymptotically normal estimate q* of the parameter q of the beta distribution.

Asymptotic Confidence Intervals for Beta Distribution Parameters Corresponding to Confidence Probabilityγ, are as follows: for the parameter p

$$(p * - C(\gamma)\sqrt{D * (p *)}; p * + C(\gamma)\sqrt{D * (p *)}), \tag{55}$$

where p* is calculated by formula (18), and D*(p*) - by formula (49) (with the replacement of the theoretical mathematical expectation and variance of a random variable X, which has a beta distribution, with the sample arithmetic mean and sample variance, respectively), and for the parameter q

$$(q * - C(\gamma)\sqrt{D * (q *)}; q * + C(\gamma)\sqrt{D * (q *)}), \tag{56}$$

where q* is calculated by formula (19), and D*(q*) - by formula (54) (with the replacement of the theoretical mathematical expectation and variance of a random variable X, which has a beta distribution, with the sample arithmetic mean and sample variance, respectively). In (55) and (56), as usual when constructing confidence intervals for asymptotically normal estimates,

$$\Phi(C(\gamma)) - \Phi(-C(\gamma)) = \gamma, \Phi(C(\gamma)) = \frac{1+\gamma}{2}, C(\gamma) = \Phi^{-1}\left(\frac{1+\gamma}{2}\right). \quad (57)$$

Asymptotic distributions of beta-distribution parameter estimates by the method of moments

Since all the moments in (49) are expressed in terms of the parameters of the beta distribution, then to calculate the estimates of the variances and confidence intervals, one can first find the estimates of the parameters using formulas (18) and (19), then calculate the estimates of the moments, and, finally, substitute them into (49). (54) - (56). In this case, formulas (4) are used for*D*(X) and

$$M(X - M(X))^3 = \frac{2pq(q-p)}{(p+q)^3(p+q+1)(p+q+2)} \quad (57)$$

[1, p.146]. Because the

$$D(X^2) = M[(X - M(X^2))^2] = M[X^4 - 2X^2M(X^2) + M(X^2)^2] = M(X^4) - M(X)^{2'} \quad (58)$$

then to calculate D(X2), along with (4), we need a formula for the fourth initial moment [1, p.145]:

$$M(X^4) = \frac{p(p+1)(p+2)(p+3)}{(p+q)(p+q+1)(p+q+2)(p+q+3)}. \quad (59)$$

Algorithm for Obtaining Point and Interval Estimates for Beta Distribution Parameters by Sample$X_1, X_2, \ldots, X_n$ consists of the following steps.

1. Calculate the sample arithmetic mean and sample variance using formulas (16) and (17), respectively.

2. Find grades*p** and q* method of moments of parameters p and q according to formulas (18) and (19), respectively,

3. Get estimates of the moments of a random variable*X*, which has a beta distribution, by replacing the values of the parameters p and q in the corresponding formulas with their estimates p* and q* of the method of moments. Namely, the estimate of the first initial moment M(X), the second initial moment M(X2) and the variance D(X) - according to the formula (4), the third central moment M(X - M(X))3 - according to the formula (57), the fourth initial moment - according to formula (59), the variance of the square of the random variable D(X2) - according to formula (58).

4. Calculate variance estimates*p** and q* according to formulas (49) and (54), respectively.

5. Find confidence intervals (i.e. obtain interval estimates) for the parameters*p*and q by formulas (55) and (56), respectively.

The third and fourth moments of the beta distribution could be estimated not with the help of parameter estimates, but directly by equating them with sample moments. However, since the values of the sample moments of the third and fourth order are unstable to outliers, calculation errors, and other deviations from the assumptions of the considered probabilistic-statistical model, it is advisable to apply the algorithm described above.

Beta distribution on a segment

To solve applied problems, a linear function of a random variable is often used*X*, which has a beta distribution:

$$Y = a + \mathrm{hX}, -\infty < a < +\infty, h > 0. \quad (60)$$

According to (1), the density of the random variable Y is positive for a < x < b, where b = a + h, and is equal to

$$f_{a,b}(x) = f(x; a, b, p, q) = \frac{1}{(b-a)^{p+q-1}B(p,q)}(x-a)^{p-1}(b+a-x)^{q-1}, a < x < b. \quad (61)$$

They say that the random variable*Y*has a beta distribution on the interval (a; b) with parameters p and q. In practice, it is assumed that the values a and b are given, and the parameters p and q must be estimated from statistical data -

from a sample$Y, Y_2,\ldots, Y_n$.It is natural to pass from this sample in accordance with (60) to the sample$X_1, X_2,\ldots, X_n$, Where

$$X_i = \frac{Y_i - a}{h} = \frac{Y_i - a}{b - a}, i = 1,2,\ldots, n. \quad (62)$$

The elements of this sample take values between 0 and 1 and have parameters p and q. The algorithm for obtaining estimates of these parameters is described above. At the first stage, it is necessary to obtain the sample arithmetic mean and sample variance for the sample (62). It's obvious that

$$\overline{X} = \frac{\overline{Y} - a}{h}, s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{(b-a)^2}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right). \quad (63)$$

The further steps of the algorithm for calculating the values of estimates of the parameters p and q of the beta distribution on the interval (a; b) do not require changes.

An example of point and interval estimation of beta distribution parameters

Let's carry out a point and interval estimation of the parameters of the beta distribution according to the data of Table 1. This table shows the values of the elements of the sample$Y, Y_2,\ldots, Y_n$ volume n = 50. Based on the analysis of a specific applied problem, the boundaries of the interval of variation of the considered random variable are established: a = 5, b = 135, then b - a = h = 130.

For the data given in Table 1, the sample arithmetic mean is$\overline{Y}$ = 57,88,and the sample variance is$\frac{1}{n}\sum_{i=1}^{50}\left(Y_i - \overline{Y}\right)^2$ = 663,00.

Table 1.

Measurement resultsduration of work

Number i	Yi value	Number i	Yi value	Number i	Yi value
1	9	18	47.5	35	63
2	17.5	19	48	36	64.5
3	21	20	50	37	65

4	26.5	21	51	38	67.5
5	27.5	22	53.5	39	68.5
6	31	23	55	40	70
7	32.5	24	56	41	72.5
8	34	25	56	42	77.5
9	36	26	56.5	43	81
10	36.5	27	57.5	44	82.5
eleven	39	28	58	45	90
12	40	29	59	46	96
13	41	thirty	59	47	101.5
14	42.5	31	60	48	117.5
15	43	32	61	49	127.5
16	45	33	61.5	50	130
17	46	34	62	-	-

After passing with the help of formulas (62) to a sample from the beta distribution with values in (0; 1), we have:

$$\overline{X} = \frac{57,88-5}{130} = 0,407, s^2 = \frac{663,00}{130^2} = \frac{663,00}{16900} = 0,0392.$$

Let us find estimates for the method of moments:

$$p * 0,407\left\{\frac{0,407\times0,593}{0,0392} - 1\right\} = 0,407 \times 5,16 = 2,10, \quad (64)$$

$$q * 0,593 \times 5,16 = 3,06. \quad (65)$$

In accordance with item 3 of the algorithm developed above, we find the estimates of the moments (indicated by asterisks) of the random variable*x*.From (4), (64) and (65) it follows that

$$M(X) * \frac{2,10}{2,10+3,06} = \frac{2,10}{5,16} = 0,407, M(X^2) * 0,407 \times \frac{3,10}{6,16} = 0,205, D(X) * \frac{2,10\times3,06}{5,16^2\times6,16} = \frac{6,43}{164,0} = 0,0392 \quad (66)$$

According to (57), (64) and (65), the estimate of the third central moment is as follows:

$$[M(X-M(X))^3]*\frac{2\times2,10\times3,06\times0,96}{5,16^3\times6,16\times7,16}=\frac{12,34}{6060}=0,00204. \quad (67)$$

In accordance with (59), (64), and (65), the estimate for the fourth initial moment is

$$[M(X^4)]*\frac{2,10\times3,10\times4,10\times5,10}{5,16\times6,16\times7,16\times8,16}=\frac{136,12}{1858,1}=0,07326. \quad (68)$$

Based on (58), (66), and (68), we conclude that the estimate for the variance of the squared random variable D(X2) is

$$D(X^2)*[M(X^4)]*-[M(X)*]^2=0,07326-0,205^2=0,031235. \quad (69)$$

Let us proceed to the calculation of estimates of variances*p** and q* according to formulas (49) and (54), respectively. Let's start by calculating the values of partial derivatives (47) and (48) used in formula (49):

$$\frac{\partial f_1}{\partial x}=\frac{2\times0,407-3\times0,407^2}{0,392}-1=\frac{0,814-0,497}{0,392}-1=0,809, \quad (70)$$

$$\frac{\partial f_1}{\partial y}=-\frac{0,407^2\times0,593}{0,0392^2}=-\frac{0,0982}{0,00154}=-63,77. \quad (71)$$

Instead of formula (49), it is easier to use its original form - formula (44), according to which, in accordance with (66), (67), (69) - (71) we have

$$D*(p*)=$$
$$=\frac{1}{50}\left((0,809)^2\times0,0392-2\times0,809\times63,77\times0,00204+63,77^2\times0,031235\right)=$$
$$\frac{1}{50}(0,0257-0,2105+127,0201)=\frac{126,835}{50}=2,537$$

Therefore, the standard deviation of the estimate of the parameter p* of the beta distribution is

$$\sqrt{D*(p*)}=\sqrt{2,537}=1,496. \quad (72)$$

Let's move on to estimating the variance of the estimate*q**. In accordance with (52) and (53) we have

$$\frac{\partial f_2}{\partial x}=\frac{1-4\times0,407+3\times0,407^2}{0,0392}+1=\frac{1-1,628+0,497}{0,0392}+1=$$
$$\frac{-0,131}{0,0392}+1=-3,342+1=-2,342, \quad (73)$$

$$\frac{\partial f_2}{\partial y}=-\frac{0,407\times0,593^2}{0,0392^2}=-\frac{0,143}{0,00154}=-92,86. \quad (74)$$

Instead of formula (54), it is easier to use its original form - formula (44), according to which, in accordance with (66), (67), (69), (74) that*D**(q*) is

$$\frac{1}{50}((-2,342)^2\times0,0392+2(-2,342)\times(-92,86)\times0,00204)+$$
$$+\frac{1}{50}((-92,86)^2\times0,031235)=\frac{1}{50}(0,215+0,887+269,3)=5,4075. \quad (75)$$

Therefore, the standard deviation of the estimate of the parameter q* of the beta distribution is

$$\sqrt{D * (q *)} = \sqrt{5{,}4075} = 2{,}325. \quad (76)$$

Find confidence intervals (i.e., obtain interval estimates) for the parameters$p$and q by formulas (55) and (56), respectively, using numerical values (64), (65).

Confidence interval for$p$with a confidence levelγhas the form

$$(2{,}10 - 1{,}496C(\gamma); 2{,}10 + 1{,}496C(\gamma)) \quad (77)$$

and for parameter q is:

$$(3{,}06 - 2{,}325C(\gamma); 3{,}06 - 2{,}325C(\gamma)). \quad (78)$$

If$C(\gamma) = 1$,which corresponds to the traditional notation$M \pm \sigma$, then according to (57)

$$\Phi(1) = \frac{1+\gamma}{2}, \gamma = 2\Phi(1) - 1 = 0{,}8413. \quad (79)$$

With such a confidence interval, the interval (77) turns into (0.604; 3.596), and (78) into (0.735; 5.385).

If we use the most common value of the confidence probability in socio-economic studies$\gamma = 0{,}95$, That$C(\gamma) = 1{,}96$, and with such a confidence interval (77) goes into ((-0.832); 5.032), and (78) - into ((-1.497); 7.617). Since the parameters of the beta distribution are positive, the left ends of the obtained intervals should be replaced by 0. Thus, p with a confidence probability$\gamma = 0{,}95$ the confidence interval for the p parameter is (0; 5.032), and for the q parameter is: (0; 7.617).

For the analyzed statistical data, the asymptotic confidence intervals calculated above are quite wide. With an increase in the sample size from the beta distribution, the confidence intervals will narrow, and the left boundaries of the intervals will become positive. The exit of the left boundaries of the intervals beyond the boundaries of the positive semiaxis is associated with the use of asymptotic (for$n \to \infty$) ratios.

The resulting confidence intervals can be used to test statistical hypotheses about the parameter values. Let the null hypothesis be:H_0: p = p0, and the alternative hypothesis H1 is its negation. Let the level of significance be givenα. The decision rule is formulated as follows: if the confidence interval for the parameter p corresponding to the confidence probability$\gamma = 1 - \alpha$, includes p0, then H0 is accepted: (i.e. the value p = p0 does not contradict the statistical data), otherwise the alternative hypothesis H1 is accepted (the value p = p0 is incompatible with the statistical data). Hypotheses about the values of the parameter q are tested in a similar way.

Based on the data in Table 1 and the above calculations, we conclude at the level of significance$\alpha = 0{,}05$that the value of the parameter p does not exceed 5.032, and the value of the parameter q does not exceed 7.617.

Beta distribution in estimating the duration of work

Let us consider an important applied problem, for the solution of which it is necessary to use point and interval estimates of the parameters of the beta distribution.

In the "Planning of development work" section of the tutorial [9, p. 170] it is said: "In network planning, in relation to forecasting the time of work execution, beta distribution is used" (with reference to [10]). For this use, the responsible agent calls the "minimum (optimistic) time*tmin*- the duration of work under the most favorable set of circumstances (this refers to the judgment, for example, this: "No matter how well everything goes, but faster than 20 days, we will not be able to complete this work")" and "maximum (pessimistic) time tmax is the duration of work under extremely unfavorable circumstances (in this case, for example, such an assessment is assumed: "No matter how unsuccessful everything is, we will do this work in any case in 30 days")" [9, p.170]. In the notation of this article, tmin = a and tmax = b. Further, the mathematical

expectation of a random variable Y - the duration of work - is proposed to be calculated by the formula

$$M(Y) = \frac{3a+2b}{5}, \quad (80)$$

and the variance of this random variable - according to the formula

$$D(Y) = \left(\frac{b-a}{5}\right)^2 \quad (81)$$

(see formulas (4.2.3) and (4.2.4) on p.171 of the textbook [9]). The same formulas are given in [11].

Let us find out for what values of the parameters of the beta distribution formulas (80) and (81) are valid. Let's move on from a random variable*Y*to a random variable X having a beta distribution with density (1):

$$X = \frac{Y-a}{b-a}. \quad (82)$$

Then, in accordance with (80)

$$M(X) = \frac{M(Y)-a}{b-a} = \frac{1}{b-a}\left(\frac{3a+2b}{5} - a\right) = \frac{2}{5}, \quad (83)$$

$$D(X) = \frac{1}{(b-a)^2} D(Y) = \frac{1}{25}. \quad (84)$$

In accordance with (4), the random variable X, which has a beta distribution with density (1), must satisfy the relations

$$M(X) = \frac{p}{p+q} = \frac{2}{5}, D(X) = \frac{\mathrm{pq}}{(p+q)^2(p+q+1)} = \frac{1}{25}. \quad (85)$$

We solve the system of equations (85). We have in accordance with (4):

$$M(X^2) = \frac{p(p+1)}{(p+q)(p+q+1)} = D(X) + (M(X))^2 = \frac{1}{25} + \frac{4}{25} = \frac{1}{5}. \quad (86)$$

From the first relation in (85) we obtain that

$$p + q = \frac{5}{2}p. \quad (87)$$

Substitute (87) into (86):

$$\frac{p(p+1)}{\frac{5}{2}p\left(\frac{5}{2}p+1\right)} = \frac{4p(p+1)}{5p(5p+2)} = \frac{4(p+1)}{5(5p+2)} = \frac{1}{5}. \quad (88)$$

Hence,

$$4(p+1) = 5p + 2, 4p + 4 = 5p + 2, p = 2. \quad (89)$$

From (87) and (89) it follows that

$$q = \frac{5}{2}p - p = \frac{3}{2}p = 3 \quad (90)$$

Why beta distribution parameters take values*p*= 2 and q = 3? In [9] there is no answer to this question.

It is known that basic research on the subject under consideration was carried out in the 1960s [12, 13]. Probabilistic model of D.I. Golenko [12] is based on the use of the beta distribution. Conducted by D.I. Golenko and his colleagues, numerous empirical studies have shown that the quantities*p*and q, averaged over a large number of network projects, are concentrated around the constant values p = 2 and q = 3. Since then, formulas (80) and (81) can be found in various publications and qualifying papers without any attempts to justify the possibility of their use .

In our opinion, the unjustified use of formulas (80) and (81) can be regarded as a violation of thebasic requirements for statistical methods of data analysis[14]. Is there any reason to believe that the distribution of development work time has not changed for more than half a century - since the 60s, and is also the same for all types of organizations and enterprises? In our opinion, there are no such grounds.

The use of the beta distribution seems quite natural. It is this parametric family that is concentrated on a finite interval (by this it stands out among all the parametric families of distributions of continuous random variables considered in probability theory and mathematical statistics) and, for different values of the parameters, covers various forms of probability distribution densities. However, point and interval estimates of distribution parameters must be calculated from statistical data corresponding to specific types of development work performed in a particular organization, at a particular enterprise. The calculation methods are detailed in this article.

Note, however, that the data given in Table. 1 are consistent with the values of the parameters adopted in [9, 12]*p*= 2 and q = 3, since these values are included in the corresponding confidence intervals. It can be expected that with a relatively small sample size (several tens of values), confidence intervals in

most real situations will include the values of the parameters p = 2 and q = 3, as a result, the use of these values and formulas (80) and (81) based on them will not lead to errors significant for the corresponding application area.

There are other approaches to the empirical estimation of the parameters of the beta distribution. One of them, given in [9, p.170-171], is based on the use, in addition to*tmin*= a and tmax = b, "the most probable time tnv - the duration of work under normal, usually occurring conditions for performing work." With this approach, instead of formulas (80) and (81), expressions for the mathematical expectation should be used

$$M(Y) = \frac{a+4t_{\text{нв}}+b}{6}, \quad (91)$$

and the variance of this random variable - according to the formula

$$D(Y) = \left(\frac{b-a}{6}\right)^2 \quad (92)$$

(see formulas (4.2.1) and (4.2.2) [9, p.171]). By tnv it is natural to understand the mode Mod(Y) of the considered random variable Y.

Let us find out what values of the parameters of the beta distribution correspond to formulas (91) and (92). As follows from (1) and (61), the modes of random variables and (see (60)) are related by the relation

$$Mod(\text{Y}) = \text{a} + (\text{b} - \text{a})\ \text{Mod}(\text{X}), \quad (93)$$

where the random variable X takes values from the interval (0; 1) and has a beta distribution (1) with parameters p and q. Because the

$$M(Y) = \frac{a+4t_{\text{нв}}+b}{6} = \frac{a+4a+4(b-a)\text{Mod}(X)+b}{6}, \quad (94)$$

That

$$M(X) = \frac{M(Y)-a}{b-a} = \frac{5a+4(b-a)\text{Mod}(X)+b-6a}{6(b-a)} = \frac{4\text{Mod}(X)+1}{6}.$$

As is known [1, 2]:

$$\text{Mod}(X) = \frac{p-1}{p+q-2} \quad (95)$$

(for p > 1, q > 1). Based on (93) and (94), we obtain a system of equations similar to (85):

$$M(X) = \frac{p}{p+q} = \frac{1}{6}\left(4\frac{p-1}{p+q-2}+1\right), D(X) = \frac{\text{pq}}{(p+q)^2(p+q+1)} = \frac{1}{36}.$$

Formulas (91) and (92) correspond to the solutions of this system of equations. We see that the parameter values $p = 2$, $q = 3$, obtained for the approach based on setting tmin and tmax, are not the solution of the system of equations corresponding to the second approach with three values of tmin, tmax and tnv. However, this was also clear from the fact that the variances for the two approaches are different. The question of choosing between approaches to apply in solving practical problems is open.

Conclusion

The development of statistical methods is analyzed in [15, part 1]. The cutting edge of scientific research in this area has moved from descriptive statistics (before 1900) to parametric mathematical statistics (1900-1933), then to non-parametric statistics (1933-1979) and statistics of non-numerical data (from 1979 to the present. In21st century most of the new research in the field of statistical methods refers specifically to the statistics of non-numerical data [16]. We are talking about the front line. In applied research, the methods of descriptive, parametric and nonparametric statistics are currently very widely used. University courses in the discipline "Probability Theory and Mathematical Statistics" are mainly devoted to parametric statistics, the main scientific results of which were obtained in the first half of the 20th century.

It is important to emphasize that actual unsolved problems remain in parametric statistics as well. They lie, it can be stated, "in the rear" of the advanced front of scientific research. One can name, for example, one-step parameter estimates that are more preferable than maximum likelihood estimates [4, Section 6.2; 17]. The system of rules for determining estimates and confidence limits for the parameters of the gamma distribution was first developed by us only in the first half of the 80s. It is partially reflected in [4] and is the main content of GOST 11.011-83 (unfortunately, now cancelled). And there is still no similar system for determining estimates and confidence limits

for the parameters of the beta distribution, although primary information about the beta distribution is available in reference books [1 - 3] and is included in detailed courses on the discipline "

In this article, we have begun to fill this gap. For the first time, the asymptotic distribution of estimates of beta-distribution parameters by the method of moments is obtained. The rules for determining estimates and confidence limits for the parameters of the beta distribution using the method of moments are given. Derivation of an algorithm for calculating point and interval estimates of beta-distribution parameters and its application for the analysis of specific statistical data are described in detail.

Research should be continued. At the next stages of the study, it is necessary to apply the method of moments to test statistical hypotheses (including checking the agreement between empirical data and the family of beta distributions, similar to how it was done for the gamma distribution in [4]). It is necessary to analyze the possibility of calculating the estimates of the maximum likelihood method and develop an algorithm for obtaining one-step estimates. The problem of the practical use of the beta distribution, including the planning of development work, more generally, the application of statistical methods of network planning and management, including the modern version of the PERT system and its varieties in the management of production and research projects, deserves attention.

References

1. Balakrishnan N., Nevzorov VB A primer of statistical distributions. - New Jersey: Wiley-Interscience, 2003. - 328 p.R.

2. Hastings N., Peacock J. Handbook of statistical distributions - M.: Statistics, 1980. - 95 p.

3. Probability and mathematical statistics: Encyclopedia / Ch. ed. Yu.V. Prokhorov. - M. : Bolshaya Ros. Encycl., 1999. - 910 p.

4. Orlov A.I. Applied statistical analysis. - M .: AI Pi Ar Media, 2022. - 812 p.

5. Kriventsov A.S., Ulyanov M.V. Interval estimation of beta-distribution parameters in determining the confidence complexity of algorithms. Izvestiya SFedU. Technical science. 2012. No. 7(132). pp. 210-220.

6. Milich V.N. Using the beta distribution in the tasks of analyzing the information content of features and improving the efficiency of the decision rule in recognizing texture images // Bulletin of the Udmurt University. Mathematics. Mechanics. Computer science. 2014. No. 3. S. 134-141.

7. Oleinikova S.A., Kirilov A.A. Numerical estimation of beta-distribution parameters // Bulletin of the Voronezh State Technical University. 2011. V. 7. No. 7. S. 209-212.

8.Borovkov A.A. Math statistics. -M.: Nauka, 1984. - 472 p.

9. Skvortsov Yu.V. Organizational and economic issues in graduation design. - M.: Higher school, 2006. - 399 p.

10. Razumov I.M., Belova L.D., Ipatov M.I., Proskuryakov A.V. Network diagrams in planning / 3rd ed., revised. and additional - M.: Higher School, 1981, - 168 p.

11. Organization and planning of machine-building production (production management): a textbook for universities / Nekrasov L. A., Postnikova E. S., Skvortsov Yu. V., Ukhanova T. V.; ed. Skvortsov Yu. V. - 3rd ed., revised. and additional - M. : Student, 2019. - 412 p.

12.Golenko D.I. Statistical methods of network planning and management. - M.: Science. Ch. ed. Phys.-Math. lit., 1968. - 400 p.

13.Kofman A., Debasey G.Network planning methods. Application of the PERT system and its varieties in the management of industrial and research projects: per. from fr. - M.: Progress, 1968. - 182 p.

14. Orlov AI Basic requirements for statistical methods of data analysis // Polythematic Online Scientific Journal of Kuban State Agrarian University. 2022. N 181. P. 316-343. – DOI 10.21515/1990-4665-181-026. – EDN OKGBOS.

15. Loiko V.I., Lutsenko E.V., Orlov A.I. High statistical technologies and system-cognitive modeling in ecology: monograph. - Krasnodar: KubGAU, 2019. - 258 p.

16.Orlov A.I. Development of mathematical research methods (2006- 2015) // Factory laboratory. material diagnostics. 2017. V.83. No. 1. Part 1. pp. 78-86.

17. Orlov A.I. Parameter estimation: one-step estimates are preferable to maximum likelihood estimates // Scientific journal of KubSAU.2015. No. 109. pp. 208 - 237.