

УДК 519.2; 316.4

DOI: 10.36871/2618-9976.2023.07-2.001

К ВОПРОСУ ОБ ОПРЕДЕЛЕНИИ ПЛОТНОСТИ ВЕРОЯТНОСТИ ТИПОВЫХ РАСПРЕДЕЛЕНИЙ ПО ЭКСПЕРИМЕНТАЛЬНЫМ ДАННЫМ

Светлана Васильевна Прокопчина¹

¹ Доктор технических наук, профессор, профессор кафедры "Системный анализ в экономике", Финансовый университет при Правительстве Российской Федерации, Москва, Россия, e-mail: svprokopchina@mail.ru

ИНФОРМАЦИЯ

Ключевые слова:

плотность распределения вероятностей
объем выборочных данных
моделирование законов распределений

АННОТАЦИЯ

В статье рассмотрены вопросы определения погрешностей при определении типа распределений. Тип распределений выражается в виде плотностей вероятностей унимодальных законов распределений системы К. Пирсона. Получены аналитические зависимости для определения необходимых объемов выборочных данных для обеспечения требуемой точности идентификации и аппроксимации этих законов распределений. Приведены конкретные аналитические зависимости и графики для определения таких законов распределений как простые законы (нормальное, равномерное, треугольное) распределения вероятностей, однопараметрических законов (гамма-распределения, логнормальное распределение, распределение Вейбулла), двухпараметрического бета-распределения.

ДЛЯ ЦИТИРОВАНИЯ: Прокопчина С.В. К вопросу об определении плотности вероятности типовых распределений по экспериментальным данным // Мягкие измерения и вычисления. 2023. № 7-2. Т. 68(2). С. 5–13; <https://doi.org/10.36871/2618-9976.2023.07-2.001>.

ON THE QUESTION OF DETERMINING THE PROBABILITY DENSITY OF TYPICAL DISTRIBUTIONS FROM EXPERIMENTAL DATA

Svetlana V. Prokopchina¹

¹ Doctor of Technical Sciences, Professor, Professor of the Department "System Analysis in Economics", Financial University under the Government of the Russian Federation, Moscow, Russia, e-mail: svprokopchina@mail.ru

ARTICLE INFO

Keywords:

Probability distribution density
Sample data volume
Modeling of distribution laws

ABSTRACT

The article discusses the issues of determining errors in determining the type of distributions. The type of distributions is expressed in the form of probability densities of unimodal distribution laws of the K. Pearson system. Analytical dependences are obtained to determine the necessary volumes of sample data to ensure the required accuracy of identification and approximation of these distribution laws. Specific analytical dependences and graphs are given for determining such distribution laws as simple laws (normal, uniform, triangular) probability distributions, one-parameter laws (gamma distributions, lognormal distribution, Weibull distribution), two-parameter beta distribution.

FOR CITATION: Prokopchina S.V. (2023) On the question of determining the probability density of typical distributions from experimental data. *Soft measurements and computing*, vol. 68(2), no. 7-2, pp. 5–13 (In Russ.); <https://doi.org/10.36871/2618-9976.2023.07-2.001>.

Введение

Как известно из теоретических положений теории вероятности и математической статистики, закон распределения является самой полной характеристикой случайных величин или случайных процессов. Самой удобной формой представления законов распределений является выражение их через аналитические зависимости, так называемых, типовых распределений, которые входят в известную систему кривых К. Пирсона [1].

Распределения этой системы делятся на три группы:

- простые распределения (нормальное, равномерное, треугольное, Рэлея, экспоненциальное);
- однопараметрические распределения (гамма-распределение, логнормальное, Вейбулла);
- двухпараметрические распределения (бета-распределение, U-образное распределение Джонсона, J-образное распределение Джонсона).

Сложными методологическими и практическими вопросами являются вопросы определения этих аналитических зависимостей по экспериментальным данным статистических выборок. Статья посвящена решению этих вопросов.

1. Постановка задачи аппроксимации ПВ типовыми распределениями

При определении значений ПВ, СВ, СЭСП методом дискретных выборок возникает два вида погрешностей [4–6]:

- погрешность смещения оценки ПВ относительно ее теоретического значения на уровне анализа x ;
- дисперсия оценки плотности вероятности.

Появление погрешности смещения вызвано конечной шириной интервала значений СВ или СЭСП, соответствующему ряду гистограммы. Этот интервал иначе называют дифференциальным коридором. В общем случае при конечной ширине дифференциального коридора даже при бесконечно большом объеме выборочных данных, используемых для получения оценки плотности вероятности, ордината, соответствующая ступени гистограммы, не равна ординате кривой ПВ, соответствующей середине интервала значений. В соответствии с этим возникает смещение ординаты ПВ на уровне x , равном середине соответствующего дифференциального коридора. Величина погрешности смещения при этом может быть записана следующим образом [4, 5]:

$$G = \Delta f(x) = \frac{1}{\Delta x} \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} f(x) dx - f(x). \quad (1)$$

Относительная погрешность смещения для каждого уровня анализа может быть вычислена по формуле:

$$\gamma' = \frac{G}{f(x)}. \quad (2)$$

При разложении выражения (2) в ряд Тейлора в окрестности точки x и учете только первых трех членов этого ряда, формула для определения относительной погрешности смещения приводится к виду [4, 5]:

$$\gamma' = \frac{(\Delta x)^2}{24} \cdot \frac{|f''(x)|}{f(x)}. \quad (3)$$

Из (1), (3) видно, что величина абсолютной и относительной погрешностей смещения зависит от уровня анализа x . Поэтому этими формулами пользоваться при выборе ширины дифференциального коридора Δx неудобно. Если величину погрешности смещения рассматривать в долях максимального значения ПА, которое соответствует уровню анализа x_m , равному моде распределения, можно записать:

$$\gamma = \frac{\Delta x^2}{24} \frac{|f''(x)|}{f(x_m)}. \quad (4)$$

Тогда, если при заданных допустимых величинах G_d и γ_d выбирать Δx для уровня, соответствующего максимальному значению второй производной функции ПВ $f''_{max}(x)$, то, очевидно, смещение оценки ПВ в остальном диапазоне изменения x будет меньше смещения соответствующего G_d и γ_d . следовательно, ширину дифференциального коридора Δx можно вычислять по формуле:

$$\Delta x = \sqrt{\frac{f(x_m) \cdot \gamma_d \cdot 24}{|f''_{max}(x)|}}. \quad (5)$$

Должны быть исследованы погрешности оценивания типовых распределений, входящих в САР, для определения уровня анализа x , соответствующего максимальной величине второй производной ПВ, можно воспользоваться результатами исследования и переписать (5) в виде:

$$\Delta x = \sqrt{\frac{f(x_m) \cdot \gamma_d \cdot 24}{K_f |f''(x_m)|}}, \quad (6)$$

где K_f – коэффициент пропорциональности между максимальным значением второй производной и значением второй производной в точке моды.

Конкретные значения коэффициента K_f для сложных распределений САР с определенными значениями параметров формы можно получить, используя графики [рис. 1](#), приведенные в работе [\[3\]](#).

Дисперсия оценки ПВ обусловлена конечностью объема выборочных данных, используемых для оценивания. Ее относительная величина для каждого уровня анализа может быть записана следующим образом в общем виде [\[4, 5, 7\]](#):

$$\delta^2 = \frac{1}{N f(x) \Delta x} - \frac{1}{N}. \quad (7)$$

Для больших объемов выборки для максимального значения функции ПВ:

$$\delta^2 = \frac{1}{N f(x_m) \Delta x}. \quad (8)$$

Как видно из сравнения (4) и (8), погрешность смещения и дисперсии противоположным образом зависят от величины дифференциального коридора. Поэтому умень-

шение погрешности смещения при оценивании ПВ путем уменьшения величины Δx приводит к увеличению погрешности за счет дисперсии оценки ПВ [4, 5].

Задаваясь допустимой величиной относительной дисперсии δ_d^2 в соответствии с (8), получаем формулу для вычисления объема выборочных данных, при котором для всех уровней анализа x величина дисперсии оценки ПВ $D_{ПВ}$ будет меньше или равной $D_{ПВд}$ допустимой:

$$N = \frac{1}{f(x_m)\Delta x\delta_d^2}. \quad (9)$$

Количество разрядов гистограммы, соответствующее определенному значению Δx , может быть вычислено по следующей формуле [4, 5]:

$$n = ENT \left[\frac{x_{max} - x_{min}}{\Delta x} \right] + 1, \quad (10)$$

где $ENT[\cdot]$ – целая часть числа;

x_{max}, x_{min} – максимальное и минимальное значение x , связанные формулой:

$$F(x_{max}) - F(x_{min}) = 0,997. \quad (10)$$

Пользуясь формулами (6), (8), (10), определим теперь величину дифференциального коридора, число разрядов гистограммы в зависимости от требуемой погрешности смещения оценки ПВ и величину относительной дисперсии оценки ПВ при различном числе дифференциальных коридоров (для $N = 1000$) при оценивании ПВ типовых аппроксимирующих распределений.

Подставляя в (6) выражения, соответствующие аналитической записи функции ПВ аппроксимирующих типовых распределений, и выражения их вторых производных в точке моды, окончательно получаем вид формул для вычисления необходимой величины дифференциального коридора, чтобы погрешность смещения оценки ПВ этих распределений была меньше или равной допустимой. Подставляя затем найденные выражения в (10), получаем формулы для определения n в зависимости от γ_d :

$$\Delta x = \sqrt{\gamma_d \cdot 24 \cdot \delta_x^2}, \quad (11)$$

где δ_x — среднеквадратичное отклонение случайной величины x .

$$n = ENT \left[\frac{6}{\sqrt{\gamma_d \cdot 24}} \right] + 1. \quad (11)$$

Формула для определения величины относительной дисперсии оценки ПВ в соответствии с (8) при уровне анализа, равном моде распределения, и в зависимости от числа дифференциальных коридоров n :

$$\delta^2 = \frac{0,417 \cdot n}{N}. \quad (12)$$

– для равномерного распределения:
погрешность смещения оценки ПВ равна нулю

$$\delta^2 = \frac{n}{N}. \quad (13)$$

- для распределения Симпсона:
погрешность смещения равна нулю

$$\delta^2 = \frac{n}{2N}. \quad (14)$$

- для распределения Рэлея:
коэффициент пропорциональности $K_f=1,2$

$$\Delta x = \sqrt{\gamma_d \cdot 22,7 \cdot \sigma_x^2}; \quad (15)$$

$$n = ENT \left[\frac{6}{\sqrt{\gamma_d \cdot 22,7}} \right] + 1; \quad (15)$$

$$\delta^2 = \frac{0,412 \cdot n}{N}. \quad (16)$$

- для экспоненциального распределения:
коэффициент пропорциональности $K_f=1$

$$\Delta x = \sqrt{\gamma_d \cdot 24 \cdot \sigma_x^2}; \quad (17)$$

$$n = ENT \left[\frac{6}{\sqrt{\gamma_d \cdot 22,7}} \right] + 1; \quad (17)$$

$$\delta^2 = \frac{0,167 \cdot n}{N}. \quad (18)$$

2. Оценка аналитических решений для простых распределений

Проводя сравнение выражений, соответствующих величинам дифференциальных коридоров, для различных простых аппроксимирующих распределений можно отметить, что при одинаковой дисперсии значений СВ или СЭСП минимальный дифференциальный коридор (максимальное число уровней анализа) для достижения одинаковой величины γ_d , необходим для распределения Рэлея. Однако в общем количество разрядов, необходимое для сохранения величины погрешности смещения оценки ПВ в допустимых пределах для всех простых распределений примерно одинаковое, (например, для $\gamma_d=0,1$, $n=5$). Относительная δ_d^2 дисперсия оценок ПВ при $N=1000$ примерно на два порядка меньше, чем относительная погрешность смещения для каждого из простых распределений с $\gamma_d \neq 0$. С максимальной относительной погрешностью за счет дисперсии оценки ПВ при фиксированном объеме выборочных данных N оцениваются значения ПВ равновероятного распределения, обладающего при прочих равных условиях минимальной погрешностью смещения среди всех простых эталонных аппроксимирующих распределений.

Таким образом, на основании проведенного анализа можно заключить, что для оценивания ПВ простых типовых аппроксимирующих распределений, входящих в САР, достаточно 5 разрядов гистограммы при $\gamma_d \geq 0,1$, что соответствует трем двоичным разрядам ПНК. Причем, общая погрешность оценивания при числе разрядов гистограммы, меньшем 6 и $N \geq 4000$ определяется погрешность смещения.

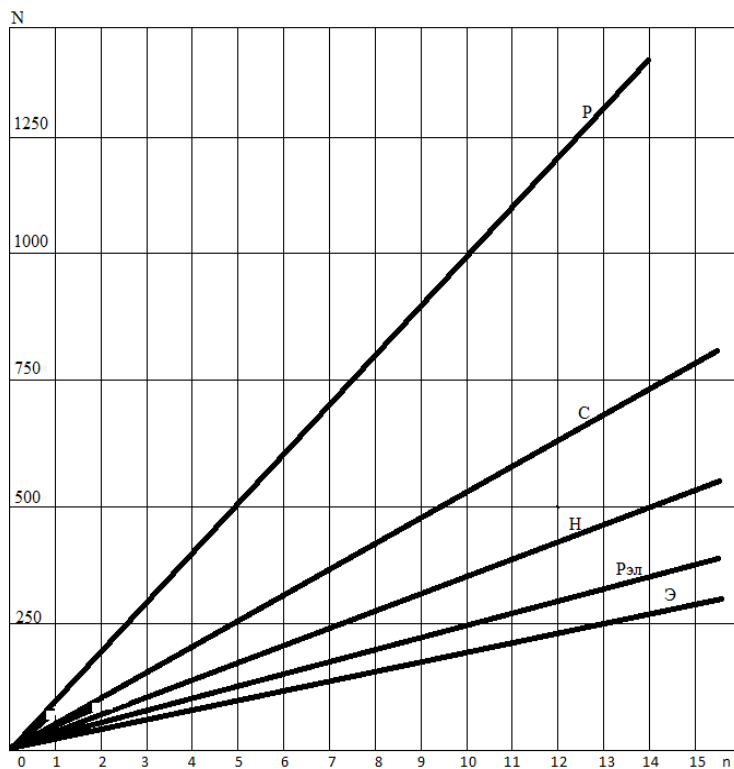
Производя преобразования, аналогичные случаю с простыми типовыми распределениями, получены следующие формулы для вычисления значений Δx , n , δ^2 для сложных типовых аппроксимирующих распределений:

– для гамма-распределения:

$$\Delta x = \sqrt{\frac{24 \cdot \gamma_d \cdot (K - 1)}{K_f}} * l; (K > 1) \quad (19)$$

$$n = ENT \left[7 \sqrt{\frac{K \cdot K_f}{24 \cdot \gamma_d \cdot (K - 1)}} \right] + 1; (K > 1) \quad (19)$$

$$\delta^2 = \frac{n \cdot \Gamma(K)}{N \cdot 7\sqrt{K} \cdot (K - 1)^{K-1} l^{-(K-1)}}; (K > 1) \quad (20)$$



Н – нормальное распределение.
Р – равномерное распределение.
С – распределение Симпсона.
Рэл – распределение Рэля.
Э – экспоненциальное распределение.

Рис. 1. Зависимость объема выборки от количества разрядов гистограммы при $\delta^2_d = 0.01$

3. Оценка аналитических решений параметрических распределений

Аналитические выражения для параметров алгоритмов, аналогичные тем, что приведены выше для простых распределений в этом разделе приводятся для сложных распределений, обладающих параметрами формы.

Они имеют следующий вид:

– для распределения Вейбулла:

$$\Delta x = \left[\sqrt{\frac{24 \cdot \gamma_d \cdot (K - 1)}{K_f \cdot K}} \right] \left(\frac{K}{K - 1} \right)^{1/K} (K - 1) * l; (K > 1) \quad (21)$$

$$n = ENT \left\{ 6 \left[\left(\frac{K}{K-1} \right)^{1/K} (K-1) \sqrt{\frac{\left\{ \Gamma \left(\frac{2}{K} + 1 \right) - \left[\Gamma \left(\frac{1}{K} + 1 \right) \right]^2 \right\} \cdot K \cdot K_f}{24 \cdot \gamma_d \cdot (K-1)}} \right] \right\} + 1; (K > 1) \quad (21)$$

$$\delta^2 = \frac{n \cdot \exp \left(\frac{K-1}{K} \right)}{N \cdot 6 \sqrt{\left\{ \Gamma \left(\frac{2}{K} + 1 \right) - \left[\Gamma \left(\frac{1}{K} + 1 \right) \right]^2 \right\} \cdot K \left(\frac{K-1}{K} \right)^{K-1/K}}}; (K > 1) \quad (22)$$

– для логарифмически-нормального распределения:

$$\Delta x = \frac{K}{l^{K^2}} \sqrt{\frac{24 \cdot \gamma_d}{K_f}}; \quad (23)$$

$$n = ENT \left[\frac{7 \cdot l^{K^2}}{K} \sqrt{\frac{l^{K^2} (l^{K^2} - 1) \cdot K_f}{24 \cdot \gamma_d}} \right] + 1; \quad (23')$$

$$\delta^2 = \frac{n \cdot K}{N \cdot 6 \cdot l^{K^2}} \sqrt{\frac{2\pi}{(l^{K^2} - 1)}}; \quad (24)$$

– для бета-распределения:

$$\Delta x = \sqrt{\frac{24 \cdot \gamma_d \cdot (K_1 - 1)(K_2 - 1)}{(K_1 + K_2 - 2)^3 K_f}} \cdot \frac{1}{M_2 - M_1}; \quad (25)$$

$$n = ENT \left[\sqrt{\frac{(K_1 + K_2 - 2) \cdot 3 \cdot K_f}{24 \cdot \gamma_d \cdot (K_1 - 1)(K_2 - 1)}} \right] + 1; \quad (25')$$

– для приближенных вычислений:

$$\delta^2 = \frac{n \cdot \Gamma(K_1) \Gamma(K_2) (K_1 + K_2 - 2)^{(K_1 + K_2 - 2)}}{N \cdot \Gamma(K_1 + K_2) (K_1 - 1)^{K_1 - 1} (K_2 - 1)^{(K_2 - 1)}}. \quad (26)$$

Результаты вычислений значений n в зависимости от γ_d и δ^2 от n и в зависимости от значений параметров формы представлены в [3].

Графики рис. 2 иллюстрируют зависимость объема выборки N от значения параметров формы K при $n=10$, $\delta_d^2=0,01$. При других значениях n и δ_d^2 , должны быть использованы пересчетные формулы (см. рис. 2). Анализ формул (19) – (26) и результатов перечисленных таблиц свидетельствуют о том, что погрешности оценивания ПВ для сложных типовых аппроксимирующих распределений зависят от значений параметров формы распределений. Кроме того, для всех этих распределений с увеличением симметричности кривой, с приближением ее формы к форме нормального распределения погрешности сме-

щения оценки ПВ уменьшаются, погрешности за счет дисперсии оценки увеличиваются и значения их приближаются к соответствующим значениям нормального распределения. Относительная величина погрешности за счет дисперсии оценки при $N \geq 1000$ и $n \leq 30$ много меньше, чем относительная величина погрешности за счет смещения оценки ПВ (примерно на 1–2 порядка). Поэтому, можно сделать вывод: так как основной вклад в общий уровень ошибок оценивания ПВ сложных типовых распределений при $N \geq 4000$ и $n < 5$ вносят ошибки, вызванные смещением оценки ПВ, то на величину последней можно ориентироваться при определении этих в случаях погрешности оценивания ПВ. В целом, для оценивания ПВ сложных распределений в диапазоне изменения параметров формы с $\gamma_d \geq 0,1$ достаточно 14 разрядов гистограммы.

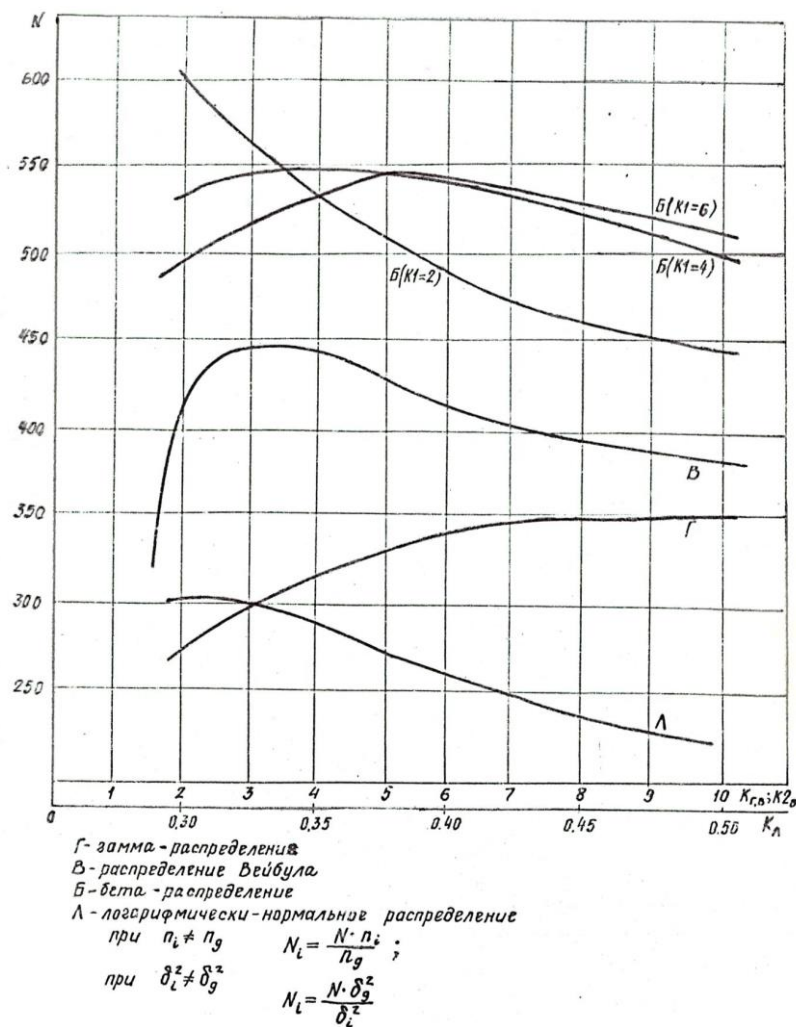


Рис. 2. Зависимость объема выборки от параметра формы сложных аппроксимирующих распределений при $n_g = 10$, $\delta_g^2 = 0,01$

При выбранном методе оценивания ПВ требуемое количество разрядов гистограммы определяет необходимое количество разрядов преобразователя напряжения, которым первоначально представлена СВ или СЭСП, в цифровой код для обеспечения заданной точности оценивания ПВ, а вместе с этим и, точности проведения этапа проверки адекватности гипотез о виде ПВ. Ориентируясь на максимально необходимое число разрядов гистограммы для соблюдения заданной точности оценивания ПВ для системы аппроксимирующих распределений в установленном диапазоне изменении параметров формы, можно сделать вывод, что наибольшее число разрядов ПНК, включая и знаковый, необхо-

димое для обеспечения точности оценки ПВ с $\gamma_d \geq 0,1$, $G_d \geq 0,1f(x_m)$ равно 5. Выбор числа разрядов гистограммы для оценивания определенного распределения с требуемой точностью может быть произведен с использованием соответствующих формул в зависимости от значений γ_d , δ_d^2 и параметров формы для сложных распределений. Таким образом, выбирается необходимое число разрядов гистограммы в соответствии с заданной точностью оценивания (заданной допустимой величиной относительной ошибки смещения оценки) могут быть также рекомендованы для использования при выборе необходимого количества разрядов гистограммы для оценивания ПВ указанных типовых распределений или проверки их по критерию согласия χ^2 .

Заключение

В статье определены аналитические зависимости для погрешностей при определении типа распределений. Тип распределений выражается в виде плотностей вероятностей унимодальных законов распределений системы К. Пирсона.

Полученные аналитические зависимости могут применяться при вероятностно-статистической обработке экспериментальных данных, а также в эконометрических задачах для определения необходимых объемов выборочных данных и обеспечения требуемой точности идентификации и аппроксимации этих законов распределений. Приведены конкретные аналитические зависимости и графики для определения таких законов распределений как простые законы (нормальное, равномерное, треугольное) распределения вероятностей, однопараметрических законов (гамма-распределения, логнормальное распределение, распределение Вейбулла), двухпараметрического бета-распределения.

Список источников

- [1] Кендалл М.Дж., Стьюарт А. Теория распределений. М.: Наука, 1966. 587 с.
- [2] Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1974. 899 с.
- [3] Прокопчина С.В. Байесовские интеллектуальные технологии в задачах моделирования законов распределения в условиях неопределенности. М.: Издательский дом "НАУЧНАЯ БИБЛИОТЕКА", 2020. 292 с.
- [4] Корн Г.А. Справочник по математике для научных работников и инженеров. М.: Наука, 1974. 831 с.
- [5] Крамер Г. Математические методы статистики. М.: Мир. 1975. 647 с.
- [6] Бендат Дж., Пирсон А. Измерение и анализ случайных процессов. М.: Мир, 1974. 463 с.

Referances

- [1] Kendall M.J., Stewart A. Theory of distributions. M.: Nauka, 1966. 587 p.
- [2] Kendall M.J., Stewart A. Statistical conclusions and connections. M.: Nauka, 1974. 899 p.
- [3] Prokopchina S.V. Bayesian intelligent technologies in problems of modeling distribution laws under uncertainty. M.: Publishing House "SCIENTIFIC LIBRARY", 2020. 292 p.
- [4] Korn G.A. Handbook of Mathematics for scientists and engineers. M.: Nauka, 1974. 831 p.
- [5] Kramer G. Mathematical methods of statistics. M.: Mir, 1975. 647 p.
- [6] Bendat J., Pearson A. Measurement and analysis of random processes. M.: Mir, 1974, 463 p.

Статья поступила в редакцию 17.07.2023; одобрена после рецензирования 24.07.2023; принята к публикации 24.07.2023.

The article was submitted 17.07.2023; approved after reviewing 24.07.2023; accepted for publication 24.07.2023.
