УДК 004.89 : 519.21

В.В. Нешитой

Вычисление закона распределения случайной величины

Дается классификация непрерывных случайных величин, строятся три системы непрерывных распределений, излагается суть устойчивого метода вычисления закона распределения (в том числе рангового) и оценок параметров, а также универсального метода моментов. Обосновываются универсальные законы рассеяния и старения публикаций.

Ключевые слова: *непрерывные случайные величины, непрерывные распределения, вычисление закона распределения и оценок параметров, рассеяние и старение публикаций*

ВВЕДЕНИЕ

В настоящее время при отыскании наилучшего аппроксимирующего распределения для описания статистического вариационного ряда используется метод выдвижения гипотез. При этом вид теоретического распределения устанавливается по форме гистограммы, что не дает однозначного результата. Например, для аппроксимации симметричной гистограммы чаще других используется нормальный закон из всего разнообразия симметричных распределений. В этом случае по выборочным значениям случайной величины вычисляются среднее их значение и среднее квадратическое отклонение. Далее эти значения принимаются в качестве оценок параметров теоретического закона (того же нормального) и осуществляется проверка степени близости теоретического и статистического распределений по критериям согласия (например, «хи-квадрат» Пирсона) при заданном уровне значимости α – обычно в пределах от 0.01 до 0.1. Величина α представляет собой вероятность отклонения верной гипотезы. Вводится также мощность критерия $1-\beta$, где β – вероятность принятия неверной гипотезы. Однако, несмотря на такие предосторожности, некоторые неверные гипотезы могут быть приняты. В итоге задача нахождения закона распределения может оказаться нерешенной даже при выдвижении нескольких гипотез подряд.

Та же ситуация возникает при аппроксимации статистических ранговых распределений, которые широко применяются в информатике, технике, математической лингвистике, библиотечной деятельности, социологии и других областях знания. Для аппроксимации таких распределений многие авторы используют законы Дж. Ципфа, С. Бредфорда (закон рассеяния публикаций), а также более общие модели – логарифмически нормальный закон и др. Метод выдвижения гипотез в этих случаях также не дает однозначных результатов. Это происходит по нескольким причинам:

- набор используемых распределений крайне ограничен;
- отсутствует классификация случайных величин;
- рассматриваются выборки недостаточного объема (особенно это касается ранговых распределений);
- попытки описать неоднородные выборки (например, ранговые распределения слов) одним теоретическим распределением;
- неудачно выбранная форма представления ранговых распределений (в системе координат «логарифм ранга – логарифм частоты»), которая несет слишком мало информации о статистическом распределении и не имеет вероятностного смысла;
- отсутствие общего подхода, т.е. предпринимаются попытки решить одну из частных задач без предварительного решения общей задачи.

В такой ситуации любой метод оценивания будет неэффективным.

Несмотря на недостатки метода выдвижения гипотез, он все еще используется в научных исследованиях и преподается в университетах. Использование этого метода обосновывается тем, что по статистическому ряду якобы невозможно установить вид теоретического распределения. Это глубокое заблуждение, поскольку любая случайная величина, если она представляет собой статистически однородную совокупность большого числа значений, несет в себе достаточное количество информации, позволяющее отличить ее от другой случайной величины. И задача

исследователя заключается в том, чтобы эту информацию извлечь из выборочных данных путем нахождения наилучшего аппроксимирующего распределения. Если же пытаться найти его путем перебора небольшого числа теоретических распределений (как правило, не более 10–15), которые обычно заложены в известные приложения, то действительно будет слишком мала вероятность правильного решения этой задачи.

Поскольку закон распределения является наиболее полной характеристикой случайной величины, то наиважнейшей задачей при обработке статистических данных является его **вычисление**, так как закон распределения позволяет осуществлять различные исследования, в том числе давать прогноз и решать множество практических задач. Приложения, которые не в состоянии решать эту задачу, перекладывают всю ответственность по отысканию закона распределения на пользователя, не предоставляя ему практически никаких гарантий точного решения этой задачи. Использование таких приложений в системах менеджмента качества может нанести огромный материальный и моральный ущерб предприятию, и в целом – экономике государства, но разработчики этих приложений не понесут никакой ответственности.

В литературе по теории вероятностей и математической статистике приводится множество методов оценивания параметров, даже по нескольку для каждого отдельного распределения (все это частные задачи), но нет надежного **метода вычисления закона распределения** по статистическим данным (т.е. не решена общая задача). А без такого метода практически бесполезно вычислять оценки параметров наугад выбранного аппроксимирующего распределения. Таким образом, **любой метод оценивания параметров должен в первую очередь обеспечить вычисление закона распределения**.

Естественно, что к разработке различных методов оценивания параметров принуждают различные факторы, например, неполнота статистических данных, разная форма их представления, но и в этом случае необходимо в первую очередь вычислить закон распределения.

В настоящей статье дается обоснование и построение методов вычисления закона распределения и оценок параметров по статистическому ряду без выдвижения гипотез и проверки каждой из них по критериям согласия, строятся три системы непрерывных четырехпараметрических распределений, излагаются два метода вычисления закона распределения и оценок параметров, разработанные автором, – общий устойчивый метод, который по точности не уступает методу наибольшего правдоподобия, и универсальный метод моментов. В обоих случаях для установления типа аппроксимирующего распределения достаточно решить систему двух уравнений с двумя неизвестными параметрами формы (k, u). Два других параметра (α, β) вычисляются по простым формулам при известных значениях двух параметров формы, которые найдены на первом этапе.

Из вышесказанного следует, что закон распределения должен и может быть вычислен по статистическому распределению, причем без выдвижения гипотез, но с использованием свойств случайной величины. Для решения этой важнейшей задачи необходимо иметь общий метод, включающий как минимум три составные части [1]:

- классификацию случайных величин;
- универсальные (обобщенные) законы распределения, а лучше – системы распределений, включающие как частные случаи множество распределений своего класса;
- хотя бы один **метод вычисления закона распределения и оценок его параметров, единый для всех систем распределений** (при условии полных статистических данных, однородности выборки, достаточном ее объеме).

Необходимые расчеты желательно выполнять в автоматизированном режиме с помощью компьютерной программы.

Наличие нескольких систем непрерывных распределений и общего метода оценивания гарантирует правильное вычисление закона распределения с весьма высокой вероятностью, близкой к единице. При этом отпадает необходимость использования критериев согласия, поскольку вычисленное распределение будет наилучшим для аппроксимации имеющихся статистических данных. Чем больше объем выборки и выше степень однородности случайной величины, тем точнее будет аппроксимация.

Рассмотрим каждую из трех отмеченных выше составных частей.

КЛАССИФИКАЦИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Анализ свойств непрерывных случайных величин показывает, что можно четко выделить по крайней мере три их класса.

Отнесем к первому классу такие случайные величины, которые могут принимать как положительные, так и отрицательные значения, а в некоторых случаях и те и другие в одном распределении, например, в системах качества – разброс отклонений от среднего размера детали в большую или меньшую сторону. Распределения таких случайных величин будем описывать **первой системой непрерывных распределений**. Эти распределения обладают характерным для них свойством: с ростом среднего значения вся кривая распределения, т.е. график плотности перемещается по горизонтальной оси без изменения своей формы.

В качестве примера можно привести статистическое распределение сотрудников некоторого учреждения по возрасту. При постоянном составе сотрудников средний возраст и возраст отдельных сотрудников с каждым годом увеличивается на единицу. Это значит, что последующие значения случайной величины (возраста) образуются из предыдущих путем ежегодного прибавления постоянной величины C=1. Для других распределений величина C может принимать другие значения, в том числе разные на одинаковых отрезках времени. Но главным здесь остается правило прибавления ее к предыдущим значениям.

Если величина C постоянна на равных отрезках времени, то среднее значение случайной величины растет во времени по линейному закону. Это еще очень важное свойство распределений первого клас-

са. Количество свойств этих распределений может пополняться. Всем перечисленным свойствам должна удовлетворять первая система непрерывных распределений.

Отнесем ко **второму классу** такие случайные величины, которые заданы на положительной полуоси. Последующие значения таких случайных величин образуются из предыдущих путем их умножения на некоторую положительную величину С. На равных отрезках времени она может принимать различные значения. Если эта величина на равных отрезках времени постоянна, то среднее (и отдельные значения) такой случайной величины растут во времени по показательному закону, а логарифм среднего – по линейному закону.

Распределения таких случайных величин будем описывать **второй системой непрерывных распределений**.

Примером такого распределения является статистическое распределение сотрудников того же учреждения по уровню заработной платы. С учетом инфляции государственные органы корректируют заработную плату, причем, как правило, путем умножения предыдущих значений на некоторую положительную величину С >1. Правда, такой метод повышения заработной платы приводит к негативным последствиям: он увеличивает расслоение общества по доходам – богатые богатеют, бедные беднеют. А отсюда и множество других проблем. Поэтому необходимо периодически использовать метод прибавления некоторой величины С, обеспечивающей приемлемый уровень минимального потребительского бюджета и в то же время – экономию бюджетных средств. Кроме того, этот метод будет сдерживать темп расслоения общества по доходам.

Другими примерами статистических распределений, которые могут быть с высокой точностью описаны второй системой непрерывных распределений, являются распределения тех же сотрудников по росту и весу [1].

Отнесем далее к **третьему классу** такие существенно положительные случайные величины, последующие значения которых образуются из предыдущих путем их возведения в некоторую степень С. Если величина С на равных отрезках времени постоянна, то среднее значение такой случайной величины растет во времени по двойному показательному закону, его логарифм – по показательному закону, а двойной логарифм – по линейному закону.

Распределения таких случайных величин будем описывать **третьей системой непрерывных распределений**. Эта система наряду со второй может быть использована в математической лингвистике, информатике, библиотечном деле для описания статистических ранговых распределений слов частотного словаря, ключевых слов.

Из анализа свойств случайных величин следует важный вывод: **распределение и динамика случайных величин взаимосвязаны**. По известному закону роста среднего значения безошибочно определяется система распределений. Отметим, что это не выдвижение гипотезы, а выбор системы распределений в зависимости от свойств случайной величины.

Зная закон распределения, можно легко решать различные задачи. Например, как изменится распределение работающих по уровню заработной платы при ее новом среднем уровне, как оптимизировать налоговую систему и т.д.

Все вышесказанное относилось к трем классам случайных величин, средние значения которых растут во времени по законам соответственно: линейному, показательному и двойному показательному. Для описания распределений этих величин используются три системы непрерывных распределений.

Но далее неизбежно возникает вопрос: какими распределениями описывать случайные величины, рост которых задается другими законами? На этот вопрос можно дать простой ответ: необходимо найти промежуточные (дополнительные) системы непрерывных распределений. Они должны находиться между первой и второй системами, второй и третьей. А это значит, что с учетом дополнительных систем будет достаточное количество теоретических распределений для описания практически всего многообразия статистических распределений. При этом аппроксимирующие распределения можно будет вычислять в большинстве случаев по двум системам – первой и второй.

Таким образом, выбор системы упрощается до предела и при этом обеспечивается вычисление наилучшего аппроксимирующего распределения, т.е. по сути – закона распределения.

ПОСТРОЕНИЕ СИСТЕМ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ

Поставим общую задачу: построить три системы непрерывных распределений, удовлетворяющие требованиям трех классов случайных величин. Только решение этой общей задачи позволит с высокой точностью аппроксимировать практически любые статистические распределения однородных случайных величин, в том числе ранговые распределения. Как правило, общая задача решается значительно проще частных задач.

Из свойств случайных величин следует, что наиболее просто можно построить вторую систему непрерывных распределений, которые заданы на положительной полуоси.

Для решения этой задачи используем **метод моделирования**. Рассмотрим самые простые распределения, плотности которых заданы **уравнением прямой**. Оно позволяет записать три плотности распределения некоторой случайной величины Т с одним параметром α [2]:

$$p(t)=\alpha(1-\alpha t/2), \quad (0<t<2/\alpha); \tag{1}$$

$$p(t)=\alpha, \quad (0<t<1/\alpha); \tag{2}$$

$$p(t)=2\alpha t, \quad \left(0<t<\sqrt{1/\alpha}\right). \tag{3}$$

Графики этих плотностей, т.е. кривые распределения, имеют вид прямой и заданы на положительной полуоси. Эти распределения записаны из условия, что площадь под кривой распределения равна единице.

Первая из трех приведенных формул представляет собой треугольное убывающее распределение. Вторая – равномерное распределение. Третья – треугольное возрастающее распределение.

Поскольку графики этих плотностей заданы уравнениями прямой, то они принадлежат одной, а именно второй системе непрерывных распределений, которая должна быть задана общей формулой – плотностью распределения с несколькими параметрами. Назовем это распределение обобщенным, или универсальным.

Итак, должно существовать некоторое универсальное распределение, частными случаями которого являются приведенные выше три плотности распределения. Теперь необходимо найти метод, который поможет выявить предполагаемую обобщенную плотность распределения.

Здесь уместно отметить, что закон распределения может быть задан не только плотностью, но и функцией распределения, которая представляет собой интеграл от плотности $F(t)=\int_0^t p(t)dt$.

При интегрировании плотностей распределения (1)–(3) появятся **новые значения параметров**, т.е. будет выявлена новая информация. Вместо этих значений введем новые параметры.

Итак, интегрируя плотности (1)–(3), найдем три функции распределения

$$F(t)=1-(1-\alpha t/2)^2; \quad (4)$$

$$F(t)=\alpha t=1-(1-\alpha t); \quad (5)$$

$$F(t)=\alpha t^2=1-(1-\alpha t^2). \quad (6)$$

Далее используем **метод обобщения**. Обобщим попарно функции распределения (4) и (5), (5) и (6) путем введения новых параметров (например, u, β) вместо показателей степени 1 и 2. В первом случае в результате обобщения функций распределения (4) и (5) будем иметь

$$F(t)=1-(1-\alpha u t)^{\frac{1}{u}}. \quad (7)$$

Во втором случае при обобщении функций распределения (5) и (6) получим

$$F(t)=1-(1-\alpha t^{\beta}). \quad (8)$$

Теперь замечаем, что в формуле (8) имеется параметр β, но его нет в формуле (7). Введем его в формулу (7). В результате получим трехпараметрическую функцию распределения

$$F(t)=1-(1-\alpha u t^{\beta})^{\frac{1}{u}}, \quad (9)$$

откуда дифференцированием по t найдем плотность распределения с тремя параметрами α, β, u:

$$p(t)=\alpha\beta t^{\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}. \quad (10)$$

Последняя плотность может быть еще более расширена за счет **введения нового параметра.** Параметр β в формуле (10) используется дважды в качестве показателя степени. Пусть это будут два разных параметра: в одном случае – параметр β, в другом – произведение kβ, где k – новый параметр. Тогда вместо (10) можем записать искомую четырехпараметрическую плотность в виде [2]

$$p(t)=Nt^{k\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}, \quad (11)$$

где N – нормирующий множитель, зависящий от четырех параметров α, β, k, u. Он вычисляется из условия, что площадь под кривой распределения равна единице.

Итак, простейшими средствами – моделирования плотностей на базе уравнения прямой, интегрирования, обобщения функций распределения, дифференцирования трехпараметрической функции распределения и, наконец, введения четвертого параметра k, – нами получена универсальная четырехпараметрическая плотность распределения, предназначенная для аппроксимации существенно положительных случайных величин (T>0), в том числе статистических ранговых распределений. Эта плотность задает **вторую систему непрерывных распределений**. Отметим, что в библиотечно-информационной деятельности она может использоваться как **универсальный закон рассеяния публикаций** [2, 3]. Приведем формулировку закона Бредфорда: «Если научные журналы расположить в порядке убывания числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп или зон, каждая из которых содержит столько же статей, что и ядро. Тогда числа журналов в ядре и последующих зонах будут относиться как $1:n:n^2$». Из этой формулировки следует, что закон рассеяния Бредфорда основан на статистических ранговых распределениях, но об их свойствах Бредфорд ничего не сообщает.

Плотность (11) можно получить и другим, но более сложным методом – на базе кривых роста новых событий и взаимосвязанных с ними законов распределения вероятностей новых событий. Кривые роста позволяют также построить систему дискретных распределений [4]. Однако приведенный выше метод построения обобщенного распределения значительно проще.

Нетрудно убедиться в том, что плотность (11) включает как частные случаи нормальный закон, законы Стьюдента, Коши, Вейбулла, Максвелла, «хи-квадрат», гамма-распределение, бета-распределение и множество других, в том числе закон Ципфа. Она с успехом может описывать гистограммы с различной формой, а также статистические ранговые распределения. Исследования автора на большом числе статистических распределений, в том числе ранговых, показали, что ни в одном случае не был вычислен закон Ципфа, хотя в различных исследованиях он используется довольно часто. Нормальным законом описывается незначительная доля статистических распределений. Следовательно, при наличии обобщенной плотности метод выдвижения гипотез лишается смысла, как и частично – критерии согласия, поскольку для однородной выборки достаточно боль-

шого объема они покажут высокую точность аппроксимации статистического распределения вычисленной четырехпараметрической плотностью.

На базе плотности (11) легко получить другие плотности как распределения функций случайного аргумента. Например, при $T = e^X$ по известной формуле $p(x) = p(t)(dt/dx)$ найдем

$$p(x) = Ne^{k\beta x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}. \qquad (12)$$

Здесь случайная величина X может быть задана на всей числовой оси, т.е. значения x могут быть как положительными, так и отрицательными. Эта плотность задает **первую систему непрерывных распределений.** Отметим, что она может использоваться в библиотечно-информационной деятельности как **универсальный закон старения публикаций** [2].

Полученные плотности содержат два параметра формы k, u. Главным из них является параметр u. От его значений зависит тип распределения [2, с. 144]. Параметр β может быть параметром формы или масштаба, а параметр α – масштаба или сдвига.

Найдем, наконец, третью систему непрерывных распределений.

Если в формуле (11) принять T=LnY, то получим

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}\left[1 - \alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (13)$$

Последняя плотность задает **третью систему непрерывных распределений**. Она может использоваться в математической лингвистике для описания ранговых распределений слов частотного словаря.

На базе полученных плотностей можно достаточно просто решать различные задачи.

Так, график плотности (12) при значениях параметра формы u ≤ 1/2 имеет три характерные точки: моду x_C и две точки перегиба x_A, x_B, расположенные на равных расстояниях по обе стороны от моды. При 1/2 < u < 1 имеются две характерные точки – x_A и x_C (для распределений с левосторонней асимметрией). При u ≥ 1 кривая не имеет характерных точек.

График плотности (11) при определенных значениях параметров имеет вид убывающей кривой. Следовательно, эта плотность может описывать ранговые распределения, представленные в системе координат $p_r = f(r)$, где r – ранг события, p_r – его относительная частота. Но такая форма представления ранговых распределений несет слишком мало информации о них, поскольку убывающая кривая не имеет никаких характерных точек. А их можно было бы использовать для **вычисления границ ядра и зон рассеяния** публикаций. Со времени окончательной формулировки С. Бредфордом своего закона рассеяния, т.е. с 1948 г. никто не предложил метода вычисления границ ядра и зон рассеяния публикаций по статистическому ранговому распределению, несмотря на то, что было предпринято множество попыток уточнения закона С. Бредфорда. Но это частная задача, решить которую невозможно без предварительного решения общей задачи – разработки систем непрерывных распределений и исследования их свойств. При наличии систем распределений эта задача решается весьма просто. Изложим метод ее решения.

Координаты трех характерных точек рангового распределения можно найти путем преобразования плотности (11) к форме плотности (12). Это достигается путем умножения левой и правой части выражения (11) на t и использования равенства $t^{\beta} = e^{\beta \ln t}$. В итоге имеем [3, 5]

$$tp(t) = Ne^{k\beta \ln t}\left(1 - \alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (14)$$

Последнее равенство также представляет собой плотность распределения. Действительно, если ввести обозначение x=lnt, то плотность p(lnt) можно получить на базе плотности p(x) как распределение функции случайного аргумента:

$$p(\ln t) = p(x)\frac{dx}{d\ln t} = p(x).$$

С учетом плотности p(x) и равенства x=lnt можем записать

$$p(\ln t) = Ne^{k\beta \ln t}\left(1 - \alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (15)$$

Из формул (14) и (15) следует соотношение между плотностями p(t) и p(lnt)

$$tp(t) = p(\ln t). \qquad (16)$$

Из тех же формул следует также равенство функций распределения

$$F(\ln t) = \int p(\ln t)d\ln t = \int tp(t)d\ln t = \int tp(t)\frac{dt}{t} = F(t),$$

т.е.

$$F(\ln t) = F(t). \qquad (17)$$

При дифференцировании последнего равенства по t из него следует равенство (16). Следовательно, плотность p(t), приведенная к форме $tp(t) = f(\ln t)$, представляет собой плотность p(x) и обладает всеми свойствами последней, т.е. при значениях параметра u ≤ 1/2 она имеет три характерные точки A, C, B. Отсюда вытекает правило: чтобы для убывающего рангового распределения найти характерные точки, его необходимо привести к форме плотности p(x), т.е. изобразить графически в системе координат $tp(t) = f(\ln t)$ [3, 5, 6] (см. рис.).

Тогда ранговое распределение будет иметь моду $\ln t_C$ и две точки перегиба $\ln t_A$ и $\ln t_B$, которые находятся на равных расстояниях от моды:

$$\ln t_C - \ln t_A = \ln t_B - \ln t_C.$$

Абсциссы этих характерных точек **приняты автором настоящей статьи в качестве границ ядра журналов и зон рассеяния** (см. рис.). Из последнего равенства следует соотношение

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n, \qquad (18)$$

которое можно принять в качестве уточненной формулировки закона рассеяния публикаций в толковании С. Бредфорда, хотя оно представляет собой новую формулировку закона рассеяния публикаций.

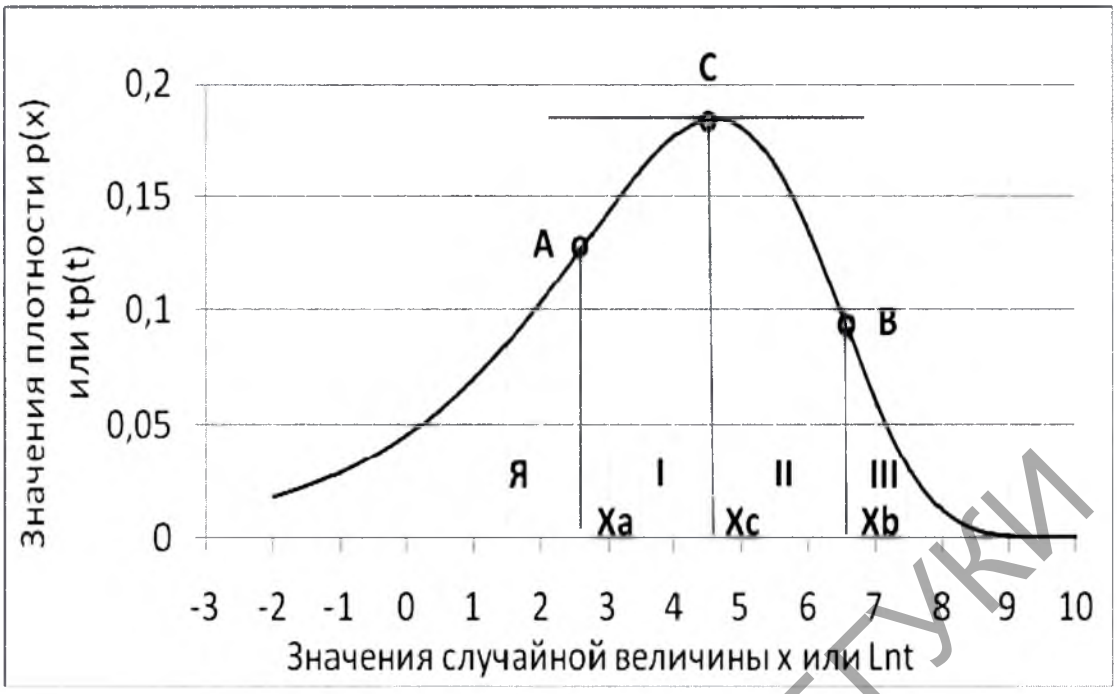


График плотностей *p(x)* и *tp(t)=f(lnt)* при равных значениях параметров

И вывод такой простой формулы оказался возможным лишь после построения обобщенных распределений! Полученное нами соотношение непротиворечиво и универсально, так как оно справедливо для всех распределений, образующих вторую систему непрерывных распределений при значениях параметра формы $u \leq 1/2$. Несмотря на это, **равенство (18) не является законом рассеяния публикаций** (как и «закон» Бредфорда), поскольку не содержит информации о количестве журналов и долях статей в каждой зоне, о том, сколько может быть зон рассеяния, как вычислить границы ядра и зон рассеяния, чему равна величина n. Вся эта информация содержится в обобщенной плотности p(t). Поэтому **вторая система непрерывных распределений по праву является универсальным законом рассеяния публикаций,** а формула (18) – лишь следствие свойств универсального закона. Она отражает соотношение между абсциссами характерных точек на кривой распределения. Значения же функции распределения в характерных точках в этой формуле не задействованы. Поэтому, не зная теоретического рангового распределения с его значениями параметров, нельзя вычислить число журналов, входящих в ядро и зоны рассеяния, величину n, а также доли статей в ядре и зонах рассеяния, которые выражаются через функцию распределения.

Таким образом, приведение плотности (11) к форме плотности (12) позволило выявить новую информацию о существовании трех характерных точек рангового распределения при значениях параметра формы $u \leq 1/2$. При $u \geq 1$ характерных точек на кривой рангового распределения не существует. К такому типу кривых относится закон Ципфа, который является частным случаем обобщенной плотности (11) с параметрами формы $u=1$, $\beta<0$. Его ни в коем случае нельзя использовать для аппроксимации статистических ранговых распределений, что убедительно показывает приведенный выше график. Представленные в виде зависимости tp(t)=f(lnt) ранговые распределения имеют моду и две точки перегиба. Для закона же Ципфа произведение ранга на относительную частоту равно постоянной величине. Следовательно, горизонтальная касательная к кривой распределения в точке С и есть закон Ципфа, а все ранговые распределения находятся под этой прямой.

Распределение (14) задает новую форму представления ранговых распределений (слов частотного словаря; журналов, упорядоченных по убыванию числа опубликованных в них статей по заданной тематике или числу обращений к ним, книг по числу их выдач; ученых в зависимости от количества публикаций или числа ссылок на них и множества других объектов исследования, упорядоченных по убыванию некоторого признака), а именно: по горизонтальной оси откладываются логарифмы рангов, а по вертикальной – произведения рангов на относительные частоты. В результате получается одновершинная кривая распределения, позволяющая быстро оценить, является ли выборка однородной и достаточен ли ее объем [2, 5], а также вычислить координаты характерных точек в случае однородной выборки.

Необходимо отметить, что для ранговых распределений характерна высокая энтропия. Предложенный метод приведения второй системы распределений к форме первой позволил уменьшить энтропию

распределений, в том числе ранговых, и извлечь информацию из ранговых распределений в виде трех характерных точек [7].

Можно утверждать, что универсальные законы старения и рассеяния публикаций, а также ранговые распределения лексических единиц, разных наименований книг, заданные обобщенными плотностями, являются фундаментальными закономерностями информатики, математической лингвистики и библиотековедения. Поскольку «…рассеяние научной информации является краеугольным камнем всей научно-информационной деятельности, а изучение этого свойства научной информации – важнейшей проблемой информатики» [8, с. 93], то эту проблему необходимо разрешать весьма серьезными средствами. К таким средствам относятся рассмотренные выше обобщенные распределения. Они могут быть также успешно использованы в теории вероятностей и математической статистике, библиометрии, наукометрии, экономике (эконометрии), социологии, в системах менеджмента качества и во всех других областях знания, где требуется высокая точность аппроксимации статистических распределений, в том числе ранговых.

Плотности (11), (12), (13) задают три основные системы непрерывных распределений [2]. Докажем, что каждая система соответствует требованиям своего класса случайных величин.

Рассмотрим **первую систему** непрерывных распределений, заданную обобщенной плотностью (12)

$$p(x) = Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}.$$

Эта система обладает тем свойством, что при увеличении всех значений случайной величины X на постоянную величину C форма кривой распределения, т.е. графика плотности p(x) не изменяется. Обозначим новое значение случайной величины X через X*, при этом

$$X^*=X+C. \qquad (19)$$

Тогда распределение новой случайной величины X* определится по формуле

$$p(x^*) = p(x)\frac{dx}{dx^*}. \qquad (20)$$

Поскольку на основании (19) dx/dx* = 1, то из формулы (20) следует равенство p(x*) = p(x). Подставляя сюда вместо p(x) плотность (12), с учетом равенства (19) получим

$$p(x^*) = Ne^{k\beta(x^*-C)}\left(1-\alpha u e^{\beta(x^*-C)}\right)^{\frac{1}{u}-1}. \qquad (21)$$

Последнюю плотность можно привести к виду

$$p(x^*) = N^* e^{k\beta x^*}\left(1-\alpha^* u e^{\beta x^*}\right)^{\frac{1}{u}-1}, \qquad (22)$$

где

$$N^* = \mathrm{N}/\mathrm{e}^{\mathrm{k}\beta C};\ \ \alpha^* = \alpha / e^{\beta C}. \qquad (23)$$

Таким образом, смещение случайной величины X на постоянную C приводит к изменению параметра сдвига α и вместе с ним нормирующего множителя N. Параметры формы k, u, β не изменяются, т.е. не изменяется форма кривой распределения, что и требовалось доказать. Поскольку случайные величины X и X* связаны функциональной зависимостью, причем с ростом X растет и X*, то их функции распределения равны

$$F(x^*) = F(x). \qquad (24)$$

Формулы (21) и (22) позволяют прогнозировать распределение случайной величины X. Чтобы рассчитать новые значения плотности распределения с учетом смещения C, в случае первой системы непрерывных распределений достаточно сместить на C значения случайной величины без изменения значений плотности распределения.

Рассмотрим далее случай, когда последующие значения случайной величины X образуются из предыдущих путем их умножения на постоянную величину C: $X^* = X \cdot C$.

Тогда $X = X^*/C$, $dx/dx^* = 1/C$. Плотность $p(x^*)$ получается из плотности $p(x)$ при прежних значениях параметров α, k, u, но при этом параметр формы β и нормирующий множитель N уменьшаются в C раз: β*=β/C, N*=N/C. С уменьшением параметра β кривая распределения становится более пологой и длинной. Плотность распределения $p(x^*)$ задается формулой

$$p(x^*) = N^* e^{k\beta^* x}\left(1-\alpha u e^{\beta^* x}\right)^{\frac{1}{u}-1}. \qquad (25)$$

Итак, умножение случайной величины X на постоянную величину C приводит к уменьшению одного из параметров формы – β. Два других параметра k, u остаются прежними. Формулу (25) также можно использовать для прогнозирования распределений первой системы при условии, когда случайная величина X увеличивается в C раз.

Перейдем ко **второй системе** непрерывных распределений.

Распределение случайной величины T задается обобщенной плотностью (11)

$$p(t) = Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}.$$

Пусть все значения случайной величины T увеличатся в C раз. Требуется найти распределение случайной величины T*=T·C. Поскольку t=t*/C, dt/dt*=1/C, то

$$p(t^*) = p(t)\frac{dt}{dt^*} = \frac{p(t)}{C} \qquad (26)$$

или

$$p(t^*) = \frac{N}{C^{k\beta}} t^{*k\beta-1}\left(1-\frac{\alpha}{C^{\beta}} u t^{*\beta}\right)^{\frac{1}{u}-1}. \qquad (27)$$

Введя обозначения

$$N^* = N/C^{k\beta},\quad \alpha^* = \alpha/C^{\beta}, \qquad (28)$$

последнюю плотность перепишем в виде

$$p(t^*) = N^* t^{*k\beta-1}\left(1-\alpha^* u t^{*\beta}\right)^{\frac{1}{u}-1}. \qquad (29)$$

Увеличение случайной величины T в C раз приводит к уменьшению параметра α и нормирующего множителя N. При этом плотность распределения p(t) уменьшается в C раз (см. формулу (26)), а произведение tp(t), а также среднее значение $\overline{tp(t)}$ остаются без изменения. Это значит, что форма кривой распределения $tp(t) = f(\ln t)$ не изменяется, поскольку не изменяются параметры формы k, u, β. При этом справедливы равенства: F(t*)=F(t), t*p(t*)=tp(t).

Пусть далее случайная величина T возводится в степень C: $T^* = T^C$. Отсюда находим

$$T = T^{*1/C}, \quad dt/dt^* = (1/C)t^{*1/C-1}.$$

Плотность p(t*) равна $p(t^*) = p(t)dt/dt^*$, или

$$p(t^*) = N^* t^{*k\beta^*-1}\left(1-\alpha u t^{*\beta^*}\right)^{\frac{1}{u}-1}, \qquad (30)$$

где $N^* = N/C$, $\beta^* = \beta/C$.

Как видим, в этом случае уменьшились в C раз нормирующий множитель и параметр формы β. Неизменными остались параметр α и параметры формы k, u.

Плотности (29) и (30) позволяют прогнозировать распределения второй системы.

Наконец, рассмотрим **третью систему** непрерывных распределений, заданную плотностью (13)

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}\left[1-\alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1}.$$

Возведем случайную величину Y в степень C и запишем равенство: $Y^* = Y^C$. Отсюда находим: $Y = Y^{*1/C}$, $dy/dy^* = (1/C)y^{*1/C-1}$. Плотность p(y*) задается формулой p(y*)=p(y)dy/dy*, или

$$p(y^*) = \frac{N^*}{y^*}(\ln y^*)^{k\beta-1}\left[1-\alpha^* u(\ln y^*)^{\beta}\right]^{\frac{1}{u}-1}, \qquad (31)$$

где $N^* = N/C^{k\beta}$, $\alpha^* = \alpha/C^{\beta}$.

Из полученных формул видно, что возведение случайной величины Y в степень C не изменяет параметров формы k, u, β. Изменяется лишь параметр α, а вместе с ним и нормирующий множитель N. Обобщенная плотность (31) позволяет прогнозировать распределения третьей системы.

Итак, все три обобщенные плотности удовлетворяют требованиям случайных величин своего класса.

Осталось найти **дополнительные системы**. Поскольку они должны служить связующими звеньями между первой и второй основными системами, второй и третьей, третьей и четвертой [2], то их необходимо получить на базе найденных выше четырехпараметрических распределений. Рассмотрим первую систему непрерывных распределений, заданную плотностью (12)

$$p(x) = Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}.$$

Пусть случайная величина X связана со случайной величиной T зависимостью X=ln(T–L), где L – параметр сдвига. Тогда первая производная $dx/dt = 1/(t-l)$ и плотность p(t) задается пятипараметрической формулой [2]

$$p(t) = N(t-l)^{k\beta-1}\left(1-\alpha u(t-l)^{\beta}\right)^{\frac{1}{u}-1}. \qquad (32)$$

ДОПОЛНИТЕЛЬНЫЕ СИСТЕМЫ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ

В результате исследований автором установлено, что **дополнительными для первой системы** непрерывных распределений могут быть две плотности

$$p(t) = N(t-l)^{k-1}\left[1-\alpha u(t-l)\right]^{\frac{1}{u}-1}, \qquad (33)$$

$$p(t) = N\left[1-\alpha u(t-\bar{t})^2\right]^{\frac{1}{u}-1}, \qquad (34)$$

которые являются частными случаями плотности (32). В первой из них параметр $\beta=1$, во второй $\beta=2$. Величина $\bar{t}$ – среднее значение. Вторая плотность задает семейство симметричных распределений. Распределения, заданные плотностью (33), также симметричны при условии $ku=1$. Две приведенные плотности задают основную часть семейства распределений Пирсона. Они являются связующим звеном первой и второй систем непрерывных распределений. Отметим, что частным случаем плотности (34) является нормальный закон при u→0.

Аналогично во **вторую систему** непрерывных распределений войдут дополнительные плотности

$$p(y) = \frac{N(\ln y - l)^{k-1}}{y}\left[1-\alpha u(\ln y - l)\right]^{\frac{1}{u}-1}, \qquad (35)$$

$$p(y) = \frac{N}{y}\left[1-\alpha u(\ln y - \overline{\ln y})^2\right]^{\frac{1}{u}-1}, \qquad (36)$$

которые следуют из плотностей (33), (34) при T=LnY. Они являются переходными между второй и третьей системами. Из плотности (36) при u→0 следует логарифмически нормальный закон.

Так же можно ввести дополнительные системы для третьей плотности.

Легко проверить, что дополнительные системы распределений соответствуют свойствам случайных величин своих классов.

В результате каждая из трех систем непрерывных распределений содержит по три плотности. Следовательно, всего имеется девять обобщенных плотностей, включающих более 50 типов распределений, которых достаточно для аппроксимации подавляющего большинства статистических распределений.

Итак, этап построения систем непрерывных распределений нами пройден. Далее необходимо всесторонне их исследовать: дать классификацию обобщенных распределений; выразить нормирующие множители всех типов распределений через их параметры; рассмотреть возможные формы кривых распределения и их зависимость от значений параметров формы; разработать методы вычисления законов распределения и оценок параметров; разработать алгоритмы и компьютерные программы; проверить по-

строенную теорию на большом статистическом материале; дать сравнительный анализ построенных моделей, методов, алгоритмов, программ и тех, что широко используются в настоящее время. Главной и наиболее сложной из этого перечня задач является разработка метода вычисления закона распределения.

ОБЩИЙ УСТОЙЧИВЫЙ МЕТОД ВЫЧИСЛЕНИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ И ОЦЕНОК ПАРАМЕТРОВ

Наконец, осталось разработать общий, единый для трех систем непрерывных распределений метод **вычисления закона распределения и оценок параметров**, причем, метод должен быть устойчивым к выбросам на концах статистического распределения.

Еще раз отметим, что существующие методы позволяют вычислять оценки параметров тех распределений, которые **заранее выбраны** в качестве нулевой гипотезы. Далее эмпирические моменты (или другие величины) приравниваются к теоретическим и решается система уравнений, число которых равно числу параметров выбранного распределения. Если после проверки по критериям согласия нулевая гипотеза не подтверждается, выбирается другое распределение и все расчеты повторяются. Такой подход для обобщенных распределений, разработанных автором данной статьи, неприемлем по двум причинам.

Во-первых, свойства случайной величины однозначно определяют нужную систему распределений – обычно первую, либо вторую. Здесь ошибиться трудно. Далее нашей задачей является **вычисление** в этой системе **закона распределения и оценок его параметров**.

Во-вторых, автором предложен метод, с помощью которого поставленная задача решается в **два этапа**. На первом этапе по двум показателям, зависящим от параметров формы k, u, устанавливается тип аппроксимирующей кривой и вычисляются оценки параметров формы **путем решения системы двух уравнений с двумя неизвестными**, что значительно упрощает расчеты. На втором этапе вычисляются оценки двух других параметров по относительно простым формулам (с учетом известных оценок параметров k, u). Заметим, что этот метод требует предварительного группирования статистических данных.

Отныне для установления теоретического закона распределения непрерывной случайной величины по ее статистическому распределению не требуется выдвижения многочисленных гипотез об аппроксимирующей кривой и проверки каждой из них по критериям согласия. Система непрерывных распределений выбирается в зависимости от свойств случайной величины, а тип распределения и оценки его параметров определяются расчетом.

При этом основная сложность заключается в разработке двух показателей для установления типа распределения и вычисления оценок параметров формы. Для разработки таких показателей система четырехпараметрических непрерывных распределений случайной величины X сводится к системе двухпараметрических распределений случайной величины Z. Затем используются взаимосвязи между случайными величинами X и Z и их плотностями распределения p(x), p(z).

Рассмотрим обобщенную плотность p(x), которая задает первую основную систему непрерывных распределений (см. формулу (25)). Введем два показателя – асимметрии B и островершинности H, которые зависят от двух параметров формы k, u. По этим показателям однозначно устанавливается тип аппроксимирующего распределения и находятся оценки параметров k, u с помощью номограммы (см.ниже).

Для обобщенной плотности p(x) показатели B, H задаются формулами

$$\left.\begin{aligned} B &= M\left[p(x)(x-M(x))\right] = f(k,u) \\ H &= S_3 / S_1^3 = f(k,u) \end{aligned}\right\}, \qquad (37)$$

где

$$S_r = M\left[p(x)\right]^r = f(\beta, \mathrm{k}, \mathrm{u}). \qquad (38)$$

Исследования показали, что величина H задана на интервале $\sqrt{2} < H < 2$, а величина B – на интервале $-1/4 < B < 1/4$.

Если вычислить для разных типов распределений значения показателей B, H при различных значениях параметров k, u, то по этим данным можно построить номограмму (бинарную сетку) [2, 3], которая применима к трем основным системам непрерывных распределений, заданным первыми плотностями. При этом плотности p(t) и p(y) должны быть приведены к форме плотности p(x), т.е. представлены в виде tp(t)=f(lnt), ylnyp(y)=f(lnlny).

Для дополнительных систем автором построена другая номограмма, которая является продолжением первой [9].

Чтобы найти закон распределения, достаточно вычислить по статистическому распределению оценки показателей B, H. Эти показатели однозначно определяют тип распределения, приведенного к форме плотности p(x). Более того, с их помощью легко находятся оценки параметров k, u непосредственно из номограммы (*Приложение 1*). Более точно они вычисляются по программе автора. Оставшиеся два параметра (α и β) вычисляются по специальным формулам [2]. Метод обеспечивает вычисление как одновершинных частотных законов распределения, так и ранговых. Здесь уместно отметить, что показатели B, H играют роль критериев согласия теоретического и статистического распределений.

При известных оценках параметров **рангового распределения** легко вычисляются координаты характерных точек по заранее выведенным формулам, что позволяет находить ядро журналов или книжного фонда, а точнее, ядро профессиональных информационных потребностей читателей t_A (на приведенном рисунке абсцисса точки $x_A = \ln t_A$, откуда $t_A = e^{x_A}$); зоны рассеяния I, II, III; оптимальный объем фонда (активная его часть) $t_B = e^{x_B}$; а также вычислять информационную полноту комплектования

ального фонда или, другими словами, вероят- удовлетворения информационных потребно- ользователей этим фондом. Например, в случае дливости закона Вейбулла ядро фонда удовле- т информационные потребности пользовате- а 31,75%, а оптимальный объем фонда – на %. Эти величины зависят от параметров ап- имирующего рангового распределения, а в ито- т статистического рангового распределения и тавляют собой накопленную долю книговыдач, дящуюся на ядро фонда t_A и оптимальный фонда t_B. В данном случае она выражена в нтах. Здесь следует уточнить, что под ядром в нашем случае понимается наиболее востре- лая его часть, как и в случае с ядром журналов.

личие систем непрерывных распределений и ов вычисления закона распределения и оценок етров позволяет утверждать, что решение мно- практических задач отныне сводится к накоп- достаточного количества статистических дан- Именно в них содержится ценная информация, я может быть извлечена путем вычисления за- аспределения и его всестороннего исследова- сли же речь идет о виде аппроксимирующего деления, то здесь вовсе не возникает вопросов – ма непрерывных распределений (первая или) выбирается в соответствии со свойствами ной величины, а тип и оценки параметров шего аппроксимирующего распределения на- я путем расчета. Поэтому некоторые задачи ваются заранее решенными, и только для вы- ния оценок параметров теоретических распре- ий требуется обработка статистических данных. ие множества теоретических задач также зна- но упрощается.

льнейшие исследования показывают, что об- стойчивый метод по точности аппроксимации тупает методу наибольшего правдоподобия шера, к тому же он значительно проще послед- Существенным преимуществом методов автора ся наличие двух показателей, позволяющих **слять** тип аппроксимирующего распределения ки параметров формы в заданной системе не- вных распределений. Метод Фишера не распо- такими показателями [7]. Их еще предстоит отать.

ЕРСАЛЬНЫЙ МЕТОД МОМЕНТОВ

тод моментов был предложен К. Пирсоном в . (назовем его классическим методом момен- лавное его достоинство – простота. Он позво- ычислять закон распределения и оценки пара- в по статистическому распределению, правда, при условии, когда наверняка известно, что ический закон входит в семейство кривых на. Следует отметить, что данное семейство рывных распределений слишком узкое и за его ами находится бесконечное множество других ических и статистических распределений. того, метод моментов Пирсона может быть ьзован лишь при условии, если существуют моменты вплоть до четвертого порядка. Этих недостатков лишен предложенный автором настоящей статьи универсальный метод моментов.

Суть его заключается в том, что он разработан для первой системы непрерывных распределений, заданной четырехпараметрической плотностью (12)

$$p(x) = Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}. \qquad (39)$$

Для вычисления закона распределения второй и третьей систем непрерывных распределений последние должны быть приведены к форме первой системы. Для приведения второй системы к форме первой необходимо использовать замену переменных X=LnT. Тогда получим плотность (14)

$$tp(t) = Ne^{k\beta \ln t}\left(1-\alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}, \qquad (40)$$

где lnt=x, tp(t)=p(x). Последняя плотность обладает всеми свойствами плотности (12). Поскольку случайная величина X имеет все моменты, то таким же свойством обладает и случайная величина LnT. Отсюда следует, что при выравнивании статистических распределений обобщенными плотностями, которые относятся к разным системам непрерывных распределений, центральные моменты вычисляются по разным формулам.

Итак, рассмотрим первую обобщенную плотность первой системы непрерывных распределений, все параметры которой больше нуля:

$$p(x) = \frac{\beta(\alpha u)^k \Gamma(k+\frac{1}{u})}{\Gamma(k)\Gamma\left(\frac{1}{u}\right)} e^{k\beta x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}. \qquad (41)$$

Введем случайную величину Z, которая связана со случайной величиной X зависимостью

$$Z = \alpha u e^{\beta X} \qquad (42)$$

и найдем ее закон распределения.

На основании (41), (42) и формулы

$$p(z) = p(x)\left(dx/dz\right)$$

имеем:

$$p(z) = \frac{\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(\frac{1}{u}\right)} z^{k-1}(1-z)^{\frac{1}{u}-1} \qquad (43)$$

Это известное бета-распределение. Оно зависит от двух параметров формы k,u.

Пусть нам известны оба этих параметра. Тогда оценку параметра α можно найти из формулы (42). Для этого прологарифмируем ее

$$\ln Z = \ln \alpha u + \beta X \qquad (44)$$

и заменим величины lnZ и X их математическими ожиданиями. В результате получим

$$M(\ln Z) = \ln \alpha u + \beta M(X), \qquad (45)$$

откуда найдем

$$\alpha u = e^{M(\ln Z) - \beta M(X)}. \qquad (46)$$

Величина M(lnZ) зависит от двух параметров k,u. Она может быть вычислена теоретически. Величина M(X) зависит от четырех параметров, но она заменяется оценкой, т.е. средней величиной $\bar{x}$, рассчитанной по статистическому распределению. Здесь уместно отметить, что произведение αu вычисляется также по формуле (46) в случае ранее рассмотренного устойчивого метода.

Теперь осталось найти формулы для вычисления оценок трех параметров: β, k, u. Для этого на базе формул (44), (45) составим равенство для центральных моментов случайных величин lnZ и X:

$$M\left[\ln Z - M(\ln Z)\right]^r = \beta^r M\left[X - M(X)\right]^r. \quad (47)$$

Вводя другие обозначения

$$\mu_r^{(z)} = M\left[\ln Z - M(\ln Z)\right]^r \qquad \mu_r^x = M\left[X - M(X)\right]^r,$$

формулу (47) перепишем в виде

$$\mu_r^z = \beta^r \mu_r^x. \quad (48)$$

При r=2 из (48) следует формула для вычисления параметра β.

$$\beta = \sqrt{\frac{\mu_2^{(z)}}{\mu_2^{(x)}}}. \quad (49)$$

Наконец, осталось найти оценки двух главных параметров: k, u. Они могут быть найдены из формулы (48). Введем два показателя – асимметрии и островершинности, зависящие от двух параметров k, u,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \quad \beta_2 = \frac{\mu_4}{\mu_2^2}. \quad (50)$$

Это – критерии Пирсона, но примененные к обобщенной плотности p(x) с четырьмя параметрами.

Показатели β_1 и β_2 могут быть вычислены по центральным моментам случайной величины lnZ либо случайной величины X. В результате получим формулы, зависящие от двух параметров k, u. На основе таких формул, найденных для распределений всех типов, строится номограмма (*Приложение 2*), позволяющая устанавливать тип аппроксимирующей кривой распределения и вычислять оценки параметров k, u по критериям β_1, β_2, вычисленным по статистическому распределению.

Универсальный метод моментов, разработанный для первой системы непрерывных распределений (плотности p(x)), может быть применен ко второй и третьей системам непрерывных распределений. Но для этого их необходимо привести к форме первой системы непрерывных распределений. Например, плотность (11) должна быть приведена к форме (40). Плотность (13) – третья система распределений – должна быть приведена к форме yp(y)lny = f(lnlny).

Отсюда следует, что центральные моменты каждой системы непрерывных распределений вычисляются по своим формулам, но при этом тип распределения и оценки параметров формы k, u находятся по одной и той же номограмме. Отметим, что часть номограммы *Приложения 2* ниже прямой $\beta_2 = 3 + 1.5\beta_1$ относится к дополнительным системам непрерывных распределений с параметром β=1.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МЕТОД ГАУССА)

Этот чрезвычайно простой метод целесообразно использовать в том случае, когда закон распределения может быть представлен уравнением прямой.

Рассмотрим закон Вейбулла, который является частным случаем обобщенного распределения (11) при $u \to 0$. Его плотность и функция распределения задаются формулами

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}; \; F(t) = 1 - e^{-\alpha t^{\beta}}.$$

Закон Вейбулла во многих случаях хорошо аппроксимирует статистические ранговые распределения.

Представим функцию распределения в виде уравнения прямой

$$\ln\ln(1/(1\text{-}F(t))) = \ln\alpha + \beta\ln t.$$

Вводя новые обозначения, можем записать

$$Y = A + \beta X. \quad (51)$$

Чтобы проверить применимость закона Вейбулла для аппроксимации статистического рангового распределения, необходимо вычислить величины

$$Y = \ln\ln(1/(1\text{-}F(t)) \text{ и } X = \ln t, \quad (52)$$

где t – ранг события, т.е. его порядковый номер в списке по убывающим относительным частотам. Если эмпирические точки ложатся на прямую (51), необходимо вычислить оценки величин $A = \ln\alpha$ и β по методу наименьших квадратов (см., например, список литературы п. 3, 6):

$$\beta = \frac{\overline{XY} - \bar{X}\cdot\bar{Y}}{\overline{X^2} - \left(\bar{X}\right)^2}; \qquad A = \bar{Y} - \beta\bar{X}.$$

Оценка параметра α равна

$$\alpha = e^{A}.$$

Для вычисления координат трех характерных точек используются формулы, справедливые в случае закона Вейбулла

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}; n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}; t_A = \frac{t_C}{n}; t_B = t_C \cdot n;$$

$$F(t_A) = 0{,}3175; F(t_C) = 0{,}6321; F(t_B) = 0{,}9271.$$

ЗАКЛЮЧЕНИЕ

Эффективность статистических методов в теоретических и прикладных исследованиях в решающей степени зависит от точности аппроксимации статистических распределений. Наибольшую точность аппроксимации можно получить при использовании теории обобщенных распределений автора настоящей статьи, в которой изложены некоторые сведения о названной теории. Кроме того, теория включает

также систему дискретных распределений [4], взаимосвязанную с системой кривых роста новых событий, номограммы для графического определения типа аппроксимирующей кривой и оценок параметров, а также серию компьютерных программ для работы с указанными системами. Применение этой теории на практике значительно облегчает задачу нахождения закона распределения по статистическим данным.

В этом случае нет необходимости выдвигать гипотезы о предполагаемом аппроксимирующем распределении. В зависимости от свойств случайной величины выбирается система непрерывных распределений (как правило, первая или вторая) и по статистическому распределению вычисляются два показателя – асимметрии и островершинности – по формулам, справедливым для данной системы. Далее они приравниваются к соответствующим теоретическим показателям, которые зависят лишь от двух параметров формы k, u, хотя обобщенная плотность содержит, как правило, четыре параметра. С помощью номограммы по двум показателям (B, H) или (β_1,β_2) устанавливается тип теоретического распределения и находятся оценки параметров k, u: в первом приближении – в ручном режиме, а более точные их значения, а также значения параметров α, β – в автоматизированном режиме по программам автора.

Следует отметить, что одной точке на номограмме с заданными значениями показателей B, H и параметров формы k, u соответствует не единственное распределение, а множество распределений с различными значениями параметров α, β. Например, во второй системе распределений одинаковые значения параметров формы k=1, u→0 имеют такие распределения, как показательное (β=1), Релея (β=2) и Вейбулла (β>0). Оценки параметров α, β вычисляются по специальным формулам с учетом найденных оценок параметров формы k, u.

Обобщенные распределения включают как частные случаи множество известных распределений, в том числе семейство кривых Пирсона, и могут претендовать на роль универсальных законов распределения не только теории вероятностей и математической статистики, но и информатики, математической лингвистики, библиотековедения, библиометрии, наукометрии, эконометрии, экономики, социологии и многих других областей знания. Применение обобщенных распределений и общего устойчивого метода вычисления закона распределения по статистическим данным гарантирует высокую экономическую эффективность статистических методов во всех практических приложениях. Так, использование обобщенных распределений в системах менеджмента качества позволяет с высокой точностью оценивать возможности технологических процессов и поддерживать их в статистически управляемом состоянии при любом законе распределения технологических погрешностей, что обеспечивает значительное снижение уровня брака.

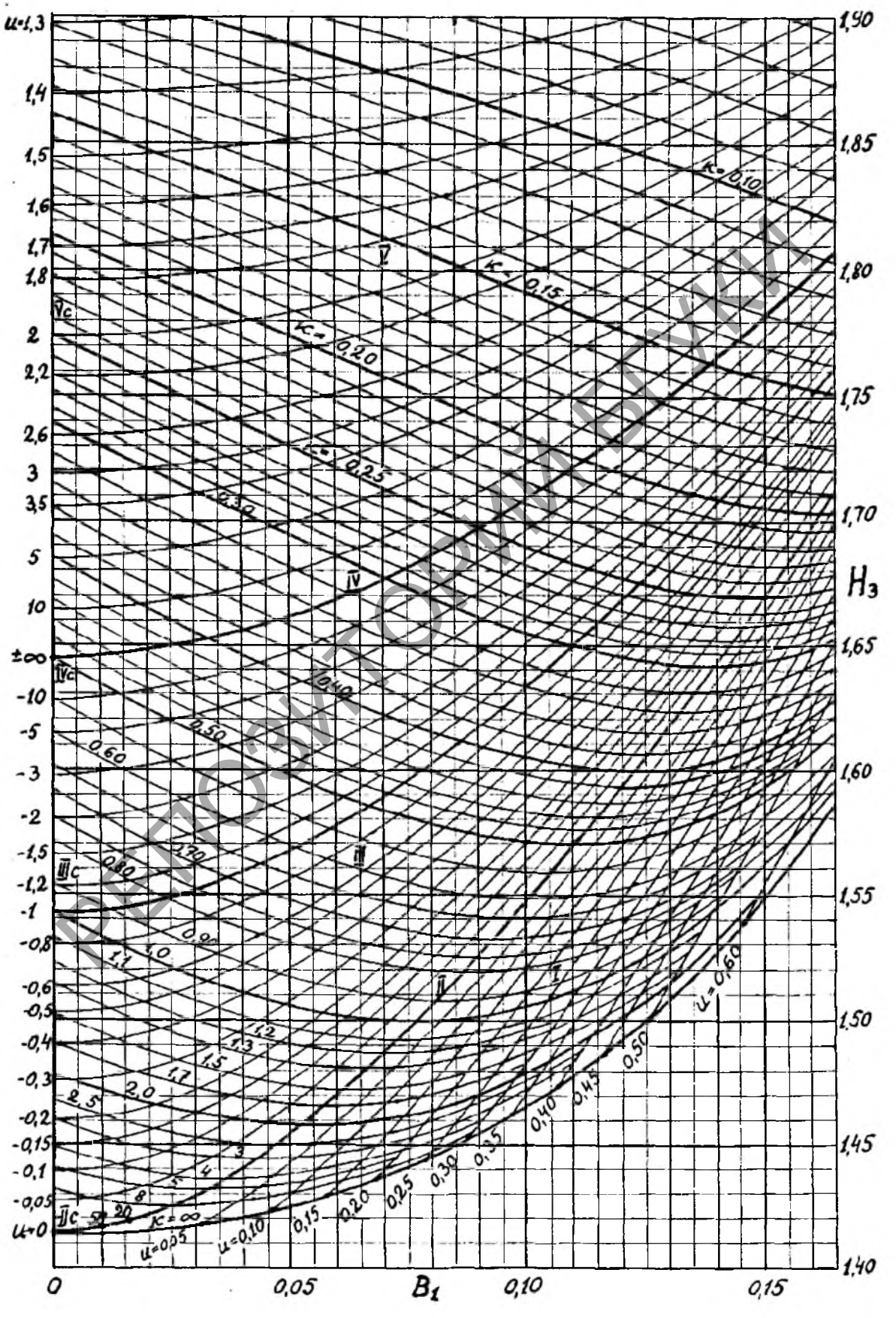
Но широкое использование теории обобщенных распределений можно реализовать лишь путем включения ее в учебные программы высших учебных заведений, что обещает дать большой экономический эффект во всех областях ее применения.

СПИСОК ЛИТЕРАТУРЫ

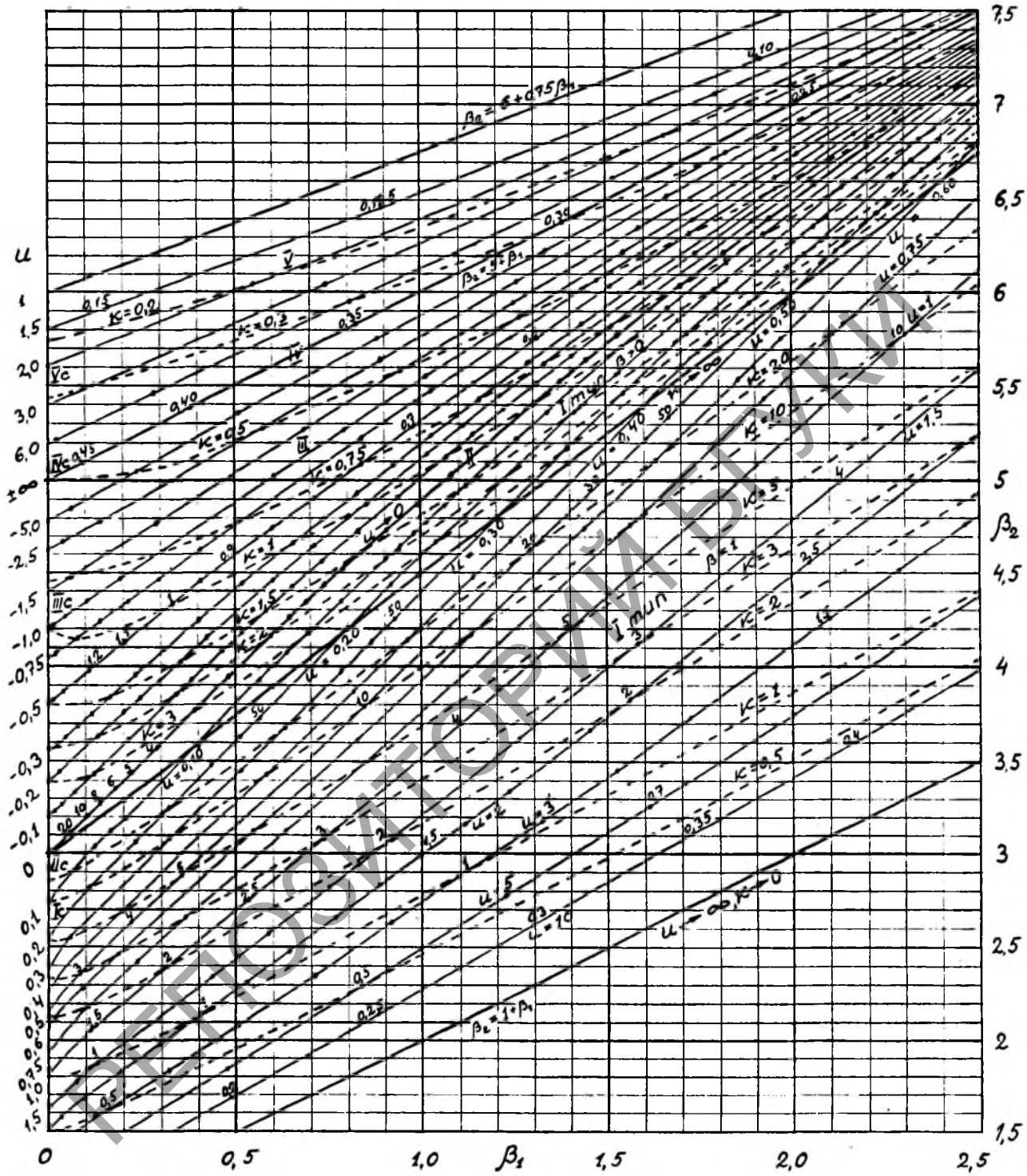
1. Нешитой В.В. Как вычислить закон распределения случайной величины? // Материалы XI Междунар. конф. «Медико-социальная экология личности: состояние и перспективы», 17–18 мая 2013г., г. Минск / отв. ред. В.А. Прокашева. – Минск: Изд. центр БГУ, 2013. – С. 484–486.
2. Нешитой В.В. Элементы теории обобщенных распределений: монография. – Минск: РИВШ, 2009. – 204 с.
3. Нешитой В.В. Законы Ципфа, Бредфорда и универсальные модели // Научно-техническая информация. Сер. 2. – 2010. – № 1. – С. 26–33; Neshitoi V.V. Zipf's and Bradford's laws and universal models // Automatic Documentation and Mathematical Linguistics. – 2010. – Vol. 44, № 1. – P. 30–37.
4. Нешитой В.В. Методы статанализа в библиотечно-информационной деятельности: вычисление дискретных распределений и кривых роста: учеб.-метод. пособие. – Минск: РИВШ, 2012. – 134 с.
5. Нешитой В.В. Форма представления ранговых распределений // Ученые записки Тартуского гос. ун-та. – 1987. – Вып. 774. – С. 123–134.
6. Нешитой В.В. Методы вычисления границ ядра и зон рассеяния публикаций // Научно-техническая информация. Сер. 2. – 2013. – №11. – С. 1–11; Neshitoi V.V. Methods for Calculating the Boundaries of the Core and Zones of Scattering of Publications // Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, № 5–6. – P. 169–179.
7. Нешитой В.В. Метод наибольшего правдоподобия, устойчивый метод и энтропия // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 27–33.
8. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики.– М.: Наука, 1968. – 756 с.
9. Нешитой В.В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие. – Минск: Бел. гос. ун-т культуры и искусств, 2010. – 61 с.

Приложение 1.

Номограмма для вычисления закона распределения и оценок параметров по общему устойчивому методу



Номограмма для вычисления закона распределения и оценок параметров по универсальному методу моментов



Материал поступил в редакцию 25.12.15.

Сведения об авторе

НЕШИТОЙ Василий Васильевич – доктор технических наук, профессор, профессор кафедры информационных ресурсов УО «Белорусский государственный университет культуры и искусств», г. Минск
e-mail: neshitoi_vv@tut.by