УДК 001.102:001.891.57

В. В. Нешитой

Законы Ципфа, Бредфорда и универсальные модели

В социально-гуманитарной сфере широкое распространение получили законы Дж. Ципфа и С. Бредфорда. Оба они представляют собой эмпирические законы, полученные путём обработки и анализа так называемых ранговых распределений. В настоящей работе они рассматриваются как частные случаи обобщенных (универсальных) непрерывных распределений, способных с высокой точностью описывать широкое разнообразие статистических распределений, в том числе ранговых. Приводятся универсальные законы рассеяния и старения публикаций.

Ключевые слова: *ранговые распределения, закон Ципфа, закон Бредфорда, универсальный закон, системы непрерывных распределений, рассеяние публикаций, старение публикаций, вычисление аппроксимирующих распределений, устойчивый метод.*

ЗАКОНЫ ЦИФРА И БРЕДФОРДА

Дж. Ципф предложил свой закон для описания относительной частоты слов частотного словаря [1]

$$p_r = \frac{k}{r}, \qquad (1)$$

где p_r — относительная частота слова с рангом r; k — параметр.

Если все разные слова (словоформы или лексемы), встретившиеся в тексте достаточно большого объёма, упорядочить по убыванию (точнее, по невозрастанию) их относительных частот, то, в зависимости от порядкового номера слова в ранжированном ряду, т. е. его ранга, относительная частота слова в первом приближении может быть оценена по формуле (1) Ципфа.

Закон Ципфа можно записать в более общей форме как закон Ципфа-Мандельброта [2]

$$p_r = \frac{k}{(r+\rho)^\gamma}. \qquad (2)$$

Несмотря на то, что последняя формула содержит два дополнительных параметра, она не способна описывать ранговые распределения с необходимой точностью.

С. Бредфорд получил свой закон путём упорядочения журналов, публикующих статьи по некоторому заданному предмету, по убыванию числа таких статей. Но Бредфорд пошёл дальше Ципфа. Он разделил все журналы на несколько зон: ядро и зоны рассеяния, — предположив при этом, что число статей в каждой зоне журналов такое же, как и в ядре. Тогда числа журналов в ядре и зонах рассеяния относятся как $1 : n : n^2$ [3]. Правда, Бредфорд не предложил конкретных формул для вычисления величины n, координат границ ядра и зон рассеяния, числа журналов в ядре и зонах рассеяния. Но его большая заслуга в том, что он дал формулировку закона рассеяния журнальных публикаций (пусть в первом приближении) и обратил внимание исследователей на эту проблему.

Последователям Ципфа и Бредфорда не удалось уточнить теоретическую модель ранговых распределений Ципфа, а также дать математически точную формулировку закона рассеяния публикаций Бредфорда. И это закономерно, поскольку исследователи строили свои модели на базе закона Ципфа и исходили из предположения о равенстве числа статей в ядре и зонах рассеяния.

Выход из создавшегося положения был найден путём разработки теории обобщённых распределений [4]. Последние заданы четырехпараметрическими плотностями и способны с высокой точностью описывать практически всё многообразие статистических распределений, в том числе ранговых. Теория обобщённых распределений включает три системы непрерывных распределений, систему дискретных распределений, взаимосвязанную с системой кривых роста новых событий, методы оценивания параметров (универсальный метод моментов и обший устойчивый метод), номограммы для графического установления типа аппроксимирующей кривой распределения и нахождения в первом приближении оценок двух параметров формы (при ручном счёте) и серию компьютерных программ для работы с указанными системами. Эта теория позволяет легко решать многие задачи.

ОБОБЩЕННЫЕ РАСПРЕДЕЛЕНИЯ

Рассмотрим три системы непрерывных распределений.

Первая система непрерывных распределений задается плотностями [4]

$$\left.\begin{aligned} p(x) &= Ne^{k\beta x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1} \\ p(t) &= N(t-l)^{k-1}[1-\alpha u(t-l)]^{\frac{1}{u}-1} \\ p(t) &= N[1-\alpha u(t-\bar{t})^2]^{\frac{1}{u}-1} \end{aligned}\right\}, \qquad (3)$$

где N — нормирующий множитель, который выражается через параметры α, β, k, u, l, $\bar{t}$. Оценки параметров вычисляются по статистическим распределениям. Эта система включает три группы симметричных распределений. Одна из них (типы $IIIc - Vc$) задана первой плотностью $p(x)$ при $k = 0.5(1 - 1/u)$. Другая (типа $Ic - Vc$) задана третьей плотностью. Сюда входят нормальный закон Стьюдента, Коши. Третья группа задана второй плотностью при условии $k = u1$.

Первая система непрерывных распределений описывает статистические распределения таких случайных величин, которые заданы на всей числовой оси и растут во времени по линейному закону.

Вторая система непрерывных распределений задается плотностями

$$\left.\begin{aligned} p(t) &= Ne^{k\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1} \\ p(y) &= \frac{N(\ln y - l)^{k-1}}{y}[1-\alpha u(\ln y - l)]^{\frac{1}{u}-1} \\ p(y) &= \frac{N}{y}[1-\alpha u(\ln y - \overline{\ln y})^2]^{\frac{1}{u}-1} \end{aligned}\right\}, \qquad (4)$$

которые могут быть получены из первой системы как распределения функций случайных аргументов: $X = \ln T$ — для первой плотности; $T = \ln Y$ — для двух других плотностей.

Частными случаями этой системы являются логарифмически нормальный закон, закон Ципфа и многие другие.

Распределения второй системы при определенных значениях параметров могут быть убывающими и, следовательно, пригодны для описания ранговых распределений (например, журналов, упорядоченных по убыванию числа помещенных в них статей по некоторому заданному предмету; книг, упорядоченных по убыванию их выдач и др.). Кроме того, вторая система непрерывных распределений может описывать статистические распределения таких существенно положительных случайных величин, которые растут во времени по показательному закону.

Третья система непрерывных распределений задается плотностями

$$\left.\begin{aligned} p(y) &= \frac{N(\ln y)^{k\beta-1}}{y}[1-\alpha u(\ln y)^{\beta}]^{\frac{1}{u}-1} \\ p(w) &= \frac{N(\ln\ln w - l)^{k-1}}{w\ln w}[1-\alpha u(\ln\ln w - l)]^{\frac{1}{u}-1} \\ p(w) &= \frac{N}{w\ln w}[1-\alpha u(\ln\ln w - \overline{\ln\ln w})^2]^{\frac{1}{u}-1} \end{aligned}\right\}. \qquad (5)$$

При определенных значениях параметров эта система также может описывать статистические ранговые распределения (например, лексических единиц — словоформ, лексем, терминов, ключевых слов). Кроме того, она может описывать статистические распределения таких случайных величин, которые растут во времени по двойному показательному закону.

Таким образом, динамика и распределение случайной величины взаимосвязаны [4].

Все распределения каждой системы можно разделить на типы в зависимости от значений параметров u, α и знака параметра β (см. рис. 1).

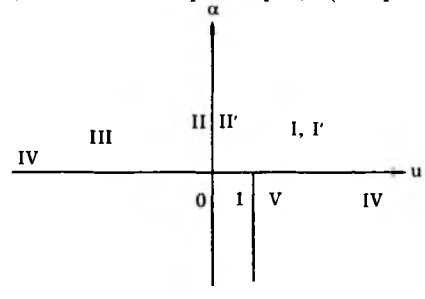

Рис. 1

КЛАССИФИКАЦИЯ РАСПРЕДЕЛЕНИЙ (ТИПЫ СО ШТРИХОМ — ПРИ $\beta < 0$)

Для первых плотностей трех систем распределений нормирующий множитель задается приведенными ниже формулами.

Типы I,I′: $N = \dfrac{\beta(\alpha u)^k \Gamma(k + 1/u)}{\Gamma(k)\Gamma(1/u)}$;

Типы II,II′: $N = \dfrac{\beta\alpha^k}{\Gamma(k)}$;

Типы III–V: $N = \dfrac{\beta(-\alpha u)^k \Gamma(1 - 1/u)}{\Gamma(k)\Gamma(1 - 1/u - k)}$.

Установим место закона Ципфа в системе непрерывных распределений. Рассмотрим распределение I′ типа, заданное первой плотностью второй системы непрерывных распределений. Оно имеет вид (при $\beta > 0$)

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

Здесь величина $t^\beta > \alpha u$, следовательно, $t > (\alpha u)^{1/\beta}$. Перенесём начало отсчёта кривой распределения в точку $t = (\alpha u)^{1/\beta}$ и запишем последнюю плотность при $u = 1$

$$p(t) = \frac{k\beta\alpha^k}{(t + \alpha^{1/\beta})^{k\beta+1}}, \qquad (6)$$

где $0 < t < \infty$. Полученная формула представляет собой закон Ципфа-Мандельброта (2).

ФОРМА КРИВЫХ РАСПРЕДЕЛЕНИЯ

Теоретические распределения, заданные тремя обобщёнными плотностями, имеют множество форм. Конкретный вид кривой распределения зависит от значений параметров формы. Рассмотрим первую систему непрерывных распределений, заданную первой плотностью $p(x)$. Исследования показывают, что кривая распределения при значениях параметра $u \le 1/2$ имеет три характерные точки: моду x_c, в которой плотность максимальна, и две точки перегиба x_A и x_B, расположенные на равных расстояниях от моды, т. е.

$$x_C - x_A = x_B - x_C. \qquad (7)$$

При $1/2 < u < 1$ кривая имеет моду и одну точку перегиба, а при $u \ge 1$ не имеет характерных точек, т. е. кривая либо убывает, либо возрастает.

Плотность $p(x)$ при значениях параметра $u < 1$, т. е. имеющая моду и точки перегиба, не может быть использована для описания ранговых распределений, поскольку они являются убывающими. Но характерные точки можно было бы использовать для вычисления границ ядра и зон рассеяния рангового распределения журналов, если первую плотность $p(t)$ второй системы непрерывных распределений преобразовать к плотности $p(x)$. Для этого умножим левую и правую части плотности $p(t)$ второй системы непрерывных распределений на t, а величину t^β запишем в виде $e^{\beta \ln t}$, что одно и то же. В результате получим

$$tp(t) = Ne^{k\beta \ln t}(1 - \alpha u e^{\beta \ln t})^{\frac{1}{u}-1}. \qquad (8)$$

Последнее равенство также представляет собой плотность распределения. Действительно, если ввести обозначение $x = \ln t$, то плотность $p(\ln t)$ 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 $x = \ln t$ имеем

$$p(\ln t) = Ne^{k\beta \ln t}(1 - \alpha u e^{\beta \ln t})^{\frac{1}{u}-1}. \qquad (9)$$

Из формул (8) и (9) следует соотношение между плотностями $p(t)$ и $p(\ln t)$

$$tp(t) = p(\ln t). \qquad (10)$$

Из тех же формул следует также равенство функций распределения

$$F(\ln t) = \int p(\ln t) d\ln t = \int tp(t) d\ln t =$$

$$= \int tp(t)\frac{dt}{t} = F(t), \text{ т. е.}$$

$$F(\ln t) = F(t). \qquad (11)$$

Следовательно, плотность $p(t)$, приведенная к форме $tp(t) = f(\ln t)$, представляет собой плотность $p(x)$ и обладает всеми свойствами последней. 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 $rp(r) = f(\ln r)$ [5]. Тогда ранговое распределение будет иметь моду $\ln r_C$ и две точки перегиба $\ln r_A$ и $\ln r_B$, которые находятся на равных расстояниях от моды:

$$\ln r_C - \ln r_A = \ln r_B - \ln r_C.$$

Из последнего равенства следует **соотношение**

$$\frac{r_C}{r_A} = \frac{r_B}{r_C} = n, \qquad (12)$$

которое может быть принято в качестве закона рассеяния публикаций в смысле Бредфорда. На базе формулы (12) можно записать соотношение между количеством журналов от начала частотного списка до точек A, C, B

$$t_A : t_C : t_B = t_A(1 : n : n^2) \qquad (13)$$

и соотношение между количеством журналов в ядре и зонах рассеяния (при $t_Я = t_A$)

$$t_Я : t_I : t_{II} = t_Я[1 : (n - 1) : (n - 1)n]. \qquad (14)$$

Легко видеть, что закон Бредфорда является комбинацией из двух точных формул (13) и (14).

Плотность $p(t)$ позволяет вычислять величину n и координаты характерных точек.

Мода t_C находится из условия $dtp(t)/dlnt = 0$ и в общем случае для распределений $I - V$ типов равна [4]

$$t_C = \left(\frac{k}{\alpha(1 + ku - u)}\right)^{1/\beta}. \qquad (15)$$

Величина n задается формулой

$$n = \left[1 + \frac{1 - u + \sqrt{[4k(1+ku-u)+(1-u)](1-u)}}{2k(1+ku-u)}\right]^{1/\beta}. \quad (16)$$

Абсциссы точек перегиба вычисляются по формулам:

$$t_A = t_C/n;\ t_B = t_C \cdot n. \quad (17)$$

Доли статей в каждой зоне и для любого другого интервала рангов вычисляются с помощью функции распределения.

Поскольку плотность $p(x)$ при $u = 1$ не имеет характерных точек, то и закон Ципфа-Мандельброта (2), приведенный к виду $tp(t) = f(\ln t)$, тоже не имеет характерных точек (потому что этот закон имеет параметр $u = 1$). Отсюда следует вывод, что **на базе закона Ципфа-Мандельброта невозможно получить закон рассеяния публикаций Бредфорда.**

УНИВЕРСАЛЬНЫЙ ЗАКОН РАССЕЯНИЯ ПУБЛИКАЦИЙ

Обобщённые распределения позволяют решать все задачи, связанные с рассеянием журнальных публикаций, старением информации, а также множество других задач. Например, в формулировке закона Бредфорда утверждается, что в ядре журналов и в последующих законах количество статей одинаково, что не соответствует действительности. Кроме того, отсутствуют формулы для вычисления границ ядра и зон рассеяния; вычисления доли статей в каждой зоне и для любого интервала рангов; отсутствуют объективные критерии для разделения ранжированного ряда журналов на ядро и зоны рассеяния.

На все эти и другие вопросы дает ответы обобщенная плотность $p(t)$. Приведенная к форме плотности $p(x)$, при $u \le 1/2$ она имеет три характерные точки: моду C и две точки перегиба A и B. **Координаты этих точек приняты автором в качестве границ ядра и зон рассеяния,** что позволило дать математически точную формулировку закона рассеяния публикаций в смысле Бредфорда в виде двух формул (13) и (14). Ранее эти формулы были получены на базе закона Вейбулла [6], который является частным случаем второй системы непрерывных распределений и следует из обобщенной плотности $p(t)$ при $k = 1,\ u \to 0$.

Три характерные точки A, C, B делят все журналы в ранжированном ряду на четыре части: ядро и три зоны рассеяния. Количество журналов, входящих в ядро, определяется равенством $t_{\text{Я}} = t_A$. Количество журналов в первой зоне равно разности $t_I = t_C - t_A$; во второй зоне — $t_{II} = t_B - t_C$. Остальные журналы относятся к третьей зоне: $t_{III} > t_B$. При этом количество журналов от начала частотного списка до точки C в n раз больше количества журналов в ядре. Количество журналов до точки B в n раз больше их количества до точки C и в n^2 раз больше, чем в ядре. Отсюда следуют формулы (13), (14). Но они не содержат информации о количестве журналов и долях статей в каждой зоне, сколько может быть зон рассеяния, чему равна величина n. Вся эта информация содержится в обобщенной плотности $p(t)$, а также в двух других плотностях второй системы непрерывных распределений. Поэтому **вторая система непрерывных распределений по праву является универсальным законом рассеяния публикаций** [7].

Журналы, входящие в ядро, содержат долю статей, равную функции распределения $F(t_A) = \int_0^{t_A} p(t)dt$. Аналогично доля статей в журналах, входящих в ядро и первую зону рассеяния, составляет $F(t_C)$, и т. д. Следовательно, доля статей в первой зоне рассеяния составляет $F(t_C) - F(t_A)$; во второй зоне — $F(t_B) - F(t_C)$, а в третьей зоне — $1 - F(t_B)$.

Величина t_B дает оценку оптимального объёма фонда, а величина $F(t_B)$ — информационную полноту комплектования фонда объёмом t_B, т. е. вероятность удовлетворения информационных потребностей пользователей этим фондом. В то же время величина $F(t_B)$ — это доля книговыдач, доля статей и т. д., содержащихся в ядре и первых двух зонах рассеяния.

В заключение следует дать обоснование использования трех характерных точек для вывода математически точной формулировки закона рассеяния публикаций в смысле Бредфорда, а также для разделения ранговых и других распределений на зоны.

Точки A, C, B являются особыми точками кривой распределения, заданной плотностью $p(x)$. Между ординатами трех характерных точек существует соотношение [4]

$$\lambda = \frac{[p(x_C)]^2}{p(x_A)p(x_B)} = \left(\frac{1-2u}{1-u}\right)^{1-\frac{1}{u}} = \left(1 + \frac{1}{1-1/u}\right)^{1-\frac{1}{u}}, \quad (18)$$

из которого видно, что **показатель** λ зависит лишь от одного параметра формы u, т. е. он **является идентификатором типа кривой распределения.** В зависимости от значений показателя λ, распределения, заданные плотностью $p(x)$, можно разделить на типы. Для I, I′ типов $e < \lambda < \infty$ (при $0 < u < 1/2$); для II, II′ типов $\lambda = e$; для III типа $2 < \lambda < e$; для IV типа $\lambda = 2$; и, наконец, для V типа $1 < \lambda < 2$. Таким образом, по ординатам трех характерных точек A, C, B может быть однозначно установлен тип кривой распределения и найдена оценка параметра u из соотношения (18) при известном λ.

Анализ плотности $p(x)$ показывает, что у распределений *II–V*, *II′*типов существуют все три точки: A, C, B. У распределений *I* типа точка перегиба B существует при $0 < u < 1/2$, точки A, C — при $0 < u < 1$. У распределений *I′* типа точка перегиба A существует при $0 < u < 1/2$, точки B, C — при $0 < u < 1$.

Если кривая рангового распределения, приведенная к форме плотности $p(x)$, имеет три характерные точки, то для такого распределения существует ядро и три (не более!) зоны рассеяния.

УНИВЕРСАЛЬНЫЙ ЗАКОН СТАРЕНИЯ ПУБЛИКАЦИЙ

Закон старения публикаций заключается в том, что число ссылок на публикации в зависимости от

их года издания вначале резко растет, затем убывает с увеличением срока давности издания. Максимальное число ссылок приходится на публикации одно-двухлетней давности.

Для описания этого закона предлагалось множество математических моделей, но задача так и не была решена (по той же причине, что и в случае закона рассеяния публикаций, т. е. из-за отсутствия подходящего универсального распределения).

Исследования автора показали, что распределение числа ссылок на публикации в зависимости от года их издания хорошо описывается первой системой непрерывных распределений, в частности, обобщенной плотностью $p(x)$ [7, 8], где x — год издания. Если за начало отсчета принять текущий год ($x = 0$), то для предыдущего года будем иметь $x = -1$ и т. д. Обобщенная плотность распределения $p(x)$ обладает тем свойством, что значения случайной величины X могут быть как положительными, так и отрицательными.

Таким образом, **наиболее общим законом старения публикаций является первая система непрерывных распределений,** заданная тремя обобщенными плотностями (3). Обобщенные плотности позволяют наиболее точно описывать статистические распределения, вычислять накопленную долю ссылок на публикации по любому заданному интервалу времени их издания, вычислять координаты трех характерных точек, как и в случае закона рассеяния, а также вычислять другие показатели, интересующие исследователя.

Абсциссы трех характерных точек для плотности $p(x)$ задаются формулами (в случае распределений I–V типов)

$$x_c = \frac{1}{\beta}\ln\frac{k}{\alpha(1+ku-u)}, \tag{19}$$

$$x_{A,B} = x_c \mp \ln n, \tag{20}$$

где величина n рассчитывается по прежней формуле (16).

РАНГОВЫЕ РАСПРЕДЕЛЕНИЯ ЛЕКСИЧЕСКИХ ЕДИНИЦ

В случае однородной совокупности лексических единиц (слов, словосочетаний, терминов, дескрипторов) их ранговые распределения хорошо описываются третьей системой непрерывных распределений [8], которая задана тремя обобщенными плотностями (5). Для вычисления типа выравнивающей кривой и оценок ее параметров статистическое распределение необходимо привести к форме плотности $p(t)$, либо $p(x)$ и воспользоваться соответствующей компьютерной программой.

Характерные точки кривых распределения могут быть использованы как естественные границы различных зон лексических единиц.

ВЫЧИСЛЕНИЕ АППРОКСИМИРУЮЩИХ РАСПРЕДЕЛЕНИЙ ПО УСТОЙЧИВОМУ МЕТОДУ

Метод оценивания параметров аппроксимирующих распределений является устойчивым, если он не чувствителен к выбросам на концах статистического распределения.

Такой метод, разработанный автором [8], излагается ниже.

Рассмотрим обобщенную плотность $p(x)$, которая задает первую основную систему непрерывных распределений. Введем два показателя — асимметрии B и островершинности H, которые зависят от двух параметров формы k, u. По этим показателям устанавливается тип аппроксимирующего распределения и находятся оценки параметров k, u. Оценки двух других параметров рассчитываются по простым формулам. Заметим, что этот метод требует предварительного группирования статистических данных.

Для обобщенной плотности $p(x)$ показатели B, H равны

$$\left.\begin{aligned} B &= M[p(x)(x - M(x))] = f(k,u) \\ H &= S_3/S_1^3 = f(k,u) \end{aligned}\right\}, \tag{21}$$

где

$$S_r = M[p(x)]^r = f(\beta, k, u). \tag{22}$$

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

Вычислим для разных типов распределений значения показателей B, H при различных значениях параметров k, u. Далее построим номограмму (рис. 2). Она применима для трех основных систем непрерывных распределений, заданных первыми плотностями. При этом плотности $p(t)$ и $p(y)$ должны быть приведены к форме плотности $p(x)$, т. е. представлены в виде $tp(t) = f(lnt)$, $ylnyp(y) = f(lnlny)$.

На номограмме распределения II, II' и IV типов представлены кривыми. Типы I, I', III, V занимают определенные области. Симметричные распределения $IIIc$, Vc типов представлены отрезками на оси OH: для $IIIc$ типа $\sqrt{2} < H < \pi^2/6$; для Vc типа $\pi^2/6 < H < 2$. Распределения IVc типа представлены точкой $H = \pi^2/6$. Распределения IIc типа также представлены точкой $H = \sqrt{2}$.

На номограмме изображены области распределений с левосторонней асимметрией, для которых $0 < B_1 < 1/4$. Сюда относится часть распределений III–V типов при $0 < k < (1 - 1/u)/2$, а также распределения I, II типов. При этом распределения приведены к форме плотности $p(x)$.

Распределения I', II' типов, а также часть распределений III–V типов при $(1 - 1/u)/2 < k < 1/u$ имеют правостороннюю асимметрию. Для них $-1/4 < B < 0$, причем для распределений I, II и I', II' типов справедливы равенства: $B' = -B$, $H' = H$.

Здесь следует отметить, что для распределений с параметром сдвига l и параметром $\beta = 1$ автором построена другая номограмма, которая является продолжением настоящей. Здесь она не приводится.

Показатели B, H однозначно определяют тип распределения, приведенного к форме плотности $p(x)$. Более того, с помощью этих показателей могут быть найдены оценки параметров k, u непосредственно из номограммы.

Для распределений III–V типов при $B < 0$ из номограммы вначале находятся оценки параметров k', u (при $B > 0$), затем вычисляется величина $k = 1 - 1/u - k'$.

Номограмма для установления типа аппроксимирующего распределения и нахождения оценок параметров k, u

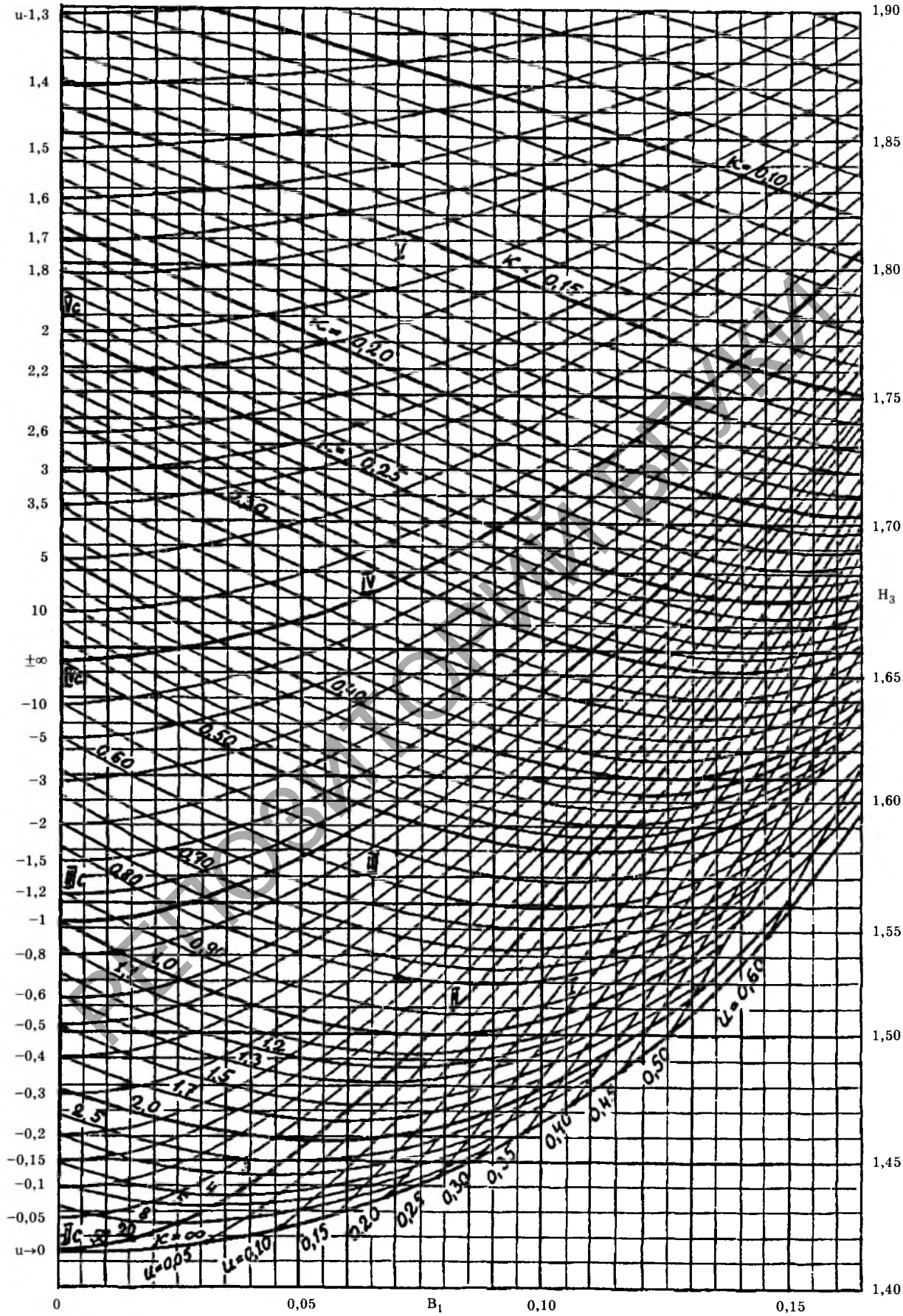

Рис. 2

Оценка параметра β для всех типов равна

$$\beta = \frac{S_1}{S_1^{(z)}}. \tag{23}$$

Оценки параметра α для распределений *II*, *II′* типов и произведения αu для остальных типов равны:

$$\left.\begin{array}{ll} \text{Типы I, I}': & \alpha u = e^{\pm}\left(\nu\right)_1^{(z)} - \beta\nu_1\right) \\ \text{Типы II, II}': & \alpha = e^{\pm}\left(\nu_1^{(z)} - \beta\nu_1\right) \\ \text{Типы III–V}: & \alpha u = -e^{\nu_1^{(z)}} - \beta\nu_1, \end{array}\right\}, \tag{24}$$

где в зависимости от типа распределения величины $\nu_1^{(z)}$ и $S_1^{(z)}$ рассчитываются по формулам:

Типы I, I′:

$$\left.\begin{array}{l} \nu_1^{(z)} = \pm\left[\Psi(k) - \Psi\left(k + \frac{1}{u}\right)\right] \\ S_1^{(z)} = \frac{1}{2\sqrt{\pi}} \frac{2\left(k + \frac{1}{u}\right) - 1}{\frac{2}{u} - 1} \cdot \frac{g(k)g\left(\frac{1}{u}\right)}{g\left(k + \frac{1}{u}\right)} \end{array}\right\}. \tag{25}$$

Типы II, II′:

$$\nu_1^{(z)} = \pm\Psi(k); \quad S_1^{(z)} = \frac{g(k)}{2\sqrt{\pi}}. \tag{26}$$

Типы III–V:

$$\left.\begin{array}{l} \nu_1^{(z)} = \Psi(k) - \Psi\left(1 - \frac{1}{u} - k\right) \\ S_1^{(z)} = \frac{1}{2\sqrt{\pi}} \frac{g(k)g\left(1 - \frac{1}{u} - k\right)}{g\left(1 - \frac{1}{u}\right)} \end{array}\right\}. \tag{27}$$

Величина

$$g(x) = \frac{\Gamma(x + 1/2)}{\Gamma(x)}$$

может быть вычислена по приближенной формуле:

$$g(x) \approx \frac{x(x + 0{,}875)}{(x + 0{,}5)\sqrt{x + 1}}. \tag{28}$$

Величина $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$, т. е. логарифмическая производная гамма-функции приближенно равна

$$\psi(x) \approx -\left(\frac{1}{x} + \frac{1}{x + 1}\right) + \ln(x + 2) - \frac{1}{2(x + 2)} - \frac{1}{12(x + 2)\hat{}2}. \tag{29}$$

Более точные формулы для вычисления величин $g(x)$, $\psi(x)$ и других приведены в монографии [4].

Для установления типа аппроксимирующего распределения и нахождения оценок параметров по устойчивому методу достаточно найти значения статистических показателей ν_1^*, S_1^*, B^*, H^* и приравнять их соответствующим теоретическим. Эти показатели для каждой системы непрерывных распределений вычисляются по-своему. Но номограмма применима ко всем трем системам непрерывных распределений.

Оценки статистических показателей в случае аппроксимирующих распределений, заданных плотностью $p(x)$, вычисляются по формулам:

$$\left.\begin{array}{l} \nu_1^* = \overline{x} = \sum\limits_{i=1}^{n} x_i p_i h_i \\ S_1^* = \sum\limits_{i=1}^{n} p_i^2 h_i, \quad S_3^* = \sum\limits_{i=1}^{n} p_i^4 h_i \\ B_1^* = \sum\limits_{i=1}^{n} x_i p_i^2 h_i - \nu_1^* S_1^*; \quad H^* = \frac{S_3^*}{(S_1^*)^3} \end{array}\right\}, \tag{30}$$

где $p_i = m_i/(Mh_i)$ — эмпирическая плотность распределения; m_i —наблюденная частота случайной величины X в i-м интервале ($i = 1, 2, \ldots, n$); $M = \sum\limits_{i=1}^{n} m_i$ — наблюденная частота во всех n интервалах (объем выборки); h_i — ширина i-го интервала; x_i — значение случайной величины X в середине i-го интервала.

В случае плотности $p(t)$ статистические показатели рассчитываются по формулам:

$$\left.\begin{array}{l} \nu_1^* = \overline{\ln t} = \sum\limits_{i=1}^{n} \ln t_i p_i h_i; \quad S_1^* = \sum\limits_{i=1}^{n} t_i p_i^2 h_i \\ S_3^* = \sum\limits_{i=1}^{n} t_i^3 p_i^4 h_i; \quad H^* = \frac{S_3^*}{(S_1^*)^3} \\ B^* = \sum\limits_{i=1}^{n} t_i \ln t_i p_i^2 h_i - \nu_i^* S_1^* \end{array}\right\}. \tag{31}$$

В случае плотности $p(y)$ — по формулам:

$$\left.\begin{array}{l} \nu_1^* = \overline{\ln\ln y} = \sum\limits_{i=1}^{n} (\ln\ln y_i) p_i h_i; \\ S_1^* = \sum\limits_{i=1}^{n} (y_i \ln y_i) p_i^2 h_i \\ S_3^* = \sum\limits_{i=1}^{n} (y_i \ln y_i)^3 p_i^4 h_i; \quad H^* = S_3^*/(S_1^*)^3 \\ B^* = \sum\limits_{i=1}^{n} (y_i \ln y_i)(\ln\ln y_i) p_i^2 h_i - \nu_1^* S_1^* \end{array}\right\}. \tag{32}$$

После вычисления эмпирических показателей B^*, H^* их следует приравнять теоретическим, с помощью номограммы установить тип аппроксимирующего распределения и найти оценки параметров k, u (по точке с координатами B, H). При известных оценках параметров k, u по приведенным выше формулам вычисляются оценки параметров β, α или произведения au. Далее остается вычислить нормирующий множитель, который вначале целесообразно прологарифмировать, затем вычислить lnN и значение $N = e^{\ln N}$. Логарифм гамма-функции можно вычислить по приближенной формуле

$$\ln\Gamma(x) \approx 0{,}91894 - \ln[x(x + 1)] + (x + 1{,}5)\ln(x + 2) - (x + 2) + \frac{1}{12(x + 2)}.$$

Подставив найденные оценки параметров в соответствующее распределение, можно рассчитать теоретические значения плотности при заданных значениях случайной величины.

ЗАКЛЮЧЕНИЕ

В итоге можно сделать вывод, что обобщенные распределения являются универсальными законами распределения не только теории вероятностей и математической статистики, но и информатики, математической лингвистики, экономики и других областей знания. При использовании обобщенных распределений исчезают ранее существовавшие барьеры на пути к новому знанию. Например, для нахождения наилучшей аппроксимирующей кривой не требуется выдвигать гипотезы о виде закона распределения. Система непрерывных распределений выбирается в зависимости от свойств случайной величины, а тип распределения и оценки параметров вычисляются по статистическому распределению. При этом вычисленная кривая распределения является наилучшей (разумеется, для принятого метода оценивания параметров). В случае однородности статической совокупности оба метода — универсальный метод моментов и устойчивый метод — дают близкие оценки параметров аппроксимирующего распределения. Наиболее точные оценки параметров получаются в случае симметричного или близкого к нему статистического распределения, приведенного к форме плотности $p(x)$.

Универсальные законы старения и рассеяния публикаций, а также ранговые распределения лексических единиц, заданные соответственно первой, второй и третьей системами непрерывных распределений, являются фундаментальными закономерностями информатики, математической лингвистики и библиотековедения.

СПИСОК ЛИТЕРАТУРЫ

1. Zipf G. K. Human behavior and the principle of least effort.— Cambridge, 1949.— 140 p.

2. Мандельброт Б. О. О рекуррентном кодировании, ограничивающем влияние помех // Теория передачи сообщений.— М., 1957.— С. 139–157.

3. Bradford S. C. Documentation.— London, 1948.— 156 p.

4. Нешитой В. В. Элементы теории обобщенных распределений.— Минск: РИВШ, 2009.— 204 с.

5. Нешитой В. В. Форма представления ранговых распределений // Ученые записки Тартуского гос. ун-та.— 1987.— Вып. 774.— С. 123–134.

6. Петренко Б. В., Нешитой В. В. Применение закона Вейбулла для расчета полноты комплектования справочно-информационного фонда // Проблемы оптимального комплектования и использования справочно-информационного фонда для принятия решений / Общ-во "Знание" Укр. ССР.— Киев, 1974.— С. 6–8.

7. Нешитой В. В. Универсальные законы рассеяния и старения публикаций // Веснік Беларус. дзярж. ун-та культуры і маст.— 2007.— № 8.— С. 128–133.

8. Нешитой В. В. Исследование статистических закономерностей текста и информационных потоков: дис. ... д-ра техн. наук.— Минск, 1987.— 505 с.

Материал поступил в редакцию 30.10.09.