РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 11 Москва 2013

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [001.102:002] : 004

В.В. Нешитой

Методы вычисления границ ядра и зон рассеяния публикаций

Рассматриваются три метода вычисления границ ядра и зон рассеяния публикаций, а точнее – координат трех характерных точек A, C, B на кривой распределения: аналитический, графический и метод наименьших квадратов. Первые два метода могут быть применены к любому статистическому ранговому распределению в случае однородной выборки. Такие распределения описываются второй системой непрерывных распределений, которая является универсальным законом рассеяния публикаций. В частном случае, если справедлив закон Вейбулла, используется метод наименьших квадратов.

Приводятся практические примеры, подтверждающие высокую точность аппроксимации статистических ранговых распределений второй системой непрерывных распределений автора.

Ключевые слова: *три характерные точки кривых распределения, границы ядра и зон рассеяния, аналитический метод, графический метод, метод наименьших квадратов, вторая система непрерывных распределений, форма представления ранговых распределений, универсальный закон рассеяния публикаций*

ВВЕДЕНИЕ

В научных исследованиях широко используются статистические ранговые распределения различных объектов – журналов, книг и других изданий, представленных в списке по убыванию частоты встречаемости; лексические единицы частотного словаря; ключевые слова, упорядоченные по убыванию частоты их использования при индексировании документов, и многие другие объекты. Порядковый номер журнала, книги, слова в этом списке называется рангом. На базе статистических ранговых распределений

РЕПОЗИТОРИЙ БГУКИ

ть множество практических задач, в том
лять границы ядра и зон рассеяния пуб-
сожалению, анализ научных работ с ис-
м ранговых распределений показывает,
ые авторы не могут вычислить теорети-
и и, следовательно, границы ядра и зон
Это можно осуществить лишь при ис-
теории обобщенных распределений.
С. Бредфорд окончательно сформулиро-
ссеяния журнальных публикаций, кото-
ется в следующем: «Если научные жур-
ложить в порядке убывания числа
в них статей по какому-либо заданному
в полученном списке можно выделить
ов, посвященных непосредственно этому
несколько групп или зон, каждая из ко-
жит столько же статей, что и ядро. Тогда
лов в ядре и последующих зонах будут
ак $1 : n : n^2$ » [1, с. 93, 94].
з этой формулировки неясно, как по ста-
у ранговому распределению вычислить
а и зон рассеяния, поскольку нет ника-
для их вычисления. Неизвестно также,
ическое ранговое распределение принято
кона рассеяния, какие точки на графике
аспределения приняты в качестве таких
ько может быть зон рассеяния, как вы-
еличина n. В формулировке С.Бредфорда
вается, что можно выделить ядро журна-
лько зон, при этом предполагается, что
в каждой зоне такое же, как и в ядре.
предположение не согласуется с факти-
ными и не обосновывается теоретически.

ДБОРА

формулировку С.Бредфорда за основу,
ли вынуждены по своему усмотрению
ть объем ядра журналов, а количество
я и их размер получаются на основании
данных из условия, что каждая зона со-
ько же статей, что и ядро. При таком
ичество зон рассеяния для одного и того
ческого рангового распределения у раз-
вателей может сильно колебаться.
в законе С.Бредфорда из того, что не
омнению – это расположение журналов
о количества опубликованных в них ста-
ному предмету, т.е. ранжирование. Ран-
еделения с наибольшей полнотой отра-
рассеяния публикаций. Но многие
ли обходят этот главный вопрос сторо-
изучения свойств статистических ранго-
елений и теоретического закона распре-
такими же свойствами они пытаются
ро и зоны рассеяния, используя только
ку С. Бредфорда 1948 года!
нию, в литературных источниках не при-
тистические ранговые распределения на-
книг, упорядоченных по убыванию час-
речаемости, недостаточно активно
работа по составлению частотных спи-
ов, упорядоченных по различным при-
знакам. Но для изучения свойств ранговых распределений можно обратиться к математической лингвистике. В этой области знания накоплено большое количество частотных словарей, в которых лексические единицы расположены в порядке убывания (точнее, невозрастания) частоты их употребления в текстах. Выработана удобная для анализа таблица частот слов, в которой приводятся необходимые сведения о частотном словаре: ранг или интервал рангов слов, их абсолютная частота, количество слов с данной частотой, накопленная абсолютная частота, относительная частота, накопленная относительная частота слов. Такая таблица приводится, как правило, в конце частотного словаря. Она позволяет представить в компактной форме частотную структуру словаря любого объема.

Для аппроксимации статистических распределений, в том числе ранговых, созданы универсальные математические модели – обобщенные распределения, разработаны методы вычисления законов распределения по статистическим данным и оценок их параметров. Предложена новая форма представления ранговых распределений, а на ее основе – критерии однородности и достаточности объема выборки. Имеются компьютерные программы для построения частотных словарей, в том числе свободно распространяемые.

Это значит, что многие **задачи по обработке статистических ранговых распределений в лингвистике, информатике, библиотечном деле давно решены**. Но, к сожалению, накопленные в этих областях знания результаты научных исследований практически не используются. Причина заключается в том, что исследователю необходимо затратить немалый труд и время на изучение и практическое применение этих результатов в своей научной работе. Значительно проще традиционно выдвигать гипотезы и проверять их по различным критериям согласия, но таким путем нельзя получить новое знание.

Упомянутый выше метод подбора размеров ядра и зон рассеяния – это тот же метод выдвижения гипотез, но без проверки результатов по критериям согласия. Такой метод в научных исследованиях неприемлем.

Аналитический метод

Итак, задано ранговое распределение журналов. Если найти теоретическое распределение, которое достаточно точно описывает статистическое ранговое распределение, то оно позволит дать математически точную формулировку закона рассеяния в смысле С.Бредфорда, т.е. этот закон должен следовать из рангового распределения как частный случай. Наиболее же общим, **универсальным законом рассеяния может быть только теоретическое ранговое распределение**, поскольку закон распределения является наиболее полной характеристикой любой случайной величины.

Многие авторы при определении границ ядра и зон рассеяния пытаются обрабатывать статистические ранговые распределения без предварительного решения общей задачи – нахождения закона распределения. Но в этом случае им приходится для аппроксимации каждого статистического распределе-

РЕПОЗИТОРИЙ БГУКИ

ния подбирать различные формулы вместо использования одного теоретического закона, но с разными значениями параметров. Кроме того, для подбора теоретической кривой используется неудачная форма представления статистических ранговых распределений в системе координат «логарифм ранга – логарифм частоты», которая несет слишком мало информации о ранговом распределении и более того – не имеет вероятностного смысла.

Таким образом, проблема рассеяния публикаций состоит в решении общей задачи – нахождении такой универсальной формулы или системы формул, которые способны с высокой точностью аппроксимировать все многообразие статистических ранговых распределений. Поскольку «…рассеяние научной информации является краеугольным камнем всей научно-информационной деятельности, а изучение этого свойства научной информации – важнейшей проблемой информатики» [1, с. 93], то эту проблему необходимо разрешать весьма серьезными средствами. Такими средствами являются обобщенные распределения, предложенные автором [2–6].

Рассмотрим первую и вторую системы непрерывных распределений, каждая из которых задана тремя обобщенными плотностями. Запишем первые плотности этих систем:

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

$$\mathrm{p(t)} = \mathrm{Nt}^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}. \qquad (2)$$

Здесь α, β, k, u – параметры распределения. Они вычисляются по статистическому распределению. N – нормирующий множитель, который выражается через параметры распределения при условии, что площадь под кривой распределения равна единице.

Первая плотность обладает тем замечательным свойством, что при значениях параметра формы $u \le 1/2$ она **имеет моду** x_C, т.е. такое значение случайной величины X, при котором плотность $p(x)$ максимальна, и **две точки перегиба** – x_A, x_B, **расположенные на равных расстояниях по обе стороны от моды**. Это такие точки на графике плотности распределения, которые отделяют выпуклую часть кривой от вогнутой или вогнутую часть от выпуклой. При $1/2<u<1$ имеются лишь две характерные точки – x_A и x_C (для распределений с левосторонней асимметрией). При $u \ge 1$ характерных точек не существует. Распределение (1) может быть задано на всей числовой оси, т.е. $-\infty < x < \infty$.

Распределение (2) задано на положительной полуоси $t > 0$. График плотности $p(t)$ может принимать различные формы. При определенных значениях параметров формы k, β, u график плотности $p(t)$ может иметь вид убывающей кривой распределения, а это значит, что плотность $p(t)$ описывает не только одновершинные статистические распределения, но и ранговые (убывающие).

Для вычисления закона распределения случайной величины по статистическому распределению, в том числе ранговому, нами разработаны два метода – универсальный метод моментов и общий устойчивый метод [4–6]. С целью упрощения расчетов и извлечения новой информации из рангового распределения оно приводится к форме одновершинного распределения [7]. Это достигается путем приведения формы второй плотности к форме первой. Для этого умножим левую и правую части плотности $p(t)$ на t, а величину t^{β} представим в виде $e^{\beta \ln t}$. Тогда плотность (2) примет вид

$$tp(t) = Ne^{k\beta \ln t}\left(1-\alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (3)$$

Сравнивая (1) и (3), можем записать:

$$tp(t) = p(x),\ lnt = x. \qquad (4)$$

Таким образом, ранговое распределение (2), приведенное к форме $tp(t) = f(lnt)$, с учетом равенств (4) представляет собой распределение (1) и обладает всеми его свойствами: оно имеет моду $\ln t_C$ и две точки перегиба $\ln t_A$ и $\ln t_B$. В этом заключается новая информация о ранговом распределении. **Примем абсциссы точек *A, C, B* обобщенных распределений в качестве границ ядра и зон рассеяния** различных объектов.

Поскольку точки $\ln t_A$ и $\ln t_B$ расположены на равных расстояниях от моды $\ln t_C$, то можем записать равенство $\ln t_C - \ln t_A = \ln t_B - \ln t_C$, откуда имеем $\ln(t_C / t_A) = \ln(t_B / t_C)$, или

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \qquad (5)$$

Формула (5) может служить уточненной формулировкой закона рассеяния публикаций в смысле С. Бредфорда, хотя она представляет собой новую формулировку закона рассеяния публикаций.

Из (5) можно записать еще две формулы:

$$\mathrm{t_A} : t_C : t_B = t_A(1 : n : n^2), \qquad (6)$$

$$t_A : t_I : t_{II} = t_A\left[1 : (n-1) : n(n-1)\right]. \qquad (7)$$

Здесь t_A, t_C, t_B – количество журналов от начала частотного списка до точек *A, C, B;* $\mathrm{t_I}, t_{II}$ – количество журналов в первой и второй зонах рассеяния. При этом $\mathrm{t_I} = t_C - t_A$, $t_{II} = t_B - t_C$. С учетом этих равенств получена формула (7) из формулы (6).

Все три формулировки, заданные формулами (5) – (7), отличаются от формулировки С. Бредфорда и уточняют ее. Однако **приведенные формулы** (как и закон Бредфорда) **не являются законом рассеяния**, так как не дают полной информации о нем. Они устанавливают лишь общие соотношения между абсциссами трех характерных точек – моды и двух точек перегиба. Но для вычисления координат этих точек требуется знание закона рангового распределения. Только при известных оценках параметров этого распределения можно получить нужную информацию – вычислить границы ядра журналов и зон рассеяния, долю статей в ядре и зонах рассеяния, а также долю статей для любого числа журналов от начала частотного списка. Она выражается через функцию распределения. В формулах же (5) – (7) функция распределения не задействована.

РЕПОЗИТОРИЙ БГУКИ

бразом, наиболее полную информацию
ангового распределения. Это значит, что
плотность (2), а точнее, **вторая система
ых распределений, заданная тремя
ми плотностями, является универ-
аконом рассеяния публикаций**. Здесь
етить, что первая система непрерывных
ий, включающая плотность (1), является
ым законом старения публикаций [2].
ые исследователи утверждают, что закон
а – это другая форма представления за-
Ципфа – $p_r = k/r$. Эта формула была
им для описания ранговых распределе-
стотного словаря. Здесь p_r – относитель-
слова с рангом r, k – параметр. Для зако-
фа произведение относительной частоты
г равно постоянной величине k, что на
истеме координат $rp_r = f(\ln r)$ изобра-
зонтальной прямой, на которой нет ника-
ных точек. Это убедительно свидетель-
ом, что законом Дж. Ципфа нельзя
овать статистические ранговые распре-
скольку в данной системе координат в
родной выборки они имеют вид одно-
кривой с двумя точками перегиба. Такие
ысокой точностью описываются второй
прерывных распределений. Здесь следует
го закон Дж.Ципфа, а в более общей
ке, с двумя дополнительными парамет-
н Эсту-Ципфа-Мандельброта, – является
учаем обобщенного распределения (2)
< 0. Но при этом условии кривая распре-
$_r = f(\ln r)$ не имеет характерных точек.
коим образом нельзя из них получить за-
ия, нельзя их использовать для аппрок-
атистических ранговых распределений,
серьезных научных исследованиях. Наи-
етическое распределение для аппрокси-
истического рангового распределения
ъ вычислено по второй системе непре-
пределений [6].
им далее закон рассеяния С.Бредфорда
$= t_A$)

$$t_я : t_1 : t_{II} = t_я(1 : n : n^2). \qquad (8)$$

ия эту формулу с формулами (6) и (7), ви-
кон рассеяния С.Бредфорда является не-
комбинацией двух правильных формул:
соответствует формуле (7), а правая –

(8) можно представить в другом виде – с
ех характерных точек, о которых
в своей формулировке закона рассеяния
. Из (8) имеем

$$: (t_C - t_я) : (t_B - t_C) = t_я(1 : n : n^2).$$

в левой части этого равенства величину

$$t_C - t_я)/t_я : (t_B - t_C)/t_я] = t_я(1 : n : n^2).$$

можем записать два равенства:
$= n$, $(t_B - t_C)/t_я = n^2$, откуда находим,
что отношение $t_C : t_я = n + 1$, а отношение $t_B : t_я = n^2 + n + 1 = (n+1)^2 - n$. В результате имеем формулу

$$t_я : t_C : t_B = t_я[1 : (n+1) : (n^2 + n + 1)]. \qquad (9)$$

В нашей точной формуле (6) первое отношение равно n, второе n^2. Поскольку эти отношения в формуле (9) не соблюдаются, то по закону С.Бредфорда точки перегиба на кривой рангового распределения $tp(t) = f(\ln t)$ должны располагаться **на разных** расстояниях от моды $\ln t_C$, а это противоречит свойствам статистических и теоретических ранговых распределений. Другое противоречие закона С.Бредфорда состоит в утверждении, что каждая зона рассеяния содержит такое же число статей, как и в ядре. Естественно, что этим двум требованиям не может удовлетворить ни одно теоретическое распределение. Поэтому все попытки усовершенствовать закон рассеяния С.Бредфорда при сохранении присущих ему противоречий оказались неудачными. И это закономерно, потому что предпринимались попытки решить частную задачу без предварительного решения общей задачи – нахождения такого теоретического рангового распределения, которое позволило бы с высокой точностью аппроксимировать широкое разнообразие статистических ранговых распределений.

Оценим погрешность формулы (9). Пусть величина n=5. Вычислим отношения $t_C : t_я$ и $t_B : t_я$. В первом случае оно равно n+1=6 , а во втором n^2+n+1=31. По точной формуле (6) имеем соответственно 5 и 25. Погрешность вычисления абсциссы точки С составила 20%, а точки В – 24%. Размер ядра в обоих случаях принят одинаковым, так как по закону С.Бредфорда его вычислить нельзя. Следует отметить, что с ростом величины n эта погрешность уменьшается.

Из полученных результатов следует, что закон рассеяния С.Бредфорда, представленный в виде формул (8) и (9), относительно близок к точным формулам (6) и (7), хотя при его формулировке им не были использованы теоретические ранговые распределения. Утверждение же С.Бредфорда о том, что число статей в ядре и зонах рассеяния одинаково, не соответствует действительности и вводит в заблуждение некоторых исследователей.

Однако огромная заслуга С.Бредфорда заключается в том, что он первым обратил внимание на явление рассеяния публикаций и побудил многих исследователей к углубленному изучению этого явления.

Вопрос этот оказался очень сложным. Универсальный закон рассеяния был найден нами лишь после разработки теории обобщенных распределений. Выше отмечалось, что таким законом является вторая система непрерывных распределений. Из обобщенной плотности (2) выводятся формулы для вычисления границ ядра и зон рассеяния.

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

$$t_C = \left(\frac{k}{\alpha(1 + ku - u)} \right)^{1/\beta}. \qquad (10)$$

РЕПОЗИТОРИЙ БГУКИ

Величина n задается формулой

$$n=\left[1+\frac{1-u+\sqrt{\left[4k\left(1+ku-u\right)+\left(1-u\right)\right]\left(1-u\right)}}{2k\left(1+ku-u\right)}\right]^{1/\beta}. \quad (11)$$

Абсциссы точек перегиба вычисляются по формулам:

$$t_A = t_C / n;\ t_B = t_C \cdot n. \quad (12)$$

Доли статей в каждой зоне и для любого другого интервала рангов вычисляются с помощью функции распределения или по статистическому ранговому распределению.

Из формул (11), (12) следует, что при заданных значениях параметров формы k, u с уменьшением параметра β величина $n = t_C / t_A = t_B / t_C$ растет. Это значит, что кривая распределения $tp(t) = f(\ln t)$ становится более широкой и пологой.

Методы вычисления аппроксимирующих четырехпараметрических распределений изложены в ряде работ [2–6], но эти методы рассчитаны на подготовленного исследователя.

Графический метод

Для нахождения границ ядра и зон рассеяния по статистическому ранговому распределению без предварительного вычисления теоретического закона можно предложить простой графический метод, который следует из анализа свойств обобщенных распределений (1) и (2). Суть этого метода покажем на примере.

Рассмотрим для примера закон Вейбулла, функция и плотность распределения которого заданы формулами

$$F(t) = 1 - e^{-\alpha t^{\beta}}, \quad (13)$$

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}. \quad (14)$$

При значениях параметра $\beta \le 1$ плотность (14) может с высокой точностью описывать некоторые статистические ранговые распределения.

Чтобы получить из рангового распределения полезную информацию, преобразуем плотность (14) к форме *tp(t) = f(lnt)* [7]:

$$tp(t) = \alpha\beta t^{\beta} e^{-\alpha t^{\beta}} = \alpha\beta e^{\beta \ln t} e^{-\alpha e^{\beta \ln t}} \quad (15)$$

С учетом равенств *tp(t) = p(x); lnt = x* плотность (15) примет вид

$$p(x) = \alpha\beta e^{\beta x} e^{-\alpha e^{\beta x}}. \quad (16)$$

Полученная формула представляет собой частный случай плотности (1) при *u→0, k=1*. Плотность (16) дает новую информацию о ранговом распределении. График этой плотности, т.е. кривая распределения содержит три характерные точки – моду *C* и две точки перегиба *A* и *B*, расположенные на равных расстояниях по обе стороны от моды. Найдем абсциссы этих точек. Продифференцируем плотность (16) по *x* и приравняем первую производную нулю. Из полученного уравнения найдем выражение для моды

$$x_C = \frac{1}{\beta}\ln\frac{1}{\alpha}. \quad (17)$$

В точках перегиба вторая производная равна нулю. Из этого условия для плотности (16) найдем

$$x_A = x_C - \frac{1}{\beta}\ln\frac{3+\sqrt{5}}{2},$$

$$x_B = x_C + \frac{1}{\beta}\ln\frac{3+\sqrt{5}}{2}.$$

Введем обозначение

$$n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \quad (18)$$

С учетом (18) последние две формулы примут более простой вид

$$x_A = x_C - \ln n; \quad (19)$$

$$x_B = x_C + \ln n. \quad (20)$$

Переходя к распределению Вейбулла, из формул (19), (20) с учетом равенства x=lnt найдем

$$\ln t_C - \ln t_A = \ln t_B - \ln t_C = \ln n,$$

откуда $\ln(t_C / t_A) = \ln(t_B / t_C) = \ln n$, или

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \quad (21)$$

Здесь

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}};\quad t_A = \frac{t_C}{n};\quad t_B = t_C \cdot n. \quad (22)$$

Формула (21), полученная на базе закона Вейбулла, совпадает с аналогичной формулой (5), полученной на базе четырехпараметрического распределения (2). Она остается также справедливой для любого частного случая распределения (2) при значениях параметра формы $u \le 1/2$. Закон Вейбулла является частным случаем распределения (2) при *u→0, k=1*. В некоторых случаях он с высокой точностью описывает статистические ранговые распределения. Но универсальным законом остается обобщенная плотность (2), которая позволяет вычислять и координаты характерных точек, и функцию распределения [4, с. 155–160] практически для любого статистического рангового распределения. В некоторых случаях статистические ранговые распределения с высокой точностью описываются дополнительными плотностями второй системы непрерывных распределений.

Предположим для определенности, что ранговое распределение задано законом Вейбулла с параметрами *α=0,1; β=0,5*. Приведем его к плотности *p(x)*, которая представлена формулой (16). Тогда для этой плотности по формулам (17)–(20) найдем: n=6,8541; $x_C = 4,6052$; $x_A = 2,6804$; $x_B = 6,53$.

Рассчитаем по формуле (16) значения плотности *p(x)* с интервалом *Δx=0,5* и сведем результаты в таблицу 1.

Построим график плотности *p(x)*, т.е. кривую распределения (рис. 1а).

На кривой распределения абсциссу точки *C* легко найти графически путем проведения горизонтальной касательной к кривой (см. рис. 1а).

РЕПОЗИТОРИЙ БГУКИ

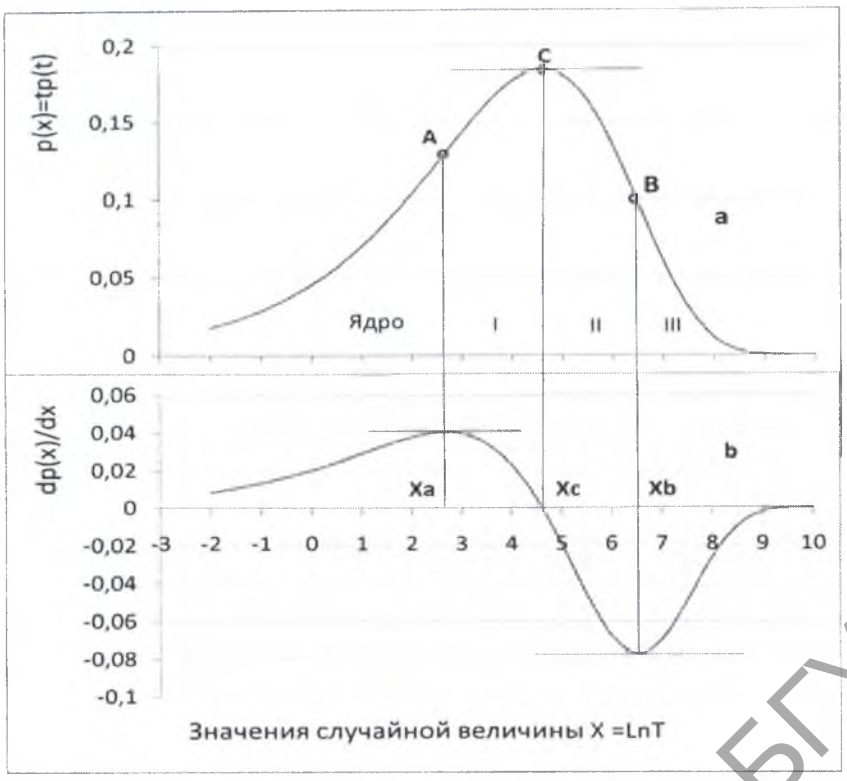

Рис. 1. Графики плотности распределения (a) и ее первой производной (b)

Таблица 1

Значения плотности и тангенса угла наклона касательной к кривой в серединах интервалов

x	p(x)	dp(x)/dx	x	p(x)	dp(x)/dx
-2	0,01773	0,008539	4	0,176464	0,023037
-1,5	0,022529	0,010732	4,5	0,18369	0,004705
-1	0,028542	0,013405	5	0,180147	-0,01966
-0,5	0,036022	0,016609	5,5	0,163655	-0,04617
0	0,045242	0,020359	6	0,134756	-0,06795
0,5	0,056465	0,024607	6,5	0,097806	-0,07722
1	0,069906	0,02919	7	0,060369	-0,06977
1,5	0,085655	0,033761	7,5	0,030263	-0,04921
2	0,103565	0,037706	8	0,011614	-0,0259
2,5	0,123099	0,040067	8,5	0,003163	-0,00951
3	0,143144	0,039496	9	0,000554	-0,00222
3,5	0,161833	0,034352	9,5	5,52E-05	-0,00029

ждения абсцисс точек перегиба воспольсвойством кривой распределения, что в *B* первая производная принимает экстречения: в точке *A* она имеет максимум, а в инимум. Вычислим тангенс угла наклона ивой к горизонтальной оси на всех интеротношение разности между значениями *(x)* на границах интервала к ширине ин-Другими словами, найдем приближенные рвой производной в серединах интервалов (в табл. 1 приведены расчетные значения производной dp(x)/dx) и построим график (см. рис. 1b).

Первая производная *dp(x)/dx* имеет максимум в точке x_A и минимум в точке x_B. Эти точки легко определить путем проведения горизонтальных касательных к кривой на рис. 1b.

Из построенного графика можно приближенно найти абсциссы трех характерных точек: $x_A = 2,7$; $x_C = 4,6$; $x_B = 6,5$.

РЕПОЗИТОРИЙ БГУКИ

Переходя к ранговому распределению, находим: $t_A = \exp(x_A) \approx 15$; $t_C = \exp(x_C) \approx 99$; $t_B = \exp(x_B) \approx 665$. Точные значения для распределения Вейбулла равны: $t_A = 14,59$; $t_C = 100$; $t_B = 685$.

Таким простым методом приближенно могут быть найдены абсциссы трех характерных точек любого статистического рангового распределения без предварительного вычисления теоретического закона распределения, но с использованием его свойств.

Значения функции распределения *F(t)* при любом заданном значении ранга *t*, в том числе в трех характерных точках, могут быть вычислены по статистическому ранговому распределению.

Используя графический метод, разные исследователи получат близкие результаты по определению границ ядра и зон рассеяния для одного и того же статистического рангового распределения.

Здесь необходимо отметить, что представленная на рис. 1а кривая теоретического распределения плавно возрастает до максимального значения и затем плавно убывает. Поэтому вычисление тангенса угла наклона отрезков кривой на всех интервалах не вызывает затруднений. Аналогичная кривая статистического распределения *tp(t) = f(lnt)* имеет многочисленные всплески и впадины, что затрудняет построение рис. 1b. Поэтому предварительно ее необходимо сгладить, например, с помощью лекала.

Отметим, что на рис. 1а горизонтальная касательная к кривой распределения *tp(t) = f(lnt)* в точке С представляет собой закон Дж.Ципфа. Отсюда следует, что этим законом невозможно описать никакое ранговое распределение.

Метод наименьших квадратов

В некоторых случаях статистическое ранговое распределение может с высокой точностью описываться законом Вейбулла, функция распределения и плотность вероятностей которого заданы формулами (13) и (14). Этот закон впервые использовал Г.Г. Белоногов для описания рангового распределения слов частотного словаря [8]. Поскольку этот закон весьма простой, его целесообразно проверять в первую очередь при отыскании подходящего рангового распределения. Для такой проверки функцию распределения необходимо преобразовать к линейному виду

$$\ln\ln(1/(1-F(t))) = \ln\alpha + \beta \ln t . \qquad (23)$$

Введем обозначения:

$$Y = \ln\ln(1/(1-F(t))), \quad X = \ln t . \qquad (24)$$

Тогда последнее уравнение запишется в виде

$$Y = \ln\alpha + \beta X . \qquad (23')$$

Для проверки применимости закона Вейбулла необходимо по статистической функции распределения вычислить значения X, Y по формулам (24) и построить график зависимости Y=f(X). Если эмпирические точки расположатся вдоль прямой (23′), то далее по методу наименьших квадратов следует вычислить оценки параметров α и β этой прямой:

$$\beta = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \left(\overline{X}\right)^2}, \quad \alpha = \exp\left(\overline{Y} - \beta\overline{X}\right). \qquad (25)$$

Для оценки тесноты линейной связи между переменными Y, X вычисляется выборочный коэффициент корреляции

$$R_{y/x} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y}, \qquad (26)$$

где средние квадратические отклонения σ_x, σ_y равны:

$$\sigma_x = \sqrt{\overline{X^2} - (\overline{X})^2}, \quad \sigma_y = \sqrt{\overline{Y^2} - (\overline{Y})^2} .$$

При этом

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i , \quad \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i ,$$

$$\overline{X^2} = \frac{1}{N}\sum_{i=1}^{N} X_i^2 , \quad \overline{Y^2} = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 ,$$

где N – количество значений случайных величин X, Y.

Абсциссы точек А, С, В для закона Вейбулла вычисляются по формулам (18), (22). Значения функции распределения в этих точках при любых значениях параметров α и β соответственно равны:

$$F(t_A) = 0,31748; \; F(t_C) = 0,63212; \; F(t_B) = 0,92705. \qquad (27)$$

Отметим, что статистические данные расположатся вдоль прямой лишь в случае однородной выборки, для которой справедлив закон Вейбулла. Однако, если попытаться описать этим законом ранговое распределение слов частотного словаря, то окажется, что первые 50–100 наиболее частых слов не подчиняются закону Вейбулла. Это в основном служебные слова. Они составляют неоднородную часть выборки. Поэтому для более точного описания таких ранговых распределений можно предварительно удалить первые 50–100 слов с последующим пересчетом рангов и относительных частот слов, получив, таким образом, однородную выборку. Если же имеется необходимость аппроксимировать ранговое распределение всех слов частотного словаря, включая служебные, то следует ввести дополнительный параметр в теоретическое распределение. Наши исследования в свое время привели к выводу, что закон Вейбулла с учетом третьего параметра (обозначим его δ) можно представить в следующем виде [9]

$$F(t) = 1 - e^{-\alpha[(t+1)^{\beta} - e^{-\delta t}]}, \qquad (28)$$

$$p(t) = \frac{\alpha\beta(t+1)^{\beta-1} + \alpha\delta e^{-\delta t}}{e^{\alpha[(t+1)^{\beta} - e^{-\delta t}]}}. \qquad (29)$$

Параметры α, β могут быть вычислены по методу наименьших квадратов по формулам (23)–(25) для рангов слов частотного словаря от 50–100 до рангов слов с частотой 2–3. Далее при известных оценках этих параметров вычисляется дополнительный параметр δ по формуле, которая следует из функции распределения (28):

$$\delta = -\frac{1}{t}\left[\ln\left((t+1)^{\beta} - \frac{1}{\alpha}\ln\frac{1}{1-F(t)}\right)\right]. \qquad (30)$$

Его можно вычислить один раз при заданной относительной частоте самого частого слова, которая равна функции распределения F(t=1).

Двухпараметрическим законом Вейбулла хорошо описываются некоторые ранговые распределения журналов, терминов, ключевых слов, образующих

РЕПОЗИТОРИЙ БГУКИ

ки однородные выборки, а трехпараметнекоторые неоднородные выборки.

им для примера *ранговое распределение тного словаря современного русского язы*построен на выборке огромного размера – оупотреблений. Количество разных лексем – лемм) составило 739930 единиц. Из них лемм с частотой 2 и более раза составило тотой 3 и более раза – 268106. В указанном иводится частотный список первых 20000 озволяет вычислить накопленные относиоты, т.е. функцию распределения для всех о 20000. При известном количестве лемм с потребления 1 и 2 раза (соответственно 930 – 360755 и 92649 = 360755 – 268106) лнительно вычислить два значения функеления:

=1–379175/135000000=0,9971913;

1–(379175+92649*2)/135000000=0,9958187.

суммарной доли употреблений всех й единице, вычитается доля употребле частотой один раз – в первом случае и аза – во втором.

Составим на базе Частотного словаря табл. 2, где приведем отдельные ранги слов ($R \geq 80$) и соответствующие этим рангам значения функции распределения (см. первые два столбца). Вычислим далее по методу наименьших квадратов оценки параметров закона Вейбулла, а также коэффициент корреляции. Параметр β оказался равным 0,309427, параметр α=0,111757. Коэффициент корреляции $R_{y/x}$=0,999789. Это значит, что эмпирическая зависимость оказалась близкой к теоретической прямой (23), которая представлена на рис. 2.

При известных оценках параметров α, β и функции распределения F(t=1)=0,035802 по формуле (30) найдем значение третьего параметра, который необходим для более точного описания наиболее частых слов: δ=0,091. Вычислим далее значения функции распределения по трехпараметрическому закону Вейбулла (см. табл. 3) и построим в полулогарифмическом масштабе график функции распределения (см. рис. 3) с учетом служебных слов, где отдельными точками показана эмпирическая функция распределения, сплошной линией – теоретическая.

Таблица 2

Расчет параметров закона Вейбулла по статистическому распределению

Функция распределения	lnr	$\ln\ln\frac{1}{1-F(r)}$					
F(r)	X	Y	XY	X^2	Y^2	Yрасч.	Fрасч.
0,354324	4,382027	-0,82678	-3,62295	19,20216	0,683558	-0,83551	0,351864
0,374545	4,60517	-0,75656	-3,48411	21,20759	0,57239	-0,76646	0,371648
0,411675	5,010635	-0,63398	-3,17665	25,10647	0,401932	-0,641	0,409488
0,458992	5,521461	-0,48724	-2,69026	30,48653	0,2374	-0,48294	0,460423
0,506106	5,991465	-0,34894	-2,09067	35,89765	0,12176	-0,3375	0,510098
0,561671	6,50129	-0,19263	-1,25236	42,26677	0,037107	-0,17975	0,566333
0,620558	7,000334	-0,03144	-0,22006	49,00468	0,000988	-0,02533	0,622802
0,681615	7,500529	0,134963	1,012291	56,25794	0,018215	0,129442	0,679603
0,740589	8,006368	0,299617	2,398842	64,10192	0,08977	0,285962	0,735798
0,792717	8,49699	0,453411	3,852626	72,19885	0,205581	0,437774	0,787594
0,838483	8,999619	0,600563	5,404838	80,99315	0,360676	0,593301	0,836338
0,875102	9,50002	0,732492	6,958688	90,25039	0,536544	0,748139	0,879133
0,898718	9,903488	0,828485	8,204889	98,07907	0,686387	0,872983	0,90874
0,995819	12,49914	1,700595	21,25597	156,2284	2,892023	1,676148	0,995228
0,997191	12,79595	1,770694	22,65771	163,7364	3,135356	1,767992	0,997146
	116,7145	3,243251	55,2088	1005,018	9,979688		
	7,780966	0,216217	3,680587	67,0012	0,665313		
0,309427	Sx=	2,541215					
0,111757	Sy=	0,786488					
0,999789	δ=	0,091					

РЕПОЗИТОРИЙ БГУИР

Из табл. 3 и графика функции распределения видно, что введение третьего параметра в закон Вейбулла позволило весьма точно аппроксимировать статистическое распределение первых 50–80 слов Частотного словаря. Ранговое распределение остальных слов хорошо описывается классическим законом Вейбулла с двумя параметрами.

Вычислим абсциссы трех характерных точек по формулам (18) и (22) при известных оценках параметров α и β: n=22,4287; $t_C = 1191$; $t_A = 53$; $t_B = 26704$. Отметим, что логарифмы рангов слов в точках A, C, B равны: 3,97187; 7,08221; 10,19256. Теоретическая функция распределения в этих точках равна: 0,31748; 0,63212; 0,92705.

Из этих расчетов следует, что ядро Частотного словаря составляют первые 53 слова. Они покрывают 31.7% текста. В первую зону рассеяния A–C входит 1191-53=1138 слов, которые покрывают 31.5% текста (63.2-31.7=31.5). Во вторую зону C–B входит 26704-1191=25513 слов, которые покрывают 29.5% текста (92.7-63.2=29.5). В третью зону рассеяния входит вся остальная лексика 739930-25513=714417 слов. Этот огромный словарь покрывает лишь 7.3% текста (100-92.7=7.3).

Рассмотрим еще один пример – *ранговое распределение периодических изданий по химии и химической технологии* [1]. Поскольку в данном случае выборка однородная, то для аппроксимации статистического рангового распределения может быть использован закон Вейбулла с двумя параметрами. Проведем необходимые расчеты по методу, изложенному выше. Результаты представлены в виде табл. 4 и рис. 4.

Эти результаты свидетельствуют о высокой точности аппроксимации законом Вейбулла статистического рангового распределения журналов, упорядоченных по убыванию опубликованных в них статей по химии и химической технологии. Коэффициент корреляции Ry/x=0,999705. Ядро образуют 88 журналов. Количество журналов до точки *C*, т. е. входящих в ядро и первую зону рассеяния, равно 552. В ядро и в первые две зоны рассеяния, т. е. до точки B входят 3469 журналов, в которых содержится 92,705% статей от их общего количества 187911. На третью зону рассеяния приходятся все остальные журналы 10850-3469=7381, и в этих журналах содержится 100-92,705=7,295 % статей.

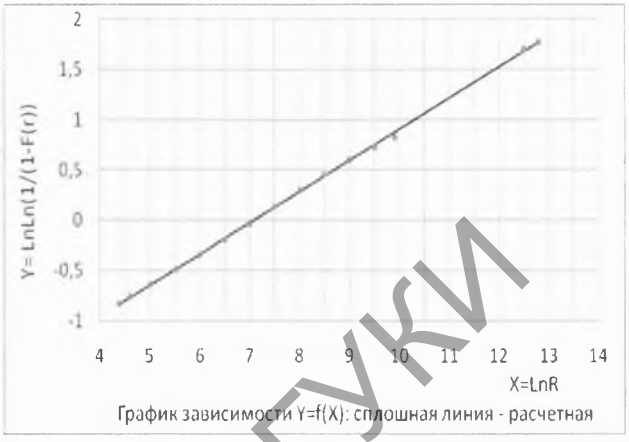

Рис. 2. Прямая Вейбулла

Рис. 3. Функция распределения Вейбулла

Таблица 3

Эмпирическая и теоретическая функции распределения наиболее частых слов

Ранги слов	Эмпирическая функция распределения	Уточненная теоретическая функция распределения
1	0,035802	0,035798
2	0,067176	0,061847
3	0,085204	0,082922
4	0,101071	0,100782
5	0,113756	0,116317
10	0,165571	0,172804
20	0,217115	0,235534
30	0,255761	0,271023
50	0,309294	0,313454

РЕПОЗИТОРИЙ БГУКИ

Таблица 4

Рассеяние журнальных публикаций по химии и химической технологии (10850 журналов, 187911 статей)

	Доля статей	lnt	$\ln\ln\frac{1}{1-F(t)}$					
	F(t)	X	Y	XY	X^2	Y^2	F(t)расч.	Yрасч.
	0,15	2,890372	-1,817	-5,25181	8,3542489	3,301489	0,1536	-1,7912
	0,25	3,912023	-1,2459	-4,87399	15,303924	1,552267	0,24772	-1,25649
	0,34	4,60517	-0,8782	-4,04426	21,207592	0,771235	0,33577	-0,89371
	0,62	6,214608	-0,033	-0,20508	38,621354	0,001089	0,61323	-0,05138
	0,75	6,907755	0,3266	2,25607	47,717083	0,106668	0,7447	0,3114
	0,85	7,600902	0,6403	4,86686	57,773718	0,409984	0,85948	0,67418
		32,13083	-3,0072	-7,25221	188,97792	6,142732		
		5,355138	-0,5012	-1,2087	31,49632	1,023789		
	0,523374	Sx=	1,67893183	tc=	551,5708	F(tc)=	0,63212	
	0,036738	Sy=	0,87896938	ta=	87,69708	F(ta)=	0,31748	
	0,999705	n=	6,28949982	tb=	3469,104	F(tb)=	0,92705	

Рис. 4. Прямая Вейбулла – рассеяние журнальных публикаций

несмотря на высокую точность аппрок-
которых ранговых распределений зако-
ла, в исследованиях по информатике и
ской лингвистике он применяется весь-
Более часто используется закон Дж.
орый вовсе нельзя применять в таких
иях. Принимая во внимание то обстоя-
го оба этих закона и множество других
астными случаями обобщенного рас-
(2), для описания различного рода ран-
еделений следует использовать вторую
прерывных распределений.

ЗАКЛЮЧЕНИЕ

При обработке статистических рядов распределения главной задачей является вычисление теоретического закона распределения. Она решается довольно просто по методам, изложенным в теории обобщенных распределений. Для аппроксимации статистических ранговых распределений используется вторая система непрерывных распределений.

На основании анализа свойств обобщенных распределений нами предлагаются математически точные формулировки закона рассеяния публикаций в

РЕПОЗИТОРИЙ БГУКИ

смысле Бредфорда. Но такие формулировки, как и закон С.Бредфорда, не могут быть приняты в качестве полноценного закона рассеяния публикаций. Универсальным законом рассеяния является вторая система непрерывных распределений, поскольку обобщенная четырехпараметрическая плотность, т. е. закон распределения, наиболее полно характеризует случайную величину.

Для вычисления закона распределения и оценок его параметров по статистическому ранговому распределению используется общий устойчивый метод. При известных оценках параметров по заранее выведенным формулам вычисляются абсциссы трех характерных точек A, C, B, которые приняты автором в качестве границ ядра и зон рассеяния. Абсциссы точек C и B, вычисленные по закону С.Бредфорда и универсальному закону, различаются на 20–25 % при условии, что величина $n = 5$, а размер ядра в обоих случаях одинаков. С ростом n эта погрешность уменьшается.

Для использования аналитического метода необходимо знать хотя бы некоторые сведения из теории обобщенных распределений. Этот метод рассчитан на подготовленного исследователя.

На базе свойств ранговых распределений нами предложен графический метод приближенного вычисления границ ядра и зон рассеяния. Он значительно проще аналитического метода, поскольку не требует вычисления закона распределения.

В случае однородных выборок некоторые статистические ранговые распределения могут быть описаны законом Вейбулла (13), (14). Оценки параметров этого закона наиболее просто вычисляются по методу наименьших квадратов. Если при этом ранговое распределение содержит неоднородную часть, например, служебные слова, то по формуле (30) следует дополнительно вычислить третий параметр δ при известных значениях параметров α, β и относительной частоты первого слова.

Проведенное исследование показало высокую точность аппроксимации некоторых статистических ранговых распределений законом Вейбулла. Однако для гарантированного вычисления наилучшего теоретического рангового распределения по статистическому ряду следует использовать вторую систему непрерывных распределений и общий устойчивый метод.

СПИСОК ЛИТЕРАТУРЫ

1. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968. – 756 с.
2. Нешитой В.В. Универсальные законы рассеяния и старения публикаций // Веснік Беларус. дзярж. ун-та культуры і маст.– 2007.– № 8.– С. 128–133.
3. Нешитой В.В. Элементы теории обобщённых распределений: монография.– Минск: РИВШ, 2009. – 204 с.
4. Нешитой В.В. Математико-статистические методы анализа в библиотечно-информационной деятельности: учеб.-метод. пособие. – Минск: БГУ культуры и искусств, 2009. – 203 с.
5. Нешитой В.В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие. – Минск: Бел. гос. ун-т культуры и искусств, 2010. – 61с.
6. 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.
7. Нешитой В.В. Форма представления ранговых распределений // Ученые записки Тартуского гос. ун-та. – 1987. – Вып. 774. – С. 123–134.
8. Белоногов Г.Г. О некоторых статистических закономерностях в русской письменной речи // Вопросы языкознания. – 1962. – № 1. – С. 100–101.
9. Нешитой В.В. Законы распределения слов в тексте и его лексическая параметризация: дис... канд. филол. наук. – Минск, 1973. – 135 с.
10. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М.: Азбуковник, 2009. – URL: http://dict.ruslang.ru/freq.php.

Материал поступил в редакцию 03.05.13.

Сведения об авторе

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

РЕПОЗИТОРИЙ БГУКИ