В. В. Нешитой,
доктор технических наук, профессор

МЕТОДЫ ВЫЯВЛЕНИЯ СКРЫТОЙ ИНФОРМАЦИИ НА ПРИМЕРЕ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

В информатике, библиотечном деле, социологии, математической лингвистике широко используются ранговые распределения. Для описания статистических ранговых распределений все еще применяется закон распределения Дж. Ципфа

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

где r – ранг, т. е. порядковый номер журнала (книги, слова и т. д.) в списке, где они упорядочены по убыванию числа опубликованных в них статей по заданной тематике (или книг по числу выдач, слов по частоте употребления в текстах и т. д.);

p_r – относительная частота статей в журнале с рангом r;

k – коэффициент (параметр распределения).

Естественно, что никакое статистическое ранговое распределение не может быть описано с достаточной точностью приведенной формулой с одним параметром. В связи с этим многими исследователями были предприняты попытки усовершенствовать модель Ципфа, но они не дали существенных результатов. Например, была предложена трехпараметрическая формула – закон Эсту-Ципфа-Мандельброта:

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

Некоторые исследователи утверждают, что закон рассеяния публикаций, сформулированный С. Бредфордом, является следствием закона Ципфа, что не соответствует действительности. Закон С. Бредфорда никоим образом не может быть получен из закона Ципфа [1].

Спрашивается, почему долгое время не удавалось найти более точные математические модели для описания ранговых распределений, а также для обоснования законов рассеяния и старения публикаций? Ответ заключается в том, что многие исследователи пытались аппроксимировать частные статистические распределения (Ципфа, Бредфорда и др.), не ставя перед собой

общей задачи, например, построить универсальную модель для аппроксимации любых, в том числе ранговых, распределений.

Только решение этой общей задачи позволит аппроксимировать практически любые статистические распределения однородных случайных величин, в том числе ранговые распределения. Как правило, общая задача решается значительно проще частных задач.

Итак, следует найти универсальное распределение, включающее как частные случаи множество известных распределений. Но как подступиться к решению этой задачи?

Используем метод моделирования. Рассмотрим самые простые распределения, плотности которых заданы уравнением прямой. Если не принимать во внимание усеченных распределений, то уравнение прямой позволяет записать три плотности распределения с одним параметром α:

$$p(t) = a(1 - at/2); \quad (0 < t < 2/a); \tag{3}$$

$$p(t) = a; \quad (0 < t < 1/a); \tag{4}$$

$$p(t) = 2at; \quad (0 < t < \sqrt{1/a}). \tag{5}$$

Графики этих плотностей, т. е. кривые распределения, имеют вид прямой. Эти распределения случайной величины Т 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 некоторому одному, покуда еще неизвестному, семейству распределений, которое должно быть задано 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 распределения. Теперь необходимо найти метод, который поможет выявить предполагаемое обобщенное распределение.

Здесь уместно отметить, что закон распределения может быть задан не только плотностью, но и функцией распределения, которая представляет собой интеграл от плотности

$$F(t) = \int p(t)dt.$$

При интегрировании плотностей распределения (3)–(5) появятся новые значения параметров, т.е. будет выявлена новая информация. Вместо этих значений введем новые параметры.

Итак, интегрируя плотности (3)–(5), найдем три функции распределения:

$$F(t)=1-(1-at/2)^2; \quad (6)$$

$$F(t)=at=1-(1-at); \quad (7)$$

$$F(t)=at^2=1-(1-at^2). \quad (8)$$

Далее используем метод обобщения. Обобщим попарно функции распределения (6) и (7), (7) и (8) путем введения новых параметров (вместо показателей степени 1 и 2). В первом случае можем записать

$$F(t)=1-(1-aut)^{\frac{1}{u}}. \quad (9)$$

Во втором случае

$$F(t)=1-(1-at^b). \quad (10)$$

Теперь замечаем, что в формуле (10) имеется параметр β, но его нет в формуле (9). Введем его в последнюю формулу. В результате получим

$$F(t)=1-(1-aut^b)^{\frac{1}{u}},$$

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

$$p(t)=abt^{b-1}(1-aut^b)^{\frac{1}{u}-1}. \quad (11)$$

Полученное распределение может быть еще более расширено за счет введения нового параметра k. Тогда вместо (11) можем записать искомое четырехпараметрическое распределение, которое задано плотностью

$$p(t)=Nt^{kb-1}(1-aut^b)^{\frac{1}{u}-1}. \quad (12)$$

Итак, простейшими средствами (моделирования плотностей на базе уравнения прямой, интегрирования, обобщения функций распределения, дифференцирования трехпараметрической функции распределения и, наконец, введения четвертого параметра k) нами получено универсальное четырехпараметрическое распределение, предназначенное для аппроксимации существенно положительных случайных величин (T>0), в том числе статистических ранговых распределений.

Исследования показали, что формула (12) включает как частные случаи законы Ципфа (1), Эсту-Ципфа-Мандельброта (2), нормальный закон, Стьюдента, Коши, Вейбулла, «хи-квадрат», гамма-распределение, Максвелла и множество других. При этом обобщенное распределение (12) является универсальным законом рассеяния публикаций [2, с. 143].

Имея формулу (12), на ее базе можно получать другие плотности как распределения функций случайного аргумента. Например, при $T = e^X$ найдем

$$p(x) = Ne^{kbx}\left(1 - aue^{bx}\right)^{\frac{1}{u}-1}. \qquad (13)$$

Здесь случайная величина X может быть задана на всей числовой оси, т. е. значения x могут быть как положительными, так и отрицательными. На базе уже этих двух распределений (12) и (13) можно достаточно просто решать различные задачи.

Так, график плотности (13) при $u<1/2$ имеет моду и две точки перегиба, расположенные на равных расстояниях от моды. График плотности (12) при определенных значениях параметров имеет вид убывающей кривой. Следовательно, эта плотность может описывать ранговые распределения. Преобразованная к форме плотности (13), т.е. представленная в виде [3]

$$tp(t) = Ne^{kb\ln t}\left(1 - aue^{b\ln t}\right)^{\frac{1}{u}-1},$$

при значениях параметра $u<1/2$ она приобретает свойства плотности (13), т. е. имеет моду и две точки перегиба, абсциссы которых могут служить границами ядра журналов и зон рассеяния.

Таким образом, приведение плотности (12) к форме плотности (13) позволило выявить новую информацию о существовании трех характерных точек.

Рассмотрим еще один пример – вычисление оценок параметров закона Вейбулла по методу наибольшего правдоподобия. Этот закон с успехом может быть использован в информатике, библиотечном деле, математической линвистике для описания ранговых распределений. Функция распределения и плотность вероятности его задаются формулами

$$F(t) = 1 - e^{-at^b}, \qquad (14)$$

$$p(t) = abt^{b-1}e^{-at^b}. \qquad (15)$$

Приведем плотность распределения Вейбулла к форме $tp(t) = f(\ln t)$, а именно

$$tp(t) = abe^{b\ln t}e^{-ae^{b\ln t}}. \qquad (16)$$

Для нахождения оценок параметров α и β распределения Вейбулла вначале прологарифмируем равенство (16)

$$\ln tp(t) = \ln a + \ln b + b\ln t - ae^{b\ln t}.$$

Запишем далее логарифмическую функцию правдоподобия как математическое ожидание логарифма произведения $tp(t)$:

$$M[\ln tp(t)] = \ln a + \ln b + bM(\ln t) - aM(t^b) \qquad (17)$$

Дифференцируя (17) по параметрам α и β и приравнивая производные нулю, найдем два уравнения правдоподобия. Решая

систему полученных уравнений, можно найти оценки параметров α и β. Итак, имеем

$$\frac{\partial M[\ln tp(t)]}{\partial a} = \frac{1}{a} - M(t^b) = 0; \qquad (18)$$

$$\frac{\partial M[\ln tp(t)]}{\partial b} = \frac{1}{b} + M(\ln t) - aM(t^b \ln t) = 0. \qquad (19)$$

Поскольку решить эту систему весьма сложно, поищем более простые формулы. Для этого рассмотрим трехпараметрическое распределение второго типа [2, с. 33]

$$p(t) = \frac{b a^k}{\Gamma(\mathrm{k})} t^{kb-1} e^{-at^b}, \qquad (20)$$

которое включает как частный случай закон Вейбулла при $k=1$. Запишем логарифмическую функцию правдоподобия для плотности (20):

$$M[\ln tp(t)] = \ln b + k \ln a - \ln \Gamma(\mathrm{k}) + \mathrm{k}b\mathrm{M}(\mathrm{lnt}) - a\mathrm{M}(\mathrm{t}^b). \quad (21)$$

Найдем из (21) дополнительное уравнение правдоподобия как частную производную по параметру k:

$$\frac{\partial M[\ln tp(t)]}{\partial k} = \ln a - y(k) + bM(\ln t) = 0. \qquad (22)$$

Преобразуем логарифмическую функцию правдоподобия (21) с учетом уравнений (18) и (22) при $k=1$:

$$M[\ln tp(t)] = \ln b + y(1) - 1 = \ln b - 1{,}577216. \qquad (23)$$

Отсюда находим

$$\ln b = M[\ln tp(t)] + 1{,}577216,$$

$$b = e^{M[\ln tp(t)]+1,577216} = 4{,}841458 e^{M[\ln tp(t)]} = 4{,}841458 e^{M(\ln t)+M[\ln p(t)]}. \quad (24)$$

Из уравнения правдоподобия (22) при $k=1$ получим

$$\ln a = y(1) - bM(\ln t) = -0{,}577216 - bM(\ln t),$$

$$a = e^{-(0,577216+bM(\ln t))} = 0{,}561459 e^{-bM(\ln t)}. \qquad (25)$$

Таким образом, использование более общего распределения, включающего как частный случай закон Вейбулла (при $k=1$), позволило найти для последнего третье (скрытое) уравнение правдоподобия, которое непосредственно из распределения Вейбулла не вытекает. В результате были получены две простые формулы для вычисления оценок параметров α, β по статистическому распределению. Для этого достаточно подставить в формулы (24), (25) вместо соответствующих математических ожиданий их оценки, т. е. средние $\overline{\ln t}$ и $\overline{\ln p(t)}$, вычисленные по статистическому распределению.

1. Нешитой, В. В. Законы Ципфа, Бредфорда и универсальные модели / В. В. Нешитой // НТИ. Сер. 2. – М., 2010. – № 1. – С. 26–33.

2. Нешитой, В. В. Элементы теории обобщенных распределений / В. В. Нешитой. – Минск : РИВШ, 2009. – 204 с.

3. Нешитой, В. В. Форма представления ранговых распределений / В. В. Нешитой // Учен. зап. Тартус. гос. ун-та. – 1987. – Вып. 774. – С. 123–134.