В. В. НЕШИТОЙ

УНИВЕРСАЛЬНЫЕ ЗАКОНЫ РАССЕЯНИЯ И СТАРЕНИЯ ПУБЛИКАЦИЙ

Для оптимизации комплектования библиотечного фонда, улучшения информационного обслуживания потребителей необходимо знать и уметь использовать на практике законы рассеяния и старения публикаций. Чтобы использовать эти законы с наибольшей эффективностью, статистические распределения необходимо с высокой точностью аппроксимировать теоретическими.

В статье обосновывается применение обобщенных четырехпараметрических распределений в качестве универсальных законов рассеяния и старения публикаций, которые наиболее точно описывают статистические распределения.

Ранговые распределения. Статистические данные, полученные в результате наблюдения, представляют собой простой статистический ряд. Его можно упорядочить либо по возрастанию значений случайной величины, либо по убыванию. В обоих случаях получим вариационный (ранжированный) ряд. Ранговые распределения находят широкое применение в информатике, математической лингвистике, социологии, библиотечном деле и других отраслях знания.

Рассмотрим, например, частотный словарь. В таком словаре разные слова упорядочены по убыванию (точнее, по невозрастанию) частоты их употребления в текстах, на базе которых построен словарь. Порядковый номер слова и есть его ранг. В качестве другого примера можно привести ранговое распределение журналов по некоторой отрасли знания (например, по химии и химической технологии), упорядоченных по убыванию числа помещенных в них статей по заданному предмету.

Для описания ранжированных рядов необходимо использовать такие теоретические распределения, которые обладают теми же свойствами, что и ранжированные ряды. Возникает вопрос, откуда взять распределения, пригодные для выравнивания статистических ранговых распределений. Чтобы решить эту проблему, необходимо либо разработать теорию ранговых распределений, либо использовать ранее построенные обобщенные распределения. Среди множества частных случаев этих распределений найдутся такие, которые с достаточной точностью могут описывать статистические ранговые распределения.

Форма представления ранговых распределений. Рассмотрим четыре системы непрерывных распределений, которые заданы обобщенными плотностями [1].

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

$$p(t)=Nt^{kb-1}\left(1-aut^{b}\right)^{\frac{1}{u}-1}, \quad (2)$$

$$p(y)=\frac{N}{y}(\ln y)^{kb-1}\left[1-au(\ln y)^{b}\right]^{\frac{1}{u}-1}, \quad (3)$$

$$p(w)=\frac{N}{w\ln w}(\ln\ln w)^{kb-1}\left[1-au(\ln\ln w)^{b}\right]^{\frac{1}{u}-1}, \quad (4)$$

где N – нормирующий множитель; α, β, k, u – параметры.

Они различаются между собой началом отсчета значений случайных величин: $X > -\infty$; $T > 0$; $Y > 1$; $W > e$. Здесь e – основание натуральных логарифмов (e = 2,71828 ...).

Статистическое ранговое распределение можно представить в виде обычной гистограммы, которую можно аппроксимировать непрерывной убывающей кривой распределения. Для описания ранговых распределений могут быть использованы три последние плотности. Первая плотность не обладает свойствами ранговых распределений.

Исследования показали [2], что ранговые распределения журналов, упорядоченных по убыванию числа помещенных в них статей по заданному предмету, хорошо описываются второй системой непрерывных распределений, которая задана обобщенной плотностью *p(t)*.

Для ранговых распределений автором предложена удобная форма представления в виде зависимости произведения rp_r от *lnr*, где p_r – доля статей по заданному предмету в журнале с рангом *r*. Преимущество такой формы представления ранговых распределений заключается в том, что убывающая кривая распределения $p_r = f(r)$ после ее приведения к форме $rp_r = f(lnr)$ в случае однородной выборки превращается в одновершинную кривую, которая описывается плотностью *p(x)* [3].

Универсальный закон рассеяния публикаций. График плотности *p(x)* при $u < 1/2$ представляет собой одновершинную кривую распределения, которая имеет три характерные точки: моду С и две точки перегиба А, В, причем эти точки расположены на равных расстояниях от моды С. И в этом состоит суть закона рассеяния публикаций в формулировке Бредфорда.

Итак, для плотности *p(x)* имеем: $x_C - x_A = x_B - x_C$. Учитывая взаимосвязи между первой и второй системами непрерывных распределений, т.е. $x = lnt$, $p(x) = tp(t)$, для плотности *p(t)* можем записать: $lnt_C - lnt_A = lnt_B - lnt_C$, откуда имеем равенство

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n.$$

Из последней формулы следует закон рассеяния публикаций, выраженный через абсциссы трех характерных точек на кривой распределения *tp(t) = f(lnt):*

$t_A : t_C : t_B = 1 : n : n^2$, (5)

$t_Я : t_I : t_{II} = 1 : (n - 1) : (n - 1)n$, (6)

где $t_Я = t_A$; $t_I = t_C - t_Я$; $t_{II} = t_B - t_C$.

Полученные формулы отличаются от формулировки С.Бредфорда закона рассеяния в виде $t_Я : t_I : t_{II} = 1 : n : n^2$. Кроме того, у С.Бредфорда предполагается, что доли статей в ядре и зонах рассеяния одинаковы. В нашем случае они различны.

Из обобщенной плотности *p(t)* следует, что доля статей в ядре журналов равна функции распределения в точке А, т.е. $F(t_A)$. Доля статей в журналах, входящих в ядро и первую зону рассеяния, составляет $F(t_C)$, а входящих в ядро и первые две зоны рассеяния – $F(t_B)$.

На базе плотности *p(t)* нетрудно найти координаты трех характерных точек и вычислить величину *n*. Абсциссы точек *A* и *B* можно рассчитать при известных значениях величин t_C и *n*. Мода t_C находится из условия *dtp(t)/dlnt = 0*. Для распределений с параметром $\beta > 0$ она равна

$$t_C = \left(\frac{k}{a\left(1 + ku - u\right)} \right)^{1/b} . \quad (7)$$

Величина *n* задается формулой

$$n = \left[1 + \frac{1 - u + \sqrt{\left[4k\left(1 + ku - u\right) + \left(1 - u\right)\right]\left(1 - u\right)}}{2k\left(1 + ku - u\right)} \right]^{1/b} . \quad (8)$$

Абсциссы точек перегиба вычисляются по формулам*:*

$t_A = t_C/n$; $t_B = t_C \cdot n$. (9)

Дальнейшие исследования показали [2], что обобщенная плотность *p(t)* при различных значениях параметров дает закон рассеяния в виде формул (5), (6), *но при этом размеры ядра и зон рассеяния, величина* n*, а также доли статей в ядре и зонах рассеяния различны* (последние у С. Бредфорда одинаковы) и зависят от значений параметров.

Поскольку наиболее полной характеристикой случайной величины является ее закон распределения, в данном случае рангового, то *наиболее общим и универсальным законом рассеяния публикаций является вторая система непрерывных распределений, заданная*

обобщенной плотностью p(t). Именно обобщенная плотность позволяет наиболее точно описывать статистические ранговые распределения журналов, вычислять накопленную долю статей в заданном числе журналов в ранжированном ряду, в том числе в характерных точках А, С, В, вычислять координаты этих точек и величину *n*, входящую в закон рассеяния. Именно обобщенная плотность позволяет дать математически точную формулировку закона рассеяния публикаций в виде формул (5), (6). Другими словами, в качестве универсального закона рассеяния публикаций выступает вторая система непрерывных распределений, а закон рассеяния в виде формул (5), (6) является лишь следствием свойств ранговых распределений, частным случаем универсального закона, отражающим соотношение между *абсциссами* характерных точек на кривой распределения. Значения же функции распределения в характерных точках в этих формулах не используются. Поэтому, не зная теоретического распределения с его значениями параметров, нельзя вычислить число журналов, входящих в ядро и зоны рассеяния, величину *n*, а также долю статей в ядре и зонах рассеяния, которая выражается через функцию распределения.

Обобщенные распределения позволяют вычислять число журналов, содержащих заранее определенную долю статей по заданному предмету. Например, в случае распределений группы А (для которых параметр $k = 1$) с функцией распределения

$$F(t) = 1 - \left(1 - a\,u t^{\,b}\right)^{\frac{1}{u}} \qquad (10)$$

имеем

$$t = \left\{ \frac{1}{a\,u} \left[1 - \left(1 - F(t)\right)^{u} \right] \right\}^{\frac{1}{b}},$$

где t – ранг журнала; $F(t)$ – накопленная относительная частота статей по заданному предмету в t журналах.

В частном случае при $u \to 0$ из формулы (10) имеем распределение Вейбулла, из которого находим

$$t = \left(\frac{1}{a} \ln \frac{1}{1 - F(t)} \right)^{\frac{1}{b}}.$$

В случае распределений группы *Б*, для которых параметр $k \neq 1$, эта задача решается с помощью соответствующих компьютерных программ.

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 (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, которая задана обобщенной плотностью $p(x)$ [2], 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 и универсальным законом старения публикаций является первая система непрерывных распределений, заданная обобщенной плотностью p(x).* Обобщенная плотность позволяет 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)* при β > 0 задаются формулами

$$x_c = \frac{1}{b} \ln \frac{k}{a\,(1+ku-u)}, \qquad x_{A,B} = x_C \pm \ln n,$$

где величина *n* рассчитывается по прежней формуле (8).

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 отраслей знания. При использовании обобщенных распределений исчезают ранее существовавшие барьеры на пути к новому знанию. Например, для нахождения наилучшей аппроксимирующей кривой распределения 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 статистической совокупности оба метода [2], разработанные специально для обобщенных распределений, – универсальный метод моментов и устойчивый метод – дают очень близкие значения оценок параметров аппроксимирующего распределения. Наиболее точные оценки параметров получаются в случае симметричного или близкого к нему статистического распределения, приведенного к форме плотности *p(x)*.

Для работы с системами непрерывных распределений автором разработана серия компьютерных программ под общим названием SNR для обоих методов оценивания параметров.

---

1. Нешитой, В. В. Методы статистического анализа на базе обобщенных распределений: учеб.-метод. пособие / В. В. Нешитой. – Мн.: Веды, 2001. – 168 с.

2. Нешитой, В. В. Исследование статистических закономерностей текста и информационных потоков: дис…докт. техн. наук / В. В. Нешитой. – Мн., 1987. – 505 с.

3. Нешитой, В. В. Форма представления ранговых распределений / В. В. Нешитой // Ученые записки Тартуского гос. ун-та, 1987. – Вып. 774.