ИСПОЛЬЗОВАНИЕ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ В СОЦИАЛЬНО-ГУМАНИТАРНОЙ СФЕРЕ

В.В.Нешитой, *доктор технических наук, Белорусский государственный университет культуры и искусств*

Научные исследования в любой области знания не обходятся без применения вероятностно-статистических методов и моделей. Ниже пойдет речь о ранговых распределениях, которые используются не только в естественных науках, но и в таких областях знания, как социология, математическая лингвистика, биология, библиотековедение, культура и др.

В естественных науках наиболее часто используются такие распределения, которые описывают взаимосвязь между значениями случайной величины и их вероятностями. К этому классу относятся непрерывные распределения (нормальное, Стьюдента, показательное, Вейбулла, хи-квадрат и др.) и дискретные распределения (гипергеометрическое, биномиальное, Пуассона, отрицательное биномиальное и др.).

В гуманитарных науках наряду с приведенными используются ранговые распределения. Они описывают статистические ранжированные ряды распределения, т.е. такие, в которых все элементы ряда упорядочены по невозрастанию их относительных или абсолютных частот появления в исследуемой выборке.

Для наглядного представления ранговых распределений традиционно строят график зависимости $p_r = f(r)$ либо $\ln p_r = \varphi(\ln r)$, где r – ранг события, т.е. его порядковый номер от начала частотного списка, p_r – относительная частота события с рангом r.

Примерами ранговых распределений являются распределение разных слов по частоте их употребления в текстах (слова упорядочены по невозрастанию частот), распределение периодических изданий по количеству помещенных в них статей по заданному предмету, распределение научных сотрудников по продуктивности, распределение книг по числу их выдачи читателям и т.д.

Следует отметить, что ранговые распределения изучены в недостаточной степени для того, чтобы с их помощью успешно решать различные практические задачи. Ранговые распределения

требуют всестороннего исследования, при этом в первую очередь необходимо решить следующие задачи:

– определить класс распределений, пригодных для описания того или иного рангового распределения;

– выработать наиболее целесообразную форму представления ранговых распределений;

– установить критерий однородности ранговых распределений и т.д., а также решить множество других задач.

Для описания статистических ранговых распределений часто используют закон Ципфа [3] $p_r = k/r$, где k – параметр (для англоязычных текстов $k \approx 0{,}1$). Ясно, что распределение с одним параметром не может с достаточной точностью описывать статистические ранговые распределения. Для этого требуются многопараметрические семейства распределений.

Три системы непрерывных распределений. Для описания практически всего многообразия статистических распределений, в том числе ранговых, автором разработаны три системы непрерывных распределений, которые заданы плотностями

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

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

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}(1-\alpha u \ln^{\beta} y)^{\frac{1}{u}-1}. \qquad (3)$$

Плотность (1) описывает статистические одновершинные распределения, заданные на всей числовой оси, т.е. $-\infty < x < \infty$. Они не могут описывать ранговые распределения, поскольку не являются убывающими.

Две другие плотности содержат частные случаи, пригодные для описания невозрастающих ранговых распределений.

Практика показала, что плотность (2) хорошо описывает ранговые распределения периодических изданий, упорядоченных по убыванию числа помещенных в них статей по заданному предмету.

Плотность (3) хорошо описывает статистические ранговые распределения однородных лексических единиц (знаменательных слов, словосочетаний, ключевых слов, терминов).

Отметим, что каждая система непрерывных распределений может быть дополнена еще двумя плотностями, что значительно расширяет их аппроксимирующие возможности [1].

Форма представления ранговых распределений. Ранговые распределения графически обычно изображают либо в координатах “ранг – частота” (по горизонтальной оси откладываются ранги, а по вертикальной – их абсолютные или относительные частоты), либо в координатах “логарифм ранга – логарифм частоты”. Такая форма представления (как в первом, так и во втором случае) не дает ясного представления о свойствах исследуемого рангового распределения, так как содержит мало информации о нем.

Ранговые распределения желательно привести в такой форме, чтобы они ничем не отличались от других распределений, например тех, которые заданы плотностью (1).

Найдем такую форму для плотности (2). Умножим левую и правую части (2) на *t*, при этом в правой части выражение t^{β} запишем в виде $e^{\beta \ln t}$, что одно и то же. В результате будем иметь равенство

$$tp(p) = Ne^{k\beta \ln t}(1 - \alpha u e^{\beta \ln t})^{\frac{1}{u}-1}. \qquad (4)$$

Сравнивая (1) и (4), видим, что *tp(t)=p(x), lnt=x.* Следовательно, плотность (2), представленная в системе координат *tp(t)=f(lnt),* преобразуется в плотность (1) и обладает всеми свойствами последней. Таким же путем нетрудно найти новую форму представления для плотности (3).

Умножим левую и правую части выражения (3) на произведение *ylny* и запишем выражение $\ln^{\beta} y$ в виде $e^{\beta \ln \ln y}$. В результате получим

$$y \ln y p(y) = Ne^{k\beta \ln \ln y}(1 - \alpha u e^{\beta \ln \ln y})^{\frac{1}{u}-1}. \qquad (5)$$

Здесь *ylnyp(y) = p(x), lnlny = x*. С учетом последних равенств формула (5) представляет собой плотность распределения *p(x),* заданную формулой (1).

Таким образом, для того, чтобы привести ранговое распределение к обычному одновершинному, достаточно плотность (2) представить в виде зависимости *tp(t)=f(lnt)*, а плотность (3) – в виде зависимости $y \ln y p(y) = \varphi(\ln \ln y)$. Приведенные к такой форме распределения (2) и (3) обретают свойства распределения (1). Последнее обладает замечательными свойствами: оно имеет одну колоколообразную форму; *k, u* может быть симметричным либо иметь право- или левостороннюю асимметрию в зависимости от значений параметров формы; кривые распределения, заданные плотностью (1), в общем случае имеют три характерные точки А, С, В, где А и В – точки перегиба, С – мода. Точки А и В расположены

на равных расстояниях от моды C. Эти точки используются автором как границы ядра и зон рассеяния (например, в случае рангового распределения журналов). Однако вычислить координаты трех характерных точек возможно лишь при условии однородности статистического рангового распределения.

Критерий однородности ранговых распределений. Предложенная форма представления ранговых распределений имеет принципиальные преимущества перед традиционной формой.

Во-первых, кривая распределения становится компактной, несмотря на то, что ранг самого редкого события может быть очень большим (до 100 000 и более), поскольку по горизонтальной оси откладываются не сами ранги, а их логарифмы – в случае плотности (2). По вертикальной оси откладываются значения произведения rp_r, которые с ростом ранга вначале возрастают, затем убывают.

Во-вторых, статистическое ранговое распределение однородных единиц проявляет себя в полной мере лишь при условии, если крайняя справа точка близка к горизонтальной оси. Только при этом условии можно вычислять аппроксимирующее непрерывное распределение.

В-третьих, статистическую кривую распределения $rp_r = f(\ln r)$ можно использовать для расчета необходимого объема выборки, а также для построения достоверного словаря заданного объема [2].

В-четвертых, статистическое ранговое распределение однородных элементов выборочной совокупности имеет одновершинную кривую распределения с закономерным возрастанием и убыванием без резких скачков. Это позволяет устанавливать однородность или неоднородность статистических ранговых распределений и при необходимости выделять неоднородную часть.

В-пятых, закон Ципфа $p_r = k/r$, приведенный к форме $rp_r = f(\ln r)$, имеет вид $rp_r=k=const$, чего на практике никогда не наблюдается. Это свидетельствует о том, что закон Ципфа не может быть использован для аппроксимации статистических ранговых распределений, хотя и является частным случаем плотностей (2) и (3).

Итак, введение новой формы представления ранговых распределений позволяет работать с ними как с обычными одновер-

шинными распределениями, для которых разработаны методы вычисления типа аппроксимирующей кривой, оценок параметров, а также координат характерных точек.

Новая форма представления ранговых распределений позволяет решать ряд важных задач, что было невозможно при традиционной форме их представления в виде “ранг – частота” или “логарифм ранга – логарифм частоты”.

Отныне ранговые распределения становятся в один ряд с обычными одновершинными распределениями, задаются одними и теми же обобщенными распределениями (2), (3), для вычисления оценок их параметров используются те же методы, что и для обычных одновершинных распределений.

_____

1. Нешитой, В.В. Элементы теории обобщенных распределений / В.В.Нешитой. – Мн.: РИВШ, 2009.

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

3. Zipf, G.K. Human behavior and the principle of least effort / G.K.Zipf. – Cambridge, 1949.