В.В.НЕШИТОЙ

ВЫЧИСЛЕНИЕ АППРОКСИМИРУЮЩИХ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

Ранговые распределения широко используются в библиотечно-информационной деятельности, например при исследовании структуры библиотечного фонда, выделении ядра журналов и зон рассеяния. При этом журналы располагаются в порядке убывания числа помещенных в них статей по заданному предмету. Для обработки таких распределений требуется другой подход, чем в случае, когда статистическое распределение представлено в виде вариационного интервального ряда или графически – в виде гистограммы.

В статье обосновывается новая форма представления статистических ранговых распределений и излагается порядок вычисления аппроксимирующих ранговых распределений.

Ранговые распределения занимают особое положение среди других распределений. Наиболее часто они используются в социально-гуманитарной сфере. Статистические ранговые распределения строятся путем упорядочения элементов выборочной совокупности по невозрастанию их частот (абсолютных или относительных). Такое упорядочение уже позволяет извлечь некоторую информацию из статистических данных.

Более существенную информацию о структуре выборочной совокупности можно получить при преобразовании рангового распределения к форме обычного одновершинного распределения. Поскольку наиболее полной характеристикой любой случайной величины является ее закон распределения, то отыскание этого закона следует считать главной задачей статистического исследования.

При решении этой задачи в настоящее время используется метод выдвижения гипотез об аппроксимирующем распределении и проверки каждой из них по критериям согласия, например по критерию “хи-квадрат” К.Пирсона. Действительно, он позволяет оценить степень близости статистического и аппроксимирующего распределений, но однозначного ответа на вопрос – является ли предлагаемое аппроксимирующее распределение *законом распределения* случайной величины – он не дает. Дело в том, что при использовании критерия “хи-квадрат” К.Пирсона принимают уровень значимости, т.е. вероятность α отклонения нулевой (выдвигаемой) гипотезы в пределах от 0,01 до 0,10. Такой низкий уровень

значимости не позволяет отклонить многие гипотезы. Если же поднять уровень значимости до 0,5–0,6, то будут отклоняться практически все гипотезы (при существующем подходе к решению этой задачи).

Для вычисления наиболее подходящего распределения необходимо использовать другой подход, который заключается в следующем. Поскольку различные статистические распределения обладают разными свойствами, они объединяются в разные классы. Для каждого класса разрабатываются универсальные (обобщенные) распределения или системы распределений. Далее в зависимости от свойств случайной величины выбирается соответствующая система непрерывных распределений, обладающая такими же свойствами, и *вычисляется* наилучшее аппроксимирующее распределение, которое можно считать законом распределения исследуемой случайной величины.

Три системы непрерывных распределений. Для описания статистических распределений, заданных на всей числовой оси, автором предложена первая система непрерывных распределений [3]:

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

где N – нормирующий множитель; $\alpha, \beta, k, u, l, \bar{t}$ – параметры.

Эта система предназначена для описания статистических распределений таких случайных величин, последующие значения которых получаются из предыдущих путем их изменения (сдвига) на всех интервалах на постоянную величину C (без изменения частот интервалов). Эти распределения содержат параметры сдвига $\alpha, l, \bar{t}$, при изменении которых аппроксимирующая кривая распределения перемещается по горизонтальной оси без изменения формы. Параметры формы k, u при сдвиге кривой не изменяются.

Из первой системы непрерывных распределений можно получить вторую систему, если принять $X=\ln T$ – для первой плотности и $T=\ln Y$ – для двух других плотностей:

$$\left.\begin{aligned} p(t) &= Nt^{k\beta-1}\left(1-\alpha ut^{\beta}\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N\left|\ln y - l\right|^{k-1}}{y}\left(1-\alpha u(\ln y - l)\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N}{y}\left(1-\alpha u(\ln y - \overline{\ln y})^2\right)^{\frac{1}{u}-1} \end{aligned}\right\}. \qquad (2)$$

Вторая система предназначена для описания статистических распределений таких неотрицательных случайных величин, последующие значения

логарифмов которых на всех интервалах получаются из предыдущих путем их изменения (сдвига) на постоянную величину *lnC*. В этом случае последующие значения случайной величины получаются из предыдущих умножением их на всех интервалах на постоянную С без изменения частот интервалов (при этом ширина интервалов и их границы увеличиваются в С раз). При определенных значениях параметров вторая система может описывать такие статистические ранговые распределения, как периодических изданий, упорядоченных по убыванию количества опубликованных в них статей по заданному предмету; книг, упорядоченных по убыванию числа выдач их читателям, и др.

Из второй системы при $T=lnY$ – для первой плотности и $lnY=lnlnW$ – для двух других плотностей следует третья система непрерывных распределений. При определенных значениях параметров третья система может описывать статистические ранговые распределения различных лексических единиц: лексем, словоформ, терминов, ключевых слов, словосочетаний и др.

Вычисление аппроксимирующих ранговых распределений. Статистические ранговые распределения обычно представляют в виде графика зависимости $p_r = f(r)$, где r – ранг события, а p_r – его относительная частота, либо в логарифмических координатах $\ln p_r = f(\ln r)$. Однако такая форма представления ранговых распределений содержит мало информации о них. Для извлечения более полной информации из статистических ранговых распределений автором предложена новая форма их представления [1], а именно:

$$rp_r = f(\ln r), \qquad (3)$$

т.е. по горизонтальной оси откладываются логарифмы рангов, а по вертикальной – произведения рангов на относительные частоты событий. В этом случае вместо убывающей кривой получается колоколообразная и, как правило, асимметричная кривая распределения. Она имеет моду С и две точки перегиба А, В.

Если ранговое распределение хорошо описывается второй системой непрерывных распределений, то в системе координат $rp_r = f(\ln r)$ оно аппроксимируется первой системой непрерывных распределений и обладает всеми свойствами последней.

Это означает, что в случае однородной выборки статистическая кривая распределения (3) будет иметь закономерное возрастание и убывание. Точки перегиба А и В будут расположены на равных расстояниях от моды.

При неоднородной выборке начало кривой распределения будет иметь несколько пиков и впадин. Последняя впадина перед закономерным ростом может служить границей, отделяющей неоднородную (левую) часть от однородной.

Для нахождения закона распределения необходимо из статистического ранжированного ряда удалить неоднородную часть (она обычно составляет от одного до нескольких десятков первых элементов), пересчитать ранги и относительные частоты событий и построить новый график зависимости (3).

Чтобы вычислить аппроксимирующее ранговое распределение, необходимо из графика снять 10–20 значений ординат rp_r при постоянной

ширине интервала $\ln r$. Сумма произведений всех ординат на ширину интервала должна быть равна единице. Дальнейшие расчеты выполняются по одной из компьютерных программ автора, которая используется в случае первой системы непрерывных распределений. Таким путем находится закон распределения для статистической кривой (3). Он задается плотностью *p(x)*, но поскольку *p(x)=tp(t)*, *x=lnt*, вычисленные оценки параметров плотности *p(x)* являются одновременно оценками параметров плотности *p(t)* (см. первые формулы систем непрерывных распределений (1) и (2)). Та же программа вычисляет координаты трех характерных точек A, C, B, которые приняты автором в качестве границ ядра и зон рассеяния [2].

Описанный метод вычисления теоретического рангового закона распределения может быть реализован при достаточно большом объеме выборки, когда последняя точка на статистической кривой распределения (3) расположена близко к горизонтальной оси.

В случае, если выборка небольшая и при этом задана эмпирическая функция распределения хотя бы при нескольких значениях рангов, можно использовать метод наименьших квадратов. Для описания ранговых распределений иногда подходит закон Вейбулла, который является частным случаем второй системы непрерывных распределений. Функция распределения и плотность вероятности его задаются формулами

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

где t может обозначать ранг события; α, β – параметры; $F(t)$ – накопленная вероятность первых t событий от начала частотного списка, т.е. функция распределения. При использовании метода наименьших квадратов функция распределения приводится к линейному виду $Y = \ln\alpha + \beta X$, где $Y=\ln\ln(1/(1-F(t)))$, $X=\ln t$. Оценки параметров β, α вычисляются по формулам

$$\beta=\frac{\overline{XY}-\bar{X}\bar{Y}}{\overline{X^2}-(\bar{X})^2}, \quad \alpha=e^{\bar{Y}-\beta\bar{X}}.$$

При известных оценках параметров аппроксимирующего рангового распределения легко вычисляются координаты характерных точек и, следовательно, ядро (журналов, книг библиотечного фонда и т.д.) и зоны рассеяния [3].

Описанный порядок исследования статистических ранговых распределений может быть применен при анализе структуры библиотечного фонда, степени его использования, при изучении читательских интересов по заданной тематической группе документов на основе анализа содержания ядра и зон рассеяния фонда.

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

2. Нешитой, В.В. Универсальные законы рассеяния и старения публи- каций / В.В.Нешитой // Веснік Бел. дзярж. ун-та культ. і маст. – 2007. – № 8. – С. 128–133.

3. Нешитой, В.В. Элементы теории обобщенных распределений: моно- графия / В.В.Нешитой. – Мн.: РИВШ, 2009. – 204 с.