УДК 801:519.2

ЗАВИСИМОСТЬ ПАРАМЕТРОВ ЛИНГВИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ ОТ ЧИСЛА ПОДВЫБОРОК И ИХ ОБЪЕМА

В. В. Нешитой

В работе автора [1] было показано, что эмпирические распределения частот отдельных слов в подвыборках одинакового объема (обозначим его через Δ) хорошо описываются обобщенным дискретным распределением I типа:

$$y_m = \left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \frac{\prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right]}{m!}\, y_{m=0},$$

$$m = 1, 2, \ldots, \qquad (1)$$

$$y_{m=0} = \frac{1}{\alpha u}\,[1+\alpha(1-u)x]^{\frac{-u}{1-u}},$$

где m — частота слова в отдельной подвыборке; y_m — количество подвыборок, в которых данное слово встретилось m раз; $x = \sum_{m \geqslant 1} m y_m$ — общая частота употребления данного слова во всех n обследованных подвыборках; α, u — параметры распределения ($\alpha u = 1/n$). Средняя частота здесь равна $M(m) = \alpha u x = x/n$.

Обобщенное распределение (1) включает как частные случаи: биномиальное распределение при $u > 1$; распределение Пуассона при $u \to 1$; отрицательное биномиальное распределение при $0 < u < 1$.

В той же работе было показано, что эмпирические распределения слов, употребляющихся в тексте независимо и случайно, при объеме подвыборки $\Delta = 1000$ словоупотреблений близки к закону Пуассона ($u \to 1$). Для зависимых слов, употребление которых связано с ситуацией, справедливо отрицательное биномиальное распределение ($0 < u < 1$).

Целью настоящей статьи является:

исследование взаимосвязей между параметрами обобщенного распределения I типа (1) и известных дискретных распределений;

исследование зависимости параметров обобщенного распределения от числа подвыборок n и их объема Δ;

установление оптимального объема подвыборки Δ, дающего для зависимых слов наибольшие отклонения параметра u от единицы.

1. ИССЛЕДОВАНИЕ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ

Установим связь обобщенного распределения I типа с известными дискретными распределениями. Для этого рассмотрим некоторые его частные случаи.

Случай 1 ($u > 1$). Связь с биномиальным распределением.

Рассмотрим отношение двух соседних ординат y_m / y_{m+1} или p_m / p_{m+1}. На основании (1) можем записать

$$\frac{y_m}{y_{m+1}} = \frac{P_m}{P_{m+1}} = \frac{m+1}{\frac{u}{u-1}-m} \cdot \frac{1-\alpha(u-1)x}{\alpha(u-1)x}. \qquad (2)$$

В то же время вероятность $p_{m,k}$ появления некоторого события A ровно m раз при k испытаниях (в нашей схеме — в подвыборке объемом $\Delta = k$) задается биномиальным законом распределения

$$P_{m,k} = \frac{k!}{m!\,(k-m)!}\, p^m q^{k-m} \quad (m = 0, 1, 2, \ldots, k;$$

$$M(m) = kp),$$

на основании которого находим

$$\frac{P_{m,k}}{P_{m+1,k}} = \frac{m+1}{k-m} \cdot \frac{q}{p}. \qquad (3)$$

Приравнивая правые части в формулах (2) и (3), получим

$$k = \Delta = \frac{u}{u-1}, \qquad (4)$$

$$p = 1-q = \alpha(u-1)x, \qquad (5)$$

или, с учетом равенств $\alpha u = 1/n$, $x/n = M(m)$,

$$p = \frac{x}{n}\left(\frac{u-1}{u}\right) = \frac{x}{\Delta n} = \frac{M(m)}{\Delta}. \qquad (5^1)$$

Из (4) можно выразить параметр u через $k = \Delta$:

$$u = \frac{k}{k-1} = \frac{\Delta}{\Delta-1}. \qquad (6)$$

Тогда параметр $\alpha = 1/nu$ выразится через величины n, k:

$$\alpha = \frac{1}{n}\left(1-\frac{1}{k}\right) = \frac{1}{n}\left(1-\frac{1}{\Delta}\right). \qquad (7)$$

Формулы (4)—(7) связывают параметры биномиального закона и распределения I типа (1).

Из формул (6), (7) следует, что *если некоторое событие A появляется независимо и случайно, то параметр u зависит только от объема подвыборки $k = \Delta$. С ростом объема подвыборки Δ параметр u уменьшается, асимптотически приближаясь к единице*, т. е. биномиальное распределение приближается к распределению Пуассона, а параметр α растет (при постоянном n) до своего предельного значения $\alpha = 1/n$. В то же время при постоянном объеме подвыборки Δ с ростом числа подвыборок n параметр α уменьшается обратно пропорционально n, а параметр u не зависит от n.

Случай 2 ($u \to 1$). Связь с законом Пуассона.

При $u \to 1$ распределение I типа (1) превращается в закон Пуассона

$$y_m = \frac{(\alpha x)^m}{m!\, \alpha e^{\alpha x}} \quad (m = 0,\ 1,\ 2,\ \ldots;\ M(m) = \alpha x), \tag{8}$$

для которого

$$\frac{y_m}{y_{m+1}} = \frac{P_m}{P_{m+1}} = \frac{m+1}{\alpha x}.$$

Это же отношение следует из (2) при $u \to 1$. Здесь величины m, k не ограничены справа. Действительно, из (4) при $u \to 1$ имеем $k = \Delta \to \infty$.

Закон Пуассона может описывать распределение n равновозможных событий, составляющих полную группу, по частоте их появления при x независимых испытаниях, либо распределение по частоте появления одного и того же события в n подвыборках равного объема, при этом вероятность появления события в каждой подвыборке одна и та же: $\alpha = 1/n$.

Случай 3 ($0 < u < 1$). Связь с отрицательным биномиальным распределением.

Для отрицательного биномиального распределения, которое можно рассматривать как распределение суммы k взаимно независимых геометрически распределенных случайных величин [2, с. 121] и которое задается формулой

$$P_{m,k} = \frac{(k+m-1)!}{m!\,(k-1)!}\, p^k q^m \left(m = 0,\ 1,\ \ldots;\ M(m) = k\frac{1-p}{p} \right),$$

данное отношение имеет вид:

$$\frac{P_{m,k}}{P_{m+1,k}} = \frac{m+1}{k+m} \cdot \frac{1}{q}. \tag{9}$$

Приравнивая правые части формул (2) и (9), найдем

$$k = \frac{u}{1-u}, \tag{10}$$

$$p = 1 - q = \frac{1}{1 + \alpha(1-u)x}, \tag{11}$$

или, с учетом равенств $\alpha = 1/nu$, $x/n = M(m)$,

$$p = \frac{1}{1 + \frac{x}{n}\frac{1-u}{u}} = \frac{1}{1 + M(m)\frac{1-u}{u}} = \frac{k}{M(m) + k}. \tag{11^1}$$

Из (10) выразим параметр u через k

$$u = \frac{k}{k+1}. \tag{12}$$

Тогда

$$\alpha = \frac{1}{nu} = \frac{k+1}{nk}, \tag{13}$$

где $n = \sum_{m \geqslant 0} y_m$.

При $k = 1$ имеем геометрическое распределение. В этом случае формула (12) дает: $u = 1/2$. *С ростом величины k параметр u растет до своего предельного значения* $u = 1$, т. е. отрицательное биномиальное распределение переходит в закон Пуассона, при этом $\alpha \to 1/n$. В то же время при постоянном значении k с ростом числа подвыборок n параметр α уменьшается обратно пропорционально n, а параметр u не зависит от n. Здесь необходимо помнить, что параметр k не равен объему подвыборки Δ, как в случае биномиального распределения.

Итак, *биномиальный и отрицательный биномиальный законы распределения вероятностей некоторого случайного события A с ростом величины k асимптотически приближаются к закону Пуассона с параметрами* $u \to 1$, $\alpha = 1/n$.

Проверим эти выводы на фактическом материале.

2. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ЗАКОН ПУАССОНА

В случае малых объемов подвыборок ($\Delta < 100$) распределение частот случайного события A описывается биномиальным законом с параметрами $u = \Delta/(\Delta - 1)$, $\alpha = 1/nu$, $x = \bar{m}n = \Delta pn$, где p — вероятность появления события A в одном испытании, n — количество подвыборок.

Рассмотрим пример распределения существительных по количеству употреблений в 400 фрагментах текста длиной $\Delta = 25$ словоупотреблений. Статистические данные, заимствованные из книги [3, с. 163], представлены в табл. 1 (столбцы 1 и 2).

Таблица 1

Распределение существительных по количеству употреблений в фрагментах текста длиной 25 словоупотреблений (n=400)

Число появлений события, m	Фактическое число подвыборок, y_m	$y_m^{\text{расч}}$ при		
		$u=1{,}03182$ $\alpha=0{,}002423$ $x=3035$	$u=25/24$ $\alpha=0{,}0024$ $x=3035$	$u \to 1$ $\alpha=0{,}0025$ $x=3035$
0	0	0,07	0,05	0,20
1	3	0,70	0,52	1,54
2	5	3,35	2,70	5,83
3	7	10,38	9,01	14,76
4	24	23,33	21,58	27,99
5	40	40,52	39,50	42,48
6	52	56,58	57,38	53,72
7	64	65,25	67,86	58,23
8	66	63,35	66,53	55,22
9	47	52,53	54,76	46,56
10	48	37,59	38,18	35,32
11	24	23,41	22,69	24,37
12	14	12,77	11,53	15,41
13	3	6,13	5,03	8,99
14	2	2,60	1,88	4,87
15	1	0,98	0,60	2,47
16		0,32	0,16	1,17
17		0,10	0,04	0,52
18		0,03	0,01	0,22
19		0,01		0,09
20				0,03

Найдем выравнивающее дискретное распределение. По данным табл. 1, имеем:

$$\alpha = \frac{1}{x^2}\sum_{m \geqslant 1} m^2 y_m - \frac{1}{x} = 0{,}002423,\quad u = \frac{1}{\alpha n} = 1{,}03182,$$

где $x = \sum_{m \geqslant 1} m y_m = 3035$, $n = \sum_{m \geqslant 0} y_m = 400$. Выравнивающим является биномиальное распределение ($u > 1$), которое относится к I типу. Расчетные значения y_m, вычисленные по формулам

$$y_{m+1} = y_m \frac{\alpha x\,[u + m(1-u)]}{[1+\alpha(1-u)\,x]\,(m+1)},\quad m = 0,\ 1,\ \ldots,$$

$$y_{m=0} = \frac{1}{\alpha u}\,[1+\alpha(1-u)\,x]^{\frac{-u}{1-u}},$$

$$y_{m=1} = x\,[1+\alpha(1-u)\,x]^{-\frac{1}{1-u}},$$

которые следуют из (1), приведены в табл. 1 (столбец 3). Они хорошо согласуются с опытными данными.

Следует отметить, что если существительные появляются в тексте независимо и случайно, то параметр u при объеме подвыборки $\Delta = 25$ должен быть равен $u = \Delta/(\Delta-1) = 25/24 = 1{,}04167$, а параметр $\alpha = 1/nu = 0{,}0024$, что весьма близко к найденным из опыта оценкам этих параметров. Расчетные значения y_m (при $\alpha = 25/24$, $\alpha = 0{,}0024$) также близки к опытным данным (см. табл. 1, столбец 4).

Наконец, в столбце 5 приведены значения y_m, рассчитанные по закону Пуассона (8) при $\alpha = 1/n = 0{,}0025$, $n = 3035$, $\bar{m} = 7{,}5875$. Как видно из табл. 1 (столбцы 4 и 5), биномиальное распределение при $\Delta = 25$, $\bar{m} = 7{,}5875$ обнаруживает некоторую близость к закону Пуассона. Однако, как показывают расчеты, с последующим ростом объема подвыборки (например, до $\Delta = 100$ словоупотреблений) сближения обоих распределений практически не наблюдается, поскольку пропорционально Δ растет средняя частота $\bar{m} = x/n$. И лишь при весьма больших значениях Δ оба распределения сближаются, что следует из формул (6), (7): при $\Delta \to \infty$ параметр $u \to 1$, а параметр $\alpha \to 1/n$ как в случае распределения Пуассона.

3. ОТРИЦАТЕЛЬНОЕ БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ЗАКОН ПУАССОНА

Воспользуемся статистическими данными К. Б. Бектаева и К. Ф. Лукьяненкова [4, с. 67—69] о распределении частот указательного местоимения *this* по подвыборкам объемом 1000, 2000 и 4000 словоупотреблений. Обработка этих рядов распределения в восьми сериях по $n_1 = 50$ подвыборок в каждой (объемом $\Delta = 1000$ словоупотреблений) и в четырех сериях по $n_2 = 100$ подвыборок дала следующие результаты: в первом случае $\bar{\alpha}_1 = 0{,}022654$, $u_1 = 1/\bar{\alpha}_1 n_1 = 0{,}88284$; во втором случае $\bar{\alpha}_2 = 0{,}011347$, $u_2 = 1/\bar{\alpha}_2 n_2 = 0{,}88127$. Отношение параметров $\bar{\alpha}_1/\bar{\alpha}_2 = 1{,}99647$, что близко к отношению чисел подвыборок $n_2/n_1 = 100/50 = 2$, как и должно быть согласно теоретическим выводам.

Исследуем далее поведение параметров α, u распределения I типа, а также k, p отрицательного биномиального распределения с ростом объема подвыборки Δ. На основании опытных данных, приведенных в той же работе, были проведены соответствующие расчеты, результаты которых сведены в табл. 2. Из этой таблицы видно, что с ростом объема подвыборки Δ в два раза параметр k растет примерно в $\sqrt{2}$ раза, т. е. значительно медленнее величины Δ. Параметр α здесь также растет (в 1,94 раза), поскольку количество подвыборок n уменьшается (в два раза), что качественно согласуется с теорией. Однако параметр u растет не так быстро, как предсказывает теория при условии постоянной вероятности p отрицательного биномиального распределения, или, что то же, при росте k пропорционально объему подвыборки Δ. Как видно из табл. 2, вероятность p с ростом Δ довольно быстро уменьшается, что и сдерживает рост параметра u.

При постоянной вероятности p (или при росте k пропорционально объему подвыборки Δ) рост параметра u был бы более быстрым, чем при уменьшении p с ростом Δ (две последние строки в табл. 2).

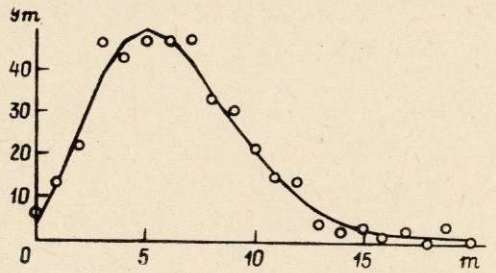
Таблица 2

Зависимость параметров дискретного распределения I типа от объема подвыборок

Параметры распределения	Значения параметров при объеме подвыборок		
	$\Delta_1 = 1000$	$\Delta_2 = 2000$	$\Delta_3 = 4000$
α	0,0028383	0,0055161	0,0107021
u	0,88081	0,90643	0,93440
n	400	200	100
x	2541	2541	2541
$\bar{m} = x/n$	6,3525	12,705	25,41
$k = \frac{u}{1-u}$	7,3900	9,6872	14,2439
$p = \frac{k}{M(m)+k}$	0,53774	0,43261	0,35919
u при $p = 0{,}53774$	0,88081	0,93663	0,96728
k при $p = 0{,}53774$	7,39	14,78	29,56

Итак, параметр u обобщенного дискретного распределения I типа не зависит от количества подвыборок n, но с ростом объема подвыборки Δ асимптотически приближается к единице. Это значит, что *с ростом объема подвыборки Δ эмпирические распределения частот отдельного слова или класса слов приближаются к закону Пуассона.*

На рисунке приведены графики эмпирического и выравнивающего распределений частот указательного



Распределение частот указательного местоимения *this* в 400 подвыборках объемом Δ_1-1000 словоупотреблений

местоимения *this* в подвыборках объемом $\Delta_1 = 1000$ словоупотреблений. Оценки параметров выравнивающего распределения даны в табл. 2 (столбец 2, первые четыре строки).

4. ОПТИМАЛЬНЫЙ ОБЪЕМ ПОДВЫБОРКИ

На основе анализа данных, приведенных в табл. 1 и 2, можно утверждать, что существует некоторый оптимальный объем подвыборки $\Delta_{опт}$, при котором распределения частот слов, употребляющихся независимо и случайно, описываются биномиальным законом (с па-

ietром $u=\Delta/\Delta(-1)\approx 1$), близким к закону Пуассо- ($u\to 1$), а при не вполне случайном употреблении — ицательным биномиальным распределением ($0<$ <1); оптимальный объем подвыборки для зависи- х слов дает наибольшие отклонения параметра u единицы (точнее, от $u=\Delta/(\Delta-1)$).

Какой объем целесообразно ограничить интервалом $\leqslant\Delta_{опт.}\leqslant 1000$ словоупотреблений.

Отклонение параметра u, выравнивающего распреде- ия от единицы, может служить показателем сте- и неравномерности употребления данного слова и в оторой мере — показателем степени семантической рузки. Поскольку параметр u не зависит от коли- тва подвыборок n, то для достижения большей точ- ти оценок этого параметра целесообразно увеличи- ь значения n, принимая объем подвыборки Δ одина- ым для всех исследуемых слов.

ЛИТЕРАТУРА

1. Нешитой В. В. Ранжирование слов по степени семантической нагрузки//НТИ. Сер. 2.—1986.—№ 4. — С. 20—25.
2. Поллард Дж. Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982.—344 с.
3. Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. Математическая лингвистика. — М.: Высшая школа, 1977.—383 с.
4. Бектаев К. Б., Лукьяненков К. Ф. О законах распределения единиц письменной речи//Статистика речи и автоматический анализ текста. — Л.: Наука, 1971. — С. 47—112.

Статья поступила в редакцию 31 октября 1986 г.