Министерство культуры Республики Беларусь
Учреждение образования
«Белорусский государственный университет
культуры и искусств»

В. В. Нешитой

ИНФОРМЕТРИЯ: МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ ИССЛЕДОВАНИЯ

Минск
БГУКИ
2017

УДК 519.2:[002.1+02]
ББК 22.17+78.60
Н 597

Р е ц е н з е н т ы :
П. В. Гляков, заведующий кафедрой информационных технологий
в культуре БГУКИ, кандидат физико-математических наук, доцент;
А. Г. Буравкин, ведущий научный сотрудник
Объединенного института проблем информатики
НАН Беларуси, кандидат технических наук, доцент

Рекомендовано к изданию научно-техническим советом учреждения образования «Белорусский государственный университет культуры и искусств» (протокол № 7 от 30.03.2017 г.)

Нешитой, В. В.

Н597 Информетрия: математические модели и методы исследования / В. В. Нешитой ; М-во культуры Респ. Беларусь, Белорус. гос. ун-т культуры и искусств. – Минск : БГУКИ, 2017. – 275 с.
ISBN 978-985-522-173-0.

Излагается разработанная автором теория обобщенных четырехпараметрических непрерывных распределений, включающая бесконечное множество распределений в виде 4-х систем непрерывных распределений, системы дискретных распределений, взаимосвязанной с системой кривых роста новых событий, новые методы вычисления законов распределения и оценок параметров – универсальный метод моментов и общий устойчивый метод, что гарантирует высокую точность аппроксимации статистических непрерывных распределений, в том числе ранговых. Впервые автором установлены универсальные законы старения публикаций (первая система непрерывных распределений) и рассеяния (вторая система непрерывных распределений), дано обоснование закона Лотки на базе дискретного распределения 3-го типа автора, показано, что закона Ципфа не существует вовсе. Вместо него предлагается 2-я система непрерывных распределений, в том числе частный ее случай – закон Вейбулла. Вводится множество новых информетрических показателей, которые могут быть использованы на практике.

УДК 519.2:[002.1+02]
ББК 22.17+78.60

ISBN 978-985-522-173-0

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

При исследовании случайных величин в математической статистике используется выборочный метод. Он заключается в том, что из генеральной совокупности отбирается выборка объемом, как правило, не менее 100 единиц. При этом она должна правильно отражать пропорции генеральной совокупности, т. е. быть представительной (репрезентативной). Только в этом случае результаты исследования выборки могут быть распространены на всю генеральную совокупность.

Чтобы извлечь информацию из выборки, которая представляет собой простой статистический ряд, необходимо упорядочить все значения исследуемой случайной величины либо по возрастанию, либо по убыванию и построить интервальный ряд распределения или ранжированный ряд. Далее вычисляются числовые характеристики случайной величины, которые используются при вычислении оценок параметров аппроксимирующего (выравнивающего) распределения. Вид последнего устанавливается как правило методом выдвижения гипотез. Ясно, что таких гипотез может быть несколько.

При вычислении оценок параметров наиболее часто используются классический метод моментов К. Пирсона и метод наибольшего правдоподобия Р. Фишера. К сожалению, недостатком этих методов является то, что они требуют составления и решения системы n уравнений с n неизвестными (по числу параметров распределения). В результате может получиться весьма сложная система уравнений, которую практически невозможно решить. При этом основная задача статисти-

ческого исследования: установление закона распределения, который наиболее полно характеризует случайную величину, – остается нерешенной.

Проблема установления закона распределения случайной величины по выборочным данным к настоящему времени до конца не решена. Исследования автора показали, что для однозначного и более точного решения этой задачи следует не выдвигать гипотезы об аппроксимирующем распределении, а вычислять его по статистическим данным. Правда, для этого необходимо иметь универсальные (обобщенные) распределения, включающие как частные случаи все или почти все известные распределения, в том числе семейство кривых К. Пирсона, а также методы и алгоритмы установления типа аппроксимирующего распределения и вычисления оценок его параметров.

Для решения таких задач автором разработана теория обобщенных распределений (ТОР) [23], которая включает четыре системы непрерывных распределений, заданные четырехпараметрическими плотностями, систему дискретных распределений, взаимосвязанную с системой кривых роста новых событий, методы и алгоритмы установления типа аппроксимирующей кривой и вычисления оценок параметров (универсальный метод моментов и общий устойчивый метод), номограммы для быстрого оценивания типа кривой распределения и нахождения в первом приближении оценок двух параметров формы (при ручном счете) и серию компьютерных программ по всем разделам теории (под общим названием SNR – для работы с системами непрерывных распределений и SDR – для работы с системой дискретных распределений, а также другие программы). Обобщенные распределения способны с высокой точностью описывать практически все многообразие статистических

распределений однородных случайных величин, в том числе ранговых.

Эта теория позволяет легко решать многие задачи. При этом в некоторых случаях могут быть использованы классические методы оценивания параметров (метод моментов, метод наименьших квадратов), а в общем случае – разработанные автором универсальный метод моментов и общий устойчивый метод. На базе этих методов могут быть разработаны другие методы. Предлагаемые методы отличаются от известных тем, что задача по вычислению типа наилучшего аппроксимирующего распределения и оценок его параметров разбивается на два этапа. На первом этапе разрабатываются два критерия, зависящие от двух параметров формы. Они позволяют вычислять тип выравнивающей кривой и оценки параметров формы путем решения системы двух уравнений с двумя неизвестными. На втором этапе вычисляются оценки двух других параметров по простым формулам (при известных оценках параметров формы).

Отныне для установления теоретического закона распределения непрерывной случайной величины по ее статистическому распределению не требуется выдвижения многочисленных гипотез о выравнивающей кривой и проверки каждой из них по критериям согласия. Система непрерывных распределений выбирается в зависимости от свойств случайной величины, а тип распределения и оценки его параметров определяются расчетом.

При этом основная сложность заключается в разработке двух критериев для установления типа выравнивающей кривой и вычисления оценок параметров формы. Для разработки таких критериев система четырехпараметрических непрерывных распределений случайной величины X сводится к системе двухпараметрических распределений случайной величины Z. Затем ис-

пользуются взаимосвязи между случайными величинами X и Z и их плотностями распределения $p(x)$, $p(z)$.

Необходимый минимум сведений для успешной работы с настоящей книгой можно найти в учебно-методическом пособии автора [22]. Там же изложены: классический и универсальный методы моментов, метод наименьших квадратов и метод максимального правдоподобия.

На практике для решения большинства задач достаточно использовать две или три системы непрерывных распределений.

1. НЕКОТОРЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. СЛУЧАЙНЫЕ СОБЫТИЯ И ИХ ВЕРОЯТНОСТИ

1.1. Случайные события. Испытания. Относительная частота и вероятность

Пусть требуется оценить качество изделий в некоторой партии объемом n. Для этого необходимо над каждым изделием провести наблюдение, т. е. осмотр, измерение, взвешивание и т. д. В теории вероятностей и математической статистике всем этим понятиям соответствует один термин – испытание.

В результате отдельного испытания изделие может быть признано либо годным, либо браком. Возможные исходы испытания в данном примере – это случайные события: A – годное изделие; B – брак. Эти события называются случайными, потому что заранее нельзя точно предсказать, какое из них наступит при следующем испытании.

Пусть после проверки всей партии изделий объемом n, т. е. после n испытаний, случайное событие A – число годных изделий – появилось n_A раз. Это значит, что относительная частота случайного события A равна

$$w_A = n_A / n .$$

Если провести несколько серий испытаний (проверить несколько партий изделий), то относительные частоты в разных сериях будут группироваться около определенного числа, которое называется вероятностью

случайного события A и обозначается $P(A)$. Как показала практика, с ростом объема партии изделий n относительные частоты теснее группируются около вероятности, т. е. обнаруживают устойчивость.

Устойчивость относительной частоты случайного события является определяющим его свойством, позволяющим использовать относительную частоту как оценку вероятности в различных практических расчетах.

1.2. Виды случайных событий

События, которые непременно происходят при каждом испытании, называются достоверными.

События, которые не могут произойти ни при каком испытании, называются невозможными.

Вероятность достоверного события равна единице, вероятность невозможного события равна нулю.

Если при осуществлении испытания может наступить хотя бы одно из двух событий A или B, то событие

$$C = A + B$$

называется суммой, или объединением событий A и B.

Два события A и B называются несовместными, если они не могут наступить вместе при одном испытании.

Случайные события образуют полную группу, если они попарно несовместны и при любом отдельном испытании непременно должно произойти одно из них.

Сумма вероятностей событий, образующих полную группу, равна единице.

Два случайных события называются противоположными, если в одном испытании появление одного из них (A) исключает появление другого ($\overline{A}$ – читается не A).

Сумма вероятностей двух противоположных событий равна единице

$$P(A) + P(\overline{A}) = 1.$$

Противоположные события образуют полную группу.

Если при осуществлении испытания может наступить и событие A, и событие В (совмещение событий А и В), то событие

$$C = A \cdot B$$

называется произведением, или пересечением событий A и B.

Два случайных события называются независимыми, если при осуществлении испытаний появление одного из них не изменяет вероятности появления другого.

1.3. Определения вероятности

Классическое определение вероятности события A – отношение числа m элементарных событий (исходов испытаний), благоприятствующих событию A, к общему числу n равновозможных элементарных событий

$$P(A) = \frac{m}{n}.$$

Статистическое определение вероятности

$$P(A) = \frac{n_A}{n},$$

где n_A – частота события A при n испытаниях.

Геометрическая вероятность

$$P(A) = \frac{S_A}{S},$$

где S_A – площадь некоторого замкнутого контура, составляющая часть площади S.

1.4. Основные формулы комбинаторики

Используются для вычисления вероятностей событий.

Перестановки – комбинации из n различных элементов, отличающиеся лишь порядком. Число перестановок вычисляется по формуле

$$P_n = n! = 1.2...n.$$

Число перестановок из n элементов по m, где каждый элемент может использоваться от 0 до m раз, равно

$$P_{n,m} = n^m.$$

Например, при $n=2$, $m=8$, $P_{n,m}=2^8=256$.

Размещения - комбинации из n различных элементов по m элементов, которые различаются либо составом элементов, либо их порядком

$$A_n^m = n(n-1)(n-2)...(n-m+1) = \frac{n!}{(n-m)!}.$$

Сочетания – комбинации из n различных элементов по m элементов, различающиеся хотя бы одним элементом

$$C_n^m = \frac{n!}{m!(n-m)!} = \frac{A_n^m}{P_m} = \frac{n(n-1)...(n-m+1)}{1.2...m} = C_n^{n-m}.$$

При этом $C_n^0 = C_n^n = 1$.

Пример.

В партии из N деталей M стандартных. Выбираются n деталей. Требуется найти вероятность того, что m деталей будут стандартными.

Решение. Общее число возможных исходов равно числу способов, которыми можно взять n деталей из N. Это число равно числу сочетаний из N по n, т. е. C_N^n.

Найдем далее число благоприятствующих исходов. Поскольку m стандартных деталей выбираются из общего их числа M, то число таких комбинаций равно C_M^m. Остальные n-m нестандартных деталей выбираются из N–M нестандартных деталей – это C_{N-M}^{n-m} комбинаций. Число благоприятствующих исходов равно произведению $C_M^m C_{N-M}^{n-m}$. Следовательно,

$$P(A=m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

Это – известное гипергеометрическое распределение.

1.5. Основные теоремы теории вероятностей

1.5.1. Теорема сложения вероятностей (несовместных событий)

Пусть A и B – несовместные события. Вероятность суммы двух несовместных событий равна сумме вероятностей этих событий

$$P(A+B) = P(A)+P(B).$$

Для нескольких несовместных событий имеем

$$P(A_1 + A_2 + ... + A_n) = P(A_1) + P(A_2) + ... + P(A_n).$$

Для совместных событий

$$P(A+B) = P(A)+P(B)-P(AB),$$

где $P(AB)$ – вероятность совместного появления событий A и B.

1.5.2. Теорема умножения вероятностей (независимых событий)

Вероятность произведения (совмещения) двух независимых событий равна произведению вероятностей этих событий

$$P(AB) = P(A)P(B).$$

Вероятность произведения двух зависимых событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную при условии, что первое имело место

$$P(AB) = P(A)P(B/A) = P(B)P(A/B).$$

Пример. В урне 2 белых и 3 черных шара. Вынимаем подряд 2 шара. Какова вероятность того, что оба шара белые, т.е. $A = A_1A_2$.

Решение. A_1 – появление белого шара при 1-м испытании; A_2 – появление белого шара при 2-м испытании

$$P(A) = P(A_1)P(A_2 / A_1) = \frac{2}{5} \quad \frac{1}{4} = 0{,}1.$$

Следствие теоремы умножения вероятностей.

Вероятность появления хотя бы одного события из событий A_1, $A_2, \ldots, A_n$, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$:

$$P(A) = 1 - q_1 \cdot q_2 \ldots \; q_n .$$

В частном случае, при $q_1 = q_2 = \ldots = q_n = q$

$$P(A) = 1 - q^n .$$

Пример. Вероятности попадания в цель каждого из трех стрелков равны: p_1=0,8; p_2=0,7; p_3=0,9. Найти вероятность хотя бы одного попадания при одном залпе.

Решение. Вероятности промахов равны: $q_1 = 1 - p_1 = 0{,}2; \quad q_2 = 0{,}3; \; q_3 = 0{,}1$. Следовательно,

$$P(A) = 1 - q_1 q_2 q_3 = 0{,}994 .$$

1.5.3. Формула полной вероятности

Следствием теорем сложения и умножения вероятностей является формула полной вероятности.

Пусть некоторое событие A может произойти вместе с одним из событий H_1, $H_2, \ldots, H_n$, причем последние образуют полную группу несовместных событий. Их называют гипотезами.

Приведем без доказательства формулу для вычисления вероятности события A. Она равна сумме произведений вероятности каждой гипотезы на вероятность события при этой гипотезе [2].

$$P(A) = \sum_{i=1}^{n} P(H_i) P(A / H_i).$$

Другими словами, **формула полной вероятности определяет средневзвешенную по всем гипотезам вероятность наступления некоторого события** A.

Пример. Есть два набора деталей. Вероятность того, что деталь первого набора стандартна, равна 0,8, а второго – 0,9. Найти вероятность того, что взятая наудачу деталь из наудачу взятого набора стандартна, т. е. $P(A)$=?

Решение. Событие H_1 – деталь взята из первого набора; $P(H_1)$=1/2.

Событие H_2 – деталь взята из второго набора; $P(H_2)$=1/2.

Далее, вероятность события A при первой гипотезе $P(A/H_1)$=0,8; при второй гипотезе $P(A/H_2)$=0,9.

Средневзвешенная вероятность события A по двум гипотезам равна

$$P(A) = 0{,}8 \cdot 0{,}5 + 0{,}9 \cdot 0{,}5 = 0{,}85.$$

1.5.4. Теорема гипотез (формула Бейеса)

Пусть имеется полная группа несовместных гипотез $H_1, \ldots, H_n$, при этом вероятности их равны $P(H_i)$.

Кроме того, известны вероятности некоторого события A, которое может произойти совместно с каждой гипотезой.

Пусть в результате опыта наступило событие A.

Тогда распределение условных вероятностей гипотез при наступлении события A задается формулой Бейеса

$$P(H_i / A) = \frac{P(H_i)P(A / H_i)}{\sum_{i=1}^{n} P(H_i)P(A / H_i)}.$$

Рассмотрим пример на формулу полной вероятности и формулу Бейеса, позаимствованный из [3].

Три станка выпускают одинаковые детали (см. табл.).

№ станка	Дневная выработка (деталей)	Вероятность гипотезы	% брака	Вероятность брака
1	m_1=600	$P(H_1)=0,6$	3	$P_1=P(A/H_1)=0,03$
2	m_2=100	$P(H_2)=0,1$	5	$P_2=P(A/H_2)=0,05$
3	m_3=300	$P(H_3)=0,3$	10	$P_3=P(A/H_3)=0,10$
	$\sum m_i = 1000$			

На складе продукция трех станков смешивается. Далее выбирается случайным образом одна деталь.

Требуется:

а) *найти вероятность того, что она бракованная*.

Здесь используется формула полной вероятности.

Вероятности гипотез равны (см. табл.):

$$P(H_1) = \frac{m_1}{m_1 + m_2 + m_3} = \frac{600}{1000} = 0,6;$$
$$P(H_2) = 0,1;\ P(H_3) = 0,3.$$

Вероятности брака при каждой гипотезе равны:

$$P(A/H_1) = 0,03; \quad P(A/H_2) = 0,05; \quad P(A/H_3) = 0,10.$$

Тогда

$$P(A) = \sum_{i=1}^{3} P(H_i)P(A/H_i) = 0,6 \cdot 0,03 + 0,1 \cdot 0,05 + 0,3 \cdot 0,1 = 0,053.$$

Формула полной вероятности в данном примере определяет средневзвешенную вероятность брака по трем станкам. Действительно, записав ее в более простом виде, находим

$$P(A) = \frac{m_1 p_1 + m_2 p_2 + m_3 p_3}{m_1 + m_2 + m_3} = \overline{p},$$

где $p_i = P(A/H_i)$; $\quad \frac{m_i}{m_1 + m_2 + m_3} = P(H_i)$.

б) *Найти вероятность того, что случайно отобранная деталь, оказавшаяся бракованной, выпущена первым станком*.

По формуле Бейеса находим

$$P(H_1 / A) = \frac{P(H_1)\ P(A/H_1)}{\sum_{i=1}^{n} P(H_i)\ P(A/H_i)} = \frac{0{,}6 \cdot 0{,}3}{0{,}6 \cdot 0{,}3 + 0{,}1 \cdot 0{,}05 + 0{,}3 \cdot 0{,}1} = \frac{0{,}018}{0{,}053} = 0{,}34 .$$

Формула Бейеса в данном примере определяет долю бракованных изделий (в общем объеме брака), изготовленных одним *i*-м станком.

$$P(H_1 / A) = \frac{m_1 p_1}{m_1 p_1 + m_2 p_2 + m_3 p_3} = \frac{600 \cdot 0{,}03}{600 \cdot 0{,}03 + 100 \cdot 0{,}05 + 300 \cdot 0{,}1} = \frac{18}{53} = 0{,}034 .$$

То же для других станков

$$P(H_2 / A) = \frac{5}{53} = 0{,}094; \qquad P(H_3 / A) = \frac{30}{53} = 0{,}566 .$$

1.6. Дискретные случайные величины

Для случайных величин приняты обозначения *X, Y, Z,…*

Возможные значения случайной величины *X* обозначаются строчными буквами $x_1, x_2, \ldots, x_n$.

Дискретной называют случайную величину, которая принимает отдельные изолированные возможные значения с определенными вероятностями (например, число отказавших приборов) в отличие от непрерывной случайной величины, которая может принимать все значения из некоторого конечного или бесконечного промежутка (например, время безотказной работы прибора).

Возможные значения прерывных (дискретных) величин могут быть заранее перечислены, а непрерывных – не могут быть перечислены.

1.6.1. Закон распределения вероятностей дискретной случайной величины

Закон распределения случайной величины – это соответствие между возможными значениями случайной величины и их вероятностями.

Его можно задать таблично, аналитически и графически:

а) табличная форма закона распределения в виде ряда распределения

X	x_1	x_2	...	x_n
p	p_1	p_2	...	p_n

б) аналитическая форма

$$p_i = f(x_i).$$

в) графическая форма – в виде многоугольника распределения.

На оси абсцисс откладываются значения случайной величины и строятся отрезки, равные по высоте вероятностям. Вершины отрезков для наглядности соединяются ломаной.

1.6.2. Числовые характеристики дискретной случайной величины

Математическое ожидание

Математическое ожидание дискретной случайной величины равно сумме произведений всех ее возможных значений на их вероятности

$$M(X) = x_1 p_1 + x_2 p_2 + ... + x_n p_n.$$

Математическое ожидание есть неслучайная (постоянная) величина.

Пример 1. Найти математическое ожидание случайной величины X по ее закону распределения:

X	3	5	2
P	0,1	0,6	0,3

Решение. $M(X) = 3 \cdot 0{,}1 + 5 \cdot 0{,}6 + 2 \cdot 0{,}3 = 3{,}9$.

Пример 2. Найти математическое ожидание числа появлений события А в одном испытании, если вероятность события A равна p.

Решение. Случайная величина X – число появлений события A в одном испытании – может принимать два значения: $x_1 = 1$ (событие наступило) с вероятностью p и $x_2 = 0$ (событие не наступило) с вероятностью $1-p=q$.

Следовательно,

$$M(X) = 1 \cdot p + 0 \cdot q = p,$$

т. е. математическое ожидание числа появлений события в одном испытании равно вероятности этого события.

Оценкой математического ожидания является среднее арифметическое наблюденных значений случайной величины.

Свойства математического ожидания

Приведем без доказательства основные свойства математического ожидания.

1. $M(C)=C$ – математическое ожидание постоянной величины C равно значению самой постоянной.

2. $M(CX)=CM(X)$ – постоянную величину можно выносить за знак математического ожидания.

3. $M(XY)=M(X)M(Y)$ – для двух независимых случайных величин математическое ожидание произведения равно произведению их математических ожиданий.

4. $M(X+Y)=M(X)+M(Y)$ – для двух случайных величин (зависимых или независимых) математическое ожидание суммы равно сумме математических ожиданий слагаемых.

Математическое ожидание числа появлений события A в n независимых испытаниях равно произведению числа испытаний n на вероятность появления события в каждом испытании p, т. е.

$$M(X)=np.$$

Доказательство свойств математического ожидания см., например, в учебнике [4].

Дисперсия дискретной случайной величины

Две случайные величины могут иметь одинаковые математические ожидания, но разное рассеяние. Это значит, что математическое ожидание полностью случайную величину не характеризует. Поэтому вводится еще одна числовая характеристика, которая называется дисперсией и характеризует рассеяние случайной величины относительно математического ожидания.

Дисперсией дискретной случайной величины X называется математическое ожидание квадрата отклонения случайной величины X от ее математического ожидания

$$D(X) = M[X - M(X)]^2.$$

Пример. Случайная величина X имеет распределение

X	1	2	5
p	0,3	0,5	0,2

Требуется вычислить дисперсию.

Имеем:

$$M(X) = 1 \cdot 0{,}3 + 2 \cdot 0{,}5 + 5 \cdot 0{,}2 = 2{,}3.$$

$$D(X) = \sum_{i=1}^{3} (x_i - M(X))^2 p_i = (1 - 2{,}3)^2 \cdot 0{,}3 + (2 - 2{,}3)^2 \cdot 0{,}5 +$$

$$+ (5 - 2{,}3)^2 \cdot 0{,}2 = 2{,}01.$$

На практике для вычисления дисперсии используют другую, более удобную формулу

$$D(X) = M(X^2) - [M(X)]^2.$$

Доказательство:

$$\mathrm{D(X)} = \mathrm{M[X - M(X)]^2} = M\left[X^2 - 2XM(X) + (M(X))^2\right] =$$

$$= M(X^2) - 2M(X)M(X) + [M(X)]^2 = M(X^2) - [M(X)]^2.$$

Свойства дисперсии

1. $D(C)=0$, т. е. дисперсия постоянной величины C равен нуля.

2. $D(CX)=C^2D(X)$ – постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат.

Доказательство:

$$D(CX)=M[CX-M(CX)]^2=M\left\{C^2[X-M(X)]^2\right\}=C^2D(X).$$

3. $D(X+Y)=D(X)+D(Y)$ – дисперсия суммы двух случайных величин равна сумме их дисперсий.

Доказательство:

$$D(X+Y)=M(X+Y)^2-[M(X+Y)]^2=M\left(X^2+2XY+Y^2\right)-[M(X)+M(Y)]^2=$$
$$=M(X^2)+2M(XY)+M(Y^2)-\left[(M(X))^2+2M(X)M(Y)+(M(Y))^2\right].$$

Так как $M(XY)=M(X)M(Y)$, то последнее равенство примет вид

$$D(X+Y)=M(X^2)-[M(X)]^2+M(Y^2)-[M(Y)]^2,$$

откуда

$$D(X+Y)=D(X)+D(Y).$$

Приведем без доказательства еще два свойства дисперсии.

4. $D(C+X)=D(X)$.

5. $D(X-Y)=D(X)+D(Y)$.

Отметим еще одно важное свойство дисперсии.

Дисперсия числа появлений события A в n независимых испытаниях равна

$$D(X)=npq.$$

Доказательство:

Найдем дисперсию числа появлений события А в одном испытании

$$D(X_1)=M(X_1^2)-[M(X_1)]^2.$$
$$M(X_1)=1\cdot p+0(1-p)=p.$$
$$M(X_1^2)=1^2\cdot p+0^2(1-p)=p.$$

Тогда $D(X_1)=p-p^2=p(1-p)=pq$.

Всего n испытаний, следовательно, $D(X)=npq$.

Дисперсия имеет размерность случайность величины в квадрате.

Среднее квадратическое отклонение

Если извлечь из дисперсии квадратный корень, получим среднее квадратическое отклонение

$$\sigma(X) = \sqrt{D(X)}.$$

Размерность величины $\sigma(X)$ та же, что и случайной величины X.

Пример. По распределению

X	2	3	10
p	0,1	0,4	0,5

требуется вычислить среднее квадратическое отклонение.

Решение.

$$M(X) = 2 \cdot 0{,}1 + 3 \cdot 0{,}4 + 10 \cdot 0{,}5 = 6{,}4.$$
$$M(X^2) = 2^2 \cdot 0{,}1 + 3^2 \cdot 0{,}4 + 10^2 \cdot 0{,}5 = 54.$$
$$D(X) = M(X^2) - [M(X)]^2 = 54 - 6{,}4^2 = 13{,}04.$$
$$\sigma(X) = \sqrt{D(X)} = 3{,}61.$$

Среднее квадратическое отклонение суммы взаимно независимых случайных величин равно

$$\sigma(X_1 + X_2 ... + X_n) = \sigma(X) = \sqrt{\sigma^2(X_1) + \sigma^2(X_2) + ... + \sigma^2(X_n)}$$

Доказательство.

Дисперсия суммы случайных величин равна

$$D(X) = D(X_1) + D(X_2) + ... + D(X_n).$$

Тогда

$$\sqrt{D(X)} = \sigma(X) = \sqrt{D(X_1) + ... + D(X_n)} = \sqrt{\sigma^2(X_1) + ... + \sigma^2(X_n)}.$$

1.6.3. Одинаково распределенные взаимно независимые случайные величины

Рассмотрим n взаимно независимых одинаково распределенных случайных величин $X_1, X_2, ..., X_n$.

Для них среднее арифметическое равно

$$\overline{X} = (X_1 + X_2 + ... + X_n)/n .$$

Докажем три положения [4]:

1. Математическое ожидание среднего арифметического *n* взаимно независимых одинаково распределенных случайных величин равно математическому ожиданию, *a* каждой из величин

$$M(\overline{X}) = a .$$

Доказательство:

$$M(\overline{X}) = M\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{M(X_1) + M(X_2) + ... + M(X_n)}{n} = \frac{na}{n} = a .$$

2. Дисперсия среднего арифметического *n* взаимно независимых одинаково распределенных случайных величин в *n* раз меньше дисперсии *D* каждой из величин:

$$D(\overline{X}) = \frac{D(X_i)}{n} .$$

Доказательство:

Так как постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат, то

$$D(\overline{X}) = D\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{D(X_1) + D(X_2) + ... + D(X_n)}{n^2} = \frac{nD(X_i)}{n^2} = \frac{D(X_i)}{n} .$$

3. $\sigma(\overline{X}) = \frac{\sigma(X_i)}{\sqrt{n}}$ (следует из п.2), т. е. среднее арифметическое *n* взаимно независимых одинаково распределенных случайных величин имеет значительно меньшее рассеяние (в $\sqrt{n}$ раз), чем каждая отдельная величина.

1.6.4. Моменты (начальные, центральные) дискретной случайной величины

Начальный момент порядка r – это математическое ожидание случайной величины X^r

$$\nu_r = M(X^r) = \sum_{i=1}^{n} x_i^r p_i .$$

Например, начальные моменты первого и второго порядков равны

$$v_1 = M(X);\ v_2 = M(X^2).$$

Центральный момент порядка r задается формулой

$$\mu_r = M[X - M(X)]^r ,$$

при этом $\mu_1 = 0; \mu_2 = D(X)$.

Центральный момент второго порядка представляет собой дисперсию.

Между начальными и центральными моментами существуют соотношения

$$\mu_2 = \nu_2 - \nu_1^2 ,$$
$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 ,$$
$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 .$$

Следовательно, формула для вычисления дисперсии может быть записана в виде

$$D(X) = \mu_2 = \nu_2 - \nu_1^2 .$$

1.6.5. Примеры законов распределения дискретных случайных величин

Гипергеометрическое распределение

$$P_{N,M}(n,m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} .$$

В партии из N изделий M стандартных ($M<N$). Из партии отбирают n изделий (без возврата).

Случайная величина m – число стандартных изделий среди n отобранных имеет гипергеометрическое распределение. Оно широко используется в статистических методах контроля качества продукции.

Биномиальный закон

Если в гипергеометрическом распределении объем партии изделий увеличивать, то гипергеометрическое распределение будет приближаться к биномиальному закону ($M/N=p$)

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}.$$

Здесь выборка – с возвращением!

Закон Пуассона

Следует из биномиального при $n\to\infty$ и малой вероятности p (величина np – постоянная)

$$P_n(m) = \frac{(np)^m}{m! e^{np}},$$

где np – среднее. При $n \le 0{,}1N$ и q (или p) $\le 0{,}1$ закон Пуассона можно использовать вместо гипергеометрического.

1.7. Непрерывные случайные величины

1.7.1. Функция распределения

Функцией распределения (интегральной функцией распределения) случайной величины X называется функция $F(x)$, значения которой равны вероятностям $P(X<x)$

$$F(x)=P(X<x)=P(-\infty<X<x).$$

Из этого определения вытекают следующие свойства функции распределения:

1. $0 \le F(x) \le 1$.

2. $F(b) \geq F(a)$ при $b>a$, т. е. функция распределения – неубывающая.

3. $P(a \leq X < b) = F(b) - F(a)$.

4. $\lim\limits_{x \to -\infty} F(x) = 0, \quad \lim\limits_{x \to \infty} F(x) = 1$, если распределение задано на всей числовой оси.

5. Если между двумя случайными величинами X и Y существует функциональная зависимость $Y=\varphi(X)$, причем, с ростом X монотонно растет и Y, то их функции распределения равны

$$F(x) = F(y) = F(\varphi(x)).$$

6. Если с ростом X величина Y монотонно убывает, то

$$F(x) = 1 - F(y) = 1 - F(\varphi(x)).$$

1.7.2. Плотность распределения

Плотностью распределения (дифференциальным законом распределения) непрерывной случайной величины X называется первая производная от функции распределения

$$p(x) = F'(x) = \frac{dF(x)}{dx}.$$

График плотности распределения называется кривой распределения.

Свойства плотности распределения.

1. $p(x) \geq 0$.

2. $\int\limits_{-\infty}^{\infty} p(x)dx = 1$.

3. $P(a \leq X < b) = \int\limits_{a}^{b} p(x)dx.$

4. $F(x) = \int\limits_{-\infty}^{x} p(x)dx.$

1. Если между двумя случайными величинами X и Y существует функциональная зависимость $Y=\varphi(X)$, то

взаимосвязь между плотностями распределения p*(x)* и *p(y)* задается формулой

$$p(x) = p(y)\left|\frac{dy}{dx}\right| = p(\varphi(x))\left|\frac{d\varphi(x))}{dx}\right| \quad .$$

Действительно, пусть с ростом *X* растет *Y*. Тогда *F(x)=F(y)=F(φ(x))* (см. свойства функции распределения). По правилу дифференцирования сложной функции находим

$$p(x) = \frac{dF(x)}{dx} = \frac{dF(y)}{dx} = \frac{dF(y)}{dy}\frac{dy}{dx} = p(y)\frac{dy}{dx}$$

или, поскольку $y = \varphi(x)$,

$$p(x) = \frac{dF(\varphi(x))}{d\varphi(x)}\frac{d\varphi(x)}{dx} = p(\varphi(x))\frac{d\varphi(x)}{dx}.$$

В случае, если с ростом X величина *Y* убывает, первая производная $dy/dx < 0$, но плотность $p(x) > 0$. Поэтому в общем случае первая производная берется по абсолютной величине.

Рассмотрим примеры.

Пример 1. Из равенства функций распределения *F(x) = F(lnx)* требуется найти соотношение между плотностями *p(x)* и *p(lnx)*.

Дифференцируя последнее равенство по *x*, имеем:

$$p(x) = \frac{dF(x)}{dx} = \frac{dF(\ln x)}{d\ln x}\frac{d\ln x}{dx} = \frac{p(\ln x)}{x},$$

откуда $xp(x) = p(lnx)$.

Пример 2. Задана плотность распределения (показательный закон)

$$p(x) = Ne^{-\alpha x} \quad (0 < x < \infty).$$

Найти: *N* – нормирующий множитель; *F*(*x*) – функцию распределения; вероятность попадания случайной величины *X* на интервал $3<x<5$.

Используем свойства плотности:

$$\int_0^\infty p(x)dx = \int_0^\infty Ne^{-\alpha x}dx = \frac{N}{-\alpha}e^{-\alpha x}\Big|_0^\infty = -\frac{N}{\alpha e^{\alpha x}}\Big|_0^\infty = \frac{N}{\alpha} = 1.$$

Отсюда $N = \alpha$, т. е.

$$p(x) = \frac{\alpha}{e^{\alpha x}}.$$

Далее функция распределения равна

$$F(x) = \int_0^x p(x)dx = \int_0^x \frac{\alpha}{e^{\alpha x}} = 1 - \frac{1}{e^{\alpha x}}.$$

Вероятность попадания случайной величины X в заданный интервал равна

$$P(3 < x < 5) = F(5) - F(3) = \frac{1}{e^{3\alpha}} - \frac{1}{e^{5\alpha}}.$$

Пусть далее некоторая случайная величина Y связана со случайной величиной X зависимостью $Y=1/X$.

Найдем функцию распределения $F(y)$ и плотность $p(y)$.

Так как здесь с ростом X величина Y убывает, то

$$F(y) = 1 - F(x) = 1/e^{\alpha x}.$$

Но $X=1/Y$, поэтому

$$F(y) = \frac{1}{e^{\alpha / y}}.$$

Отсюда плотность $p(y)$ равна

$$p(y) = \frac{dF(y)}{dy} = \frac{\alpha}{y^2 e^{\alpha / y}}.$$

Плотность $p(y)$ можно найти непосредственно по плотности $p(x)$.

Поскольку $X=1/Y$, $dx/dy=-1/y^2$, то

$$p(y) = p(x)\left|\frac{dx}{dy}\right| = \frac{\alpha}{e^{\alpha x}}\frac{1}{y^2} = \frac{\alpha}{y^2 e^{\alpha / y}}.$$

1.7.3. Числовые характеристики непрерывных случайных величин

Для возможно более полного и всестороннего описания случайных величин используют различные показатели. К ним относятся:

характеристики положения – математическое ожидание, мода, медиана;

характеристики вариации – дисперсия, среднее квадратическое отклонение, коэффициент вариации;

характеристики формы распределения – коэффициенты асимметрии и островершинности, которые выражаются через моменты.

Математическое ожидание

случайной величины X задается интегралом

$$M(X) = \int_{-\infty}^{\infty} xp(x)dx .$$

Свойства математического ожидания непрерывной случайной величины те же, что и дискретной случайной величины.

Мода – такое значение случайной величины, при котором плотность максимальна.

Медиана *(Me)* случайной величины X определяется соотношением

$$P(X<Me)=P(X>Me).$$

Она делит площадь под кривой распределения пополам.

Дисперсия

непрерывной случайной величины X задается формулой

$$D(X) = M[X - M(X)]^2 = \int_{-\infty}^{\infty} (x - m_x)^2 p(x)dx ,$$

где $m_x = M(X).$

Коэффициент вариации

$$V = \frac{\sigma(X)}{m_x} \cdot 100\%$$

– выраженное в процентах отношение среднего квадратического отклонения случайной величины X к ее математическому ожиданию.

Центральные моменты

r-го порядка ($r = 2,3,4$) задаются формулой

$$\mu_r = M[X - M(X)]^r = \int_{-\infty}^{\infty} (x - m_x)^r f(x)dx.$$

Заметим, что $\mu_0 = 1; \quad \mu_1 = 0$.

Коэффициент асимметрии (скошенность)

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ или } \beta_1 = \frac{\mu_3}{\mu_2^{3/2}}.$$

Коэффициент островершинности (эксцесс)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ или } \beta_2 = \frac{\mu_4}{\mu_2^2} - 3.$$

1.7.4. Примеры непрерывных распределений

Нормальный закон

Нормальный закон распределения задается плотностью

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad a = M(X), \quad \sigma^2 = D(X).$$

Кривая распределения имеет симметричную колоколообразную форму и характеризуется показателями: $\beta_1=0$; $\beta_2=3$.

Вероятность попадания случайной величины X, распределенной по нормальному закону, на интервал $\alpha < x < \beta$ определяется по формуле

$$P(\alpha < x < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right),$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{t^2/2} dt$ – функция Лапласа. Здесь величина

$$t = \frac{x - a}{\sigma}$$

представляет собой выраженное в долях «сигма» отклонение случайной величины *X* от центра распределения *a*.

В зависимости от значения t вероятность попадания случайной величины X на заданный интервал $m_x \pm t\sigma$ равна:

При t = 1 P = 0,6827.

При t = 2 P = 0,9545.

При t = 3 P = 0,9973.

Таким образом, вероятность выхода значений случайной величины X за пределы 3σ очень мала и равна 1–0,9973=0,0027. Это значит, что из 1000 значений случайной величины X, распределенной по нормальному закону, в среднем только три могут выйти за границы трех стандартных отклонений (правило «трех сигма»). Это «правило» используется во многих практических расчетах, например, при статистическом анализе точности технологических процессов.

1.7.4.1. Показательный закон

Плотность вероятности и функция распределения задаются формулами

$$p(x) = \alpha e^{-\alpha x};\ F(x) = 1 - e^{-\alpha x}.$$

Это – один из простейших однопараметрических законов распределения.

1.7.4.2. Закон Вейбулла

Плотность вероятности и функция распределения задаются формулами

$p(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}$; $F(x) = 1 - e^{-\alpha x^{\beta}}$. Из закона Вейбулла при $\beta = 1$ следует показательный закон, а при $\beta = 2$ – распределение Релея.

2. МЕТОДЫ ПОСТРОЕНИЯ ОБОБЩЕННЫХ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ

2.1. Построение системы непрерывных распределений методом обобщения

Рассмотрим три простейших распределения: равномерное, треугольное убывающее и треугольное возрастающее [19].

В случае равномерной плотности функция распределения задается формулой

$$F(t)= \alpha t=1-(1-\alpha t). \quad (2.1)$$

В случае треугольного убывающего распределения получим

$$F(t)=1-\left(1-\frac{\alpha}{2}t\right)^2 \quad (2.2)$$

Для треугольного возрастающего распределения имеем

$$F(t)=\alpha t^2=1-(1-\alpha t^2). \quad (2.3)$$

Обобщим попарно функции распределения (2.1), (2.2) и (2.1), (2.3) путем введения новых параметров.

В первом случае получим

$$F(t)=1-(1-\alpha u t)^{\frac{1}{u}} \quad (2.4)$$

Во втором случае

$$F(t)=1-(1-\alpha t^{\beta}). \quad (2.5)$$

Теперь замечаем, что в формуле (2.5) имеется параметр β, но его нет в формуле (2.4). Введем его в последнюю формулу. В результате получим

$$F(t) = 1 - (1 - \alpha u t^{\beta})^{\frac{1}{u}} \qquad (2.6)$$

откуда дифференцированием по t найдем плотность распределения

$$p(t) = \alpha \beta t^{\beta-1} (1 - \alpha u t^{\beta})^{\frac{1}{u}-1} \qquad (2.7)$$

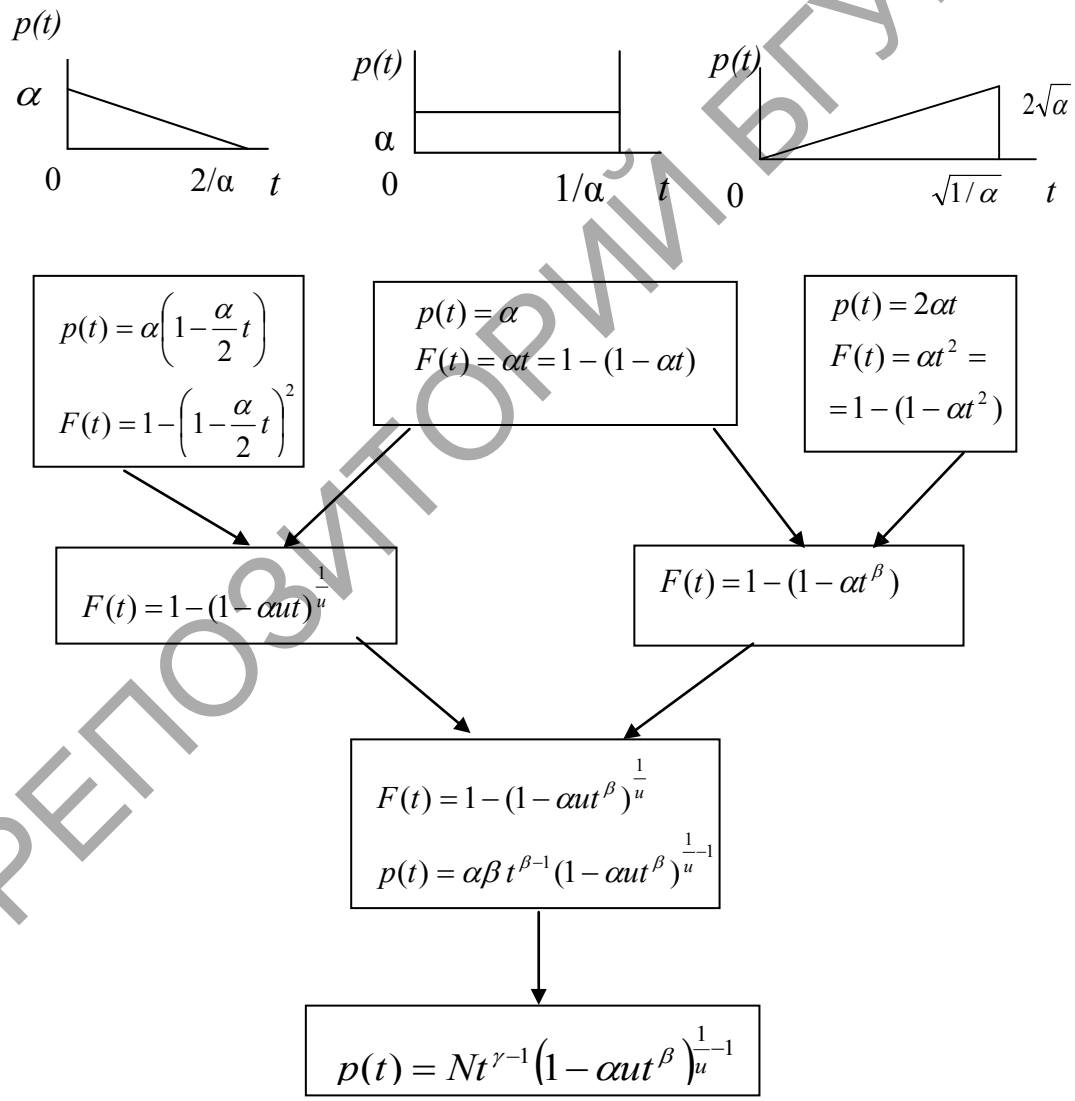


Рис. 2.1. Последовательность обобщения простейших непрерывных распределений

Последняя плотность может быть еще более расширена за счет введения нового параметра формы. Параметр β в формуле (2.7) используется дважды в качестве показателя степени. Пусть это будут два разных параметра. Тогда вместо (2.7) можем записать [23]

$$p(t) = Nt^{\gamma-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1} \qquad (2.8)$$

В итоге получена обобщенная плотность распределения с четырьмя параметрами α, β, γ, u. Нормирующий множитель N выражается через эти параметры из условия нормировки

$$\int_0^{\infty} p(t)dt = 1$$

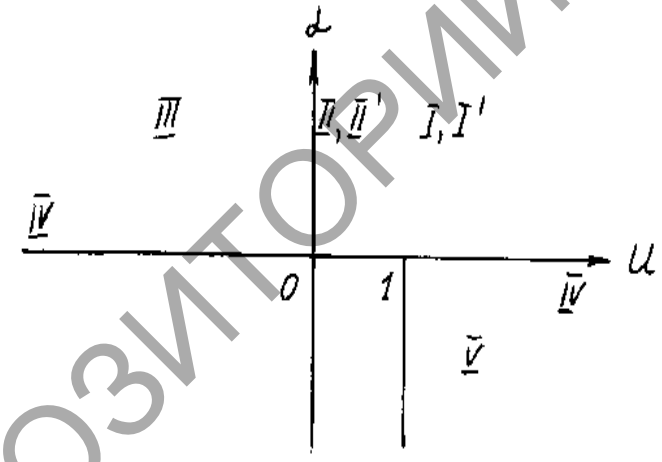


Рис. 2.2. Классификация распределений
(типы со штрихом – при β, γ < 0).

2.2. Классификация обобщенных распределений

В зависимости от значений параметров α, u, а также от знака параметров β, γ распределения, заданные обобщенной плотностью (2.8), можно разделить на типы (см. рис. 2.2.).

В таблице 2.1. приведены значения параметров распределений разных типов.

Таблица 2.1

Классификация распределений

Тип кривой	Параметры кривой		
	u	α	$k=\gamma/\beta$
I, I′	$0<u<\infty$	$\alpha > 0$	$0<k<\infty$
II, II′	$u\to 0$		
III	$-\infty<u<\infty$		$0 < k < 1-\frac{1}{u}$
IV	$u\to\pm\infty$	$\alpha u < 0$	
V	$1<u<\infty$	$\alpha < 0$	

Все распределения можно разбить на две большие группы: А и Б.

В группу А входят распределения с параметрами $\beta=\gamma$, или $\gamma/\beta=k=1$. Они задаются формулами (2.6) и (2.7).

В группу Б входят распределения, заданные обобщенной плотностью (2.8). В этом случае функция распределения, т. е. интеграл

$$F(t)=\int_0^t p(t)dt$$

как правило, не выражается конечным числом элементарных функций.

Отметим, что из плотности (2.8) при $\beta = 2$, $\gamma = 1$ следует группа симметричных распределений.

Симметричны также распределения I типа с параметрами $\beta = 1$, $\gamma=1/u$.

Приведем все существующие типы распределений обеих групп (см. табл. 2.2–2.4).

Таблица 2.2

Распределения группы А

Тип кривой	Функция распределения	Плотность распределения	Границы кривой
I	$F(t)=1-\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}}$	$p(t)=\alpha\beta t^{\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}$	$0<t<(\frac{1}{\alpha u})^{\frac{1}{\beta}}$ $(u>0)$
II	$F(t)=1-e^{-\alpha t^{\beta}}$	$p(t)=\alpha\beta t^{\beta-1}e^{-\alpha t^{\beta}}$	$0<t<\infty$ $(u\to 0)$
III	$F(t)=1-\frac{1}{(1-\alpha u t^{\beta})^{-1/u}}$	$p(t)=\frac{\alpha\beta t^{\beta-1}}{\left(1-\alpha u t^{\beta}\right)^{1-\frac{1}{u}}}$	$0<t<\infty$ $(u<0)$
I′	$F(t)=\left(1-\frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}$	$p(t)=\frac{\alpha\beta}{t^{\beta+1}}\left(1-\frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}-1}$	$(\alpha u)^{1/\beta}<t<\infty$ $(u>0)$
II′	$F(t)=e^{-\alpha/t^{\beta}}$	$p(t)=\frac{\alpha\beta}{t^{\beta+1}e^{\alpha/t^{\beta}}}$	$0<t<\infty$ $(u\to 0)$
III′	$F(t)=\frac{1}{\left(1-\frac{\alpha u}{t^{\beta}}\right)^{-1/u}}$	$p(t)=\frac{\alpha\beta}{t^{\beta+1}}\frac{1}{\left(1-\frac{\alpha u}{t^{\beta}}\right)^{1-\frac{1}{u}}}$	$0<t<\infty$ $(u<0)$

Таблица 2.3

Распределения группы Б

Тип кривой	Плотность распределения	Границы кривой
I	$p(t)=\frac{\beta(\alpha u)^{k}\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(\frac{1}{u}\right)}t^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}$	$0<t<(\frac{1}{\alpha u})^{\frac{1}{\beta}}$ $(u>0)$
I′	$p(t)=\frac{\beta(\alpha u)^{k}\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(\frac{1}{u}\right)}\frac{1}{t^{k\beta+1}}\left(1-\frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}-1}$	$(\alpha u)^{1/\beta}<t<\infty$ $(u>0)$

Тип кривой	Плотность распределения	Границы кривой
II	$p(t)=\frac{\beta\alpha^{k}}{\Gamma(k)}t^{k\beta-1}e^{-\alpha t^{\beta}}$	$0<t<\infty$ $(u\to 0)$
II′	$p(t)=\frac{\beta\alpha^{k}}{\Gamma(k)}\frac{1}{t^{k\beta+1}e^{\alpha/t^{\beta}}}$	$0<t<\infty$ $(u\to 0)$
III-V	$p(t)=\frac{\beta(-\alpha u)^{k}\Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1-\frac{1}{u}-k\right)}\frac{t^{k\beta-1}}{\left(1-\alpha u t^{\beta}\right)^{1-\frac{1}{u}}}$	$0<t<\infty$ $(\alpha u<0)$

Таблица 2.4

Группа симметричных распределений

Тип кривой	Плотность симметричного распределения	Границы кривой
Ic	$p(t)=\frac{\sqrt{\alpha u}\,\Gamma\left(\frac{1}{2}+\frac{1}{u}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{1}{u}\right)}\left(1-\alpha u t^{2}\right)^{\frac{1}{u}-1}$	$-\sqrt{\frac{1}{\alpha u}}<t<\sqrt{\frac{1}{\alpha u}}$
IIc	$p(t)=\sqrt{\frac{\alpha}{\pi}}\,e^{-\alpha t^{2}}$	$-\infty<t<\infty$
IIIc-Vc	$p(t)=\frac{\sqrt{-\alpha u}\,\Gamma\left(1-\frac{1}{u}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}-\frac{1}{u}\right)}\frac{1}{\left(1-\alpha u t^{2}\right)^{1-\frac{1}{u}}}$	$-\infty<t<\infty$

2.3. Распределения функций случайного аргумента

Из обобщенной плотности (2.8) можно получить другие распределения как функции случайного аргумента.

Если две случайные величины X, T связаны между собой функциональной зависимостью X=f(T), причем с ростом X растет T, то вероятность P(X < x) = F(x) должна быть равна вероятности P(T < t) = F(t), т. е.

$$F(x) = F(t). \qquad (2.9)$$

Найдем зависимость между плотностями распределения p(x) и p(t).

По правилу дифференцирования сложной функции из (2.9) имеем

$$p(x)=\frac{dF(x)}{dx}=\frac{dF(t)}{dt}\frac{dt}{dx}=p(t)\frac{dt}{dx} \qquad (2.10)$$

Воспользуемся последней формулой для нахождения других обобщенных плотностей.

Пусть между двумя случайными величинами T, X существует взаимосвязь $T=e^{X}$. Тогда $dt/dx=e^{x}$ и, следовательно,

$$p(x)=p(t)\frac{dt}{dx}=N\,e^{\gamma x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1} \qquad (2.11)$$

Характерной особенностью этой обобщенной плотности является то, что кривые III-V типов при $k=\frac{\gamma}{\beta}=\frac{1}{2}\left(1-\frac{1}{u}\right)$ являются симметричными. Если $T=\ln Y$,

то таким же путем получим еще одну обобщенную плотность

$$p(y)=\frac{N}{y}(\ln y)^{\gamma-1}\left[1-\alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (2.12)$$

Кривые распределения, заданные тремя обобщенными плотностями p(x), p(t), p(y), имеют разнообразную форму. Например, для кривой I типа, заданной плотностью p(t), существуют формы начала и конца кривой, которые представлены ниже.

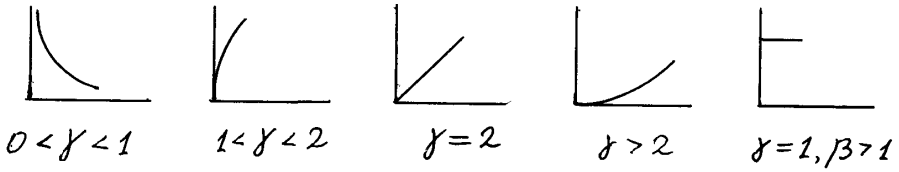


Рис. 2.3. Формы начала кривой в зависимости от значений параметра γ=kβ

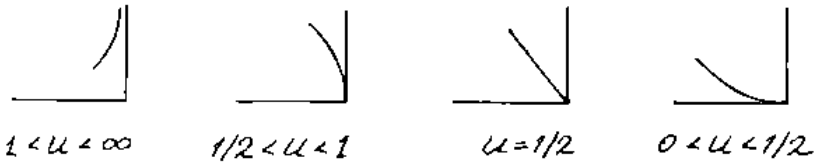


Рис. 2.4. Формы конца кривой в зависимости от значений параметра u

2.4. Три основные и три дополнительные системы непрерывных распределений В. В. Нешитого

Полученные выше обобщенные плотности распределения

$$\left.\begin{aligned} p(x) &= Ne^{\gamma x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1} \\ p(t) &= Nt^{\gamma-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N(\ln y)^{\gamma-1}}{y}\left[1-\alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} \end{aligned}\right\} \qquad (2.13)$$

образуют три основные системы непрерывных распределений В. Нешитого.

Вторая основная система непрерывных распределений, заданная плотностью p(t), представлена в таблицах 2.2 и 2.3.

Введем в плотность p(t) дополнительный параметр сдвига l и перепишем ее в виде

$$p(t) = N(t-l)^{\gamma-1}\left[1-\alpha u(t-l)^{\beta}\right]^{\frac{1}{u}-1} \qquad (2.14)$$

На основании плотности (2.14) можно получить три дополнительные системы непрерывных распределений.

Первая дополнительная система непрерывных распределений в общем случае задается формулой (2.14) при $|\beta| = 1$, а в случае симметричных распределений – при $\beta = 2$, $\gamma = 1$.

Ее легко получить из второй основной системы непрерывных распределений. Для этого достаточно в табл. 2.3 принять |β| = 1, t^{β} заменить на $t-l$, а в табл. 2.4 заменить величину t^2 на $(t-l)^2$.

Для обозначения типов кривых дополнительной системы непрерывных распределений будем использовать двузначный код, записанный арабскими цифрами через точку: 1.1, 1.1′, 2.1 и т.д., где первая цифра обозначает тип кривой, а вторая (единица) указывает на то, что параметр β=1; единица со штрихом соответствует параметру $\beta=-1$.В большинстве случаев в тексте используется единое обозначение типов, но при необходимости указывается, что |β| = 1. В таблице 2.5 приведены существующие типы первой дополнительной системы непрерывных распределений.

Из симметричных распределений приведен один нормальный закон. Распределения типа 1.1 при k = 1/u также являются симметричными. Первая дополнительная система непрерывных распределений представляет собой основную часть семейства кривых К. Пирсона.

Таблица 2.5

Первая дополнительная система непрерывных распределений

Тип кривой	Плотность распределения (k=γ/β=γ)	Границы кривой
1.1	$p(t)=\frac{(\alpha u)^k\,\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(\frac{1}{u}\right)}(t-l)^{k-1}\left[1-\alpha u(t-l)\right]^{\frac{1}{u}-1}$	$l<t<\frac{1}{\alpha u}+l$
1.1′	$p(t)=\frac{(\alpha u)^k\,\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(\frac{1}{u}\right)}\frac{1}{(t-l)^{k+1}}\left(1-\frac{\alpha u}{t-l}\right)^{\frac{1}{u}-1}$	$t>\alpha u+l$

Тип кривой	Плотность распределения (k=γ/β=γ)	Границы кривой
2.1	$p(t)=\frac{\alpha^k}{\Gamma(k)}\frac{(t-l)^{k-1}}{e^{\alpha(t-l)}}$	$t>l$
2.1′	$p(t)=\frac{\alpha^k}{\Gamma(k)}\frac{1}{(t-l)^{k+1}e^{\alpha/(t-l)}}$	$t>l$
3.1	$p(t)=\frac{(-\alpha u)^k\,\Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1-\frac{1}{u}-k\right)}\frac{(t-l)^{k-1}}{[1-\alpha u(t-l)]^{1-\frac{1}{u}}}$	$t>l$
IIc	$p(t)=\sqrt{\frac{\alpha}{\pi}}e^{-\alpha(t-l)^2}$	$-\infty<t<\infty$

Вторая дополнительная система непрерывных распределений получается из первой при t = ln y и прежних значениях параметров β, γ. При этом обобщенная плотность имеет вид

$$p(y)=\frac{N(\ln y-l)^{\gamma-1}}{y}\left[1-\alpha u(\ln y-l)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (2.15)$$

Третья дополнительная система непрерывных распределений получается из второй при $y=\ln w$

$$p(w)=\frac{N(\ln\ln w-l)^{\gamma-1}}{w\ln w}\left[1-\alpha u(\ln\ln w-l)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (2.16)$$

3. КЛАССИЧЕСКИЕ МЕТОДЫ ОЦЕНИВАНИЯ ПАРАМЕТРОВ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ

Методы оценивания параметров обобщенных непрерывных распределений

При исследовании случайных величин в математической статистике используется выборочный метод. Он заключается в том, что из генеральной совокупности отбирается выборка объемом, как правило, не менее 100 единиц. При этом она должна правильно отражать пропорции генеральной совокупности, т. е. быть представительной (репрезентативной). Только в этом случае результаты исследования выборки могут быть распространены на всю генеральную совокупность.

Чтобы извлечь информацию из выборки, которая представляет собой простой статистический ряд, необходимо упорядочить все значения исследуемой случайной величины либо по возрастанию, либо по убыванию и построить интервальный ряд распределения или ранжированный ряд. Далее вычисляются числовые характеристики случайной величины и по ним – аппроксимирующий закон распределения и оценки его параметров. Закон распределения является наиболее полной характеристикой случайной величины.

3.1. Метод наименьших квадратов

Этим методом могут быть найдены оценки параметров распределений группы А.

Рассмотрим распределения I – III типов группы *A*. Преобразуем функцию распределения

$$F(t) = 1 - \left(1 - \alpha\, ut^{\beta}\right)^{\frac{1}{u}}$$

к уравнению прямой

$$\ln \frac{1 - [1 - F(t)]^{u}}{u} = \ln \alpha + \beta \ln t \qquad (3.1.1)$$

Построив по эмпирической функции распределения график зависимости (3.1.1) (при известной оценке параметра *u*) и убедившись, что опытные точки рассеиваются вдоль прямой, по методу наименьших квадратов найдем оценки величин $\ln \alpha, \beta$. Введем обозначения:

$$Y = \ln \frac{1 - [1 - F(t)]^{u}}{u}, \qquad X = \ln t.$$

Тогда вместо формулы (3.1) запишем

$$Y = \ln \alpha + \beta X. \qquad (3.1.2)$$

Оценки параметров $\ln \alpha, \beta$ (при заданном значении параметра *u*) по методу наименьших квадратов будут равны

$$\beta = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}, \qquad (3.1.3)$$

$$\ln \alpha = \frac{1}{n}(\Sigma Y - \beta \Sigma X). \qquad (3.1.4)$$

Для оценки тесноты связи между переменными *Y, X* при различных значениях параметра *u* вычисляется выборочный коэффициент корреляции

$$r_{y/x} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}. \qquad (3.1.5)$$

В качестве оценки параметра *u* следует принять то его значение, при котором коэффициент корреляции по модулю ближе к единице.

Аналогично приводятся к уравнению прямой функции распределения остальных типов.

Тип II: $F(t)=1-e^{-\alpha t^{\beta}}$, $\ln\ln\dfrac{1}{1-F(t)}=\ln\alpha+\beta\ln t$.

Вводя обозначения $Y=\ln\ln\dfrac{1}{1-F(t)}$, $X=\ln t$, получим уравнение прямой (3.1.2).

Тип II′: $F(t)=e^{-\alpha/t^{\beta}}$, $\ln\ln\dfrac{1}{F(t)}=\ln\alpha-\beta\ln t$.

Типы I′, III′: $F(t)=\left(1-\dfrac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}$, $\ln\dfrac{1-[F(t)]^{u}}{u}=\ln\alpha-\beta\ln t$.

Из рассмотренных примеров видно, что главная трудность здесь заключается в выборе подходящего значения параметра *u*. Его можно найти путем подбора и вычисления при каждом значении *u* коэффициента корреляции. Однако имеется возможность оценить его более простым и быстрым методом.

Если построить кривую распределения в форме $tp(t)=f(\ln t)$ и график функции распределения $F(t)=\varphi(\ln t)$, то мода $\ln t_c$, т. е. точка, в которой произведение *tp(t)* максимально, равна

$$\ln t_c=\frac{1}{\beta}\ln\frac{1}{\alpha},$$

откуда $t_c=(1/\alpha)^{1/\beta}$. Подставив значение t_c в функцию распределения, получим [22]

$$F(t_c)=1-(1-\alpha u t^{\beta})^{\frac{1}{u}}=1-(1-u)^{\frac{1}{u}}. \qquad (3.1.6)$$

Последняя формула справедлива для распределений I-III типов группы *A*. Для распределений I′-III′ типов справедливо равенство

$$F(t_c) = (1-u)^{\frac{1}{u}}. \qquad (3.1.7)$$

В таблице 3.1 приведены значения $F(t_c)$, рассчитанные по формулам (3.1.6), (3.1.7).

Таблица 3.1

Значение функции распределения $F(t_c)$

Параметр *u*	$F(t_c)^{I-III}$	$F(t_c)^{I'-III'}$	Тип кривой
1	1	0	I, I′
0,9	0,9226	0,0774	
0,8	0,8663	0,1337	
0,7	0,8209	0,1791	
0,6	0,7828	0,2172	
0,5	0,7500	0,2500	
0,4	0,7211	0,2789	
0,3	0,6954	0,3046	
0,2	0,6723	0,3277	
0,1	0,6513	0,3487	
0	0,6321	0,3679	II, II′
-0,2	0,5981	0,4019	III-III′
-0,4	0,5688	0,4312	
-0,6	0,5431	0,4569	
-0,8	0,5204	0,4796	
-1,0	0,5000	0,5000	
-1,5	0,4571	0,5429	
-2	0,4226	0,5774	
-2,5	0,3941	0,6059	
-3	0,3700	0,6300	
-4	0,3313	0,6687	
-5	0,3012	0,6988	
-10	0,2132	0,7868	
-20	0,1412	0,8588	
-30	0,1082	0,8918	
-∞	0	1	

На основании полученных результатов можно рекомендовать следующий порядок установления типа выравнивающего распределения группы *А* и нахождения оценок параметров на примере плотности *p(t)*.

1. Выбрать за начало отсчета значений случайной величины *T* начало кривой распределения.

2. Найти эмпирическую моду $\ln t_c^*$ кривой распределения $tp(t) = p(\ln t)$.

3. Найти эмпирическое значение функции распределения в точке *С* и приравнять теоретическому.

4. С помощью таблицы 3.1 определить два значения параметра *и* (в предположении, что выравнивающее распределение относится либо к I-III, либо к I′-III′ типам).

5. По двум значениям параметра *и* определить два типа возможных выравнивающих распределений.

6. Для обоих типов распределений путем построения графиков проверить, ложатся ли опытные точки на прямые.

7. В качестве выравнивающего принять наиболее подходящее распределение.

Таким же способом могут быть найдены оценки параметров распределений группы *А*, заданных плотностями р*(x), p(y)*. При этом плотность *p(y)* должна быть приведена к форме $yp(y)\ln y = p(\ln\ln y)$.

3.2. Метод наибольшего правдоподобия

Покажем применение этого метода на примере распределений I типа группы *Б*

$$p(t) = \frac{\beta(\alpha u)^k \Gamma\left(k + \frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(\frac{1}{u}\right)} t^{k\beta-1}\left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}-1}, \quad k = \gamma/\beta \,. (3.2.1)$$

Примем в качестве логарифмической функции правдоподобия величину $\ln L = M\left[\ln tp(t)\right]$ [31].

Вначале логарифмируем плотность *p(t)* (лучше – произведение *tp(t)):*

$$\ln tp(t) = \ln\beta + k\ln\alpha u + \ln\Gamma\left(k+\frac{1}{u}\right) - \ln\Gamma(k) - \ln\Gamma\left(\frac{1}{u}\right) + k\beta\ln t + \left(\frac{1}{u}-1\right)\ln\left(1-\alpha u t^{\beta}\right). \quad (3.2.2)$$

Далее находим математическое ожидание величины $\ln tp(t)$

$$\ln L = M\left[\ln tp(t)\right] = \ln\beta + k\ln\alpha u + \ln\Gamma\left(k+\frac{1}{u}\right) - \ln\Gamma(k) - \ln\Gamma\left(\frac{1}{u}\right) + k\beta M(\ln t) + \left(\frac{1}{u}-1\right)M\left[\ln\left(1-\alpha u t^{\beta}\right)\right]. \quad (3.2.3)$$

Уравнения правдоподобия находятся из условий**:**

$$\frac{\partial\ln L}{\partial\alpha}=0;\ \ \frac{\partial\ln L}{\partial\beta}=0;\ \ \frac{\partial\ln L}{\partial k}=0;\ \ \frac{\partial\ln L}{\partial u}=0\ .$$

Приняв обозначение $\frac{d}{dk}\ln\Gamma(k)=\Psi(k)$ для логарифмической производной гамма-функции, или иначе пси-функции, из (3.2.3) найдем

$$\left.\begin{aligned} &\frac{k}{\alpha}-(1-u)M\left(\frac{t^{\beta}}{1-\alpha u t^{\beta}}\right)=0\\ &\frac{1}{\beta}+kM(\ln t)-\alpha(1-u)M\left(\frac{t^{\beta}\ln t}{1-\alpha u t^{\beta}}\right)=0\\ &\ln\alpha u+\Psi\left(k+\frac{1}{u}\right)-\Psi(k)+\beta M(\ln t)=0\\ &\Psi\left(\frac{1}{u}\right)-\Psi\left(k+\frac{1}{u}\right)-M\left[\ln\left(1-\alpha u t^{\beta}\right)\right]=0 \end{aligned}\right\} \quad (3.2.4)$$

Здесь последнее уравнение приведено к более простой форме с учетом первого уравнения. Оценки параметров могут быть найдены путем решения системы четырех уравнений правдоподобия – (3.2.4). При этом соответствующие математические ожидания заменяются

их оценками, которые вычисляются по статистическому распределению. Однако для нахождения оценок таких величин, как $M\left[t^{\beta}/(1-\alpha u t^{\beta})\right]$ и др. необходимо знать значения параметра β и произведения αu, оценки которых предстоит найти. Кроме того, предварительно необходимо знать тип выравнивающего распределения, а метод наибольшего правдоподобия не предлагает критериев для его установления.

Эти обстоятельства сильно ограничивают возможности использования метода наибольшего правдоподобия для нахождения оценок параметров обобщенных выравнивающих распределений.

3.3. Классический метод моментов

Метод пригоден для оценивания параметров обобщенных распределений с параметром $|\beta| = 1$**,** т. е. в случае трех дополнительных систем непрерывных распределений, заданных плотностями (2.14)–(2.16). Причем, эти плотности должны быть представлены в виде

$$yp(y) = p(\ln y),\ \text{w}\ln w\ p(w) = p(\ln\ln w).$$

Метод подробно рассмотрен в статьях учебных пособиях автора теории обобщенных распределений [15, 21, 22, 23]. Поскольку он может быть использован лишь в случаях распределений с параметром β=1, здесь его рассматривать не будем.

4. УНИВЕРСАЛЬНЫЙ МЕТОД МОМЕНТОВ ВЫЧИСЛЕНИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ И ОЦЕНОК ПАРАМЕТРОВ

4.1. Универсальный метод моментов

За пределами применимости классического метода моментов остается широкий класс распределений, для которых не существует моментов высоких порядков. Оценки параметров таких распределений могут быть найдены по универсальному методу моментов, который впервые был описан автором в 1983 г. и опубликован в депонированной рукописи «Построение и исследование системы дискретных распределений» деп. в БЕЛНИИНТИ 17.07.1985, № 931, 71 с.

Основное отличие этого метода от классического метода моментов заключается прежде всего в том, что он применяется к распределениям, заданным обобщенной плотностью *p(x)*. Другие плотности должны быть приведены к этой форме. Например, вместо плотности p(t), которую представим в виде (при $\gamma = k\beta$)

$$p(t) = Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \qquad (4.1)$$

используется плотность $tp(t) = p(\ln t)$, т. е.

$$tp(t) = p(\ln t) = Ne^{k\beta \ln t}\left(1-\alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (4.2)$$

Здесь последнее равенство получено из предыдущего путем умножения на *t* обеих его частей и использования записи $e^{\beta \ln t}$ вместо t^{β}, что одно и то же.

Введем далее обозначение *lnt* = *x*. Тогда последнее равенство примет вид

$$p(x) = Ne^{k\beta x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}, \qquad (4.3)$$

т. е. получили обобщенную плотность *p*(*x*).

Если плотность *p(t)* привести к форме $tp(t) = p(\ln t)$, то она будет обладать всеми свойствами плотности *p(x)*.

Плотность

$$p(y) = \frac{N(\ln y)^{k\beta-1}}{y}\left[1 - \alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} \qquad (4.4)$$

также приводится к форме плотности *p*(*x*).

Умножим обе части последней формулы на произведение $y \ln y$, а величину $(\ln y)^{\beta}$ запишем в виде $e^{\beta \ln\ln y}$. В результате получим

$$y \ln y\ p(y) = Ne^{k\beta \ln\ln y}\left(1 - \alpha u e^{\beta \ln\ln y}\right)^{\frac{1}{u}-1}. \qquad (4.5)$$

Приняв далее обозначения

$\ln\ln y = x$, $y \ln y\ p(y) = p(\ln\ln y) = p(x)$, получим плотность (4.3).

Далее так же, как и в классическом методе моментов, центральные моменты μ_2, μ_3, μ_4, а также показатели асимметрии $\beta_1 = \mu_3^2 / \mu_2^3$ и островершинности $\beta_2 = \mu_4 / \mu_2^2$ выражаются через параметры обобщенного распределения (4.3). При этом показатели β_1 и β_2 зависят лишь от двух параметров формы ($k = \gamma/\beta$, u) и в зависимости от их значений распределения разделяются на типы.

Приравнивая далее эмпирические значения показателей β_1^*, β_2^* теоретическим β_1 и β_2 устанавливаем тип выравнивающей кривой распределения и находим оценки двух параметров формы *k, u*. Оценки двух других параметров – α, β (или произведения αu) вычисляются по простым формулам при известных оценках параметров *k, u* .

Отметим, что статистические центральные моменты *r*-го порядка в зависимости от вида плотности выравнивающего распределения вычисляются по формулам:

– в случае обобщенной плотности $p(x)$

$$\mu_r^* = \frac{1}{M}\sum_{i=1}^{n}\left(x_i - \nu_1^*\right)^r m_i \text{, где } \nu_1^* = \overline{x} = \frac{1}{M}\sum_{i=1}^{n} x_i m_i \text{ ;}$$

– в случае обобщенной плотности *p(t),* которая приводится к форме $tp(t) = p(\ln t)$,

$$\mu_r^* = \frac{1}{M}\sum_{i=1}^{n}\left(\ln t_i - \nu_1^*\right)^r m_i \text{, где } \nu_1^* = \overline{\ln t} = \frac{1}{M}\sum_{i=1}^{n}\ln t_i m_i \text{ ;}$$

– в случае обобщенной плотности *p(y),* которая приводится к форме $y\ln yp(y) = p(\ln\ln y)$,

$$\mu_r^* = \frac{1}{M}\sum_{i=1}^{n}\left(\ln\ln y_i - \nu_1^*\right)^r m_i ,$$

где

$$\nu_1^* = \overline{\ln\ln y} = \frac{1}{M}\sum_{i=1}^{n}\left(\ln\ln y_i\right) m_i .$$

Здесь n – число интервалов группирования статистических данных; m_i – частота i-го интервала; $M = \sum m_i$ – объем выборки.

Это обеспечивает единый порядок установления типа выравнивающего распределения и нахождения оценок параметров для трех основных систем непрерывных распределений.

Эти же моменты используются для оценивания параметров трех дополнительных систем непрерывных распределений, т. е. в случае классического метода моментов.

Рассмотрим для примера распределения III–V типов, заданные плотностью

$$p(x) = \frac{\beta(-\alpha u)^k \Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1-\frac{1}{u}-k\right)} \frac{e^{k\beta x}}{\left(1-\alpha u e^{\beta x}\right)^{1-\frac{1}{u}}} \qquad (4.6)$$

Этими распределениями можно дополнить первую (дополнительную) систему распределений (см. табл. 2.5) при условии $B^2 - 4AC < 0$ [см., напр. 19 Классический метод моментов].

Выразим центральные моменты (2–4)-го порядков и начальный момент 1-го порядка (математическое ожидание) через параметры распределения (4.6).

Используя теорию производящих функций, для обобщенной плотности (4.6) получим:

$$\left.\begin{aligned} \nu_1 &= \frac{1}{\beta}\left[\Psi(k) - \Psi(k') - \ln(-\alpha u)\right] \\ \mu_2 &= \frac{1}{\beta^2}\left[\Psi'(k) + \Psi'(k')\right] \\ \mu_3 &= \frac{1}{\beta^3}\left[\Psi''(k) - \Psi''(k')\right] \\ \mu_4 &= 3\mu_2^2 + \frac{1}{\beta^4}\left[\Psi'''(k) + \Psi'''(k')\right] \end{aligned}\right\} \qquad (4.7)$$

где $k'=1-1/u-k$. В этом случае существуют все моменты для всех распределений, в том числе для распределения Коши, потому что вычисляются моменты не самой случайной величины, а их логарифмов. В этом состоит преимущество универсального метода моментов перед классическим методом моментов.

Показатели асимметрии β_1 и островершинности β_2 равны

$$\left.\begin{aligned} \beta_1 &= \frac{\left[\Psi''(k) - \Psi''(k')\right]^2}{\left[\Psi'(k) + \Psi'(k')\right]^3} \\ \beta_2 &= 3 + \frac{\Psi'''(k) + \Psi'''(k')}{\left[\Psi'(k) + \Psi'(k')\right]^2} \end{aligned}\right\} \qquad (4.8)$$

Заменяя показатели β_1, β_2 их оценками, из системы двух уравнений (4.8) можно найти оценки двух параметров *k, u*, предварительно установив по тем же показателям тип выравнивающей кривой.

Это – большое преимущество перед методом наибольшего правдоподобия, который требует решения четырех уравнений с четырьмя неизвестными, причем при условии, когда тип распределения заранее задан.

Для нахождения оценок параметров α и β (или произведения αu) введем случайную величину Z, которая связана со случайной величиной X зависимостью $Z = -\alpha u e^{\beta X}$ (см. формулу (4.6)) и рассмотрим ее логарифм

$$\boldsymbol{\ln Z = \beta X + \ln(-\alpha u)}.$$

Это уравнение является базой для построения универсального метода моментов.

Найдем математическое ожидание логарифма случайной величины Z

$$M(\ln Z) = \beta M(X) + \ln(-\alpha u).$$

Из последней формулы следует, что

$$M(\mathrm{X}) = \frac{1}{\beta}[M(\ln Z) - \ln(-\alpha u)],$$

$$M[X - M(X)]^r = \frac{1}{\beta^r} M[\ln Z - M(\ln Z)]^r$$

или $\mu_r = \mu_r^{(Z)} / \beta^r$.

С учетом полученных равенств первые две формулы из четырех формул (4.7) можно переписать в виде

$$\left.\begin{aligned} \nu_1 &= \frac{1}{\beta}\left[\nu_1^{(z)} - \ln(-\alpha u)\right] \\ \mu_2 &= \frac{\mu_2^{(z)}}{\beta^2} \end{aligned}\right\}, \qquad (4.7')$$

где $\nu_1^{(z)} = M(\ln Z) = \Psi(k) - \Psi(1 - 1/u - k)$ – математическое ожидание случайной величины $\ln Z$; $\mu_2^{(z)} = \Psi'(k) + \Psi'\left(1 - \frac{1}{u} - k\right)$ –

центральный момент второго порядка случайной величины $\ln Z$; $\Psi'(k)$ – первая производная пси-функции $\Psi(k)=\frac{d}{dk}\ln\Gamma(k)$.

На основании (4.7′) оценки параметра β и произведения αu равны

$$\beta=\sqrt{\frac{\mu_2^{(z)}}{\mu_2}}, \tag{4.9}$$

$$\alpha u=-e^{\nu_1^{(z)}-\beta\nu_1}, \tag{4.10}$$

где $\nu_1=M(X)$. При вычислении оценок β и αu центральный момент второго порядка случайной величины X следует заменить его оценкой μ_2^* (выборочной дисперсией), а $M(X)$ – выборочным средним $\nu_1^*=\bar{x}$.

Аналогично выводятся формулы для оценок параметров распределений других типов, заданных плотностью $p(x)$.

Так, в случае распределений II, II′ типов имеем

$$\beta=\sqrt{\frac{\mu_2^{(z)}}{\mu_2}};\qquad \alpha=e^{\pm\left(\nu_1^{(z)}-\beta\nu_1\right)}, \tag{4.11}$$

где $\nu_1^{(z)}=\pm\Psi(k);\ \mu_2^{(z)}=\Psi'(k)$.

Здесь знак "+" относится ко II типу, а "–" – ко II′ типу.

В случае распределений I, I′ типов

$$\beta=\sqrt{\frac{\mu_2^{(z)}}{\mu_2}};\qquad \alpha u=e^{\pm\left(\nu_1^{(z)}-\beta\nu_1\right)}, \tag{4.12}$$

где $\nu_1^{(z)}=\pm\left[\Psi(k)-\Psi\left(k+\frac{1}{u}\right)\right]$,

$\mu_2^{(z)}=\Psi'(k)-\Psi'\left(k+\frac{1}{u}\right)$.

При расчетах по универсальному методу моментов необходимо уметь вычислять с заданной точностью значения гамма-функции и ее логарифмических производных. Ниже приводятся приближенные формулы для их вычисления.

Логарифм гамма-функции находится по формуле

$$\left.\begin{aligned}\ln\Gamma(x) \approx \frac{\ln 2\pi}{2} - \sum_{S=x}^{x+n} \ln s + \left(x+n+\frac{1}{2}\right)\ln(x+n) - \\ -(x+n) + \frac{1}{12(x+n)} - \frac{1}{360(x+n)^3} + \frac{1}{1260(x+n)^5} + \ldots\end{aligned}\right\} \quad (4.13)$$

Логарифмические производные гамма-функции на основании (4.13) равны:

$$\left.\begin{aligned}\Psi(x) = \frac{d}{dx}\ln\Gamma(x) \approx -\sum_{s=x}^{x+n}\frac{1}{s} + \ln(x+n) + \frac{1}{2(x+n)} - \frac{1}{12(x+n)^2} + \\ + \frac{1}{120(x+n)^4} - \frac{1}{252(x+n)^6} + \ldots \\ \Psi'(x) = \sum_{s=x}^{x+n}\frac{1}{s^2} + \frac{1}{x+n} - \frac{1}{2(x+n)^2} + \frac{1}{6(x+n)^3} - \frac{1}{30(x+n)^5} + \ldots \\ \Psi''(x) = -2\sum_{s=x}^{x+n}\frac{1}{s^3} - \frac{1}{(x+n)^2} + \frac{1}{(x+n)^3} - \frac{1}{2(x+n)^4} + \frac{1}{6(x+n)^6} + \ldots \\ \Psi'''(x) = 6\sum_{s=x}^{x+n}\frac{1}{s^4} + \frac{2}{(x+n)^3} - \frac{3}{(x+n)^4} + \frac{2}{(x+n)^5} - \frac{1}{(x+n)^7} + \ldots\end{aligned}\right\} \quad (4.14)$$

Для облегчения различных расчетов в Приложении 1 дана таблица значений функций $\Gamma(x), \Psi(x), \Psi'(x)$ при $x = 0{,}1 \div 4$ с шагом 0,01.

Точность приведенных формул тем выше, чем больше сумма $x+n$. Для приближенных расчетов на калькуляторе можно принять $n=2$, а при более точных расчетах на ПЭВМ – $n=4$.

Рассчитаем для разных типов распределений значения показателей β_1, β_2 при различных значениях параметров *k, u.* Далее в системе координат (β_1, β_2) отметим области для распределений разных типов (см. рис. 4.1).

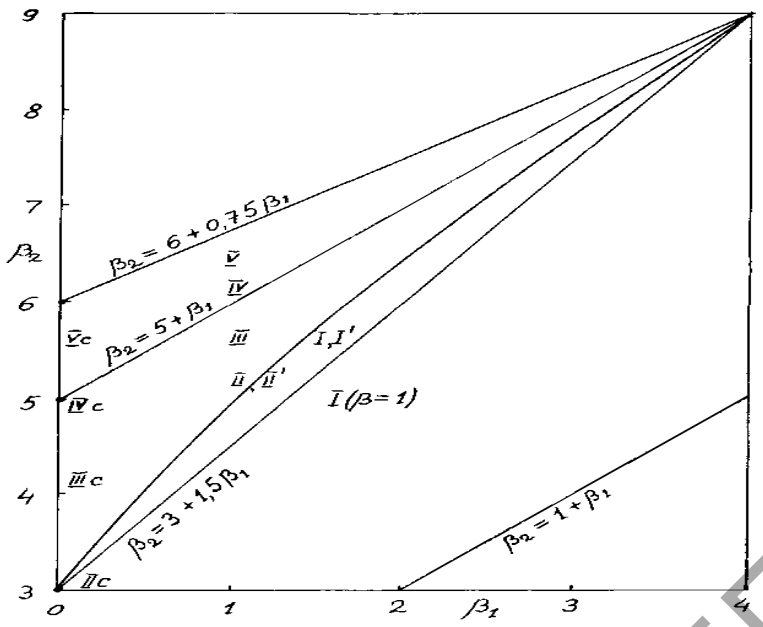


Рис. 4.1. Классификация распределений, заданных обобщенной плотностью $p(x) = Ne^{k\beta x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}$, по критериям β_1, β_2

Как видно из рисунка, распределения II, II′ типов представлены кривой, распределения IV типа – прямой $\beta_2 = 5 + \beta_1$. Распределения V типа лежат ниже прямой $\beta_2 = 6 + 0.75\beta_1$, а распределения I, I′ типов – выше прямой $\beta_2 = 3 + 1.5\beta_1$

Симметричные распределения III типа с параметрами формы $k = 0{,}5(1 - 1/u)$ представлены отрезком $3 < \beta_2 < 6$ оси ординат. С ростом параметра *k* распределения IIIс типа, а также II, II′ типов приближаются к нормальному закону, для которого $\beta_1 = 0, \quad \beta_2 = 3$.

В заключение отметим, что ниже прямой $\beta_2 = 3 + 1{,}5\beta_1$ находится область распределений I типа, заданных плотностью (3.10) с параметром β=1.

Для быстрого установления типа выравнивающей кривой и нахождения оценок параметров *k, u* по методу моментов автором построена номограмма (Приложение 2). Она строилась для распределений с левосторон-

ней асимметрией, у которых центральный момент 3-го порядка $\mu_3 < 0$.

Если статистическое распределение имеет правостороннюю асимметрию($\mu_3 > 0$), то в случае распределений III-V типов вначале с помощью номограммы находятся оценки параметров *u, k′,* затем вычисляется оценка параметра *k*:

$$k = 1 - \frac{1}{u} - k' . \qquad (4.15)$$

Аналогично для распределений I типа (β=1) при $\mu_3 > 0$ вначале по номограмме находятся оценки параметров *u′, k′*, затем вычисляются оценки параметров *u, k*:

$$u = \frac{1}{k'}; \;\; k = \frac{1}{u'} . \qquad (4.16)$$

Построенная номограмма состоит из двух частей. Верхняя часть (выше прямой $\beta_2 = 3 + 1.5\beta_1$) относится к распределениям с плотностью *p(x)* или к распределениям с плотностью *p(t), p(y),* которые приведены соответственно к форме $tp(t) = p(\ln t); y \ln y \; p(y) = p(\ln \ln y)$.

Нижняя часть номограммы относится к распределениям I типа с плотностью *p(t)* при β=1 (т.е. типа 1.1) и является продолжением верхней части. Прямой $\beta_2 = 3 + 1{,}5\beta_1$ при $\mu_3 > 0$ представлены распределения второго типа с параметром $\beta = 1$ (гамма-распределения).

Это дает возможность расширить основные системы непрерывных распределений за счет включения в них распределений типов 1.1 и 2.1, которые относятся к дополнительным системам непрерывных распределений (с параметром β=1).

Тогда первая (основная) система непрерывных распределений SRN1 в общем случае будет включать три обобщенные плотности

$$\left.\begin{aligned} p(x) &= Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1} \\ p(t) &= N(t-l)^{k-1}\left[1-\alpha u(t-l)\right]^{\frac{1}{u}-1} \\ p(t) &= N\left[1-\alpha u(t-\bar{t})^2\right]^{\frac{1}{u}-1} \end{aligned}\right\} \qquad (4.17)$$

Первая система непрерывных распределений включает две группы симметричных распределений: одна из них (типы IIIc-Vc) задана плотностью $p(x)$ при $k=0{,}5(1-1/u)$, другая (типы Ic-Vc) – плотностью $p(t)$ при $\beta=2,\ \gamma=1$. Кроме того, симметричные распределения Ic типа описываются также плотностью $p(t)$ с параметром сдвига l при $ku = 1$.

Первая система непрерывных распределений может быть также задана двумя плотностями (без последней) или даже одной плотностью $p(x)$.

Аналогично во вторую основную систему непрерывных распределений SRN2 войдут обобщенные плотности

$$\left.\begin{aligned} p(t) &= Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N(\ln y-l)^{k-1}}{y}\left[1-\alpha u(\ln y-l)\right]^{\frac{1}{u}-1} \\ p(y) &= \frac{N}{y}\left[1-\alpha u(\ln y-\overline{\ln y})^2\right]^{\frac{1}{u}-1} \end{aligned}\right\} \qquad (4.18)$$

которые получены из первой системы как распределения функций случайных аргументов: $X=lnT$ – для первой плотности; $T=lnY$ – для двух других плотностей.

Наконец, в третью основную систему непрерывных распределений SNR3 войдут обобщенные плотности

$$\left.\begin{aligned} p(y) &= \frac{N(\ln y)^{k\beta-1}}{y}\left[1-\alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} \\ p(w) &= \frac{N(\ln\ln w-l)^{k-1}}{w\ln w}\left[1-\alpha u(\ln\ln w-l)\right]^{\frac{1}{u}-1} \\ p(w) &= \frac{N}{w\ln w}\left[1-\alpha u(\ln\ln w-\overline{\ln\ln w})^2\right]^{\frac{1}{u}-1} \end{aligned}\right\} \qquad (4.19)$$

Вторая и третья основные системы непрерывных распределений также могут быть заданы либо двумя плотностями (без третьей), либо одной первой плотностью распределения.

Для нахождения оценок параметров трех основных систем непрерывных распределений по методу моментов автором созданы **программы** **SNR1MM**, **SNR2MM**, **SNR3MM**.

Номограмма, представленная в Приложении 2, остается справедливой для трех основных систем непрерывных распределений.

4.2. Законы распределения суммы независимых случайных величин

Системы непрерывных распределений, заданные обобщенными плотностями, а также методы оценивания параметров, доведенные до программной реализации, позволяют более просто решать различные задачи.

Пусть, например, требуется установить закон распределения суммы n независимых одинаково распределенных случайных величин $X=X_1+X_2+\ldots+X_n$. Среднее каждой случайной величины равно ν_1.

Распределение случайной величины X_i может быть задано как аналитически, так и таблично.

Поэтому для нахождения закона распределения суммы n независимых случайных величин, т. е. композиции n распределений, можно использовать общий метод.

Для этого достаточно вычислить моменты суммы n независимых случайных величин $\nu_{1(n)}$, $м_{2(n)}$, $м_{3(n)}$, $м_{4(n)}$, а также показатели $в_{1(n)}$ и $в_{2(n)}$ по известным моментам случайной величины X_i.

Далее по методу моментов (универсальному или классическому) с помощью программы устанавливается тип выравнивающей кривой и находятся оценки параметров.

Пусть моменты случайной величины X_i известны. Обозначим их соответственно ν_1, $м_2$, $м_3$, $м_4$.

Тогда среднее суммы n независимых случайных величин будет равно

$$\nu_{1(n)} = \sum_{i=1}^{n} \nu_{1i} \quad . \tag{4.20}$$

Если случайные величины X_i равны и подчиняются одному и тому же закону распределения, то

$$\nu_{1(n)} = n\nu_1 . \tag{4.21}$$

Найдем далее центральный момент второго порядка суммы n независимых одинаково распределенных случайных величин $м_{2(n)}$.

Начнем с рассмотрения суммы двух независимых случайных величин:

$$\mu_2(X+Y) = M\left[(X+Y)-\left(m_x+m_y\right)\right]^2 = M\left[\left(X-m_x\right)+\left(Y-m_y\right)\right]^2 ,$$

где $m_x = \nu_1(x),\ \ m_y = \nu_1(y)$.

Обозначим для краткости $X-m_x = x, Y-m_y = y$. Тогда

$$\mu_2(X+Y) = M(x+y)^2 = M\left(x^2+2xy+y^2\right) = M\left(x^2\right)+2M(xy)+M\left(y^2\right).$$

Поскольку $M(xy) = M(x)M(y)$, последнее выражение можно представить в виде

$$\mu_2(X+Y) = M\left(X-m_x\right)^2 + 2M\left(X-m_x\right)M\left(Y-m_y\right)+M\left(Y-m_y\right)^2,$$

или $\mu_2(X+Y) = \mu_2(X)+2\mu_1(X)\mu_1(Y)+\mu_2(Y)$.

Но центральный момент первого порядка равен нулю. Поэтому второе слагаемое здесь равно нулю, и последняя формула примет вид

$$\mu_2(X+Y)=\mu_2(X)+\mu_2(Y). \qquad (4.22)$$

На основании рассмотренного примера можно сформулировать следующее правило: при возведении в *r*-ю степень суммы случайных величин $x=X-m_x$; $y=Y-m_y$, ... в итоге следует учесть только те члены, которые не содержат первых степеней сомножителей, так как их математические ожидания равны нулю.

Используя это правило, найдем центральный момент второго порядка суммы трех случайных величин

$$\mu_2(X+Y+Z)=M(x+y+z)^2,$$

где $x=X-m_x$; $y=Y-m_y$; $z=Z-m_z$.

Итак,

$$\mu_2(X+Y+Z)=M(x+y+z)^2=M\left(x^2+y^2+z^2+\ldots\right).$$

Здесь не записаны члены, математические ожидания которых равны нулю. Следовательно,

$$\mu_2(X+Y+Z)=\mu_2(X)+\mu_2(Y)+\mu_2(Z). \qquad (4.23)$$

В случае суммы *n* независимых одинаково распределенных случайных величин

$$\mu_{2(n)}=n\mu_2. \qquad (4.24)$$

Найдем далее выражение для центрального момента третьего порядка суммы *n* независимых одинаково распределенных случайных величин.

Рассмотрим вначале сумму двух независимых случайных величин

$$\mu_3(X+Y)=M(x+y)^3=M\left(x^3+3x^2y+3xy^2+y^3\right),$$

откуда

$$\mu_3(X+Y)=\mu_3(X)+\mu_3(Y). \qquad (4.25)$$

Аналогично для суммы трех случайных величин имеем

$$\mu_3(X+Y+Z) = M(x+y+z)^3 = M\left[(x+y+z)^2(x+y+z)\right] =$$
$$= M\left[\left(x^2+y^2+z^2+2xy+2xz+2yz\right)(x+y+z)\right] = M\left(x^3+y^3+z^3+...\right)$$

Остальные члены в квадратных скобках равны нулю.

Таким образом,

$$\mu_3(X+Y+Z) = \mu_3(X)+\mu_3(Y)+\mu_3(Z)\ . \qquad (4.26)$$

Для суммы n независимых одинаково распределенных случайных величин

$$\mu_{3(n)} = n\mu_3 . \qquad (4.27)$$

И, наконец, найдем выражение для центрального момента четвертого порядка суммы n независимых одинаково распределенных случайных величин.

Начнем с суммы двух случайных величин

$$\mu_4(X+Y) = M(x+y)^4,$$

где по-прежнему $x = X - m_x; y = Y - m_y$. Итак,

$$\mu_4(X+Y) = M(x+y)^4 = M\left[(x+y)^3(x+y)\right] =$$
$$= M\left[(x^3+3x^2y+3xy^2+y^3)(x+y)\right] = M(x^4+3x^2y^2+3x^2y^2+y^4+...) =$$
$$= M(x^4)+6M(x^2)M(y^2)+M(y^4).$$

Отсюда имеем

$$\mu_4(X+Y) = \mu_4(X)+\mu_4(Y)+6\mu_2(X)\mu_2(Y). \qquad (4.28)$$

Если $X=Y$, то $\mu_{4(n=2)} = 2\mu_4 + 6\mu_2^2$. (4.29)

Найдем далее центральный момент четвертого порядка суммы трех случайных величин

$$\mu_4(X+Y+Z) = M(x+y+z)^4 = M\left[(x+y+z)(x+y+z)\right]^2 =$$
$$= M(x^2+y^2+z^2+2xy+2xz+2yz)^2 =$$
$$= M(x^4+y^4+z^4+6x^2y^2+6x^2z^2+6y^2z^2+...),$$

откуда

$$\mu_4(X+Y+Z)=\mu_4(X)+\mu_4(Y)+\mu_4(Z)+ \\ +6\mu_2(X)\mu_2(Y)+6\mu_2(X)\mu_2(Z)+6\mu_2(Y)\mu_2(Z). \qquad (4.30)$$

Если $X=Y=Z$, то

$$\mu_{4(n=3)}=3\mu_4+18\mu_2^2. \qquad (4.31)$$

На основании полученных ранее формул можно записать общее выражение для центрального момента 4-го порядка суммы n независимых одинаково распределенных случайных величин.

$$\mu_{4(n)}=n\mu_4+3n(n-1)\mu_2^2. \qquad (4.32)$$

Действительно, произведение $3n(n-1)$ при n=2 равно 6, а при n=3 равно 18.

Таким образом, моменты суммы n независимых одинаково распределенных случайных величин $X=\sum X_i$ равны

$$\left.\begin{aligned} \nu_{1(n)}&=n\nu_1 \\ \mu_{2(n)}&=n\mu_2 \\ \mu_{3(n)}&=n\mu_3 \\ \mu_{4(n)}&=n\mu_4+3n(n-1)\mu_2^2 \end{aligned}\right\} \qquad (4.33)$$

и легко вычисляются по моментам отдельной случайной величины X_i. Далее по известным моментам можно найти выравнивающее распределение суммы n независимых одинаково распределенных случайных величин.

При этом найденное выравнивающее распределение может совпадать с композицией законов распределения слагаемых (например, в случае n показательных законов), но может и не совпадать с ней (например, если случайные величины распределены по закону равномерной плотности). Это связано с тем, что моменты не определяют полностью распределения.

4.3. Центральная предельная теорема для трех систем непрерывных распределений

Используя теорию производящих функций, можно показать, что для первой системы непрерывных распределений, заданной обобщенной плотностью

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

моменты суммы n независимых одинаково распределенных случайных величин $X = \sum X_i$ связаны с моментами случайной величины X_i формулами (4.33).

Производящая функция в этом случае есть

$$M\left[e^{(X-\nu_1)t}\right],$$

где t – вспомогательный параметр.

Выразим на основании формул (4.33) показатели асимметрии и островершинности распределения суммы n независимых одинаково распределенных случайных величин $\beta_{1(n)}$ и $\beta_{2(n)}$ через аналогичные показатели отдельной случайной величины X_i, т.е. β_1 и β_2:

$$\left.\begin{aligned}\beta_{1(n)} &= \frac{\mu_{3(n)}^2}{\mu_{2(n)}^3} = \frac{n^2\mu_3^2}{n^3\mu_2^3} = \frac{\beta_1}{n}\\ \beta_{2(n)} &= \frac{\mu_{4(n)}}{\mu_{2(n)}^2} = 3 + \frac{\beta_2 - 3}{n}\end{aligned}\right\} \qquad (4.34)$$

Из формул (4.34) немедленно следует центральная предельная теорема теории вероятностей (для первой системы непрерывных распределений): распределение суммы n независимых одинаково распределенных случайных величин $X = \sum X_i$ с ростом n приближается к нормальному закону

$$p(x) = \sqrt{\frac{\alpha}{\pi}}\, e^{-\alpha\left(x-\nu_{1(n)}\right)^2}, \qquad (4.35)$$

для которого β_1=0, β_2=3. При этом на номограмме (Приложение 2) точка с координатами $(\beta_{1(n)}, \beta_{2(n)})$ с ростом n перемещается по прямой от точки (β_1, β_2) исходного распределения (случайной величины X_i) к точке (0,3) нормального закона, оценки параметров которого равны

$$\left.\begin{aligned} \nu_{1(n)} &= n\nu_1 \\ \alpha &= \frac{1}{2\mu_{2(n)}} = \frac{1}{2n\mu_2} \end{aligned}\right\}. \qquad (4.36)$$

Формулы (4.33), (4.34) позволяют также переходить от распределения суммы n независимых одинаково распределенных случайных величин к распределению отдельной случайной величины X_i.

Полученные выше результаты остаются в силе и для обобщенной плотности

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

т. е. в случае второй системы непрерывных распределений, если ее привести к форме плотности $p(x)$, т. е. представить в виде

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

Моменты случайной величины $\ln T_i$ будут задаваться формулами

$$\nu_1 = M\left(\ln T_i\right),$$
$$\mu_r = M\left(\ln T_i - \nu_1\right)^r.$$

Формулировка центральной предельной теоремы несколько изменится: распределение суммы логарифмов n независимых одинаково распределенных случайных величин $\ln T = \sum \ln T_i$ с ростом n приближается к нормально-

му закону, а произведение n случайных величин $T=T_1T_2\ldots T_n$ – к логарифмически нормальному закону

$$p(t)=\sqrt{\frac{\alpha}{\pi}}\frac{1}{t}e^{-\alpha\left(\ln t-\nu_{1(n)}\right)^2}, \qquad (4.37)$$

для которого $\beta_1=0$, $\beta_2=3$. Оценки параметров $\nu_{1(n)}$, α задаются формулами (4.36).

И, наконец, в случае третьей системы непрерывных распределений, заданных обобщенной плотностью

$$p(Y)=\frac{N(\ln Y)^{k\beta-1}}{Y}\left[1-\alpha u(\ln Y)^{\beta}\right]^{\frac{1}{u}-1},$$

полученные выше результаты остаются справедливыми, если ее также привести к форме плотности $p(x)$, т. е. представить в виде

$$Yp(Y)\ln Y=Ne^{k\beta\ln\ln Y}\left(1-\alpha u e^{\beta\ln\ln Y}\right)^{\frac{1}{u}-1}.$$

Тогда моменты случайной величины $ln\,ln\,Y_i$ будут задаваться формулами

$$\nu_1=M\left(\ln\ln Y_i\right),$$
$$\mu_r=M\left(\ln\ln Y_i-\nu_1\right)^r.$$

Центральная предельная теорема сформулируется в виде: распределение суммы двойных логарифмов n независимых одинаково распределенных случайных величин $\ln\ln Y=\sum\ln\ln Y_i$ с ростом n приближается к нормальному закону, произведение $\ln Y=\ln Y_1\ln Y_2\ldots\ln Y_n$ – к логарифмически нормальному закону, а величина $Y=e^{\ln Y_1\ln Y_2\ldots\ln Y_n}$ – к двойному логарифмически нормальному закону

$$p(Y)=\sqrt{\frac{\alpha}{\pi}}\frac{1}{Y\ln Y}e^{-\alpha\left(\ln\ln Y-\nu_{1(n)}\right)^2}, \qquad (4.38)$$

для которого β_1=0, β_2=3.

Здесь также оценки параметров $\nu_{1(n)}$, α задаются формулами (4.36).

4.4. Законы распределения среднего выборочного

Рассмотрим *n* случайных величин, распределенных по одному и тому же закону, заданному, например, обобщенной плотностью *p(x)* (или *p(t)*, или *p(y)*).

Найдем закон распределения среднего выборочного $\overline{X}$, (или $\overline{\ln T}$, или $\overline{\ln\ln Y}$) по заданному закону распределения случайной величины *X* (или $\ln T$, или *lnlnY*).

Для решения этой задачи достаточно найти центральные моменты (2–4)-го порядков среднего арифметического *n* независимых одинаково распределенных случайных величин и вычислить показатели асимметрии и островершинности. Далее по универсальному или классическому методу моментов с помощью программы легко устанавливается тип искомого распределения и вычисляются оценки его параметров.

Известно, что математическое ожидание среднего арифметического *n* независимых случайных величин с одинаковыми средними равно среднему отдельной случайной величины

$$M(\overline{X})=\nu_1. \qquad (4.39)$$

Найдем далее центральные моменты выборочного среднего. Для этого вначале докажем, что постоянный множитель можно выносить за знак центрального момента *r*-го порядка, возведя его в *r*-ю степень:

$$\mu_r(CX)=C^r\mu_r(X). \qquad (4.40)$$

Действительно,

$$\mu_r(CX) = M\left[CX - M(CX)\right]^r = C^r M\left[X - M(X)\right]^r = C^r \mu_r(X).$$

На основании (4.40) имеем:

$$\mu_2(\overline{X}) = m_{2(n)} = \mu_2\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{\mu_2\left(X_1 + X_2 + \ldots + X_n\right)}{n^2}.$$

Поскольку для n независимых одинаково распределенных случайных величин справедливо равенство (4.24)

$$\mu_{2(n)} = n\mu_2,$$

то из предыдущего выражения получим

$$\mu_2(\overline{X}) = m_{2(n)} = \frac{n\mu_2}{n^2} = \frac{\mu_2}{n}. \qquad (4.41)$$

Найдем центральный момент третьего порядка выборочного среднего. На основании равенства (4.40) можем записать

$$\mu_3(\overline{X}) = m_{3(n)} = \mu_3\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^3}\mu_3\left(X_1 + X_2 + \ldots + X_n\right).$$

Поскольку $\mu_{3(n)} = n\mu_3$ (см. формулу (4.27)), то из предыдущего выражения получим

$$\mu_3(\overline{X}) = m_{3(n)} = \frac{n\mu_3}{n^3} = \frac{\mu_3}{n^2}. \qquad (4.42)$$

Для центрального момента четвертого порядка среднего выборочного можем записать

$$\mu_4(\overline{X}) = m_{4(n)} = \mu_4\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^4}\mu_4\left(X_1 + X_2 + \ldots + X_n\right).$$

Поскольку для суммы n независимых одинаково распределенных случайных величин справедливо равенство (4.32)

$$\mu_{4(n)} = n\mu_4 + 3n(n-1)\mu_2^2,$$

то из предыдущей формулы получим

$$\mu_4(\overline{X}) = m_{4(n)} = \frac{1}{n^3}\left(\mu_4 + 3(n-1)\,\mu_2^2\right). \qquad (4.43)$$

Таким образом, моменты среднего арифметического n независимых одинаково распределенных случайных величин выражаются через моменты отдельной случайной величины посредством формул

$$\left.\begin{aligned} M(\overline{X}) &= \nu_1 \\ m_{2(n)} &= \frac{\mu_2}{n} \\ m_{3(n)} &= \frac{\mu_3}{n^2} \\ m_{4(n)} &= \frac{1}{n^3}\left(\mu_4 + 3(n-1)\mu_2^2\right) \end{aligned}\right\} \qquad (4.44)$$

Из формул (4.33) и (4.44) найдем взаимосвязь между моментами суммы и среднего арифметического n независимых одинаково распределенных случайных величин:

$$\left.\begin{aligned} M(\overline{X}) &= \frac{\nu_{1(n)}}{n} = \nu_1 \\ m_{r(n)} &= \frac{\mu_{r(n)}}{n^r} \end{aligned}\right\} \qquad (4.45)$$

Из (4.45) следует, что показатели асимметрии β_1 и островершинности β_2 для распределений суммы и среднего арифметического n независимых одинаково распределенных случайных величин одни и те же.

Зная центральные моменты, а также показатели асимметрии и островершинности, по универсальному или классическому методу моментов нетрудно найти закон распределения среднего арифметического и, следовательно, вычислить доверительные границы для среднего выборочного при заданной доверительной вероятности и любом заданном значении n

С ростом n распределение среднего выборочного $\overline{x}$ (или $\overline{\ln T}$, или $\overline{\ln \ln Y}$) асимптотически приближается к нормальному закону (центральная предельная теорема в случае первой системы непрерывных распределений).

В случае второй системы – распределение среднего геометрического случайной величины T

$$\overline{T}_{геом.} = \left(T_1\, T_2 \ldots T_n\right)^{\frac{1}{n}} = e^{\frac{\Sigma \ln Ti}{n}} = e^{\overline{\ln T}}$$

с ростом n приближается к логарифмически нормальному закону.

В случае третьей системы – распределение среднего геометрического логарифмов отдельных случайных величин

$$\overline{\ln Y}_{геом} = \left(\ln Y_1 \cdot \ln Y_2 \ldots \ln Y_n\right)^{\frac{1}{n}}$$

с ростом n приближается к логарифмически нормальному закону, а величина

$$Y = e^{\left(\ln Y_1 \cdot \ln Y_2 \ldots \ln Yn\right)^{\frac{1}{n}}}$$

– к двойному логарифмически нормальному закону.

При $n > 100$ распределение среднего выборочного $\overline{x}$ (или $\overline{\ln T}$, или $\overline{\ln \ln Y}$) можно считать нормальным.

Действительно, если распределение случайной величины X (или $\ln T$, или $ln\, ln\, Y$) характеризуется показателями $0 < \beta_1 < 4, \quad 3 < \beta_2 < 9$ (в случае трех основных систем непрерывных распределений), то распределение среднего арифметического $n = 100$ независимых одинаково распределенных случайных величин на основании формул (4.34) будет иметь показатели

$$0 < \beta_1(n) < 0{,}04; \quad 3 < \beta_{2(n)} < 3{,}06, \qquad (4.46)$$

т. е. близкие к нормальному закону.

Отметим, что формулы (4.34) и (4.44) позволяют переходить от распределения среднего арифметического n независимых одинаково распределенных случайных величин к распределению отдельной случайной величины.

Полученные результаты вскрывают характер изменения показателей $\beta_{1(n)}$, $\beta_{2(n)}$ с изменением числа n независимых одинаково распределенных случайных величин: с ростом n точка с координатами $(\beta_{1(n)};\ \beta_{2(n)})$ приближается к точке (0; 3) по кратчайшему пути, т. е. по прямой

$$\beta_{2(n)} = 3 + \frac{\beta_2 - 3}{\beta_1}\beta_{1(n)}, \qquad (4.47)$$

которая следует из (4.34).

При этом степень приближения к нормальному закону при больших n зависит от исходных значений показателей β_1, β_2 случайной величины X_i.

Установим минимально необходимое значение n, при котором распределение среднего арифметического n независимых одинаково распределенных случайных величин можно считать нормальным.

Рассмотрим рис. 4.2. На нем точкой $A(\beta_1; \beta_2)$ обозначено распределение случайной величины X_i.

Распределение среднего арифметического n независимых одинаково распределенных случайных величин представлено точкой $B(\beta_{1(n)};\ \beta_{2(n)})$.

Заштрихованная площадь – это область нормального закона. Она ограничена двумя вертикальными и двумя наклонными отрезками прямых, тангенс угла наклона которых к горизонтальной оси принят равным 1,75.

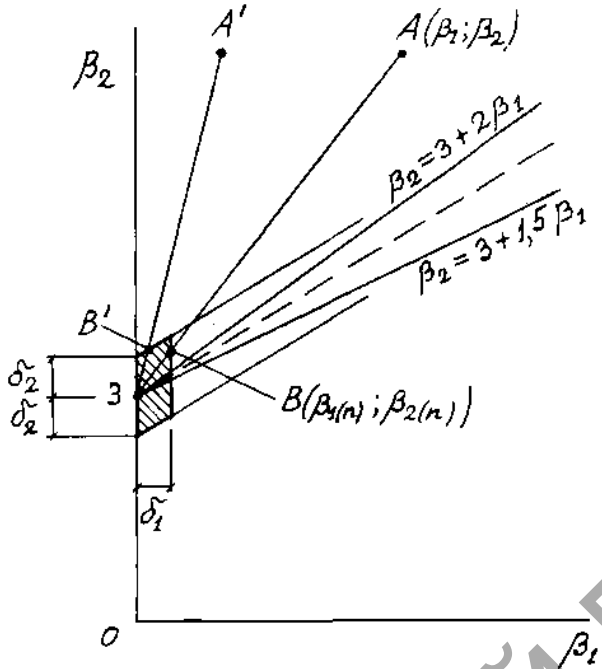


Рис. 4.2. Фрагмент номограммы в области нормального закона

Область нормального закона может быть задана и другими способами.

Если при некотором значении n точка B (или B') попадает на границу или внутрь заштрихованной области, то закон распределения среднего арифметического можно считать нормальным.

Пусть точка находится на правой вертикальной границе области нормального закона (см. рис. 4.2). При этом условии минимально необходимое значение n можно найти из формулы

$$\beta_{1(n)} = \frac{\beta_1}{n}$$

при

$$\beta_{1(n)} = \delta_1:$$

$$n_1 = \frac{\beta_1}{\delta_1}. \qquad (4.48)$$

Если точка B займет положение B', то она будет являться точкой пересечения двух наклонных прямых, которые задаются уравнениями

$$\beta_{2(n)} = 3 + \frac{\beta_2 - 3}{n};\ \beta_{2(n)} = 3 + \delta_2 + 1{,}75\beta_{1(n)} = 3 + \delta_2 + 1{,}75\frac{\beta_1}{n}.$$

Приравнивая правые части этих уравнений, найдем

$$n_2 = \frac{|\beta_2 - 3 - 1{,}75\beta_1|}{\delta_2}. \qquad (4.49)$$

Из двух полученных значений n выбираем большее.

Если в формулах (4.48), (4.49) принять $\delta_1 = \delta_2 = 0{,}05$, то значения параметра u выравнивающих распределений по абсолютной величине будут несколько меньше 0,02 (для нормального закона $u \to 0$).

Рассмотрим пример.

Пусть случайная величина X_i описывается распределением II′ типа:

$$p(x) = \frac{\alpha^k}{\Gamma(k)} \frac{1}{x^{k+1} e^{\alpha/x}} \qquad (4.50)$$

с параметрами: α = 12; k = 5. Тогда теоретические моменты будут равны (см. формулы (4.24), (4.27) при $u \to 0$, $\gamma = k$):

$$\nu_1 = \frac{\alpha}{k-1} = 3,$$

$$\mu_2 = \frac{\alpha^2}{(k-1)^2 (k-2)} = 3,$$

$$\mu_3 = \frac{4\alpha\mu_2}{(k-1)(k-3)} = 18,$$

$$\mu_4 = \frac{3\alpha}{(k-1)(k-4)}\left(\frac{\alpha\mu_2}{k-1} + 2\mu_3\right) = 405.$$

Показатели асимметрии и островершинности на основании (4.37) равны:

$$\beta_1 = \frac{16(k-2)}{(k-3)^2} = 12;\ \ \beta_2 = \frac{3(k-2)(k+5)}{(k-3)(k-4)} = 45$$

Найдем по формулам (4.48), (4.49) минимально необходимое значение n (объем выборки), при котором распределение среднего арифметического n независимых одинаково распределенных случайных величин можно считать нормальным.

При $\delta_1 = \delta_2 = 0{,}05$ имеем:

$$n_1 = \frac{\beta_1}{\delta_1} = \frac{12}{0{,}05} = 240; \quad n_2 = \frac{\beta_2 - 3 - 1{,}75\beta_1}{\delta_2} = \frac{45 - 3 - 1{,}75 \cdot 12}{0{,}05} = 420.$$

Принимаем $n = 420$.

На практике часто ограничиваются небольшими значениями n $(n = 25 \div 30)$.

Пусть $n = 25$. Тогда распределение среднего арифметического n независимых одинаково распределенных случайных величин будет иметь моменты (см. формулу (4.44):

$$m_{2(n)} = \frac{\mu_2}{n} = \frac{3}{25} = 0{,}12,$$

$$m_{3(n)} = \frac{\mu_3}{n^2} = \frac{18}{25^2} = 0{,}0288,$$

$$m_{4(n)} = \frac{1}{n^3}\left(\mu_4 + 3(n-1)\mu_2^2\right) = \frac{1}{25^3}\left(405 + 3 \cdot 24 \cdot 3^2\right) = 0{,}067392.$$

Показатели асимметрии и островершинности равны:

$$\beta_{1(n)} = \frac{m_{3(n)}^2}{m_{2(n)}^3} = \frac{0{,}0288^2}{0{,}12^3} = 0{,}48,$$

$$\beta_{2(n)} = \frac{m_{4(n)}}{m_{2(n)}^2} = \frac{0{,}067392}{0{,}12^2} = 4{,}68.$$

Контроль:

$$\beta_{1(n)} = \frac{\beta_1}{n} = \frac{12}{25} = 0{,}48; \quad \beta_{2(n)} = 3 + \frac{\beta_2 - 3}{n} = 3 + \frac{45 - 3}{25} = 4{,}68.$$

Выравнивающее распределение среднего арифметического при n=25 может быть описано обобщенной плотностью

$$p(x) = Ne^{\gamma x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}$$

(точнее, распределением III типа первой системы непрерывных распределений) с параметрами

$$\alpha u = -8{,}036098E-7;\ \beta = 5{,}062564;\ \gamma = 8{,}556006;\ u = -0{,}668744$$

и нормирующим множителем N = 3,194777E–10.

Зададим доверительную вероятность P = 0,9973 и найдем по одной из программ (например, $SNR11M97$) доверительный интервал для среднего арифметического: $2{,}038459 < \bar{x} < 4{,}50299$. Его ширина составляет $7{,}114489\,S(\bar{x})$.

Для сравнения по той же программе найдем доверительный интервал при условии справедливости нормального закона: Его ширина составляет $6\,S(\bar{x})$, т. е. ошибка в определении границ доверительного интервала по нормальному закону оказалась существенной – ($1{,}960769 < \bar{x} < 4{,}039222$).

Полученные выше формулы (4.48) и (4.49) можно использовать также для вычисления минимально необходимого значения n, при котором распределение среднего арифметического логарифмов n независимых одинаково распределенных случайных величин можно считать нормальным.

Используя универсальный метод моментов, для рассмотренной выше плотности II′ типа с параметрами α=12, k=γ=5, β=1, приведенной к форме $xp(x) = p(\ln x)$, найдем:

$$M[\ln(x)] = \nu_1 = -\frac{1}{\beta}[\Psi(k) - \ln\alpha] = 0{,}978787;$$

$$\mu_2 = \frac{1}{\beta^2}\Psi'(k) = 0{,}22132;$$

$$\mu_3 = -\frac{1}{\beta^3}\Psi''(k) = 0{,}04879;$$

$$\mu_4 = 3\mu_2^2 + \frac{1}{\beta^4}\Psi'''(k) = 0{,}16838.$$

Показатели асимметрии и островершинности равны:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0{,}2196;\ \beta_2 = \frac{\mu_4}{\mu_2^2} = 3{,}4375.$$

Далее по формулам (5.48) и (5.49 имеем (при $\delta_1 = \delta_2 = 0{,}05$)

$$n_1 = \frac{\beta_1}{\delta_1} = 4{,}392;\ n_2 = \frac{\beta_2 - 3 - 1{,}75\beta_1}{\delta_2} = 1{,}064.$$

Принимаем $n = 5$.

Итак, если случайная величина X задана плотностью (4.50), то распределение среднего арифметического логарифма случайной величины X близко к нормальному закону уже при $n = 5$, а распределение среднего арифметического самой случайной величины X близко к нормальному закону при значительно большей величине n ($n = 420$).

Если же центральные моменты высоких порядков не существуют (например, $\mu_4 \to \infty$), то и величина $n \to \infty$, т. е. распределение среднего арифметического случайной величины X ни при каком n не приближается к нормальному закону.

5. ОБЩИЙ УСТОЙЧИВЫЙ МЕТОД

Проверка показала, что универсальный метод моментов в принципе решает задачу оценивания параметров обобщенных распределений. Однако существенным его недостатком является неустойчивость, поскольку эмпирические моменты высоких порядков (μ_3^*, μ_4^*) сильно зависят от значений частот на концах распределения.

Поэтому автором обобщенных распределений был разработан общий устойчивый метод оценивания параметров [13–20], который по точности не уступает методу наибольшего правдоподобия, но значительно проще последнего.

Здесь так же, как и в случае универсального метода моментов, вводятся два показателя – асимметрии B и островершинности H, которые зависят от двух параметров формы $k=\gamma/\beta$, u. По этим показателям устанавливается тип выравнивающей кривой распределения и находятся оценки параметров k, u. Оценки двух других параметров рассчитываются по простым формулам.

Достоинством метода является его устойчивость, т. е. он мало чувствителен к выбросам на концах статистического распределения.

К недостаткам его следует отнести то, что для оценивания параметров выравнивающей кривой он требует группирования статистических данных, так же, как и метод наибольшего правдоподобия.

Если обобщенное распределение задано плотностью $p(x)$, то показатели B, H равны

$$\left.\begin{aligned} B &= M\left[p(x)(x-M(x))\right] = f(k,u) \\ H &= S_3 / S_1^3 = f(k,u) \end{aligned}\right\}, \qquad (5.1)$$

где

$$S_r = M\left[p(x)\right]^r = f(\beta,k,u). \qquad (5.2)$$

Исследования показали, что величина H задана на интервале $\sqrt{2} < H < 2$, а величина B – на интервале – $(-1/4<B<1/4)$.

Вычислим для разных типов распределений значения показателей B, H при различных значениях параметров k, u. Далее построим номограмму (Приложение 3). Она справедлива для трех основных систем непрерывных распределений, заданных первыми плотностями. При этом они должны быть приведены к форме плотности $p(x)$.

На номограмме распределения II, II′ и IV типов представлены кривыми. Типы I, I′, III, V занимают определенные области. Симметричные распределения IIIc, Vc типов представлены отрезками на оси OH: для IIIc типов $\sqrt{2} < H < \pi^2/6$; для Vc типа $\pi^2/6 < H < 2$. Распределения IVc типа представлены точкой $H = \pi^2/6$. Распределения IIc типа также представлены точкой $H = \sqrt{2}$.

На номограмме изображены области распределений с левосторонней асимметрией, для которых $0 < B_1 < 1/4$. Сюда относится часть распределений III-V типов при $0<k<(1-1/u)/2$, а также распределения I, II типов. При этом распределения приведены к форме плотности $p(x)$.

Распределения I′, II′ типов, а также часть распределений III-V типов при $(1-1/u)/2<k<1-1/u$ имеют правостороннюю асимметрию. Для них $-1/4<B<0$, причем для распределений I, II и I′, II′ типов справедливы равенства: $B' = -B, H' = H$.

Таким образом, показатели B, H однозначно определяют тип распределения, приведенного к форме плот-

ности $p(x)$. Более того, с помощью этих показателей могут быть найдены оценки параметров u, k непосредственно из номограммы.

Для распределений III-V типов при $B < 0$ из номограммы вначале находятся оценки параметров k', u (при $B > 0$), затем вычисляется величина $k=1-1/u-k'$.

Оценка параметра β для всех типов равна [21]

$$\beta = \frac{S_1}{S_1^{(z)}}. \tag{5.3}$$

Тогда $\gamma = k\beta$.

Оценки параметра α для распределений II, II′ типов и произведения αu для остальных типов равны [21]:

$$\left.\begin{array}{ll} \text{Типы I, I}': & \alpha u = e^{\pm\left(\nu - \beta\nu_1^{(z)}\right)} \\ \text{Типы II, II}': & \alpha = e^{\pm\left(\nu_1^{(z)} - \beta\nu_1\right)} \\ \text{Типы III - V}: & \alpha u = -e^{\nu_1^{(z)} - \beta\nu_1} \end{array}\right\} \tag{5.4}$$

где в зависимости от типа распределения величины $\nu_1^{(z)}$ и $S_1^{(z)}$ рассчитываются по формулам:

Типы I, I′:

$$\left.\begin{array}{l} \nu_1^{(z)} = \pm\left[\Psi(k) - \Psi\left(k+\frac{1}{u}\right)\right] \\ S_1^{(z)} = \frac{1}{2\sqrt{\pi}} \frac{2\left(k+\frac{1}{u}\right)-1}{\frac{2}{u}-1} \cdot \frac{g(k)g\left(\frac{1}{u}\right)}{g\left(k+\frac{1}{u}\right)} \end{array}\right\} \tag{5.5}$$

Типы II, II′:

$$\nu_1^{(z)} = \pm\Psi(k); \quad S_1^{(z)} = \frac{g(k)}{2\sqrt{\pi}} \tag{5.6}$$

Типы III-V:

$$\left.\begin{aligned} \nu_1^{(z)} &= \Psi(k) - \Psi\left(1-\frac{1}{u}-k\right) \\ S_1^{(z)} &= \frac{1}{2\sqrt{\pi}} \frac{g(k)\; g\left(1-\frac{1}{u}-k\right)}{g\left(1-\frac{1}{u}\right)} \end{aligned}\right\} \qquad (5.7)$$

Величина

$$g(k) = \frac{\Gamma\left(k+\frac{1}{2}\right)}{\Gamma(k)} \qquad (5.8)$$

может быть вычислена по приближенным формулам:

– при $x > 4$

$$g(x) \approx \sqrt{x}\, e^{-\frac{1}{8x}+\frac{1}{192x^3}-\frac{1}{640x^5}+\ldots}; \qquad (5.9)$$

– при $0 < x \leq 4$

$$g(x) = \frac{g(x+n)}{\prod_{i=1}^{n}\left[1+\frac{1}{2(x+i-1)}\right]}, \qquad (5.10)$$

где $n \geq 4$;

$$g(x+n) = \sqrt{x+n}\, e^{\frac{1}{8(x+n)}+\frac{1}{192(x+n)^3}-\ldots} \qquad (5.11)$$

Для облегчения расчетов в Приложении 1 приводятся также значения функции $g(x)$.

Для установления типа выравнивающей кривой распределения и нахождения оценок параметров по общему устойчивому методу достаточно найти значения статистических показателей ν_1^*, S_1^*, B^*, H^* и приравнять их соответствующим теоретическим. Эти показатели для каждой системы непрерывных распределений вычисляются по-своему. Но номограмма применима ко всем трем системам непрерывных распределений.

Оценки статистических показателей в случае выравнивающих распределений, заданных плотностью $p(x)$, вычисляются по формулам:

$$\left.\begin{aligned} \nu_1^* &= \bar{x} = \sum_{i-1}^{n} x_i p_i h_i \\ S_1^* &= \sum_{i-1}^{n} p_i^2 h_i,\ S_3^* = \sum_{i-1}^{n} p_i^4 h_i \\ B_1^* &= \sum_{i-1}^{n} x_i p_i^2 h_i - \nu_1^* S_1^*;\ \text{H}^* = \frac{S_3^*}{(S_1^*)^3} \end{aligned}\right\} \qquad (5.12)$$

где $p_i = m_i/(Mh_i)$ – эмпирическая плотность распределения; m_i – наблюденная частота случайной величины X в i-ом интервале $(i = 1,2...,n); M = \sum_{i=1}^{n} m_i$ – наблюденная частота во всех n интервалах (объем выборки); h_i – ширина i-го интервала; x_i – значение случайной величины X в середине i-го интервала.

Формулы (5.12) можно выразить через абсолютные частоты m_i:

$$\left.\begin{aligned} \nu_1^* &= \bar{x} = \sum_{i-1}^{n} x_i \frac{m_i}{M} \\ S_1^* &= \sum_{i-1}^{n} \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i};\ S_3^* = \sum_{i-1}^{n} \left(\frac{m_i}{M}\right)^4 \frac{1}{h_i^3} \\ B^* &= \sum_{i-1}^{n} x_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} - \nu_1^* S_1^*;\ H^* = \frac{S_3^*}{(S_1^{(*)})^3} \end{aligned}\right\} \qquad (5.13)$$

Показатель островершинности H^* при h_i = const примет вид

$$H^* = M^2 \frac{\sum_{i=1}^{n} m_i^4}{\left(\sum_{i=1}^{n} m_i^2\right)^3}, \qquad (5.14)$$

т. е. ширина интервала не входит в формулу (5.14). Отсюда следует вывод, что ширину интервала группирования статистических данных лучше принимать постоянной (по крайней мере для распределений, близких к симметричным).

Если выравнивающее распределение задано обобщенной плотностью $p(t)$, статистические показатели рассчитываются по формулам:

$$\left.\begin{aligned} &\nu_1^* = \overline{\ln t} = \sum_{i-1}^{n} \ln t_i \frac{m_i}{M};\ \ S_1^* = \sum_{i-1}^{n} t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} \\ &S_3^* = \sum_{i-1}^{n} t_i^3 \left(\frac{m_i}{M}\right)^4 \frac{1}{h^3};\ H_3^* = \frac{S_3^*}{\left(S_1^*\right)^3} \\ &B_1^* = \sum_{i-1}^{n} t_i \ln t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} - \nu_1^* S_1^* \end{aligned}\right\} \qquad (5.15)$$

При h_i = const

$$H^* = M^2 \frac{\sum_{i=1}^{n} t_i^3 m_i^4}{\left(\sum_{i=1}^{n} t_i m_i^2\right)^3}. \qquad (5.16)$$

Для установления типа выравнивающей кривой и нахождения оценок параметров по общему устойчивому методу автором созданы программы SNR1, SNR2, SNR3.

В заключение отметим, что общий устойчивый метод основан на взаимосвязи между законами распределения случайных величин X и Z.

Запишем обобщенную плотность $p(x)$

$$p(x) = Ne^{k\beta x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}.$$

Пусть для определенности параметр $u > 0$.

Введем случайную величину

$$Z = \alpha u e^{\beta X}. \qquad (5.17)$$

Тогда плотность $p(z)$ будет равна

$$p(z) = p(x)\frac{dx}{dz}.$$

Поскольку на основании (5.17)

$$x = \frac{1}{\beta}(\ln z - \ln \alpha u),$$

то

$$\frac{dx}{dz} = \frac{1}{\beta z},\ p(z) = \frac{p(x)}{\beta z}, \qquad (5.18)$$

откуда имеем замечательное равенство

$$\beta z p(z) = p(x). \qquad (5.19)$$

На его базе строится общий устойчивый метод оценивания параметров.

Поскольку плотность *p(z)* является функцией двух параметров формы $k = \gamma / \beta, u$, то последняя формула позволяет ввести критерии, зависящие от этих двух параметров.

Запишем на основании формулы (5.19) следующее равенство:

$$\beta^r M[zp(z)]^r = M[p(x)]^r.$$

Введем обозначения

$$M[zp(z)]^r = S_r^{(z)}; M[p(x)]^r = S_r.$$

Тогда последнее равенство перепишется в виде

$$\beta^r \boldsymbol{S}_r^{(z)} = \boldsymbol{S}_r. \qquad (5.20)$$

Формула (5.20) позволяет найти значение параметра β (например, при *r* =1), а также получить критерий островершинности, зависящий от двух параметров *k, u.* Для этого необходимо взять отношение S_2 / S_1^2 либо S_3 / S_1^3. Последнее оказалось наиболее подходящим.

Таким путем был получен показатель островершинности *H*.

Показатель асимметрии B найден из условия, чтобы для симметричных распределений он был равен нулю и в то же время использовал ранее введенные величины. Такой показатель может иметь вид

$$B = M[xp(x)] - M(x)M[p(x)]$$

или

$$\boldsymbol{B = M[p(x)(x - M(x))]}.$$

Покажем, что он зависит от двух параметров *k, u*.

Поскольку $p(x) = \beta z p(z),\ x = \frac{1}{\beta}(\ln z - \ln \alpha u)$, то

$$B = M\left[\beta z p(z)\left(\frac{1}{\beta}\ln z - \frac{1}{\beta}M(\ln z\right)\right] = M\left[z p(z)(\ln z - M(\ln z)\right] = f(k,u).$$

По показателям *B, H* строится номограмма, позволяющая устанавливать тип выравнивающей кривой распределения и находить оценки параметров *k, u*. Оценка параметра β вычисляется по величинам S_1 и $S_1^{(z)}$. Оценка параметра α или произведения α*u* вычисляется по тем же формулам, что и в случае универсального метода моментов.

Если в качестве показателей асимметрии и островершинности использовать величины

$$B = F(x_c) - 0{,}5,\ H = \frac{p(x_c)}{M\left[p(x)\right]},$$

где x_c – мода, то можно построить аналогичную номограмму для установления типа выравнивающей кривой распределения и нахождения в первом приближении оценок параметров *k, u* по координатам одной характерной точки *C* и среднему значению плотности *p*(*x*).

6. РАНГОВЫЕ РАСПРЕДЕЛЕНИЯ В БИБЛИОТЕЧНО-ИНФОРМАЦИОННОЙ ДЕЯТЕЛЬНОСТИ

6.1. Ранговые распределения

Статистические данные, полученные в результате наблюдения, представляют собой простой статистический ряд. Чтобы извлечь из этого ряда информацию, его упорядочивают либо по возрастанию значений случайной величины, либо по убыванию. В обоих случаях получим вариационный (ранжированный) ряд.

Статистические распределения, в том числе ранговые, широко используются в научных исследованиях. Анализ этих распределений позволяет ученым совершать открытия. Так, в физике была введена постоянная Планка, в химии Д.И. Менделеевым построена Периодическая система элементов, в информатике С. Бредфордом сформулирован закон рассеяния публикаций по периодическим изданиям.

При ранжировании статистических данных открываются возможности извлечения новой информации, изучения структуры выборки, вычисления различных показателей, более полно характеризующих исследуемую случайную величину.

Наличие обобщенных распределений для описания статистических вариационных рядов открывает перед исследователем новые перспективы.

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. Спрашивается, 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.

6.2. Форма представления ранговых распределений

Статистическое ранговое распределение можно представить в виде обычной гистограммы, которую можно аппроксимировать непрерывной убывающей кривой распределения. Для большей наглядности статистического рангового распределения обычно строят график зависимости $\ln p_r = f(\ln r)$, где p_r – относительная частота слова частотного словаря с рангом r или доля статей по заданному предмету в журнале с рангом r.

Однако принятая форма представления ранговых распределений несет слишком мало информации о статистическом распределении. На таком графике колебания

частот мало заметны, поскольку последние изображены в логарифмическом масштабе. Кроме того, такое преобразование кривой распределения не имеет вероятностного смысла, а построенная таким путем статистическая кривая не является ранговым распределением.

В связи с вышесказанным целесообразно перейти к другой форме представления ранговых распределений, а именно, $rp_r = f(lnr)$ [14, 16]. По оси ординат будем откладывать произведение ранга слова (журнала) на его относительную частоту (или долю статей), а по оси абсцисс – натуральный логарифм ранга. Указанная зависимость представляет собой **закон распределения**, что дает возможность исследовать статистические ранговые распределения. Площадь под этой кривой распределения, как и положено, равна единице.

График зависимости $rp_r = f(lnr)$ имеет принципиальные преимущества перед традиционной формой представления ранговых распределений. Во-первых, он описывается первой системой непрерывных распределений (плотностью $p(x)$). В данном случае $rp_r = p(x)$; $lnr = x$. Во-вторых, на такой кривой видны колебания самих частот (по оси ординат), а не их логарифмов. В-третьих, статистические ранговые распределения однородных случайных величин имеют одновершинную кривую распределения (этим свойством обладает обобщенная плотность $p(x)$ при $u<1$). Это позволяет устанавливать однородность или неоднородность статистических ранговых распределений, выделять неоднородную часть, а также решать другие задачи [15].

6.3. Универсальный закон рассеяния публикаций

Глубокое изучение любой дисциплины, допускающей применение количественных методов исследования, должно сопровождаться построением и использованием математических или вероятностно-статистических мо-

делей. Так, в информатике и математической лингвистике широко известны такие математические модели, как закон Дж.Ципфа

$$p_r = \frac{k}{r^{\gamma}}, \qquad (6.1)$$

применяемый для описания ранговых распределений слов частотного словаря, а также журналов, упорядоченных по убыванию числа помещенных в них статей по заданному предмету; закон С.Бредфорда рассеяния публикаций; закон старения публикаций и др. К сожалению, каждый из этих законов, как правило, используется сам по себе, без взаимосвязи с другими законами, т. е. без указания его места в ряду других, более общих законов распределения [13–21], которые детально исследованы и прошли большую практическую апробацию. В таком случае не мог бы использоваться «закон» Ципфа, да и закон Бредфорда имел бы точную математическую формулировку. Отметим, что закон Ципфа никогда не подтверждался на статистических ранговых распределениях, графики которых построены в системе координат $rp_r = f(\ln r)$, тем не менее его поддержали многие математики, причем без предварительной его проверки на различных частотных словарях, изо всех сил пытались дать обоснование этому «закону». Были потрачены многие годы разных исследователей на несуществующий закон. Дело дошло до абсурда – один рецензент не дал рекомендации на публикацию моей статьи по теории обобщенных распределений из-за того, что в ней не был «найден» закон Ципфа и дал мне совет обратиться к работам авторов, которые проявляли активность по обоснованию этого несуществующего закона.

Для аппроксимации различных статистических распределений, в том числе ранговых, могут использовать-

ся построенные автором системы непрерывных распределений, поскольку они включают как частные случаи множество известных распределений, в том числе указанные выше.

С помощью обобщенных распределений можно описать практически любое статистическое распределение, если оно представляет собой однородную совокупность значений непрерывной случайной величины.

Так, первая система хорошо описывает распределение первоисточников по числу цитирований в зависимости от года издания (закон старения публикаций), а также распределение технологических погрешностей, распределение работников некоторой организации по возрасту.

Вторая система описывает ранговые распределения журналов, упорядоченных по убыванию числа помещенных в них статей по заданному предмету. Из этой же системы выводится математически точная формулировка закона рассеяния публикаций в смысле С. Бредфорда. Она описывает также распределение слов словаря, фраз и предложений по длине, распределение работающих по уровню заработной платы.

Третья система описывает ранговые распределения знаменательных (полнозначных) слов частотного словаря, а также частотных словарей дескрипторов, терминов.

Четвертая система описывает распределение простых чисел.

Закон Ципфа входит как частный случай во вторую и третью системы непрерывных распределений. Закон Вейбулла, который также используется в математической лингвистике и информатике, относится ко второй системе распределений группы А. Из второй системы следуют основные распределения семейства К.Пирсона.

Таким образом, в результате разработки теории обобщенных распределений информатика, математическая лингвистика, информетрия, квалиметрия, системы ме-

неджмента качества и т. д. приобрели мощный математический аппарат, позволяющий решать множество задач на более высоком уровне.

Решим на базе обобщенных распределений наиболее важные проблемы в информатике: дадим математически точную формулировку закона рассеяния публикаций в толковании С. Бредфорда, а также установим в самом общем виде законы рассеяния и старения публикаций.

В 1948 г. С. Бредфорд дал окончательную формулировку открытого им в 1934г. закона рассеяния публикаций в периодических изданиях. Приведем формулировку этого закона, позаимствованную из книги [9, с. 178, 179]: «Если научные журналы расположить в порядке убывания числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп или зон, каждая из которых содержит столько же статей, что и ядро. Тогда числа журналов в ядре и в последующих зонах будут относиться как $1 : n : n^2$».

Несмотря на некоторую неопределенность этой формулировки, С. Бредфорду удалось отразить в ней суть закона рассеяния публикаций, по крайней мере в первом приближении.

Все последующие попытки других исследователей по совершенствованию модели С.Бредфорда оказались безуспешными. И это закономерно, поскольку исследователи строили свои модели в основном на законе Ципфа и предположении о равенстве числа статей в ядре журналов и зонах рассеяния.

Математически точную формулировку закона рассеяния можно дать лишь на базе универсальных распределений, которые с высокой точностью описывают статистические ранговые распределения журналов по раз-

личным отраслям знания. Эти распределения для каждого статистического вариационного ряда имеют свои параметры.

Исследования показали [13–21], что статистические ранговые распределения журналов, упорядоченных по убыванию числа помещенных в них статей по заданному предмету, хорошо аппроксимируются обобщенной плотностью

$$p(t)=Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \qquad (6.2)$$

или в более общем случае – второй системой непрерывных распределений, заданной тремя обобщенными плотностями (4.1), (4.3), (4.4).

Однако убывающая кривая «ранг – относительная частота», т. е. $p_r = f(r)$ не имеет никаких особых точек, которые позволили бы дать математически точную формулировку закона рассеяния публикаций. Поэтому автором введена другая форма представления ранговых распределений, а именно: $rp_r = f(lnr)$ [16]. Ранее было показано, 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

$$p(x)=Ne^{k\beta x}\left(1-\alpha u e^{k\beta x}\right)^{\frac{1}{u}-1}. \qquad (6.3)$$

Другими словами, такое преобразование распределений второй системы сводит их к распределениям первой системы, т. е. плотность $p(t)$ преобразуется к плотности $p(x)$. Действительно, если умножить обе части плотности (6.2) на величину t и записать выражение t^{β} в виде $e^{\beta lnt}$, что одно и то же, то из плотности $p(t)$ получим плотность $p(x)$. График этой плотности, т. е. кривая распределения имеет три характерные точки: моду С и две точки перегиба А и В. При этом точки перегиба

расположены на равных расстояниях от моды С – и в этом, по нашему мнению, состоит суть закона рассеяния в толковании Бредфорда! Примем эти точки в качестве границ ядра и зон рассеяния.

Итак, для плотности *p(x)* имеем

$$x_C - x_A = x_B - x_C. \qquad (6.4)$$

Учитывая взаимосвязи между первой и второй системами непрерывных распределений, т. е. $x = lnt$, $p(x) = tp(t)$, для плотности *p(t)* можем записать

$$lnt_C - lnt_A = lnt_B - lnt_C, \qquad (6.5)$$

откуда имеем равенство

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \qquad (6.6)$$

Точки А, С, В делят все журналы в ранжированном ряду на четыре части: ядро и три зоны рассеяния. Количество журналов, входящих в ядро, определяется равенством $t_Я = t_A$. Количество журналов в первой зоне $t_I = t_C - t_A$; во второй зоне $t_{II} = t_B - t_C$. Остальные журналы относятся к III зоне: $t_{III} > t_B$.

Теперь можно дать математически точную формулировку закона рассеяния публикаций. Она несколько отличается от формулировки Бредфорда $(t_Я : t_I : t_{II} = 1 : n : n^2)$.

Из формулы (6.6) следует, что между количеством наименований журналов **от начала частотного списка до точек А, С, В** имеется соотношение:

$$t_A : t_C : t_B = t_A(1 : n : n^2) \qquad (6.7)$$

В то же время между количеством наименований журналов **в ядре и последующих зонах имеется другое соотношение (при** $(t_Я = t_A)$**)**

$$t_Я : t_I : t_{II} = t_A(1 : (n-1) : (n-1)n). \qquad (6.8)$$

Как видим, формулировка Бредфорда является комбинацией из двух точных формул (6.7) и (6.8). При этом

из закона Бредфорда неясно, как определяется число журналов, образующих ядро, какая доля статей содержится в нем, сколько может быть зон рассеяния, чему равна величина *n*. Обобщенная плотность *p(t)* дает возможность однозначно ответить на все эти вопросы.

Журналы, входящие в ядро, содержат долю статей, равную функции распределения в точке А, т. е. $F(t_A)$. Аналогично доля статей в журналах, входящих в ядро и первую зону рассеяния, составляет $F(t_C)$, и т. д., следовательно, доля статей в первой зоне рассеяния составляет $F(t_C) - F(t_A)$; во второй зоне $F(t_B) - F(t_C)$, а в третьей зоне – $1 - F(t_B)$.

Количество зон рассеяния, как правило, равно трем. Но при определенных значениях параметров аппроксимирующей плотности *p(t)* оно может быть меньше.

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 $dtp(t)/dlnt = 0$ и в общем случае для распределений I-V типов равна

$$t_C = \left(\frac{k}{\alpha(1+ku-u)} \right)^{1/\beta}. \qquad (6.9)$$

Величина *n* задается формулой

$$n = \left[1 + \frac{1-u \mp \sqrt{[4k(1+ku-u)+(1-u)](1-u)}}{2k(1+ku-u)} \right]^{1/\beta}. \qquad (6.10)$$

В формуле (6.10) в числителе знак «минус» относится к распределениям 5-го типа. Поскольку такие ранговые распределения встречаются весьма редко, во многих статьях автора этот знак опущен.

Абсциссы точек перегиба вычисляются по формулам:

$$t_A = t_C/n;\ t_B = t_C \cdot n.$$

Рассмотрим один частный случай. Ранговые распределения журналов часто описываются законом Вейбулла с функцией распределения

$$F(t)=1-e^{-\alpha t^{\beta}}, \qquad (6.11)$$

которая следует из плотности p(t) при $u \to 0$, $k = 1$. Тогда из формул (6.9) и (6.10) при $k = 1$ имеем равенства:

$$t_c=\left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}, \quad n=\left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \qquad (6.12)$$

При этом значения функции распределения в трех характерных точках независимо от значений параметров равны: $F(t_A) = 0.3175$; $F(t_C) = 0.6321$; $F(t_B) = 0.9271$. Это значит, что в ядро журналов входит 32% от всех статей по данному предмету; в ядро и первую зону рассеяния – 63% статей, а в ядро и первые две зоны – 93% статей. По зонам рассеяния доли статей распределяются так: первая зона рассеяния содержит 31% статей; вторая зона – 30% статей. На третью зону приходится лишь 7% статей. Между числом наименований журналов в ядре и последующих зонах справедливо общее соотношение (6.8).

Отсюда следует, что для более полного удовлетворения информационных потребностей специалистов справочно-информационный фонд должен комплектоваться по крайней мере теми журналами, которые образуют ядро и первые две зоны рассеяния. Количество таких журналов равно t_B при этом полнота комплектования фонда $F(t_B) = 0.93$ (под полнотой комплектования фонда понимается вероятность удовлетворения запросов потребителей информации этим фондом). Величина t_B может характеризовать некоторый оптимальный объем справочно-информационного фонда с точки зрения полноты его комплектования при ограниченных материальных ресурсах.

Дальнейшие исследования показали [13–15], что любая обобщенная плотность, входящая во вторую систему непрерывных распределений, дает закон рассеяния в виде формул (6.7), (6.8), 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 в общем случае зависят как от вида распределения, так и 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 тремя обобщенными плотностями (4.1.), (4.3), (4.4). Именно обобщенные плотности позволяют 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 обобщенные плотности позволяют дать математически точную формулировку закона рассеяния публикаций в виде формул (6.7), (6.8), справедливых для всех убывающих распределений второй системы. Другими словами, в качестве универсального закона рассеяния публикаций выступает вторая система непрерывных распределений, а закон рассеяния в виде формул (6.7), (6.8) является лишь следствием свойств ранговых распределений, частным случаем универсального закона, отражающим соотношение между абсциссами характерных точек на кривых распределения. Значения же функции распределения в характерных точках в этих формулах не задействованы. 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.

Если же за основу принять формулировку закона рассеяния С. Бредфорда, то нетрудно прийти к выводу, что «не существует универсальной математической модели, пригодной для описания распределения публикаций и журналов вне зависимости от их тематической принадлежности» [39, с. 57].

Цитата приведена из Справочника библиографа 2002г. издания. Далее авторы Справочника иллюстрируют сказанное известной таблицей, в которой приводятся ранги журналов, публикующих статьи по химии и химической технологии, и процент статей от общего числа библиографированных.

Но эти табличные данные очень хорошо описываются законом Вейбулла, который следует из обобщенной плотности (4.1) при $k = 1$, $u \to 0$, т. е.

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}} . \qquad (6.13)$$

Функция распределения $F(t)$, обозначающая в данном случае суммарную долю библиографированных статей из t первых журналов частотного списка, задается формулой (6.11), при этом параметры закона Вейбулла приближенно равны: $\alpha = 0{,}036$; $\beta = 0{,}53$ [22, 32].

Обобщенная плотность $p(t)$, заданная формулой (6.2), и ее частный случай – распределение Вейбулла (6.13) дают один и тот же закон рассеяния в виде формул (6.7), (6.8), но значения функции распределения в трех характерных точках в обоих случаях различны и зависят от вида распределения и значений параметров.

Таким образом, универсальная математическая модель, о которой говорится в Справочнике, существует и задается обобщенной плотностью $p(t)$, а в общем случае – второй системой непрерывных распределений.

Более того, из этих распределений как следствие их свойств вытекают формулы (6.7) и (6.8), которые уточняют закон рассеяния С. Бредфорда.

Удовлетворить же требованиям закона С. Бредфорда в виде $t_Я : t_I : t_{II} = 1 : n : n^2$ при условии, что число статей в ядре журналов и зонах рассеяния одинаково, действительно не может никакое теоретическое распределение.

Остается только признать в качестве универсального закона рассеяния публикаций вторую систему непрерывных распределений (по крайней мере ту ее часть, которая хорошо описывает ранговые распределения). Тогда все противоречия снимаются.

Обобщенные распределения позволяют также вычислять число журналов, содержащих заранее определенную долю статей по заданному предмету. Например, в случае распределений группы А I-III типов с функцией распределения

$$F(t) = 1 - \left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}}, \qquad (6.14)$$

имеем

$$t = \left\{ \frac{1}{\alpha u} \left[1 - (1 - F(t))^{u} \right] \right\}^{\frac{1}{\beta}}, \qquad (6.15)$$

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 (6.14) имеем распределение Вейбулла, из которого находим

$$t = \left(\frac{1}{\alpha} \ln \frac{1}{1 - F(t)} \right)^{\frac{1}{\beta}}. \qquad (6.16)$$

В случае распределений группы Б эта задача решается с помощью соответствующих компьютерных программ.

Поскольку вторая система непрерывных распределений хорошо описывает статистические распределения

работающих по уровню заработной платы, ее можно также использовать для естественной и объективной классификации работающих с помощью характерных точек с целью оптимизации уровня подоходного налога, который может быть различным для разных зон.

Необходимо также отметить, что вторая система непрерывных распределений, в частности, плотность *p(t)* хорошо описывает такие статистические распределения, как слов по длине (в словаре), фраз по количеству словоупотреблений, словосочетаний по длине, терминов по длине и др., а также ранговые распределения научных сотрудников по продуктивности [14, 18, 20, 21, 29].

В заключение следует дать обоснование использования трех характерных точек для вывода математически точной формулировки закона рассеяния публикаций в смысле Бредфорда.

Точки A, C, B являются особыми точками кривой распределения, заданной плотностью *p(x)*. Между ординатами трех характерных точек существует соотношение

$$\lambda = \frac{[p(x_C)]^2}{p(x_A)p(x_B)} = \left(\frac{1-2u}{1-u}\right)^{1-\frac{1}{u}} = \left(1+\frac{1}{1-1/u}\right)^{1-\frac{1}{u}}, \qquad (6.17)$$

из которого видно, что показатель λ зависит лишь от одного параметра формы *u*, т.е. он является идентификатором типа кривой распределения. В зависимости от значений показателя λ распределения, заданные плотностью *p(x)*, можно разделить на типы. Для I,I′ типов $e < \lambda < \infty$ (при $0 < u < 1/2$); для II,II′ типов $\lambda = e$; для III типа $2 < \lambda < e$; для IV типа $\lambda = 2$; и наконец для V типа $1 < \lambda < 2$. Таким образом, по ординатам трех характерных точек A, C, B может быть однозначно установлен тип кривой распределения и найдена оценка параметра *u* из соотношения (6.17) при известном λ.

Анализ плотности *p(x)* показывает, что у распределений II-V, II′ типов существуют все три точки *A, C, B*. У распределений I типа точка перегиба *B* существует при $0 < u < 1/2$, точки *A, C* – при $0 < u < 1$. У распределений Iг типа точка перегиба *A* существует при $0 < u < 1/2$, точки *B, C* – при $0 < u < 1$.

Если кривая рангового распределения, приведенная к форме плотности *p(x),* имеет три характерные точки, то для такого распределения существует ядро и три (не более!) зоны рассеяния.

6.4. 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 (по той же причине, что и в случае закона рассеяния публикаций, т.е. 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, в частности, обобщенной плотностью p(x) [13,14,17], 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. Обобщенные плотности позволяют 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)* задаются формулами (в случае распределений I-V типов)

$$x_c = \frac{1}{\beta} \ln \frac{k}{\alpha(1 + ku - u)}, \qquad (6.18)$$

$$x_{A,B} = x_C \mp \ln n, \qquad (6.19)$$

где величина *n* рассчитывается по прежней формуле (6.10).

6.5. Ранговые распределения лексических единиц

Для вычисления аппроксимирующей кривой распределения, заданной плотностью p(t), ее целесообразно привести к форме распределений с плотностью (p(x), т.е. привести к первой системе непрерывных распределений. Между указанными плотностями существуют соотношения:

$p(x) = tp(t); x = \ln t;$ величина t соответствует рангу r статистического распределения.

Воспользуемся данными «Частотного словаря русского языка» (ЧСРЯ) [41]. Он построен на выборке объемом X=1056382 словоупотребления, объем словаря У=39268 разных слов. Предполагая, что аппроксимирующее распределение задается плотностью p(t), которая относится ко второй системе непрерывных распре-

делений, построим по «Таблице распределения частот», приведенной в [41, с. 895–915], график зависимости

$$rp_r = f(\ln r), \tag{6.20}$$

т. е. приведем ранговое распределение к форме плотности p(x). Здесь $p_r = m_r / X; m_r, p_r$ – абсолютная и относительная частоты слова с рангом r.

Путем такого же преобразования приводится к этой форме теоретическое распределение (p(t): $tp(t) = f(\ln t)$.

При построении указанного графика в целях выявления характерных особенностей статистического распределения и достижения наибольшей наглядности и точности графического изображения необходимо соблюдать следующие правила [14]:

– начальный участок графика (для слов с рангами $1 \le r \le 50$) строится в соответствии с формулой $(r - 0.5)p_r = f(\ln(r - 0.5))$, которая учитывает дискретность статистического распределения слов по частоте их употребления в выборке и непрерывность аппроксимирующего распределения. Здесь принимается, что относительные частоты сосредоточены в серединах интервалов 0-1, 1-2 и т. д.;

– конечный участок графика (для слов с частотами $1 \le m \le 50$) строится в соответствии с формулой

$$r\frac{m_r + m_{r+1}}{2X} = f(\ln r), \tag{6.21}$$

где частоты двух соседних слов с рангами r и r + 1, как правило, различаются на единицу, при этом учитывается средняя их частота $\overline{m} = (m_r + m_{r+1})/2$. Из формулы (6.21) следует, что последняя справа точка по построению имеет ординату

$$rp_r = \frac{Y}{2X},$$

поскольку в этом случае $r = Y, m_r = 1, m_{r+1} = 0$;

– средняя часть графика строится в соответствии с формулой (6.20);

– для рангов 1 < r < (10–15) строятся все точки подряд; в остальных случаях точки строятся по рангам, возрастающим в геометрической прогрессии со знаменателем 1.25–1.5;

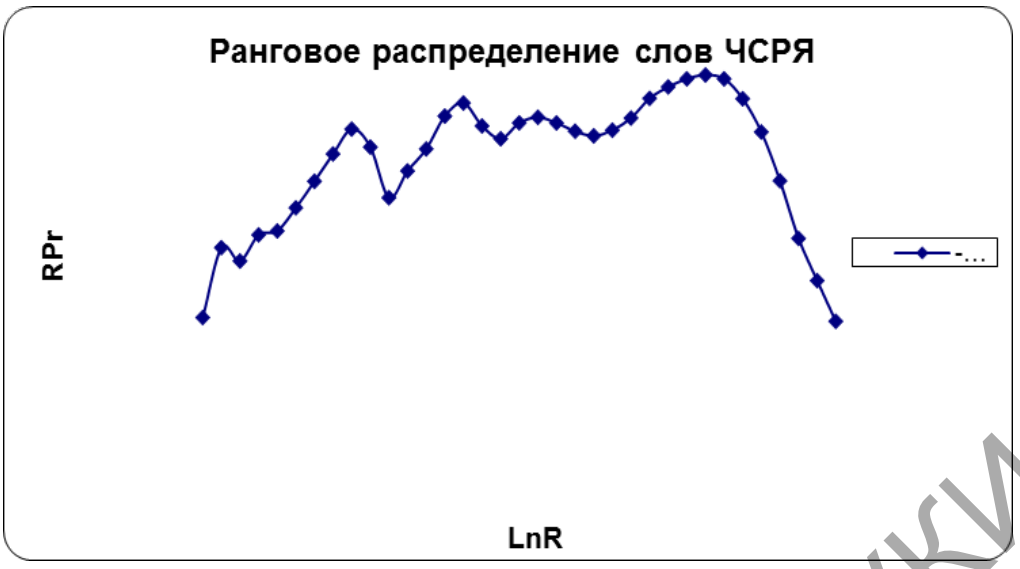
– полученные точки соединяются отрезками прямых.

Для построения графика зависимости (6.20) по данным ЧСРЯ составим таблицу 6.1. Для ее составления достаточно выписать из «Таблицы распределения частот» ранги, которые приведены в графе 5 – «накопленное число слов абсолютное» и абсолютные частоты (графа 2). По этим статистическим данным найдем значения $\ln r$, rp_r (графы 3, 4 табл.). При этом будем придерживаться сформулированных выше правил.

Таблица 6.1

Статистические данные для построения кривой распределения в системе координат ($\ln r; rp_r$)

r	m	lnr	rm/x
0,5	42854	-0,69315	0,020283
1,5	36266	0,405465	0,051496
2,5	19228	0,916291	0,045504
3,5	17261	1,252763	0,057189
4,5	13839	1,504077	0,058952
5,5	13307	1,704748	0,069282
6,5	13185	1,871802	0,081128
7,5	13143	2,014903	0,093311
8,5	12975	2,140066	0,104401
9,5	10719	2,251292	0,096396
10,5	7425	2,351375	0,073801



На *рис. 6.1* представлена статистическая кривая распределения $rp_r = f(\ln r)$ слов по частоте их употребления в текстах по данным ЧСРЯ. Отличительной особенностью этой кривой является ее неправильная форма, что свидетельствует о сложной статистической структуре словаря.

Кривые распределения, построенные в этой же системе координат по другим частотным словарям, имеют похожую форму, причем, левая часть кривых, как правило, имеет высокие гребни несмотря на большие частоты первых слов частотных словарей. Правая часть кривых распределения всегда более плавная. Распределения обычно имеют левостороннюю асимметрию (левая часть кривой длиннее правой).

Анализ формы построенной кривой распределения позволяет сделать некоторые полезные выводы.

Во-первых, многовершинность кривой распределения свидетельствует о неоднородности лексического состава частотного словаря русского языка. Так и должно быть, поскольку в частотном словаре представлены как знаменательные (полнозначные), так и служебные слова.

Во-вторых, построенный график позволяет выделить неоднородную часть: последняя впадина перед закономерным ростом и последующим убыванием кривой

имеет абсциссу *lnr = 4.6–4.9*, что соответствует рангам 100–135. Таким образом, около 100 первых слов частотного словаря представляют собой неоднородную часть. В основном это служебные слова. Всю остальную лексику можно считать однородной (по крайней мере в первом приближении).

В-третьих, объемы выборок для разных типов текстов должны быть различными для того, чтобы закон распределения разных слов по частоте их употребления в текстах проявился в одинаковой степени. Для выполнения этого условия можно, например, потребовать, чтобы крайние справа точки имели примерно одинаковые значения ординат, которые рассчитываются по формуле (см. ниже). Чем меньше эта ордината, тем надежнее может быть установлен теоретический закон распределения и точнее оценены его параметры.

Величина, определяемая указанной формулой, а также ее отношение к наибольшей высоте кривой распределения, уменьшается с ростом объема выборки и, следовательно, характеризует ее размер.

Найдем отношение ординаты крайней справа точки к наибольшей ординате кривой распределения и запишем неравенство

$$\frac{Y}{2X(rp_r)_{\max}} \le \delta,$$

где δ – некоторое наперед заданное число, например, 0.1–0.2. Отсюда найдем минимально необходимое отношение объема выборки к объему словаря

$$\frac{X}{Y} \ge \frac{1}{2\delta(rp_r)_{\max}}.$$

В-четвертых, статистическую кривую распределения $rp_r = f(\ln r)$ можно использовать для расчета минимально необходимого объема выборки при построении достоверного словаря заданного объема. Пусть достоверная

частота m_r и наибольший ранг r, т.е. объем словаря, заданы. Тогда из равенства

$$rp_r = r\frac{m_r}{X} \qquad (6.22)$$

находим необходимый объем выборки X

$$X = \frac{r \cdot m_r}{rp_r}.$$

Произведение rp_r, входящее в формулу (6.22), берется из графика зависимости $rp_r = f(\ln r)$. Например, при *r = 2500 (lnr = 7.8)*, *r = 5000 (lnr =8.5)*, *r = 10000 (lnr = 9.2)* из графика находим соответственно rp_r*= 0.123; 0.105; 0.08*. Пусть далее m_r*= 30*. Тогда необходимый объем выборки будет равен: в первом случае *X = 2500*30/0.123 = 610000*; в остальных случаях соответственно *1430000; 3750000*.

Как видно из приведенных расчетов, между объемом достоверного словаря и необходимым для его построения объемом выборки нет линейной зависимости: объем выборки растет значительно быстрее объема достоверного словаря.

В случае однородной совокупности лексических единиц (слов, словосочетаний, терминов, дескрипторов) их ранговые распределения хорошо описываются второй и третьей системами непрерывных распределений [14], которые заданы тремя обобщенными плотностями. Для вычисления типа выравнивающей кривой и оценок ее параметров статистическое распределение необходимо привести к форме плотности *p(t)* либо *p(x)* и воспользоваться соответствующей компьютерной программой.

Характерные точки кривых распределения могут быть использованы как естественные границы различных зон лексических единиц (служебных слов, общеупотребительной лексики, отраслевой, межотраслевой).

В итоге можно сделать вывод, что обобщенные распределения являются универсальными законами распределения не только теории вероятностей и математической статистики, но и информатики, информетрии, наукометрии, квалиметрии, эконометрики, математической лингвистики, экономики и других областей знания. При использовании обобщенных распределений исчезают ранее существовавшие барьеры на пути к новому знанию. Например, **для нахождения наилучшей аппроксимирующей кривой 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.

6.6. Методы вычисления границ ядра и зон рассеяния публикаций

В научных исследованиях широко используются статистические ранговые распределения различных объектов. Это журналы, разные наименования книг, представленные в списке по убыванию частоты обращения к ним; лексические единицы частотного словаря; ключевые слова, упорядоченные по убыванию частоты их использования при индексировании документов, и многие другие объекты исследования. Порядковый номер журнала, книги, слова в этом списке называется рангом. На базе статистических ранговых распределений можно решать множество практических задач, в том числе вычислять границы ядра и зон рассеяния публикаций. К сожалению, анализ научных работ с использованием ранговых распределений показывает, что некоторые авторы не могут вычислить теоретический закон и, следовательно, границ ядра и зон рассеяния. Это можно осуществить лишь при использовании теории обобщенных распределений.

В 1948 г. С. Бредфорд окончательно сформулировал свой закон рассеяния журнальных публикаций, который заключается в следующем (см. формулировку в п. 6.3).

Однако из этой формулировки неясно, как по статистическому ранговому распределению вычислить границы ядра и зон рассеяния, поскольку нет никаких формул для их вычисления. Неизвестно также, какое теоретическое ранговое распределение принято за основу закона рассеяния, какие точки на графике рангового распределения приняты в качестве таких границ, сколько может быть зон рассеяния, как вычисляется величина n. В формулировке С. Бредфорда лишь указывается, что можно выделить ядро журналов и несколько зон, при этом предполагается, что число статей в каждой зоне такое же, как и в ядре. Однако это предположение не согласуется с фактическими данными и не обосновывается теоретически.

Метод подбора

Приняв формулировку С. Бредфорда за основу, исследователи вынуждены по своему усмотрению устанавливать объем ядра журналов, а количество зон рассеяния и их размер получаются на основании фактических данных из условия, что каждая зона содержит столько же статей, что и ядро. При таком подходе количество зон рассеяния для одного и того же статистического рангового распределения у разных исследователей может сильно колебаться.

Главное в законе С. Бредфорда из того, что не подлежит сомнению – это расположение журналов по убыванию числа опубликованных в них статей по заданному предмету, т.е. ранжирование журналов. Ранговые распределения с наибольшей полнотой отражают суть рассеяния публикаций. Но многие исследователи обходят этот главный вопрос стороной. Вместо изучения

свойств статистических ранговых распределений и нахождения теоретического закона распределения с такими же свойствами они пытаются выделить ядро и зоны рассеяния, используя только словесную формулировку С. Бредфорда 1948 года!

К сожалению, в литературных источниках не приводятся статистические ранговые распределения разных наименований книг, упорядоченных по убыванию частоты их выдач, недостаточно активно проводится работа по составлению частотных списков журналов, упорядоченных по различным признакам. Но для изучения свойств ранговых распределений можно обратиться к математической лингвистике. В этой области знания накоплено большое количество частотных словарей, в которых лексические единицы расположены в порядке убывания (точнее, невозрастания) частоты их употребления в текстах. Выработана удобная для анализа таблица частот слов, в которой приводятся необходимые сведения о частотном словаре: ранг или интервал рангов слов, их абсолютная частота, количество слов с данной частотой, накопленная абсолютная частота, относительная частота, накопленная относительная частота слов. Такая таблица приводится, как правило, в конце частотного словаря. Она позволяет представить в компактной форме частотную структуру словаря любого объема.

Для аппроксимации статистических распределений, в том числе ранговых созданы универсальные математические модели – обобщенные распределения, разработаны методы вычисления законов распределения по статистическим данным и оценок их параметров [15, 21]. Предложена новая форма представления ранговых распределений, а на ее основе – критерии однородности и достаточности объема выборки. Имеются компьютерные программы для построения частотных словарей, в том числе свободно распространяемые.

Это значит, что многие задачи по обработке статистических ранговых распределений в лингвистике, информатике, библиотечном деле давно решены. Но, к сожалению, накопленные в этих областях знания результаты научных исследований практически не используются. Причина заключается в том, что для их использования исследователю необходимо затратить немалый труд и время на изучение и практическое применение этих результатов в своей научной работе. Значительно проще традиционно выдвигать гипотезы и проверять их по различным критериям согласия, но таким путем нельзя получить новое знание.

Упомянутый выше метод подбора размеров ядра и зон рассеяния – это тот же метод выдвижения гипотез (фактически метод угадывания), но без проверки результатов по критериям согласия. Такой метод в научных исследованиях неприемлем.

Итак, задано ранговое распределение тех же журналов. Если найти теоретическое распределение, которое достаточно точно описывает статистическое ранговое распределение, то оно позволит дать математически точную формулировку закона рассеяния в смысле С. Бредфорда, т. е. этот закон должен следовать из рангового распределения как частный случай. Наиболее же общим, универсальным законом рассеяния может быть только теоретическое ранговое распределение, поскольку закон распределения является наиболее полной характеристикой любой случайной величины.

Как правило, все авторы при определении границ ядра и зон рассеяния пытаются обрабатывать статистические ранговые распределения без предварительного решения общей задачи – нахождения закона распределения. Но в этом случае им приходится для аппроксимации каждого статистического распределения подбирать различные формулы [11] вместо использования одного

теоретического закона, но с разными значениями параметров. Кроме того, для подбора теоретической кривой используется неудачная форма представления статистических ранговых распределений в системе координат «логарифм ранга – логарифм частоты», которая несет слишком мало информации о ранговом распределении и более того, не имеет вероятностного смысла.

Таким образом, проблема рассеяния публикаций состоит в решении общей задачи – нахождении такой универсальной формулы или системы формул, которые способны с высокой точностью аппроксимировать все многообразие статистических ранговых распределений. Поскольку «…рассеяние научной информации является краеугольным камнем всей научно-информационной деятельности, а изучение этого свойства научной информации – важнейшей проблемой информатики» [9, с. 93], то эту проблему необходимо разрешать весьма серьезными средствами. Такими средствами являются обобщенные распределения автора [14, 15, 17–21].

Рассмотрим первую и вторую системы непрерывных распределений, каждая из которых задана тремя обобщенными плотностями. Запишем первые плотности этих систем:

$$p(x) = Ne^{k\beta x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1}, \tag{6.23}$$

$$\mathrm{p(t)} = \mathrm{Nt}^{\mathrm{k}\beta\text{-}1}\left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}-1}. \tag{6.24}$$

Здесь α, β, k, u – параметры распределения. Они вычисляются по статистическому распределению. N – нормирующий множитель, который выражается через параметры распределения при условии, что площадь под кривыми распределения равна единице.

Первая плотность обладает тем замечательным свойством, что при значениях параметра формы $u \leq 1/2$ она имеет моду x_C, т. е. такое значение случайной величи-

ны X, при котором плотность p(x) максимальна, и две точки перегиба – x_A, x_B, расположенные на равных расстояниях по обе стороны от моды. Это такие точки на графике плотности распределения, которые отделяют выпуклую часть кривой от вогнутой или вогнутую часть от выпуклой. При 1/2<u<1 имеются лишь две характерные точки – x_A и x_C (для распределений с левосторонней асимметрией). При u ≥ 1 характерных точек не существует. Распределение (6.23) может быть задано на всей числовой оси, т. е. – ∞ <x <∞.

Распределение (6.24) задано на положительной полуоси t> 0. График плотности p(t) может принимать различные формы. При определенных значениях параметров формы k, β, u график плотности p(t) может иметь вид убывающей кривой распределения, а это значит, что плотность p(t) описывает не только одновершинные статистические распределения, но и ранговые (убывающие).

Для вычисления закона распределения случайной величины по статистическому распределению, в том числе ранговому, автором разработаны два метода – универсальный метод моментов и общий устойчивый метод [21]. С целью упрощения расчетов и извлечения новой информации из рангового распределения оно приводится к форме одновершинного распределения [14, 18]. Это достигается путем приведения второй плотности к форме первой. Для этого умножим левую и правую части плотности p(t) на t, а величину t^{β} представим в виде $e^{\beta \ln t}$. Тогда эта плотность примет вид

$$tp(t) = Ne^{k\beta \ln t}\left(1 - \alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (6.25)$$

Сравнивая (6.23) и (6.25), можем записать:

$$tp(t) = p(x),\ \ln t = x. \qquad (6.26)$$

Таким образом, ранговое распределение (p(t), приведенное к форме tp(t) = f(lnt), представляет собой распределение (p(x) и обладает всеми его свойствами: оно имеет моду $\ln t_C$ и две точки перегиба $\ln t_A$ и $\ln t_B$. В этом заключается новая информация о ранговом распределении. Примем абсциссы точек А, С, В обобщенных распределений в качестве границ ядра и зон рассеяния различных объектов.

Поскольку точки $\ln t_A$ и $\ln t_B$ расположены на равных расстояниях от моды $\ln t_C$, то можем записать равенство $\ln t_C - \ln t_A = \ln t_B - \ln t_C$, откуда имеем $\ln(t_C / t_A) = \ln(t_B / t_C)$, или

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n \quad . \tag{6.27}$$

Формула (6.27) может служить уточненной формулировкой закона рассеяния публикаций в смысле С. Бредфорда, хотя она представляет собой новую формулировку закона рассеяния публикаций.

Из (6.27) можно записать еще две формулы:

$$t_A : t_C : t_B = t_A(1 : n : n^2), \tag{6.28}$$

$$t_A : t_I : t_{II} = t_A[1 : (n-1) : n(n-1)]. \tag{6.29}$$

Здесь t_A, t_C, t_B – количество журналов от начала частотного списка до точек А, С, В; t_1, t_{II} – количество журналов в первой и второй зонах рассеяния. При этом $t_1 = t_C - t_A$, $t_{II} = t_B - t_C$. С учетом этих равенств получена формула (6.29) из формулы (6.28).

Все три формулировки, заданные формулами (6.27) – (6.29), отличаются от формулировки С. Бредфорда и уточняют ее. Однако приведенные формулы (как и закон Бредфорда) не являются законом рассеяния, так как не дают полной информации о нем. Они устанавливают лишь общие соотношения между абсциссами трех характерных точек – моды и точек перегиба. Но для вычисления координат этих точек требуется знание теоре-

тического рангового закона распределения. Только при известных оценках параметров этого распределения можно получить нужную информацию – вычислить границы ядра журналов и зон рассеяния, долю статей в ядре и зонах рассеяния, а также долю статей для любого числа журналов от начала частотного списка. Она выражается через функцию распределения. В формулах же (6.27) – (6.29) функция распределения не задействована.

Таким образом, наиболее полную информацию дает закон рангового распределения. Это значит, что обобщенная плотность (p(t), а точнее, вторая система непрерывных распределений, заданная тремя обобщенными плотностями, является универсальным законом рассеяния публикаций. Здесь уместно отметить, что первая система непрерывных распределений, включающая плотность (p(x), является универсальным законом старения публикаций [20].

Некоторые исследователи утверждают, что закон С. Бредфорда – это другая форма представления закона Дж. Ципфа – $p_r = k/r$. Эта формула была предложена им для описания ранговых распределений слов частотного словаря. Здесь p_r – относительная частота слова с рангом r, k – параметр. Для закона Дж. Ципфа произведение относительной частоты слова на ранг равно постоянной величине k, что на графике в системе координат $rp_r = f(\ln r)$ изображается горизонтальной прямой, на которой нет никаких характерных точек. Это убедительно свидетельствует о том, что законом Дж. Ципфа нельзя аппроксимировать статистические ранговые распределения, поскольку в данной системе координат в случае однородной выборки они имеют вид одновершинной кривой с двумя точками перегиба. Такие кривые с высокой точностью описываются второй системой непрерывных распределений. Здесь следует отметить, что закон Дж. Ципфа, а в более общей формулировке, с двумя дополнительными параметрами – закон Эсту-Ципфа-

Мандельброта – является частным случаем обобщенного распределения (p(t) при u=1, β < 0. Но при этом условии кривая распределения $rp_r = f(\ln r)$ не имеет характерных точек. Поэтому никоим образом нельзя из них получить закон рассеяния, нельзя их использовать для аппроксимации статистических ранговых распределений, тем более в серьезных научных исследованиях. Наилучшее теоретическое распределение для аппроксимации статистического рангового распределения должно быть вычислено по второй системе непрерывных распределений [17].

Рассмотрим далее закон рассеяния С.Бредфорда (при r=t, $t_Я = t_A$)

$$t_Я : t_I : t_{II} = t_Я(1 : n : n^2). \qquad (6.30)$$

Сравнивая эту формулу с формулами (6.28) и (6.29), видим, что закон рассеяния С.Бредфорда является неправильной комбинацией двух правильных формул: левая часть соответствует формуле (6.29), а правая – формуле (6.28).

Формулу (6.29) можно представить в другом виде – с учетом трех характерных точек, о которых С.Бредфорд в своей формулировке закона рассеяния не упоминал. Из (6.29) имеем

$$t_Я : (t_C - t_Я) : (t_B - t_C) = t_Я(1 : n : n^2).$$

Вынесем в левой части этого равенства величину $t_Я$ за скобки

$$t_Я[1 : (t_C - t_Я)/t_Я : (t_B - t_C)/t_Я] = t_Я(1 : n : n^2).$$

Теперь можем записать два равенства: $(t_C - t_Я)/t_Я = n$, $(t_B - t_C)/t_Я = n^2$, откуда находим, что отношение $t_C : t_Я = n + 1$, а отношение $t_B : t_Я = n^2 + n + 1 = (n+1)^2 - n$. В результате имеем формулу

$$t_Я : t_C : t_B = t_Я[1 : (n+1) : (n^2 + n + 1)]. \qquad (6.31)$$

В нашей точной формуле (6.28) первое отношение равно n, второе n^2. Поскольку эти отношения в формуле (6.29) не соблюдаются, то по закону С. Бредфорда точки перегиба на кривой рангового распределения $tp(t) = f(\ln t)$ должны располагаться на разных расстояниях от моды $\ln t_C$, а это противоречит свойствам статистических и теоретических ранговых распределений. Другое противоречие закона С. Бредфорда состоит в утверждении, что каждая зона рассеяния содержит такое же число статей, как и в ядре. Естественно, что этим двум требованиям не может удовлетворить ни одно теоретическое распределение. Поэтому все попытки усовершенствовать закон рассеяния С. Бредфорда при сохранении присущих ему противоречий оказались неудачными. И это закономерно, потому что предпринимались попытки решить частную задачу без предварительного решения общей задачи – нахождения такого теоретического рангового распределения, которое позволило бы с высокой точностью аппроксимировать широкое разнообразие статистических ранговых распределений.

Оценим погрешность формулы (6.29). Пусть величина n=5. Вычислим отношения $t_C : t_Я$ и $t_B : t_Я$. В первом случае оно равно n+1=6 , а во втором $n^2 + n + 1 = 31$. По точной формуле (6.28) имеем соответственно 5 и 25. Погрешность вычисления абсциссы точки С составила 20%, а точки В – 24%. Размер ядра в обоих случаях принят одинаковым, так как по закону С.Бредфорда его вычислить нельзя. Следует отметить, что с ростом величины n эта погрешность уменьшается.

Из полученных результатов следует, что закон рассеяния С.Бредфорда, представленный в виде формул (6.28) и (6.29), относительно близок к точным формулам (6.26) и (6.27), хотя при его формулировке им не были использованы теоретические ранговые распределения. Утверждение же С. Бредфорда о том, что число статей в

ядре и зонах рассеяния одинаково, не соответствует действительности и приводит в заблуждение некоторых исследователей.

Однако огромная заслуга С. Бредфорда заключается в том, что он первым обратил внимание на явление рассеяния публикаций и побудил многих исследователей к углубленному изучению этого явления.

Вопрос этот оказался очень сложным. Универсальный закон рассеяния был найден автором лишь после разработки теории обобщенных распределений. Выше отмечалось, что таким законом является вторая система непрерывных распределений. Из обобщенной плотности) p(t) выводятся формулы для вычисления границ ядра и зон рассеяния.

Мода t_C находится из условия $\mathrm{dtp(t)/dlnt} = 0$ и в общем случае для распределений I-V типов равна [19]

$$t_C = \left(\frac{k}{\alpha(1+ku-u)} \right)^{1/\beta}. \tag{6.32}$$

Величина n задается формулой

$$n = \left[1 + \frac{1-u \mp \sqrt{[4k(1+ku-u)+(1-u)](1-u)}}{2k(1+ku-u)} \right]^{1/\beta}. \tag{6.33}$$

Абсциссы точек перегиба вычисляются по формулам:

$$\mathrm{t_A} = t_C / n;\ t_B = t_C \cdot n. \tag{6.34}$$

Доли статей в каждой зоне и для любого другого интервала рангов вычисляются с помощью функции распределения или по статистическому ранговому распределению.

Из формул (6.11), (6.12) следует, что при заданных значениях параметров формы k, u с уменьшением параметра β величина $\mathrm{n} = \mathrm{t_C} / t_A = t_B / t_C$ растет. Это значит, что кривая распределения $\mathrm{tp}(t) = f(\ln t)$ становится более широкой и пологой.

Методы вычисления аппроксимирующих четырехпараметрических распределений изложены в ряде работ [24, 25], но эти методы рассчитаны на подготовленного исследователя.

Для нахождения границ ядра и зон рассеяния по статистическому ранговому распределению без предварительного вычисления теоретического закона можно предложить простой графический метод [25], который следует из анализа свойств обобщенных распределений (6.1) и (6.2). Суть этого метода покажем на примере.

Рассмотрим для примера закон Вейбулла, функция и плотность распределения которого заданы формулами

$$F(t) = 1 - e^{-\alpha t^{\beta}}, \tag{6.35}$$

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}. \tag{6.36}$$

При значениях параметра $\beta \leq 1$ плотность (6.36) может с высокой точностью описывать некоторые статистические ранговые распределения.

Чтобы получить из рангового распределения полезную информацию, преобразуем плотность (6.36) к форме tp(t) = f(lnt) [25]:

$$tp(t) = \alpha\beta t^{\beta} e^{-\alpha t^{\beta}} = \alpha\beta e^{\beta \ln t} e^{-\alpha e^{\beta \ln t}}. \tag{6.37}$$

С учетом равенств tp(t) = p(x); lnt = x плотность (6.37) примет вид

$$p(x) = \alpha\beta e^{\beta x} e^{-\alpha e^{\beta x}}. \tag{6.38}$$

Полученная формула представляет собой частный случай плотности (6.25) при u→0, k=1. Плотность (6.38) дает новую информацию о ранговом распределении. График этой плотности, т.е. кривая распределения содержит три характерные точки – моду С и две точки перегиба А и В, расположенные на равных расстояниях по обе стороны от моды. Найдем абсциссы этих точек.

Продифференцируем плотность (6.38) по x и приравняем первую производную нулю. Из полученного уравнения найдем выражение для моды

$$x_C = \frac{1}{\beta}\ln\frac{1}{\alpha}. \qquad (6.39)$$

В точках перегиба вторая производная равна нулю. Из этого условия для плотности (6.38) найдем

$$x_A = x_C - \frac{\ln n}{\beta}, \quad x_B = x_C + \frac{\ln n}{\beta},$$

где

$$n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \qquad (6.40)$$

С учетом (6.40) последние две формулы примут более простой вид

$$x_A = x_C - \ln n; \qquad (6.41)$$

$$x_B = x_C + \ln n. \qquad (6.42)$$

Переходя к распределению Вейбулла, из формул (6.19), (6.20) с учетом равенства x=lnt найдем

$$\ln t_C - \ln t_A = \ln t_B - \ln t_C = \ln n,$$

откуда $\ln(t_C / t_A) = \ln(t_B / t_C) = \ln n$, или

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \qquad (6.43)$$

Здесь

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}; \quad t_A = \frac{t_C}{n}; \quad t_B = t_C \cdot n. \qquad (6.44)$$

Формула (6.44), полученная на базе закона Вейбулла, совпадает с аналогичной формулой (6.5), полученной на

базе четырехпараметрического распределения (6.24). Она остается также справедливой для любого частного случая распределения (6.24) при значениях параметра формы u ≤ 1/2. Закон Вейбулла является частным случаем распределения (6.24) при u→0, k=1. Во многих случаях он с высокой точностью описывает статистические ранговые распределения. Но универсальным законом остается обобщенная плотность (6.24), которая позволяет вычислять и координаты характерных точек, и функцию распределения [17, 20] практически для любого статистического рангового распределения. В некоторых случаях статистические ранговые распределения с высокой точностью описываются дополнительными плотностями второй системы непрерывных распределений.

Предположим для определенности, что ранговое распределение задано законом Вейбулла с параметрами α=0,1; β=0,5. Приведем его к плотности p(x), которая представлена формулой (6.23). Тогда для этой плотности по формулам (6.17)–(6.20) найдем: n=6,8541; $x_C = 4{,}6052$; $x_A = 2{,}6804$; $x_B = 6{,}53$.

Рассчитаем по формуле (6.38) значения плотности p(x) с интервалом Δx=0,5 и сведем результаты в таблицу 6.1.

Построим график плотности p(x), т. е. кривую распределения (рис. 6.1а).

На кривой распределения абсциссу точки C легко найти графически путем проведения горизонтальной касательной к кривой (см. рис. 6.1а).

Таблица 6.2

Значения плотности и тангенса угла наклона касательной к кривой в серединах интервалов

x	p(x)	dp(x)/dx	x	p(x)	dp(x)/dx
-2	0,01773	0,008539	4	0,176464	0,023037
-1,5	0,022529	0,010732	4,5	0,18369	0,004705
-1	0,028542	0,013405	5	0,180147	-0,01966
-0,5	0,036022	0,016609	5,5	0,163655	-0,04617
0	0,045242	0,020359	6	0,134756	-0,06795
0,5	0,056465	0,024607	6,5	0,097806	-0,07722
1	0,069906	0,02919	7	0,060369	-0,06977
1,5	0,085655	0,033761	7,5	0,030263	-0,04921
2	0,103565	0,037706	8	0,011614	-0,0259
2,5	0,123099	0,040067	8,5	0,003163	-0,00951
3	0,143144	0,039496	9	0,000554	-0,00222
3,5	0,161833	0,034352	9,5	5,52E-05	-0,00029

Для нахождения абсцисс точек перегиба воспользуемся тем свойством кривой распределения, что в точках A и B первая производная принимает экстремальные значения: в точке A она имеет максимум, а в точке B – минимум. Вычислим тангенс угла наклона отрезков кривой к горизонтальной оси на всех интервалах как отношение разности между значениями плотности p(x) на границах интервала к ширине интервала Δx. Другими словами, найдем приближенные значения первой производной в серединах интервалов (в таблице приведены расчетные значения производной dp(x)/dx) и построим график (см. рис. 6.1b).

Первая производная dp(x)/dx имеет максимум в точке x_A и минимум в точке x_B. Эти точки легко определить

путем проведения горизонтальных касательных к кривой на рис.6.1b.

Из построенного графика можно приближенно найти абсциссы трех характерных точек: $x_A = 2{,}7;\ x_C = 4{,}6;\ x_B = 6{,}5$.

Переходя к ранговому распределению, находим: $t_A = \exp(x_A) \approx 15$; $t_C = \exp(x_C) \approx 99$; $t_B = \exp(x_B) \approx 665$. Точные значения для распределения Вейбулла равны: $t_A = 14{,}59;\ t_C = 100;\ t_B = 685.$

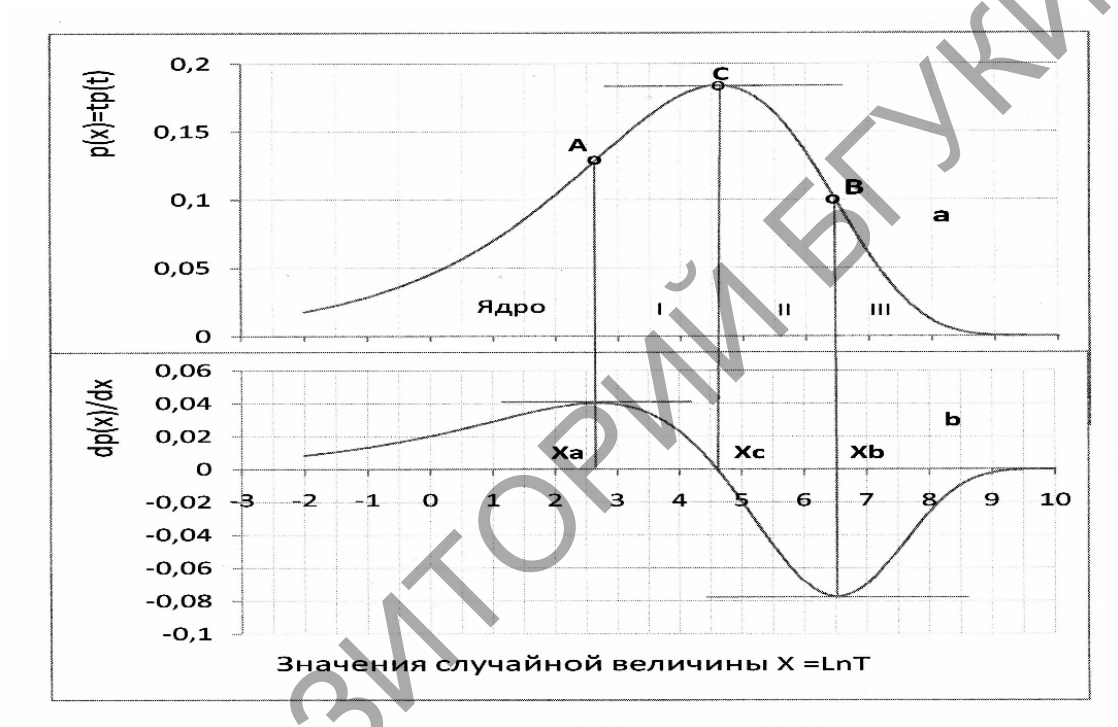


Рис. 6.1. Графики плотности распределения (6.1a) и ее первой производной (6.1b)

Таким простым методом приближенно могут быть найдены абсциссы трех характерных точек любого статистического рангового распределения без предварительного вычисления теоретического закона распределения, но с использованием его свойств.

Значения функции распределения F(t) при любом заданном значении ранга t, в том числе в трех характерных точках могут быть вычислены по статистическому ранговому распределению.

Используя графический метод, разные исследователи получат близкие результаты по определению границ ядра и зон рассеяния для одного и того же статистического рангового распределения.

Здесь необходимо отметить, что представленная на рис. 6.1a теоретическая кривая распределения плавно возрастает до максимального значения и затем плавно убывает. Поэтому вычисление тангенса угла наклона отрезков кривой на всех интервалах не вызывает затруднений. Аналогичная статистическая кривая распределения tp(t) = f(lnt) имеет многочисленные всплески и впадины, что затрудняет построение рис. 6.1b. Поэтому предварительно ее необходимо сгладить, например, с помощью лекала.

Отметим, что на рис. 6.1a горизонтальная касательная к кривой распределения tp(t) = f(lnt) в точке C представляет собой закон Дж. Ципфа. Отсюда следует, что этим законом невозможно описать никакое ранговое распределение.

Метод наименьших квадратов.

В некоторых случаях статистическое ранговое распределение может с высокой точностью описываться законом Вейбулла, функция распределения и плотность вероятностей которого заданы формулами (6.35) и (6.36). Этот закон впервые использовал Г. Г. Белоногов для описания рангового распределения слов частотного словаря [1]. Поскольку этот закон весьма простой, его целесообразно проверять в первую очередь при отыскании подходящего рангового распределения. Для такой проверки функцию распределения необходимо преобразовать к линейному виду

$$\ln\ln(1/(1-F(t))) = \ln\alpha + \beta\ln t. \qquad (6.45)$$

Введем обозначения:

$$Y = \ln\ln(1/(1 - F(t))),\ X = \ln t. \tag{6.46}$$

Тогда последнее уравнение запишется в виде

$$Y=\ln\alpha + \beta X. \tag{6.45′}$$

Для проверки применимости закона Вейбулла необходимо по статистической функции распределения вычислить значения X, Y по формулам (6.46) и построить график зависимости Y=f(X). Если эмпирические точки расположатся вдоль прямой (6.45г), то далее по методу наименьших квадратов следует вычислить оценки параметров α и β этой прямой:

$$\beta = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \left(\overline{X}\right)^2},\quad \alpha = \exp\left(\overline{Y} - \beta\overline{X}\right). \tag{6.47}$$

Для оценки тесноты линейной связи между переменными Y, X вычисляется выборочный коэффициент корреляции

$$R_{y/x} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y}, \tag{6.48}$$

где средние квадратические отклонения σ_x, σ_y равны:

$$\sigma_x = \sqrt{\overline{X^2} - (\overline{X})^2},\ \sigma_y = \sqrt{\overline{Y^2} - (\overline{Y})^2}.$$

При этом

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i,\quad \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i,\quad \overline{X^2} = \frac{1}{N}\sum_{i=1}^{N} X_i^2,\quad \overline{Y^2} = \frac{1}{N}\sum_{i=1}^{N} Y_i^2,$$

где N – количество значений случайных величин X, Y.

Абсциссы точек A, C, B для закона Вейбулла вычисляются по формулам (6.18), (6.22). Значения функции распределения в этих точках при любых значениях параметров α и β соответственно равны:

$$F(t_A) = 0{,}31748;\ F(t_C) = 0{,}63212;\ F(t_B) = 0{,}92705. \tag{6.49}$$

Отметим, что статистические данные расположатся вдоль прямой лишь в случае однородной выборки, для которой справедлив закон Вейбулла. Однако если попытаться описать этим законом ранговое распределение слов частотного словаря, то окажется, что первые 50–100 наиболее частых слов не подчиняются закону Вейбулла. Это в основном служебные слова. Они составляют неоднородную часть выборки. Поэтому для более точного описания таких ранговых распределений можно предварительно удалить первые 50–100 слов с последующим пересчетом рангов и относительных частот слов, получив таким образом однородную выборку. Если же имеется необходимость аппроксимировать ранговое распределение всех слов частотного словаря, включая служебные, то следует ввести дополнительный параметр в теоретическое распределение. Исследования автора в свое время привели к выводу, что закон Вейбулла с учетом третьего параметра (обозначим его δ) можно представить в следующем виде [13]

$$\mathrm{F(t)} = 1 - \mathrm{e}^{-\alpha[(\mathrm{t}+1)^{\beta} - e^{-\delta t}]}, \tag{6.50}$$

$$\mathrm{p(t)} = \frac{\alpha[\beta(\mathrm{t}+1)^{\beta-1} + \delta e^{-\delta t}]}{e^{\alpha[(t+1)^{\beta} - e^{-\delta t}]}}. \tag{6.51}$$

Параметры α, β могут быть вычислены по методу наименьших квадратов по формулам (7.46)–(7.47) для рангов слов частотного словаря от 50–100 до рангов слов с частотой 2–3. Далее при известных оценках этих параметров вычисляется дополнительный параметр δ по формуле, которая следует из функции распределения (6.50):

$$\delta = -\frac{1}{\mathrm{t}}\left[\ln\left((t+1)^{\beta} - \frac{1}{\alpha}\ln\frac{1}{1-F(t)}\right)\right]. \tag{6.52}$$

Его можно вычислить один раз при заданной относительной частоте самого частого слова, которая равна функции распределения F(t=1).

Двухпараметрическим законом Вейбулла хорошо описываются некоторые ранговые распределения журналов, терминов, ключевых слов, образующих статистически однородные выборки, а трехпараметрическим – некоторые неоднородные выборки.

Рассмотрим для примера ранговое распределение слов «Частотного словаря современного русского языка» [8]. Он построен на выборке огромного размера – 135 млн. словоупотреблений. Количество разных лексем (в источнике – лемм) составило 739930 единиц. Из них количество лемм с частотой 2 и более раза составило 360755, с частотой 3 и более раза – 268106. В указанном источнике приводится частотный список первых 20000 лемм, что позволяет вычислить накопленные относительные частоты, т.е. функцию распределения для всех рангов от 1 до 20000. При известном количестве лемм с частотами употребления 1 и 2 раза (соответственно 379175 = 739930 – 360755 и 92649 = 360755 – 268106) можно дополнительно вычислить два значения функции распределения:

F(360755)=1–379175/135000000=0,9971913;

F(268106)=1–(379175+92649*2)/135000000=0,9958187.

Здесь из суммарной доли употреблений всех лемм, равной единице, вычитается доля употреблений лемм с частотой один раз – в первом случае и один и два раза – во втором.

Составим на базе Частотного словаря таблицу 2. В ней приведены отдельные ранги слов ($R \geq 80$) и соответствующие этим рангам значения функции распределения (см. первые два столбца). Вычислим далее по методу наименьших квадратов оценки параметров закона Вейбулла, а также коэффициент корреляции. Параметр

β оказался равным 0,309427, параметр α=0,111757. Коэффициент корреляции Ry/x=0,999789. Это значит, что эмпирическая зависимость оказалась близкой к теоретической прямой (6.45), которая представлена на рис. 6.2.

При известных оценках параметров α, β и функции распределения F(t=1)=0,035802 по формуле (6.30) найдем значение третьего параметра, который необходим для более точного описания наиболее частых слов: δ=0,091. Вычислим далее значения функции распределения по трехпараметрическому закону Вейбулла (см. табл. 6.3) и построим в полулогарифмическом масштабе график функции распределения (см. рис. 6.3) с учетом служебных слов. Отдельными точками показана эмпирическая функция распределения, сплошной линией – теоретическая.

Таблица 6.3

Расчет параметров закона Вейбулла по статистическому распределению

R	Fr	LnR	Ln(Ln(1/(1-Fr)))					
Ранги слов	Функция распределения.	X	Yэмп.	XY	X^2	Y^2	Yрасч.	Fрасч.
80	0,354324	4,382027	-0,82678	-3,62295	19,20216	0,683558	-0,83551	0,351864
100	0,374545	4,60517	-0,75656	-3,48411	21,20759	0,57239	-0,76646	0,371648
150	0,411675	5,010635	-0,63398	-3,17665	25,10647	0,401932	-0,641	0,409488
250	0,458992	5,521461	-0,48724	-2,69026	30,48653	0,2374	-0,48294	0,460423
400	0,506106	5,991465	-0,34894	-2,09067	35,89765	0,12176	-0,3375	0,510098
666	0,561671	6,50129	-0,19263	-1,25236	42,26677	0,037107	-0,17975	0,566333
1097	0,620558	7,000334	-0,03144	-0,22006	49,00468	0,000988	-0,02533	0,622802
1809	0,681615	7,500529	0,134963	1,012291	56,25794	0,018215	0,129442	0,679603
3000	0,740589	8,006368	0,299617	2,398842	64,10192	0,08977	0,285962	0,735798
4900	0,792717	8,49699	0,453411	3,852626	72,19885	0,205581	0,437774	0,787594
8100	0,838483	8,999619	0,600563	5,404838	80,99315	0,360676	0,593301	0,836338
13360	0,875102	9,50002	0,732492	6,958688	90,25039	0,536544	0,748139	0,879133
20000	0,898718	9,903488	0,828485	8,204889	98,07907	0,686387	0,872983	0,90874
268106	0,995819	12,49914	1,700595	21,25597	156,2284	2,892023	1,676148	0,995228
360755	0,997191	12,79595	1,770694	22,65771	163,7364	3,135356	1,767992	0,997146

R	Fr	LnR	Ln(Ln(1/(1-Fr)))					
Summa		116,714	3,243251	55,2088	1005,018	9,979688		
Srednee		7,78096	0,216217	3,680587	67,0012	0,665313		
Beta=	0,309427	Sx=	2,541215					
Alfa=	0,111757	Sy=	0,786488					
Ry/x=	0,999789	д=	0,091					

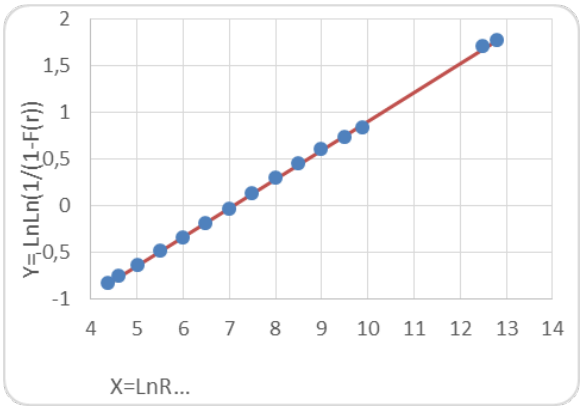


Рис. 6.2. Прямая Вейбулла

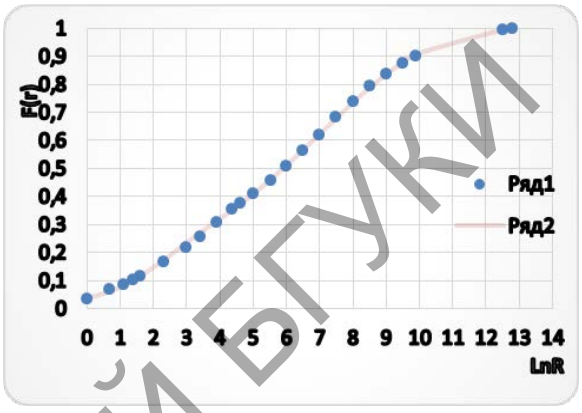


Рис. 6.3. Функция распределения Вейбулла

Таблица 6.4

Эмпирическая и теоретическая функции распределения наиболее частых слов

Ранги слов	Эмпирическая функция распределения	Уточненная теоретическая функция распределения
1	0,035802	0,035798
2	0,067176	0,061847
3	0,085204	0,082922
4	0,101071	0,100782
5	0,113756	0,116317
10	0,165571	0,172804
20	0,217115	0,235534
30	0,255761	0,271023
50	0,309294	0,313454

Из таблицы 6.4 и графика функции распределения видно, что введение третьего параметра в закон Вейбулла позволило весьма точно аппроксимировать статистическое распределение первых 50–80 слов Частотного словаря. Ранговое распределение остальных слов хорошо описывается классическим законом Вейбулла с двумя параметрами.

Вычислим абсциссы трех характерных точек по формулам (6.18) и (6.22) при известных оценках параметров α и β: n=22,4287; $t_C = 1191$; $t_A = 53$; $t_B = 26704$. Отметим, что логарифмы рангов слов в точках А, С, В равны: 3,97187; 7,08221; 10,19256. Теоретическая функция распределения в этих точках равна: 0,31748; 0,63212; 0,92705.

Из этих расчетов следует, что ядро Частотного словаря составляют первые 53 слова. Они покрывают 31.7% текста. В первую зону рассеяния А–С входит 1191–53=1138 слов, которые покрывают 31.5% текста (63.2–31.7=31.5). Во вторую зону С–Входит 26704–1191=25513 слов, которые покрывают 29.5% текста (92.7–63.2=29.5). В третью зону рассеяния входит вся остальная лексика 739930–25513=714417 слов. Этот огромный словарь покрывает лишь 7.3% текста (100–92.7=7.3).

Рассмотрим еще один пример – ранговое распределение периодических изданий по химии и химической технологии [9]. Поскольку в данном случае выборка однородная, то для аппроксимации статистического рангового распределения может быть использован закон Вейбулла с двумя параметрами. Проведем необходимые расчеты по методу, изложенному выше. Результаты представлены в виде табл. 6.5 и рис. 6.4.

Таблица 6.5

Рассеяние журнальных публикаций по химии и химической технологии
(10850 журналов, 187911 статей)

К-во журн.	Доля стат.	lnt	lnln1/(1-F(t))					
t	F(t)	X	Y	XY	X^2	Y^2	F(t)рас.	Yрасч.
18	0,15	2,890372	-1,817	-5,25181	8,3542489	3,301489	0,1536	-1,7912
50	0,25	3,912023	-1,2459	-4,87399	15,303924	1,552267	0,24772	-1,25649
100	0,34	4,60517	-0,8782	-4,04426	21,207592	0,771235	0,33577	-0,89371
500	0,62	6,214608	-0,033	-0,20508	38,621354	0,001089	0,61323	-0,05138
1000	0,75	6,907755	0,3266	2,25607	47,717083	0,106668	0,7447	0,3114
2000	0,85	7,600902	0,6403	4,86686	57,773718	0,409984	0,85948	0,67418
Sum		32,13083	-3,0072	-7,25221	188,97792	6,142732		
Sredn		5,355138	-0,5012	-1,2087	31,49632	1,023789		
Beta=	0,523374	Sx=	1,67893183	tc=	551,5708	F(tc)=	0,63212	
Alfa=	0,036738	Sy=	0,87896938	ta=	87,69708	F(ta)=	0,31748	
Ry/x=	0,999705	n=	6,28949982	tb=	3469,104	F(tb)=	0,92705	

Рис. 6.4. Прямая Вейбулла – рассеяние журнальных публикаций

Они свидетельствуют о высокой точности аппроксимации законом Вейбулла статистического рангового распределения журналов, упорядоченных по убыванию опубликованных в них статей по химии и химической технологии. Коэффициент корреляции Ry/x=0,999705. Ядро образуют 88 журналов. Количество журналов до точки С, т. е. входящих в ядро и первую зону рассеяния, равно 552. В ядро и первые две зоны рассеяния, т. е. до точки В входит 3469 журналов, в которых содержится 92,705% статей от их общего количества 187911. На третью зону рассеяния приходятся все остальные журналы 10850–3469=7381, и в этих журналах содержится 100–92,705=7,295 процентов статей.

Однако, несмотря на высокую точность аппроксимации некоторых ранговых распределений законом Вейбулла, в исследованиях по информатике и математической лингвистике он применяется весьма редко. Более часто используется закон Дж. Ципфа, который вовсе нельзя применять в таких исследованиях. Принимая во внимание то обстоятельство, что оба этих закона и множество других являются частными случаями обобщенного распределения (6.2), для описания различного рода ранговых распределений следует использовать вторую систему непрерывных распределений.

При обработке статистических рядов распределения главной задачей является вычисление теоретического закона распределения. Она решается довольно просто по методам, изложенным в теории обобщенных распределений. Для аппроксимации статистических ранговых распределений используется вторая система непрерывных распределений.

На основании анализа свойств обобщенных распределений предлагаются математически точные формулировки закона рассеяния публикаций в смысле Бредфорда. Но такие формулировки, как и закон С. Бредфорда,

не могут быть приняты в качестве полноценного закона рассеяния публикаций. Универсальным законом рассеяния является вторая система непрерывных распределений, поскольку обобщенная четырехпараметрическая плотность, т.е. закон распределения наиболее полно характеризует случайную величину.

Для вычисления закона распределения и оценок его параметров по статистическому ранговому распределению используется общий устойчивый метод. При известных оценках параметров по заранее выведенным формулам вычисляются абсциссы трех характерных точек A, C, B, которые приняты автором в качестве границ ядра и зон рассеяния. Абсциссы точек C и B, вычисленные по закону С. Бредфорда и универсальному закону, различаются на 20–25 процентов при условии, что величина n = 5, а размер ядра в обоих случаях одинаков. С ростом n эта погрешность уменьшается.

Для использования аналитического метода необходимо знать хотя бы некоторые сведения из теории обобщенных распределений. Этот метод рассчитан на подготовленного исследователя.

На базе свойств ранговых распределений предложен графический метод приближенного вычисления границ ядра и зон рассеяния. Он значительно проще аналитического метода, поскольку не требует вычисления закона распределения.

В случае однородных выборок многие статистические ранговые распределения могут быть описаны законом Вейбулла (6.35), (6.36). Оценки его параметров наиболее просто вычисляются по методу наименьших квадратов. Если при этом ранговое распределение содержит неоднородную часть, например, служебные слова, то по формуле (6.52) следует дополнительно вычислить третий параметр δ при известных значениях параметров α, β и относительной частоты первого слова.

Проведенное исследование показало высокую точность аппроксимации некоторых статистических ранговых распределений законом Вейбулла. Однако для гарантированного вычисления наилучшего теоретического рангового распределения по статистическому ряду следует использовать вторую систему непрерывных распределений и общий устойчивый метод.

7. СИСТЕМЫ КРИВЫХ РОСТА. МЕТОДЫ ОЦЕНИВАНИЯ ПАРАМЕТРОВ

7.1. Вероятностная модель текста и ее исследование

Одним из наиболее эффективных методов изучения статистических закономерностей такого сложного объекта, каким является текст, написанный человеком, является метод построения моделей. Текст в первом приближении можно рассматривать как случайную последовательность словоупотреблений. В этой весьма упрощённой модели текста не учтены грамматические и семантические связи, существующие между словами. Однако, как показывают исследования, в реальном тексте эти связи не могут оказать существенного влияния на характер некоторых количественных закономерностей текста.

После выявления характера этих закономерностей (на основе исследования упрощённой модели текста) и опытной проверки полученных результатов можно будет построить более точную модель, учитывающую грамматические и семантические связи между словами реального текста, и, более того, найти для них количественную меру.

В качестве вероятностной модели текста будем рассматривать один класс случайных функций, описывающих статистическую зависимость между количеством произведенных испытаний и количеством наступивших при этом разных событий.

Пусть имеется n несовместных событий (n разных слов в некотором вероятностном словаре) $A_1, A_2, ..., A_n$, составляющих полную группу, причем, вероятности их заданы и равны $p_1, p_2, ..., p_n$. Пусть далее производятся независимые испытания, в каждом из которых может наступить любое из n разных событий, например, событие A_k с вероятностью p_k. Если произвести достаточно большое число испытаний, то отдельные события наступят более одного раза. Условимся считать новым любое из n разных событий при первом его появлении от начала испытаний. Тогда число наступивших разных событий будет равно числу новых событий.

В этой схеме испытаний нас будет интересовать математическое ожидание (среднее значение) числа наступивших разных событий $M[Y]$ при осуществлении X испытаний, т. е. математическое ожидание случайной функции $M[Y(X)]$.

Результаты испытаний можно представить на графике. Будем откладывать по оси абсцисс число произведенных испытаний X, а по оси ординат – $M[Y(X)]$. Построенные таким образом точки для наглядности можно соединить отрезками прямых. Получим выпуклую ломаную линию – графическое изображение математического ожидания случайной функции.

Эту ломаную можно аппроксимировать непрерывной плавной кривой $y = f(x)$, которую будем называть кривой роста новых событий.

Найдем математическое ожидание числа разных (новых) событий, наступающих при X испытаниях. Оно задается известной формулой В. М. Калинина [6, с. 246]

$$M[Y(X)] = \sum_{k=1}^{n} \left[1 - (1 - p_k)^X\right]. \qquad (7.1.1)$$

Математическое ожидание числа разных событий, наступающих ровно m раз при X испытаниях (частотный спектр), определится по другой формуле В. М. Калинина

$$M[Y_m(X)] = \sum_{k=1}^{n} C_X^m p_k^m (1-p_k)^{X-m} . \qquad (7.1.2)$$

В последней формуле выражение под знаком суммы представляет собой вероятность появления k-го события m раз при X испытаниях, т. е. биномиальный закон распределения.

Важной величиной, характеризующей скорость роста новых событий, является вероятность появления какого-нибудь нового события $A^н = \sum_{k=1}^{n} A_k^н$ при (X+1)-м испытании (которая равна математическому ожиданию числа новых событий, наступающих при этом испытании)

$$P(A^н, X+1) = \sum_{k=1}^{n} p_k (1-p_k)^X \qquad (7.1.3)$$

Введем еще одну величину – среднее значение вероятностей новых событий, которые могут наступить при одном X-м испытании [33,34]. Обозначим его $\overline{p}_x$

$$\overline{p}_x = \sum_{k=1}^{n} p_k^2 (1-p_k)^{X-1} \qquad (7.1.4)$$

Тогда накопленная вероятность новых событий, наступивших при X испытаниях, будет рана сумме средних вероятностей $\overline{p}_x$

$$\overline{F}(X) = \sum_{i=1}^{X} \overline{p}_i = \sum_{i=1}^{X} \sum_{k=1}^{n} p_k (1-p_k)^{i-1} = 1 - \sum_{k=1}^{n} p_k (1-p_k)^X = 1 - P(A^н, X+1). \qquad (7.1.5)$$

Если вероятности отдельных событий малы, а число испытаний достаточно большое, то вероятности p_k целесообразно аппроксимировать непрерывной плотностью $p(t)$, удовлетворяющей условию

$$\int_{k-1}^{k} p(t)dt = p_k ,$$

а формулы (7.1.1) – (7.1.5) представить в виде [14]

$$y = \int_0^n \left(1 - e^{-xp(t)}\right) dt , \qquad (7.1.6)$$

$$y_m = C_x^m \int_0^n [p(t)]^m e^{-(x-m)p(t)} dt , \qquad (7.1.7)$$

$$P(A^{n},\mathrm{x}) = \int_0^n \mathrm{p(t)e}^{-\mathrm{xp(t)}} dt = \frac{dy}{dx}, \tag{7.1.8}$$

$$\overline{p}(x) = \int_0^n [p(t)]^2 e^{-xp(t)} dt = -\frac{d^2 y}{dx^2}, \tag{7.1.9}$$

$$\overline{F}(x) = \int_0^n \overline{p}(x) dx = 1 - \frac{dy}{dx}. \tag{7.1.10}$$

Формулу (7.1.7) при $x \to \infty$ и ограниченных значениях m можно несколько упростить. Действительно,

$$C_x^m = \frac{x!}{m!(x-m)!} = \frac{x(x-1)...[x-(m-1)]}{m!} = \frac{x^m}{m!}\prod_{i=0}^{m-1}\left(1-\frac{i}{x}\right),$$

что при $x \to \infty$ дает $C_x^m \approx x^m / m!$ В то же время при $x \to \infty$ (произведение $xp(t)$ конечно) имеем

$$\lim_{x\to\infty} e^{(x-m)p(t)} = \lim_{x\to\infty} e^{x(1-m/x)p(t)} = e^{xp(t)}.$$

Следовательно,

$$y_m = \frac{1}{m!}\int_0^n [xp(t)]^m e^{-xp(t)} dt. \tag{7.1.7′}$$

Как видно из формул (7.1.8) и (7.1.9), вероятность появления нового события при x произведенных испытаниях равна значению первой производной в точке (x; y) кривой роста новых событий $y = f(x)$ (7.1.6), а средняя плотность $\overline{p}(x)$– второй производной, взятой со знаком «минус».

Обозначим далее через $\overline{p}_j$ среднее значение вероятностей новых событий, которые могут наступить j–ми от начала испытаний (j – порядковый номер нового события), а через $\overline{p}(y)$– среднюю плотность распределения вероятностей новых событий, аппроксимирующую вероятности $\overline{p}_j$. Тогда

$$\overline{p}(y) = \overline{p}(x)\frac{dy}{dx} = -\frac{d^2 y}{dx^2}\cdot\frac{dx}{dy} = -\frac{d}{dy}\left(\frac{dy}{dx}\right), \tag{7.1.11}$$

или с учетом (7.1.6) [14]

$$\overline{p}(y) = \int_0^n \frac{[p(t)]^2}{e^{xp(t)}} dt \Big/ \int_0^n \frac{p(t)}{e^{xp(t)}} dt; \qquad (7.1.11')$$

$$\overline{F}(y) = \int_0^y \overline{p}(y) dy = 1 - \frac{dy}{dx}. \qquad (7.1.12)$$

Из формул (7.1.9) – (7.1.12) видно, что закон распределения вероятностей новых событий в отличие от плотности $p(t)$ однозначно определяется кривой роста новых событий, что дает возможность по системе кривых роста строить систему непрерывных распределений.

Формулы (7.1.10) – (7.1.12) позволяют находить кривую роста новых событий по заданной функции распределения $\overline{F}(x)$ или $\overline{F}(y)$:

$$y = \int [1 - \overline{F}(x)] dx + C; \qquad (7.1.13)$$

$$x = \int \frac{dy}{1 - \overline{F}(y)} + C, \qquad (7.1.14)$$

где постоянная интегрирования С находится из условия: y=0 при x=0.

Если известны обе функции распределения, то кривая роста находится непосредственно из равенства $\overline{F}(x) = \overline{F}(y)$.

Отметим, что при больших x и малых m между зависимостями (7.1.6) и (7.1.7′) существует взаимосвязь, установленная В. М. Калининым [6, с. 247]

$$\frac{d^m y}{dx^m} = (-1)^{m+1} \frac{m!}{x^m} y_m. \qquad (7.1.15)$$

Из (7.1.15) следует, что функция $y = f(x)$ бесконечно дифференцируема. Это позволяет строить ее разложение в ряд Тэйлора. В. М. Калинин получает таким образом формулу, позволяющую восстанавливать кривую роста новых событий по заданному частотному спектру выборки x_0

$$y = y_0 - \sum_{m \geq 1} \left(1 - \frac{x}{x_0}\right)^m y_m , \qquad (7.1.16)$$

где y_0 – число разных событий в выборке объемом x_0; y_m – число m-разовых событий в выборке x_0; y – ожидаемое среднее число разных событий в произвольной выборке $x\,(x < x_0)$.

С другой стороны, если известна кривая роста новых событий $y = f(x)$, то по формуле

$$y_m = (-1)^{m+1} \frac{x^m}{m!} \frac{d^m y}{dx^m}, \qquad (7.1.17)$$

которая следует из (7.1.15), можно рассчитать частотный спектр (статистическую структуру выборки), т. е. число событий с частотой появления 0, 1, …, m раз.

Если известна средняя плотность $\bar{p}(x)$, то частотный спектр рассчитывается по формуле автора [14]

$$y_m = (-1)^m \frac{x^m}{m!} \frac{d^{m-2} \bar{p}(x)}{dx^{m-2}}, \qquad (7.1.18)$$

которая следует из (7.1.17), (7.1.13). При этом плотность $\bar{p}(x)$ должна быть убывающей функцией, не имеющей точек перегиба.

Полученные формулы позволяют по одному известному закону распределения вероятностей новых событий построить системы непрерывных распределений и кривых роста новых событий, а также систему дискретных распределений [21].

7.2. Построение систем кривых роста и непрерывных распределений новых событий

Пусть все n событий, составляющих полную группу, имеют равные вероятности $p_k = 1/n = \alpha$. Следовательно, плотность p(t), аппроксимирующая вероятности p_k, также постоянна: $p(t) = \alpha$. В этом случае формула (7.1.11′) дает

$$\overline{p}(y) = \alpha . \qquad (7.2.1)$$

Итак, один закон распределения вероятностей новых событий задан. Восстановим по этому закону кривую роста новых событий $y = f(x)$ и среднюю плотность $\overline{p}(x)$, для чего используем формулы (7.1.12), (7.1.14), (7.1.9). Функция распределения здесь равна $\overline{F}(y) = \alpha y = 1 - (1 - \alpha y)$. Тогда согласно (7.1.14) имеем $x = \frac{1}{\alpha}\ln(1 - \alpha y)$, откуда

$$y = \frac{1}{\alpha}\left(1 - \frac{1}{e^{\alpha x}}\right). \qquad (7.2.2)$$

Далее по формуле (7.1.9) находим

$$\overline{p}(x) = \frac{\alpha}{e^{\alpha x}}, \qquad (7.2.3)$$

т. е. получили второй закон распределения.

Пусть теперь средняя плотность $\overline{p}(y)$ задается формулой

$$\overline{p}(y) = \frac{\alpha}{e^{\alpha y}}. \qquad (7.2.3')$$

Проделав ту же последовательность операций, что и в первом случае, найдем

$$y = \frac{1}{\alpha}\ln(1 + \alpha x), \qquad (7.2.4)$$

$$\overline{p}(x) = \frac{\alpha}{(1 + \alpha x)^2}. \qquad (7.2.5)$$

На следующем этапе средняя плотность $\overline{p}(y)$ будет задаваться формулой

$$\overline{p}(y) = \frac{\alpha}{(1 + \alpha y)^2} \qquad (7.2.5')$$

и т. д.

Этих результатов достаточно, чтобы сделать обобщение. Оно достигается путем введения нового параметра u, при определенных значениях которого из общих

формул будут следовать рассмотренные выше частные случаи. Итак, обобщая (7.2.1), (7.2.3′), (7.2.5′), получаем

$$\overline{p}(y) = \alpha(1-\alpha u y)^{\frac{1}{u}-1}. \qquad (7.2.6)$$

Далее на основании (7.2.3), (7.2.5) можем записать

$$\overline{p}(x) = \alpha[1-\alpha(u-1)x]^{\frac{1}{u-1}-1}. \qquad (7.2.7)$$

Интегрируя плотности (7.2.6), (7.2.7), найдем функции распределения

$$\overline{F}(y) = 1-(1-\alpha u y)^{\frac{1}{u}}, \qquad (7.2.8)$$

$$\overline{F}(x) = 1-[1-\alpha(u-1)x]^{\frac{1}{u-1}}. \qquad (7.2.9)$$

Приравнивая правые части функций распределения новых событий, найдем формулу для кривой роста новых событий:

$$y = \frac{1}{\alpha u}\left[1-(1-\alpha(u-1)x)^{\frac{u}{u-1}}\right] = \frac{1}{\alpha u}\left[1-(1+\alpha(1-u)x)^{\frac{-u}{1-u}}\right]. \qquad (7.2.10)$$

Рассмотрим двухпараметрические плотности (7.2.6), (7.2.7). Будем относить распределения, а также кривые роста к I типу при u > 0, ко II типу – при $u \to 0$, к III типу – при u < 0. Следует отметить, что при переходе от плотности $\overline{p}(y)$ к плотности $\overline{p}(x)$ параметр u уменьшается на единицу.

7.3. Построение обобщенных непрерывных распределений

Рассмотрим функцию распределения (7.2.8). Пусть случайная величина Y связана со случайной величиной T зависимостью $Y = T^{\beta}$. Тогда функция распределения случайной величины T на основании (7.2.8) будет иметь вид

$$F(t) = 1-\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}}. \qquad (7.3.1)$$

Дифференцируя ее по t, найдем трехпараметрическую плотность распределения случайной величины T

$$p(t) = \alpha\beta t^{\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}. \qquad (7.3.2)$$

Структура последней формулы позволяет еще более расширить семейство непрерывных распределений за счет введения дополнительного параметра k

$$p(t) = Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}. \qquad (7.3.3)$$

Имея четырехпараметрическую плотность $p(t)$, на ее базе можно получать другие плотности как функции случайного аргумента. Например, при $T = e^{X}$ плотность $p(x) = p(t)\dfrac{dt}{dx}$, или

$$p(x) = Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}. \qquad (7.3.4)$$

При $T = \ln Y$ плотность $p(y)$ задается формулой

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}\left(1-\alpha u(\ln y)^{\beta}\right)^{\frac{1}{u}-1}. \qquad (7.3.5)$$

При $y = \ln w$

$$p(w) = \frac{N(\ln\ln w)^{k\beta-1}}{w\ln w}\left[1-\alpha u(\ln\ln w)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (7.3.6)$$

В итоге имеем четыре системы непрерывных распределений – это основные системы. Каждая из четырех систем отличается от других тем, что соответствующая случайная величина имеет свое начало отсчета:

$X > -\infty; T > 0; Y > 1; W > e$.

Итак, метод построения обобщенных распределений по кривым роста новых событий привел к тем же результатам, что и при использовании метода обобщения.

Рассмотрим четыре основные системы кривых роста.

7.4. Система I кривых роста

Эти кривые описываются формулой (7.1.6)

$$y = \int_0^n \left(1 - e^{-x\,p(t)}\right) dt,$$

где y – математическое ожидание числа разных событий (например, разных слов), наступающих при x испытаниях (появляющихся в выборке объемом x словоупотреблений); p(t) – непрерывная плотность распределения, аппроксимирующая вероятности p_k (k=1,2,…,n) разных событий, составляющих полную группу. Эти события могут быть упорядочены по любому правилу, в том числе по убыванию (невозрастанию) вероятностей. В последнем случае зависимость между порядковым номером события (т. е. его рангом r) и вероятностью p_r, будет представлять собой ранговый закон распределения вероятностей разных событий. Таким образом, форма аппроксимирующей кривой распределения, заданной плотностью p(t), может быть различной.

С учетом рангового распределения формула (7.1.6) может быть записана в виде

$$y = \int_0^n \left(1 - e^{-xp_r}\right) dr. \qquad (7.4.1)$$

Поскольку здесь интегрирование осуществляется по всем рангам, то от порядка следования вероятностей p_r величина y не зависит.

Недостатком этой системы кривых роста является то, что интеграл (7.4.1), как правило, не выражается конечным числом элементарных функций. Кроме того, необходимо знать параметры закона распределения вероятностей разных событий, который может быть задан обобщенными плотностями p(t) или p(y).

Достоинством же системы I кривых роста является то, что по известной кривой роста новых событий y=f(x) можно рассчитать весь частотный спектр при любом за-

данном числе испытаний x, т. е. количество разных событий y_m с частотой m (m=1,2,…). При этом $\sum_{m\geq 1} y_m = y, \sum_{m\geq 1} m y_m = x$. Если задана кривая роста новых событий y=f(x), то частотный спектр рассчитывается по формуле В. М. Калинина (7.1.17). Если задана плотность p(t), то используется формула (7.1.7) или (7.1.7′).

При небольшом числе событий, составляющих полную группу, для вычисления математического ожидания числа наступивших разных событий при x испытаниях используется формула (7.1.1)

$$M[Y(X)] = \sum_{k=1}^{n} \left[1 - (1 - p_k)^X\right].$$

Рассмотрим пример.

Из таблицы случайных чисел отбирается x цифр. Необходимо установить зависимость среднего значения числа разных цифр y от объема выборки x. В данном случае можно воспользоваться формулой (7.1.1). Пусть x=10. Тогда при $p_k = 0{,}1, n = 10$ эта формула дает

$$M[Y(X)] = 10(1 - 0{,}9^{10}) = 6{,}513.$$

Найдем для сравнения величину y по общей формуле (7.1.6). В случае равномерной плотности распределения $p(t) = \alpha, n = 1/\alpha$ последняя формула принимает вид

$$y = \frac{1}{\alpha}\left(1 - \frac{1}{e^{\alpha x}}\right), \qquad (7.4.2)$$

откуда при α=0,1 находим: y=6,321.

Относительная ошибка формулы (7.4.2) составила 3%. С ростом количества разных событий n, составляющих полную группу, она уменьшается.

7.5. Система II (а,б) кривых роста. Кривая роста простых чисел

Эти кривые в общем виде задаются уравнением

$$\frac{dy}{dx} = 1 - \overline{F}(x) = 1 - \overline{F}(y), \qquad (7.5.1)$$

где $\overline{F}(x), \overline{F}(y)$ – функции распределения вероятностей новых событий.

Пусть эти функции распределения задаются формулами (7.2.8), (7.2.9). Тогда система IIа кривых роста будет описываться дифференциальным уравнением

$$\frac{dy}{dx} = (1 - \alpha u y)^{\frac{1}{u}} = [1 - \alpha(u-1)x]^{\frac{1}{u-1}}, \qquad (7.5.2)$$

решением, которого будет уравнение (1.3.10)

$$y = \frac{1}{\alpha u}\left[1 - (1 - \alpha(u-1)x)^{\frac{u}{u-1}}\right], \qquad (7.5.3)$$

где x – число произведенных испытаний; y – число наступивших разных событий.

Формула (7.5.3) позволяет выразить зависимость x от y:

$$x = \frac{1}{\alpha(u-1)}\left[1 - (1 - \alpha u y)^{\frac{u-1}{u}}\right]. \qquad (7.5.4)$$

При известных функциях распределения $\overline{F}(x)$ или $\overline{F}(y)$ кривые роста, относящиеся к системе IIа, в общем виде задаются формулами (7.2.8), (7.2.9)

$$y = \int [1 - \overline{F}(x)]dx + C,$$

$$x = \int \frac{dy}{1 - \overline{F}(y)} + C.$$

В отличие от формулы (7.1.1) формула (7.5.3) позволяет в явном виде выразить частотный спектр, т. е. количество событий с частотой 0, 1, 2 …

С другой стороны, по заданному частотному спектру может быть восстановлена кривая роста новых событий (7.5.3), найдены оценки её параметров и при необходи-

мости рассчитан новый частотный спектр при любом заданном числе испытаний x. Это значит, что с помощью формулы (7.5.3) можно прогнозировать кривую роста новых событий, а по системе дискретных распределений рассчитывать новый частотный спектр, т. е. прогнозировать дискретное распределение.

Для нахождения кривой роста новых событий можно также использовать закон их распределения, заданный либо функциями распределения $\overline{F}(x)$, $\overline{F}(y)$, либо средними плотностями $\overline{p}(x), \overline{p}(y)$.

Обе плотности являются невозрастающими, поскольку при осуществлении испытаний новые события наступают в среднем в порядке убывания их вероятностей. В связи с этим приведенные здесь формулы остаются справедливыми и в том случае, если новые события наступают строго по порядку убывания их вероятностей. Последнее условие выполняется на простых числах.

Действительно, новые простые числа в натуральном ряду чисел появляются в порядке возрастания: 2,3,5,… Этот порядок совпадает с порядком убывания их вероятностей. Так, если все натуральные числа от 1 до X, где X – достаточно большое натуральное число, представить в виде произведений простых чисел, подсчитать частоту употребления каждого простого числа и ранжировать их по убыванию частоты употребления, то получим тот же порядок: 2,3,5,…

Установленная взаимосвязь между кривой роста и законом распределения вероятностей новых событий позволяет решать различные задачи. Так, известно, что кривая роста простых чисел, т. е. количества простых чисел y, меньших натурального числа x, задается интегральным логарифмом (формулой П. Л. Чебышева) [5, с. 41]

$$y = \int_2^x \frac{dX}{\ln X}, \quad (X = x + 2, x \geq 0). \tag{7.5.5}$$

Найдем закон распределения вероятностей простых чисел. Дифференцируя (7.5.5) по X и принимая во внимание равенство (7.5.1), получим выражение для вероятности появления нового простого числа

$$\frac{dy}{dx} = \frac{1}{\ln X} = 1 - \overline{F}(X),$$

Откуда

$$\overline{F}(X) = 1 - \frac{1}{\ln X}, \qquad (7.5.6)$$

$$\overline{p}(X) = \frac{d\overline{F}(X)}{dX} = \frac{1}{X(\ln X)^2}, \qquad (7.5.7)$$

$$\overline{p}(y) = \overline{p}(X)\frac{dX}{dy} = \frac{1}{X \ln X}. \qquad (7.5.8$$

Из (7.5.6) следует, что величина X> e; X=e+x, где x – натуральное число.

Установим место закона распределения простых чисел в системе непрерывных распределений. Запишем двойное логарифмическое распределение, заданное обобщенной плотностью

$$p(w) = \frac{N(\ln\ln w)^{k\beta-1}}{w \ln w}\left[1 - \alpha u(\ln\ln w)^{\beta}\right]^{\frac{1}{u}-1}.$$

Рассмотрим частный случай этого распределения при u→0 (II тип)

$$p(w) = \frac{N(\ln\ln w)^{k\beta-1}}{w \ln w} e^{-\alpha(\ln\ln w)^{\beta}}, \quad (e < w < \infty), \qquad (7.5.9)$$

где нормирующий множитель N задается формулой

$$N = \frac{\beta\alpha^{k}}{\Gamma(\mathrm{k})}.$$

При значениях параметров α, β, k, равных единице, плотность (7.5.9) принимает вид

$$p(w) = \frac{1}{w(\ln w)^2}, \quad (e < w < \infty).$$

Последняя плотность распределения совпадает со средней плотностью $\bar{p}(x)$, заданной формулой (7.5.9). Следовательно, закон распределения простых чисел является частным случаем двойного логарифмического распределения (7.3.6), относится ко II типу группы А (поскольку u→0, k=1), характеризуется параметрами α=1, β=1 и задан на интервале e <X <∞ (X=e+x). С ростом X от e до ∞ плотность $\bar{p}(X)$ убывает от значения 1/e (при X=e) до нуля (при X→∞).

Закон распределения простых чисел можно также получить как частный случай из другой обобщенной плотности. Запишем логарифмическое распределение

$$p(Y)=\frac{N}{Y}(\ln Y)^{k\beta-1}\left[1-\alpha u(\ln Y)^{\beta}\right]^{\frac{1}{u}-1}$$

и рассмотрим Iґ тип (β <0; α>0; u>0)

$$p(Y)=\frac{\beta(\alpha u)^{k}\Gamma(k+1/u)}{\Gamma(k)\Gamma(1/u)}\frac{1}{Y(\ln Y)^{k\beta+1}}\left[1-\frac{\alpha u}{(\ln Y)^{\beta}}\right]^{\frac{1}{u}-1}. \qquad (7.5.10)$$

Здесь $\alpha u<(\ln Y)^{\beta}<\infty$, или $(\alpha u)^{1/\beta}<\ln Y<\infty$.

Если все параметры распределения (7.5.10) равны единице, то оно имеет вид

$$p(Y)=\frac{1}{Y(\ln Y)^{2}},\quad (e<Y<\infty),$$

т. е. совпадает с распределением (7.5.7).

Таким образом, закон распределения простых чисел является также частным случаем обобщенной логарифмической плотности p(y), относится к I′ типу группы А (поскольку u>0, k=1), характеризуется параметрами β=1; u=1; α=1 и задан на интервале e <Y <∞.

Тот факт, что один из фундаментальных законов природы – закон распределения простых чисел – входит как частный случай в обобщенные распределения автора, свидетельствует об их широких возможностях в деле познания законов природы. Другими словами, обоб-

щенные распределения – это не искусственные построения, а математические модели законов природы.

Формула для кривой роста простых чисел должна быть записана в виде

$$y = \int_{e}^{X} \frac{dX}{\ln X}. \qquad (7.5.11)$$

Она отличается от формулы П.Л.Чебышева (7.5.5) лишь значением нижней границы интегрирования (число e вместо 2). Но при X=2 кривая роста не удовлетворяет условию

$$\frac{dy}{dx} = 1 \text{ при } x \to 0.$$

Интегрируя (7.5.11), найдем (см. Г. Д. Двайт Таблицы интегралов – изд-во Наука, М, 1966, с. 120, п. 617)

$$y = \ln\ln X + \sum_{S=1}^{\infty} \frac{(\ln X)^S}{S \cdot S!} - \sum_{S=1}^{\infty} \frac{1}{S \cdot S!}, \qquad (7.5.12)$$

где

$$\sum_{S=1}^{\infty} \frac{1}{S \cdot S!} = 1{,}317902...$$

Расчетные (по формуле (7.5.12)) и эмпирические значения y при заданных значениях x, взятые из [5, с. 13], приведены в табл. 7.5.1. Формула (7.5.12) дает завышенные значения y по сравнению с эмпирической кривой роста простых чисел, однако с ростом x=X–e относительная ошибка убывает.

Таблица7.5.1

Эмпирическая и расчетная кривые роста простых чисел

x	уэмп.	урасч.	Относительная ошибка в %
2	1	1,561229	56,12
5	3	3,221949	7,40
100	25	28,81957	15,28
1000	168	176,1080	4,83
10000	1229	1244,537	1,26

x	уэмп.	урасч.	Относительная ошибка в %
100000	9592	9628,149	0,38
1000000	78498	78625,78	0,16
10000000	664579	664916,3	0,05
100000000	5761455	5762208	0,01

7.6. Система III кривых роста

Эти кривые задаются формулами

$$y = y_0 F(t), \tag{7.6.1}$$

$$y = y_0 [1 - F(t)], \tag{7.6.2}$$

где y_0 – некоторый параметр.

Свойства кривых роста этой системы полностью определяются свойствами функции распределения $F(t)$.

Формула (7.6.1) может описывать, например, количество разных статей по определенной теме, опубликованных в первых t журналах, при условии, что последние упорядочены по убыванию количества таких статей. Количество заболеваний при эпидемиях за время t от начала эпидемии, а также множество других кривых, в том числе кривых роста числа отказов за время t в теории надежности.

Система III кривых роста является более широкой, чем система II. Поэтому рассмотрим ее более детально.

Пусть в формуле (7.6.1) функция $F(t)$ представляет собой функцию распределения I′ типа группы A, т. е. задается формулой

$$F(t) = \left(1 - \frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}, \quad (t^{\beta} > \alpha u). \tag{7.6.3}$$

Снимем ограничения, накладываемые на параметры распределения (3.3.3), и запишем кривую роста (7.6.1) в виде

$$y = y_0 (1 + \alpha u t^{\beta})^{\frac{1}{u}}. \tag{7.6.4}$$

Подставляя в ту же формулу (7.6.1) функцию распределения III типа группы А, т. е.

$$F(t) = 1 - \frac{1}{\left(1 + \alpha |u| t^{\beta}\right)^{\frac{1}{|u|}}},$$

получим следующую кривую роста

$$y = y_0 \left[1 - \frac{1}{\left(1 + \alpha u t^{\beta}\right)^{1/u}} \right]. \qquad (7.6.5)$$

Кривые роста (7.6.4), (7.6.5) имеют большое разнообразие форм. Исследования показывают, что кривая (7.6.4) при одинаковых знаках параметров α, β растет (при α, β>0 – от y_0 до ∞, а при α, β<0 – от 0 до y_0), а при разных знаках – убывает (при α>0, β<0 – от ∞ до y_0, а при α<0, β>0 – от y_0 до 0). Кривая роста (7.6.5) имеет точку перегиба при условиях

$$\frac{1-\beta}{\alpha(\beta - u)} > 0, \quad \beta \frac{1-u}{\beta - u} > 0.$$

Параметр u в уравнении (7.6.4) может быть, как положительным, так и отрицательным. Рассмотрим случай, когда параметр u→0. При этом условии уравнение (7.6.4) примет вид

$$y = y_0 e^{\alpha t^{\beta}}. \qquad (7.6.6)$$

При β=1 из (7.6.6) имеем показательный закон роста с постоянным темпом роста

$$T_p = \frac{y_t}{y_{t-1}} = e^{\alpha} = 1 + \frac{\alpha}{1!} + \frac{\alpha^2}{2!} + \ldots \qquad (7.6.7)$$

Рассмотрим мгновенный темп прироста [7, с. 291] $\tau = d \ln y / dt$. В данном случае он равен

$$\tau = \frac{d \ln y}{dt} = \alpha \beta t^{\beta - 1} \qquad (7.6.8)$$

При β=1 мгновенный темп прироста равен α. При β>1 с ростом t он растет, а при β<1 – убывает. Следователь-

но, параметр β здесь является показателем ускорения или замедления темпа прироста (и темпа роста) кривой (7.6.6).

Исследования показывают, что в зависимости от значений параметров α, β семейство кривых, заданных уравнением (7.6.6), имеет формы, представленные в таблице 7.6.1.

Похожие формы имеют кривые при u>0 или u<0, которые заданы общим уравнением (7.6.4).

В некоторых случаях, например, при нахождении оценок параметров уравнения (7.6.4), его целесообразно представить в другом виде

$$y = \left[y_0^u + y_0^u \alpha u t^\beta\right]^{\frac{1}{u}}.$$

Вводя далее обозначения

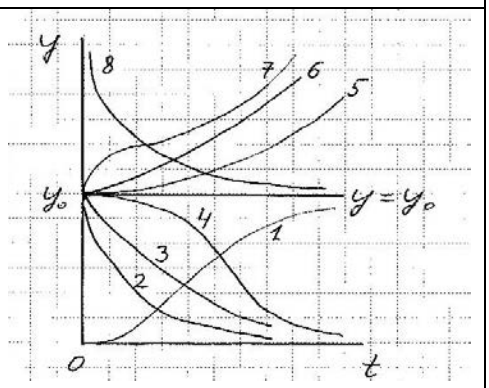
$$y_0^u = A; \quad \alpha u y_0^u = B; \quad \beta = C,$$

запишем окончательно

$$y = \left(A + Bt^C\right)^{\frac{1}{u}}. \qquad (7.6.9)$$

Таблица 7.6.1

Графики семейства кривых, заданных уравнением (7.6.4)

N формулы	Значения параметров		Формы кривых
	α	β	
1	α <0	β<0	
2		0<β<1	
3		β=1	
4		β>1	
5	α >0	β>1	
6		β=1	
7		0<β<1	
8		β<0	

В частном случае, при u→0, из (7.6.4) имеем

$$y = e^{\ln y_0 + \alpha t^{\beta}},$$

или

$$y = e^{A+Bt^{C}}, \qquad (7.6.10)$$

где $A = \ln y_0; \quad B = \alpha; \quad C = \beta$.

Уравнение (7.6.10) при C=1представляет собой экспоненту

$$y = e^{A+Bt}, \qquad (7.6.11)$$

темп роста которой не зависит от t. Поскольку для реальных кривых роста с увеличением t он изменяется, введем в формулу (7.6.11) новый параметр u, способный учитывать эти изменения

$$y = e^{(A+Bt)^{\frac{1}{u}}}. \qquad (7.6.12)$$

При u→0 вместо (7.6.12) будем иметь

$$y = e^{e^{A+Bt}}. \qquad (7.6.13)$$

Здесь по показательному закону растет логарифм величины y, т. е.

$$\ln y = e^{A+Bt}.$$

Полученные выше формулы будем относить к третьей системе кривых роста. Они обладают широкими возможностями по выравниванию различных эмпирических кривых, в том числе динамических рядов, поскольку включают как частные случаи множество известных моделей роста, имеющих практическое применение.

Однако они не описывают всего разнообразия кривых роста и динамических рядов, встречающихся на практике. Поэтому рассмотрим еще одно семейство кривых

третьей системы, частными случаями которого являются логистическая кривая и кривая Гомпертца.

Для построения такого семейства кривых воспользуемся все тем же уравнением (7.6.4), в котором положим $t = e^x$. В результате получим

$$y = y_0\left(1 + \alpha u e^{\beta x}\right)^{\frac{1}{u}}, \tag{7.6.14}$$

или

$$y = \left(A + Be^{\beta x}\right)^{\frac{1}{u}}, \tag{7.6.15}$$

где $A = y_0^u, \quad B = \alpha u y_0^u$.

Последнее уравнение можно также представить в виде

$$y = \left(A + BC^x\right)^{\frac{1}{u}}. \tag{7.6.16}$$

При u→0 уравнение (7.6.14) принимает вид

$$y = y_0 e^{\alpha e^{\beta x}} \tag{7.6.17}$$

или после введения других обозначений

$$y = y_0 A^{C^x}, \tag{7.6.18}$$

т. е. имеем кривую Гомпертца.

График функции (7.6.14) имеет точку перегиба с координатами

$$x_c = \frac{1}{\beta}\ln\frac{1}{(-\alpha)}; \quad y_c = y_0(1-u)^{\frac{1}{u}}. \tag{7.6.19}$$

Точка перегиба существует при α<0, u<1, причем, при β>0 кривая убывает, а при β<0 – растет. При α>0 кривая не имеет точки перегиба: при β>0 она растет, а при β<0 – убывает.

7.7. Система IV (а, б, в) кривых роста

Система IV кривых роста строится на основе обобщенных распределений (7.3.3), (7.3.4), (7.3.5). Вводя другие обозначения переменных и освобождаясь от ограничений, накладываемых на параметры кривых рас-

пределения, можем записать следующие уравнения для описания различного рода кривых роста (здесь γ=kβ)

$$IVa: \quad y = Nx^{\gamma-1}\left(1-\alpha u x^{\beta}\right)^{\frac{1}{u}-1}; \tag{7.7.1}$$

$$IVб: \quad y = Ne^{\gamma x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}; \tag{7.7.2}$$

$$IVв: \quad \ln y = N(\ln x)^{\gamma-1}\left(1-\alpha u \ln^{\beta} x\right)^{\frac{1}{u}-1}. \tag{7.7.3}$$

Система IV кривых роста является наиболее широкой системой, включающей кривые самой разнообразной формы. Она включает также некоторые кривые, принадлежащие другим системам. Часть кривых, являющихся кривыми распределения, нами была рассмотрена ранее.

Формулы (7.7.1)–(7.7.3) включают как частные случаи множество известных кривых, в том числе прямую, экспоненту, параболы различных степеней, кривую Гомпертца, логистическую кривую и др.

В диссертационной докторской работе автора [14] исследуются формы кривых (7.7.1)–(7.7.3), приводятся графики при различных значениях параметров формы.

Приведенные формулы могут быть использованы для описания числа наступивших разных событий y в зависимости от числа произведенных испытаний x.

В этом случае подходящими будут формулы:

$$y = x\left(1-\alpha u x^{\beta}\right)^{\frac{1}{u}}, \tag{7.7.4}$$

$$\ln y = \ln x\left(1-\alpha u \ln^{\beta} x\right)^{\frac{1}{u}}, \tag{7.7.5}$$

которые следуют из (7.7.1) и (7.7.3) при N=1, γ=2.

Последние две формулы с высокой точностью описывают кривые роста новых слов в тексте, а также другие кривые. Например, формула (7.7.5) хорошо описывает

кривую роста простых чисел. С учетом оценок параметров она имеет вид

$$\ln y = \frac{\ln x}{\left(1 + \frac{94{,}504404}{(\ln x)^{1{,}234022}}\right)^{0{,}131579}}. \qquad (7.7.6)$$

Относительная погрешность значений y, вычисленных по формуле (7.7.6) при 1000<x<100.000.000 находится в пределах -0,0022<δ<0,0022, т. е. не превышает 0,22%. При x=100 y=24,9 вместо 25, т. е. относительная погрешность равна -0,4%.

Несмотря на то, что формула (7.7.6) не удовлетворяет условию

$$\frac{dy}{dx} = 1 \text{ при } x \to 0,$$

она дает весьма высокую точность аппроксимации в широком интервале значений x.

7.8. Вычисление оценок параметров кривых роста

Для нахождения оценок параметров кривых роста по статистическим данным используется, как правило, метод наименьших квадратов.

Рассмотрим третью систему кривых роста, в частности, уравнение

$$y = y_0\left(1 + \alpha u t^{\beta}\right)^{\frac{1}{u}}. \qquad (7.8.1)$$

Здесь четыре параметра: y_0, α, β, u. Оценку величины y_0 можно взять непосредственно из эмпирической кривой роста при t=0 либо из сглаженной кривой. Остается найти оценки трех параметров – α, β, u. Для этого преобразуем уравнение (7.8.1) к линейному виду (при α>0):

$$\ln \frac{(y/y_0)^u - 1}{u} = \ln \alpha + \beta \ln t. \qquad (7.8.2)$$

Если ввести обозначения

$$Y = \ln\frac{(y/y_0)^u - 1}{u}; \quad A = \ln\alpha; \quad B = \beta; \quad T = \ln t,$$

то последняя формула примет вид

$$Y = A + BT.$$

Вычислив по методу наименьших квадратов оценки параметров A, B

$$B = \frac{N\sum TY - \sum T\sum Y}{N\sum T^2 - (\sum T)^2} = \frac{\overline{TY} - \overline{T}\cdot\overline{Y}}{\overline{T^2} - (\overline{T})^2}, \qquad (7.8.3)$$

$$A = \frac{1}{N}(\sum Y - B\sum T) = \overline{Y} - B\overline{T}, \qquad (7.8.4)$$

где N – число точек (объем выборки), найдем далее оценку параметра α:

$$\alpha = e^A.$$

Основная трудность здесь заключается в нахождении оценки параметра u. Она может быть определена с помощью ПЭВМ путем перебора значений этого параметра с заданным шагом, например, Δu=0,1 или 0,01 и т. д.

В качестве оценки параметра u можно принять то его значение, при котором коэффициент линейной корреляции

$$R = \frac{N\sum TY - \sum T\sum Y}{\sqrt{N\sum T^2 - (\sum T)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} = \frac{\overline{TY} - \overline{T}\cdot\overline{Y}}{\sqrt{\overline{T^2} - (\overline{T})^2}\sqrt{\overline{Y^2} - (\overline{Y})^2}} \qquad (7.8.5)$$

по модулю максимален, т. е. ближе к единице.

Кроме того, необходимо рассмотреть три случая: α>0; α<0; α→0.

При α→0 кривая роста (7.8.4) принимает вид (см. ниже (7.8.6)). В этом случае для нахождения оценок параметров α, β, y_0 можно воспользоваться следующим приемом.

Выразим зависимость y_{nt} (т. е. при значении $t^* = nt$) от y_t, где n>1:

$$y_{nt} = y_0 e^{\alpha (nt)^\beta} = y_0 \left(e^{\alpha t^\beta} \right)^{n^\beta} = y_0 \left(\frac{y_t}{y_0} \right)^{n^\beta} = y_0^{1-n^\beta} y_t^{n^\beta}, \qquad (7.8.6)$$

или

$$y_{nt} = a y_t^b, \qquad (7.8.7)$$

где $a = y_0^{1-n^\beta}$; $b = n^\beta$.

Определив оценки параметров α, β равенства (7.8.6), что можно сделать по уравнению прямой

$$\ln y_{nt} = \ln a + b \ln y_t,$$

находим далее оценки параметров y_0 и β:

$$y_0 = a^{1/(1-b)}; \qquad (7.8.8)$$

$$\beta = \frac{\ln b}{\ln n}. \qquad (7.8.9)$$

Оценка параметра α определится по формуле

$$\alpha = \frac{1}{S} \sum_{i=1}^{S} \frac{1}{t_i^\beta} \ln \frac{y_i}{y_0}, \qquad (7.8.10)$$

где S – количество значений Y, участвующих в расчете.

Оценки параметров α, β в формуле (7.8.6) при известном значении y_0 проще найти путем построения графика прямой

$$\ln \ln \frac{y}{y_0} = \ln \alpha + \beta \ln x$$

при α>0, либо прямой

$$\ln \ln \frac{y_0}{y} = \ln |\alpha| + \beta \ln x$$

при α<0.

Найдем далее оценки параметров уравнений (7.6.16) и (7.6.10):

$$y = \left(A + Bt^C \right)^{\frac{1}{u}}; \quad y = e^{A + Bt^C}.$$

Приняв обозначения

$$y^{u} = Y; \quad \ln y = Y,$$

последние два уравнения представим в виде

$$Y = A + Bt^{C}. \tag{7.8.11}$$

Теперь осталось найти оценки параметров A, B, C. Значение параметра u будем считать заданным.

Найдем из формулы (7.6.10) зависимость Y_{nt} от Y_t (при n>1):

$$Y_{nt} = A + Bn^{C} t^{C} = A + n^{C}\left(Bt^{C}\right).$$

Подставляя сюда значение Bt^{C} из (7.6.10), получим

$$Y_{nt} = n^{C} Y_t - \left(n^{C} - 1\right)A. \tag{7.8.12}$$

В этом уравнении два неизвестных параметра: A, C. Их можно найти по методу наименьших квадратов.

Найдем оценку параметра C:

$$n^{C} = \frac{S\sum Y_t Y_{nt} - \sum Y_t \sum Y_{nt}}{S\sum Y_t^{2} - \left(\sum Y_t\right)^{2}},$$

Откуда

$$C = \frac{1}{\ln n} \ln \frac{S\sum Y_t Y_{nt} - \sum Y_t \sum Y_{nt}}{S\sum Y_t^{2} - \left(\sum Y_t\right)^{2}}, \tag{7.8.13}$$

где S – количество значений Y_{nt}, участвующих в расчете. Величину n целесообразно принимать равной 1,5 или 2.

При известной оценке параметра C легко найти оценки параметров A, B уравнения (7.6.10)

$$B = \frac{N\sum Y_t t^{C} - \sum Y_t \sum t^{C}}{N\sum t^{2C} - \left(\sum t^{C}\right)^{2}}, \tag{7.8.14}$$

$$A = \frac{1}{N}\left(\sum Y_t - B\sum t^{C}\right), \tag{7.8.15}$$

где N – количество всех точек на эмпирической кривой роста.

Оценки параметров A, B, C вычисляются при различных значениях параметра u. В итоге выбирается тот вариант, который обеспечивает наибольший по модулю коэффициент корреляции или наименьшую остаточную дисперсию

$$D(Y)=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^*-Y_i\right)^2, \qquad (7.8.16)$$

где Y_i^*, Y_i – соответственно эмпирическое и расчетное значения Y.

При этом могут использоваться и другие критерии, например

$$DS=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i^*-Y_i}{Y_i}\right)^2. \qquad (7.8.17)$$

Найдем, наконец, оценки параметров уравнений (7.6.16) и (7.6.10):

$$y=\left(A+BC^x\right)^{\frac{1}{u}}; \quad y=y_0e^{\alpha e^{\beta x}}.$$

Перепишем последнее уравнение в виде

$$y=e^{A+BC^x}. \qquad (7.8.18)$$

Приняв обозначения $y^u=Y;\ \ln y=Y$ соответственно для первого и второго случаев, представим уравнения (7.7.16), (7.7.17) в единой форме

$$Y=A+BC^x. \qquad (7.8.19)$$

Теперь осталось найти оценки трех параметров: A, B, C. Значение параметра u будем считать заданным.

Найдем из (7.7.19) зависимость Y_{x+1} от Y_x

$$Y_{x+1}=A+BC^xC.$$

Поставляя сюда значение $BC^x=Y-A$ из (7.7.18), получим

$$Y_{x+1}=CY_x+A(1-C), \qquad (7.8.20)$$

т.е. имеем линейное уравнение, из которого по методу наименьших квадратов найдем оценку параметра C:

$$C = \frac{(N-1)\sum_{i=1}^{N-1} Y_{x_i} Y_{x_{i+1}} - \sum_{i=1}^{N-1} Y_{x_i} \sum_{i=1}^{N-1} Y_{x_{i+1}}}{(N-1)\sum_{i=1}^{N-1} Y_{x_i}^2 - \left(\sum_{i=1}^{N-1} Y_{x_i}\right)^2}. \qquad (7.8.21)$$

Далее при известной оценке параметра C из уравнения (7.8.19) таким же образом находим оценки параметров B, A:

$$B = \frac{N\sum_{i=1}^{N} Y_{x_i} C^{x_i} - \sum_{i=1}^{N} Y_{x_i} \sum_{i=1}^{N} C^{x_i}}{N\sum_{i=1}^{N} C^{2x_i} - \left(\sum_{i=1}^{N} C^{x_i}\right)^2}, \qquad (7.8.22)$$

$$A = \frac{1}{N}\left(\sum_{i=1}^{N} Y_{x_i} - B\sum_{i=1}^{N} C^{x_i}\right). \qquad (7.8.23)$$

Для нахождения оценок параметров A, B, C необходимо предварительно задать значение параметра u. Его целесообразно принимать с шагом 0,25 и менее, например, 0,05; 0,01.

При этом для оценки точности (качества) аппроксимации могут быть использованы рассмотренные выше критерии R, D(Y), DS, заданные формулами (7.8.4), (7.8.15), (7.8.16).

Аналогично находятся оценки параметров кривых роста, относящихся к IV системе.

Рассмотрим для примера формулу (3.4.4)

$$y = x\left(1 - \alpha u x^{\beta}\right)^{\frac{1}{u}}.$$

Преобразуем ее к уравнению прямой

$$\ln\frac{1-(y/x)^u}{u} = \ln\alpha + \beta\ln x.$$

Пусть $\ln\frac{1-(y/x)^u}{u} = Y;\quad \ln\alpha = A;\quad \ln x = X.$

Тогда имеем прямую $Y = A + BX$, оценки параметров которой легко находятся по методу наименьших квадратов.

7.9. Вычисление доверительных границ

Для вычисления нижней и верхней доверительных границ аппроксимирующей кривой роста используем свойства распределений.

Распределения, относящиеся к первой системе непрерывных распределений, описывают такие случайные величины, последующие значения которых получаются из предыдущих путем их изменения (сдвига) на некоторую постоянную величину, при этом средние значения таких случайных величин растут во времени по линейному закону [23].

Следовательно, рассеяние случайной величины относительно некоторой прямой может быть описано первой системой непрерывных распределений, а в частом случае – нормальным законом, если речь идет о рассеянии средних значений случайной величины относительно некоторой прямой.

Найдем для примера 90%-ые доверительные границы для кривой роста, заданной формулой (7.6.14)

$$y = y_0\left(1 + \alpha u t^{\beta}\right)^{\frac{1}{u}}.$$

Приведенная к уравнению прямой, она имеет вид

$$\ln\frac{(y/y_0)^u - 1}{u} = \ln\alpha + \beta\ln t. \qquad (7.9.1)$$

Рассеяние эмпирических значений случайной величины

$$Y = \ln\frac{(y/y_0)^u - 1}{u}$$

относительно прямой $\overline{Y} = \ln\alpha + \beta\ln t$ будет описываться первой системой непрерывных распределений, в частности, нормальным законом.

В случае необходимости распределение случайной величины y при заданных значениях t можно найти по распределению случайной величины Y по известной формуле

$$p(y) = p(Y)\frac{dY}{dy}.$$

Запишем формулы для вычисления верхней и нижней доверительных границ случайной величины y. На основании уравнения прямой (3.6.1) имеем

$$\ln\frac{\left(\frac{y_{в,н}}{\mathrm{y}_0}\right)^{u} - 1}{u} = \ln\alpha + \beta\ln t \pm ZS. \qquad (7.9.2)$$

Отсюда находим

$$y_{в,н} = \mathrm{y}_0\left(1 + \alpha_{в,н}\mathrm{ut}^{\beta}\right)^{\frac{1}{u}}, \qquad (7.9.3)$$

где

$$\alpha_в = \alpha\mathrm{e}^{\mathrm{ZS}}; \quad \alpha_н = \alpha/\mathrm{e}^{\mathrm{ZS}}.$$

Величина Z зависит от доверительной вероятности P и числа степеней свободы n, которое связано с числом точек n. При малых n величина Z определяется по таблицам распределения Стьюдента.

При P=0,9 величину Z можно рассчитать по формуле [24, с. 174]

$$Z = \frac{1{,}6448}{\left(1 - \frac{0{,}41611}{(n-2)^{1{,}0396}}\right)^{2{,}5}}, \qquad (7.9.4)$$

которая получена автором путем аппроксимации табличных данных по формуле (7.6.14).

Величина S – это среднее квадратическое отклонение эмпирических значений случайной величины Y от прямой $\overline{Y} = \ln\alpha + \beta\ln t$.

Произведение ZS является показателем точности аппроксимации при заданной надежности (доверительной вероятности) P.

Аналогично вычисляются доверительные границы в случае других кривых роста.

7.10. Системы кривых роста в теории надежности

Одним из практических приложений систем кривых роста является теория надежности. При этом использование той или иной системы зависит от плана испытаний на надежность.

7.10.1. Система IIa кривых роста

Система IIa кривых роста, заданная уравнением (7.5.3)

$$y = \frac{1}{\alpha u}\left[1 - (1 - \alpha(u-1)x)^{\frac{u}{u-1}}\right],$$

может использоваться для описания числа разных отказавших изделий $y = \sum y_m$ в зависимости от накопленного числа отказов $x = \sum m y_m$ всех N испытываемых изделий при условии, что испытания производятся по планам [NRr], [NRT], [NR(r,T)], где N обозначает количество испытываемых изделий; R – отказавшие изделия заменяются новыми; испытания прекращают при числе отказов r (суммарном по всем позициям) или при истечении времени испытаний или наработки T в каждой позиции.

Каждое из N испытываемых изделий занимает свою позицию (стенд, испытательную площадку, место книги на полке и т. д.). Следовательно, величина y_m – это число разных позиций, на каждой из которых произошло m отказов. Величина $x = \sum m y_m = r.$

Приведенная выше формула позволяет рассчитывать частотный спектр, т. е. вычислять количество разных

изделий (разных позиций), отказавших 0,1,2,…,m раз, а также прогнозировать кривую роста новых отказавших изделий и частотный спектр.

7.10.2. Система III кривых роста

Формула (7.6.1)

$$y = y_0 F(t)$$

может описывать количество отказавших изделий за время t, а формула (7.6.2)

$$y = y_0[1 - F(t)]$$

– количество работающих изделий на момент времени t при условии, когда на испытания поставлено N изделий, отказавшие изделия не восстанавливаются и не заменяются, испытания продолжаются до отказа r изделий или до момента времени T (планы испытаний [Nur], [NUT], [NU(r, T)]) [12].

В этом случае формула (7.6.1) может быть записана в виде

$$\frac{r_t}{N} = F(t), \qquad (7.10.2.1)$$

где r_t – число отказавших изделий на момент времени t; N – число испытываемых изделий; F(t) – теоретическая функция распределения наработки до отказа, которая в случае распределений I-III типов группы А может быть задана формулой

$$F(t) = 1 - \left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}}, \qquad (7.10.2.2)$$

а в случае распределений Iг-IIIг типов группы А – формулой

$$F(t) = \left(1 - \frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}. \qquad (7.10.2.3)$$

Оценки параметров аппроксимирующих распределений (и кривых роста) находятся по методу наименьших квадратов.

В случае распределений I-III типов на основании (7.10.2.1) и (7.10.2.2) имеем

$$\frac{r_t}{N} = 1 - \left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}}. \qquad (7.10.2.4)$$

Приведем последнее уравнение и виду

$$\ln \frac{1 - \left(1 - r_t / N\right)^u}{u} = \ln \alpha + \beta \ln t, \qquad (7.10.2.5)$$

т. е. получили уравнение прямой

$$Y = \ln \alpha + \beta X,$$

где $Y = \ln \frac{1 - \left(1 - r_t / N\right)^u}{u}$; $X = \ln t$.

Аналогичные формулы можно получить для распределений I′-III′ типов:

$$\frac{r_t}{N} = \left(1 - \frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}, \qquad (7.10.2.6)$$

откуда

$$\ln \frac{1 - \left(r_t / N\right)^u}{u} = \ln \alpha - \beta \ln t. \qquad (7.10.2.7)$$

После нахождения оценок параметров легко вычислить необходимые показатели надежности.

Так, вероятность отказа на момент времени t вычисляется по формулам (7.10.2.2) или (7.10.2.3).

Тогда вероятность безотказной работы будет равна

$$P(t) = 1 - F(t). \qquad (7.10.2.8)$$

Интенсивность отказов [12]

$$\lambda(t) = \frac{p(t)}{1 - F(t)}, \qquad (7.10.2.9)$$

где $p(t) = dF(t)/dt$ – плотность распределения наработки до отказа.

Средняя наработка до отказа

$$\bar{t} = \int_0^\infty tp(t)dt = \int_0^\infty P(t)dt\,. \qquad (7.10.2.10)$$

В зависимости от типа распределения она задается формулами:

$$I:\quad \bar{t} = \left(\frac{1}{\alpha u}\right)^{\frac{1}{\beta}} \frac{\Gamma(1+1/\beta)\Gamma(1+1/u)}{\Gamma(1+1/\beta+1/u)};$$

$$II:\quad \bar{t} = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}} \Gamma\left(1+\frac{1}{\beta}\right);$$

$$III:\quad \bar{t} = \left(-\frac{1}{\alpha u}\right)^{\frac{1}{\beta}} \frac{\Gamma(1+1/\beta)\Gamma(-1/u-1/\beta)}{\Gamma(-1/u)};$$

$$I':\quad \bar{t} = (\alpha u)^{\frac{1}{\beta}} \frac{\Gamma(1-1/\beta)\Gamma(1+1/u)}{\Gamma(1-1/\beta+1/u)};$$

$$II':\quad \bar{t} = (\alpha)^{\frac{1}{\beta}} \Gamma\left(1-\frac{1}{\beta}\right).$$

Найдем далее **мгновенный темп прироста отказов.** Пусть кривая роста отказов в общем виде задается функцией

$$r_t = NF(t)\,. \qquad (7.10.2.11)$$

Тогда мгновенный темп прироста отказов определится по формуле

$$\tau_{np} = \frac{\mathrm{dlnr_t}}{dt}\,. \qquad (7.10.2.12)$$

Логарифмируя (7.10.2.11), находим

$$\ln r_t = \ln N + \ln F(t)\,,$$

$$\tau_{np} = \frac{\mathrm{dlnr_t}}{dt} = \frac{p(t)}{F(t)}\,. \qquad (7.10.2.13)$$

Пусть далее кривая убывания исправных элементов задается в общем виде функцией

$$n_t = N[1-F(t)]. \qquad (7.10.2.14)$$

В этом случае величина

$$\tau_{np} = \frac{\mathrm{dlnn_t}}{dt} = -\frac{p(t)}{1-F(t)}. \qquad (7.10.2.15)$$

Таким образом, темп прироста (точнее, убывания) исправных элементов численно равен интенсивности отказов λ(t) (см. формулу (7.10.2.9)).

Приведенные выше формулы позволяют вычислять **наработку t до r_t отказавших изделий.**

В зависимости от типа распределения имеем:

$$I, III: \quad t = \left(\frac{1-\left(1-r_t/N\right)^u}{\alpha u}\right)^{\frac{1}{\beta}};$$

$$II: \quad t = \left(\frac{1}{\alpha}\ln\frac{N}{N-r_t}\right)^{\frac{1}{\beta}};$$

$$I': \quad t = \left(\frac{\alpha u}{1-(r_t/N)^u}\right)^{\frac{1}{\beta}};$$

$$II': \quad t = \left(\frac{\alpha}{\ln(N/r_t)}\right)^{\frac{1}{\beta}}.$$

Для вычисления наиболее подходящей кривой роста отказов, оценок параметров и показателей надежности автором разработана программа KROT (кривая роста отказов).

Выше были построены четыре системы кривых роста, рассмотрены методы оценивания параметров.

Первые две системы предназначены для описания кривых роста новых событий.

За основу построения кривых роста в первом случае (система I) принята плотность распределения вероятностей разных событий, во втором случае (система II) – функции распределения вероятностей новых событий $\overline{F}(y), \overline{F}(x)$.

За основу построения системы III кривых роста принята функция распределения вероятностей разных событий. Здесь к кривой роста не предъявляется никаких особых требований. Эти кривые имеют более разнообразные формы, чем кривые первых двух систем.

Система IV кривых роста строится на основе обобщенных плотностей $p(x), p(t), p(y)$ при условии снятия ограничений на параметры. Эта система кривых роста имеет самые разнообразные формы.

Приведены две кривые роста простых чисел. Первая из них получена на базе формулы П. Л. Чебышева, а вторая – на базе системы IV в кривых роста.

Рассмотрен метод оценивания доверительных границ для различных кривых роста с использованием обобщенных распределений.

В случае прямолинейной зависимости $y = f(x)$ доверительные границы при заданной доверительной вероятности рассчитываются по первой системе непрерывных распределений.

При показательном законе роста используется вторая система непрерывных распределений, а при двойном показательном законе – третья система.

При любом другом законе роста уравнение кривой необходимо преобразовать к прямолинейному виду и для нахождения доверительных границ полученной прямой использовать первую систему непрерывных распределений.

Наконец, рассмотрены возможности использования систем кривых роста в теории надежности.

8. ПРИМЕНЕНИЕ КРИВЫХ РОСТА ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ ТЕКСТА

8.1. Кривые роста новых слов в выборке

Под выборкой из некоторой совокупности текстов будем понимать такой условный текст, в котором разные слова появляются независимо и случайно.

Для описания кривых роста новых слов в выборке можно использовать формулы, полученные в предыдущих главах.

При известном законе распределения вероятностей однородных лексических единиц (слов, словосочетаний, дескрипторов), заданном, например, логарифмической плотностью распределения (7.3.5)

$$p(Y) = \frac{N(\ln Y)^{k\beta-1}}{Y}\left[1-\alpha u(\ln Y)^{\beta}\right]^{\frac{1}{u}-1},$$

можно рассчитать кривую роста новых слов в выборке $y = f(x)$, где x – объем выборки, y – объем словаря (y=Y-1). Расчет осуществляется по формуле (7.4.1)

$$y = \int_1^n \left(1 - e^{-xp(Y)}\right) dY .$$

Система I кривых роста позволяет рассчитывать частотный спектр. При этом используется формула

$$y_{m,x} = \frac{1}{m!}\int_1^n \frac{[xp(Y)]^m}{e^{xp(Y)}}\, dY .$$

Система IIа кривых роста

Система IIа кривых роста, как и система I, может быть использована для описания роста однородных лексических единиц.

Если известна статистическая структура выборки, т. е. заданы значения m, y_m $(m = 1,2,...)$, то кривая роста новых слов в выборке приближенно может быть описана формулой (7.5.3)

$$y = \frac{1}{\alpha u}\left[1 - (1 - \alpha(u-1)x)^{\frac{u}{u-1}}\right],$$

которая задает систему IIа кривых роста. При этом оценки параметров α, u могут быть рассчитаны по статистическому дискретному распределению или по трем величинам: $x, y, y_{m=1}$. С другой стороны, при известных оценках параметров α, u система IIа кривых роста позволяет рассчитывать частотный спектр при любом объеме выборки x.

Рассмотренные выше формулы справедливы в случае строгого соответствия статистического распределения теоретическому, что наблюдается далеко не всегда из-за наличия в частотном словаре группы слов, составляющих неоднородную часть. В связи с этим, а также с целью максимального облегчения расчетов представляется целесообразным для описания кривых роста новых слов в выборке использовать простые приближенные формулы, точность которых может быть оценена на реальных выборках. Их можно получить различными методами. При решении этой и других подобных задач широкие возможности предоставляют построенные автором системы кривых роста. Отберем из этих систем наиболее подходящие формулы.

Система IIб кривых роста

Эта система кривых роста получается из системы IIа введением замены $x = \ln X,\ y = \ln Y$. Она задается формулой (7.5.14)

$$\ln Y = \frac{1}{\alpha u}\left[1 - (1 - \alpha(u-1)\ln X)^{\frac{u}{u-1}}\right].$$

При различных значениях параметра u из (7.5.14) можно получить ряд приближенных формул для описания кривых роста новых слов в выборке, однако удобными для практических расчетов будут лишь те формулы, структура которых позволит в явном виде выразить параметр α через переменные X и Y. Формула (7.5.14) удовлетворяет этому требованию при u=2; 1/2; -1.

Проверка показала, что подходящей является формула при u=1/2:

$$Y = X^{1/\left(1+\frac{\alpha}{2}\ln X\right)}. \qquad (8.1.1)$$

Из (8.1.1) легко найти зависимость параметра α от X и Y

$$\alpha = \frac{2}{\ln X}\left(\frac{\ln X}{\ln Y} - 1\right). \qquad (8.1.2)$$

Поскольку $\ln X / \ln Y \geq 1$, то из (8.1.2) следует, что α≥0, причем, как было показано ранее, $\alpha = \bar{p}(Y = 1)$, т. е. теоретически параметр α представляет собой среднее значение вероятностей разных слов.

Параметр α, вычисленный по опытным значениям X, Y, не является постоянной величиной, однако он изменяется закономерно. Если построить график зависимости α от lnX, то получим прямую, которая задается формулой

$$\alpha = \alpha_0 + k \ln X \qquad (8.1.3)$$

или

$$\alpha = \alpha_0 + k' \lg X \qquad (8.1.3')$$

с начальной ординатой α_0 и угловым коэффициентом k (или k′=2,3026).

В таблице 8.1.1 приведены опытные значения X, Y для французского языка по данным П. Гиро (Guiraud P. Les caracteres Statistiques du vocabulaire. Paris, 1954. 116 p.). Если вычислить по формуле (8.1.2) значения параметра α и построить график зависимости α от lgX (см. рис. 8.1.1), то окажется, что опытные точки рассеяны вдоль прямой $\alpha = -0{,}0060 + 0{,}0125 \lg X$ с начальной ординатой $\alpha_0 = -0{,}0060$ и угловым коэффициентом $k' = 0{,}0125$. Значение α<0 связано с приближенностью формулы (8.1.1). Но уже при X>3 параметр α>0.

Таблица 8.1.1

Зависимость объема словаря Y от объема выборки X для французского языка

$\lg X$	X	Y	$\alpha_{u=1/2}$	$Y_{расч}$
3,301	2000	800	0,0361	815
3,602	4000	1250	0,0393	1260
3,778	6000	1580	0,0417	1598
3,903	8000	1900	0,0424	1878
4,000	10000	2120	0,0440	2119
4,146	14000	2505	0,0461	2523
4,301	20000	3040	0,0474	3009
4,602	40000	4200	0,0510	4122
5,000	100000	6000	0,0562	5928
5,301	200000	7500	0,0603	7510

В той же таблице приведены расчетные значения Y, вычисленные по формуле (8.1.1) при найденных оценках параметров α_0, k' (параметры прямой определены по графику). Статистические и расчетные данные различаются не более чем на $\pm 1{,}88\%$.

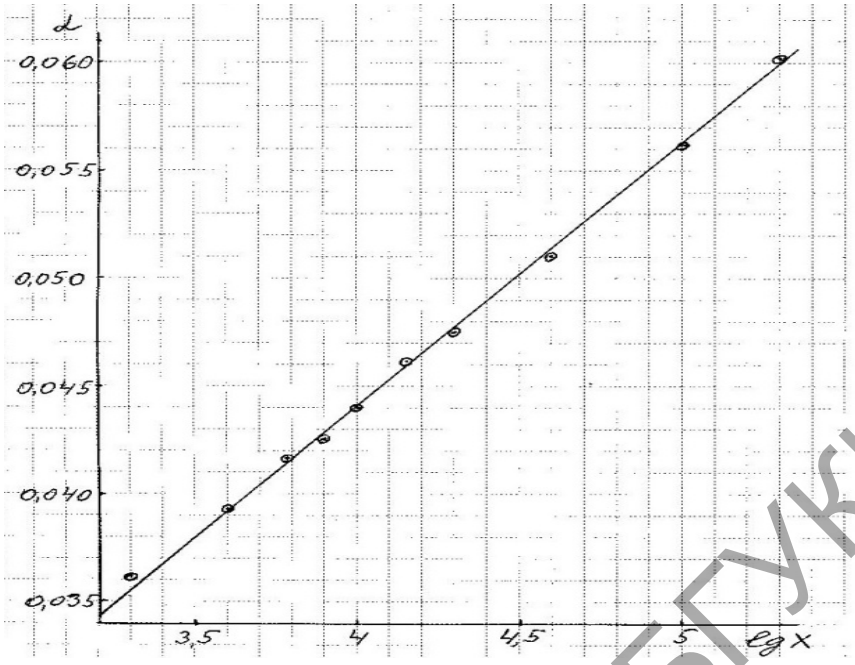


Рис. 8.1.1. Зависимость параметра $\alpha_{u=1/2}$ от lgx для французского языка

Несмотря на то, что в рассмотренном примере формула (8.1.1) с учетом (8.1.3′) оказалась подходящей для аппроксимации кривой роста новых слов в выборке, она имеет существенный недостаток: вначале растет до своего наибольшего значения $Y_{\max} = e^{1/(\alpha_0/2+\sqrt{2k})}$, которое достигается при $X = e^{\sqrt{2/k}}$, затем убывает. В качестве аппроксимирующей может служить восходящая часть кривой. Для рассмотренного случая $Y_{\max} = 19570$ (при $X = 2{,}167 \cdot 10^8$). Эту формулу практически можно использовать при объемах выборки $(10^3 \div 10^4) < X < (10^6 \div 10^7)$. При этом оценки параметров выборки α_0, k могут быть най-

дены не только по графику прямой (8.1.3), но и более просто – по трем величинам $X, Y, Y_{m=1}$, т. е. объемам выборки, словаря и числу слов с частотой m=1.

В соответствии с формулой В. М. Калинина (1.2.15) имеем

$$\frac{dY}{dX} = \frac{Y_{m=1}}{X}. \tag{8.1.4}$$

С другой стороны, дифференцируя (8.1.1) с учетом (8.1.3) по X, после преобразований получим

$$\frac{dY}{dX} = \frac{Y}{X}\left(\frac{\ln Y}{\ln X}\right)^2\left(1 - \frac{k}{2}\ln^2 X\right). \tag{8.1.5}$$

Приравнивая правые части двух последних формул, найдем

$$k = \frac{2}{\ln^2 X}\left[1 - \frac{Y_{m=1}}{Y}\left(\frac{\ln X}{\ln Y}\right)^2\right]. \tag{8.1.6}$$

Далее на основании (8.1.3) можем записать

$$\alpha_0 = \alpha - k\ln X, \tag{8.1.7}$$

или с учетом (4.1.2)

$$\alpha_0 = \frac{2}{\ln X}\left(\frac{\ln X}{\ln Y} - 1\right) - k\ln X. \tag{8.1.7′}$$

По формуле (8.1.5) можно рассчитать вероятность появления нового слова, которая равна первой производной в точке (X; Y) кривой роста новых слов в выборке.

Найдем еще одну формулу, параметр α которой выражается в явном виде через переменные X, Y. Для этого используем приближенное равенство

$$1 + \alpha\ln X \approx e^{\alpha\ln X} = X^{\alpha}, \tag{8.1.8}$$

которое тем точнее, чем меньше α.

На основании (8.1.8) из формулы (8.1.1) имеем

$$Y = X^{1/X^{\alpha/2}}, \tag{8.1.9}$$

откуда

$$\alpha = \frac{2}{\ln X} \ln \frac{\ln X}{\ln Y}, \tag{8.1.10}$$

причем, $\alpha = \alpha_0 + k \ln X$. Здесь параметр $u \approx 1{,}25$, поскольку величина α приближенно равна среднему из двух значений, рассчитанных при u=2, u=1/2.

Оценки параметров выборки здесь равны

$$k = \frac{2}{\ln^2 X}\left(1 - \ln \frac{\ln X}{\ln Y} - \frac{Y_{m=1}}{Y} \cdot \frac{\ln X}{\ln Y}\right), \tag{8.1.11}$$

$$\alpha_0 = \frac{2}{\ln X} \ln \frac{\ln X}{\ln Y} - k \ln X. \tag{8.1.12}$$

Отметим, что формулы (8.1.1), (8.1.9) могут быть получены как частные случаи из системы IV в кривых роста, которая задана уравнением (3.4.3), при N=1, ß=1, γ=2 и соответствующих значениях параметров α, u.

Система IVa кривых роста

Рассмотрим уравнение (8.4.1) – $y = Nx^{\gamma-1}\left(1 - \alpha u x^{\beta}\right)^{\frac{1}{u}-1}$. Кривые III и V типов при N=1, γ=2 обладают основными свойствами кривой роста новых событий. Они задаются уравнением

$$y = x\left(1 - \alpha u x^{\beta}\right)^{\frac{1}{u}-1}, \tag{8.1.13}$$

где 0<ß<u/(u-1). Для кривых III типа –∞ <u <0, поэтому величина ß определена на интервале 0<ß<1. Для кривых V типа 1<u <∞, следовательно,

1<ß<∞.

Проверка показала, что уравнение (8.1.13) может описывать кривые роста новых слов на интервале

$(10^3 \div 10^4) < x < (10^7 \div 10^8)$, при этом параметр $u \approx -1$ (0<β<1/2).

Формула (8.1.13) при u= –1 имеет вид [33

$$y = \frac{x}{(1+\alpha x^{\beta})^2}. \qquad (8.1.14)$$

Для нахождения оценок параметров α, β по методу наименьших квадратов преобразуем выражение (8.1.14) к уравнению прямой

$$\ln\left(\sqrt{\frac{x}{y}} - 1\right) = \ln\alpha + \beta \ln x. \qquad (8.1.15)$$

Оценки параметров выборки α, β можно найти по трем величинам: $x, y, y_{m=1}$.

Продифференцируем (8.1.14) по x:

$$\frac{dy}{dx} = \frac{1+\alpha(1-2\beta)x^{\beta}}{(1+\alpha x^{\beta})^3} \qquad (4.1.16)$$

или с учетом (8.1.14)

$$\frac{dy}{dx} = \frac{1+(1-2\beta)\left(\sqrt{\frac{x}{y}}-1\right)}{\left(\frac{x}{y}\right)^{3/2}}. \qquad (8.1.17)$$

Из (8.1.4) и (8.1.17) найдем

$$\beta = \frac{1-\frac{y_{m=1}}{y}}{2\left(1-\sqrt{\frac{y}{x}}\right)} = \frac{\sqrt{\frac{x}{y}}\left(1-\frac{y_{m=1}}{y}\right)}{2\left(\sqrt{\frac{x}{y}}-1\right)}. \qquad (8.1.18)$$

Тогда из (8.1.14) при известной оценке параметра β имеем

$$\alpha = \frac{1}{x^{\beta}}\left(\sqrt{\frac{x}{y}} - 1\right). \qquad (8.1.19)$$

Исследуем некоторые свойства кривой роста новых слов (8.1.14). Найдем долю одноразовых слов $y_{m=1}/y$. На основании (8.1.4) можем записать

$$\frac{y_{m=1}}{y} = \frac{x}{y} \cdot \frac{dy}{dx} = \frac{d \ln y}{d \ln x}. \qquad (8.1.20)$$

Из формулы (8.1.20) следует, что логарифмическая производная $d \ln y / d \ln x$ от кривой роста новых слов равна доле одноразовых слов.

С учетом (8.1.14), (8.1.16) последняя формула дает

$$\frac{y_{m=1}}{y} = 1 - \frac{2\alpha\beta x^{\beta}}{1 + \alpha x^{\beta}}. \qquad (8.1.21)$$

График зависимости $y_{m=1} / y$ от $\ln x$ имеет точку перегиба, которую обозначим через С. Из условия $d^2 (y_{m=1} / y) / d(\ln x)^2 = 0$ найдем координаты этой точки:

$$x_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}, \quad y_C = \frac{x_C}{4}, \quad \frac{y_{m=1}}{y_C} = 1 - \beta. \qquad (8.1.22)$$

Точку С легко найти с помощью графика прямой (8.1.15), поскольку

$$\ln\left(\sqrt{\frac{x_C}{y_C}} - 1\right) = 0.$$

Из (8.1.21) при $x \to \infty$ имеем равенство

$$\lim_{x \to \infty} \frac{y_{m=1}}{y} = 1 - 2\beta. \qquad (8.1.23)$$

Поскольку при $x \to \infty$ доля слов с частотой m=1 равна нулю, то из равенства (8.1.23) следует, что в этом предельном случае параметр ß→1/2. Это – верхняя его граница. Определим нижнюю границу ß. Для этого найдем долю слов с частотой m=2.

На основании формулы В.М.Калинина (7.1.17) можем записать

$$\frac{y_{m=2}}{y} = -\frac{x^2}{2y} \cdot \frac{d^2 y}{dx^2}$$

или с учетом (8.1.14)

$$\frac{y_{m=2}}{y} = \frac{\alpha\beta x^{\beta}\left[1+\beta+\alpha(1-2\beta)x^{\beta}\right]}{\left(1+\alpha x^{\beta}\right)^2}.$$

Отношение $y_{m=2}/y$ при некотором значении x имеет максимум. Координаты этой точки, найденные из условия $d(y_{m=2}/y)/dx = 0$, равны

$$x = \left(\frac{\beta+1}{\alpha(5\beta-1)}\right)^{\frac{1}{\beta}}, \qquad (8.1.24)$$

$$\left(\frac{y_{m=2}}{y}\right)_{\max} = \frac{(\beta+1)^2}{12}. \qquad (8.1.25)$$

Формула (8.1.24) при условии $x < \infty$ дает: β>1/5.

Таким образом, параметр ß имеет границы 0,2<β<0,5. При этих условиях максимальная доля слов с частотой m=2 изменяется в пределах

$$0{,}12 < \left(\frac{y_{m=2}}{y}\right)_{\max} < 0{,}1875,$$

а доля слов с частотой m=1 в точке С – в пределах

$$0{,}5 < \frac{y_{m=1}}{y_C} < 0{,}8.$$

Из формул (8.1.16), (8.1.22) следует, что скорость роста словаря в точке С зависит от параметра β:

$$\left(\frac{dy}{dx}\right)_C = \frac{1-\beta}{4}.$$

Чем меньше β, тем интенсивнее растет словарь. Следовательно, параметр β может служить показателем лексического разнообразия выборки.

Выясним далее лингвистический смысл параметра β. Для этого найдем значение функции распределения $\overline{F}(x)$ при x=1 $(\overline{F}(x=1) = \overline{F}(y=1) = \overline{p} = \sum_{r\geq 1} p_r^2)$. На основании (4.1.16) имеем

$$\overline{F}(x=1)=1-\frac{1+\alpha(1-2\beta)}{(1+\alpha)^3}\approx\frac{2\alpha(1+\beta)}{1+3\alpha}=\overline{p},$$

откуда $\alpha=\overline{p}/[2(1+\beta)-3\overline{p}]$. Произведение 2(1+β) заключено на интервале $(2{,}4\div 3)$. Следовательно, параметр α в зависимости от параметра ß заключен на интервале $\alpha=(0{,}417\div 0{,}375)\overline{p}$.

Система IVв кривых роста

Рассмотрим частный случай уравнения (7.7.3) при N=1, γ=2

$$\ln Y=\ln X\left(1-\alpha u\ln^{\beta}X\right)^{\frac{1}{u}-1}. \qquad (8.1.26)$$

Проверка показала, что последнее уравнение хорошо описывает кривую роста новых слов в выборке, при этом параметр u имеет большой разброс, но наиболее часто принимает значения на интервале от нуля до ½.

Пусть u→0. Тогда из (8.1.26) имеем равенство

$$\ln Y=\frac{\ln X}{e^{\alpha\ln^{\beta}X}}, \qquad (8.1.27)$$

которое может быть приведено к прямой

$$\ln\ln\frac{\ln X}{\ln Y}=\ln\alpha+\beta\ln\ln X. \qquad (8.1.28)$$

Параметры α,β находятся либо путем построения по опытным значениям X_i, Y_i графика зависимости (8.1.28), либо рассчитываются по значениям трех величин $X, Y, Y_{m=1}$:

$$\beta=\frac{1-\frac{Y_{m=1}}{Y}\cdot\frac{\ln X}{\ln Y}}{\ln\frac{\ln X}{\ln Y}}, \qquad (8.1.29)$$

$$\alpha=\frac{1}{\ln^{\beta}X}\ln\frac{\ln X}{\ln Y}. \qquad (8.1.30)$$

Пусть далее $u = 1/2$. Тогда из (8.1.26) получим

$$\ln Y = \left(1 - \frac{\alpha}{2}\ln^{\beta} X\right)\ln X,$$

откуда

$$Y = X^{1-\frac{\alpha}{2}\ln^{\beta} X} = \frac{X}{e^{\frac{\alpha}{2}(\ln X)^{\beta+1}}}, \qquad (8.1.31)$$

т. е. имеем формулу Ю. А. Тулдавы [44]. В обозначениях автора она выглядит несколько иначе

$$Y = \frac{X}{e^{a(\ln X)^{b}}}.$$

Достоинством формул (8.1.27), (8.1.31) является то, что они хорошо описывают кривые роста новых слов практически от начала координат до весьма больших значений X $(X = 10^{7} \div 10^{8})$ словоупотреблений. Однако при дальнейшем увеличении X поведение этих кривых не соответствует поведению кривых роста новых слов.

В приведенных выше формулах параметры выборки зависят как от типа текстов, по которым построена данная выборка, так и от ее объема. В связи с этим полученные формулы остаются справедливыми только в пределах заданной выборки, т. е. их нельзя использовать для экстраполяции кривой роста новых слов на выборку большего объема.

Проверим работу этих формул на конкретных примерах.

Восстановим кривую $y = f(x)$ на основе опытных данных по точной формуле (7.1.16), а также рассчитаем ее по приближенным формулам, при этом параметры выборки α_0, k, а также α, β определим по трем величинам: $X, Y, Y_{m=1}$.

В табл. 8.1.2 (графа 2) приведены результаты расчетов по формуле (7.1.16) на основе опытных данных частотного словаря немецкого языка и по приближенным формулам систем кривых роста IIб, IVa, IVв (графы 3–6):

$$IIб: \quad \mathrm{Y} = \mathrm{X}^{1/\mathrm{X}^{\alpha/2}} \quad (u = 1{,}25), \quad \alpha = \alpha_0 + k \ln X;$$

$$Y = X^{1\Big/\left(1+\frac{\alpha}{2}\ln X\right)} \quad (u = 1/2), \quad \alpha = \alpha_0 + k \ln X;$$

$$IVa: \quad y = \frac{x}{\left(1 + \alpha x^{\beta}\right)^2} \quad (u = -1);$$

$$IVв: \quad \ln \mathrm{Y} = \frac{\ln \mathrm{X}}{\mathrm{e}^{\alpha \ln^{\beta} \mathrm{X}}} \quad (u \to 0).$$

Объем выборки здесь равен X=10910777, объем словаря Y=258173 и количество одноразовых слов $Y_{m=1} = 126862$ [8]. Параметры выборки для каждой аппроксимирующей кривой приведены в соответствующей графе табл. 8.1.2.

Аналогичные расчеты выполнены по данным «Частотного словаря русского языка» (ЧСРЯ). Здесь X=1056382, Y=39268, $Y_{m=1} = 13379$ [48].

Из табл. 4.1.2 видно, что все формулы достаточно точно описывают кривую роста новых слов в выборке. Но в первом случае (ЧС немецкого языка) наименьшую точность показала формула при $u \approx 1{,}25$ (система IIб), а во втором случае (ЧСРЯ) – формула из той же системы при u=0,5.

Полученные результаты свидетельствуют о том, что параметр u изменяется от выборки к выборке. А это значит, что для более точного описания кривой роста новых слов в выборке необходимо использовать общие формулы (8.1.13), (8.1.26), каждая из которых содержит по три параметра. Для приближенных же расчетов удовлетворительные результаты дают рассмотренные выше формулы, имеющие по два параметра, причем, в приведенных примерах они определены лишь по трем величинам: $X, Y, Y_{m=1}$.

Таблица 8.1.2

Кривые роста новых слов, восстановленные по двум частотным словарям и рассчитанные по формулам систем кривых роста IIб, IVa, IVв

Объем выборки X	Объем словаря Y по ЧС	IIб		IVa	IVв
		u=1,25	u=0,5	u= -1	u→0
ЧС немецкого языка					
1091	625	600	619	603	626
2000	1019	979	1010	1000	1019
5000	2118	2004	2066	2092	2082
10911	3878	3609	3716	3819	3737
30000	–	7503	7699	8009	7725
100000	17800	17118	17462	18207	17485
300000	–	34735	35215	36459	35220
1091078	77870	75440	75952	77420	75913
3273233	140810	139063	139377	140230	139293
10910777	258173	258167	258232	258173	258083
α_0		0,02029	0,01622	–	–
k		0,0007485	0,001287	–	–
б		–	–	0,04223	0,005719
в		–	–	0,3005	1,3741
ЧС русского языка					
1000	625	650	713	645	646
3000	1510	1534	1674	1571	1526
10000	3637	3618	3889	3790	3604
30000	7384	7296	7682	7670	7275
105638	14815	14703	15095	15215	14683
211276	20700	20597	20890	21028	20582
528191	30495	30455	30544	30600	30450

Объем выборки X	Объем словаря Y по ЧС	IIб		IVa	IVв
		u=1,25	u=0,5	u= -1	u→0
1056382	39268	39268	39269	39268	39269
α_0		- 0,001633	- 0,01485	–	–
k		0,002935	0.004306	–	–
б		–	–	0,01451	0,001262
в		–	–	0,4084	2,0418

8.2. Кривые роста новых слов в связном тексте

Рассмотренные в п. 8.1 формулы справедливы для случайно составленной выборки, т. е. для такого условного текста, в котором разные слова появляются независимо и случайно. В реальном же тексте лексико-грамматические связи накладывают определенные ограничения на сочетаемость слов. Однако они не могут существенно изменить характер кривой роста новых слов в связном тексте по сравнению с соответствующей кривой в выборке, хотя изменяют значения параметров. Поэтому полученные формулы могут быть также использованы для описания кривых роста новых слов в связных лексически однородных текстах. Условимся обозначать параметры текста теми же символами, что и параметры выборки, а в случае необходимости добавлять индексы «т» или «в».

Проверка показала, что тексты, относящиеся к одному типу, имеют близкие значения параметра k (или в). При этом чем богаче данный текст в лексическом отношении, тем ниже располагается прямая $\alpha = \alpha_0 + k \ln X$ (для кривых системы IIб), т. е. для такого текста меньше начальная ордината α_0. Таким образом, параметр k может служить показателем лексического разнообразия данного типа текстов, а параметр α_0 – отдельно взятого текста (произведения).

Параметры текста в отличие от параметров выборки не зависят от объема текста и, следовательно, полученные формулы могут использоваться для экстраполяции кривой роста новых слов в тексте.

Заметим, что достоверные значения параметров текста можно получить лишь по такой кривой роста, которая построена на основе лексически однородных текстов. В особенности это замечание относится к смешанным текстам, когда в одну совокупность входят отрезки текстов разных типов. В этом случае при определении объема словаря y_i в подвыборке x_i требуется, чтобы в данную подвыборку входили отрезки связных текстов всех обрабатываемых типов и в тех же пропорциях, в каких они содержатся во всей исследуемой совокупности текстов. При соблюдении этих условий не разрушается связность слов в тексте, поскольку словоупотребления отбираются из текстов в их естественном порядке, и в то же время выполняется требование однородности, так как в каждой точке $(x_i; y_i)$ кривой роста новых слов пропорции между текстами разных типов одни и те же.

Здесь следует отметить, что параметры k, β в рассмотренных выше формулах в случае связных текстов меньше соответствующих параметров выборки. В связи с этим рассмотрим еще две формулы, которые на выборках нередко дают несколько завышенный рост словаря в начале кривой, но в случае связного текста более точно описывают рост словаря.

Первая формула относится к системе IIб кривых роста [33]

$$Y = X^{\frac{1}{\sqrt{1+\alpha \ln X}}}, \qquad (8.2.1)$$

где $\alpha = \alpha_0 + k \ln X$, причем, параметр α в явном виде выражается через переменные X, Y

$$\alpha = \frac{1}{\ln X}\left[\left(\frac{\ln X}{\ln Y}\right)^2 - 1\right] \qquad (8.2.2)$$

и примерно соответствует параметру $u \approx -1/4$.

Параметры текста α_0, k, входящие в формулу (8.2.1), находятся по нескольким точкам (не менее двух) на прямой $\alpha = \alpha_0 + k \ln X$. Значения параметров для некоторых текстов приводятся в работе автора [13].

Из (4.2.1) следует, что кривая роста новых слов имеет горизонтальную асимптоту $Y_{\max} = e^{\sqrt{1/k}}$. Чем меньше значение параметра k, тем больше величина $Y_{\max}$, т. е. параметр k является показателем лексического разнообразия текста.

При использовании формулы (8.2.1) для описания роста словаря в выборке параметры α_0, k могут быть оценены по количеству одноразовых слов. Дифференцируя (8.2.1) по X (при $\alpha = \alpha_0 + k \ln X$), после преобразований получим

$$\frac{dY}{dX} = \frac{Y}{X}\left(\frac{\ln Y}{\ln X}\right)^3\left(1 + \frac{\alpha_0}{2}\ln X\right). \qquad (8.2.3)$$

Далее в соответствии с формулой В. М. Калинина (7.1.15) имеем

$$\frac{dY}{dX} = \frac{Y_{m=1}}{X}.$$

Приравнивая правые части двух последних формул, найдем

$$\alpha_0 = \frac{2}{\ln X}\left[\frac{Y_{m=1}}{Y}\left(\frac{\ln X}{\ln Y}\right)^3 - 1\right]. \qquad (8.2.4)$$

Тогда

$$k = \frac{\alpha - \alpha_0}{\ln X} = \frac{1}{\ln^2 Y} - \frac{1}{\ln^2 X} - \frac{\alpha_0}{\ln X}. \qquad (8.2.5)$$

Вторая формула относится к системе IVб кривых роста и следует из общей формулы (8.1.26) при u = – 1:

$$\ln Y = \frac{\ln X}{\left(1 + \alpha \ln^\beta X\right)^2}. \qquad (8.2.6)$$

Параметры текста находятся из уравнения прямой

$$\ln\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right) = \ln\alpha + \beta \ln\ln X \qquad (8.2.7)$$

по методу наименьших квадратов на основе статистических данных.

Параметры выборки рассчитываются по формулам

$$\beta = \frac{1 - \frac{Y_{m=1}}{Y} \cdot \frac{\ln X}{\ln Y}}{2\left(1 - \sqrt{\frac{\ln Y}{\ln X}}\right)}, \qquad (8.2.8)$$

$$\alpha = \frac{1}{\ln^\beta X}\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right). \qquad (8.2.9)$$

Следует отметить, что формула (8.2.6) может быть получена из формулы (8.1.14) заменой величин x, y их логарифмами.

Итак, кривые роста новых слов в выборке и связном тексте могут быть описаны одними и теми же формулами, которые к тому же могут быть приведены к уравнениям прямой. Эти свойства кривых роста оказываются весьма полезными для решения различных задач. Рассмотрим некоторые из них.

8.2.1. Оценка степени аналитичности языка

Показателем степени аналитичности языка принято считать коэффициент, равный отношению количества

лексем к количеству словоформ частотного списка. Степень аналитичности языка тем выше, чем ближе этот показатель к единице.

Существенным недостатком этого показателя является его зависимость от объема текста. Знание аналитической зависимости между объемами словаря и текста позволяет по-другому измерять степень аналитичности языка.

Пусть кривая роста новых слов в тексте описывается формулой (8.1.14)

$$y = \frac{x}{\left(1 + \alpha x^{\beta}\right)^2}.$$

Преобразуем ее к уравнению прямой (8.1.15)

$$\ln(\sqrt{x / y} - 1) = \ln\alpha + \beta \ln x.$$

Значение параметра β этой прямой зависит от выбора единицы подсчета количества разных слов, в качестве которой может быть принята словоформа (сл.) или лексема (л.). При этом для одного и того же текста $\beta_л > \beta_{сл}$, в то время как $\alpha_л \approx \alpha_{сл}$. Последнее равенство позволяет ввести показатель степени аналитичности языка, который не зависит от объема текста:

$$\Delta\beta_a = \frac{v_л - v_{сл}}{\ln X}, \qquad (8.2.10)$$

где $v_л, v_{сл}$ рассчитываются по формуле

$$v = \ln\left(\sqrt{\frac{x}{y}} - 1\right), \qquad (8.2.11)$$

при этом $v_л = \ln\alpha_л + \beta_л \ln x$; $v_{сл} = \ln\alpha_{сл} + \beta_{сл} \ln x$.

Чем ближе $\Delta\beta_a$ к нулю, т. е. чем меньше угол между двумя прямыми на рис. 8.2.1, тем выше степень аналитичности языка.

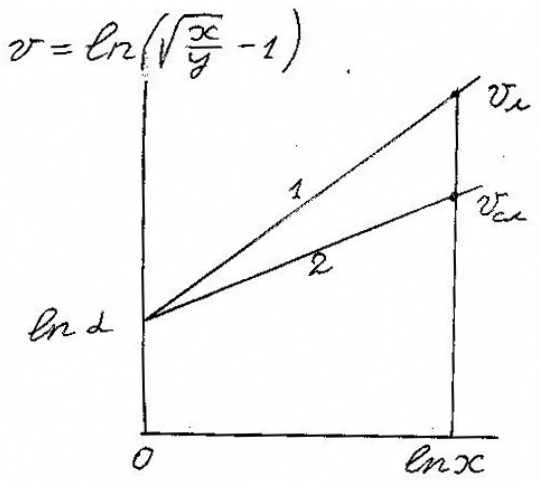


Рис. 8.2.1. Оценка степени аналитичности языка

В табл. 8.2.1 приведены значения $\Delta\beta_a$ для четырех языков, рассчитанные по формулам (8.2.10), (8.2.11) для текстов по электронике. Более подробная таблица приводится в работе автора [14, с. 396].

Полученные результаты позволяют осуществлять переход от кривой роста новых словоформ в тексте к кривой роста новых лексем. Для этого достаточно воспользоваться формулами

$$\alpha_л = \alpha_{сл}; \quad \beta_л = \beta_{сл} + \Delta\beta_a. \tag{8.2.12}$$

Таблица 8.2.1

Степень аналитичности языков (тексты по электронике

Язык, источник	Объем текста x	Объем словаря		$\Delta\beta_a$
		$y_{сл}$	$y_л$	
Русский	200894	21468	6826	0,0627
Румынский	200000	14292	5708	0,0479
Французский	100000	8108	4527	0,0336
Английский	200000	10582	7160	0,0202

8.2.2. Оценка степени связности слов в лексически однородном тексте

Так как в связном тексте лексико-грамматические связи накладывают определенные ограничения на сочетаемость слов, то естественно предположить (и это под-

тверждается опытными данными), что число разных слов в отрезке сплошного текста в среднем будет меньше, чем в случайно составленной выборке равного объема, взятой из достаточно большой совокупности лексически однородных текстов.

Эту разницу в объемах словарей можно использовать для оценки степени связности слов в тексте, причем, она оказывается независимой от объема текста.

Пусть для некоторого целого произведения из опыта известны объемы всего текста X, словаря Y, одноразовых слов $Y_{m=1}$, а также несколько промежуточных точек $(x_i; y_i)$ на кривой роста новых слов, которая описывается формулой (8.1.14)

$$y = \frac{x}{\left(1 + \alpha x^{\beta}\right)^2}.$$

По этим данным найдем параметры текста α_T, β_T путем построения графика зависимости (8.1.15)

$$v_T = \ln(\sqrt{x_T / y_T} - 1) = \ln \alpha_T + \beta \ln x_T.$$

На рис. 8.2.2 он представлен первой прямой.

Для этого же произведения можно рассчитать параметры выборки β_B, α_B по формулам (8.1.18), (8.1.19). Эти параметры будут относиться к такой кривой роста, которая получилась бы при случайном отборе словоупотреблений из данного произведения и подсчете количества разных слов. На рис. 8.2.2 график зависимости (8.1.15) для выборки представлен второй прямой.

Многочисленные расчеты по статистическим данным показали, что параметры α_T и α_B находятся в отношении

$$\frac{\alpha_T}{\alpha_B} = \eta_{св} \approx 2, \qquad (8.2.13)$$

которое может быть использовано в качестве показателя степени связности слов в тексте. Из (8.2.13) следует, что параметр $\alpha_T = 2\alpha_B$.

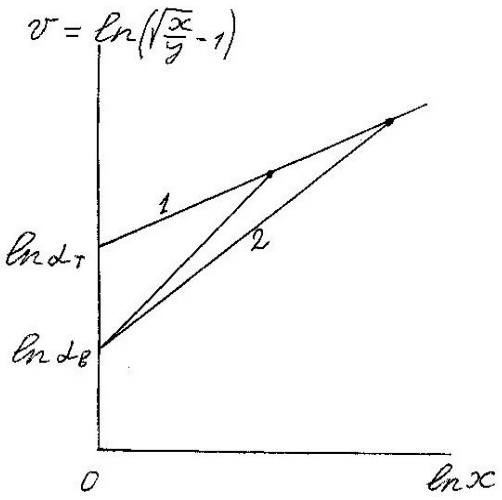


Рис. 8.2.2. Оценка степени связности слов в тексте

Параметры β_B и β_T связаны соотношением

$$\beta_T = \beta_B - \frac{\ln 2}{\ln x}. \qquad (8.2.14)$$

Формулы (8.2.12), (8.2.14) позволяют по параметрам выборки (которые легко вычислить по трем величинам $x, y, y_{m=1}$) найти параметры текста.

8.2.3. Оценка лексической близости двух связных текстов. Автоматическая классификация текстов

В информатике и лингвистике при решении практических задач часто требуется оценка степени близости (связи) словарей двух сравниваемых текстов. Она необходима, например, при автоматической классификации текстов (документов), при статистическом анализе индивидуальных и функциональных стилей и т. д.

Ниже предлагается метод оценки лексической близости двух статистически однородных текстов, основанный на использовании аналитической зависимости между объемами текста x и словаря $y = f(x)$.

Пусть кривая роста новых слов в связном тексте описывается формулой (8.1.14). Объединим два равных по объему текста с одинаковыми параметрами α, β и исследуем поведение обоих параметров этого объединенного текста. Рассмотрим два крайних случая.

Случай 1. Параметры α, β объединенного текста равны соответствующим параметрам объединяемых текстов, а его словарь содержит такое же количество разных слов, сколько их было бы в каждом из объединяемых текстов при удвоенной его длине

$$y_{12\rho=1} = f(2x_1) = \frac{2x_1}{\left(1+\alpha(2x_1)^{\beta}\right)^2}. \qquad (8.2.15)$$

В этом случае степень лексической близости двух текстов (будем измерять ее некоторым показателем ρ) равна единице, а точка с координатами $(\ln x_{12}; v_{12\rho=1})$ лежит на прямой 1 (см. рис. 8.2.3), поскольку

$$v_{12\rho=1} = \ln\left(\sqrt{\frac{2x_1}{y_{12\rho=1}}} - 1\right) = \ln\alpha + \beta\ln 2x_1 = v_1 + \beta\ln 2.$$

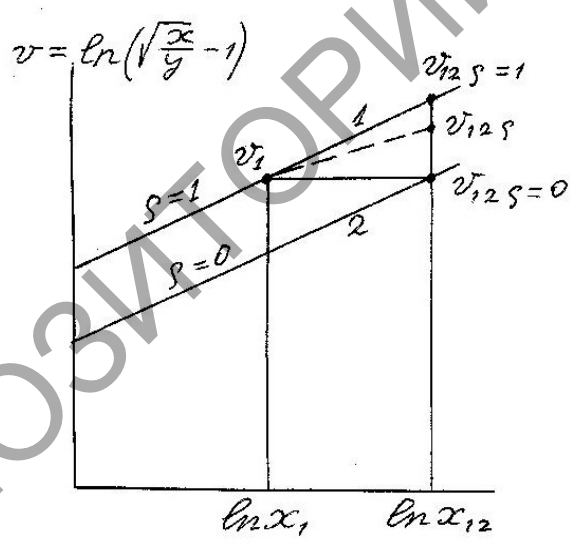


Рис. 8.2.3. Оценка лексической близости двух текстов

Случай 2. Словари обоих текстов не содержат общих слов. Это значит, что степень лексической близости таких текстов равна нулю $(\rho = 0)$. В этом случае объем словаря объединенного текста равен удвоенному объему словаря одного из объединяемых текстов

$$y_{12\rho=0} = 2y_1 = \frac{2x_1}{\left(1+\alpha x_1^{\beta}\right)^2}, \qquad (8.2.16)$$

$$v_{12\rho=0} = \ln\left(\sqrt{\frac{2x_1}{2y_1}} - 1\right) = v_1 = \ln\alpha + \beta \ln x_1,$$

т. е. прямая 2 параллельна прямой 1 и расположена ниже ее на расстоянии $\beta \ln 2$.

Из рис. 8.2.3 видно, что величина $v_{12\rho}$, характеризующая объем словаря реального объединенного текста, ограничена: $v_{12\rho=0} < v_{12\rho} < v_{12\rho=1}$. Следовательно, показатель ρ можно измерять отношением

$$\rho = \frac{v_{12\rho} - v_{12\rho=0}}{v_{12\rho=1} - v_{12\rho=0}} = \frac{v_{12\rho} - v_{12\rho=0}}{\beta \ln 2}. \qquad (8.2.17)$$

Из условия параллельности прямых 1 и 2 следует, что показатель ρ не зависит от объема текстов x_i, но оба они должны быть равными между собой.

При небольших размерах текстов (x <40000 словоупотреблений) кривая роста новых слов довольно точно аппроксимируется более простой формулой

$$y = \frac{x}{1 + \alpha x^{\beta}}. \qquad (8.2.18)$$

Ее можно привести к виду

$$v = \ln\left(\frac{x}{y} - 1\right) = \ln\alpha + \beta \ln x. \qquad (8.2.19)$$

Оценки параметров α, β находятся по эмпирической кривой роста новых слов. Для определения показателя ρ по формуле (8.2.17) необходимо вычислить входящие в нее величины, которые на основании (8.2.18) и (8.2.19) при $x_{12} = 2x_1$ равны

$$v_{12\rho} = \ln\left(\frac{x_{12}}{y_{12\rho}} - 1\right);$$

$$v_{12\rho=0} = \ln\left(\frac{x_{12}}{y_{12\rho=0}} - 1\right) = \ln\left(\frac{2x_1}{2y_1} - 1\right) = v_1$$

или

$$v_{12\rho=0} = \ln\alpha + \beta \ln x_1;$$

$$v_{12\rho=1} = \ln\alpha + \beta \ln x_{12} = v_1 + \beta \ln 2.$$

Задавая пороговое значение показателя ρ, можно классифицировать тексты (документы) по тематическим группам, включая в одну группу близкие по лексическому составу тексты.

8.2.4. Определение полноты словаря

При составлении частотного словаря по текстам разных типов (смешанным текстам) возникает вопрос об установлении определенных пропорций между ними. При этом естественно исходить из условия, чтобы тексты каждого типа одинаково полно отражали свою лексику. Критерием полноты в данном случае может служить накопленная средняя вероятность у разных слов, которые встретились в тексте данного типа объемом x словоупотреблений. Она определяется по формуле

$$\overline{F}(y) = 1 - \frac{dy}{dx},$$

где dy/dx – вероятность появления нового слова в тексте объемом x.

Если для описания кривой роста новых слов используется формула (8.1.14), то

$$\overline{F}(y) = 1 - \frac{1 + (1-2\beta)\left(\sqrt{x/y} - 1\right)}{(x/y)^{\frac{3}{2}}}. \qquad (8.2.20)$$

При использовании зависимости (8.1.27) имеем

$$\overline{F}(Y) = 1 - \frac{Y}{X} \cdot \frac{\ln Y}{\ln X}\left(1 - \beta \ln \frac{\ln X}{\ln Y}\right). \qquad (8.2.21)$$

Задавая для всех типов текстов величину $\overline{F}(Y)$ одинаковой (равной, например, 0,95), из опыта можно установить объем текста каждого типа, достаточный для достижения заданной полноты словаря.

Чтобы не вычислять оценок параметров для каждого типа текста, можно воспользоваться либо приближенной формулой

$$\overline{F}(Y) = 1 - \frac{Y}{X}\left(\frac{\ln Y}{\ln X}\right)^3, \qquad (8.2.22)$$

которая следует из (8.2.3) при $\alpha_0 = 0$, либо использовать формулу

$$F(y) = 1 - \frac{y_{m=1}}{x},$$

справедливую для случайной выборки.

Итак, рассмотрены примеры применения кривых роста новых событий при статистическом анализе связного текста и случайной выборки из него. На основе полученных результатов введены четыре показателя.

Показатель степени аналитичности языка, который позволяет переходить от кривой роста новых словоформ к кривой роста новых лексем.

Показатель степени связности слов в лексически однородном тексте, который позволяет по параметрам выборки находить параметры текста.

Показатель лексической близости двух связных текстов, с помощью которого можно выявлять близкие по содержанию тексты.

Показатель полноты словаря $\overline{F}(Y)$, т. е. накопленной средней вероятности Y разных слов, которые встретились в тексте объемом X словоупотреблений.

На базе введенных автором понятий «нового события», «кривой роста новых событий», «законов распределения вероятностей новых событий» и установленных

взаимосвязей между ними был разработан алгоритм порождения кривых роста и законов распределения вероятностей новых событий, который позволил построить систему кривых роста и систему непрерывных распределений новых событий.

Путем дальнейшего обобщения законов распределения вероятностей новых событий были построены четырехпараметрические обобщенные распределения, сгруппированные в четыре основные и три дополнительные системы непрерывных распределений. Они включают как частные случаи множество широко известных непрерывных распределений.

По системе кривых роста новых событий с помощью формулы В. М. Калинина, которая устанавливает взаимосвязь между кривой роста новых событий и частотным спектром, построена система дискретных распределений. Она включает как частные случаи биномиальный закон, Пуассона, отрицательный биномиальный, логарифмический и др. Дана классификация распределений, исследована форма полигона распределения в зависимости от значений параметров, разработаны методы нахождения оценок параметров.

Показано, что параметр u может служить показателем степени неравномерности появления отдельного события в выборках одинакового объема.

Система дискретных распределений в совокупности с системой кривых роста новых событий позволяет прогнозировать рост новых событий, а также рассчитывать частотный спектр на выборке любого объема, что было бы невозможно при использовании для этих целей отдельных дискретных распределений.

Для выравнивания и прогнозирования различного рода кривых роста и динамических рядов построены системы кривых роста, описаны методы оценивания параметров, вычисления доверительных интервалов при за-

данной доверительной вероятности с учетом свойств кривых роста.

Найдены простые приближенные формулы для описания кривых роста новых слов в связном тексте и случайной выборке, формулы для выравнивания динамических рядов. Получены две формулы для описания кривой роста простых чисел.

Выработаны показатели для оценки степени связности слов в тексте, степени аналитичности языка, степени лексической близости двух связных текстов, причем эти показатели не зависят от размеров текста. Получены простые формулы для вычисления полноты словаря.

Рассмотрены примеры применения кривых роста в теории надежности.

Для каждой системы распределений (непрерывных и дискретных), а также кривых роста автором разработаны соответствующие программы. Они вычисляют тип наилучшей аппроксимирующей кривой, выдают ее уравнение.

9. ПОСТРОЕНИЕ СИСТЕМЫ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ ПО КРИВЫМ РОСТА НОВЫХ СОБЫТИЙ

9.1. Моделирование кривой роста и статистической структуры словаря ключевых слов

Для аппроксимации статистических зависимостей между количеством произведенных испытаний и количеством наступивших разных событий автором разработана система кривых роста, заданная двухпараметрической формулой [21]

$$y = \frac{1}{\alpha u}\left[1-\left(1-\alpha(u-1)x\right)^{\frac{u}{u-1}}\right], \qquad (9.1)$$

где y – количество наступивших разных событий (разных слов, в том числе ключевых, наименований книг, запросов и т. д.); x – количество произведенных испытаний (объем выборки в словоупотреблениях, число книговыдач, число абоненто-запросов и т. д.).

Последняя формула включает систему кривых роста, которые можно разделить на типы в зависимости от значений параметра u.

При u>0 имеем кривые роста I типа. Они задаются формулой (9.1). В частности, при u→1 из (8.1) следует формула

$$y = \frac{1}{\alpha}\left(1-\frac{1}{e^{\alpha x}}\right). \qquad (9.2)$$

При u→0 из (1) следует кривая II типа

$$y = \frac{1}{\alpha}\ln(1+\alpha x). \qquad (9.3)$$

При u<0 имеем кривую III типа. Она задается той же формулой (9.1).

Между кривой роста разных событий и статистической структурой выборки существует взаимосвязь, установленная В. М. Калининым [6]

$$y_m = (-1)^{m+1} \frac{x^m}{m!} \frac{d^m y}{dx^m}. \qquad (9.4)$$

По формуле (9.4) можно рассчитать частотный спектр, или статистическую структуру выборки, т. е. количество событий с частотой появления 1, 2,…, m раз, если задана кривая роста разных событий y = f(x), причем такая кривая должна быть бесконечно дифференцируемой, что следует из формулы В. М. Калинина (9.4).

Формулы (9.1) и (9.4) позволяют также построить систему дискретных распределений.

9.2. Построение системы дискретных распределений

9.2.1. Распределения I типа (u > 0).

Продифференцируем выражение (8.1) m раз по x и подставим m-ю производную в (9.4). В результате получим формулу, позволяющую вычислять число событий с частотой m, т. е. y_m при числе испытаний x

$$y_m = \frac{y_{m=0}}{m!} \left(\frac{\alpha u x}{1+\alpha(1-u)x} \right)^m \prod_{i=0}^{m-1} \left[1+i\left(\frac{1}{u} - 1 \right) \right], \quad m = 1, 2, \ldots, \qquad (9.5)$$

где

$$y_{m=0} = \frac{1}{\alpha u} \left[1 + \alpha (1-u) x \right]^{\frac{u}{u-1}}. \qquad (9.6)$$

В данном случае число разных событий, наступающих при x испытаниях, ограничено: 0<y<1/ αu, причем, 1/ αu = n (величина n – это число разных событий, со-

ставляющих полную группу; сумма вероятностей этих событий равна единице).

Разделив величину y_m на n, получим выражение для вероятности наступления событий ровно m раз при x испытаниях: $p_m = y_m / n$ (при этом удобно разделить на n величину $y_{m=0}$):

$$p_m = \frac{p_{m=0}}{m!}\left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right], \quad m=1,2,\ldots, \tag{9.7}$$

где

$$p_{m=0} = \left[1+\alpha(1-u)x\right]^{\frac{u}{u-1}}. \tag{9.8}$$

Исследования показали, что частными случаями распределения I типа (7) являются: биномиальное – при u>1; Пуассона – при u → 1; отрицательное биномиальное – при 0<u<1 (в том числе геометрическое распределение – при u = ½).

9.2.2. Распределения II типа (u → 0).

В данном случае кривая роста разных событий задается формулой (9.3), на основании которой и формулы В. М. Калинина (9.4) имеем

$$y_m = \left(\frac{\alpha x}{1+\alpha x}\right)^m \frac{1}{\alpha m}, \quad m=1,2,\ldots \tag{9.9}$$

Разделив (9.9) на (9.3), получим

$$\frac{y_m}{y} = p_m = \left(\frac{\alpha x}{1+\alpha x}\right)^m \frac{1}{m\ln(1+\alpha x)}. \tag{9.10}$$

Последнее распределение известно как распределение Фишера по логарифмическому ряду и находит широкое применение в биологии [37].

9.2.3. Распределения III типа (- ∞ < u < ∞).

Кривая роста разных событий задается общей формулой (9.1). Из (9.1) и (8.4) имеем

$$y_m = \frac{y_{m=1}}{m!}\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1}\prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m=2,3,..., \tag{9.11}$$

где

$$y_{m=1} = x\left[1+\alpha(1-u)x\right]^{\frac{1}{u-1}}. \tag{9.12}$$

Разделив y_m на y, получим выражение для вероятности p_m

$$p_m = \frac{p_{m=1}}{m!}\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1}\prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m=2,3,..., \tag{9.13}$$

где

$$p_{m=1} = \frac{-\alpha u x}{\left(1+\alpha(1-u)x\right)\left[1-\left(1+\alpha(1-u)x\right)^{\frac{u}{1-u}}\right]}. \tag{9.14}$$

9.3. Оценивание параметров дискретных распределений

Чтобы решить эту задачу, вначале необходимо установить тип распределения. Это можно сделать двумя способами, для этого автором разработаны два метода: графический и аналитический.

В первом случае используются *графики зависимости* $\ln y_m = f(\ln m)$, которые представлены на рис. 9.1. Из него следует, что кривая дискретного распределения 2-го типа при малых частотах опускается по прямой под углом 45 градусов; кривая 1-го типа расположена выше кривой 2-го типа, а кривая 3-го типа – ниже этой кривой [15]

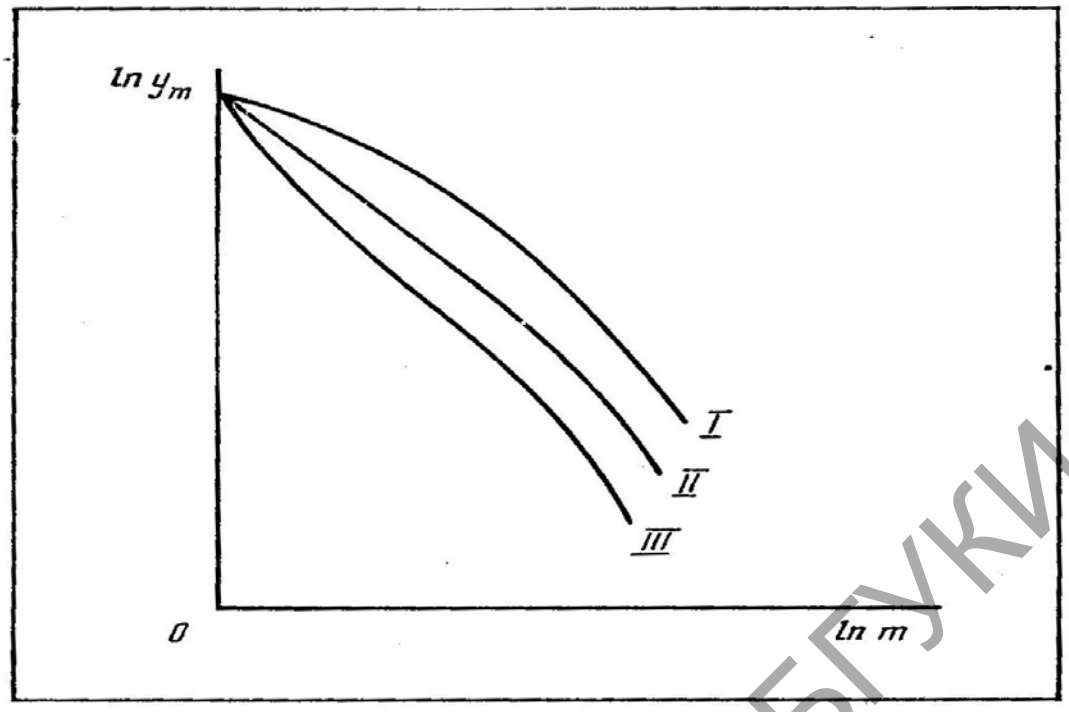


Рис. 9.1. Графики зависимости $\ln y_m = f(\ln m)$
для дискретных распределений 1–3-го типов

Аналитический метод. В этом случае используется критерий HD, который вычисляется по формуле [21]

$$HD = \frac{\frac{x}{y}\ln\frac{x}{y_{m=1}}}{\frac{x}{y_{m=1}} - 1}. \qquad (9.15)$$

При HD = 1 выравнивающее распределение относится ко II-му типу. При HD <1 – к I-му типу. При HD> 1 – к III-му типу.

Далее рассчитываются оценки параметров α, u. В случае распределений I типа

$$\alpha = \sum_{m\geq 1}\left(\frac{m}{x}\right)^2 y_m - \frac{1}{x}, \qquad (9.16)$$

$$u = \frac{1}{\alpha n}, \qquad (9.17)$$

где

$$x = \sum_{m\geq 1} m y_m, \quad n = \sum_{m\geq 0} y_m.$$

В случае распределений II типа оценка единственного параметра α находится методом простых итераций по формуле, которая следует из (9.3)

$$\alpha_{i+1} = \frac{1}{y}\ln(1+\alpha_i x), \tag{9.18}$$

где $y = \sum_{m\geq 1} y_m$; α_i – значение параметра α на предыдущем

шаге итерации. В качестве первого приближения можно принять $\alpha_1 = 1/y_{m=1}$.

В случае распределений III типа (а также I типа) оценка параметра u может быть найдена методом итераций по формуле

$$u_{i+1} = (1-u_i)\frac{x}{y}\frac{1-\left(\frac{y_{m=1}}{x}\right)^{u_i}}{\left(\frac{x}{y_{m=1}}\right)^{1-u_i}-1}. \tag{9.19}$$

Тогда оценка параметра α равна

$$\alpha = \frac{1}{uy}\left[1-\left(\frac{y_{m=1}}{x}\right)^{u}\right] = \frac{1}{(1-u)x}\left[\left(\frac{x}{y_{m=1}}\right)^{1-u}-1\right]. \tag{9.20}$$

Таким образом, для оценивания параметров α, u достаточно знать три величины: x, y, $y_{m=1}$.

Отметим, что формулы (9.16), (9.19), (9.20) справедливы для распределений трех типов.

При известных оценках параметров α, u вычисляются теоретические значения y_m и сравниваются со статистическими. Расчет осуществляется по рекуррентной формуле

$$y_{m+1} = y_m \frac{\alpha x[u+m(1-u)]}{[1+\alpha(1-u)x](m+1)}, \tag{9.21}$$

которая справедлива для распределений всех трех типов. Вначале по формуле (9.12) вычисляется количество событий с частотой m = 1, т. е. $y_{m=1}$. Далее по формуле

(9.21) последовательно находятся значения y_{m+1} при m = 1, 2 и т. д. В случае распределений I типа дополнительно вычисляется величина $y_{m=0}$ по формуле (9.6).

9.4. Кривая роста и статистическая структура словаря ключевых слов

Построенная система дискретных распределений, взаимосвязанная с системой кривых роста разных событий, позволяет легко решать многие задачи. Рассмотрим пример.

За некоторое время эксплуатации БелРАСНТИ (Белорусская Республиканская автоматизированная система научно-технической информации, которая разрабатывалась в БелНИИНТИ – Белорусском научно-исследовательском институте научно-технической информации и технико-экономических исследований ГОСПЛАНА БССР) при индексировании документов по автомобильному транспорту было употреблено y=3786 разных ключевых слов при общей их частоте употребления x = 147644. Количество ключевых слов с частотой m = 1 составило $y_{m=1}$ = 1518. По этим трем величинам требуется рассчитать:

– тип аппроксимирующего дискретного закона распределения;

– оценки параметров u, α дискретного распределения и кривой роста разных ключевых слов;

– кривую роста разных ключевых слов.

– частотный спектр ключевых слов, т. е. количество ключевых слов с частотой употребления 1, 2, …, m раз.

Установим тип аппроксимирующего дискретного распределения. Для этого вычислим по формуле (9.15) критерий HD. Он оказался равным 1,854.

Поскольку HD>1, то искомое распределение относится к III типу. Далее по формулам (9.19), (9.20) вычисляем оценки параметров u, α: u = -0,601736; α =

0,00645759. Частотный спектр описывается формулой (9.21).

Кривая роста разных ключевых слов описывается уравнением (9.1), которое при найденных оценках параметров u, α примет вид

$$y = 257.35\left[(1 + 0.0103435\,x)^{0.375677} - 1\right] \qquad (9.22)$$

Рассчитанные по формуле (9.22) значения y приведены в таблице 9.1, графа 3. В той же таблице в графе 2 даны значения y, восстановленные по частотному спектру ключевых слов с помощью формулы В.М.Калинина [6]

$$y = y_0 - \sum_{m \geq 1} \left(1 - \frac{x}{x_0}\right)^m y_m, \qquad (9.23)$$

где x, y – текущие значения объемов выборки и словаря ($x < x_0$; $y < y_0$; $x_0 = 147644$; $y_0 = 3786$). В таблице приводится также расчетное количество ключевых слов с частотой употребления один и два раза. Все расчеты выполнены по программе автора SDR99.

Таблица 9.1

Зависимость количества разных ключевых слов y от объема выборки x

Объем выборки x	Количество разных ключевых слов y		Количество ключевых слов с частотой	
	По частотному спектру	По формуле (9.22)	Один раз $Y_{m=1}$	Два раза $Y_{m=2}$
10000	1222	1218	549	170
20000	1671	1654	714	222
30000	1988	1967	833	259
40000	2240	2220	928	289
50000	2454	2436	1010	315
80000	2963	2955	1205	376

Объем выборки x	Количество разных ключевых слов y		Количество ключевых слов с частотой	
	По частотному спектру	По формуле (9.22)	Один раз $Y_{m=1}$	Два раза $Y_{m=2}$
100000	3239	3236	1311	409
147644	3786	3786	1518	474
200000		4274	1702	531
500000		6135	2401	749
1000000		8037	3116	972
1500000		9401	3628	1133
2000000		10503	4042	1262

Анализ данных таблицы 9.1 показывает, что максимальная относительная ошибка формулы (9.22) составила около 1%, и это при условии, когда оценки параметров u, α вычислялись всего лишь по трем величинам: x, y, $y_{m=1}$.

Формула (9.1) позволяет прогнозировать рост словаря ключевых слов в зависимости от количества заиндексированных документов D, а также полноту словаря, что важно знать при ведении информационно-поискового тезауруса. Для этого достаточно в формуле (9.1) заменить величину x на произведение Dh, где h – глубина индексирования. Ее можно оценить количеством ключевых слов, приходящихся в среднем на один поисковый образ документа.

Полноту словаря ключевых слов будем измерять вероятностью непоявления нового ключевого слова в точке с координатами (x;y) кривой роста, т. е. функцией распределения вероятностей новых событий $\overline{F}(y)$, при этом $\overline{F}(y)=\overline{F}(x)$. Новым будем считать любое слово при первом его употреблении для индексирования документа.

Полнота словаря рассчитывается по формуле (9.12).

$$\overline{F}(y)=\overline{F}(x)=1-\frac{dy}{dx}, \qquad (9.24)$$

где первая производная dy/dx равна вероятности появления нового слова.

Дифференцируя выражение (9.1) и подставляя первую производную в (9.24), получим

$$\overline{F}(y)=1-(1-\alpha u y)^{\frac{1}{u}},$$

$$\overline{F}(x)=1-(1-\alpha(u-1)x)^{\frac{1}{u-1}}.$$

Последние две формулы позволяют вычислять значения величин y, x, при которых будет достигнута заданная полнота $\overline{F}(y)=\overline{F}(x)$:

$$y=\frac{1}{\alpha u}\left[1-(1-\overline{F}(y))^{u}\right],$$

$$x=\frac{1}{\alpha(u-1)}\left[1-(1-\overline{F}(x))^{u-1}\right].$$

Полноту словаря объемом y можно также выразить через число ключевых слов с частотой употребления один раз [12]

$$\overline{F}(y)=1-\frac{y_{m=1}}{x},$$

где $y_{m=1}$ можно взять из частотного словаря ключевых слов. Эта формула следует из (9.24) и (9.4) при m = 1.

В таблице 9.2 приведены частоты употребления и количество ключевых слов с указанной частотой. Статистические и расчетные данные, вычисленные по программе SDR99, достаточно близки между собой.

Система дискретных распределений, взаимосвязанная с системой кривых роста разных событий, может быть использована во всех тех случаях, когда речь идет о последовательности независимых испытаний и частота появления разных событий подчиняется одному из дискретных законов, описанных в настоящем разделе.

Использование системы непрерывных распределений наряду с системой дискретных распределений, а также кривых роста и компьютерных программ, т. е. использование теории обобщенных распределений в целом позволяет описать все многообразие статистических распределений и кривых роста, которые встречаются в библиотечно-информационной деятельности.

С помощью математико-статистических моделей, наиболее точно аппроксимирующих статистические закономерности, из библиотечной (и любой другой) статистики может быть извлечена наиболее полная, объективная и ценная информация. При этом теория требует наличия определенных статистических данных. Так, при статистическом учете количества книговыдач, количества абоненто-запросов и т. д. совершенно необходимо вести учет количества разных наименований выданных книг, разных запросов и т. д. Только в этом случае может быть построена статистическая кривая роста разных событий. Анализ такой кривой дает объективную информацию, необходимую при решении различных задач. Это оптимизация комплектования фонда, оценка его полноты, анализ использования, оценка состояния и прогнозирование.

Таблица 9.2

Количество ключевых слов y_m с заданной частотой m

Частота m	Количество слов по факту		Количество слов по расчету	
	Y_m	Сумма Y_m	Y_m	Сумма Y_m
1	2	3	4	5
1	1518	1518	1518	1518
2	450	1968	473,6	1991,6
3	229	2197	256,2	2247,8
4	144	2341	168,0	2415,8
5	136	2477	121,7	2537,5
6	119	2596	93,7	2631,2
7	77	2673	75,3	2706,5
8	66	2739	62,3	2768,8

Частота m	Количество слов по факту		Количество слов по расчету	
	Y_m	Сумма Y_m	Y_m	Сумма Y_m
9	72	2811	52,7	2821,5
10	55	2866	45,4	2866,9
11	47	2913	39,7	2906,6
12	38	2951	35,2	2941,8
13	41	2992	31,4	2973,2
14	27	3019	28,3	3001,5
15	25	3044	25,7	3027,2
16 и >	742	3786	758,8	3786

Использование системы непрерывных распределений позволяет вычислять наилучшее аппроксимирующее распределение, в том числе ранговое, находить универсальные законы рассеяния и старения публикаций. На базе универсального закона рассеяния можно дать математически точную формулировку закона Бредфорда, вычислить границы ядра и зон рассеяния, доли статей в каждой зоне.

По всем разделам теории обобщенных распределений автором созданы компьютерные программы, которые апробированы на большом статистическом материале в течение длительного времени – с 1990 г.

Использование теории обобщенных распределений гарантирует высокую экономическую эффективность статистических методов во всех практических приложениях, в том числе в библиотечно-информационной деятельности, в системах управления качеством, в научных исследованиях.

Возьмем из табл. 9.2 расчетные данные о частоте ключевых слов и их количестве, вычисленном по дискретному распределению 3-го типа автора. По этим данным построим график зависимости $\mathrm{LNYm} = \mathrm{f(LNm)}$.

Из построенного графика видно, что эта зависимость близка к уравнению прямой, т. е. фактически имеем **закон Лотки.** Но теоретическая указанная зависимость имеет точку перегиба, что четко видно на представлен-

ном рис. 9.1. И, следовательно, в окрестности этой точки действительно можно провести прямую, но только на некотором ограниченном с двух сторон от этой точки интервале. При увеличении частоты m точки все дальше рассеиваются от указанной прямой и не ложатся на прямую. В доказательсто этого факта приведем еще один график (рис. 9.2) из учебного пособия автора [27]. В этом случае статистические данные полные – учтены словосочетания с частотами от 1 до 107.

Итак, закон Лотки получил теоретическое обоснование – он является следствием свойств дискретного распределения 3-го типа В. Нешитого.

Однако это не значит, что закон Лотки выполняется всегда. В моей диссертационной работе [14] рассмотрен случай дискретного статистического распределения научных работников по продуктивности по данным Л. С. Козачкова, которые подчиняются дискретному распределению первого типа, и следовательно, график зависимости $\mathrm{LNYm} = \mathrm{f(LNm)}$ имеет вид выпуклой кверху кривой (см. рис. 9.1 верхняя кривая).

В то же время закон Ципфа не подтверждается системами непрерывных распределений автора, а также статистическими ранговыми распределениями, которые в системе координат $rp_r = f(\ln r)$ всегда имеют колоколообразную форму с модой и двумя точками перегиба, а по закону Ципфа должна быть прямая, которая не содержит никаких характерных точек; но именно их наличие позволяет вычислять границы ядра и зон рассеяния. При использовании словесной формулировки от 1948 г. закона рассеяния Бредфорда еще никому не удалось точно вычислить границы ядра и зон рассеяния, хотя это легко сделать по закону Вейбулла.

Таким образом, для точной аппроксимации статистических распределений (непрерывных и дискретных) необходимо использовать Теорию обобщенных распределений.

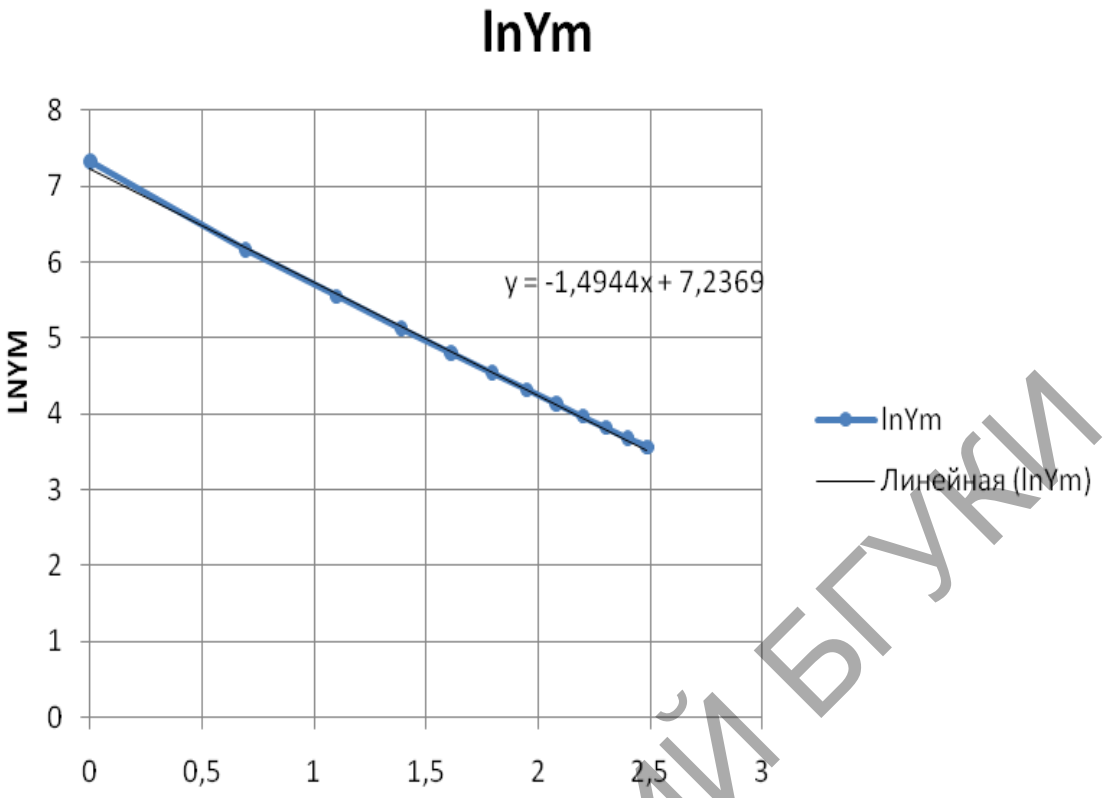


Рис. 9.1. Диаграмма Лотки $\mathrm{Ym} = \frac{\mathrm{Y}_1}{m^{1,4944}}$;

$$\mathrm{LNYm} = \mathrm{LNY}_1 - 1{,}4944\,LNm$$

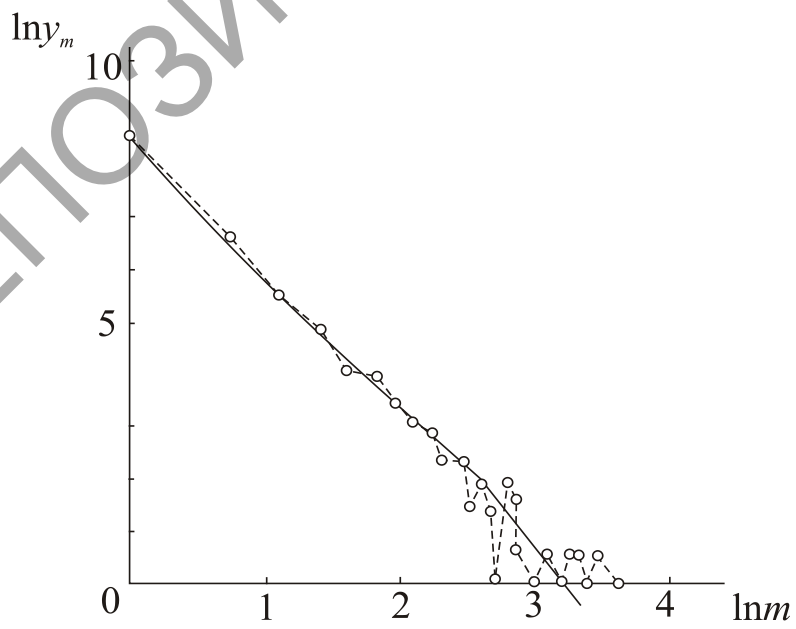


Рис. 9.2. Распределение словосочетаний в английском газетном тексте по дискретному распределению 3-го типа В. Нешитого

10. МЕТОД НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ, УСТОЙЧИВЫЙ МЕТОД И ЭНТРОПИЯ

10.1. Метод наибольшего правдоподобия

Для нахождения оценок параметров аппроксимирующих распределений Р. Фишер в 1912 г. предложил метод наибольшего (максимального) правдоподобия. Суть его сводится к тому, что наступившие события имели наибольшую вероятность наступить при заданном комплексе условий. Вероятность совместного наступления событий при условии их независимости равна произведению вероятностей наступивших событий. Она также будет максимальной. Это произведение называется функцией правдоподобия, т. е. $L = \prod_{i=1}^{n} f(x_i, \theta_j)$.

Чтобы упростить расчеты по нахождению оценок параметров θ_j аппроксимирующих распределений, рассматривают не произведение вероятностей наступивших событий, а сумму логарифмов вероятностей этих событий, которая также будет максимальной. Таким путем получают логарифмическую функцию правдоподобия

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta_j). \qquad (10.1)$$

Дифференцируя ее по параметрам θ_j и приравнивая частные производные нулю, получают систему уравнений правдоподобия, решая которую, находят оценки параметров. Однако здесь следует отметить, что в случае распределений с тремя и тем более с четырьмя параметрами получаются весьма сложные уравнения, решение которых сопряжено с большими трудностями.

10.2. Модифицированный метод наибольшего правдоподобия

В качестве логарифмической функции правдоподобия для непрерывных распределений может быть принято математическое ожидание логарифма плотности распределения [10.1], которое также будет максимальным, т. е.

$$M[\ln p(x)] = \ln L = (\ln L)_{\max}. \qquad (10.2)$$

Использование этой функции правдоподобия позволяет значительно проще решать различные задачи, например, вычислять оценки параметров.

Рассмотрим пример на вычисление M[lnp(x)]. Логарифмическую функцию правдоподобия можно вычислить двумя способами. Первый традиционный способ – это ее вычисление путем интегрирования

$$M[\ln p(x)] = \int [\ln p(x)] p(x) dx. \qquad (10.3)$$

Пусть плотность p(x) задается четырехпараметрической формулой [15]

$$p(x) = \frac{\beta(\alpha u)^k \Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})} e^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}, \qquad (10.4)$$

которая относится к первому типу первой системы непрерывных распределений.

Тогда логарифмическая функция правдоподобия будет выражаться интегралом

$$M[\ln p(x)] = \int \left[\ln \frac{\beta(\alpha u)^k \Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})} + k\beta x + (1/u - 1)\ln\left(1-\alpha u e^{\beta x}\right) \right] p(x) dx,$$

где плотность p(x) задается формулой (10.4).

При втором способе используется метод дифференцирования [31].

Прологарифмируем плотность распределения (10.4)

$$\ln p(x) = \ln\beta + k\ln\alpha u + \ln\Gamma(\mathrm{k}+1/\mathrm{u}) - \ln\Gamma(\mathrm{k}) - \ln\Gamma(1/\mathrm{u}) + \mathrm{k}\beta\mathrm{x} + (1/\mathrm{u} - 1)\ln\left(1 - \alpha u \mathrm{e}^{\beta \mathrm{x}}\right).$$

На основании последнего равенства запишем логарифмическую функцию правдоподобия

$$\ln L = M[\ln p(x)] = \ln\beta + k\ln\alpha u + \ln\Gamma(\mathrm{k}+1/\mathrm{u}) - \ln\Gamma(\mathrm{k}) - \ln\Gamma(1/\mathrm{u}) + \mathrm{k}\beta\mathrm{M}(\mathrm{x}) + \\ + (1/\mathrm{u}-1)\mathrm{M}\left[\ln\left(1-\alpha u e^{\beta \mathrm{x}}\right)\right]. \quad (10.5)$$

Возьмем от функции (10.5) частные производные по параметрам α, β, k, u и приравняем их нулю. В результате получим четыре уравнения правдоподобия с четырьмя неизвестными параметрами. Решив систему уравнений правдоподобия, найдем оценки параметров по методу наибольшего правдоподобия. Итак, частные производные по параметрам задаются системой уравнений правдоподобия

$$\begin{cases} \dfrac{\partial \ln L}{\partial \alpha} = \dfrac{k}{\alpha} + \left(\dfrac{1}{u}-1\right)M\left(\dfrac{-ue^{\beta x}}{1-\alpha u e^{\beta x}}\right) = 0 \\ \dfrac{\partial \ln L}{\partial \beta} = \dfrac{1}{\beta} + kM(x) + \left(\dfrac{1}{u}-1\right)M\left(\dfrac{-\alpha u e^{\beta x}x}{1-\alpha u e^{\beta x}}\right) = 0 \\ \dfrac{\partial \ln L}{\partial k} = \ln\alpha u + \psi\left(k+\dfrac{1}{u}\right) - \psi(k) + \beta M(x) = 0 \\ \dfrac{\partial \ln L}{\partial u} = \dfrac{k}{u} + \psi\left(k+\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right) - \psi\left(\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right) + \left(-\dfrac{1}{u^2}\right)M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right] + \left(\dfrac{1}{u}-1\right)M\left(\dfrac{-\alpha u e^{\beta x}}{1-\alpha u e^{\beta x}}\right) = 0 \end{cases} \quad (10.6)$$

Полученные уравнения можно несколько упростить. Так, из первого уравнения правдоподобия имеем

$$k = \alpha(1-u)M\left(\frac{e^{\beta x}}{1-\alpha u e^{\beta x}}\right). \quad (10.7)$$

Умножив четвертое уравнение правдоподобия на u, получим

$$k + \psi\left(k+\frac{1}{u}\right)\left(-\frac{1}{u}\right) - \psi\left(\frac{1}{u}\right)\left(-\frac{1}{u}\right) + \left(-\frac{1}{u}\right)M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right] - \alpha(1-u)M\left(\frac{e^{\beta x}}{1-\alpha u e^{\beta x}}\right) = 0.$$

С учетом (10.6) последнее равенство примет вид

$$\psi\left(k+\frac{1}{u}\right) - \psi\left(\frac{1}{u}\right) + M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right] = 0. \quad (10.8)$$

Перепишем систему уравнений правдоподобия для первого типа распределений, заданных плотностью (10.3)

$$\begin{cases} k-\alpha(1-u)M\left(\dfrac{e^{\beta x}}{1-\alpha u e^{\beta x}}\right)=0 \\ \dfrac{1}{\beta}+kM(x)-\alpha(1-u)M\left(\dfrac{xe^{\beta x}}{1-\alpha u e^{\beta x}}\right)=0 \\ \ln \alpha u+\psi\left(k+\dfrac{1}{u}\right)-\psi(k)+\beta M(x)=0 \\ \psi\left(k+\dfrac{1}{u}\right)-\psi\left(\dfrac{1}{u}\right)+M\left[\ln\left(1-\alpha u e^{\beta x}\right)\right]=0 \end{cases} \tag{10.9}$$

С учетом двух последних уравнений системы (10.9) логарифмическая функция правдоподобия запишется в окончательном виде [29]

$$\ln L = M\left[\ln p(x)\right]=\ln\beta+\ln\frac{\Gamma(\mathrm{k}+1/u)}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})}+k\left[\psi(k)-\psi\left(k+\frac{1}{u}\right)\right]+\left(\frac{1}{u}-1\right)\left[\psi\left(\frac{1}{u}\right)-\psi\left(k+\frac{1}{u}\right)\right] \tag{10.10}$$

Таким образом, логарифмическая функция правдоподобия получена достаточно простым способом без интегрирования.

Из (10.10) следует, что логарифмическая функция правдоподобия зависит от трех параметров: β, k, u. Она может быть вычислена по этой формуле для распределений первого типа не только первой системы непрерывных распределений, но и второй и третьей систем, если их привести к форме плотности p(x), т. е. представить в виде tp(t)=f(lnt), yp(y)lny=f(lnlny).

10.2.1. Оценивание параметров по методу наибольшего правдоподобия

Рассмотрим случайную величину

$$Z=\alpha u e^{\beta X} \tag{10.11}$$

и найдем плотность распределения случайной величины Z по известной формуле *p(z)=p(x)(dx/dz)=p(x)/(dz/dx)*.

Первая производная от z по x равна $dz/dx = \alpha u \beta e^{\beta x}$. Тогда из плотности (10.3) получим

$$p(z) = \frac{\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})} z^{k-1}(1-z)^{\frac{1}{u}-1}, \qquad (10.12)$$

т. е. имеем известное бета-распределение.

Приведем плотность (10.12) к форме плотности (10.3), т. е. представим ее в форме *zp(z)=f(lnz)*:

$$zp(z) = \frac{\Gamma(\mathrm{k}+1/\mathrm{u})}{\Gamma(\mathrm{k})\Gamma(1/\mathrm{u})} e^{k\ln z}\left(1-e^{\ln z}\right)^{\frac{1}{u}-1}.$$

Запишем для последней плотности логарифмическую функцию правдоподобия

$$\ln L_{(z)} = M\left[\ln zp(z)\right] = \ln\Gamma\left(\mathrm{k}+\frac{1}{\mathrm{u}}\right) - \ln\Gamma(\mathrm{k}) - \ln\Gamma\left(\frac{1}{\mathrm{u}}\right) + kM(\ln z) + \left(\frac{1}{u}-1\right)M\left[\ln(1-z)\right]. \qquad (10.13)$$

Найдем уравнения правдоподобия.

$$\begin{cases} \dfrac{\partial \ln L_{(z)}}{\partial k} = \psi\left(k+\dfrac{1}{u}\right) - \psi(k) + M(\ln z) = 0 \\ \dfrac{\partial \ln L_{(z)}}{\partial u} = \psi\left(k+\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right) - \psi\left(\dfrac{1}{u}\right)\left(-\dfrac{1}{u^2}\right) + \left(-\dfrac{1}{u^2}\right)M\left[\ln(1-z)\right] = 0 \end{cases}$$

Из первого уравнения правдоподобия имеем

$$M(\ln z) = \psi(k) - \psi\left(k+\frac{1}{u}\right). \qquad (10.14)$$

Из второго уравнения правдоподобия находим

$$M\left[\ln(1-z)\right] = \psi\left(\frac{1}{u}\right) - \psi\left(k+\frac{1}{u}\right). \qquad (10.15)$$

Логарифмическая функция правдоподобия (10.12) с учетом формул (10.13) и (10.14) перепишется в виде

$$\ln L_{(z)} = M\left[\ln zp(z)\right] = \ln\frac{\Gamma(k+1/u)}{\Gamma(k)\Gamma(1/u)} + k\left[\psi(k) - \psi\left(k+\frac{1}{u}\right)\right] + \left(\frac{1}{u}-1\right)\left[\psi\left(\frac{1}{u}\right) - \psi\left(k+\frac{1}{u}\right)\right] \qquad (10.16)$$

Из сопоставления функций правдоподобия (9) и (15) следует, что первая из них зависит от трех параметров β, k, u, а вторая – только от двух параметров k, u.

На основании формул (10.10) и (10.16) можем записать равенство

$$M[\ln p(x)] = \ln \beta + M[\ln zp(z)], \qquad (10.17)$$

которое справедливо также для других типов распределений.

Равенство (10.17) позволяет вычислять оценку параметра β при известных оценках двух параметров формы k, u

$$\beta = e^{M[\ln p(x)] - M[\ln zp(z)]} = \frac{e^{M[\ln p(x)]}}{e^{M[\ln zp(z)]}} = \frac{\overline{p(x)}_{геом.}}{\overline{\mathrm{zp(z)}}_{геом.}}. \qquad (10.18)$$

Перепишем далее третье уравнение правдоподобия системы (8) с учетом формулы (9.13)

$$\beta M(x) = M(\ln z) - \ln \alpha u.$$

Отсюда найдем оценку произведения αu:

$$\alpha u = e^{M(\ln z) - \beta M(x)}. \qquad (10.19)$$

Входящие в формулы (10.18), (10.19) величины *M(lnzp(z))* и *M(lnz)* зависят от двух параметров формы k, u. Они вычисляются по формулам (10.16) и (10.14). Другие величины – *M(lnp(x))* и *M(x)* – в общем случае зависят от четырех параметров, но эти величины вычисляются по статистическому распределению.

Таким образом, остается невыясненным вопрос о вычислении типа аппроксимирующего распределения и оценок параметров k, u по двум показателям, зависящим от этих параметров. Нахождение таких показателей является наиважнейшей задачей любого метода оценивания. Успешное решение этой задачи позволяет отказаться от выдвижения гипотез о виде аппроксимирующего распределения и проверки каждой из них по критериям со-

гласия. Наличие таких показателей позволяет вычислять наилучшее аппроксимирующее распределение из трех систем непрерывных распределений автора без выдвижения гипотез, легко решать другие задачи.

В случае метода наибольшего правдоподобия такие показатели можно получить из равенства (10.17) в виде центральных моментов второго-четвертого порядков

$$\mu_r = M\left[\ln p(x) - M\left(\ln p(x)\right)\right]^r = M\left[\ln zp(z) - M\left(\ln zp(z)\right)\right]^r, \qquad (10.20)$$

которые зависят от двух параметров формы k, u. Если эти показатели окажутся неподходящими, можно воспользоваться показателями B, H устойчивого метода для вычисления типа аппроксимирующего распределения и нахождения оценок параметров k, u. Оценки остальных двух параметров (β, α или произведения αu) вычисляются по простым формулам, например, в случае распределений первого типа – по формулам (10.18) и (10.19).

10.2.2. Устойчивый метод

Устойчивым называется метод оценивания параметров, который не чувствителен к выбросам на концах статистического распределения.

Рассмотрим плотности (10.3) и (10.11). Случайные величины X и Z связаны соотношением (10.11)

$$Z = \alpha u e^{\beta X}.$$

Базой устойчивого метода является равенство, устанавливающее взаимосвязь между плотностями *p(x)* и *p(z)*. Найдем его.

Поскольку $p(z) = p(x)(dx/dz), \quad dx/dz = 1/\beta z,$ то $p(z) = p(x)/\beta z$, откуда и следует искомое равенство

$$p(x) = \beta z p(z). \qquad (10.21)$$

Запишем на основе (9.20) новые равенства

$$M\left[p(x)\right]^r = \beta^r M\left[zp(z)\right]^r \qquad (10.22)$$

или

$$S_r^{(x)} = \beta^r S_r^{(z)} . \qquad (10.23)$$

При известных оценках параметров k, u оценка параметра ß устойчивого метода задается равенством (при r=1 в формулах (10.22), (10.23))

$$\beta = \frac{M\left[p(x)\right]}{M\left[zp(z)\right]} = \frac{S_1^{(x)}}{S_1^{(z)}} . \qquad (10.24)$$

Здесь математическое ожидание плотности p(x) заменяется средним его значением, которое вычисляется по статистическому распределению.

Логарифмируя равенство (10.21) и переходя к математическим ожиданиям, получим формулу (10.17), которая является базой метода наибольшего правдоподобия. Из формулы (10.17) оценка наибольшего правдоподобия параметра β равна отношению средних геометрических значений величин *p(x)* и *zp(z)*. Отсюда следует, что оценки параметра β, вычисленные по обоим методам, должны быть практически одинаковыми. Оценка параметра α (или произведения αu) вычисляется в обоих методах по одним и тем же формулам.

Итак, устойчивый метод близок к методу наибольшего правдоподобия, но в то же время он значительно проще последнего. Устойчивый метод обладает тем несомненным преимуществом перед методом наибольшего правдоподобия, что для него разработаны два показателя (асимметрии В и островершинности Н), с помощью которых по заранее построенной бинарной сетке (номограмме) легко вычисляются аппроксимирующие распределения и оценки двух параметров формы k, u [14–21]. Для метода наибольшего правдоподобия такие показатели еще предстоит разработать. С другой сторо-

ны, метод наибольшего правдоподобия позволяет легко выражать через параметры распределения математическое ожидание логарифма плотности распределения. Например, для плотности p(x) (см. формулу (10.3)) величина M[lnp(x)] задается формулой (10.10). Эту величину можно использовать как естественный критерий близости статистического распределения и вычисленного закона распределения. По найденным оценкам параметров β, k, u следует вычислить теоретическое значение величины M[lnp(x)] и сравнить его с эмпирическим значением $\overline{\ln p(x)}$, рассчитанным непосредственно по статистическому распределению. Оба значения должны практически совпадать.

Энтропия

В качестве меры неопределенности системы принята энтропия. В случае непрерывных распределений она представляет собой математическое ожидание логарифма плотности, взятое с обратным знаком. Другими словами, энтропия – это взятая с обратным знаком логарифмическая функция правдоподобия (10.2), т. е.

$$H_x = -M[\ln p(x)]. \qquad (10.25)$$

Рассмотрим единицы измерения энтропии. В приведенной формуле логарифм плотности взят по основанию e=2.71828…. В данном случае в качестве единицы измерения энтропии принят «нат». При основании 10 единица измерения энтропии называется «дит», а при основании 2 – «бит».

С энтропией связано понятие информации. Количество полученной информации о системе уменьшает энтропию системы на это количество информации. Если о системе известно все, то количество информации равно энтропии этой системы, т. е. $I_x = H_x$ [6].

Рассмотрим один способ извлечения информации из ранговых распределений. Пусть ранговое распределение задано плотностью

$$p(t) = \frac{\beta\alpha^k}{\Gamma(\mathrm{k})} t^{k\beta-1} e^{-\alpha t^\beta}, \qquad (10.26)$$

которая относится ко второму типу второй системы непрерывных распределений автора. Запишем логарифмическую функцию правдоподобия

$$\ln L = M\left[\ln p(t)\right] = \ln\beta + k\ln\alpha - \ln\Gamma(\mathrm{k}) + (\mathrm{k}\beta - 1)\mathrm{M}(\mathrm{lnt}) - \alpha\mathrm{M}(\mathrm{t}^\beta). \qquad (10.27)$$

Для того чтобы выразить ее через параметры распределения, найдем уравнения правдоподобия:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{k}{\alpha} - M\left(t^\beta\right) = 0;$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\beta} + kM(\ln t) - \alpha M\left(t^\beta \ln t\right) = 0;$$

$$\frac{\partial \ln L}{\partial k} = \ln\alpha - \psi(k) + \beta M(\ln t) = 0.$$

Из первого и последнего уравнений правдоподобия имеем

$$k = \alpha M\left(t^\beta\right), \qquad (10.28)$$

$$M(\ln t) = \frac{1}{\beta}\left[\psi(k) - \ln\alpha\right]. \qquad (10.29)$$

Подставляя значения величин k, M(lnt) в формулу (9.26), найдем

$$\ln L = M\left[\ln p(t)\right] = \ln\beta - \ln\Gamma(\mathrm{k}) + \mathrm{k}\psi(\mathrm{k}) - \mathrm{k} - \mathrm{M}(\mathrm{lnt}). \qquad (10.30)$$

Следовательно, энтропия плотности (9.25) равна

$$H_{p(t)} = -M\left[\ln p(t)\right] = M(\ln t) + \ln\Gamma(\mathrm{k}) + \mathrm{k} - \mathrm{k}\psi(\mathrm{k}) - \ln\beta. \qquad (10.31)$$

Пусть параметры рангового распределения (10.26) равны: α=0,5; β=0,4; k=2. Вычислим его энтропию. Для этого найдем вначале M(lnt) по формуле (10.29)

$M(\ln t)=\frac{1}{\beta}[\psi(k)-\ln\alpha]=\frac{1}{0,4}[\psi(2)-\ln 0,5]=2,789829$. Здесь значение пси-функции Ψ(2)=0,4227843 взято из таблицы [23, с. 53]. Тогда

$H_{p(t)}=-M[\ln p(t)]=2,773655+\ln\Gamma(2)+2-2\psi(2)-\ln 0,4=4,860552$ (нат).

Преобразуем далее плотность (10.26) к форме соответствующей плотности p(x) первой системы непрерывных распределений, т. е. представим плотность p(t) в виде tp(t)=f(lnt), где tp(t)=p(x), lnt=x

$$tp(t)=\frac{\beta\alpha^{k}}{\Gamma(\mathrm{k})}e^{k\beta\ln t}e^{-\alpha e^{\beta\ln t}}. \qquad (10.32)$$

Запишем для нее логарифмическую функцию правдоподобия

$$\ln L=M[\ln tp(t)]=\ln\beta+k\ln\alpha-\ln\Gamma(\mathrm{k})+\mathrm{k}\beta\mathrm{M}(\mathrm{lnt})-\alpha\mathrm{M}\left(\mathrm{e}^{\beta\mathrm{lnt}}\right). \qquad (10.33)$$

Выраженная через параметры распределения, она равна

$$\ln L=M[\ln tp(t)]=\ln\beta-\ln\Gamma(\mathrm{k})+\mathrm{k}\psi(\mathrm{k})-\mathrm{k}. \qquad (10.34)$$

Следовательно, энтропия плотности (9.31) равна

$$H_{tp(t)}=-M[\ln tp(t)]=\ln\Gamma(\mathrm{k})+\mathrm{k}-\mathrm{k}\psi(\mathrm{k})-\ln\beta. \qquad (10.35)$$

Сравнивая это равенство с (10.31), имеем

$$H_{tp(t)}=H_{p(t)}-M(\ln t), \qquad (10.36)$$

т. е. энтропия (степень неопределенности) плотности tp(t)=p(lnt)=p(x) оказалась меньше энтропии плотности p(t) на величину M(lnt)=2,7898287 и составила $H_{tp(t)}=2,0707223$ (нат), т. е. меньше энтропии $H_{p(t)}$ в 2,347 раза.

Таким образом, приведение второй системы непрерывных распределений к форме первой системы уменьшает энтропию второй системы.

Аналогично приведение третьей системы непрерывных распределений к форме второй системы (или первой) также уменьшает их энтропию.

Что касается ранговых (убывающих) распределений, то здесь наиболее ярко проявляется уменьшение энтропии, или появление новой информации в виде трех характерных точек: моды и двух точек перегиба, – которые позволяют объективно разделить ранговое распределение на ядро и три зоны рассеяния [14, 15, 18]. Действительно, убывающее ранговое распределение, будучи представленным в виде графика зависимости tp(t)=f(lnt), превращается в одномодальную кривую распределения с тремя характерными точками, которые нельзя обнаружить непосредственно на убывающей кривой.

Два метода оценивания параметров (универсальный метод моментов и общий устойчивый метод) разработаны автором для первой системы непрерывных распределений, заданной плотностью p(x), а другие системы непрерывных распределений при нахождении оценок параметров по этим методам приводятся к первой системе. Это преобразование уменьшает энтропию распределений второй и третьей систем и позволяет находить оценки их параметров методом, пригодным для первой системы непрерывных распределений.

Действительно, метод моментов не может быть применен непосредственно к плотности (10.26), где значения случайной величины T возводятся в степень β. Но после преобразования той же плотности к виду (10.32) параметр β уже не является степенью случайной величины T, при этом вычисляются моменты не самой случайной величины T, а ее логарифма. Фактически в этом случае находятся оценки параметров плотности

$$p(x) = \frac{\beta\alpha^k}{\Gamma(\mathrm{k})} e^{k\beta x} e^{-\alpha e^{\beta x}},$$

где x=lnt, p(x)=tp(t)=p(lnt) [30]. Найденные оценки являются также оценками параметров исходной плотности p(t), которая задана формулой (10.26).

В заключение следует отметить, что оба метода оценивания параметров четырехпараметрических распределений базируются на одном и том же равенстве

$$p(x) = \beta z p(z),$$

где плотность p(x) зависит от четырех параметров, а плотность p(z) – от двух.

В методе наибольшего правдоподобия используется логарифмическая форма приведенного равенства.

Устойчивый метод позволяет вычислять закон распределения на базе четырехпараметрических систем непрерывных распределений с помощью показателей асимметрии B и островершинности H. Оценки последних находятся по статистическому распределению и зависят от двух параметров формы. Тип распределения и оценки параметров формы находятся по заранее построенной бинарной сетке (номограмме) [25].

В методе наибольшего правдоподобия такие показатели отсутствуют. Поэтому вид аппроксимирующего распределения подбирается традиционным методом – путем выдвижения гипотез.

Логарифмическая функция правдоподобия, заданная формулой $\ln L = M[\ln p(x)]$ и взятая с обратным знаком, представляет собой энтропию.

При преобразовании распределений второй системы к форме первой системы энтропия плотности уменьшается, т.е. появляется новая информация.

В главе 10 исследована близость двух методов оценивания параметров непрерывных распределений – наибольшего правдоподобия Р. Фишера и устойчивого метода автора. Показано, что оба метода базируются на одном и том же равенстве $p(x) = \beta z p(z)$, *устанавли-*

вающем взаимосвязь между обобщенной плотностью p(x) и двухпараметрической плотностью p(z). В методе наибольшего правдоподобия используется логарифм этого равенства. Показывается, что логарифмическая функция правдоподобия, взятая с обратным знаком, представляет собой энтропию распределения. При преобразовании распределений второй системы к форме первой энтропия плотности уменьшается, т. е. появляется новая информация.

11. ПРОГНОЗИРОВАНИЕ РАСПРЕДЕЛЕНИЙ

Знание характерных особенностей каждой системы непрерывных распределений позволяет осуществлять **прогноз распределения**.

Рассмотрим вторую систему непрерывных распределений.

11.1. Вторая система непрерывных распределений

Пусть распределение случайной величины T задано обобщенной плотностью

$$p(t) = Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}$$

с известными оценками параметров. Пусть далее известно, что все значения t_i (i = 1,2,…, n) увеличатся в C раз.

Требуется найти распределение случайной величины $T^*=T\cdot C$.

Поскольку $t=t^*/C$, $dt/dt^*=1/C$, то

$$p(t^*) = p(t)\frac{dt}{dt^*} = \frac{p(t)}{C} \qquad (11.1.1)$$

или

$$p(t^*) = \frac{N}{C^{k\beta}}t^{*k\beta-1}\left(1-\frac{\alpha}{C^{\beta}}ut^{*\beta}\right)^{\frac{1}{u}-1}. \qquad (11.1.2)$$

Введя обозначения

$$N^* = \frac{N}{C^{k\beta}},\ \alpha^* = \frac{\alpha}{C^{\beta}}, \qquad (11.1.3)$$

Плотность (11.1.2) можно переписть в виде

$$p(t^*) = N^* t^{*k\beta-1}\left(1-\alpha^* u t^{*\beta}\right)^{\frac{1}{u}-1}. \qquad (11.1.4)$$

Увеличение случайной величины T в C раз приводит к уменьшению параметра α и нормирующего множителя N. При этом **плотность распределения *p(t)* уменьшается в *C* раз, а произведение *tp(t)*, а также среднее значение** $\overline{tp(t)}$ **остаются без изменения**. Это значит, **что форма кривой распределения** $tp(t) = p(\ln t)$, а также характеризующие ее показатели не изменяются, при этом остается справедливым равенство $F(t^*) = F(t)$.

Таким же путем найдем, что **при увеличении в *C* раз случайной величины *Y* вторая и третья плотности второй системы непрерывных распределений примут вид**

$$p(y^*) = \frac{N}{y^*}\left(\ln y^* - l^*\right)^{k-1}\left[1-\alpha u\left(\ln y^* - l^*\right)\right]^{\frac{1}{u}-1} \qquad (11.1.5)$$

где

$$l^* = l + \ln C, \qquad (11.1.6)$$

$$p(y^*) = \frac{N}{y^*}\left[1-\alpha u\left(\ln y^* - \overline{\ln y^*}\right)^2\right]^{\frac{1}{u}-1}, \qquad (11.1.7)$$

где

$$\overline{\ln y^*} = \overline{\ln y} + \ln C. \qquad (11.1.8)$$

Рассмотрим характерный пример. В табл. 11.1.1. приведены статистические данные о распределении населения СССР по среднедушевому совокупному доходу [Аргументы и факты. № 32. 1989 г.] за 1980, 1985 и 1988 г.

Таблица 11.1.1.

Распределение населения страны по среднедушевому совокупному доходу, в % к итогу (расчет по данным обследования 90 тыс. семейных бюджетов)

Интервал в руб.	Год		
	1980	1985	1988
До 50	7,3	4,3	2,9
50-75	18,5	13,6	9,7
75-100	23,2	19,8	15,7
100-125	19,5	19,3	17,6
125-150	13,2	15,0	15,7
150-175	8,2	10,4	12,2
175-200	4,7	6,7	9,0
200-250	4,1	6,9	10,1
>250	1,3	4,0	7,1
Итого	**100%**	**100%**	**100%**

Найдем выравнивающее распределение по статистическому интервальному ряду за 1980 г. По программе **SNR2V97** имеем: [16]

$$\overline{ln\,t} = 4{,}596399;\ B = 0{,}021588;\ H = 1{,}441297\,.$$

Выравнивающее распределение задается обобщенной плотностью

$$p(t) = \frac{Nt^{\gamma-1}}{(1-\alpha u t^{\beta})^{1-1/u}} \qquad (11.1.9)$$

и относится к III типу. Оценки параметров и нормирующего множителя равны:

$$\alpha u = -5{,}23884\cdot 10^{-4};\ \beta = 1{,}399491;\ \gamma = 5{,}020293;$$

$$u = -0{,}0793593;\ N = 3{,}852922\cdot 10^{-9}.$$

Рассчитаем по формуле (11.1.9) при известных оценках параметров значения плотности *p(t)* в серединах интервалов. Для того, чтобы сравнить выравнивающее и статистическое распределения, умножим значения плотности *p(t)* на ширину интервала (h = 25). Получим относительные частоты интервалов.

Результаты расчетов сведены в табл. 11.1.2 графы 1,2.

Таблица 11.1.2

Расчетные относительные частоты интервалов

Интервал в руб.	Год		
	1980 (выравнивание)	1985 (прогноз)	1988 (прогноз)
0- 25	0,0019	0,0010	0,0006
25- 50	0,0688	0,0418	0,0264
50- 75	0,1871	0,1340	0,0959
75-100	0,2306	0,1952	0,1590
100-125	0,1948	0,1936	0,1789
125-150	0,1339	0,1544	0,1608
150-175	0,0816	0,1079	0,1257
175-200	0,0464	0,0695	0,0897
200-225	0,0254	0,0425	0,0602
225-250	0,0136	0,0252	0,0389
250-275	0,0072	0,0146	0,0245
275-300	0,0039	0,0084	0,0152
300-325	0,0021	0,0049	0,0093

Найдем далее по данным таблицы 8.3.1 средние значения логарифмов среднедушевого дохода за 1985 и 1988 гг. Они оказались равными соответственно 4,732163 и 4,850433, откуда находим, что среднедушевой доход в 1985 г. вырос в среднем в 1,145358 раза, а в 1988 г. – в 1,289216 раза по сравнению с 1980 годом.

Чтобы спрогнозировать распределение среднедушевого дохода населения на 1985 и 1988 г. по его распределению за 1980 г., достаточно вычислить новые значения нормирующего множителя *N* и произведения *αu* (см. формулы (8.3.13), где $k\beta = \gamma$):

$N^* = N / C^{\gamma}$; $\alpha^* u = \alpha u / C^{\beta}$.

При C = 1,145358 они равны:

$N^* = 1{,}94934 \cdot 10^{-9}$; $\alpha^* u = -4{,}332588 \cdot 10^{-4}$.

При C = 1,289216 они равны:

$N^* = 1{,}076274 \cdot 10^{-9}$; $\alpha^* u = -3{,}671431 \cdot 10^{-4}$.

Оценки параметров ß, γ, *u* – те же, что и в выравнивающем распределении дохода за 1980 г.

В табл. 11.1.2 приведены расчетные значения относительных частот интервалов (прогноз) на 1985 и 1988 г. (графы 3, 4).

Ожидаемые распределения на 1985 и 1988 г. можно также получить с помощью Программы по статистическому распределению за 1980 г., если середину и ширину каждого интервала увеличить в *C* раз, оставив без изменения частоту (долю) интервала, что вытекает из формулы (8.3.11).

Из нее же следует, что на базе статистического интервального ряда распределения населения по среднедушевому совокупному доходу за 1980 г. (см. табл. 11.1.1) можно построить новый (ожидаемый) интервальный ряд распределения с учетом коэффициента роста С. **Для этого достаточно увеличить в *C* раз границы, середину и ширину всех интервалов, оставив без изменения долю интервалов. Тогда значения эмпирической плотности в связи с ростом ширины интервалов уменьшатся в *C* раз**.

Полученное распределение и будет ожидаемым на некоторый период упреждения *r*, когда среднедушевой совокупный доход вырастет в *C* раз по сравнению с 1980-м годом.

На рис. 11.1.1 (а, б, в) представлены гистограммы, построенные по статистическим данным, и непрерывные кривые $m(t) = p(t) \cdot h$ при h=25 (на рис. а **– выравнивающая кривая**, полученная непосредственно по статистическому интервальному ряду за 1980 г.; **на рис. б, в** –

прогнозируемые непрерывные кривые). Расчетные и статистические данные находятся в хорошем согласии между собой.

Чтобы оценить уровень жизни населения, необходимо сопоставить приведенные интервальные ряды распределения (эмпирические или расчетные) с **минимальным потребительским бюджетом** в одни и те же моменты времени.

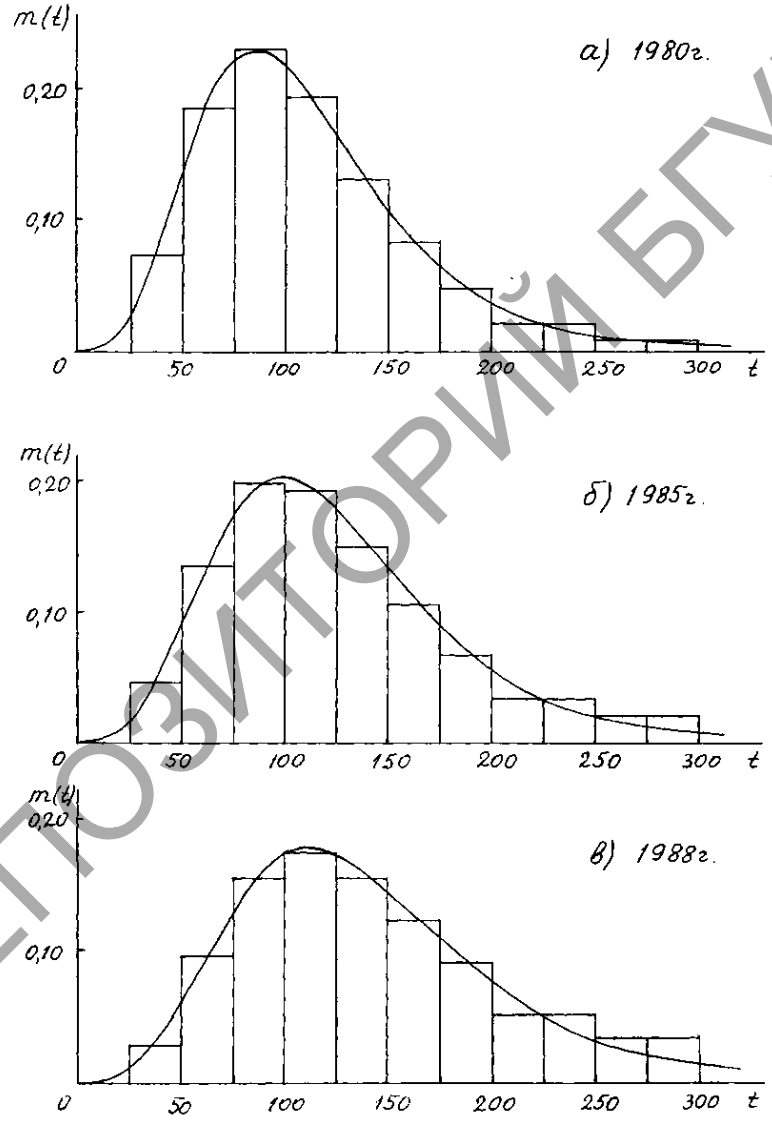


Рис. 11.1.1. Статистические (представлены гистограммами) и теоретические распределения среднедушевого дохода населения (на **рис. а изображено выравнивающее непрерывное распределение**; на **рис. б, в – прогнозируемые распределения**)

Знание закона распределения совокупного дохода позволяет также осуществлять различные расчеты, в том числе давать **прогноз**.

Выше отмечалось, что с ростом величины t в C раз ($t^* = t \cdot C$) функция распределения не изменяется, т. е. $F(t^*) = F(t)$.

Это дает возможность прогнозировать долю населения с заданным совокупным доходом, например, до одного минимального потребительского бюджета (МПБ), двух МПБ, S МПБ.

Пусть **совокупный доход** населения (СД) растет по **показательному закону** с темпом роста $Q_{СД}$ в единицу времени r (год, месяц), а МПБ – с темпом роста $Q_{МПБ}$.

Пусть далее доля населения с совокупным доходом $СД_0$ на момент времени t_0 равна $F(t_0) = \alpha$.

Минимальный потребительский бюджет на тот же момент времени равен $МПБ_0$.

Найдем период времени r, через который будет выполняться равенство (11.1.10) [52].

$$S \cdot МПБ = СД \ . \qquad (11.1.10)$$

Через r единиц времени S-кратный МПБ будет равен

$$S \cdot МПБ = S \cdot МПБ_0 \cdot Q_{МПБ}^r , \qquad (11.1.11)$$

а совокупный доход равен

$$СД = СД_0 \cdot Q_{СД}^r \ . \qquad (11.1.12$$

Приравнивая два последних равенства, найдем

$$r = \frac{ln \frac{S \cdot МПБ_0}{СД_0}}{ln \frac{Q_{СД}}{Q_{МПБ}}} \ . \qquad (11.1.13)$$

Проанализируем полученную формулу.

Пусть **темп роста совокупного дохода превышает темп роста минимального потребительского бюджета**. Тогда знаменатель в формуле (11.1.13) будет больше

нуля, а величина r будет равна периоду времени, через который доля населения α будет иметь совокупный доход, равный S-кратному минимальному потребительскому бюджету. При **этом структура совокупности изменяется в лучшую сторону, т. е. уменьшается доля населения с низким совокупным доходом**.

При равенстве темпов роста совокупного дохода и минимального потребительского бюджета величина $r = \infty$, т. е. структура совокупности не улучшается (не уменьшается доля населения с низким совокупным доходом).

Для стабильного улучшения распределения совокупного дохода необходимо, чтобы темп роста совокупного дохода (или темп прироста, т. е. темп роста, уменьшенный на единицу) **постоянно опережал темп роста (или прироста) минимального потребительского бюджета**.

Закон распределения совокупного дохода, а также заработной платы – это зеркало экономики.

В совокупности с законами роста минимального потребительского бюджета и среднемесячной заработной платы он представляет собой весьма чувствительный инструмент, позволяющий оценивать состояние на данный момент времени и прогнозировать различные экономические явления и процессы.

Отметим, что этот вопрос более детально был в свое время рассмотрен автором в статье «Анализ распределения и динамики заработной платы в строительстве», опубликованной в газете «Строительство и недвижимость от 14.03.2000г, с. 4». На основе анализа сложившейся на то время ситуации с заработной платой в строительстве мною было рассчитано, что доля работающих с заработной платой ниже одного минимального потребительского бюджета снизится в строительстве до 20% через 18 лет, т. е. в 2018г.

11.2. Показатели стабильности и качества выборки

Для уверенного прогнозирования структуры выборки, которое осуществляется посредством прогнозирования выравнивающей кривой распределения, полученной в некоторый базовый момент времени, необходимо иметь оценку степени стабильности ранжированного статистического ряда распределения. Прогнозирование можно осуществлять лишь при условии неизменности порядка следования элементов в ранжированном (вариационном) ряду, т. е. при неизменности рангов элементов с течением времени.

Действительно, увеличение отдельных значений x_i на постоянную величину C или умножение их на ту же величину не меняет порядок их следования в вариационном ряду и, как было показано ранее, не приводит к изменению параметров формы выравнивающих кривых.

Следовательно, для измерения степени связи двух ранжированных рядов распределения, а точнее, одного и того же ряда, но в разные моменты времени, может быть использован коэффициент ранговой корреляции Спирмена [37]

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \qquad (11.2.1)$$

где di – разность между значениями рангов одного и того же элемента выборки в двух ранжированных рядах; n – объем выборки.

Коэффициент ранговой корреляции Спирмена может служить показателем степени стабильности вариационных рядов. Успешное прогнозирование структуры выборки посредством выравнивающих распределений возможно лишь при условии, когда коэффициент ранговой корреляции Спирмена близок к единице.

Рассмотрим далее показатели качества выборки. Одним из таких показателей может служить величина S1, которая в зависимости от системы непрерывных распределений задается формулами: [16]

$$S_1^{(x)} = \overline{p(x)}; \quad S_1^{(t)} = \overline{tp(t)}; \quad S_1^{(y)} = \overline{yp(y)\ln y} \qquad (11.2.2)$$

Чем выше показатель S1, тем лучше качество выборки. Действительно, пусть в случае второй системы непрерывных распределений величина ti обозначает срок службы однородных изделий в i-ом интервале, либо наработку до отказа и т. д., а pi – долю изделий с данным сроком службы. Тогда качество изделий будет тем выше, чем большая их доля имеет наибольший срок службы.

Вместо величины S1 можно также использовать произведение $\gamma_1\varsigma_1$, которое в зависимости от системы непрерывных распределений задается формулами:

$$\nu_1 S_1^{(x)} = \bar{x} \cdot \overline{p(x)}; \ \nu_1 S_1^{(t)} = \overline{\ln t} \cdot \overline{tp(t)}; \ \nu_1 S_1^{(y)} = \overline{\ln\ln y} \cdot \overline{yp(y)\ln y} \qquad (11.2.3)$$

Введем еще один показатель, который обозначим R(P), например, R(P=0,9), где P – вероятность попадания случайной величины на заданный интервал, ограниченный верхним и нижним уровнями. В зависимости от системы непрерывных распределений этот показатель задается формулами:

$$R^{(x)}(P) = x_B - x_H, \qquad (11.2.4)$$

$$R^{(t)}(P) = \ln t_B - \ln t_H, \qquad (11.2.5)$$

$$R^{(y)}(P) = \ln\ln y_B - \ln l \qquad (11.2.6)$$

(Вместо показателей (11.2.5), (11.2.6) могут использоваться показатели

$$R^{(t)}(P) = \frac{t_в}{t_н}, \qquad (11.2.7)$$

$$R^{(y)}(P)=\frac{\ln y_в}{\ln y_н} \qquad (11.2.8)$$

С улучшением качества выборки показатели (11.2.4) – (11.2.8) уменьшаются.

Показатель R(P) непосредственно связан еще с одним показателем качества выборки – дисперсией случайных величин $X, \ln T, \ln\ln Y$:

$$D(X) = M\left(X - \bar{x}\right)^2, \qquad (11.2.9)$$

$$D(\ln T) = M\left(\ln t - \overline{\ln t}\right)^2, \qquad (11.2.10)$$

$$D(\ln\ln Y) = M\left(\ln\ln y - \overline{\ln\ln y}\right)^2. \qquad (11.2.11)$$

Дисперсия также уменьшается с улучшением качества выборки.

С помощью приведенных показателей можно отслеживать динамику качественных изменений статистических распределений исследуемых экономических и других показателей.

12. СТАТИСТИЧЕСКИЙ АНАЛИЗ ТОЧНОСТИ И СТАБИЛЬНОСТИ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССОВ НА БАЗЕ ОБОБЩЕННЫХ РАСПРЕДЕЛЕНИЙ

Статистический анализ технологических процессов требуется для решения различных задач, в том числе [см. 51 – Методика организации внедрения статистических методов контроля качества продукции на промышленном предприятии. – М.: Изд-во стандартов, 1977. – 40 с.]:

– статистической оценки технологической точности производственного оборудования во время эксплуатации, перед сдачей в ремонт, после ремонта, при подготовке к внедрению статистического регулирования и в других необходимых случаях;

– статистической оценки технологической точности нового оборудования;

– корректировки конструкторского допуска на основе статанализа результатов испытаний опытных образцов и серийных изделий;

– устранения несоответствия между заданной точностью и возможностями реального технологического процесса;

– контроля качества механических свойств металлов;

– установления необходимости ремонта оборудования;

– определения качества выполненного ремонта оборудования;

– сравнительной оценки точности вариантов технологического процесса, оборудования и оснастки;

– сравнительной оценки точности режимов обработки; еще следует добавить:

– оценки эффективности управляющих воздействий.

Статистический анализ заключается в выявлении закона распределения производственных погрешностей при производстве продукции, нахождении оценок его параметров и вычислении необходимых показателей, характеризующих состояние технологического процесса.

Для установления закона распределения контролируемого параметра необходимо отобрать не менее 100 единиц продукции (по ряду мгновенных выборок при неизменной настройке технологического процесса) и измерить значения контролируемого параметра. Далее следует выбрать подходящую систему непрерывных распределений и по соответствующей программе найти выравнивающее распределение и оценки его параметров.

Найденный закон **распределения случайной величины является наиболее полной ее характеристикой**. Более того, он позволяет рассчитывать показатели состояния технологического процесса (при условии его статистической управляемости, когда устранены грубые отклонения от нормы).

Одним из таких показателей является **показатель точности процесса** (коэффициент рассеяния). Он вычисляется по формуле [50, с. 32,52, с. 6, 53].

$$K_T = \frac{\omega}{\delta} = \frac{X_в - X_н}{T_в - T_н}, \qquad (12.1.1)$$

где $\omega = X_в - X_н$ – ширина **поля рассеяния (технологический допуск)**; X_B, X_H – верхняя и нижняя **границы поля рассеяния**; д = $T_B - T_H$ –ширина **поля допуска** (**конструкторский допуск**); T_B, T_H – верхняя и нижняя **гра-**

ницы поля допуска. В зарубежной литературе используется индекс $C_p = 1/K_T$.

Ширина поля рассеяния вычисляется при условии, что 99,73% значений контролируемого параметра находятся внутри границ поля рассеяния. Для нормального закона щ = 6S, где S – выборочное среднее квадратическое отклонение.

При использовании обобщенных распределений ширину поля рассеяния будем определять из того же условия, т. е. $P=F(x_B)-F(x_H)=0{,}9973$, при этом значения функции распределения

$F(X_H)=0{,}00135$, $F(X_B)=1-0{,}00135=0{,}99865$.

Чем меньше ширина поля рассеяния, тем точнее технологический процесс. Однако один показатель меры точности – коэффициент $K_т$ – не в полной мере характеризует технологический процесс.

В случае, когда центр статистического распределения $\bar{x}$ смещен относительно середины поля допуска T_0, т. е. имеются систематические погрешности, процесс может не обеспечивать изготовление бездефектной продукции.

Систематические погрешности характеризуются **коэффициентом смещения** (точности настройки), или **показателем уровня настройки** [52, с.6]

$$K_н = \frac{E}{\delta} = \frac{\left|\bar{X} - T_0\right|}{T_в - T_н}, \qquad (12.1.2)$$

где $\bar{x}$ – среднее выборочное значение контролируемого параметра; $T_0=(T_в+T_н)/2$ – середина поля допуска.

В заключение статанализа вычисляется предполагаемый уровень брака q, выраженный в процентах. Он находится по формуле (см. рис. 12.1.1)

$$q = 100\% - \left[F\left(T_в\right) - F\left(T_н\right)\right]\cdot 100\%. \qquad (12.1.3)$$

При этом брак на нижней границе поля допуска равен

$$q_н = F\left(T_н\right)\cdot 100\%, \qquad (12.1.4)$$

а на верхней границе

$$q_в = [1 - F(T_в)] \cdot 100\% . \qquad (12.1.5)$$

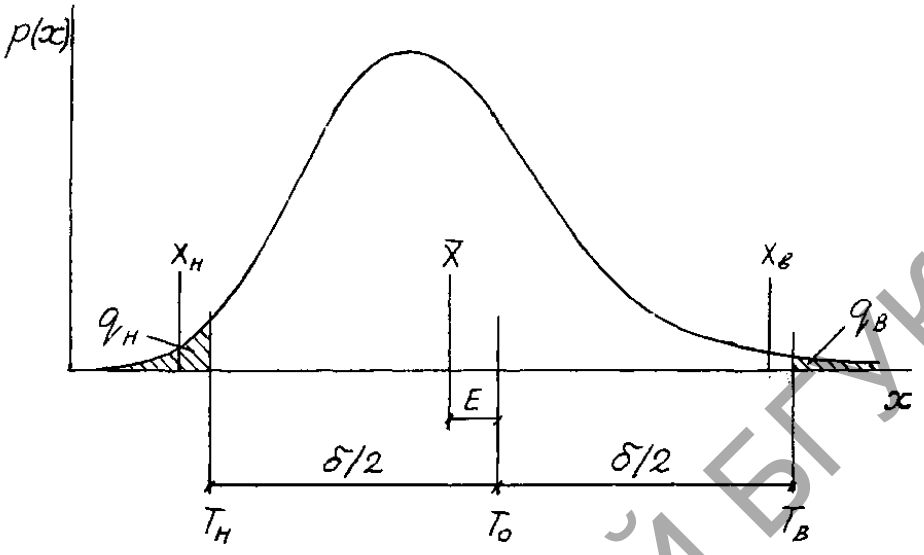


Рис. 12.1.1. Показатели состояния технологического процесса

Отметим, что при условии, когда показатель уровня настройки $K_Н$=0 (при этом E = 0), а показатель точности $K_Т$ = 1, предполагаемый процент брака $q = q_н + q_в$ не превышает 0,27%. С ростом $K_Т$ он увеличивается.

В большинстве случаев достаточно обеспечить $K_Т \leq 0{,}75$. Для ответственных технологических процессов необходимо иметь $K_Т = 0{,}4 \div 0{,}6$ [50, с. 32].

Закон распределения контролируемого показателя качества наиболее полно характеризует технологический процесс.

Если за некоторое время закон распределения не изменился, т. е. не изменились его параметры, например, центр распределения и среднее квадратическое отклонение, или изменились в допустимых пределах, **то процесс считается стабильным.**

Изменение значений центра распределения и среднего квадратического отклонения за допустимые

пределы требует переналадки процесса, так как он утратил стабильность.

Нестабильность технологического процесса по уровню настройки принято характеризовать коэффициентом смещения настройки [49, с. 96]

$$K_{см} = \frac{\overline{X}(t) - \overline{X}_0}{\delta}, \qquad (12.1.6)$$

где $\overline{X}_0, \overline{X}(t)$ – начальное и конечное (на момент времени t) значения центра распределения.

Нестабильность технологического процесса по рассеянию характеризуют коэффициентом межналадочной стабильности [49, с. 97]

$$K_{м.с.} = \frac{S(t)}{S_0}, \qquad (12.1.7)$$

где $S_0, S(t)$ – начальное и конечное (на момент времени t) значения среднего квадратического отклонения.

Поскольку статистические распределения часто имеют асимметрию, то показатель уровня настройки целесообразно вычислять не по формуле (12.1.2), в которую входит центр распределения $\overline{X}$, а по формуле

$$K_н = \frac{T_0 - X_0}{T_в - T_н} = \frac{a}{T_в - T_н}, \qquad (12.1.8)$$

где

$$a = T_0 - X_0, \qquad (12.1.9)$$

T_0, X_0 – середины полей допуска и рассеяния; они вычисляются по формулам

$$T_0 = \frac{T_в + T_н}{2}; \; X_0 = \frac{X_в + X_н}{2}. \qquad (12.1.10)$$

С учетом (12.1.10) формула (12.1.8) может быть представлена в виде

$$K_н = \frac{(T_в + T_н) - (X_в + X_н)}{2(T_в - T_н)}. \qquad (12.1.11)$$

Рассмотрим случай, когда коэффициент точности $K_T \le 1$.

Пусть $T_H = X_H$, т. е. **нижние границы конструкторского и технологического допусков совпадают**. При этом условии на нижней границе конструкторского допуска брак не будет превышать допустимого значения. Формула (12.1.11) примет вид

$$K_H = \frac{T_B - X_B}{2(T_B - T_H)} = \frac{T_B - T_H + T_H - X_B}{2(T_B - T_H)}.$$

Но $T_H = X_H$, следовательно,

$$K_H = \frac{(T_B - T_H) - (X_B - X_H)}{2(T_B - T_H)} = \frac{1}{2}(1 - K_T).$$

Пусть далее $T_B = X_B$. В этом случае на основании (12.1.11) имеем

$$K_H = \frac{T_H - X_H}{2(T_B - T_H)} = \frac{X_B - X_H + T_H - X_B}{2(T_B - T_H)}.$$

Заменяя X_B на T_B, получим

$$K_H = \frac{(X_B - X_H) - (T_B - T_H)}{2(T_B - T_H)} = \frac{1}{2}(K_T - 1) = -\frac{1}{2}(1 - K_T).$$

Следовательно, **условие, при котором брак не превышает допустимого уровня, задается неравенством**

$$-\frac{1}{2}\left(1 - K_T\right) \le K_H \le \frac{1}{2}\left(1 - K_T\right),$$

или

$$0 \le \left|K_H\right| \le \frac{1}{2}\left(1 - K_T\right). \qquad (12.1.12)$$

Для регулирования ТП необходимо установить исправленное среднее значение контролируемого параметра, которое вычисляется по формуле

$$\overline{X}_{испр} = \overline{X} + a. \qquad (12.1.13)$$

Последняя формула следует из равенства

$T_0 = X_0 + a$ (см. см. формулу (12.1.9)

При $K_T > 1$ необходимо принять все меры для уменьшения поля рассеяния контролируемого параметра.

Установим связь между объемом выборки n и показателями K_T, K_H.

Для того **чтобы гарантировать выпуск бездефектной продукции, необходимо, чтобы поле конструкторского допуска было шире поля рассеяния как минимум на величину** $6\sigma_{\bar{x}} = 6\sigma_x / \sqrt{n}$, где $\sigma_{\bar{x}}$ – среднее квадратическое отклонение среднего арифметического.

Итак, пусть

$$T_в - T_н = X_в - X_н + \frac{6\sigma_x}{\sqrt{n}}. \qquad (12.1.14)$$

Тогда **коэффициент точности в общем случае будет задаваться формулой**

$$K_T \le \frac{X_в - X_н}{X_в - X_н + \frac{6\sigma_x}{\sqrt{n}}}. \qquad (12.1.15)$$

В частном случае, **если случайная величина X распределена по нормальному закону**, из (12.1.15) имеем

$$K_T = \frac{6\sigma_x}{6\sigma_x + \frac{6\sigma_x}{\sqrt{n}}} = \frac{1}{1 + \frac{1}{\sqrt{n}}} = \frac{\sqrt{n}}{\sqrt{n}+1}. \qquad (12.1.16)$$

Тогда **коэффициент уровня настройки** на основании (12.1.12) и (12.1.16) **будет равен**

$$0 \le |K_H| \le \frac{1}{2(\sqrt{n}+1)}. \qquad (12.1.17)$$

При $n = 100$ имеем: $K_T \le 0{,}909;\ 0 \le |K_H| \le 0{,}04545$.

При $n = 25$ $K_T \le 5/6 = 0{,}833;\ 0 \le |K_H| \le 0{,}0833$.

При $n = 9$ $K_T \le 0{,}75;\ 0 \le |K_H| \le 0{,}125$.

При этих значения K_T и K_H брак не превышает предельного уровня.

Из формулы (12.1.14) можно найти требуемое значение среднего квадратического отклонения (в случае нормального закона)

$$\sigma_x = \frac{T_в - T_н}{6\left(1 + \frac{1}{\sqrt{n}}\right)}, \qquad (12.1.18)$$

а также требуемое значение допуска при заданном y_x

$$T_в - T_н = 6\sigma_x\left(1 + \frac{1}{\sqrt{n}}\right) \qquad (12.1.19)$$

ЗАКЛЮЧЕНИЕ

Эффективность статистических методов в теоретических и прикладных исследованиях в решающей степени зависит от точности аппроксимации статистических распределений. Наибольшую точность аппроксимации можно получить при использовании теории обобщенных распределений автора, некоторые сведения о которой изложены выше. Кроме того, теория включает также систему дискретных распределений [24], взаимосвязанную с системой кривых роста новых событий, универсальный метод моментов, номограммы для графического определения типа аппроксимирующей кривой и оценок параметров для обоих методов оценивания параметров, а также серию компьютерных программ для работы с указанными системами. Применение этой теории на практике значительно облегчает задачу нахождения закона распределения по статистическим данным.

В этом случае нет необходимости выдвигать гипотезы о предполагаемом аппроксимирующем распределении. В зависимости от свойств случайной величины выбирается система непрерывных распределений (как правило, первая или вторая) и по статистическому распределению вычисляются два показателя – асимметрии B и островершинности H по формулам, справедливым для данной системы. Далее они приравниваются соответствующим теоретическим показателям, которые зависят лишь от двух параметров формы k, u, хотя обобщенная плотность содержит как правило четыре параметра. С помощью номограммы по двум показателям

B, H устанавливается тип теоретического распределения и находятся в первом приближении оценки параметров k, u – в ручном режиме, либо более точные их значения, а также параметров α, β – в автоматизированном режиме по программам автора.

Следует отметить, что одной точке на номограмме с заданными значениями показателей B, H и параметров формы k, u соответствует не единственное распределение, а множество распределений с различными значениями параметров α, β. Например, во второй системе распределений одинаковые значения параметров формы k=1, u→0 имеют такие распределения, как показательное (β=1), Релея (β=2) и Вейбулла (β>0). Оценки параметров α, β вычисляются по специальным формулам с учетом найденных оценок параметров формы k, u.

Обобщенные распределения включают как частные случаи множество известных распределений, в том числе семейство кривых К. Пирсона, и могут претендовать на роль универсальных законов распределения не только теории вероятностей и математической статистики, но и информатики, математической лингвистики, библиотековедения, библиометрии, информетрии, экономики, социологии и других отраслей знания. Их применение и общего устойчивого метода вычисления закона распределения по статистическим данным гарантирует высокую экономическую эффективность статистических методов во всех практических приложениях. Так, использование обобщенных распределений в системах менеджмента качества позволяет с высокой точностью оценивать возможности технологических процессов и поддерживать их в статистически управляемом состоянии при любом законе распределения технологических погрешностей, что обеспечивает значительное снижение уровня брака.

Но широкое использование теории обобщенных распределений можно реализовать лишь путем включения ее в учебные программы высших учебных заведений, в том числе гуманитарных, что обещает дать большой экономический эффект во всех областях ее применения.

На базе введенных автором понятий «нового события», «кривой роста новых событий», «законов распределения вероятностей новых событий» и установленных взаимосвязей между ними был разработан алгоритм порождения кривых роста и законов распределения вероятностей новых событий, который позволил построить систему кривых роста и систему непрерывных распределений новых событий.

Путем дальнейшего обобщения законов распределения вероятностей новых событий были построены четырехпараметрические обобщенные распределения, сгруппированные в четыре основные и три дополнительные системы непрерывных распределений. Они включают как частные случаи множество широко известных непрерывных распределений.

К сожалению, следует отметить, что в большинстве случаев в системах менеджмента качества продукции ипользуется нормальный закон, а не универсальные четырехпараметрические распределения, что наносит огромный урон производству, потому что нормальный закон часто показывает брак, до 10 раз превышающий фактический, что вынуждает инженеров по качеству вносить корректировки в технологический процесс, которые далеко не всегда приводят к уменьшению брака. Использование обобщенных распределений в системах менеджмента качества освободило бы инженеров по качеству от ненужных хлопот и придало бы больше оптимизма и творчества в их деятельности. Но здесь дело в том, что высокие чиновники от науки считают нецелесообразным внедрять в учебный процесс в технических

вузах Теорию обобщенных распределений, в которой предлагается не выдвигать гипотезы об аппроксимирующем распределении и проверять их по критериям согласия, а вычислять теоретические распределения на основании свойств случайных величин.

По системе кривых роста новых событий с помощью формулы В. М. Калинина, которая устанавливает взаимосвязь между кривой роста новых событий и частотным спектром, построена система дискретных распределений. Для ее построения предварительно необходимо было найти бесконечно дифференцируемую кривую роста новых событий. И такая кривая роста с двумя параметрами α, u автором теории обобщенных распределений была найдена. Она включает как частные случаи дискретные распределения 3-х типов: 1-й тип – это биномиальный закон с параметром u>1, Пуассона (u→1) и отрицательный биномиальный (0<u<1); 2-й тип–распределение Фишера по логарифмическому ряду (параметр $u \to 0$), а также новый закон 3-го типа с параметром u<0. Дана классификация распределений, исследована форма полигона [24] распределения в зависимости от значений параметров, разработаны методы нахождения оценок параметров. Показано, что закон Лотки является следствием свойств дискретного закона распределения 3-го типа В. Нешитого. Таким образом дано теоретическое обоснование закона Лотки. В то же время **показано, что закона Ципфа в формулировке его автора rp_r = const не существует вовсе.**

Установлено и проверено на практике, что **параметр u может служить показателем степени неравномерности появления отдельного события в выборках одинакового объема и, следовательно, показателем степени семантической нагрузки слова** [24, с. 58–68]. Для **семантически нагруженных слов параметр u дискретного распределения меньше единицы**

Система дискретных распределений в совокупности с системой кривых роста новых событий позволяет прогнозировать рост новых событий, а также рассчитывать частотный спектр на выборке любого объема, что было бы невозможно при использовании для этих целей отдельных дискретных распределений, известных в теории вероятностей, которые не содержат двух параметров α, u, а также объема выборки x.

Для выравнивания и прогнозирования различного рода кривых роста и динамических рядов построены системы кривых роста, описаны методы оценивания параметров, вычисления доверительных интервалов при заданной доверительной вероятности с учетом свойств кривых роста.

Найдены простые приближенные формулы для описания кривых роста новых слов в связном тексте и случайной выборке, формулы для выравнивания динамических рядов. Получена эмпирическая формула для описания кривой роста простых чисел, которая с весьма высокой точностью (до 0,2%) аппроксимирует количество простых чисел p(X) при 1000<X<100.000.000, т. е. при X от одной тысячи до ста миллионов! При X=100 погрешность формулы составляет 0,4%.

Выработаны показатели для оценки степени связности слов в тексте, степени аналитичности языка, степени лексической близости двух связных текстов, причем эти показатели не зависят от размеров текста. Получены простые формулы для вычисления полноты словаря.

Рассмотрены примеры применения кривых роста в теории надежности.

Для каждой системы распределений (непрерывных и дискретных), а также кривых роста автором разработаны соответствующие программы. Они вычисляют тип наилучшей аппроксимирующей кривой, выдают ее

уравнение и точечные оценки параметров, вычисляют значения плотности и функции распределения, квантили, процентили, доверительные вероятности и доверительные интервалы, координаты моды и точек перегиба, строят кривую распределения (или кривую роста), вычисляют показатели качества продукции, в том числе – ожидаемый процент брака, а также решают другие задачи.

Обобщенные распределения и кривые роста, а также методы оценивания параметров, доведенные до программной реализации, значительно увеличивают вероятность правильного вычисления типа аппроксимирующей кривой. Это позволяет на более высоком уровне точности моделировать различного рода статистические закономерности в различных областях знания, в том числе в библиотечно-информационной деятельности, наукометрии, информетрии, квалиметрии, эконометрии, в системах менеджмента качества, а также решать широкий класс других задач, связанных со статистической обработкой данных, их анализом и прогнозированием.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Белоногов, Г. Г. О некоторых статистических закономерностях в русской письменной речи / Г. Г. Белоногов // Вопросы языкознания. –1962. – № 1. – С. 100–101.

2. Вентцель, Е. С. Теория вероятностей / Е. С. Вентцель. – М. : Наука, 1969. – 576 с.

3. Герасимович, А. И. Математическая статистика / А. И. Герасимович, Я. И. Матвеева. – Минск : Выш. шк., 1978. – 200 с.

4. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высш. шк., 1977. – 479 с.

5. Гуров, С. П. П. Л. Чебышев / С. П. Гуров, Н. А. Хромиенков, К. В. Чебышева. – М. : Просвещение, 1979. – 111 с.

6. Калинин, В. М. Некоторые статистические закономерности математической лингвистики / В. М. Калинин // Проблемы кибернетики. – М., 1964. – Вып. 11. – С. 246–255.

7. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. – М. : ФИЗМАТЛИТ, 2006. – 816 с.

8. Ляшевская, О. Н. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) [Электронный ресурс] / О. Н. Ляшевская. – Режим доступа : http ://dict.ruslang.ru/ freq.php.

9. Михайлов, А. И. Основы информатики / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М. : Наука, 1968. – 756 с.

10. Михайлов, А. И. Научные коммуникации и информатика / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М. : Наука, 1976. – 436 с.

11. Мицевич, Т. А. Исследование структуры потоков научно-технической информации по машиностроению / Т. А. Мицевич // НТИ. Сер. 2. – 1975. – № 5. – С. 3–16.

12. Надежность технических систем : справочник / Ю. К. Беляев [и др.] / под ред. И. А. Ушакова. – М. : Радио и связь, 1985. – 608 с.

13. Нешитой, В. В. Законы распределения слов в тексте и его лексическая параметризация : дис. ... канд. филол. наук / В. В. Нешитой. – Минск, 1973. – 135 л.

14. Нешитой, В. В. Исследование статистических закономерностей текста и информационных потоков : дис. … д-ра техн. наук / В. В. Нешитой. – Минск, 1987. – 505 л.

15. Нешитой, В. В. Статистические модели в биологии / В. В. Нешитой // Кибернетика. – 1987. – № 6. – С. 91–96.

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

17. Нешитой, В. В. Математические модели роста словаря и информационных потоков / В. В. Нешитой // Ученые записки Тартус. гос. ун-та. – Тарту, 1989. – Вып. 872. – С. 83–102.

18. Нешитой, В. В. Статистический анализ и регулирование технологических процессов на базе обобщенных распределений с параметром сдвига : метод. рекоменд. / В. В. Нешитой. – Минск : БелГИСС, 2000. – 38 с.

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

20. Нешитой, В. В. Универсальные законы рассеяния и старения публикаций / В. В. Нешитой // Весн. Беларус. дзярж. ун-та культуры і мастацтваў. – 2007. – № 8. – С. 128–133.

21. Нешитой, В. В. Моделирование кривой роста и статистической структуры словаря ключевых слов / В. В. Нешитой // Весн. Беларус. дзярж. ун-та культуры і мастацтваў. – 2008. – № 9. – С. 123–132.

22. Нешитой, В В. Статистическое моделирование библиотечного фонда / В. В. Нешитой // НТБ. – М., 2009. – С. 36–46.

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

24. Нешитой, В. В. Математико-статистические методы анализа в библиотечно-информационной деятельности : учеб.-метод. пособие / В. В. Нешитой. – Минск : БГУКИ, 2009. – 203 с.

25. Нешитой, В.В. Законы Ципфа, Бредфорда и универсальные модели / В. В. Нешитой // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2010. –

№ 1. – С. 26–33 ; Neshitoi V. V. Zipfŕs and Bradfordŕs Laws and Universal Models / V. V. Neshitoi // Automatic Documentation and Mathematical Linguistics. – 2010. – Vol. 44, № 1. – P. 30–37.

26. Нешитой, В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений : учеб.-метод. пособие / В. В. Нешитой. – Минск : БГУКИ, 2010. – 61 с.

27. Нешитой, В. В. Методы статанализа в библиотечно-информационной деятельности : вычисление дискретных распределений и кривых роста : учеб.-метод. пособие / В. В. Нешитой. – Минск : РИВШ, 2012. – 134 с.

28. Нешитой, В. В. Графический метод вычисления границ ядра и зон рассеяния книжного фонда / В. В. Нешитой // Кніга ў фарміраванні духоўнай культуры і дзяржаўнасці беларускага народа : XVIII Міжнар. Кірыла-Мяфодзіеўскія чытанні, прысвечаныя Дням славянскага пісьменства і культуры, Мінск, 16–18 мая 2012 г. : у 2 т. / Беларус. дзярж. ун-т культуры і мастацтваў ; рэдкал.: М. А. Мажэйка (адказ. рэд.) [і інш.]. – Мінск, 2012. – Т. 1 : Кніга як грамадская з’ява і аснова духоўнасці. – С. 252–257.

29. Нешитой, В. В. Метод наибольшего правдоподобия, устойчивый метод и энтропия / В. В. Нешитой // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2012. – № 5. – С. 27–33.

30. Нешитой, В. В. Статистические методы анализа использования библиотечного фонда / В. В. Нешитой, Б. В. Петренко. – Вестн. Библиотечной Ассамблеи Евразии РГБ. – 2013. – № 2. – С. 84–86.

31. Нешитой, В. В. Как вычислить закон распределения случайной величины? / В. В. Нешитой // Медико-социальная экология личности: состояние и перспективы : материалы XI Междунар. конф., Минск, 17–18 мая 2013 г. / редкол.: В. А. Прокашева (отв. ред.) [и др.]. – Минск, 2013. – С. 484–486.

32. Нешитой, В. В. Методы вычисления границ ядра и зон рассеяния публикаций / В. В. Нешитой // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2013. – № 11. – С. 1–11 ; Neshitoi, V. V. Methods for Calculating the Boundaries of the Core and Zones of Scattering of Publications / V. V. Neshitoi // Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, № 5–6. – P. 169–179.

33. Нешитой, В. В. Наукометрический анализ документных потоков на базе ранговых моделей / В. В. Нешитой // Навуковы пошук у сферы сучаснай культуры і мастацтва : матэрыялы навук. канф., Мінск, 28 лістап. 2013 г. / М-ва культуры Рэсп. Беларусь, Беларус. дзярж. ун-т культуры і мастацтваў ; рэдкал.: Ю. П. Бондар (старш.) [і інш.]. – Мінск, 2014. – С. 206–210.

34. Петренко, Б. В. Применение закона Вейбулла для расчета полноты комплектования справочно-информационного фонда / Б. В. Петренко, В. В. Нешитой // Проблемы оптимального комплектования и использования справочно-информационного фонда для принятия решений / Общество «Знание» Украинской ССР. – Киев, 1974. – С. 6–8.

35. Пиотровский, Р. Г. Информационные измерения языка / Р. Г. Пиотровский. – Л., 1968. – 116 с.

36. Пиотровский, Р. Г. Математическая лингвистика / Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская. – М. : Высш. шк. – Л., 1977. – 383 с.

37. Поллард Дж. Справочник по вычислительным методам статистики / Дж. Поллард ; пер. с англ. – М. : Финансы и статистика, 1982. – 344 с.

38. Прудников А. П. Интегралы и ряды. Элементарные функции. / А. П. Прудников, Ю. А. Брычков, О. И. Маричев. – М. : Наука, 1981. –800 с.

39. Справочник библиографа / Е. Н. Буринская [и др.] ; науч. ред. А. Н. Ванеев, В. А. Минкина. – СПб. : Профессия, 2002. – 527 с.

40. Хайтун, С. Д. Наукометрия. Состояние и перспективы / С. Д. Хайтун. – М : Наука, 1983. – 344 с.

41. Частотный словарь русского языка / под ред. Л. Н. Засориной. – М. : Русский язык, 1977. – 936 с.

42. Ягур В. Е. Новый подход к статистическому анализу биомедицинских данных / В. Е. Ягур, В. В. Нешитой, И. И. Саливон // Здравоохранение. – 2009. – № 8. – С. 8–13.

43. Bradford, S. C. Documentation / S. C. Bradford. – London, 1948. – 156 p.

44. Brookes, B. C. The Derivation and Application of the Bradford-Zipf Distribution / B. C. Brookes // Journal of Documentation. – 1968. – Vol. 24, № 4. – P. 247–265.

45. Brookes, B.C. Bradford's Law and the Bibliography of Science / B. C. Brookes // Nature. – 1969. – № 9. – P. 953–956.

46. Vickeri, B. C. Bradford's Law of Scattering / B. C. Vickeri // Journal of Documentation. – 1948. – Vol. 4, № 3. – P. 198–203.

47. Zipf, G. K. Human Behaviour and the Principle of Least Effort / G. K. Zipf. – Cambridge : Addison-Wesley Press, 1949. – 573 p.

Литература
по контролю качества продукции

48. СТ СЭВ 3946-82. Внедрение статистических методов анализа, регулирования технологических процессов и статистических методов приемочного контроля.

49. Гончаров, Э. Н. Контроль качества продукции / Э. Н. Гончаров, В. В. Козлов, Е. Д. Круглова. – М. : Изд-во стандартов, 1987. – 120 с.

50. Кутузов, В. А. Статистические методы в управлении качеством продукции : учеб. пособие / В. А. Кутузов. – М. : АНХ при Совете Министров СССР, 1983. – 49 с.

51. Методика организации внедрения статистических методов контроля качества продукции на промышленном предприятии. – М. : Изд-во стандартов, 1977. – 40 с.

52. Нешитой, В. В. Анализ распределения и динамики заработной платы в строительстве / В. В. Нешитой // Строительство и недвижимость. – 2000. – № 13. – С. 7.

53. Нешитой, В. В. Обобщенные распределения в системах управления качеством / В. В. Нешитой // Строительство и недвижимость. 2000. – № 11 (245). – С. 6.

*54.Нешитой, В. В.*Применение обобщенных распределений в системах управления качеством / В. В. Нешитой // Новости. Стандартизация и сертификация. – 2004. – № 1. – С. 54–58.

55. Нешитой, В. В. Пути повышения эффективности статистических методов в системах управления качеством / В. В. Нешитой // Новости. Стандартизация и сертификация. – 2001. – № 2. – С. 44–46.

56. Нешитой, В. В. Современные методы .статанализа и качество продукции / В. В. Нешитой // Человек и экономика. – 1997. – № 11–12. – С. 28–29.

57. Фридлендер, И. Г. Управляющий контроль качества продукции на рабочих местах : справочник / И. Г. Фридлендер, Э. И. Жученко. – Л. : Машиностроение, Ленингр. отд-ние, 1988. – 118 с.

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений функций Γ(x), Ψ(x), Ψ′(x), g(x)

x	Γ(x)	Ψ(x)	Ψ′(x)	g(x)
0,01	99,432585	-100,56088	10001,621	0,017483
0,02	49,44221	-50,54479	2501,598	0,034502
0,03	32,785	-33,86225	1112,687	0,05108
0,04	24,46095	-25,51327	626,5537	0,067241
0,05	19,47009	-20,49784	401,5324	0,083006
0,06	16,14573	-17,14929	279,2893	0,098394
0,07	13,7736	-14,75333	205,5729	0,113424
0,08	11,99657	-12,9528	157,7215	0,128114
0,09	10,61622	-11,54929	124,9089	0,142479
0,1	9,513508	-10,42375	101,4333	0,156535
0,11	8,612686	-9,500423	84,05954	0,170294
0,12	7,863252	-8,728789	70,8414	0,183771
0,13	7,230242	-8,073882	60,55102	0,196978
0,14	6,688686	-7,510723	52,3827	0,209926
0,15	6,220273	-7,020993	45,79	0,222626
0,16	5,811269	-6,590953	40,39171	0,235089
0,17	5,451174	-6,210094	35,9153	0,247324
0,18	5,131821	-5,870243	32,1618	0,259339
0,19	4,846763	-5,564946	28,98315	0,271145
0,2	4,590844	-5,28904	26,26738	0,282749
0,21	4,359888	-5,038344	23,92849	0,294158
0,22	4,150482	-4,809438	21,89961	0,30538
0,23	3,959804	-4,599496	20,12804	0,316421
0,24	3,785504	-4,40616	18,57186	0,327289

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
0,25		3,62561		-4,227454		17,19733		0,337989
0,26		3,47845		-4,0617		15,97709		0,348527
0,27		3,342604		-3,907473		14,88874		0,35891
0,28		3,216852		-3,763547		13,91381		0,369141
0,29		3,100143		-3,62887		13,03697		0,379226
0,3		2,991569		-3,502524		12,24536		0,38917
0,31		2,890336		-3,383714		11,52821		0,398978
0,32		2,795751		-3,271742		10,87638		0,408654
0,33		2,707206		-3,165995		10,28209		0,418201
0,34		2,624163		-3,065931		9,738684		0,427625
0,35		2,546147		-2,971071		9,240459		0,436928
0,36		2,472735		-2,880988		8,782481		0,446115
0,37		2,40355		-2,795301		8,360474		0,455189
0,38		2,338256		-2,713671		7,970718		0,464153
0,39		2,276549		-2,63579		7,609962		0,47301
0,4		2,21816		-2,561385		7,275357		0,481764
0,41		2,162841		-2,490205		6,964395		0,490417
0,42		2,110371		-2,422025		6,674866		0,498972
0,43		2,060549		-2,356642		6,404809		0,507432
0,44		2,013193		-2,29387		6,152487		0,515799
0,45		1,968136		-2,233539		5,91635		0,524076
0,46		1,925227		-2,175494		5,695018		0,532266
0,47		1,884326		-2,119593		5,487251		0,540369
0,48		1,845306		-2,065707		5,291942		0,54839
0,49		1,808051		-2,013716		5,108092		0,556329
0,5		1,772454		-1,96351		4,934802		0,56419
0,51		1,738415		-1,914988		4,771259		0,571973

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
0,52		1,705844		-1,868055		4,616728		0,57968
0,53		1,674656		-1,822625		4,470543		0,587315
0,54		1,644773		-1,778618		4,332097		0,594877
0,55		1,616124		-1,735959		4,200839		0,60237
0,56		1,588641		-1,694579		4,076267		0,609794
0,57		1,562263		-1,654413		3,957921		0,617151
0,58		1,53693		-1,615401		3,845381		0,624443
0,59		1,51259		-1,577487		3,738261		0,631671
0,6		1,489192		-1,540619		3,63621		0,638837
0,61		1,46669		-1,504747		3,538901		0,645941
0,62		1,445038		-1,469826		3,446036		0,652986
0,63		1,424197		-1,435813		3,35734		0,659973
0,64		1,404128		-1,402667		3,272557		0,666902
0,65		1,384795		-1,370349		3,191454		0,673775
0,66		1,366164		-1,338826		3,113813		0,680594
0,67		1,348204		-1,308062		3,039432		0,687359
0,68		1,330884		-1,278027		2,968125		0,694071
0,69		1,314177		-1,24869		2,899718		0,700732
0,7		1,298055		-1,220024		2,834049		0,707342
0,71		1,282495		-1,192001		2,770969		0,713902
0,72		1,267473		-1,164596		2,710339		0,720415
0,73		1,252966		-1,137786		2,652027		0,726879
0,74		1,238954		-1,111548		2,595912		0,733297
0,75		1,225417		-1,085861		2,54188		0,739669
0,76		1,212335		-1,060704		2,489825		0,745996
0,77		1,199692		-1,036058		2,439647		0,752279
0,78		1,187471		-1,011905		2,391253		0,758518

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
0,79		1,175655		-0,988227		2,344556		0,764715
0,8		1,16423		-0,965009		2,299474		0,770871
0,81		1,153181		-0,942233		2,255929		0,776985
0,82		1,142494		-0,919885		2,213849		0,783059
0,83		1,132157		-0,897951		2,173166		0,789094
0,84		1,122158		-0,876417		2,133816		0,795089
0,85		1,112484		-0,855271		2,095738		0,801047
0,86		1,103124		-0,834499		2,058874		0,806967
0,87		1,094069		-0,814089		2,023172		0,812849
0,88		1,085308		-0,794031		1,988581		0,818696
0,89		1,076831		-0,774314		1,955052		0,824507
0,9		1,068629		-0,754927		1,92254		0,830283
0,91		1,060693		-0,73586		1,891002		0,836024
0,92		1,053016		-0,717104		1,860398		0,841731
0,93		1,045588		-0,698649		1,830688		0,847405
0,94		1,038403		-0,680487		1,801837		0,853045
0,95		1,031453		-0,66261		1,773809		0,858654
0,96		1,024732		-0,645008		1,746573		0,86423
0,97		1,018232		-0,627676		1,720096		0,869775
0,98		1,011947		-0,610604		1,694349		0,87529
0,99		1,005872		-0,593786		1,669304		0,880773
1		1		-0,577216		1,644934		0,886227
1,01		0,994326		-0,560885		1,621214		0,891651
1,02		0,988844		-0,544789		1,598118		0,897046
1,03		0,98355		-0,528921		1,575625		0,902412
1,04		0,978438		-0,513275		1,553712		0,90775
1,05		0,973504		-0,497845		1,532357		0,913061

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
1,06		0,968744		-0,482626		1,511542		0,918343
1,07		0,964152		-0,467612		1,491246		0,923599
1,08		0,959725		-0,452799		1,471452		0,928828
1,09		0,955459		-0,438182		1,452142		0,93403
1,1		0,951351		-0,423755		1,433299		0,939207
1,11		0,947396		-0,409514		1,414908		0,944358
1,12		0,94359		-0,395455		1,396952		0,949484
1,13		0,939931		-0,381574		1,379418		0,954585
1,14		0,936416		-0,367866		1,362292		0,959661
1,15		0,933041		-0,354327		1,345559		0,964713
1,16		0,929803		-0,340953		1,329208		0,969741
1,17		0,9267		-0,327741		1,313226		0,974746
1,18		0,923728		-0,314687		1,297601		0,979727
1,19		0,920885		-0,301788		1,282322		0,984685
1,2		0,918169		-0,28904		1,267377		0,989621
1,21		0,915576		-0,276439		1,252757		0,994534
1,22		0,913106		-0,263984		1,238452		0,999425
1,23		0,910755		-0,251669		1,224452		1,004294
1,24		0,908521		-0,239494		1,210747		1,009141
1,25		0,906402		-0,227454		1,197329		1,013967
1,26		0,904397		-0,215546		1,184189		1,018772
1,27		0,902503		-0,203769		1,17132		1,023557
1,28		0,900718		-0,192119		1,158712		1,028321
1,29		0,899042		-0,180594		1,146359		1,033064
1,3		0,897471		-0,169191		1,134253		1,037787
1,31		0,896004		-0,157908		1,122388		1,042491
1,32		0,89464		-0,146742		1,110755		1,047175

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
1,33		0,893378		-0,135692		1,09935		1,05184
1,34		0,892216		-0,124755		1,088165		1,056485
1,35		0,891151		-0,113928		1,077194		1,061112
1,36		0,890185		-0,10321		1,066431		1,06572
1,37		0,889314		-0,092599		1,055872		1,070309
1,38		0,888537		-0,082092		1,04551		1,07488
1,39		0,887854		-0,071688		1,03534		1,079433
1,4		0,887264		-0,061385		1,025357		1,083968
1,41		0,886765		-0,05118		1,015556		1,088486
1,42		0,886356		-0,041073		1,005932		1,092986
1,43		0,886036		-0,031061		0,996481		1,097469
1,44		0,885805		-0,021143		0,987198		1,101934
1,45		0,885661		-0,011316		0,978079		1,106383
1,46		0,885604		-0,001581		0,96912		1,110815
1,47		0,885633		0,0080665		0,960316		1,11523
1,48		0,885747		0,0176263		0,951665		1,119629
1,49		0,885945		0,0271003		0,943161		1,124012
1,5		0,886227		0,03649		0,934802		1,128379
1,51		0,886592		0,0457968		0,926584		1,13273
1,52		0,887039		0,0550221		0,918504		1,137065
1,53		0,887568		0,0641673		0,910557		1,141385
1,54		0,888178		0,0732337		0,902742		1,145689
1,55		0,888868		0,0822226		0,895054		1,149978
1,56		0,889639		0,0911352		0,887491		1,154253
1,57		0,89049		0,0999728		0,880051		1,158512
1,58		0,89142		0,1087366		0,872729		1,162756
1,59		0,892428		0,1174278		0,865524		1,166986

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
1,6		0,893515		0,1260475		0,858432		1,171201
1,61		0,894681		0,1345968		0,851451		1,175402
1,62		0,895924		0,1430768		0,844579		1,179588
1,63		0,897244		0,1514887		0,837813		1,183761
1,64		0,898642		0,1598335		0,831151		1,187919
1,65		0,900117		0,1681121		0,82459		1,192064
1,66		0,901668		0,1763256		0,818129		1,196195
1,67		0,903296		0,184475		0,811765		1,200313
1,68		0,905001		0,1925612		0,805495		1,204417
1,69		0,906782		0,2005852		0,799319		1,208508
1,7		0,908639		0,2085479		0,793233		1,212586
1,71		0,910572		0,2164501		0,787236		1,216651
1,72		0,912581		0,2242929		0,781326		1,220702
1,73		0,914665		0,232077		0,775502		1,224741
1,74		0,916826		0,2398032		0,769761		1,228768
1,75		0,919063		0,2474725		0,764102		1,232781
1,76		0,921375		0,2550855		0,758523		1,236783
1,77		0,923763		0,2626432		0,753022		1,240771
1,78		0,926227		0,2701462		0,747598		1,244748
1,79		0,928767		0,2775954		0,742249		1,248713
1,8		0,931384		0,2849914		0,736974		1,252665
1,81		0,934076		0,2923351		0,731771		1,256606
1,82		0,936845		0,2996271		0,726639		1,260534
1,83		0,93969		0,3068681		0,721577		1,264451
1,84		0,942612		0,3140589		0,716582		1,268357
1,85		0,945611		0,3212		0,711655		1,272251
1,86		0,948687		0,3282922		0,706792		1,276133

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
1,87		0,95184		0,335336		0,701994		1,280004
1,88		0,955071		0,3423323		0,697259		1,283864
1,89		0,958379		0,3492814		0,692585		1,287713
1,9		0,961766		0,3561842		0,687972		1,291551
1,91		0,965231		0,3630411		0,683418		1,295378
1,92		0,968774		0,3698527		0,678923		1,299193
1,93		0,972397		0,3766197		0,674485		1,302999
1,94		0,976099		0,3833426		0,670103		1,306793
1,95		0,979881		0,390022		0,665776		1,310577
1,96		0,983743		0,3966583		0,661503		1,31435
1,97		0,987685		0,4032522		0,657284		1,318113
1,98		0,991708		0,4098042		0,653116		1,321866
1,99		0,995813		0,4163147		0,649		1,325608
2		1		0,4227843		0,644934		1,32934
2,01		1,004269		0,4292136		0,640917		1,333062
2,02		1,008621		0,4356028		0,636949		1,336775
2,03		1,013056		0,4419527		0,633029		1,340477
2,04		1,017576		0,4482636		0,629155		1,344169
2,05		1,022179		0,454536		0,625328		1,347851
2,06		1,026868		0,4607703		0,621546		1,351524
2,07		1,031643		0,466967		0,617808		1,355187
2,08		1,036503		0,4731266		0,614113		1,358841
2,09		1,041451		0,4792494		0,610462		1,362485
2,1		1,046486		0,485336		0,606853		1,366119
2,11		1,051609		0,4913866		0,603285		1,369745
2,12		1,056821		0,4974018		0,599758		1,373361
2,13		1,062123		0,5033819		0,596272		1,376967

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
2,14		1,067514		0,5093274		0,592824		1,380565
2,15		1,072997		0,5152385		0,589416		1,384154
2,16		1,078572		0,5211158		0,586045		1,387733
2,17		1,084239		0,5269596		0,582712		1,391304
2,18		1,089999		0,5327702		0,579416		1,394865
2,19		1,095853		0,538548		0,576157		1,398418
2,2		1,101802		0,5442934		0,572933		1,401963
2,21		1,107848		0,5500068		0,569744		1,405498
2,22		1,113989		0,5556884		0,56659		1,409025
2,23		1,120228		0,5613387		0,56347		1,412543
2,24		1,126566		0,5669579		0,560383		1,416053
2,25		1,133003		0,5725465		0,557329		1,419554
2,26		1,13954		0,5781046		0,554308		1,423047
2,27		1,146179		0,5836327		0,551319		1,426532
2,28		1,15292		0,5891311		0,548361		1,430008
2,29		1,159764		0,5946		0,545434		1,433476
2,3		1,166712		0,6000399		0,542537		1,436936
2,31		1,173765		0,6054509		0,539671		1,440388
2,32		1,180925		0,6108334		0,536834		1,443832
2,33		1,188193		0,6161877		0,534027		1,447268
2,34		1,195569		0,621514		0,531248		1,450696
2,35		1,203054		0,6268127		0,528497		1,454116
2,36		1,210651		0,6320841		0,525774		1,457528
2,37		1,21836		0,6373283		0,523078		1,460933
2,38		1,226181		0,6425457		0,52041		1,464329
2,39		1,234117		0,6477366		0,517768		1,467718
2,4		1,242169		0,6529012		0,515153		1,4711

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
2,41		1,250338		0,6580397		0,512563		1,474474
2,42		1,258625		0,6631525		0,509998		1,47784
2,43		1,267032		0,6682398		0,507459		1,481199
2,44		1,275559		0,6733018		0,504945		1,48455
2,45		1,284209		0,6783387		0,502455		1,487894
2,46		1,292982		0,6833509		0,499988		1,491231
2,47		1,301881		0,6883386		0,497546		1,49456
2,48		1,310906		0,6933019		0,495127		1,497883
2,49		1,320058		0,6982412		0,492731		1,501198
2,5		1,32934		0,7031566		0,490358		1,504506
2,51		1,338753		0,7080484		0,488007		1,507806
2,52		1,348299		0,7129169		0,485678		1,5111
2,53		1,357978		0,7177621		0,483371		1,514387
2,54		1,367794		0,7225843		0,481085		1,517666
2,55		1,377746		0,7273839		0,478821		1,520939
2,56		1,387837		0,7321608		0,476578		1,524205
2,57		1,398069		0,7369155		0,474355		1,527464
2,58		1,408443		0,741648		0,472152		1,530717
2,59		1,418961		0,7463586		0,46997		1,533962
2,6		1,429625		0,7510475		0,467807		1,537201
2,61		1,440436		0,7557148		0,465664		1,540433
2,62		1,451396		0,7603608		0,46354		1,543659
2,63		1,462508		0,7649856		0,461435		1,546878
2,64		1,473773		0,7695896		0,459349		1,55009
2,65		1,485193		0,7741727		0,457281		1,553296
2,66		1,496769		0,7787352		0,455232		1,556495
2,67		1,508505		0,7832774		0,4532		1,559688

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
2,68		1,520402		0,7877993		0,451187		1,562875
2,69		1,532461		0,7923012		0,449191		1,566055
2,7		1,544686		0,7967832		0,447212		1,569229
2,71		1,557078		0,8012455		0,445251		1,572396
2,72		1,569639		0,8056882		0,443306		1,575558
2,73		1,582371		0,8101116		0,441378		1,578713
2,74		1,595277		0,8145158		0,439466		1,581862
2,75		1,608359		0,818901		0,437571		1,585005
2,76		1,62162		0,8232673		0,435692		1,588141
2,77		1,635061		0,8276149		0,433829		1,591272
2,78		1,648685		0,831944		0,431981		1,594396
2,79		1,662494		0,8362546		0,430149		1,597515
2,8		1,676491		0,840547		0,428332		1,600628
2,81		1,690678		0,8448213		0,42653		1,603734
2,82		1,705058		0,8490776		0,424743		1,606835
2,83		1,719633		0,8533162		0,422971		1,60993
2,84		1,734407		0,8575371		0,421214		1,613019
2,85		1,749381		0,8617405		0,41947		1,616102
2,86		1,764558		0,8659266		0,417741		1,61918
2,87		1,779941		0,8700954		0,416026		1,622251
2,88		1,795533		0,8742472		0,414325		1,625317
2,89		1,811337		0,878382		0,412638		1,628378
2,9		1,827355		0,8825		0,410964		1,631432
2,91		1,843591		0,8866013		0,409303		1,634482
2,92		1,860047		0,8906861		0,407656		1,637525
2,93		1,876726		0,8947544		0,406021		1,640563
2,94		1,893632		0,8988065		0,4044		1,643595

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
2,95		1,910767		0,9028425		0,402791		1,646622
2,96		1,928135		0,9068624		0,401195		1,649644
2,97		1,945739		0,9108664		0,399612		1,65266
2,98		1,963583		0,9148547		0,39804		1,65567
2,99		1,981668		0,9188273		0,396481		1,658676
3		2		0,9227843		0,394934		1,661675
3,01		2,018581		0,926726		0,393399		1,66467
3,02		2,037415		0,9306524		0,391875		1,667659
3,03		2,056505		0,9345635		0,390364		1,670643
3,04		2,075854		0,9384597		0,388863		1,673622
3,05		2,095468		0,9423408		0,387374		1,676596
3,06		2,115349		0,9462072		0,385897		1,679564
3,07		2,1355		0,9500588		0,38443		1,682527
3,08		2,155927		0,9538958		0,382974		1,685485
3,09		2,176632		0,9577183		0,38153		1,688438
3,1		2,19762		0,9615264		0,380096		1,691386
3,11		2,218895		0,9653203		0,378672		1,694329
3,12		2,240461		0,9690999		0,377259		1,697266
3,13		2,262321		0,9728655		0,375857		1,700199
3,14		2,284481		0,9766171		0,374465		1,703127
3,15		2,306944		0,9803548		0,373083		1,70605
3,16		2,329715		0,9840788		0,371711		1,708968
3,17		2,352798		0,9877891		0,370349		1,711881
3,18		2,376197		0,9914858		0,368996		1,714789
3,19		2,399918		0,995169		0,367654		1,717692
3,2		2,423965		0,9988389		0,366321		1,72059
3,21		2,448343		1,0024955		0,364998		1,723484

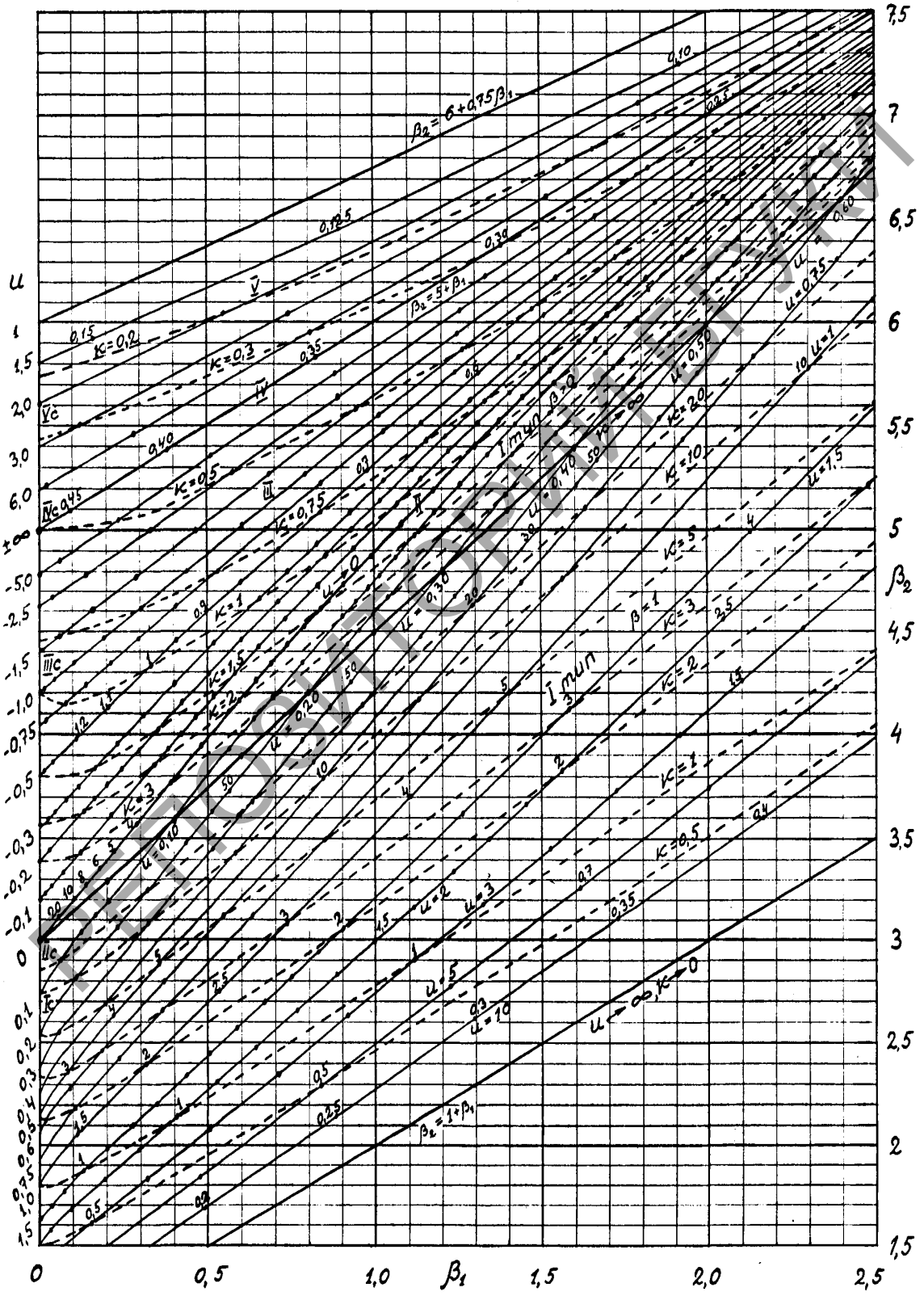
x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
3,22		2,473056		1,0061389		0,363684		1,726373
3,23		2,498109		1,0097692		0,36238		1,729257
3,24		2,523508		1,0133865		0,361084		1,732136
3,25		2,549257		1,0169909		0,359798		1,735011
3,26		2,575361		1,0205825		0,358521		1,737881
3,27		2,601826		1,0241614		0,357253		1,740746
3,28		2,628657		1,0277276		0,355994		1,743607
3,29		2,655859		1,0312813		0,354743		1,746463
3,3		2,683437		1,0348225		0,353502		1,749314
3,31		2,711398		1,0383513		0,352268		1,752161
3,32		2,739747		1,0418679		0,351044		1,755003
3,33		2,768489		1,0453722		0,349827		1,757841
3,34		2,797631		1,0488645		0,348619		1,760674
3,35		2,827178		1,0523446		0,347419		1,763502
3,36		2,857136		1,0558129		0,346228		1,766327
3,37		2,887512		1,0592692		0,345044		1,769146
3,38		2,918311		1,0627138		0,343869		1,771962
3,39		2,949541		1,0661466		0,342701		1,774772
3,4		2,981206		1,0695678		0,341541		1,777579
3,41		3,013315		1,0729775		0,340389		1,780381
3,42		3,045873		1,0763756		0,339245		1,783179
3,43		3,078887		1,0797624		0,338108		1,785972
3,44		3,112365		1,0831378		0,336979		1,788761
3,45		3,146312		1,086502		0,335857		1,791546
3,46		3,180737		1,089855		0,334743		1,794327
3,47		3,215645		1,0931969		0,333636		1,797103
3,48		3,251046		1,0965278		0,332536		1,799875

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
3,49		3,286945		1,0998476		0,331443		1,802643
3,5		3,323351		1,1031566		0,330358		1,805407
3,51		3,360271		1,1064548		0,329279		1,808166
3,52		3,397713		1,1097422		0,328208		1,810921
3,53		3,435686		1,113019		0,327143		1,813673
3,54		3,474196		1,1162851		0,326085		1,81642
3,55		3,513252		1,1195407		0,325034		1,819163
3,56		3,552863		1,1227858		0,32399		1,821902
3,57		3,593037		1,1260205		0,322952		1,824636
3,58		3,633783		1,1292449		0,321921		1,827367
3,59		3,675109		1,132459		0,320896		1,830094
3,6		3,717024		1,1356628		0,319878		1,832817
3,61		3,759537		1,1388566		0,318866		1,835535
3,62		3,802658		1,1420402		0,317861		1,83825
3,63		3,846396		1,1452138		0,316861		1,840961
3,64		3,890761		1,1483774		0,315868		1,843668
3,65		3,935761		1,1515312		0,314882		1,846371
3,66		3,981407		1,1546751		0,313901		1,84907
3,67		4,027709		1,1578092		0,312926		1,851765
3,68		4,074677		1,1609336		0,311957		1,854456
3,69		4,122321		1,1640484		0,310995		1,857143
3,7		4,170652		1,1671535		0,310038		1,859827
3,71		4,21968		1,1702492		0,309087		1,862506
3,72		4,269417		1,1733353		0,308142		1,865182
3,73		4,319873		1,176412		0,307202		1,867854
3,74		4,37106		1,1794794		0,306268		1,870523
3,75		4,422988		1,1825374		0,30534		1,873187

x		Γ(x)		Ψ(x)		Ψ′(x)		g(x)
3,76		4,475671		1,1855862		0,304417		1,875848
3,77		4,529118		1,1886258		0,3035		1,878505
3,78		4,583343		1,1916562		0,302588		1,881158
3,79		4,638358		1,1946775		0,301682		1,883808
3,8		4,694174		1,1976898		0,300781		1,886454
3,81		4,750805		1,2006932		0,299886		1,889096
3,82		4,808264		1,2036876		0,298995		1,891735
3,83		4,866563		1,2066731		0,29811		1,89437
3,84		4,925715		1,2096498		0,29723		1,897001
3,85		4,985735		1,2126177		0,296356		1,899629
3,86		5,046636		1,2155769		0,295486		1,902253
3,87		5,108431		1,2185275		0,294621		1,904874
3,88		5,171136		1,2214694		0,293762		1,907491
3,89		5,234764		1,2244027		0,292907		1,910104
3,9		5,29933		1,2273275		0,292058		1,912714
3,91		5,364849		1,2302439		0,291213		1,91532
3,92		5,431336		1,2331518		0,290373		1,917923
3,93		5,498807		1,2360514		0,289538		1,920523
3,94		5,567278		1,2389426		0,288707		1,923118
3,95		5,636763		1,2418255		0,287882		1,925711
3,96		5,707281		1,2447002		0,287061		1,9283
3,97		5,778846		1,2475668		0,286245		1,930885
3,98		5,851476		1,2504251		0,285433		1,933467
3,99		5,925188		1,2532754		0,284626		1,936046
4		6		1,2561177		0,283823		1,938621

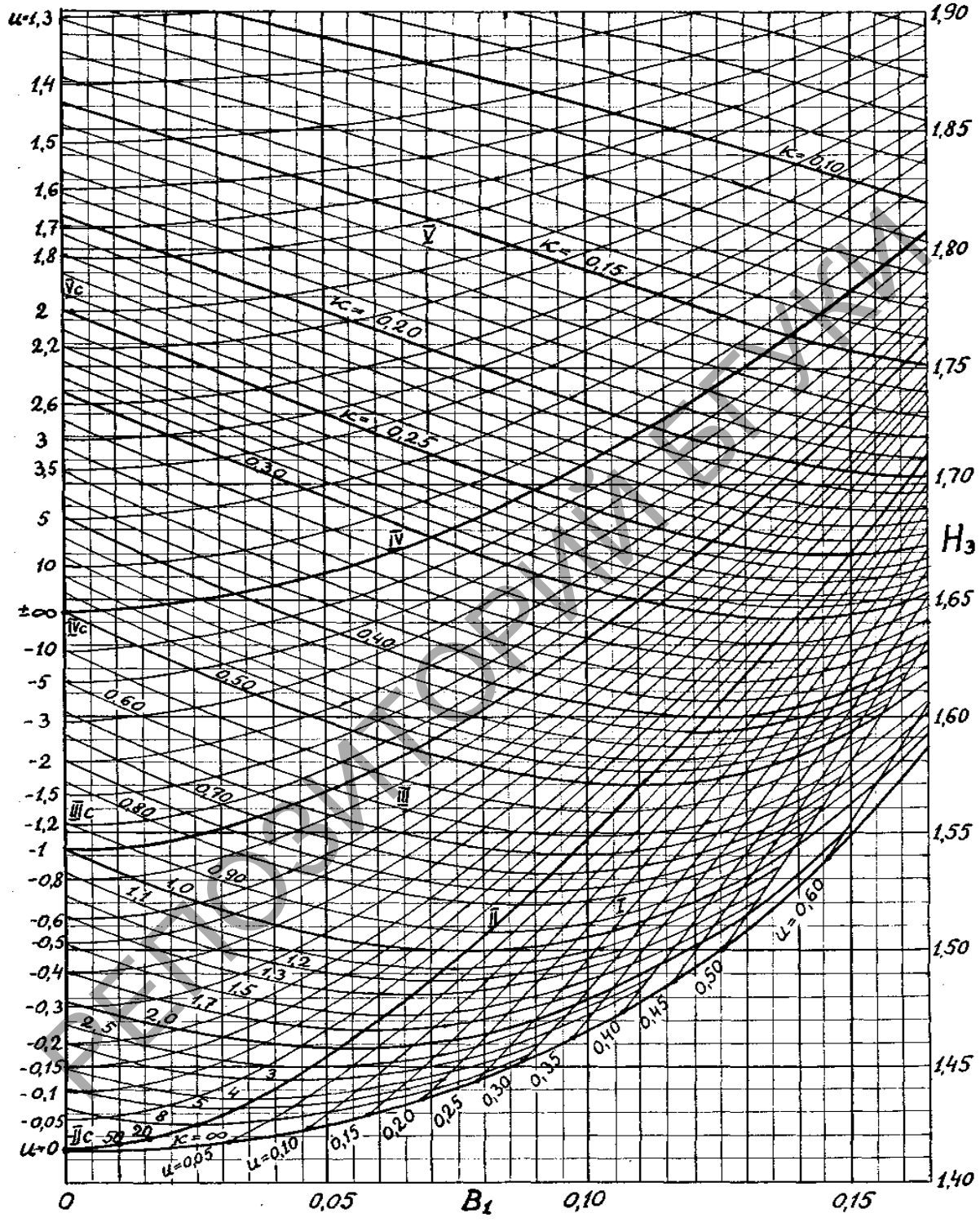
Приложение 2

Номограммы для установления типа аппроксимирующего распределения и нахождения о ценок параметров k, u: по методу моментов



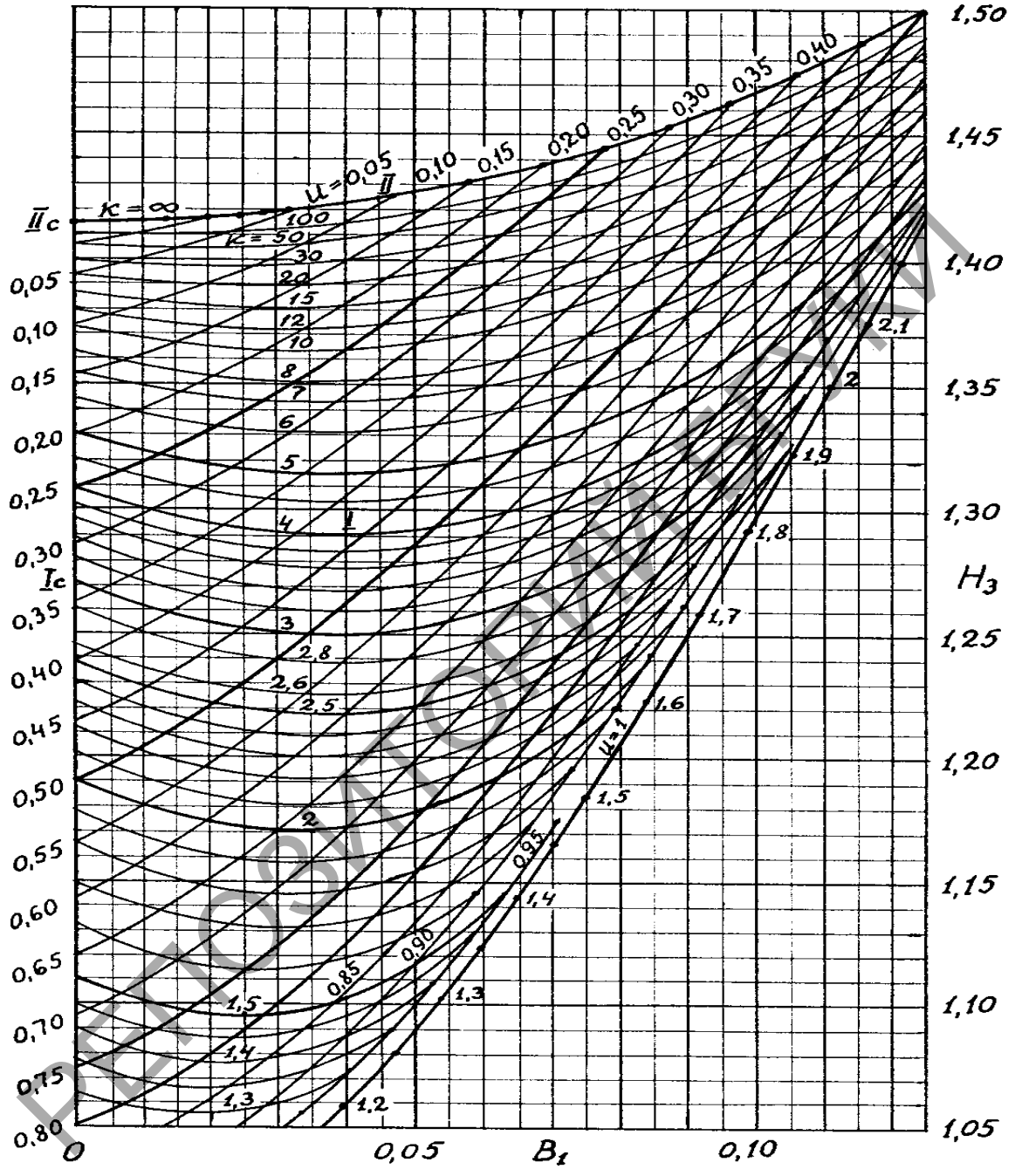
Приложение 3

Номограмма по устойчивому методу



Приложение 3А

По устойчивому методу при β=1



Приложение 4

Таблицы для вычисления показателей статистических распределений

Таблица 1

Первая система непрерывных распределений

x_i	m_i	h_i	$x_i \frac{m_i}{M}$	$\left(\frac{m_i}{M}\right)^2 \frac{1}{h_i}$	$\left(\frac{m_i}{M}\right)^4 \frac{1}{h_i^3}$	$x_i\left(\frac{m_i}{M}\right)^2 \frac{1}{h_i}$
Сумма	M		$\nu_1 = \bar{x}$	S_1	S_3	$B0 = B + \nu_1 S_1$

Таблица 2

Вторая система непрерывных распределений

t_i	m_i	h_i	$(\ln t_i)\frac{m_i}{M}$	$t_i\left(\frac{m_i}{M}\right)^2 \frac{1}{h_i}$	$t_i^3\left(\frac{m_i}{M}\right)^4 \frac{1}{h_i^3}$	$t_i(\ln t_i)\left(\frac{m_i}{M}\right)^2 \frac{1}{h_i}$
Сумма	M		$\nu_1 = \overline{\ln t}$	S_1	S_3	$B0 = B + \nu_1 S_1$

Научное издание

Нешитой Василий Васильевич

ИНФОРМЕТРИЯ:
МАТЕМАТИЧЕСКИЕ МОДЕЛИ
И МЕТОДЫ ИССЛЕДОВАНИЯ

Редактор О. М. Соколова
Технический редактор Л. Н. Мельник

Подписано в печать 2017. Формат 60x84 $^{1}/_{16}$.
Бумага офисная. Ризография.
Усл. печ. л. 15,95. Уч.-изд. л. 13,12. Тираж экз. Заказ .

Издатель и полиграфическое исполнение:
учреждение образования
«Белорусский государственный университет культуры и искусств».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/177 от 12.02.2014.
ЛП № 02330/456 от 23.01.2014.
Ул. Рабкоровская, 17, 220007, г. Минск.