Учреждение образования
«Белорусский государственный университет культуры и искусств»

Факультет информационно-документных коммуникаций
Кафедра информационно-аналитической деятельности

СОГЛАСОВАНО
Заведующий кафедрой
_____Н.А. Яцевич
___ _____2022 г.

СОГЛАСОВАНО
Декан факультета
_____Ю.Н. Галковская
___ _____2022 г.

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

# СТАТИСТИКА БИБЛИОТЕЧНО-ИНФОРМАЦИОННОЙ ДЕЯТЕЛЬНОСТИ

для специальности 1-23 01 11 Библиотечно-информационная деятельность (по направлениям),
направлению специальности 1-23 01 11-01 Библиотечно-информационная деятельность (менеджмент)

Составитель: И.Л. Белоновская, старший преподаватель кафедры информационно-аналитической деятельности

Рассмотрено и утверждено
на заседании Совета университета 21.06.2022
протокол № 12



2022

**СОСТАВИТЕЛЬ**

*И. Л. Белоновская*, старший преподаватель кафедры информационно-аналитической деятельности учреждения образования «Белорусский государственный университет культуры и искусств»

**РЕЦЕНЗЕНТЫ:**

*Научный Совет* государственного учреждения «Центральная научная библиотека имени Якуба Коласа Национальной академии наук Беларуси»;

*В. А. Касап*, профессор кафедры информационных ресурсов и коммуникаций учреждения образования «Белорусский государственный университет культуры и искусств», кандидат педагогических наук, доцент

**РЕКОМЕНДОВАН К УТВЕРЖДЕНИЮ:**

Кафедрой информационно-аналитической деятельности учреждения образования «Белорусский государственный университет культуры и искусств» (протокол от № );

Советом факультета информационно-документных коммуникаций учреждения образования «Белорусский государственный университет культуры и искусств» (протокол от _____ №___)

Советом Белорусского государственного университета культуры и искусств (протокол от_____ № ____)

## СОДЕРЖАНИЕ

# 1. ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебно-методический комплекс разработан для студентов факультета информационно-документных коммуникаций в соответствии с требованиями образовательного стандарта по специальности 1-23 01 11 Библиотечно-информационная деятельность (по направлениям).

Актуальность изучения дисциплины обусловлена необходимостью использования статистических методов анализа и обработки информации в библиотечно-информационной деятельности. Для этого требуется использование математических моделей.

Использование таких моделей в учебном процессе и практической работе требует от студента и специалиста ознакомления с базовыми понятиями высшей математики, такими как функция, виды функций, производная функции, производные элементарных функций, правила дифференцирования суммы и разности функций, их произведения, частного, дифференцирование сложных функций. Эти задачи студент должен уметь грамотно решать.

Целью учебной дисциплины является формирование у студентов общего подхода к вопросам построения универсальных вероятностных моделей для моделирования библиотечного фонда, описания статистических закономерностей информационных потоков, методов оценивания параметров и овладение современными методами статистического анализа.

Задачей изучения дисциплины является овладение студентами умениями и навыками использования различных математико-статистических методов анализа библиотечно-информационной деятельности.

В результате изучения дисциплины студенты должны знать:

– известные и новые методы оценивания параметров;

– сущность различных подходов к вычислению по статистическим данным выравнивающих распределений и кривых роста;

– элементы теории обобщённых распределений;

– статистические модели, которые могут быть использованы в библиотечно-информационной деятельности, в том числе универсальные законы рассеяния и старения публикаций.

Выпускники в пределах своей специальности должны уметь:

– использовать статистические методы для построения моделей библиотечных процессов;

вычислять законы распределения по статистическим данным;

– вычислять оценки параметров универсальных законов рассеяния и старения публикаций;

– прогнозировать статистические закономерности текста и информационных потоков.

– извлекать и анализировать информацию из библиотечной статистики.

Основой теоретического раздела учебно-методического комплекса «Статистика библиотечно-информационной деятельности» является издание, соответствующее программе дисциплины и обеспечивающее материалами, необходимыми для теоретического изучения и освоения дисциплины: Нешитой, В. В. Информетрия: математические модели и методы исследования / В. В. Нешитой. – Минск : БГУКИ, 2017. – 274 с., также размещенное в репозитории университета как электронный образовательный ресурс.

Практический раздел представлен рабочими материалами, заданиями для практических работ, методическими рекомендациями к их выполнению в объеме, определенной учебной программой. Также в раздел включена тематика семинарских занятий, порядок их проведения, рекомендуемая литература.

Раздел контроля знаний включает задания для самостоятельной работы, перечень контрольных вопросов к зачету.

Вспомогательный раздел включает типовую учебную программу по учебной дисциплине, учебно-методическую карту учебной дисциплины, списки основной и дополнительной литературы.

На изучение дисциплины «Статистические методы библиотечно-информационной деятельности» отводится 52 часа, из них 28 часов аудиторных занятий, в том числе 18 часов лекций, 10 часов практических занятий и 6 часов – самостоятельная работа студентов. Курс рассчитан на один семестр. Форма контроля – зачёт.

## 2. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

Для изучения и освоения материала по учебной дисциплине «Статистика библиотечно-информационной деятельности» рекомендуется использовать образовательные ресурсы, размещенные в репозитории учреждения образования «Белорусский государственный университет культуры и искусств».

Нешитой, В. В. Информетрия: математические модели и методы исследования / В. В. Нешитой. – Минск: БГУКИ, 2017. – 274 с.; То же [Электронный ресурс]. – Режим доступа: http://repository.buk.by/123456789/15166 . – Дата доступа: 18.05.2022.

КОНСПЕКТ ЛЕКЦИЙ

### Тема 1. Функции одной переменной. Производная функции

Производная простых функций.

К понятию производной приводят многие задачи, например, нахождение уравнения касательной к некоторой кривой в заданной точке, вычисление скорости изменения функции и ускорения в заданной точке, вычисление сечения бруса, имеющего наибольшую жесткость или прочность при его изготовлении из круглого бревна и многие другие.

Геометрический и физический смысл производной – это тангенс угла наклона касательной к кривой в некоторой точке или скорость движения.

Поизводной функции y=f(x) по аргументу x называется предел отношения приращения функции к приращению аргумента при условии, что последнее стремится к нулю. Производная функции y=f(x) обозначается через $y'$ или $f'(x)$.

Операция отыскания производной $f'(x)$ данной функции f(x) называется дифференцированием этой функции. Исходя из определения производной, можно получать формулы для вычисления производных элементарных функций. Найдем для примера производную функции $y = x^2$. Дадим аргументу приращение $\Delta x$. Тогда будем иметь $y + \Delta y = (x + \Delta x)^2 = x^2 + 2x\Delta x + (\Delta x)^2$. Отсюда найдем:

$\Delta y = x^2 + 2x\,\Delta x + (\Delta x)^2 - x^2 = 2x\Delta x + (\Delta x)^2$.

Далее находим предел отношения $\Delta y/\Delta x$ при $\Delta x \to 0$

$\lim \frac{\Delta y}{\Delta x} = 2x + \Delta x = 2x.$

Если найти таким же путем производную от $x^3$, то получим $(x^3)' = 3x^2$, а в общем случае производная от $x^n = nx^{n-1}$

Приведем небольшую таблицу производных некоторых функций.
Производная постоянной величины равна нулю, потому что прямая y=c параллельна горизонтальной оси, а тангенс нуля градусов равен нулю.

$c' = 0$,

$x' = 1$, ( прямая y=x – это биссектрисса координатного ула, а тангенс 45 градусов равен единице.

$(x^n)' = nx^{n-1}$

$(\ln x)' = \frac{1}{x}, \; x > 0$

$(e^x)' = e^x$,

$(a^x)' = a^x \ln a$

$(\sin x)' = \cos x$

$(\cos x)' = -\sin x$

$(tgx)' = \frac{1}{\cos^2 x}$

$(ctgx)' = -\frac{1}{\sin^2 x}$

Правила дифференцирования

Постоянный множитель необходимо выносить за знак производной, например $y = cx^3$. Тогда

$y' = (cx^3)' = c(3x^2)$.

$y=2x; \; y' = 2.$

Производная суммы или разности функций равна сумме или разности производных слагаемых:

$y = (2x + 5x^2 + 3)'$

$y' = (2x)' + (5x^2)' + 3' = 2 + 5 * 2x + 0 = 2 + 10x.$

Производная произведения

$y=u*v; \; y' = u'v + uv'$

$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2}$ Производная частного

Производные сложных функций

$y' = 3(ax^5 + bx^3 + c)^2 \cdot (ax^5 + bx^3 + c)' =$

$= 3(ax^5 + bx^3 + c)^2 \cdot (5ax^4 + 3bx^2).$

Логарифмическое дифференцирование.

Пример $y = u^v$

Логарифмируем последнее равенство

$\ln y = v \ln u$

Берем производные от обеих частей последнего равенства

$\frac{y'}{y} = v' \ln u + v \frac{u'}{u}$

Далее умножаем обе части равенства на y и, заменяя y на $u^v$, получим

$y' = vu^{v-1} \cdot u' + u^v \cdot \ln u \cdot v'$.

## Тема 2. Некоторые понятия теории вероятностей и математической статистики

Случайные события и их вероятности.

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.

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 А или В, 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.

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 $S$.

Основные теоремы теории вероятностей.

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 $B$.

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

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

*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: $q_1=1-p_1=0{,}2$; $q_2=0{,}3$; $q_3=0{,}1$. Следовательно,

$$P(A)=1-q_1q_2q_3=0{,}994.$$

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

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

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.

2.6.1. 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

| 1 | 2 | … | *n* |
|---|---|---|---|
| $p_1$ | $p_2$ | … | $p_n$ |

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.

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:

,1 ,6 ,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 A 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.*

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

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

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

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

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)

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

2.7.4. 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

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

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 $3\sigma$ 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.

Закон Вейбулла

Плотность вероятности и функция распределения задаются формулами

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

Из закона Вейбулла при $\beta = 1$ следует показательный закон, а при $\beta = 2$ - распределение Релея.

## Тема 3. Методы построения обобщенных непреывных распределений

Построение системы непрерывных распределений методом обобщения.

Рассмотрим три простейших распределения: равномерное, треугольное убывающее и треугольное возрастающее [8.

В случае равномерной плотности функция распределения задается формулой

$$F(t)= \alpha t=1-(1-\alpha t). \quad (3.1)$$

В случае треугольного убывающего распределения получим

$$F(t)=1-\left(1-\frac{\alpha}{2}t\right)^2 \quad (3.2)$$

Для треугольного возрастающего распределения имеем

$$F(t)=\alpha t^2=1-(1-\alpha t^2). \quad (3.3)$$

Обобщим попарно функции распределения (3.1), (3.2) и (3.1), (3.3) путем введения новых параметров.

В первом случае получим

$$F(t)=1-(1-\alpha u t)^{\frac{1}{u}}. \quad (3.4)$$

Во втором случае

$$F(t)=1-(1-\alpha t^{\beta}). \quad (3.5)$$

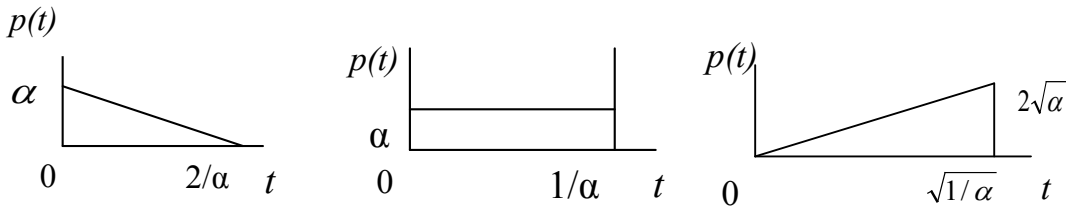
Теперь замечаем, что в формуле (3.1.4.) имеется параметр u, но его нет в формуле (3.1.5). Введем его в последнюю формулу. В результате получим

$$F(t)=1-(1-\alpha u t^{\beta})^{\frac{1}{u}} \quad (3.6)$$

откуда дифференцированием по t найдем плотность распределения

$$p(t)=\alpha\beta t^{\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1} \quad (3.7)$$





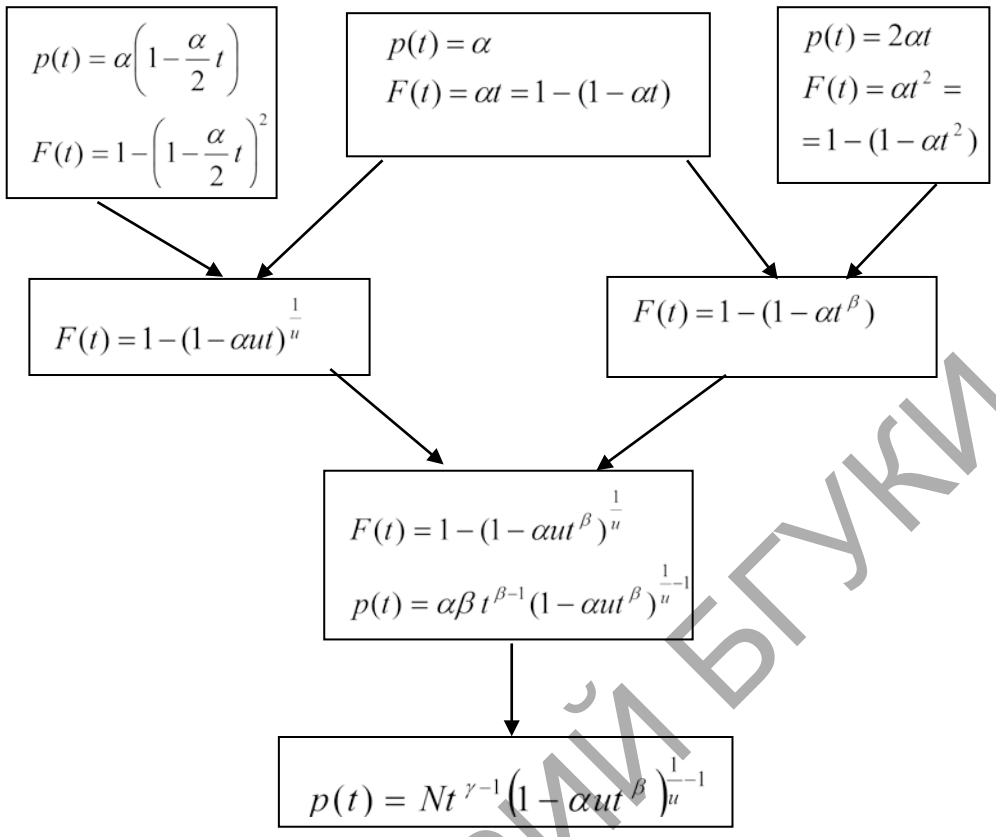


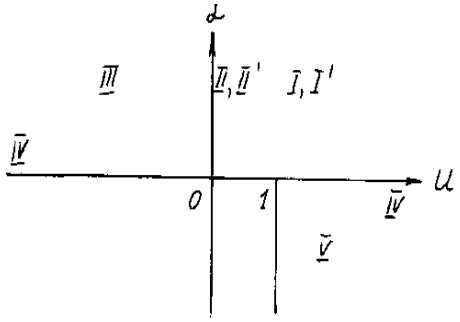

Рис. 3.1 Последовательность обобщения простейших непрерывных распределений.

Последняя плотность может быть еще более расширена за счет введения нового параметра формы. Параметр β в формуле (3.7) используется дважды в качестве показателя степени. Пусть это будут два разных параметра. Тогда вместо (3.7) можем записать

$$p(t)=Nt^{\gamma-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}. \qquad (3.8)$$

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

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




ис. 3.2 Классификация распределений (типы со штрихом – при $\beta, \gamma < 0$).

3.2. Классификация обобщенных распределений

В зависимости от значений параметров α, u, а также от знака параметров β, γ распределения, заданные обобщенной плотностью (3.8), можно разделить на типы (см. рис. 3.2.).

В таблице 3.1. приведены значения параметров распределений разных типов.

Таблица 3.1 Классификация распределений

| Тип кривой | Параметры кривой | | |
|---|---|---|---|
| | u | α | k=γ/β |
| I, I′ | 0<u<∞ | α > 0 | 0<k<∞ |
| II, II′ | u→0 | | |
| III | -∞<u<∞ | | $0 < k < 1 - \frac{1}{u}$ |
| IV | u→±∞ | αu < 0 | |
| V | 1<u<∞ | α < 0 | |

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

В группу А входят распределения с параметрами β=γ, или γ/β=k=1. Они задаются формулами (3.6) и (3.7).

В группу Б входят распределения, заданные обобщенной плотностью (3.8). 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 (3.8) при β = 2, γ = 1 следует группа симметричных распределений.

Симметричны также распределения I типа с параметрами β = 1, γ=1/u.

Приведем все существующие типы распределений обеих групп (см. табл. 3.2– 3.4).

Таблица 3.2, 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 3.3, 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′ | $p(t)=\frac{\beta(\alpha u)^{k}\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(\frac{1}{u}\right)}\frac{1}{t^{k\beta+1}}\left(1-\frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}-1}$ | $(\alpha u)^{1/\beta}<t<\infty$ $(u>0)$ |
| II | $p(t)=\frac{\beta\alpha^{k}}{\Gamma(k)}t^{k\beta-1}e^{-\alpha t^{\beta}}$ | $0<t<\infty$ $(u\to 0)$ |
| II′ | $p(t)=\frac{\beta\alpha^{k}}{\Gamma(k)}\frac{1}{t^{k\beta+1}e^{\alpha/t^{\beta}}}$ | $0<t<\infty$ $(u\to 0)$ |

| Тип кривой | Плотность распределения | Границы кривой |
|---|---|---|
| III-V | $p(t)=\dfrac{\beta(-\alpha u)^k\,\Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1-\frac{1}{u}-k\right)}\dfrac{t^{k\beta-1}}{\left(1-\alpha u t^{\beta}\right)^{1-\frac{1}{u}}}$ | $0<t<\infty$ $(\alpha u<0)$ |

Таблица 3.4, Группа симметричных распределений

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

3.3. Распределения функций случайного аргумента

Из обобщенной плотности (3.8) 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.

$$F(x) = F(t). \qquad (3.9)$$

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

По правилу дифференцирования сложной функции из (6.4.1) имеем

$$p(x)=\frac{dF(x)}{dx}=\frac{dF(t)}{dt}\frac{dt}{dx}=p(t)\frac{dt}{dx}. \qquad (3.10)$$

Воспользуемся последней формулой для нахождения других обобщенных плотностей.

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

$$p(x)=p(t)\frac{dt}{dx}=N\,e^{\gamma x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}. \qquad (3.11)$$

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

$$p(y) = \frac{N}{y}(\ln y)^{\gamma-1}\left[1 - \alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} . \qquad (3.12)$$

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

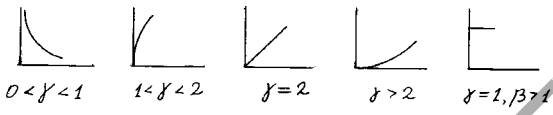
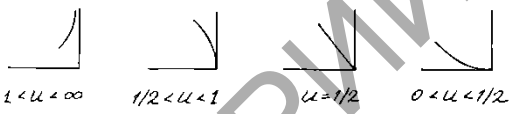



Рис. 3.4 Формы конца кривой в зависимости от значений параметра u.




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

Полученные выше обобщенные плотности распределения

$$\left.\begin{aligned} p(x) &= Ne^{\gamma x}\left(1 - \alpha u e^{\beta x}\right)^{\frac{1}{u}-1} \\ p(t) &= Nt^{\gamma-1}\left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}-1} \\ p(y) &= \frac{N(\ln y)^{\gamma-1}}{y}\left[1 - \alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} \end{aligned}\right\} \qquad (3.13)$$

образуют три основные системы непрерывных распределений В. Нешитого.

Вторая основная система непрерывных распределений, заданная плотностью p(t), представлена в таблицах 3. 2 и 3.3.

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

$$p(t) = N(t-l)^{\gamma-1}\left[1 - \alpha u(t-l)^{\beta}\right]^{\frac{1}{u}-1} . \qquad (3.14)$$

На основании плотности (3.4.2) можно получить три дополнительные системы непрерывных распределений.

Первая дополнительная система непрерывных распределений в общем случае задается формулой (3.4.2) при |β| = 1, а в случае симметричных распределений – при β = 2, γ = 1.

Ее легко получить из второй основной системы непрерывных распределений. Для этого достаточно в табл.3.2.3 принять |β| = 1, tβ заменить на $t-l$, а в табл. 3.2.3 заменить величину t2 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 3.4.1 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 3.5Первая дополнительная система непрерывных распределений

| Тип кривой | Плотность распределения (k=γ/β=γ) | Границы кривой |
|---|---|---|
| 1.1 | $p(t)=\dfrac{(\alpha u)^k\,\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(\frac{1}{u}\right)}(t-l)^{k-1}\left[1-\alpha u(t-l)\right]^{\frac{1}{u}-1}$ | $l<t<\dfrac{1}{\alpha u}+l$ |
| 1.1′ | $p(t)=\dfrac{(\alpha u)^k\,\Gamma\left(k+\frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(\frac{1}{u}\right)}\dfrac{1}{(t-l)^{k+1}}\left(1-\dfrac{\alpha u}{t-l}\right)^{\frac{1}{u}-1}$ | $t>\alpha u+l$ |
| 2.1 | $p(t)=\dfrac{\alpha^k}{\Gamma(k)}\dfrac{(t-l)^{k-1}}{e^{\alpha(t-l)}}$ | $t>l$ |
| 2.1′ | $p(t)=\dfrac{\alpha^k}{\Gamma(k)}\dfrac{1}{(t\quad l)^{k+1}e^{\alpha/(t\quad l)}}$ | $t>l$ |
| 3.1 | $p(t)=\dfrac{(-\alpha u)^k\,\Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1-\frac{1}{u}-k\right)}\dfrac{(t-l)^{k-1}}{\left[1-\alpha u(t-l)\right]^{1-\frac{1}{u}}}$ | $t>l$ |

| Тип кривой | Плотность распределения (k=γ/β=γ) | Границы кривой |
|---|---|---|
| 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

$$p(y)=\frac{N(\ln y-l)^{\gamma-1}}{y}\left[1-\alpha u(\ln y-l)^{\beta}\right]^{\frac{1}{u}-1}. \quad (3.15)$$

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

$$p(w)=\frac{N(\ln\ln w-l)^{\gamma-1}}{w\ln w}\left[1-\alpha u(\ln\ln w-l)^{\beta}\right]^{\frac{1}{u}-1}. \quad (3.16)$$

## Тема 4. Классические методы оценивания параметров непреывных 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 объёмом, 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.

### Метод наименьших квадратов

Этим методом могут быть найдены оценки параметров распределений группы *A*.

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

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

к уравнению прямой

$$\ln \frac{1 - [1 - F(t)]^{u}}{u} = \ln \alpha + \beta \ln t \quad . \tag{4.1}$$

Построив по эмпирической функции распределения график зависимости (4.1.1) (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 (4.1.1) запишем

$$Y = \ln \alpha + \beta X \, . \tag{4.2}$$

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

$$\beta = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^{2} - (\Sigma X)^{2}}, \tag{4.3}$$

$$\ln \alpha = \frac{1}{n}(\Sigma Y - \beta \Sigma X). \tag{4.4}$$

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

$$r_{y/x} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^{2} - (\Sigma X)^{2}}\sqrt{n\Sigma Y^{2} - (\Sigma Y)^{2}}}. \tag{4.5}$$

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

Вводя обозначения $Y = \ln \ln \frac{1}{1 - F(t)}$, $X = \ln t$, получим уравнение прямой (4.2).

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

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

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 [9]

$$F(t_c) = 1 - (1 - \alpha u t^{\beta})^{\frac{1}{u}} = 1 - (1-u)^{\frac{1}{u}}. \qquad (4.6)$$

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

$$F(t_c) = (1-u)^{\frac{1}{u}}. \qquad (4.7)$$

В таблице 4.1.1 приведены значения $F(t_c)$, рассчитанные по формулам (4.1.6), (4.1.7).

Таблица 4.1. Значение функции распределения $F(t_c)$

| Параметр *u* | $F(t_c)^{I\text{-}III}$ | $F(t_c)^{I'\text{-}III'}$ | 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 *C* и приравнять теоретическому.

4. С помощью таблицы 4.1. определить два значения параметра *u* (в предположении, что выравнивающее распределение относится либо к I-III, либо к I′-III′ типам).

5. По двум значениям параметра *u* 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 образом могут быть найдены оценки параметров распределений группы *А*, заданных плотностями *p(x), p(y).* При этом плотность *p(y)* должна быть приведена к форме $yp(y)\ln y = p(\ln\ln y)$.

Метод наибольшего правдоподобия

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

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

Примем в качестве логарифмической функции правдоподобия величину $\ln L = M[\ln tp(t)]$ [11].

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

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

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

$$\ln L = M\left[\ln tp(t)\right] = \ln\beta + k\ln\alpha u + \ln\Gamma\left(k+\frac{1}{u}\right) - \ln\Gamma(k) - \\ - \ln\Gamma\left(\frac{1}{u}\right) + k\beta M(\ln t) + \left(\frac{1}{u}-1\right)M\left[\ln\left(1-\alpha u t^{\beta}\right)\right]. \qquad (4.8)$$

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 (4.2.1) найдем

$$\left.\begin{aligned} &\frac{k}{\alpha} - (1-u)M\left(\frac{t^{\beta}}{1-\alpha u t^{\beta}}\right) = 0 \\ &\frac{1}{\beta} + kM(\ln t) - \alpha(1-u)M\left(\frac{t^{\beta}\ln t}{1-\alpha u t^{\beta}}\right) = 0 \\ &\ln\alpha u + \Psi\left(k+\frac{1}{u}\right) - \Psi(k) + \beta M(\ln t) = 0 \\ &\Psi\left(\frac{1}{u}\right) - \Psi\left(k+\frac{1}{u}\right) - M\left[\ln\left(1-\alpha u t^{\beta}\right)\right] = 0 \end{aligned}\right\} \qquad (4.9)$$

Здесь последнее уравнение приведено к более простой форме с учетом первого уравнения.

Оценки параметров могут быть найдены путем решения системы четырех уравнений правдоподобия – (4.2.2). 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 $\alpha u$, 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.

### 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 (6.5.2) – (6.5.4). При этом плотности (6.5.3), (6.5.4) должны быть приведены к форме плотности (6.5.2), т.е. представлены в виде

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

Рассмотрим обобщенную плотность (3.14) при $|\beta| = 1$, которую запишем в виде [9]

$$p(t) = N(t - a)^{\gamma - 1}\left[1 - \alpha u(t - a)^{\beta}\right]^{\frac{1}{u}-1}. \qquad (4.10)$$

Выразим параметры распределения (4.10) через центральные моменты. Для этого представим его в дифференциальной форме (при $|\beta| = 1$)

$$\frac{1}{p(t)}\frac{dp(t)}{dt} = \frac{\alpha(1+\gamma u - 2u)\,t + 1 + 2a\alpha u - (\gamma + a\alpha\gamma u + a\alpha)}{\alpha u t^2 - (1 + 2a\alpha u)t + a(1 + a\alpha u)}. \qquad (4.11)$$

Перепишем далее уравнение (4.3.2) в виде

$$\left[\alpha u t^2 - (1 + 2a\alpha u)\,t + a(1 + a\alpha u)\right]dp(t) = \\ = \left[\alpha(1 + \gamma u - 2u)\,t + 1 + 2a\alpha u - (\gamma + a\alpha\gamma u + a\alpha)\right]\,p(t)dt.$$

Умножим обе части последнего равенства на $t^r$ и проинтегрируем на бесконечном интервале (левую часть интегрируем по частям). В результате получим

$$t^r\left[\alpha u t^2 - (1 + 2a\alpha u)\,t + a(1 + a\alpha u)\right]p(t)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\left[(r+2)\alpha u t^{r+1} - \right.$$
$$\left. - (r+1)(1 + 2a\alpha u)t^r + ra(1 + a\alpha u)t^{r-1}\right]p(t)dt = \qquad .$$
$$= \int_{-\infty}^{\infty}\left\{\alpha(1 + \gamma u - 2u)t^{r+1} + \left[1 + 2a\alpha u - (\gamma + a\alpha\gamma u + a\alpha)\right]t^r\right\}p(t)dt.$$

Здесь первое слагаемое обращается в нуль на концах распределения, поскольку значения плотности $p(t) \to 0$.

Если начало координат перенесено в центр распределения $\nu_1$, то переменная $t$ обозначает отклонение случайной величины от ее среднего значения и поэтому интегралы вида

$$\int_{\infty}^{\infty} t^r p(t)dt,$$

входящие в последнее уравнение, представляют собой центральные моменты $\mu_r$ распределения (7.3.1) при $\beta = 1$.

Следовательно, последнее уравнение можно представить в виде

$$ra(1 + a\alpha u)\mu_{r-1} - (2ra\alpha u + \gamma + a\alpha\gamma u + a\alpha)\mu_r + \alpha(1 + \gamma u + ru)\mu_{r+1} = r\mu_r. \qquad (4.12)$$

Учитывая, что $\mu_0 = 1, \mu_1 = 0$, из (4.3.3) при $r = 0, 1, 2, 3$ найдем

$$\left.\begin{array}{llll} 0 & -(\gamma + a\alpha\gamma u + a\alpha) & +0 & = 0 \\ \alpha(1 + a\alpha u) & -0 & +\alpha(1 + \gamma u + u)\mu_2 & = 0 \\ 0 & -4a\alpha u\mu_2 & +\alpha(1 + \gamma u + 2u)\mu_3 & = 2\mu_2 \\ \alpha(1 + a\alpha u)\mu_2 & -6a\alpha u\mu_3 & +\alpha(1 + \gamma u + 3u)\mu_4 & = 3\mu_3 \end{array}\right\} \qquad (4.13)$$

Из первого уравнения системы уравнений (4.13) получим

$$a = -\frac{\gamma}{\alpha(1 + \gamma u)} = -\nu_1, \qquad (4.14)$$

т.е. параметр *a* по абсолютной величине равен математическому ожиданию случайной величины *T*.

Решая далее систему уравнений (4.3.4), найдем значения параметров распределений I-III типов

$$\left.\begin{aligned} &Aa^2 + Ba + C = 0;\ a_{12} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A};\\ &u = -\frac{AD}{6CE};\ \alpha = -\frac{6aCE}{D^2};\quad \gamma = \frac{6a^2E}{D};\ 1 = \bar{t} + a \end{aligned}\right\} \qquad (4.15)$$

где *l* – параметр сдвига (см. табл. 6.5.1);

$$\left.\begin{aligned} &A = 2\mu_2\mu_4 - 6\mu_2^3 - 3\mu_3^2\\ &B = \mu_3(3\mu_2^2 + \mu_4)\\ &C = \mu_2(4\mu_2\mu_4 - 3\mu_3^2)\\ &D = C - a^2A = 2c + aB = -a(2aA + B)\\ &E = \mu_2\mu_4 - \mu_2^3 - \mu_3^2 \end{aligned}\right\} \qquad (4.16)$$

Если разделить величины *A,..., E* на $\mu_2^3$ и принять обозначения $\beta_1 = \mu_3^2/\mu_2^3, \beta_2 = \mu_4/\mu_2^2$, введенные К. Пирсоном для показателей асимметрии и островершинности, то получим:

$$\left.\begin{aligned} &A^* = 2\beta_2 - 3\beta_1 - 6\\ &B^* = \frac{\mu_3}{\mu_2}(3 + \beta_2)\\ &C^* = \mu_2(4\beta_2 - 3\beta_1)\\ &D^* = C^* - a^2A^* = 2c^* + aB^* = -a(2aA^* + B^*)\\ &E^* = \beta_2 - \beta_1 - 1 \end{aligned}\right\} \qquad (4.17)$$

Величины *A*,...,E** могут использоваться в формулах (4.3.6) вместо величин *A,...,E*.

Выразим величины $\beta_1$ и $\beta_2$ через параметры распределения (4.3.1) при $|\beta| = 1$. Уравнение (4.3.3) с учетом (4.3.5) позволяет записать рекуррентную формулу для центральных моментов распределений I-III типов

$$\mu_{r+1} = r\frac{\gamma\mu_{r-1} + \alpha(1 - \gamma u)(1 + \gamma u)\mu_r}{[\alpha(1 + \gamma u)]^2(1 + \gamma u + ru)}. \qquad (4.18)$$

Из (4.3.9) при $r = 1,2,3$ имеем [9]

$$\left.\begin{aligned} &\mu_2 = \frac{\gamma}{[\alpha(1 + \gamma u)]^2(1 + \gamma u + u)}\\ &\mu_3 = \frac{2(1 - \gamma u)\mu_2}{\alpha(1 + \gamma u)(1 + \gamma u + 2u)}\\ &\mu_4 = 3\frac{\gamma\mu_2 + \alpha(1 - \gamma u)(1 + \gamma u)\mu_3}{[\alpha(1 + \gamma u)]^2(1 + \gamma u + 3u)} \end{aligned}\right\} \qquad (4.19)$$

Выразим с помощью формул (7.3.10) показатели $\beta_1$ и $\beta_2$ через параметры формы *k, u* распределений I-III типов:

$$\left.\begin{aligned}\beta_1 &= \frac{4(1-\gamma u)^2(1+\gamma u+u)}{\gamma(1+\gamma u+2u)^2}\\ \beta_2 &= 3\frac{1+\gamma u+u}{1+\gamma u+3u}\left[1+\frac{2(1-\gamma u)^2}{\gamma(1+\gamma u+2u)}\right]\end{aligned}\right\} \qquad (4.20)$$

Обозначим первый сомножитель в формуле для $\beta_2$ через $L$ и назовем его "критерием $L$" [9]:

$$L = 3\frac{1+\gamma u+u}{1+\gamma u+3u}. \qquad (4.21)$$

Величину $L$ можно выразить через показатели $\beta_1$, $\beta_2$. Используя формулы (7.3.6), (7.3.8), из (7.3.12) найдем

$$L = \frac{4\beta_2 - 3\beta_1}{4+\beta_1}. \qquad (4.22)$$

Из (4.3.13) следует, что при $\beta_1 = 0$ справедливо равенство $L = \beta_2$.

Таким образом, в случае симметричных распределений критерий $L$ есть не что иное, как показатель островершинности.

Из (4.20) следует, что критерий $L$ в случае распределений I типа задан на интервале $1<L<3$; для распределений II типа $L=3$, а для распределений III типа $L>3$.

Поскольку показатель асимметрии $\beta_1=0$ при $\gamma u = 1$, что видно из (4.27), то распределения I типа при условии $\gamma = 1/u$ являются симметричными. Для них критерий $L$ (обозначим его $L_c$) равен

$$L_c = 3\frac{2+u}{2+3u}. \qquad (4.23)$$

Из формул (4.18) и (4.3.20) следует, что центральный момент 4-го порядка и критерий $L$ существуют при условии $(\gamma+3)\ u > -1$. А это значит, что по классическому методу моментов может быть найдена лишь незначительная часть выравнивающих распределений III типа, для которых выполняется неравенство

$$u > -\frac{1}{\gamma+3}. \qquad (4.24)$$

Например, при $\gamma=5$ параметр $u > -0{,}125$. Все остальные распределения III типа, а также распределения IV и V типов остаются за пределами применимости классического метода моментов.

## Тема 5. Универсальный метод моментов вычисления закона распределения и оценок параметров

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 1983г.

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 (3.8), которую представим в виде (при $\gamma = k\beta$)

$$p(t) = Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \qquad (5.1)$$

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

$$tp(t) = p(\ln t) = Ne^{k\beta \ln t}\left(1-\alpha u e^{\beta \ln t}\right)^{\frac{1}{u}-1}. \qquad (5.2)$$

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

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

$$p(x) = Ne^{k\beta x}\left(1-\alpha u e^{\beta x}\right)^{\frac{1}{u}-1}, \qquad (5.3)$$

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

$$p(y) = \frac{N(\ln y)^{k\beta-1}}{y}\left[1-\alpha u(\ln y)^{\beta}\right]^{\frac{1}{u}-1} \qquad (5.4)$$

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

$$y \ln y\ p(y) = Ne^{k\beta \ln \ln y}\left(1-\alpha u e^{\beta \ln \ln y}\right)^{\frac{1}{u}-1}. \qquad (5.5)$$

Приняв далее обозначения $\ln \ln y = x$, $y \ln y\ p(y) = p(\ln \ln y) = p(x)$, получим плотность (5.3).

Далее так же, как и в классическом методе моментов, центральные моменты $\mu_2, \mu_3, \mu_4$, 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 (5.3). При этом показатели $\beta_1$ и $\beta_2$ 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 $\beta_1^*, \beta_2^*$ теоретическим $\beta_1$, $\beta_2$, устанавливаем тип выравнивающей кривой распределения и находим оценки двух параметров формы *k, u*. Оценки двух других параметров – β, α (или произведения α*u*) 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 $p(x)$

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

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

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

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

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

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

$$p(x) = \frac{\beta(-\alpha u)^k \Gamma\left(1 - \frac{1}{u}\right)}{\Gamma(k)\Gamma\left(1 - \frac{1}{u} - k\right)} \frac{e^{k\beta x}}{\left(1 - \alpha u e^{\beta x}\right)^{1-\frac{1}{u}}} \qquad (5.6)$$

Как отмечалось выше, этими распределениями можно дополнить первую (дополнительную) систему распределений (см. табл.6.1) при условии $B^2 - 4AC < 0$.

Выразим центральные моменты (2 – 4)-го порядков и начальный момент 1-го порядка (математическое ожидание) через параметры распределения (5.6).

Используя теорию производящих функций, для обобщений плотности (5.6) получим:

$$\left.\begin{aligned} \nu_1 &= \frac{1}{\beta}\left[\Psi(k) - \Psi(k') - \ln(-\alpha u)\right] \\ \mu_2 &= \frac{1}{\beta^2}\left[\Psi'(k) + \Psi'(k')\right] \\ \mu_3 &= \frac{1}{\beta^3}\left[\Psi''(k) - \Psi''(k')\right] \\ \mu_4 &= 3\mu_2^2 + \frac{1}{\beta^4}\left[\Psi'''(k) + \Psi'''(k')\right] \end{aligned}\right\} \qquad (5.7)$$

где $k'=1-1/u-k$.

Показатели асимметрии $\beta_1$ и островершинности $\beta_2$ равны

$$\left.\begin{aligned} \beta_1 &= \frac{\left[\Psi''(k) - \Psi''(k')\right]^2}{\left[\Psi'(k) + \Psi'(k')\right]^3} \\ \beta_2 &= 3 + \frac{\Psi'''(k) + \Psi'''(k')}{\left[\Psi'(k) + \Psi'(k')\right]^2} \end{aligned}\right\} \qquad (5.8)$$

Заменяя показатели $\beta_1$, $\beta_2$ их оценками, из системы двух уравнений (5.4.8) 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 β и α (или произведения α*u*) введем случайную величину *Z*, которая связана со случайной величиной *X* зависимостью $Z = -\alpha u e^{\beta X}$ (см. формулу (5.6)) 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

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

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

или $\mu_r = \mu_r^{(Z)} / \beta^r$.

С учетом полученных равенств первые две формулы из четырех формул (5.4.7) можно переписать в виде

$$\left.\begin{aligned} \nu_1 &= \frac{1}{\beta}\left[\nu_1^{(z)} - \ln(-\alpha u)\right] \\ \mu_2 &= \frac{\mu_2^{(z)}}{\beta^2} \end{aligned}\right\}, \qquad (5.7')$$

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 (7.4.7′) оценки параметра β и произведения α*u* равны

$$\beta = \sqrt{\frac{\mu_2^{(z)}}{\mu_2}}, \qquad (5.9)$$

$$\alpha u = -e^{\nu_1^{(z)} - \beta\nu_1}, \qquad (5.10)$$

где $\nu_1 = M(X)$. При вычислении оценок β и α*u* центральный момент второго порядка случайной величины *X* следует заменить его оценкой $\mu_2^*$ (выборочной дисперсией), а *M(X)* – выборочным средним $\nu_1^* = \bar{x}$.

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

## Тема 6. 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 ($\mu_3^*$, $\mu_4^*$) сильно зависят от значений частот на концах распределения.

Поэтому автором обобщенных распределений был разработан общий устойчивый метод оценивания параметров , 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

$$\left.\begin{aligned} B &= M\left[p(x)(x-M(x))\right] = f(k,u) \\ H &= S_3 / S_1^3 = f(k,u) \end{aligned}\right\}, \qquad (6.1)$$

где

$$S_r = M\left[p(x)\right]^r = f(\beta,k,u). \qquad (6.2)$$

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 плотности $p(x)$. 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

$$\beta = \frac{S_1}{S_1^{(z)}}. \qquad (6.3)$$

Тогда $\gamma = k\beta$.

Оценки параметра $\alpha$ для распределений II, II′ типов и произведения $\alpha u$ для остальных типов равны [12]:

$$\left.\begin{array}{ll} \text{Типы I, I}': & \alpha u = e^{\pm\left(\nu_1^{(z)} - \beta\nu_1\right)} \\ \text{Типы II, II}': & \alpha = e^{\pm\left(\nu_1^{(z)} - \beta\nu_1\right)} \\ \text{Типы III - V}: & \alpha u = -e^{\nu_1^{(z)} - \beta\nu_1} \end{array}\right\} \qquad (6.4)$$

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

Типы I, I′:

$$\left.\begin{array}{l} \nu_1^{(z)} = \pm\left[\Psi(k) - \Psi\left(k + \dfrac{1}{u}\right)\right] \\ S_1^{(z)} = \dfrac{1}{2\sqrt{\pi}} \dfrac{2\left(k + \dfrac{1}{u}\right) - 1}{\dfrac{2}{u} - 1} \cdot \dfrac{g(k)g\left(\dfrac{1}{u}\right)}{g\left(k + \dfrac{1}{u}\right)} \end{array}\right\} \qquad (6.5)$$

Типы II, II′:

$$\nu_1^{(Z)} = \pm\Psi(k); \quad S_1^{(Z)} = \frac{g(k)}{2\sqrt{\pi}} \qquad (6.6)$$

Типы III-V:

$$\left.\begin{array}{l} \nu_1^{(z)} = \Psi(k) - \Psi\left(1 - \dfrac{1}{u} - k\right) \\ S_1^{(z)} = \dfrac{1}{2\sqrt{\pi}} \dfrac{g(k)\ g\left(1 - \dfrac{1}{u} - k\right)}{g\left(1 - \dfrac{1}{u}\right)} \end{array}\right\} \qquad (6.7)$$

Величина

$$g(k) = \frac{\Gamma\left(k + \dfrac{1}{2}\right)}{\Gamma(k)} \qquad (6.8)$$

может быть вычислена по приближенным формулам:

- при $x > 4$

$$g(x) \approx \sqrt{x}\, e^{-\frac{1}{8x} + \frac{1}{192x^3} - \frac{1}{640x^5} + \ldots}; \qquad (6.9)$$

- при $0 < x < 4$

$$g(x)=\frac{g(x+n)}{\prod_{i=1}^{n}\left[1+\frac{1}{2(x+i-1)}\right]}, \tag{6.10}$$

где $n \geq 4$;

$$g(x+n)=\sqrt{x+n}\,e^{-\frac{1}{8(x+n)}+\frac{1}{192(x+n)^3}-\ldots} \tag{6.11}$$

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 $\nu_1^*, S_1^*, B^*, H^*$ 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:

$$\left.\begin{aligned} &\nu_1^*=\bar{x}=\sum_{i-1}^{n}x_i p_i h_i \\ &S_1^*=\sum_{i-1}^{n}p_i^2 h_i,\ S_3^*=\sum_{i-1}^{n}p_i^4 h_i \\ &B_1^*=\sum_{i-1}^{n}x_i p_i^2 h_i-\nu_1^* S_1^*;\ \mathrm{H}^*=\frac{S_3^*}{(S_1^*)^3} \end{aligned}\right\} \tag{6.12}$$

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 (6.12) можно выразить через абсолютные частоты $m_i$:

$$\left.\begin{aligned} &\nu_1^*=\bar{x}=\sum_{i-1}^{n}x_i\frac{m_i}{M} \\ &S_1^*=\sum_{i-1}^{n}\left(\frac{m_i}{M}\right)^2\frac{1}{h_i};\ S_3^*=\sum_{i-1}^{n}\left(\frac{m_i}{M}\right)^4\frac{1}{h_i^3} \\ &B^*=\sum_{i-1}^{n}x_i\left(\frac{m_i}{M}\right)^2\frac{1}{h_i}-\nu_1^* S_1^*;\ H^*=\frac{S_3^*}{(S_1^*)^3} \end{aligned}\right\} \tag{6.13}$$

Показатель островершинности $H^*$ при $h_i$ = const примет вид

$$H^* = M^2 \frac{\sum_{i=1}^{n} m_i^4}{\left(\sum_{i=1}^{n} m_i^2\right)^3}, \qquad (6.14)$$

т.е. ширина интервала не входит в формулу (6.14). 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:

$$\left.\begin{aligned} \nu_1^* &= \overline{\ln t} = \sum_{i-1}^{n} \ell n t_i \frac{m_i}{M};\ \ S_1^* = \sum_{i-1}^{n} t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} \\ S_3^* &= \sum_{i-1}^{n} t_i^3 \left(\frac{m_i}{M}\right)^4 \frac{1}{h^3};\ H^* = \frac{S_3^*}{\left(S_1^*\right)^3} \\ B^* &= \sum_{i-1}^{n} t_i \ln t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} - \nu_1^* S_1^* \end{aligned}\right\} \qquad (6.15)$$

При $h_i$ = const

$$H^* = M^2 \frac{\sum_{i=1}^{n} t_i^3 m_i^4}{\left(\sum_{i=1}^{n} t_i m_i^2\right)^3}. \qquad (6.16)$$

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

$$Z = \alpha u e^{\beta X}. \qquad (6.17)$$

Тогда плотность *p(z)* будет равна

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

Поскольку на основании (6.17)

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

то

$$\frac{dx}{dz} = \frac{1}{\beta z},\ p(z) = \frac{p(x)}{\beta z}, \qquad (6.18)$$

откуда имеем замечательное равенство

$$\beta z p(z) = p(x). \qquad (6.19)$$

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 (6.19) 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

$$\boldsymbol{\beta^r s_r^{(z)} = s_r}. \qquad (6.20)$$

Формула (7.5.20) 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

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

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 $\alpha u$ вычисляется по тем же формулам, что и в случае универсального метода моментов.

Если в качестве показателей асимметрии и островершинности использовать величины

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

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*).

## Тема 7. Ранговые распределения в библиотечно-информационной деятельности

### 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, наиболее 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.

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 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.

В связи с вышесказанным целесообразно перейти к другой форме представления ранговых распределений, а именно, $rp_r = f(lnr)$. По оси ординат будем откладывать произведение ранга слова (журнала) на его относительную частоту (или долю статей), а по оси абсцисс – натуральный логарифм ранга.

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.

Универсальный закон рассеяния публикаций

Глубокое изучение любой дисциплины, допускающей применение количественных методов исследования, должно сопровождаться построением и использованием математических или вероятностно-статистических моделей. Так, в информатике и математической лингвистике широко известны такие математические модели, как закон Дж.Ципфа

$$p_r = \frac{k}{r^{\gamma}}, \qquad (7.1)$$

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.

Такими общими законами могут служить построенные В. В. Нешитым 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 и математическая лингвистика 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 смысле С. Бредфорда, а также установим в самом общем виде законы рассеяния и старения публикаций.

В 1948г. С. Бредфорд дал окончательную формулировку открытого им в 1934г. закона рассеяния публикаций в периодических изданиях. Приведем формулировку этого закона: «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 различным отраслям знания. Эти распределения для каждого статистического вариационного ряда имеют свои параметры.

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

$$p(t)=Nt^{k\beta-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1} \qquad (7.2)$$

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: $rp_r = f(lnr)$. 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

$$p(x)=Ne^{k\beta x}\left(1-\alpha u e^{k\beta x}\right)^{\frac{1}{u}-1}. \qquad (7.3)$$

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 (7.2) на величину $t$ и записать выражение $t^{\beta}$ в виде $e^{\beta \ln t}$, 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

$$x_C - x_A = x_B - x_C . \qquad (7.4)$$

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

$$lnt_C - lnt_A = lnt_B - lnt_C , \qquad (7.5)$$

откуда имеем равенство

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \qquad (7.6)$$

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

$$t_A : t_C : t_B = t_A(1 : n : n^2) \qquad (7.7)$$

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

$$t_Я : t_I : t_{II} = t_A(1 : (n-1) : (n-1)n.. \qquad (7.8)$$

Как видим, формулировка Бредфорда является комбинацией из двух точных формул (7.7) и (7.8). 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

$$t_C = \left( \frac{k}{\alpha(1 + ku - u)} \right)^{1/\beta}. \qquad (7.9)$$

Величина *n* задается формулой

$$n = \left[ 1 + \frac{1 - u \mp \sqrt{[4k(1 + ku - u) + (1 - u)](1 - u)}}{2k(1 + ku - u)} \right]^{1/\beta}. \qquad (7.10)$$

В формуле (7.10) 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

$$F(t)=1-e^{-\alpha t^{\beta}}, \qquad (7.11)$$

которая следует из формулы (2.2.6) при $u \to 0$. Тогда из формул (7.9) и (7.10) при $k = 1$ имеем равенства:

$$t_c=\left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}, \quad n=\left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \qquad (7.12)$$

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 (7.8).

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.

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 $p(x)$, 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)

$$x_c = \frac{1}{\beta} \ln \frac{k}{\alpha(1 + ku - u)}, \tag{7.13}$$

$$x_{A,B} = x_C \mp \ln n, \tag{7.14}$$

где величина *n* рассчитывается по прежней формуле (7.10).

## Тема 8. Построение системы дискретных распределений

Построение системы дискретных распределений по кривым роста новых событий.

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 [18]

$$y = \frac{1}{\alpha u}\left[1 - \left(1 - \alpha(u-1)x\right)^{\frac{u}{u-1}}\right], \tag{8.1}$$

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 (8.1). В частности, при u→1 из (8.1) следует формула

$$y = \frac{1}{\alpha}\left(1 - \frac{1}{e^{\alpha x}}\right). \tag{8.2}$$

При u→0 из (1) следует кривая II типа

$$y = \frac{1}{\alpha}\ln(1+\alpha x). \qquad (8.3)$$

При u < 0 имеем кривую III типа. Она задается той же формулой (8.1).

Между кривой роста разных событий и статистической структурой выборки существует взаимосвязь, установленная В.М. Калининым [3]

$$y_m = (-1)^{m+1}\frac{x^m}{m!}\frac{d^m y}{dx^m}. \qquad (8.4)$$

По формуле (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).

Формулы (8.1) и (8.4) позволяют также построить систему дискретных распределений.

***Построение системы дискретных распределений***

*Распределения I типа (u > 0).*

Продифференцируем выражение (8.1) m раз по x и подставим m-ю производную в (8.4). В результате получим формулу, позволяющую вычислять число событий с частотой m, т.е. $y_m$ при числе испытаний x

$$y_m = \frac{y_{m=0}}{m!}\left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right], \quad m = 1, 2, \ldots, \qquad (8.5)$$

где

$$y_{m=0} = \frac{1}{\alpha u}\left[1+\alpha(1-u)x\right]^{\frac{u}{u-1}}. \qquad (8.6)$$

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, составляющих 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 $y_{m=0}$ ):

$$p_m = \frac{p_{m=0}}{m!}\left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right], \quad m = 1, 2, \ldots, \qquad (8.7)$$

где

$$p_{m=0} = \left[1+\alpha(1-u)x\right]^{\frac{u}{u-1}}. \qquad (8.8)$$

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).

*Распределения II типа (u → 0).* В данном случае кривая роста разных событий задается формулой (8.3), на основании которой и формулы В.М. Калинина (8.4) имеем

$$y_m = \left(\frac{\alpha x}{1+\alpha x}\right)^m \frac{1}{\alpha m}, \quad m = 1,2,... \qquad (8.9)$$

Разделив (8.9) на (8.3), получим

$$\frac{y_m}{y} = p_m = \left(\frac{\alpha x}{1+\alpha x}\right)^m \frac{1}{m\ln(1+\alpha x)}. \qquad (8.10)$$

Последнее распределение известно как распределение Фишера по логарифмическому ряду и находит широкое применение в биологии.

*Распределения III типа (- ∞ < u < ∞).* Кривая роста разных событий задается общей формулой (8.1). Из (8.1) и (8.4) имеем

$$y_m = \frac{y_{m=1}}{m!}\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1} \prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m = 2,3,..., \qquad (8.11)$$

где

$$y_{m=1} = x\left[1+\alpha(1-u)x\right]^{\frac{1}{u-1}}. \qquad (8.12)$$

Разделив $y_m$ на $y$, получим выражение для вероятности $p_m$

$$p_m = \frac{p_{m=1}}{m!}\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1} \prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m = 2,3,..., \qquad (8.13)$$

где

$$p_{m=1} = \frac{-\alpha u x}{(1+\alpha(1-u)x)\left[1-(1+\alpha(1-u)x)^{\frac{u}{1-u}}\right]}. \qquad (8.14)$$

Оценивание параметров дискретных распределений.

Для установления типа аппроксимирующего дискретного распределения введем критерий

$$HD = \frac{\frac{x}{y}\ln\frac{x}{y_{m=1}}}{\frac{x}{y_{m=1}} - 1}. \qquad (8.15)$$

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. Их можно найти по методу моментов. В случае распределений I типа

$$\alpha = \sum_{m\geq 1}\left(\frac{m}{x}\right)^2 y_m - \frac{1}{x}, \qquad (8.16)$$

$$u = \frac{1}{\alpha n}, \qquad (8.17)$$

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 (8.3)

$$\alpha_{i+1} = \frac{1}{y}\ln(1+\alpha_i x), \qquad (8.18)$$

где $y = \sum_{m\geq 1} y_m; \quad \alpha_i$ - 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

$$u_{i+1} = (1-u_i)\frac{x}{y}\frac{1-\left(\frac{y_{m=1}}{x}\right)^{u_i}}{\left(\frac{x}{y_{m=1}}\right)^{1-u_i}-1}. \qquad (8.19)$$

Тогда оценка параметра α равна

$$\alpha = \frac{1}{uy}\left[1-\left(\frac{y_{m=1}}{x}\right)^u\right] = \frac{1}{(1-u)x}\left[\left(\frac{x}{y_{m=1}}\right)^{1-u}-1\right]. \qquad (8.20)$$

Таким образом, для оценивания параметров α, u достаточно знать три величины: x, y, $y_{m=1}$.

Отметим, что формулы (8.16), (8.19), (8.20) 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

$$y_{m+1} = y_m \frac{\alpha x[u + m(1-u)]}{[1 + \alpha(1-u)x](m+1)}, \qquad (8.21)$$

которая справедлива для распределений всех трех типов. Вначале по формуле (8.12) вычисляется количество событий с частотой m = 1, т.е. $y_{m=1}$ . Далее по формуле (8.21) 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 (8.6).

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 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 (15) критерий HD. Он оказался равным 1,854.

Поскольку HD>1, то искомое распределение относится к III типу. Далее по формулам (8.19), (8.20) вычисляем оценки параметров u, α: u = -0,601736; α = 0,00645759. Частотный спектр описывается формулой (8.21).

Кривая роста разных ключевых слов описывается уравнением (1), которое при найденных оценках параметров u, α примет вид

$$y = 257.35\left[(1 + 0.0103435\,x)^{0.375677} - 1\right] \qquad (8.22)$$

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 настоящей работе.

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

Таблица 8.1 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 |
| 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 |

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.

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.

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. 8.1 расчетные данные о частоте ключевых слов и их количестве, вычисленном по дискретному распределению 3-го типа автора. По этим данным построим график зависимости $LNYm = f(LNm)$.

Из построенного графика видно, что эта зависитмость близка к уравнению прямой, т.е. фактически имеем закон Лотки. Но теоретическая указанная зависимость имеет точку перегиба. 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 доказательство этого факта приведем еще один график (рис. 8.2). В этом случае статистические данные полные – учтены словосочетания с частотами от1 до 107.

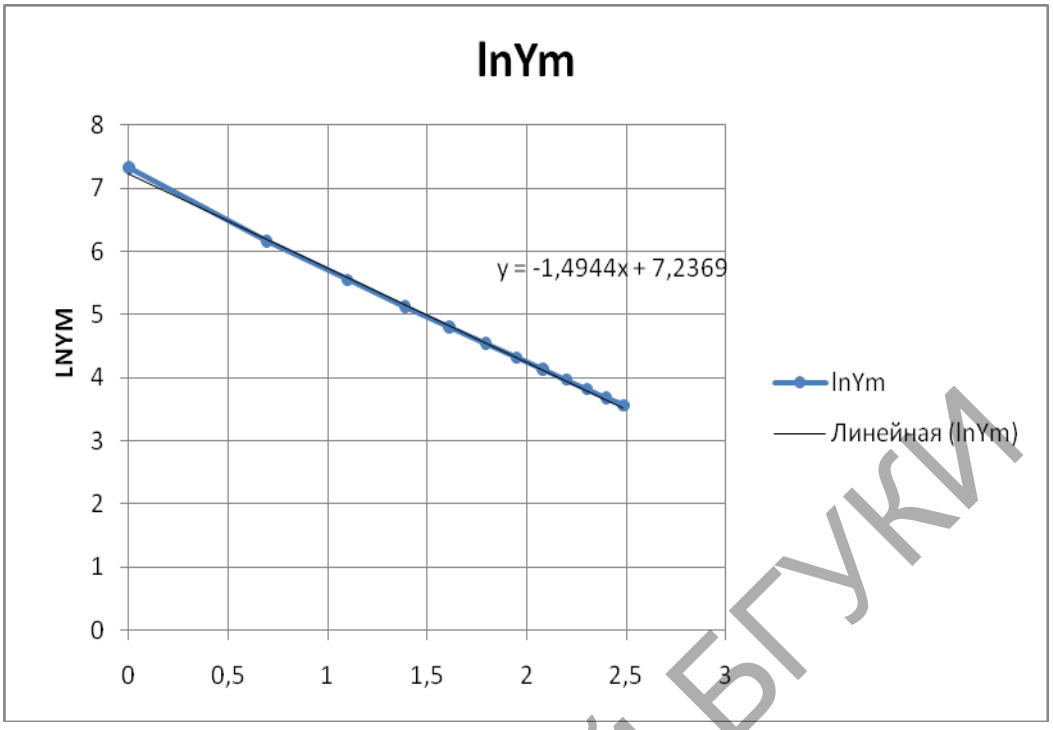



**Рисунок 8.1 Диаграмма Лотки** $\mathrm{Ym} = \frac{\mathrm{Y}_1}{m^{1,4944}}$; $\mathrm{LNYm} = \mathrm{LNY}_1 - 1{,}4944\,LNm$

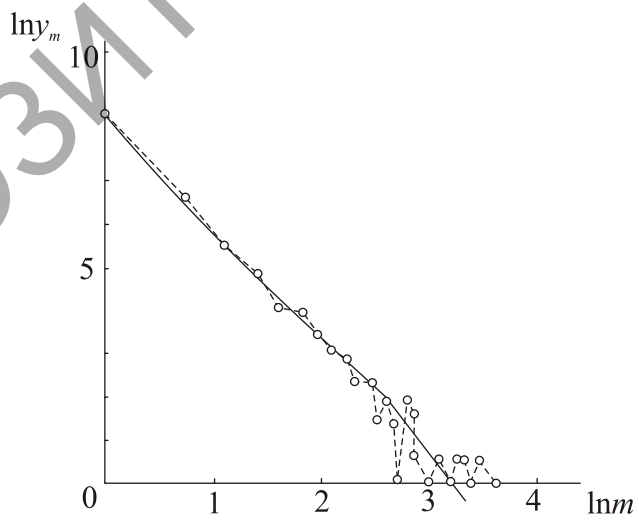



**Рис. 8.2. Распределение словосочетаний в английском газетном тексте по дискретному распределению 3-го типа**

# 3. ПРАКТИЧЕСКИЙ РАЗДЕЛ

## 3.1. МАТЕРИАЛЫ К СЕМИНАРСКИМ ЗАНЯТИЯМ

Семинар № 1
Тема: «Статистическое исследование»

1. Этапы статистического исследования.
2. Генеральная и выборочная совокупности.
3. Простой статистический ряд. Вариационный ряд. Интервальный ряд распределения.
4. Группировка статистических данных.
5. Представление статистических данных.
6. Статистические показатели.

Библиографический список литературы:

1. Логинова, С. Л. Общая теория статистики: конспект лекций / С. Л. Логинова. – Екатеринбург: Изд-во Рос. гос. проф.-пед. ун-та, 2011. –90 с.
2. Мотульский, Р. С. Библиотечная статистика: проблемы и решения / Р. С. Мотульский // Библиотечное дело – XXI век. – 2002. – № 3. – С. 28 – 64.
3. Нешитой, В. В. Информетрия: математические модели и методы исследования / В. В. Нешитой. – Минск : БГУКИ, 2017. – 274 с.; То же [Электронный ресурс]. – Режим доступа: http://repository.buk.by/123456789/15166 . – Дата доступа: 18.05.2022.

*При подготовке к семинару студенты могут обращаться к другим профильным источникам, самостоятельно подбирать информацию для подготовки к предлагаемым вопросам.*

Семинар № 2.

Тема: «Методы количественного анализа документных информационных потоков»

1. Становление методов количественного исследования документного потока.
2. Закон обратного квадрата А. Лотки.
3. Закон распределения Дж. Ципфа.
4. Закон рассеяния С. К. Брэдфорда.

5. Закон ранговых распределений Вейбулла.

Библиографический список литературы:

1. 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.

2. Нешитой, В. В. Информетрия: математические модели и методы исследования / В. В. Нешитой. – Минск : БГУКИ, 2017. – 274 с.; То же [Электронный ресурс]. – Режим доступа: http://repository.buk.by/123456789/15166 . – Дата доступа: 18.05.2022.

3. Нешитой, В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие / В. В. Нешитой. – Минск : РИВШ, 2012. – 134 с.

*При подготовке к семинару студенты могут обращаться к другим профильным источникам, самостоятельно подбирать информацию для подготовки к предлагаемым вопросам.*

## 3.2. ТЕМАТИКА И МЕТОДИКА ВЫПОЛНЕНИЯ ПРАКТИЧЕСКИХ РАБОТ

**Практическая работа 1**. Библиометрический анализ информационных ресурсов

*Цель работы:* Закрепить теоретические знания по проведению библиографического анализа

*Задание:* Разработать структуру библиометрического анализа для изучения информационных ресурсов.

*Методика выполнения*: Студент должен выбрать раздел реферативного журнала «Информатика», для которого будет разрабатывать критерии библиометрического анализа. Сформулировать цель анализа источников, определить критерии библиометрического анализа.

По разработанным критерием провести библиометрический анализ раздела.

*Выполнение заданий контролируется преподавателем в ходе занятия, а также проверяется письменная работа.*

**Практическая работа 2.** Методы количественного анализа, основывающиеся на исследовании вторичных источников информации.

*Цель работы*: Закрепить теоретические знания и сформировать умения практического применения закона рассеяния публикаций по изданиям С. К. Брэдфорда.

*Задание:* Выявить закономерность рассеяния публикаций.

*Методика выполнения:* Студент получает (выбирает их предложенных) тему и реферативный журнал ВИНИТИ. На основе анализа публикаций необходимо в показать рассеяние количество журналов и статей согласно формулировке Бредфорда

$t_я : t_I : t_{II} = 1 : n : n$

В заключении необходимо сделать выводы о тенденциях рассеяния или концентрации потока документов в периодических изданиях. Результаты анализа представить в письменнойй форме.

*Выполнение заданий контролируется преподавателем в ходе занятия, а также проверяется письменная работа.*

**Практическая работа 3**. Статистические моделирование библиотечного фонда

*Цель работы:* Закрепить теоретические знания и сформировать умения практического применения метода ранговых распределений Вейбулла для вычисления информационной полноты комплектования библиотечного фонда, а также оценка его оптимального объема.

*Задание:* Выявить границу ядра и зоны 1-го рассеивания и зны оптимального объема фонда библиотеки.

*Методика выполнения:* Студент получает в программе Excel статистические данные использования информационных ресурсов библиотеки.

Информационные ресурсы упорядочиваются по убыванию частоты обращений к ним (фактическая частота). Каждому наименованию информационного ресурса присвоен ранг, порядковый номер документа в списке по убывающим частотам. Далее необходимо определить накопленную относительную частоту, функцию распределения F(r), логарифм рангов F(r)+ r, сумму абсолютной частоты, функцию распределений lg(lg(1/1-F(r))). Статистические ранговые распределения представить в виде графика.

Высчислить следующие параметры:

Мода $t_c$ (ядро и первая зона рассеяния):
$t_c = (1/\alpha)^{(1/\beta)}$.

Величина n, где n – отношение оптимального объема фонда информационных ресурсов к ядру и первой зоне рассеяния документов:

$\left( 2 \right)$

Абсцисса точки А, параметры зоны ядра:
$t_A = \overline{n}$

Абсцисса точки В, оптимальный объем информационных ресурсов:
$t_B = t_C$ x n

По итогу работы предоставить показатели F($t_A$) = 0,3175, или ≈32%; F($t_C$) = 0,6321, или ≈63%; F($t_B$) = 0,92705, или ≈93%. Дать анализ информационных источников третьей зоны рассеивания.

*Выполнение заданий контролируется преподавателем в ходе занятия, а также проверяется письменная работа.*

## 4. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

### 4.1. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ОРГАНИЗАЦИИ И ВЫПОЛНЕНИЮ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Самостоятельная работа студентов позволяет закрепить знания, полученные на занятиях по учебной дисциплине «Статистические методы библиотечно-информационной деятельности», систематизировать информацию, расширить представления о статистических методах анализа библиотечно-информационной деятельности. Контроль за самостоятельной работой студентов осуществляется в ходе выполнения и проверки практических работ.

Критериями оценки результатов самостоятельной работы студентов являются:

– уровень усвоения студентом учебного материала;

– умение студента использовать теоретические знания при выполнении практических задач;

–уровень сформированности общих и профессиональных компетенций;

– умение использовать статистические методы;

– уровень оформления работы.

## 4.2. СОДЕРЖАНИЕ И ФОРМЫ КОНТРОЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Организация самостоятельной работы студентов предусматривает работу с научной и учебно-методической литературой, подготовку к семинарским занятиям и зачету.

| № | тема | содержание задания | вид контроля |
|---|---|---|---|
| **1** | Некоторые понятия теории вероятностей и математической статистики | Создание терминологического словаря по основным терминам курса | Проверка терминологических словарей |
| **2** | Классические методы оценивания параметров непрерывных распределений | Разработка статистической модели библиотечного фонда | Проверка письменной работы |
| **3** | Методы количественного анализа, основывающиеся на вторичных источниках информации | Сравнительный анализ методов количественного анализа | Проверка письменной работы |
| **4** | Ранговые распределения в библиотечно-информационной деятельности | Изучение литературы по теме. Анализ понятийного аппарата. | Информационное сообщение на семинаре. |

## 4.3. ВОПРОСЫ К ЗАЧЕТУ

1. Понятие математического ожидания случайной функции, нового события и кривой роста новых событий.
2. Связь кривой роста с законами распределения вероятностей разных и новых событий.
3. Формула В.М. Калинина для расчёта статистической структуры выборки по кривой роста новых событий.
4. Формула В.М. Калинина для восстановления кривой роста новых событий по статистической структуре выборки.
5. Порядок построения системы кривых роста и непрерывных распределений новых событий.
6. Методы построения универсальных (обобщённых) непрерывных распределений.
7. Систематизация и представление статистических данных.
8. Показатели стабильности и качества статистической выборки.
9. Семейство кривых К. Пирсона.
10. Три системы непрерывных распределений В.В. Нешитого.
11. Ранговые распределения. Закон Дж. Ципфа в семействе ранговых распределений.
12. Форма представления ранговых распределений.
13. Характерные точки кривых распределения и связь их с законами рассеяния публикаций.
14. Методы оценивания параметров: метод моментов.
15. Метод наибольшего правдоподобия.
16. Метод наименьших квадратов.
17. Универсальный метод моментов.
18. Применение системы непрерывных распределений в библиотечно-информационной деятельности.
19. Универсальный закон рассеяния публикаций.
20. Закон рассеяния С. К. Брэдфорда.
21. Закон обратного квадрата А. Лотки.
22. Универсальный закон старения публикаций.
23. Методы вычисления границ ядра и зон рассеяния публикаций.
24. Закон ранговых распределений В. Вейбулла
25. Методы построения системы дискретных распределений.
26. Классификация дискретных распределений.
27. Порядок вычисления по статистическим данным дискретного закона распределения и оценок параметров.

28. Критерий степени неравномерности появления событий.
29. Прогнозирование кривой роста новых событий и частотного спектра.
30. Расчёт достоверной части частотного словаря.

# 5.ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

## 5.1. УЧЕБНАЯ ПРОГРАММА

## ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Целью учебной дисциплины является формирование у студентов общего подхода к вопросам построения универсальных вероятностных моделей для моделирования библиотечного фонда, описания статистических закономерностей информационных потоков, методов оценивания параметров и овладение современными методами статистического анализа на базе обобщённых распределений и кривых роста.

Задачей изучения дисциплины является обучение студентов навыкам использования различных методов математико-статистического анализа в библиотечно-информационной деятельности.

Содержание учебной дисциплины предусматривает формирование следующих компетенций:

Академические компетенции:

АК-1. Уметь применять базовые научно-теоретические знания для решения теоретических и практических задач.

АК-2. Владеть системным и сравнительным анализом.

АК-3. Владеть исследовательскими навыками.

АК-4. Уметь работать самостоятельно.

АК-6. Владеть междисциплинарным подходом при решении проблем.

АК-7. Иметь навыки, связанные с использованием технических устройств, управлением информацией.

АК-8. Обладать навыками устной и письменной коммуникации.

АК-9. Уметь учиться, повышать свою квалификацию в течение всей жизни.

Социально-личностные компетенции:

СЛК-2. Быть способным к социальному взаимодействию.

СЛК-3. Обладать способностью к межличностным коммуникациям.

Профессиональные компетенции:

ПК-1. Выполнять библиотечно-информационные технологические процессы в среде современных автоматизированных библиотечно-информационных систем.

ПК-12. Изучать, анализировать и внедрять мироваой опыт иновационной деятельности библиотек и информационных центров.

ПК-14. Разрабатыать методические материалы и рекомендации, организационно-технологическую документацию.

ПК-32. Создавать аналитические и информационные продукты и услуги.

В результате изучения дисциплины студенты должны знать:

– известные и новые методы оценивания параметров;

- сущность различных подходов к нахождению по статистическим данным выравнивающих распределений и кривых роста;
- элементы теории обобщённых распределений;
- математико-статистические модели, которые могут быть использованы в библиотечно-информационной деятельности, в том числе универсальные законы рассеяния и старения публикаций.

Выпускники в пределах своей специальности должны уметь:

- использовать математико-статистические методы для построения моделей библиотечных процессов;
- вычислять законы распределения по статистическим данным;
- вычислять зоны рассеяния публикаций и этапы старения публикаций;
- прогнозировать статистические закономерности текста и информационных потоков.
- извлекать информацию из библиотечной статистики.

На изучение дисциплины «Статистические методы библиотечно-информационной деятельности» отводится 52 часа, из них 28 часов аудиторных занятий, в том числе 18 часов лекций и 10 часов практических занятий. Форма отчётности и контроля – зачёт.

## СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

*Введение*

Предмет учебной дисциплины, его цель, задачи и место в системе профессиональной подготовки специалистов библиотечно-информационной сферы.

Связь учебной дисциплины с другими учебными дисциплинами информационно-документного цикла. Объем, структура, содержание и порядок изучения учебной дисциплины. Формы самостоятельной работы. Система средств диагностики. Характеристика информационно-методического обеспечения учебной дисциплины.

*Тема 1. Функции одной переменной. Производная функция*

Функция. Способы задания функций. Основные элементарные функции и их графики. Гамма-функция. Понятие эмпирической функции.

Задачи, приводящие к понятию производной. Определение и смысл производной. Касательная. Вычисление производных. Производные элементарных функций. Правила дифференцирования.

*Тема 2. Некоторые понятия теории вероятностей и математической статистики*

Предмет теории вероятностей. Случайные события. Испытания. Относительная частота и вероятность. Виды случайных событий. Определение вероятности. Основные теоремы теории вероятностей. Непрерывные и дискретные случайные величины и их законы распределения. Числовые характеристики случайных величин.

Предмет математической статистики. Генеральная и выборочная совокупности. Простой статистический ряд. Вариационный ряд. Полигон. Гистограмма. Кумулятивная кривая. Кривая распределения. Плотность и функция распределения.

*Тема 3. Методы построения обобщенных непрерывных распределений*

Понятие математического ожидания случайной функции, нового события и кривой роста новых событий. Связь кривой роста с законами распределения вероятностей разных и новых событий. Установление статистической структуры выборки по кривой роста новых событий.

Восстановление кривой роста новых событий по статистической структуре выборки.

Формулы В. Калинина. Построение непрерывных распределений по методам: К. Пирсона, обобщения, как распределения функций случайных аргументов. Классификация распределений и кривых роста.

*Тема 4. Классические методы оценивания параметров непрерывных распределений*

Методы оценивания параметров обобщенных непрерывных распределений. Выборочный метод при исследовании случайных величин в математической статистике. Репрезентативность выборки. Порядок упорядочивания значений исследуемой случайной величины.

Метод наименьших квадратов.

Метод наибольшего правдоподобия.

Классический метод моментов.

Критерии для установления типа выравнивающей кривой по методу моментов.

*Тема 5. Универсальный метод моментов вычисления закона распределения и оценок параметров*

Универсальный метод моментов, его отличия от классического метода моментов. Критерии для установления типа выравнивающей кривой. Вычисление оценок параметров. Расширение трёх систем непрерывных распределений. Законы распределения суммы независимых случайных величин. Использование общего метода для нахождения закона распределения суммы $n$ независимых случайных величин. Центральная предельная теорема теории вероятностей. Простейшее доказательство.

*Тема 6. Общий устойчивый метод вычисления закона распределения и оценок параметров*

Неустойчивость как существенный недостаток оценивания параметров обобщенных распределений с помощью универсального метода моментов. Понятие устойчивости.

Общий устойчивый метод оценивания параметров В. В. Нешитого. Достоинства и недостатки метода. Критерии для установления типа выравнивающей кривой по устойчивому методу. Номограмма (типы кривых в координатах). Вычисление оценок параметров.

*Тема 7. Ранговые распределения в библиотечно-информационной деятельности*

Статистические распределения. Ранговые распределения. Использование статистических распределений, в том числе ранговых в научных исследованиях.

Форма представления ранговых распределений. Критерий однородности ранговых распределений. Выделение неоднородной части.

Законы распределения документальных информационных потоков. Универсальный закон рассеяния публикаций. Универсальный закон старения публикаций. Методы вычисления границ ядра и зон рассеяния публикаций. Оптимальный объем библиотечного фонда.

*Тема 8. Построение системы дискретных распределений, оценивание параметров*

Построение системы дискретных распределений по системе непрерывных распределений. Построение системы дискретных распределений по кривым роста новых событий на основе формулы В.Калинина. Оценивание параметров дискретных распределений. Графический метод установления типа распределения. Аналитический метод установления принципа распределения. Классификация дискретных распределений. Критерий степени неравномерности появления событий. Ранжирование слов по степени семантической нагрузки. Порядок установления типа выравнивающего распределения и нахождения оценок параметров.

## УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА ДИСЦИПЛИНЫ

| № п/п | Названия разделов и тем | Количество аудиторных часов | | | Количество часов УСР | Форма контроля знаний |
|---|---|---|---|---|---|---|
| | | Лекции | Практические занятия | Семинарские занятия. | | |
| 1. | Введение | 0,5 | | | | |
| 2. | **Тема 1.** Функции одной переменной. Производная функции. | 1,5 | | | | фронтальный опрос |
| 3. | **Тема 2.** Некоторые понятия теории вероятностей и математической статистики. | 2 | | | 6 | фронтальный опрос, проверка самостоятельной работы |
| 4. | **Тема 3.** Методы построения обобщенных непрерывных распределений (в т.ч. по кривым роста новых событий) | 4 | | 2 | 4 | фронтальный опрос, проверка самостоятельной работы |
| 5. | **Тема 4.** Классические методы оценивания параметров непрерывных распределений | 2 | | | 2 | выступление на семинарском занятии, проверка самостоятельной работы |
| 6. | **Тема 5.** Универсальный метод моментов вычисления закона распределения и оценок параметров | 2 | | | 2 | фронтальный опрос, проверка самостоятельной работы |
| 7. | **Тема 6.** Общий устойчивый метод вычисления закона распределения и оценок параметров | 2 | | | 4 | фронтальный опрос, проверка самостоятельной работы |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8. | **Тема 7.** Ранговые распределения в библиотечно - информационной деятельности | 2 | 4 | 2 | 4 | выступление на семинарском занятии, фронтальный опрос, проверка самостоятельной работы |
| 9. | **Тема 8.** Построение системы дискретных распределений, оценивание параметров | 2 | 2 | | 2 | фронтальный опрос, проверка самостоятельной работы |
| | **всего** | **18** | **6** | **4** | **24** | |

## ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

### Литература

*Основная*

1. Нешитой, В. В. Информетрия: математические модели и методы исследования / В. В. Нешитой. – Минск : БГУКИ, 2017. – 274 с.; То же [Электронный ресурс]. – Режим доступа: http://repository.buk.by/123456789/15166 . – Дата доступа: 18.05.2022.

2. Нешитой, В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие / В. В. Нешитой. – Минск : РИВШ, 2012. – 134 с.

3. Сазанова, Е. В. Общая теория статистики : учеб. пособие / Е. В. Сазанова. — Архангельск : САФУ, 2018. — 173 с.

*Дополнительная*

1. Белоновская, И. Л. Статистическое моделирование фондов школьной библиотеки / И. Л. Белоновская // Веснік Беларускага дзяржаўнага ўніверсітэта культуры і мастацтваў. – 2018. – № 1 (29). – С. 163–168.

2. Логинова, С. Л. Общая теория статистики: конспект лекций / С. Л. Логинова. – Екатеринбург: Изд-во Рос. гос. проф.-пед. ун-та, 2011. –90 с.

3. Мотульский, Р. С. Библиотечная статистика: проблемы и решения / Р. С. Мотульский // Библиотечное дело – XXI век. – 2002. – № 3. – С. 28 –64.

4. 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.

5. Нешитой, В.В. Математико-статистические методы анализа в библиотечно-информационной деятельности : учеб.-метод. пособие / В. В. Нешитой. – Минск: БГУ культуры и искусств, 2009. –203 с.

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

7. 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.

8. Нешитой, В.В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений : учеб.- метод. пособие / В. В. Нешитой. – Минск: Бел. гос. ун-т культуры и искусств, 2010. –61 с.

9. 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.

10. Нешитой, В.В. Моделирование кривой роста и статистической структуры словаря ключевых слов / В.В.Нешитой // Веснік Беларус. дзярж. ун-та культ. і мастацтв. – 2008. – №9. – С. 123-132.

11. Нешитой, В. В. Статистические методы анализа использования библиотечного фонда / В. В. Нешитой, Б. В. Петренко – Вестник Библиотечной Ассамблеи Евразии РГБ. – Москва, 2013. – №2. – С.84-86.

12. Нешитой, В.В. Статистическое моделирование библиотечного фонда/ В.В. Нешитой // Науч.и техн. б-ки. – Москва, 2009. – С.36-46.

13. Нешитой, В.В. Элементы теории обобщенных распределений: монография / В.В. Нешитой. – Минск: РИВШ, 2009. – 204 с.

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

15. Поллард, Дж. Справочник по вычислительным методам статистики / Дж. Поллард; пер. с англ. – М.: Финансы и статистика, 1982. – 344 с.

16. Шилов, В. В. Моделирование библиотечных фондов: история с математикой / В. В. Шилов, Т. В. Корткова // Науч.и техн. б-ки. – 2006. – № 12. – С. 6– 9.

## Рекомендованные средства диагностики результатов учебной деятельности студентов

Уровень учебныхдостижений студентов может быть определен с помощью следующих средств диагностики:

- выступления на семинарских занятиях;
- фронтальный опрос;
- проверка и обсуждение в группах результатов самостоятельных работ;
- проверка качества выполнения практических заданий;
- проверка заданий по основным разделам дисциплины.

## Методические рекомендации по организации и выполнению самостоятельной работы студентов

Самостоятельная работа студентов осуществляется при подготовке к семинарам, выполнении практических заданий.

При подготовке к семинарским занятиям студенты должны самостоятельно изучить источники. На основе анализа изученной литературы студенты готовят выступления. Семинарские занятия дают возможность закрепить теоретические знания по дисциплине, а так же способствуют развитию аналитического мышления.

Индивидуальные задания направлены на формирование умений применения статистических методов. В процессе выполнения практических заданий студент на основе полученных знаний должен решить задачу, требующую умения применять полученные знания.



## Характеристика рекомендуемых методов преподавания

Основными методами и технологиями преподавания, отвечающим целям и задачам изучения дисциплины, являются:

- деятельный метод, обеспечивающий не только формирование знаний, но и способов мышления и деятельности;
- метод проблемного обучения (проблемное изложение, вариативное изложение), реализуемый на лекционных занятиях;
- метод моделирования конкретных ситуаций;

- исследовательский метод обучения, позволяющий приобрести навыки критического мышления, самостоятельного решения поставленных учебных задач;
- коммуникативные педагогические технологии (дискуссия,
- диалог, работа в группах, обмен мнениями по результатам работы);
- информационно-коммуникационные технологии, в которых используются мультимедийные презентации, электронные образовательные ресурсы.

**Примерный перечень заданий для самостоятельной работы студентов**

1. Некоторые понятия теории вероятностей и математической статистики
2. Классические методы оценивания параметров непрерывных распределений Методы вычисления границ ядра и зон рассеяния публикаций.
3. Методы количественного анализа, основывающиеся на вторичных источниках информации
4. Ранговые распределения в библиотечно-информационной деятельности