Бирин Дмитрий Анатольевич – ФГУП НИИ «Квант»; e-mail: melnikov@linfotech.ru; 125438, г. Москва, 4-й Лихачевский пер., 15; зам. директора.

Пересыпкин Владимир Анатольевич – ФГУП «НТЦ «Орион»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, 38, стр. 1; научный консультант; к.т.н.

Мельников Сергей Юрьевич – ООО «Линфо»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, 38, стр. 1; тел.: +79037222824; зам. директора; к.ф.-м.н.

Писарев Илья Александрович – Южный федеральный университет; e-mail: melnikov@linfotech.ru; 347922, г. Таганрог, ул. Чехова, 2; инженер.

Цопкало Николай Николаевич – e-mail: melnikov@linfotech.ru; с.н.с.; к.т.н.

Birin Dmitrij Anatol'evich – FGUP NII «Kvant»; e-mail: melnikov@linfotech.ru; 125438, Moscow, 4-j Lihachevskij per., 15; deputy Director.

Peresypkin Vladimir Anatol'evich – FGUP "NTC "Orion""; e-mail: melnikov@linfotech.ru; 127018, Moscow, Obrazcova street, 38, str. 1; research consultant; cand. of eng. sc.

Melnikov Sergey Yur'evich – OOO "Lingvisticheskie I informatsionye tehnologii" (Limited Liability Company); e-mail: melnikov@linfotech.ru; 127018, Moscow, Obrazcova street, 38, str. 1; deputy Director; cand. of phys.-math. sc.

Pisarev Il'ya Aleksandrovich – Southern Federal University; e-mail: melnikov@linfotech.ru; 347922, Taganrog, Chekhova street, 2; engineer.

Copkalo Nikolaj Nikolaevich – e-mail: melnikov@linfotech.ru; senior scientist; cand. of eng. sc.

УДК 519.224.22 DOI 10.23683/2311-3103-2018-8-114-135

А.К. Мельников

ПРИМЕНЕНИЕ ТОЧНЫХ И ПРЕДЕЛЬНЫХ ПРИБЛИЖЕНИЙ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ЗНАЧЕНИЙ СТАТИСТИК ПРИ РЕШЕНИИ ЗАДАЧИ ПО ОБРАБОТКЕ ТЕКСТОВ

Рассматривается применение предельных и точных приближений распределения вероятностей значений статистик для решения задачи по отбору текстов с определенными статистическими свойствами. Для отбора текстов с равновероятным распределением входящих в них знаков используется статистический критерий согласия, в котором в качестве эталонного распределения тестовой статистики используются его различные приближения. В качестве предельных приближений используются предельные распределения, а в качестве точных приближений - Δ–точные распределения, которые отличающиеся от точных распределений не более чем на заданную величину Δ. Приведены результаты расчета Δ–точных распределений, показаны их отличия от значений предельных распределений для разных статистик. Рассмотрено понятие эффективности обработки по выделению равновероятных текстов, отражающее долю ложно отобранных текстов. Проведено сравнение значений эффективности обработки при применении точных и предельных приближений эталонных распределений тестовых статистик. Показано, что значение эффективности обработки не уменьшается, а во многих случаях растет при применении точного приближения вместо предельного. На основе анализа относительной эффективностью критериев и методов исследования их асимптотического поведение при различных ограничениях, для сравнение статистических критериев, использующих одинаковую тестовую статистику но разные её эталонные распределения вводится понятие относительной эффективности распределения, показывающее во сколько раз увеличится количество ложно отобранных текстов при применении в качестве эталонного распределения критерия того или иного распределения. Показана функциональная связь между понятиями эффективность обработки и относительная эффективность распределений. В условиях доступности высокопроизводительных вычислительных средств, позволяющих проводить расчеты Δ–точных распреде-

лений для интересующих параметров длины и мощности алфавита текстов, доказано утверждение об относительной эффективности распределений, позволяющее из множества распределений тестовой статистики выбрать эталонное распределение критерия при котором эффективность обработки будет наибольшая. Приведены примеры значений относительной эффективности точных и предельных приближений.

Вероятность; тестовая статистика; критерий; эталонное распределение; точное распределение; предельное распределение; эффективность обработки; относительная эффективность распределения; вычислительная сложность метода; производительность многопроцессорной вычислительной системы.

A.K. Melnikov

APPLICATION OF EXACT AND LIMIT APPROXIMATIONS OF STATISTICS PROBABILITY DISTRIBUTIONS FOR THE PROBLEM OF TEXT PROCESSING

In the paper we consider application of limit and exact approximations of statistics probability distributions for the problem of selection of texts with specific statistical properties. For selection of texts with equiprobable distribution of their symbols we use the statistical fitting criterion. Here, as a standard distribution of the test statistic we use its various approximations. As extreme approximations we use limit distributions, and as exact approximations we use Δ*–exact distributions. The difference between* Δ*–exact distributions and exact distributions does not exceed the specified* Δ*. We present the calculation results of* Δ*–exact distributions, show their variations from the values of limit distributions for different statistics. We consider the notion of processing efficiency for selection of equiprobable texts, which shows the part of wrong selected texts. We compare the processing efficiency for exact and limit approximations of standard distributions of test statistics. We have proved that the processing efficiency does not decreasing, but in many cases it is increasing, if the exact approximation is used instead of the extreme one. To compare the statistical criteria which are based on the same test statistic and different standard distributions, we introduce a concept of the distribution relative efficiency which shows the fold increase of the number of wrong selected texts for the criterion of one or another distribution used as a standard distribution. We show the functional connection between the concepts "processing efficiency" and "relative efficiency" of distributions. Owing to availability of high-performance computing facilities, which can be used for calculation of* Δ*–exact distributions for such parameters as the length and capacity of the text alphabet, we have proved the statement about relative efficiency of distributions. Owing to the statement it is possible to select a standard distribution of the criterion (with the highest processing efficiency) from the set of distributions of the test statistic. In addition we give the examples of the values of relative efficiency for exact and extreme approximations.*

Probability; test statistics, criterion; standard distribution; exact distribution; limit distribution; processing efficiency; relative efficiency of distribution; computational complexity of method; performance of multiprocessor computer system.

Введение. В информационных задачах обработки текстов [1], нередко для отбора из исходного текстового массива текстов подлежащих дальнейшей обработке применяются статистические критерии согласия с некоторым вероятностным распределением, определяющим свойства отбираемых текстов. В качестве критерия отбора текстов, знаки в которых распределены случайным равновероятным образом, применяются критерии согласия с равновероятным распределением.

При построении критерия согласия с равновероятным распределением в качестве базового (эталонного) распределения критерия применяются как точные распределения вероятностей значений используемой статистики критерия или тестовой статистики [2] – точные распределения так и их предельные приближения. Применение точных распределений ограничивается возможностью их расчета для параметров анализируемых текстов. Когда анализируются тексты, для параметров которых точные распределения рассчитаны быть не могут, применяются предельные приближения этих распределений.

Эффективность применения критериев отбора во многом зависит от применяемого эталонного распределения. Использование в качестве базового распределения критерия распределения, сколь угодно близкого к точному распределению, увеличивает эффективность критерия отбора по сравнению с применением в нём предельного приближения этого распределения.

Постановка задачи. Пусть из массива, состоящего из M текстов $T_{n,N}(j)$ длины n каждый, содержащих знаки алфавита $A_N = \{a_1, \ldots, a_N\}$ мощности N,

$$T_{n,N}(j) = \{ t_1(j), \ldots, t_n(j) \}, \; j=\overline{1, M}$$

необходимо отобрать подмассив текстов, являющихся реализациями случайных выборок длины n из равновероятного распределения на алфавите мощности N. Далее отобранные тексты необходимо подвергнуть углубленной обработке с получением положительного или отрицательного результата. При этом в подмассив нам необходимо отобрать не более $\overline{M} << M$ текстов, так как из-за ограничений, накладываемых на производительность имеющихся для проведения дальнейшей углубленной обработки вычислительных средств большего, чем $\overline{M}$ количества текстов мы обработать не сможем.

На этапе углубленной обработки результат обработки каждого из $\overline{M}$ текстов может быть положительным, а может быть и отрицательным. Будем предполагать, что положительный результат углубленной обработки будет получен в том и только в том случае, когда отобранный текст содержит равновероятное распределение входящих в него знаков, т.е. отобран в подмассив правильно. Определим число положительных результатов обработки $\overline{M}$ текстов отобранного подмассива через R^+. Из определения R^+ видно, что $R^+ \le \overline{M}$.

Под эффективностью обработки, аналогично [4], будем понимать величину ω, равную отношению числа положительных результатов углубленной обработки отобранных текстов R^+ к общему числу отобранных текстов $\overline{M}$

$$\omega = \frac{R^+}{\overline{M}}.$$

Из определения эффективности видно что, $0 \le \omega \le 1$, а максимум $\omega = 1$ достигается при $R^+ = \overline{M}$.

Простейшим способом решения поставленной задачи является сортировка всех M текстов $\{T_{n,N}(j) \mid j=\overline{1, M}\}$ по признаку равновероятности и выбор из отсортированного массива для обработки первых $\overline{M}$ текстов и их последующая углубленная обработка.

Рассмотрим ситуацию когда M текстов даны нам не все сразу, а поступают последовательно, за определенный период времени. Необходимо, обрабатывая последовательно каждый поступающий текст, принимать решение о его принадлежности к подмассиву для углубленной обработки, или отвергать.

Данный подход к построению процедуры последовательной обработки поступающих текстов был впервые сформулирован академиком РАН Боровковым А.А. еще в 60-х годах 20 века на конференции по методам прикладной статистики и частично представлен им в [4, 5].

Отбор текстов с равновероятным распределением знаков производится с помощью применения к каждому из M текстов критерия согласия с равновероятным распределением [6], использующим некоторую тестовую статистику, далее просто статистику S_n текста длины n, являющуюся функцией от h_i частот встречаемости

знаков (исходов) текста a_i из алфавита A_N мощности $N - S_n = f(n, N)$ и базовое (эталонное) распределение вероятностей значений используемой статистики (распределение) $- P\{ S_n \geq x \}$.

Также сделаем предположение, что на первоначальном этапе отбора при применении критерия отбора к каждому из M текстов мы сможем отобрать в подмассив $\overline{M}$ текстов. Отобранные $\overline{M}$ текстов будут содержать, как $\overline{M'}$ текстов отобранных правильно, с равновероятным распределением входящих в них знаков, так и $\overline{M''}$ ошибочно отобранных текстов.

$$\overline{M} = \overline{M'} + \overline{M''}.$$

Заметим, что по принятым предположениям, дальнейшая углубленная обработка текстов из числа ложно отобранных, не даст положительного результата, а значит $R^+ = \overline{M'}$. Подробно двухэтапная процедура отбора и обработки текстов описана в [7].

При принятых предположениях о возможности отбора общего количества текстов $\overline{M}$ и количества правильно $\overline{M'}$ и ошибочно $\overline{M''}$ отобранных текстов эффективность обработки при применении в критерии согласия статистики S_n и её распределения $P\{ S_n \geq x \} - \omega$ принимает вид $\omega(P\{ S_n \geq c \})$:

$$\omega\left(P\{S_n \geq x\}\right) = \frac{\overline{M'}}{\overline{M}} = \frac{\overline{M'}}{\overline{M'} + \overline{M''}}.$$

Величину $\overline{M''}$ определяет размер применяемого критерия – α [2]. Размер критерия α связан с разделяющей константой критерия c через вероятность распределения значений статистики $S_n - P\{ S_n \geq x \}$ соотношением [8]

$$P\{ S_n \geq c \} = \alpha.$$

Число текстов $\overline{M''}$, ошибочно отобранных как тексты с равновероятным распределением знаков оценивается как

$$\overline{M''} \cong \overline{M} \times \alpha,$$

а тогда применяя тождественные преобразования имеем, что эффективность обработки $\omega\left(P\{S_n \geq c\}\right)$ при применения распределения статистики $S_n - P\{ S_n \geq c \}$ с использованием разделяющей константы c принимает вид и равна

$$\omega(\alpha, P\{S_n \geq x\}) = \frac{\overline{M'}}{\overline{M}} = \frac{\overline{M} - \overline{M} \times \alpha}{\overline{M}} = \frac{\overline{M} \times (1-\alpha)}{\overline{M}} = 1 - \alpha. \qquad (1)$$

Подробное построение критерия согласия для принятия решения о равновероятном распределении знаков текста проведено в работе [9]. В зависимости от параметров выборки (n,N) применяются либо точные распределения, либо их предельные приближения [9]. Область параметров применения точных распределений определяется возможностями по производительности вычислительных средств, применяемых для их расчетов [10]. Область применения предельных приближений распределений определяется из результатов, полученных Фишером в [11], Крамером в [6] и Кендаллом в [12]. В работе [13] показано, что существует область параметров (n,N), для которой не могут быть рассчитаны точные распределения, а предельные приближения распределения применяться не могут – так называемая об-

ласть неопределенности. В условиях невозможности расчета точных распределений для параметров из области неопределенности до настоящего момента использовались предельные приближения распределений. Предложенный в [9] обобщенный статистический метод анализа текстов, дает возможность использовать для параметров из области неопределенности распределения сколь угодно близко приближенные к их точным значениям и позволяет строить критерии с наименьшим уровнем значимости α, что дает при их использовании в процедуре обработки текстов наибольшую эффективность, позволяющую экономить дорогостоящий вычислительный ресурс.

Целью данной работы является сравнение результатов расчета распределений вероятностей значений статистик близких к их точным распределениям с их предельными приближениями и сравнение эффективности применяемых для отбора текстов критериев согласия, использующих в качестве базовых распределений распределения близкие к точным распределениям и критериев, использующих в качестве базовых распределений их предельные приближения.

Пример расчета распределений вероятностей значений статистик близких к их точным распределениям. В [10], на основании предположения о производительности доступного для проведения вычислений точных распределений вычислительного ресурса в $П_{вс}=10^{16}$ операций в секунду, и времени расчета T=30 дней или 2 592 000 секунд была рассчитана область параметров (n,N) возможного расчета и, соответственно, возможного применения точных распределений. В той же работе, опираясь на утверждение Р.А. Фишера [11] о возможности применения предельных приближений распределений при $k \geq 5$, где $k = n/N$, была построена область применения предельных приближений распределений (предельных распределений).

Диаграмма области параметров точных и предельных распределений приводится на рис. 1.

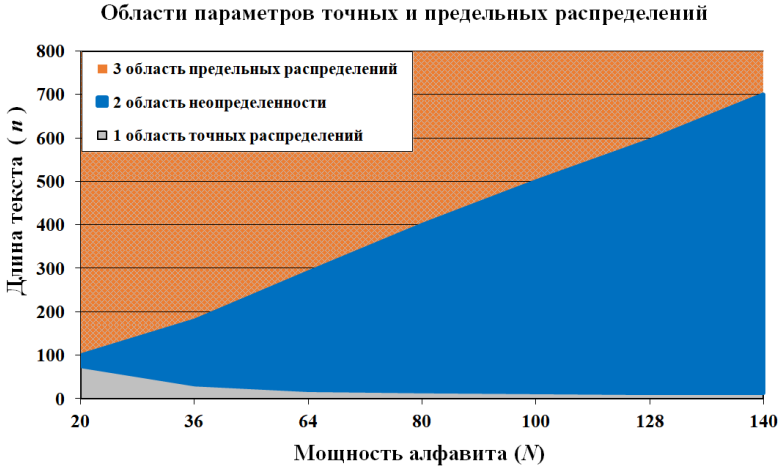

Рис. 1. Диаграмма области точных и предельных распределений, разграниченные область неопределенности.

В работе [9], ввиду невозможности расчета точных распределений значений статистик для выборок с параметрами (n,N) из области неопределенности, было предложено пользоваться распределениями, отличающимися от точных не более чем на заранее заданную величину Δ или Δ–точными распределениями. Для примера расчета Δ–точных распределений была выбрана точка из области неопреде-

лённости (50,26). В соответствии с методикой расчета Δ–точных распределений, частично определенной в [9, 14] и полностью сформулированной в [15] с точностью $\Delta=10^{-5}$ были рассчитаны Δ–точные распределения статистик, предельным приближением распределений которых является хи-квадрат распределение $\chi^2_{(N-1)}$ с (N-1) степенью свободы. Были рассчитаны Δ–точные распределения вероятностей значений статистики S_n хи-квадрат – χ_n

$$\chi_n = \sum_{i=1}^{N} \frac{(h_i - np_i)^2}{np_i},$$

предложенной в [16] и исследуемой в [17, 18], статистики S_n максимального правдоподобия – λ_n [2]

$$\lambda_n = 2 \times \sum_{i=1}^{N} h_i \times \ln \frac{h_i}{np_i}$$

и статистика Матуситы – m_n [19]

$$m_n = 4n \times \sum_{i=1}^{N} \left(\sqrt{\frac{h_i}{n}} - \sqrt{p_i} \right)^2,$$

где h_i – частота встречаемости знака (исхода) a_i, n – длина текста (объем выборки), N – число исходов полиномиальной схемы (мощность алфавита A_N) и p_i – вероятность a_i-го исхода.

В соответствии с выбранной точностью $\Delta=10^{-5}$ и методикой расчета [15], по рекуррентным формулам (8)-(10) была рассчитана вероятность статистики максимальной частоты M_n

$$M_n = \max_{i=1}^{N} h_i \cdot$$

Полученное значение вероятности статистики максимальной частоты

$$P\{M_{50} < 12\} = 0{,}9999992$$

определило границы расчета распределений рассматриваемых статистик χ_{50}, λ_{50} и m_{50}. Распределения вероятностей значений рассматриваемых статистик считались в условиях, соответствующих выбранной точности $\Delta=10^{5}$ по следующим формулам:

$$\chi_{50} = \frac{26}{50} \times \sum_{\nu=0}^{12} \mu_\nu \times \left(\nu - \frac{50}{26} \right)^2,$$

$$\lambda_{50} = 2 \times \sum_{\nu=0}^{12} \mu_\nu \times \nu \times \ln\left(\frac{26}{50} i \right),$$

$$m_{50} = 200 \times \sum_{\nu=0}^{12} \mu_\nu \times \left(\sqrt{\frac{\nu}{50}} - \sqrt{\frac{1}{26}} \right)^2,$$

где μ_v определяется как число целочисленных положительных решений уравнения $h_1 + \ldots + h_N = n$, для которых $h_i = v$, что значительно снизило трудоемкость вычислений.

Сравнение рапределений близких к точным и их предельных приближений. Рассчитанные Δ–точные распределения значений статистик χ_{50}, λ_{50} и m_{50}., отличающиеся от их точных распределений не более чем на $\Delta=10^{-5}$ и для сравнения распределение случайной величины $\chi^2_{(25)}$, имеющей хи–распределение с 25 степенями свободы из [6] приведены в табл. 1 и на рис. 2.

Таблица 1

Распределение вероятностей значений статистики хи-квадрат, максимального правдоподобия, Матуситы при n=50, N=26

C	Распределение статистик			$P\{\chi^2_{(25)} \geq c\}$
	$P\{\chi_{50} \geq c\}$	$P\{\lambda_{50} \geq c\}$	$P\{m_{50} \geq c\}$	
5	0,999998	0,999998	0,999998	0,99999
10	0,998045	0,998774	0,998872	0,99665
15	0,948410	0,978363	0,997236	0,94138
20	0,741043	0,884125	0,968972	0,74683
25	0,433904	0,678525	0,922736	0,46237
30	0,233301	0,409904	0,824062	0,22429
35	0,089475	0,195760	0,705474	0,08820
40	0,029834	0,074559	0,544615	0,02916
45	0,009142	0,022358	0,379837	0,00864
50	0,002679	0,005458	0,253562	0,00213
55	0,000910	0,001123	0,139760	0,00051
60	0,000286	0,000196	0,078224	0,00011
65	0,000083	0,000029	0,036184	0,00002
70	0,000025	0,000004	0,016434	0,00000

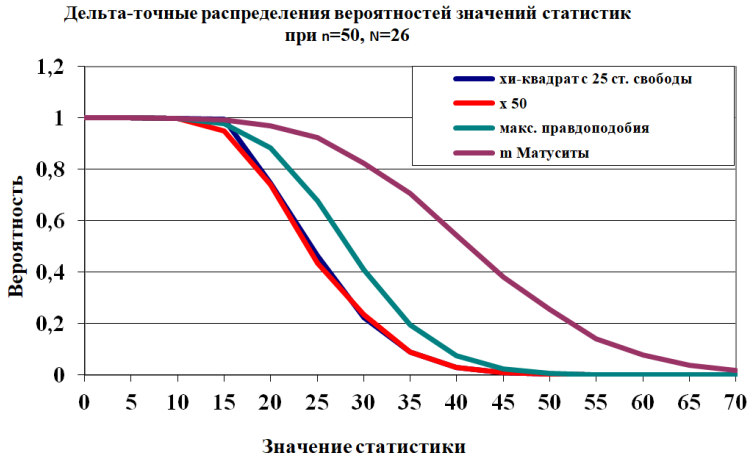

Рис. 2. Графики Δ–точных распределений значений статистик хи-квадрат, максимального правдоподобия, Матуситы и хи-квадрат распределения с 25 степенями свободы

Сравнивая значения вероятностей статистики Матуситы $P\{m_{50} \geq c\}$ и значения вероятностей хи-квадрат распределения с 25 степенями свободы $P\{\chi^2_{(25)} \geq c\}$, приведенные в табл. 1 и на рис. 1 можно сделать вывод, что для оценки вероятностей распределения Матуситы $P\{m_{50} \geq c\}$ нельзя пользоваться её приближением через $\chi^2_{(25)}$.

В табл. 2 и на рис. 3 приведены значения вероятностей статистики хи-квадрат $P\{\chi_{50} \geq c\}$ и хи-квадрат распределение с 25 степенями свободы $P\{\chi^2_{(25)} \geq c\}$ в центральной области значений при $30 \leq c \leq 40$.

Таблица 2

Распределение статистики хи-квадрат в центральной области при *n*=50, *N*=26

C	$P\{\chi_{50} \geq c\}$	$P\{\chi^2_{(25)} \geq c\}$
30	0,233301	0,22429
32	0,162366	0,15801
34	0,109813	0,10791
35	0,089475	0,08820
36	0,072489	0,07160
37	0,058441	0,05774
38	0,046891	0,04626
39	0,037475	0,03684
40	0,029834	0,02916
41	0,023676	0,023565

Рис. 3. Графики Δ–точного распределений значений статистик хи-квадрат и хи-квадрат распределения с 25 степенями свободы в центральной области

Сравнивая значения $P\{\chi_{50} \geq c\}$ и $P\{\chi^2_{(25)} \geq c\}$, приведенные в табл. 1 и 2 и на рис. 2 и 3 видим, что в центральной области при $30 \leq c \leq 40$ хорошим приближением для $P\{\chi_{50} \geq c\}$ является $P\{\chi^2_{(25)} \geq c\}$, т.е. предельное распределение может использоваться при построении статистических процедур.

В табл. 3 и на рис. 4 приведены значения статистики хи-квадрат $P\{\chi_{50} \geq c\}$ и хи-квадрат распределение с 25 степенями свободы $P\{\chi^2_{(25)} \geq c\}$ в области больших уклонений при c≥55.

Таблица 3

Распределение статистики хи-квадрат в области больших уклонений при *n*=50, *N*=26.

C	$P\{\chi_{50} \geq c\}$	$P\{\chi^2_{(25)} \geq c\}$	$P\{\chi_{50} \geq c\} / P\{\chi^2_{(25)} \geq c\}$
58	0,000470	0,00020	2,35
60	0,000286	0,000110	2,60
62	0,000175	0,000060	2,91
64	0,000107	0,000030	3,56
66	0,000073	0,000020	3,67
68	0,000040	0,000010	4,0
70	0,000025	0,000000	-
72	0,000015	0,000000	-
74	0,000009	0,000000	-

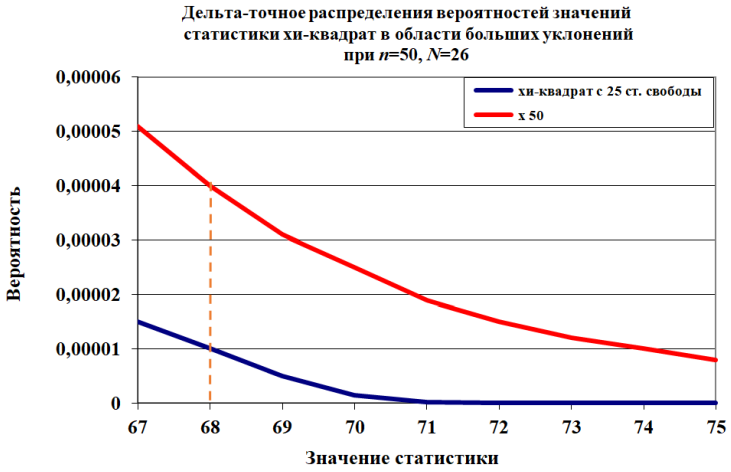

Рис. 4. Графики Δ–точного распределений значений статистик хи-квадрат и хи-квадрат распределения с 25 степенями свободы в области больших уклонений

Сравнивая значения $P\{\chi_{50} \geq c\}$ и $P\{\chi^2_{(25)} \geq c\}$, приведенные в табл. 1 и 3 и на рис. 1 и 3 в

области больших уклонений при c≥55 видим, что приближением $P\{\chi^2_{(25)} \geq c\}$ для $P\{\chi_{50} \geq c\}$ пользоваться нельзя, так как $P\{\chi_{50} \geq c\} / P\{\chi^2_{(25)} \geq c\} \approx 4$ (выделено в табл. 3).

В табл. 4 и на рис. 5 приведены значения вероятности статистики максимального правдоподобия $P\{\lambda_{50} \geq c\}$ и хи-квадрат распределение с 25 степенями свободы $P\{\chi^2_{(25)} \geq c\}$.

Таблица 4

Распределение статистики максимального правдоподобия в центральной области при *n*=50, *N*=26

C	$P\{\lambda_{50} \geq c\}$	$P\{\chi^2_{(25)} \geq c\}$	$P\{\lambda_{50} \geq c\} / P\{\chi^2_{(25)} \geq c\}$
11	0,997308	0,99295	1,004
12	0,994747	0,98657	1,008
14	0,987166	0,96173	1,026
15	0,978363	0,94138	1,039
16	0,966557	0,91483	1,056
17	0,951760	0,88179	1,079
19	0,912228	0,79712	1,144
25	0,678525	0,46237	1,467
37	0,136333	0,05774	2,361
41	0,058790	0,02356	2,495
42	0,046796	0,01797	2,603
45	0,022358	0,00864	2,586
48	0,009644	0,00373	2,584
65	0,000029	0,000025	-
66	0,000020	0,000020	-
67	0,000013	0,000015	-

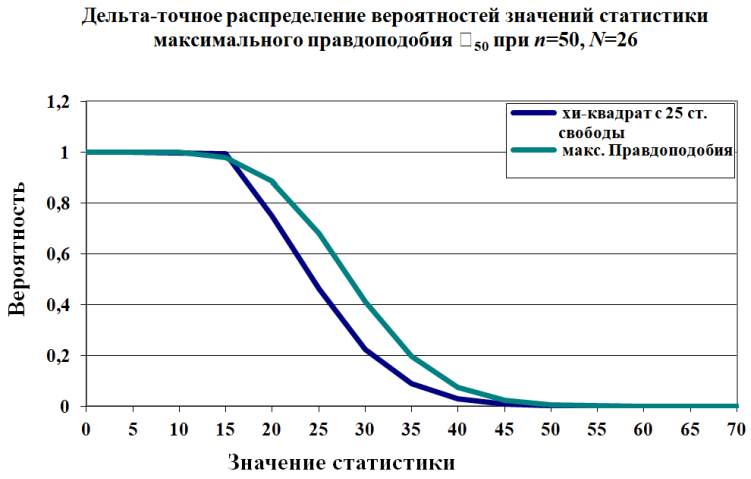

Рис. 5. Графики Δ–точного распределений значений статистик максимального λ_{50} правдоподобия и хи-квадрат распределения с 25 степенями свободы

Сравнивая значения $P\{\lambda_{50} \geq c\}$ и $P\{\chi^2_{(25)} \geq c\}$, приведенные в табл. 4 и на рис. 5 видим:

♦ в области малых отклонений при $c \leq 15$ хорошим приближением для $P\{\lambda_{50} \geq c\}$ является $P\{\chi^2_{(25)} \geq c\}$;

♦ в центральной области при $20 \leq c \leq 50$ нельзя пользоваться приближением $P\{\lambda_{50} \geq c\}$ через $P\{\chi^2_{(25)} \geq c\}$. В этой области $P\{\lambda_{50} \geq c\} / P\{\chi^2_{(25)} \geq c\} \cong 2{,}5$;

♦ в области больших уклонений при $c \geq 66$ хорошим приближением для $P\{\lambda_{50} \geq c\}$ является $P\{\chi^2_{(25)} \geq c\}$.

В табл. 5 и на рис. 6 приведены значения вероятности статистики Матуситы $P\{m_{50} \geq c\}$ и хи-квадрат распределение с 25 степенями свободы $P\{\chi^2_{(25)} \geq c\}$.

Таблица 5

Распределение вероятностей значений статистики хи-квадрат, максимального правдоподобия, Матуситы при n=50, N=26

C	$P\{m_{50} \geq c\}$	$P\{\chi^2_{(25)} \geq c\}$	$P\{m_{50} \geq c\} / P\{\chi^2_{(25)} \geq c\}$
5	0,999999	0,99999	1,00
10	0,998872	0,99665	1,00
15	0,997236	0,94138	1,00
20	0,968972	0,74683	1,29
25	0,922736	0,46237	1,99
30	0,824062	0,22429	3,67
35	0,705474	0,08820	7,99
40	0,544615	0,02916	18,68
45	0,379837	0,00864	43,99
50	0,253562	0,00213	119,72
55	0,139760	0,00051	274,03
60	0,078224	0,00011	711,13
65	0,036184	0,00002	1447,36

Рис. 6. Графики Δ–точного распределений значений статистик Матуситы и хи-квадрат распределения с 25 степенями свободы

Сравнивая значения $P\{m_{50} \geq c\}$ и $P\{\chi^2_{(25)} \geq c\}$, приведенные в табл. 5 можно сделать вывод, что:

- в области малых отклонений при $c \leq 15$ хорошим приближением для $P\{m_{50} \geq c\}$ является $P\{\chi^2_{(25)} \geq c\}$;
- в центральной области при $20 \leq c \leq 50$ и области больших уклонений $55 \leq c$ нельзя пользоваться приближением $P\{m_{50} \geq c\}$ через $P\{\chi^2_{(25)} \geq c\}$.

Сравнение эффективности обработки при применении в статистических критериях точных и предельных приближений распределения вероятностей значений тестовых статистик. Пусть исходя из общего числа отобранных текстов $\overline{M}$ и ограничения на величину ложно отобранных текстов выбран размер критерия α. Тогда используя точное приближение распределения тестовой статистики $S_n - P_\Delta\{S_n \geq x\}$ из условия

$$P_\Delta\{S_n \geq c\} = \alpha$$

получаем разделяющую константу c критерия, основанного на точном приближении распределения статистики S_n

$$c = P_\Delta^{-1}\{\alpha\}.$$

Если в качеств базового распределения для решения данной задачи мы использовали предельное приближение распределения статистики $S_n - P_{\lim}\{S_n \geq x\}$, то значение разделяющей константы c_1 критерия, основанного на предельном приближении распределения статистики S_n

$$c_1 = P_{\lim}^{-1}\{\alpha\}. \qquad (2)$$

Из свойств не отрицательности, не возрастания и монотонности функций распределения $P_\Delta\{S_n \geq x\}$, $P_{\lim}\{S_n \geq x\}$ и согласно данным приведенным в таблицах 1, 2, 3. 4

$$P_\Delta\{S_n \geq x\} \geq P_{\lim}\{S_n \geq x\} \geq 0, \text{ для } \forall x \geq 0$$

и из

$$P_\Delta\{S_n \geq c\} = P_{\lim}\{S_n \geq c_1\} = \alpha$$

следует, что $c_1 \leq c$.

Применение решающего правила с разделяющей константой критерия c_1 к выборке имеющей точное распределение $P_T\{S_n \geq x\}$ приведет к изменению размера критерия до $\alpha_2 = P_T\{S_n \geq c_1\}$. Точное распределение $P_T\{S_n \geq x\}$ нам не известно из-за большой вычислительной сложности его расчета [10], но мы можем вычислять Δ–точные распределения $P_\Delta\{S_n \geq x\}$, такие что

$$\left| P_\Delta\{S_n \geq x\} - P_T\{S_n \geq x\} \right| \leq \Delta, \text{ для } \forall x \geq 0.$$

Но применение разделяющей константы критерия c_1 при использовании в качестве базового распределения критерия точного приближения $P_\Delta\{S_n \geq x\}$ определяет размер критерия из условия

$$P_\Delta\{S_n \geq c_1\} = \alpha_1,$$

которое используя (2) может быть представлено в виде

$$P_{\Delta}\{S_n \geq P_{\lim}^{-1}\{\alpha\}\} = \alpha_1. \quad (3)$$

Повторно используя свойства не отрицательности, не возрастания и монотонности функций распределения $P_{\Delta}\{S_n \geq x\}$, $P_{\lim}\{S_n \geq x\}$ и согласно данным приведенным в таблицах 1, 2, 3. 4

$$P_{\Delta}\{S_n \geq x\} \geq P_{\lim}\{S_n \geq x\} \geq 0, \text{ для } \forall x \geq 0$$

и тогда из того что $c_1 \leq c$ следует, что

$$\alpha_1 \geq \alpha. \quad (4)$$

Из свойств точного приближения, в качестве которого используется Δ–точное распределение следует что,

$$\left| P_{\Delta}\{S_n \geq x\} - P_T\{S_n \geq x\} \right| \leq \Delta, \text{ для } \forall x \geq 0$$

и следовательно $\left| \alpha_1 - \alpha_2 \right| \leq \Delta$.

Приведенные рассуждения иллюстрируется на рис. 7.

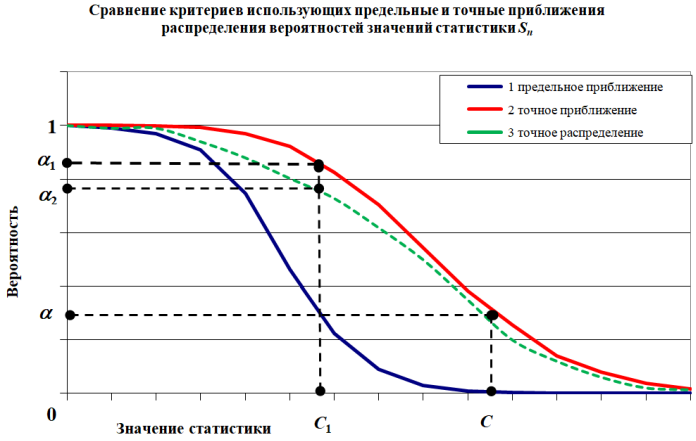

Рис. 7. Иллюстрация результатов применения в критериях обработки текстов точных и предельных приближений распределений используемой статистики, где 1 – предельное приближение распределения, 2 – точное приближение распределения, 3 – точное распределение

Применение выбранной при использовании предельного приближения распределения разделяющая константа критерия c_1 приведет к увеличению размера критерия с α до α_2, и к увеличению числа ложно отобранных текстов в α_2 / α раз, что на практике при ограничениях на $\Delta \approx 10^{-5}$ и выполнении условия

$$\left| P_{\Delta}\{S_n \geq x\} - P_{\lim}\{S_n \geq x\} \right| >> \Delta, \text{ для } \forall x \geq 0$$

может оцениваться как

$$\alpha_2 / \alpha \geq \alpha_1 - \Delta / \alpha \cong \alpha_1 / \alpha.$$

Для статистики хи-квадрат увеличение числа ложно отобранных будет до 4 раз (см. табл. 3 и рис. 4), для статистики максимального правдоподобия до 2,5 раз (см. табл. 4 и рис 5). Использование предельного приближения распределения для статистики Матуситы приводит к недостоверному результату, так как число ложно отобранных текстов увеличивается многократно (см. табл. 5), что согласуется с

выводом раздела данной статьи о сравнении распределений, утверждающим, что для оценки вероятностей распределения Матуситы $P\{m_{50} \geq c\}$ нельзя пользоваться её приближением предельным приближением через $\chi^2_{(25)}$.

Возвращаемся к вопросу эффективность обработки $\omega(\alpha, P\{S_n \geq x\})$ при применении в критерии согласия статистики S_n и её распределения $P\{ S_n \geq x \}$. Задавая размер критерия α, используя соотношение

$$c = P_{\Delta}^{-1}\{\alpha\}$$

получаем разделяющую константу c, тогда согласно (1) эффективность обработки при использовании точного приближения равна

$$\omega(\alpha, P_{\Delta}\{S_n \geq x\}) = 1 - P_{\Delta}\{S_n \geq c\} = 1 - \alpha .$$

Использование предельного приближения $P_{\lim}\{S_n \geq x\}$ определяет разделяющую константу как

$$c_1 = P_{\lim}^{-1}\{\alpha\}.$$

Используя предыдущие рассуждения и (3) можем сказать, что

$$\omega(\alpha, P_{\lim}\{S_n \geq x\}) = 1 - (P_{\Delta}\{S_n \geq P_{\lim}^{-1}\{\alpha\}\}) = 1 - P_{\Delta}\{S_n \geq c_1\} = 1 - \alpha_1 .$$

Для сравнения эффективности обработки при применении разных приближений распределений используемой в критериях одинаковой статистики S_n исследуем их разность

$$\omega(\alpha, P_{\Delta}\{S_n \geq x\}) - \omega(\alpha, P_{\lim}\{S_n \geq x\}) = 1 - \alpha - 1 + \alpha_1 = \alpha_1 - \alpha \geq 0 . \qquad (4)$$

Выражение (5) согласно (4) всегда неотрицательно и следовательно эффективность обработки текстов при применении точных приближений распределений используемых статистик будет не хуже, чем при применении предельных приближений, а как показывает практика и в несколько раз лучше.

Уделив внимание эффективности обработки текстов, построенной на применении критериев согласия и разных приближений распределений вероятностей значений используемых в них статистик кратко обратимся к истории теории сравнения статистических критериев.

В шестидесятые восьмидесятые годы 20 столетия, когда производительность вычислительных средств не позволяла вычислять точные распределения $P_T\{ S_n \geq c \}$ для значений параметров текстов (n, N) $n \geq 10$, $N \geq 15$, в качестве меры сравнения критериев, использующих различные статистики $S_n = f_n = f(n, N)$ и $S_n = g_n = g(n, N)$, предлагалось использовать отношение

$$e\{P(f_n, g_n)\} = P\{f_n \geq c\} / P\{g_n \geq c\}, \qquad (5)$$

которое называют относительной эффективностью критериев, асимптотическое поведение которого исследовалось при разных условиях.

В случае когда $P\{f(n, N) \geq c\}$ и $P\{g(n, N) \geq c\}$ «сближаются» говорили об эффективности в смысле Питмена

$$P\{f_n \geq c\} \cong P\{g_n \geq c\}.$$

При стремлении ошибки 1-го или 2-го рода каждого из критериев к нулю, говорили об эффективности в смысле Бахадура или Ходжеса-Лемана. Вычисление предела выражения (5) при стремлении ошибок 1-го и 2-го рода к нулю связывают с именем Чернова. Подробная классификация поведения асимптотической относительной эффективности критериев приведена в работах А.Ф. Ронжина [3, 20].

В аспекте дальнейшего изучения поведения критериев согласия предлагается рассмотреть относительную оценку критериев использующих одну и ту же тестовую статистику S_n, но в качестве эталонных распределений критерия разные её распределения $P_1\{S_n \geq x\}$ и $P_2\{S_n \geq x\}$ – $e_{1,2}(\alpha)$, показывающую во сколько раз увеличится количество ложно отобранных текстов при применении в качестве эталонного распределения критерия распределения $P_2\{S_n \geq x\}$ по сравнению с использованием в качестве эталонного распределения $P_1\{S_n \geq x\}$. Эту оценку $e_{1,2}(\alpha)$ предлагаем назвать относительной эффективностью применения распределения вероятностей значений используемой в критерии статистики или сокращенно относительной эффективностью распределения (ОЭР). Оценка ОЭР $e_{1,2}(\alpha)$ выражается следующим соотношением

$$e_{1,2}(\alpha) = P_1\{S_n \geq P_2^{-1}\{P_1\{S_n \geq c\}\} / P_1\{S_n \geq c\}, \quad (6)$$

где $P_1\{S_n \geq c\} = \alpha$, а под $P_2^{-1}\{P_1\{S_n \geq c\}\}$ понимается значение разделяющей константы c_1 распределения $P_2\{S_n \geq x\}$, при которой размер критерия α, при применении распределения $P_2\{S_n \geq x\}$ совпадает с размером критерия α при применении распределения $P_1\{S_n \geq x\}$, но с другой разделяющей константой c:

$$P_2^{-1}\{P_1\{S_n \geq c\}\} = \{ c_1 \mid P_2\{S_n \geq c_1\} = P_1\{S_n \geq c\} \}. \quad (7)$$

В нашем случае, используя разработанный в [9, 15] метод расчета распределений сколь угодно близких к их точным распределениям - Δ–точных распределений, можно применить понятия ОЭПР для сравнения критериев, использующих одинаковые статистики, но в качестве базовых распределений разные приближения этих распределений, в частности предельные приближения и точные приближения – Δ–точные распределения.

Теперь, используя (6) и (7), в качестве меры сравнении критериев, использующих в качестве эталонных распределений точные $P_1\{S_n \geq x\} = P_\Delta\{S_n \geq x\}$ или предельные $P_2\{S_n \geq x\} = P_{\lim}\{S_n \geq x\}$ приближения этих распределений, можно предложить следующее соотношение

$$e_{\Delta,\lim}(\alpha\}) = P_\Delta\{S_n \geq P_{\lim}^{-1}\{P_\Delta\{S_n \geq c\}\} / P_\Delta\{S_n \geq c\}, \quad (8)$$

где под $P_\Delta\{S_n \geq c\}$ понимаем Δ–точное распределение статистики S_n – точное приближение, под $P_{\lim}\{S_n \geq c\}$ её предельное приближение, а под $P_{\lim}^{-1}\{P_\Delta\{S_n \geq c\}\}$, аналогично (7), понимается значение разделяющей константы c_1 предельного приближения, при которой размер критерия α, при применении предельного приближения совпадает с размером критерия при применении точного приближения, но с другой разделяющей константой c:

$$P_{\lim}^{-1}\{P_\Delta\{S_n \geq c\}\} = \{ c_1 \mid P_{\lim}\{S_n \geq c_1\} = P_\Delta\{S_n \geq c\} \}, \quad (9)$$

равенство размеров критерия при применении различных приближений распределения статистики S_n проиллюстрировано на рис. 7.

Предложенное соотношение (8) имеет смысл рассматривать в области неопределенности [10], которая, как показано в [10] и [13], не исчезает с прогнозируемым ростом производительности вычислительных средств.

Теперь, имея понятие эффективности обработки $\omega\ (\alpha, P_\Delta\{S_n \ge x\})$ (1) и относительной эффективностью приближений распределений $e_{\Delta,\,\lim}\ (P\{S_n \ge x\})$ (8) можем определить их связь. Разность эффективности обработки при применении разных приближений распределения одной статистики согласно (4), используя (8) можно выразить как

$$\omega(\alpha, P_\Delta\{S_n \ge x\}) - \omega(\alpha, P_{\lim}\{S_n \ge x\}) = 1 - \alpha - 1 + \alpha_1 = \alpha_1 - \alpha = \alpha \times (e_{\Delta,\,\lim}(\alpha) - 1) \cdot \quad (10)$$

Выражение (10) показывает связь между понятиями эффективности обработки текстов при использовании статистического критерия и относительной эффективности применения приближений распределения вероятностей значений используемой в критерии тестовой статистики.

Используя отношение (10) и соотношения между точными и предельными распределениями, полученные из анализа результатов сравнения рассчитанных Δ–точных распределений и их предельных приближений, можно сформулировать следующее утверждение.

Утверждение 1. Для любого заданного уровня значимости статистического критерия согласия α и применяемого в нём эталонного распределения вероятностей значений используемой тестовой статистики S_n (эталонного распределения) $P_i\{S_n \ge x\}$ не превосходящего её Δ–точного распределения $P_\Delta\{S_n \ge x\})$, эффективность обработки текстов в смысле доли ложно отобранных текстов

$$\omega(\alpha, P_i\{S_n \ge x\}) = 1 - \alpha, \textit{ при } P_i\{S_n \ge c_i\} = \alpha$$

не уменьшится при применении в качестве эталонного распределения критерия согласия, использующегося в качестве критерия отбора текстов, распределения имеющего меньшую относительную эффективность относительно Δ–точного распределения $P_\Delta\{S_n \ge x\})$ – точного приближения.

$$\omega(\alpha, P_i\{S_n \ge x\}) \ge \omega(\alpha, P_j\{S_n \ge x\}) \textit{ для } \forall i,\ j, \textit{для которых } \ e_{\Delta,i}(\alpha) \le e_{\Delta,j}(\alpha) \cdot$$

Доказательство.

Для доказательства рассмотрим разность значений эффективности обработки при применении двух разных распределений $P_i\{S_n \ge x\})$ и $P_j\{S_n \ge x\})$

$$\omega(\alpha, P_i\{S_n \ge x\}) - \omega(\alpha, P_j\{S_n \ge x\}). \quad (11)$$

При отличии рассматриваемых распределений друг от друга и от Δ–точного распределения

$$P_i\{S_n \ge x\} \ne P_j\{S_n \ge x\} \ne P_\Delta\{S_n \ge x\}$$

и из условий утверждения $P_i\{S_n \ge c_i\} = \alpha$ следует, что $\exists c_i$, c_j и c_Δ, для которых

$$P_i\{S_n \ge c_i\}) = \alpha,$$

$$P_j\{S_n \ge c_j\}) = \alpha,$$

$$P_\Delta\{S_n \ge c_\Delta\}) = \alpha$$

и

$$c_\Delta \ne c_i \ne c_j.$$

Значит из определения эффективности обработки следует, что

$$\omega(\alpha, P_i\{S_n \ge x\}) = 1 - \alpha, \textit{ при } P_i\{S_n \ge c_i\} = \alpha \cdot$$

Для исследования поведения разности (11) прибавим и отнимем от неё значение эффективности обработки при применении точного приближения – Δ–точного распределения $P_\Delta\{S_n \ge x\}$, тогда

$$\omega(\alpha, P_i\{S_n \ge x\}) - \omega(\alpha, P_j\{S_n \ge x\}) = \tag{12}$$

$$= \omega(\alpha, P_\Delta\{S_n \ge x\}) - \omega(\alpha, P_j\{S_n \ge x\}) - (\,\omega(\alpha, P_\Delta\{S_n \ge x\}) - \omega(\alpha, P_i\{S_n \ge x\})\,) =$$

используя (10) продолжим равенство

$$= \alpha \times (e_{\Delta,\,j}(\alpha) - 1) - \alpha \times (e_{\Delta,\,i}(\alpha) - 1) = \alpha \times (e_{\Delta,\,j}(\alpha) - e_{\Delta,\,i}(\alpha)) \ge 0$$

по условию утверждения $e_{\Delta,i}(\alpha) \le e_{\Delta,j}(\alpha)$ и следовательно исследуемая разность (12) больше нуля, что означает что эффективность обработки не уменьшается при применении распределения имеющего меньшую относительную эффективность относительно Δ–точного распределения.

Утверждение доказано.

На основе доказанного утверждения можно сформулировать следующее следствие, позволяющее делать выбор распределения $P_i\{S_n \ge x\}$ тестовой статистики S_n из множества доступных её распределений $\{\, P_1\{S_n \ge x\}, P_2\{S_n \ge x\}, \ldots. P_Z\{S_n \ge x\}\,\}$, при применении которого в качестве эталонного распределения критерия эффективность обработки $\omega(\alpha, P_i\{S_n \ge x\})$ будет наибольшей.

Следствие.

Для заданного уровня значности α эффективность обработки $\omega(\alpha, P_i\{S_n \ge x\})$ при применении в качестве эталонного распределения тестовой статистики S_n критерия согласия распределения $P_i\{S_n \ge x\}$, выбранного из множества распределений $\{\, P_1\{S_n \ge x\}, P_2\{S_n \ge x\}, \ldots., P_Z\{S_n \ge x\}\,\}$ при выполнении условия $\{\, P_j\{S_n \ge x\} \le P_\Delta\{S_n \ge x\} \mid j = \overline{1, Z}\,\}$, будет не превосходить $\omega(\alpha, P_i\{S_n \ge x\})$: $\{\, \omega(\alpha, P_i\{S_n \ge x\}) \ge \omega(\alpha, P_j\{S_n \ge x\} \mid j = \overline{1, Z}\,\}$ тогда, когда относительная эффективность $e_{\Delta,i}(\alpha)$ выбранного распределения $P_i\{S_n \ge x\}$ относительно Δ–точного распределения $P_\Delta\{S_n \ge x\}$

$$e_{\Delta,i}\,(\alpha\}) = P_\Delta\{S_n \ge P_i^{-1}\{P_\Delta\{S_n \ge c\}\} / P_\Delta\{S_n \ge c\}$$

не будет превышать относительной эффективности всех распределений рассматриваемого множества относительно Δ–точного распределения

$$\{\, e_{\Delta,i}(\alpha) \le e_{\Delta,\,j}(\alpha) \mid j = \overline{1, Z}\,\}.$$

Следствие говорит о том, что для наибольшей эффективности обработки, в смысле минимизации числа ложно отобранных текстов, из множества распределений необходимо выбирать распределение наиболее близкое по значению к Δ–точному распределению либо само Δ–точное распределения, если оно присутствует в рассматриваемом множестве.

Доказательство.

Доказательство производится путем последовательного применения утверждения к распределениям рассматриваемого множества распределений $\{\, P_\upsilon\{S_n \ge x\} \mid \upsilon = \overline{1, Z}\,\}$ и расчета для каждого из рассматриваемых распределений

его относительной эффективности относительно Δ–точного распределения $P_\Delta\{S_n \ge x\}$ – $e_{\Delta,\upsilon}(\alpha\}$

$$e_{\Delta,\upsilon}(\alpha\}) = P_\Delta\{S_n \ge P_\upsilon^{-1}\{P_\Delta\{S_n \ge c\}\} / P_\Delta\{S_n \ge c\}.$$

Минимальное значение $e_{\Delta,i}(\alpha\}$ из рассчитанных значений $\{e_{\Delta,\ \upsilon}(\alpha) \mid \upsilon = \overline{1,Z}\}$ в соответствии с утверждением и будет указывать на искомое распределение $P_i\{S_n \ge x\}$.

Следствие доказано.

На основе утверждения и следствия о выборе эталонных распределений тестовой статистики для увеличения эффективности обработки текстов может быть получена соответствующая методика, разработка которой будет являться дальнейшим направлением исследований.

В заключении приводятся графики относительной эффективности применение точных и предельных приближений распределений статистики хи-квадрат – χ_n и статистики максимального правдоподобия – λ_n.

Рис. 8. Значение относительной эффективности применения точного и предельного приближений статистики χ_n хи-квадрат

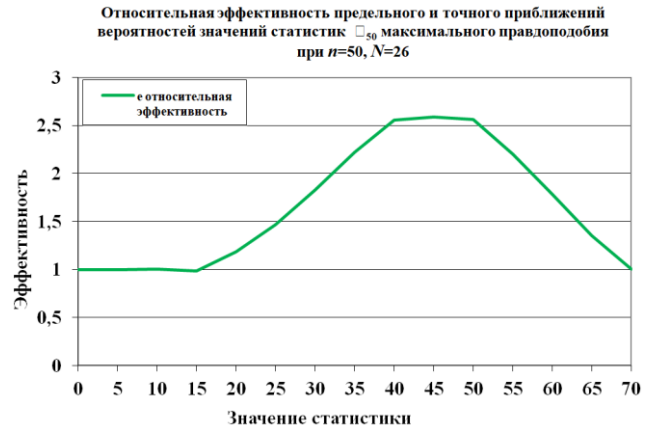

Рис. 9. Значение относительной эффективности применения точного и предельного приближений статистики λ_n максимального правдоподобия

Заключение и выводы. В работе рассматривается задача обработки текстов, сводящаяся к выбору текстов с определенными свойствами из последовательно поступающих текстов, в частности текстов с равновероятным распределением входящих в них знаков. Рассматривается решение задачи статистическим методом с помощью использования статистического критерия согласия с равновероятным распределением. Для оценки результатов обработки вводится понятие эффективности обработки, отражающее долю ложно отобранных текстов.

В условиях невозможности расчета точных распределений для интересующих параметров текста, таких как его длина и мощность алфавита, при построении статистического критерия, решающего задачу обработки текстов, в качестве эталонного распределения критерия рассматривается использование предельных и точных его приближений.

В интересах оценки результатов применения различных приближений распределений в статистических критериях и их влияния на эффективность обработки в работе проведены сравнения значений предельных приближений распределений вероятностей значений статистик, часто используемых при построении критериев статистической обработки текстов, и значений их точных приближений, рассчитанных с помощью разработанной ранее методики расчета Δ–точных распределений, отличающихся от точных распределений не более чем на заранее заданную величину. На примерах показано, что во многих случаях значения вероятностей в предельных приближениях на много отличаются от их значений в точных приближений и поэтому не могут использоваться в качестве эталонных распределений в критериях обработки текстов, так как это намного увеличивает долю ложно отобранных текстов. Доказано, что использование в критерии с равновероятным распределением его точного приближения не уменьшает по сравнению с использованием придельного приближения, а во многих случаях и увеличивает эффективность обработки в смысле уменьшения доли ложно отобранных текстов.

На основании анализа предыдущих методов сравнения критериев и для оценки использования в критерии согласия различных распределений одной и той же статистики вводится новое понятие относительной эффективности распределений, показывающее во сколько раз увеличится количество ложно отобранных текстов при применении в качестве эталонного распределения критерия одного из рассматриваемых распределений. Введенное понятие относительной эффективности распределений использовано для оценки применения в критерии отбора равновероятных текстов точных и предельных приближений распределений используемой статистики. Показана связь между эффективностью обработки текстов, построенной на применении в качестве эталонного распределения того или иного приближения этого распределения, с относительной эффективностью приближений этих распределений.

На основании сравнения значений предельных приближений распределений с результатами расчета точных приближений распределений показано, что точные приближения имеют не меньшую чем предельные приближения относительную эффективность и их использование в критерии отбора в качестве эталонного распределения не приводит к уменьшению, а во многих случаях и к увеличению, эффективности обработки. Доказано утверждение, что при заданном уровне значимости эффективность обработки текстов больше тогда, когда обработка построена на применении критерия отбора, использующего в качестве эталонного распределения распределение, имеющее меньшее отличие, а значит и меньшую относительную эффективность относительно Δ–точного распределения. На основе доказанного утверждения сформулировано следствие, позволяющее делать выбор из

множества распределений такого распределения, применение которого даст наибольшую эффективность обработки, по сравнению с применением других распределений рассматриваемого множества.

Приведенные примеры значений распределений тестовых статистик, использующихся в ходе анализа текстов, и результаты их анализа, которые подтверждают сделанные теоретические выводы.

Благодарности

Автор выражает глубокую благодарность доктору физико-математических наук, профессору Ронжину А.Ф. за постоянное внимание к работе и её обсуждение.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Чеповский А.М.* Информационные модели в задачах обработки текстов на естественных языках. – М.: Национальный открытый университет «ИНТУИТ», 2015. – 228 с. – ISBN 978-5-9556-0176-2.
2. *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику. – М.: ЛЕНАРД, 2017. – 608 с. – ISBN 978-5-9710-4535-9.
3. *Ронжин А.Ф.* Эффективность типа Чернова для критериев согласия, основанных на эмпирических функциях распределения // Теория вероятности и ее применение. – 1985. – 30:2. – С. 378-381.
4. *Боровков А.А.* Вероятностные процессы в теории массового обслуживания. – М.: Наука, 1972. – 367 с.
5. *Боровков А.А.* Математическая статистика. – Новосибирск: Изд-во ИМ СОРАН, Наука, 1997. – 772 с.
6. *Крамер Г.* Математические методы статистики. – М.: Мир, 1975. – 648 с.
7. *Мельников А.К.* Применение точных распределений в процедуре двухэтапной обработки текстов // Обозрение прикладной и промышленной математики. – 2018. – Т. 25. – Вып. 2. *В печати*. – https://tvp.ru/conferen/vsppmXIX/repso051.pdf (дата обращения 19.07.2018).
8. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. – М.: Книжный дом "ЛИБРОКОМ", 2014. – 352 с. – ISBN 978-5-397-04141-6.
9. *Мельников А.К., Ронжин А.Ф.* Обобщенный статистический метод анализа текстов, основанный на расчете распределений вероятности значений статистик // Информатика и её применения. – 2016. – Т. 10. – Вып. 4. – С. 89-95. – ISSN 1992-2264.
10. *Мельников А.К.* Сложность расчета точных распределений вероятности симметричных аддитивно разделяемых статистик и область применения предельных распределений // Доклады ТУСУР. – Томск, 2017. – Т. 20, № 4. – С. 126-130. – ISSN 1818-0442.
11. *Фишер Р.А.* Статистические методы для исследователей. – М.: Госстатиздат, 1958. – 73 с.
12. *Кендалл М.Г., Стьюарт А*. Теория распределений. – М.: Наука, 1966. – 302 с.
13. *Зелюкин Н.Б., Мельников А.К.* Сложность расчета точных распределений вероятности значений статистик и область применения предельных распределений // Электронные средства и системы управления: Материалы докладов XIII Междунар. науч.-практ. конф. (29 ноября – 1 декабря 2017 г.): в 2 ч. – Ч. 2. – Томск: В-Спектр, 2017. – С. 84-90. – https://storage.tusur.ru/files/115115/2017-2.pdf (дата обращения 13.07.2018).
14. *Мельников А.К.* Методика расчета распределений вероятностей значений статистик, близких к их точным распределениям // Обозрение прикладной и промышленной математики. – 2017. – Т. 24. – Вып. 5. – http://tvp.ru/conferen/vsppmXVIII/kisso075.pdf (дата обращения 13.07.2018).
15. *Мельников А.К.* Методика расчета распределения вероятностей значений симметричных аддитивно разделяемых статистик, приближенных к их точному распределению // Научный вестник НГТУ. – 2018. – № 1 (70). – С. 153-166. – ISBN 1814-1196. Doi: 10.17212/1814-1196-2018-1-153-166.
16. *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling // Philos. Mag. Ser. 5. – 1900. – Vol. 50, No. 302. – P. 157-170.
17. *Neyman F., Pearson E.S.* On the use and interpretation of certain test criteria for purposes of statistical inference // Biometrika. – 1928. – Vol. 20-A. − P. 175-240, 264-299.

18. *Smith P.F., Rae D.S., Manderscheid R.W., Silbergeld S.* Exact and approximate distributions of the chi-squared statistic for equiprobability // Commun. Statist. – 1979. – B. 8 (2). – No. 1. – P. 131-149.
19. *Matusita K.* Decision rules, based on the distanse, for problems of fit tu o samples, and estimation // Ann. Math. Stat. – 1955. – Vol. 26. – P. 631-640.
20. *Ронжин А.Ф.* Асимптотическая локальная относительная эффективность (АЛОЭ) критериев согласия // Тезисы докладов Всесоюзной конференции «Вероятностные методы в дискретной математике». – Петрозаводск, 1983. – С. 70-71.

REFERENCES

1. *Chepovskiy A.M.* Informatsionnye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh [Information models in tasks of processing of natural language texts]. Moscow: Natsional'nyy otkrytyy universitet «INTUIT», 2015, 228 p. ISBN 978-5-9556-0176-2.
2. *Ivchenko G.I., Medvedev Yu.I.* Vvedenie v matematicheskuyu statistiku [Introduction to mathematical statistics]. Moscow: LENARD, 2017, 608 p. ISBN 978-5-9710-4535-9.
3. *Ronzhin A.F.* Effektivnost' tipa CHernova dlya kriteriev soglasiya, osnovannykh na empiricheskikh funktsiyakh raspredeleniya [Tchernov's efficiency for fitting criteria based on empirical functions of distribution], *Teoriya veroyatnosti i ee primenenie* [Probability theory and its application], 1985, 30:2, pp. 378-381.
4. *Borovkov A.A.* Veroyatnostnye protsessy v teorii massovogo obsluzhivaniya [Stochastic processes in queueing theory]. Moscow: Nauka, 1972, 367 p.
5. *Borovkov A.A.* Matematicheskaya statistika [Mathematical statistics]. Novosibirsk: Izd-vo IM SORAN, Nauka, 1997, 772 p.
6. *Kramer G.* Matematicheskie metody statistiki [Mathematical methods of statistics]. Moscow: Mir, 1975, 648 p.
7. *Mel'nikov A.K.* Primenenie tochnykh raspredeleniy v protsedure dvukhetapnoy obrabotki tekstov [Application of exact distributions in the procedure of two-step text processing], *Obozrenie prikladnoy i promyshlennoy matematiki* [Review of applied and industrial mathematics], 2018, Vol. 25, Issue 2. In print. Available at: https://tvp.ru/conferen/vsppmXIX/repso051.pdf (accessed 19 July 2018).
8. *Ivchenko G.I., Medvedev Yu.I.* Matematicheskaya statistika [Mathematical statistics]. Moscow: Knizhnyy dom "LIBROKOM", 2014, 352 p. ISBN 978-5-397-04141-6.
9. *Mel'nikov A.K., Ronzhin A.F.* Obobshchennyy statisticheskiy metod analiza tekstov, osnovannyy na raschete raspredeleniy veroyatnosti znacheniy statistik [A generalized statistical method of analyzing texts based on the calculation of probability distributions of values of statistics], *Informatika i ee primeneniya* [Informatics and its applications], 2016, Vol. 10, Issue 4, pp. 89-95. ISSN 1992-2264.
10. *Mel'nikov A.K.* Slozhnost' rascheta tochnykh raspredeleniy veroyatnosti simmetrichnykh additivno razdelyaemykh statistik i oblast' primeneniya predel'nykh raspredeleniy [The complexity of calculating the exact probability distributions of symmetric additive-separated statistics and the application of limit distributions], *Doklady TUSUR* [Proceedings of Tomsk State University of Control Systems and Radioelectronics]. Tomsk, 2017, Vol. 20, No. 4, pp. 126-130. ISSN 1818-0442.
11. *Fisher R.A.* Statisticheskie metody dlya issledovateley [Statistical methods for researchers]. Moscow: Gosstatizdat, 1958, 73 p.
12. *Kendall M.G., St'yuart A.* Teoriya raspredeleniy [Distribution theory]. Moscow: Nauka, 1966, 302 p.
13. *Zelyukin N.B., Mel'nikov A.K.* Slozhnost' rascheta tochnykh raspredeleniy veroyatnosti znacheniy statistik i oblast' primeneniya predel'nykh raspredeleniy [Slozhnost' rascheta tochnykh raspredeleniy veroyatnosti znacheniy statistik i oblast' primeneniya predelnykh raspredeleniy], *Elektronnye sredstva i sistemy upravleniya: Materialy dokladov XIII Mezhdunar. nauch.-prakt. konf. (29 noyabrya – 1 dekabrya 2017 g.)* [Electronic facilities and control systems: reports of the XIIIth International scientific and practical], 29th November – 1st December, 2017]: In 2 part. Part 2. Tomsk: V-Spektr, 2017, pp. 84-90. Available at: https://storage.tusur.ru/files/115115/2017-2.pdf (accessed 13 July 2018).

14. *Mel'nikov A.K.* Metodika rascheta raspredeleniy veroyatnostey znacheniy statistik, blizkikh k ikh tochnym raspredeleniyam [Calculation methodology of approximate-to-exact distribution of statistics probabilities], *Obozrenie prikladnoy i promyshlennoy matematiki* [Review of applied and industrial mathematics], 2017, Vol. 24, Issue 5. Available at: http://tvp.ru/conferen/vsppmXVIII/kisso075.pdf (accessed 13 July 2018).
15. *Mel'nikov A.K.* Metodika rascheta raspredeleniya veroyatnostey znacheniy simmetrichnykh additivno razdelyaemykh statistik, priblizhennykh k ikh tochnomu raspredeleniyu [Processing complexity for exact probability distributions of symmetrical additively partitioned statistics and application area of limit distributions], *Nauchnyy vestnik NGTU* [Science bulletin of the Novosibirsk state technical university], 2018, No. 1 (70), pp. 153-166. ISBN 1814-1196. Doi: 10.17212/1814-1196-2018-1-153-166.
16. *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling, *Philos. Mag. Ser. 5*, 1900, Vol. 50, No. 302, pp. 157-170.
17. *Neyman F., Pearson E.S.* On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika*, 1928, Vol. 20-A, pp. 175-240, 264-299.
18. *Smith P.F., Rae D.S., Manderscheid R.W., Silbergeld S.* Exact and approximate distributions of the chi-squared statistic for equiprobability, *Commun. Statist*., 1979, B. 8 (2), No. 1, pp. 131-149.
19. *Matusita K.* Decision rules, based on the distanse, for problems of fit tu o samples, and estimation, *Ann. Math. Stat*., 1955, Vol. 26, pp. 631-640.
20. *Ronzhin A.F.* Asimptoticheskaya lokal'naya otnositel'naya effektivnost' (ALOE) kriteriev soglasiya [Asymptotic local relative efficiency (ALRE) of fitting criteria], *Tezisy dokladov Vsesoyuznoy konferentsii «Veroyatnostnye metody v diskretnoy matematike»* [Reports of All-USSR conference “Probabilistic methods in discrete mathematics”]. Petrozavodsk, 1983, pp. 70-71.

Статью рекомендовала к опубликованию д.т.н. А.В. Никитина.

Мельников Андрей Кимович – НТЦ ЗАО «ИнформИнвестГрупп»; e-mail: ak@iigroup.ru; 117587, Москва, Варшавское шоссе, д. 125, стр. 17; тел.: 84952870035; к. т. н.; доцент ВАК; г.н.с.

Melnikov Andrey Kimovitch – STC CLSC «InformInvestGroup»; e-mail: ak@iigroup.ru; 125, Varshavskoye road, building 17, Moscow, 117587, Russia; phone: +74952870035; cand. of eng. sc.; associate professor of SAC; chief research officer.

УДК 004.932 DOI 10.23683/2311-3103-2018-8-135-145

А.М. Абасова, Л.К. Бабенко

ЗАЩИТА АВТОРСКИХ ПРАВ НА ИЗОБРАЖЕНИЕ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МОРФОЛОГИЧЕСКОЙ ОБРАБОТКИ*

В настоящей статье рассматривается вопрос о внедрении цифровых водяных знаков в области изображения, которые наименее вероятно подвергнутся модификации и, следовательно, подходят для обеспечения эффективной защиты авторских прав, с учетом типа деструктивных воздействий, характерных при их нарушении. В качестве контейнера выступает цветное изображение, а в качестве цифрового водяного знака – текст, содержащий знак охраны авторских прав. Для внедрения выбираются блоки переднего плана, так как согласно проведенному исследованию именно они представляют ценность изображения, что особенно характерно для коммерческих фотографий. Поиск данных блоков для внедрения осуществляется с помощью маркирования с использованием методов математической морфологии. Также в статье на примере показана способность структурного элемента выполнять роль ключевой информации. Предложено использовать геометрический центр каждого найденного блока переднего плана для внедрения цифрового водяного

* Работа выполнена при поддержке гранта РФФИ № 18-07-01347.