Сибирский государственный университет
телекоммуникаций и информатики

ОБРАБОТКА ИНФОРМАЦИИ
И
МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

**РОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ**

МАТЕРИАЛЫ КОНФЕРЕНЦИИ

Новосибирск
2019

978-5-91434-048-0

СОДЕРЖАНИЕ

ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Аненков А.Д. Реализация алгоритма Дж. Брука для неблокирующей операции Allgather в библиотеке LibNBC. 5
Коротецкий И.А., Ракитский А.А. Теоретический метод для оценки и сравнения производительности процессоров на базе архитектуры MIPS. 11
Крюкова Л.П. Исследование метода программной конвейеризации циклов. 16
Курносов М.Г. Определение оптимальных параметров сегментации сообщений в алгоритмах широковещательной передачи стандарта MPI. 22
Новиков П.Л., Павский К.В., Баранов А.А. Расчет потенциального рельефа структурированных подложек кремния методом молекулярной динамики − моделирование с использованием параллельных алгоритмов. 28
Ткачёва Т.А. Анализ древовидных алгоритмов операции трансляционного обмена в стандарте MPI. 32
Токмашева Е.И. Исследование эффективности алгоритма Butterfly глобальной редукции на вычислительном кластере с сетью Gigabit Ethernet. 37
Фульман В.О. Анализ производительности векторных и индексных производных типов данных. 43

ИНФОРМАТИКА И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Подсекция НГТУ

Блинов П.Ю., Лемешко Б.Ю. К вопросу ранжирования множества критериев проверки отклонения от равномерного закона. 47
Глухов Г.И. Моделирование мимики человека. 54
Гриф А.М. Подход к определению взаимовлияния добывающих и нагнетательных скважин в динамике их работы. 59
Кочнев А.В., Волкова В.М. Анализ работы алгоритма взвешенного попарного объединения и его вариаций для кластеризации в графах. 66
Кулабухова С.О., Чубич В.М. Процедура активной параметрической идентификации детерминированных нелинейных непрерывно-дискретных систем. 71
Лемешко Б.Ю., Лемешко С.Б., Веретельникова И.В. Критерии проверки статистических гипотез при анализе больших выборок. 79
Осинцева Е.А., Чимитова Е.В. Информационная матрица Фишера для винеровской деградационной модели с учетом объясняющих переменных. 92
Поверин Д.В., Постовалов С.Н. Исследование вероятности обнаружения новых геномных ассоциаций при комбинировании результатов полногеномного анализа ассоциаций. 100
Ступаков И.М., Кондратьева Н.С., Зеленский А.В. О способах совместного учета остаточной намагниченности и вихревых токов при численном моделировании ускорительных магнитов. 114
Филоненко П.А., Постовалов С.Н. Анализ процесса статистического контроля качества при производстве колбасной продукции. 120
Четвертакова Ю.С., Черникова О.С. Исследование нелинейных непрерывно-дискретных систем с применением квадратно-корневого сигма-точечного фильтра. 128

Бастрыкин И.А. Использование технологий дополненной реальности для проведения физических экспериментов. 132
Бочкарев Б.В., Ракитский А.А. Исследование методов информационного анализа электрокардиосигналов. 138
Воронина П.Е., Муртазина М.Ш. Онтологический подход к поддержке процесса разработки программных продуктов по методологии BDD. 147
Емельянов В.В., Полетайкин А.Н. Влияние на товарооборот представленности уникальных номенклатурных позиций товаров на полке гипермаркета. 152
Захарова Т.Э. Экспериментальное обоснование «единой кривой» повреждаемости. 156
Ляхов О.А. Таксономический анализ в маршрутизации транспорта. 162
Павлова У.В., Ракитский А.А. Исследование возможности применения автоматов для прогнозирования временных рядов. 168
Полетайкин А.Н., Данилова Л.Ф. Информационная управляющая система построения компетентностной модели профессиональной образовательной программы. 173
Ракитский А.А., Дьячкова И.С. Система дистанционного анонимного голосования. 179
Сычев В.А., Полетайкин А.Н., Войновский В.А. Модель нечеткого оценивания мероприятий по обеспечению безопасности дорожного движения. 183
Токтошов Г.Ы. Об одной задаче мультикритериальной оптимизации сетей инженерных коммуникаций. 192

Критерии проверки статистических гипотез при анализе больших выборок

Б. Ю. Лемешко, С. Б. Лемешко, И. В. Веретельникова[1]

В работе рассмотрены методы построения оценок при анализе больших данных (Big Data). Демонстрируется влияние на результаты выводов по критерию $\chi 2$ Пирсона выбора числа интервалов и способа группирования. Показывается, как влияет на распределения статистик непараметрических критериев согласия ограниченная точность представления данных в больших выборках. Даются рекомендации по применению критериев для анализа больших выборок.
Показано, что на распределения статистик критериев однородности законов, а также однородности средних и однородности дисперсий влияет неравноточность представления данных в сравниваемых выборках.

Ключевые слова: Big Data, оценивание параметров, проверка гипотез, критерии согласия, критерии однородности, статистическое моделирование.

1. Введение

Вопросы применения статистических методов к анализу больших массивов данных (Big Data) в последние годы вызывают большой интерес. В связи со стремительным накоплением гигантских объёмов информации возникают потребности в анализе накапливаемых данных, в поиске, извлечении и использовании скрытых в них закономерностей, в том числе вероятностных. Естественно, что для анализа больших данных пытаются применять методы и критерии из обширного арсенала классической математической статистики, используя, в том числе, популярные программные системы статистического анализа. При попытках применения для анализа больших данных классического аппарата прикладной математической статистики, как правило, сталкиваются со специфическими проблемами, ограничивающими возможности корректного применения этого аппарата.

В настоящей работе мы будем касаться только методов и критериев, связанных с анализом одномерных случайных величин, реальные проблемы которых нам наиболее знакомы. Можно рассмотреть, по крайней мере, три ситуации, где рост размерности выборок вызывает проблемы в применении методов или критериев.

Во-первых, вследствие “проклятия размерности” хорошо зарекомендовавшие себя методы и алгоритмы становятся неэффективными. В частности, возникают проблемы с вычислением оценок параметров. При использовании методов оценивания, оперирующих негруппированными данными, с ростом размерности анализируемых выборок кардинально растут вычислительные затраты, ухудшается сходимость итерационных алгоритмов, используемых при нахождении оценок. Существенным фактором оказывается неробастность некоторых видов оценок. Естественным способом разрешения данной ситуации видится применение методов оценивания, предусматривающих группирование данных [1]. Но в этом случае возникают вопросы, как использование оценок по группированным данным отразится на свойствах критериев проверки гипотез, в которых будут использоваться эти оценки. Например,

[1] Работа выполнена при поддержке Министерства образования и науки РФ в рамках государственной работы «Обеспечение проведения научных исследований» (№ 1.4574.2017/6.7) и проектной части государственного задания (№ 1.1009.2017/4.6).

как это отразится на распределениях статистик непараметрических критериев согласия при проверке сложных гипотез, которые существенно зависят от метода оценивания параметров [2, 3, 4, 5]?

Во-вторых, многие популярные критерии проверки статистических гипотез не приспособлены даже для анализа выборок порядка тысячи наблюдений, так как информация о распределениях статистик этих критериев при справедливости проверяемой гипотезы представлена лишь краткими таблицами критических значений для некоторых объёмов выборок n. По грубой оценке, такого рода критериев более 80%. Заметим, что возможность применения таких критериев при "разумных" величинах n легко разрешается статистическим моделированием распределений статистик при данном n и справедливости проверяемой гипотезы H_0, которое может осуществляться в интерактивном режиме в ходе статистического анализа [6, 7]. Построенное в результате N имитационных экспериментов эмпирическое распределение $G_N(S_n|H_0)$ статистики S критерия далее может использоваться для оценки достигнутого уровня значимости p_{value} по значению статистики S^*, вычисленному по анализируемой выборке.

В-третьих, применение критериев проверки гипотез, для которых известны предельные (асимптотические) распределения статистик, с ростом объёмов выборок всегда приводит к отклонению даже справедливой проверяемой гипотезы. Это характерно, например, для критериев согласия, для множества специальных критериев, применяемых для проверки гипотез о принадлежности выборок нормальному, равномерному и показательному законам и т.п. В [8] показано, что корни этой проблемы связаны не только и не столько с ростом вычислительных затрат, сколько с ограниченной точностью представления анализируемых данных (с ограниченной точностью измерений). Аналогичная проблема препятствует корректности применения к большим выборкам критериев проверки гипотез об однородности (однородности законов, однородности дисперсий, в меньшей степени однородности средних). Как будет показано, в случае критериев однородности причина кроется в неравноточности измерений в анализируемых выборках.

2. Оценивание параметров законов распределения

Оценки параметров законов могут находиться различными методами. Наилучшими асимптотическим свойствами обладают оценки максимального правдоподобия (ОМП), вычисляемые в результате максимизации функции правдоподобия

$$\hat{\theta} = \arg\max_{\theta} \prod_{j=1}^{n} f(x_j, \theta), \tag{1}$$

или её логарифма, где θ – неизвестный параметр (в общем случае векторный), $f(x,\theta)$ – функция плотности закона распределения, $x_1, x_2, ..., x_n$ – выборка, по которой вычисляется оценка $\hat{\theta}$. Для некоторых законов распределения ОМП параметров получаются в виде просто вычисляемых статистик от элементов выборок, но в большинстве случаев находятся в результате использования некоторого итерационного метода.

При вычислении *MD*-оценок (оценок минимального расстояния) по θ минимизируется некоторая мера близости (расстояние) $\rho(F(x,\theta), F_n(x))$ между теоретическим $F(x,\theta)$ и эмпирическим $F_n(x)$ распределениями. *MD*-оценки находятся в процессе решения задачи

$$\hat{\theta} = \arg\min_{\theta} \rho(F(x,\theta), F_n(x)). \tag{2}$$

В качестве мер близости можно использовать, например, статистики непараметрических критериев согласия (Колмогорова, Крамера–Мизеса–Смирнова, Андерсона–Дарлинга, Купера, Ватсона и других [9]).

При относительно малых объёмах выборок могут использоваться *L*-оценки параметров, представляющие собой некоторые линейные комбинации порядковых статистик (элементов вариационного ряда $X_{(1)} < X_{(2)} < ... < X_{(n)}$, построенного по выборке $x_1, x_2, ..., x_n$).

ОМП параметров законов распределения, как правило, не являются робастными. Наличие аномальных наблюдений или ошибочность предположения о виде закона приводят к построению моделей с функциями распределения, неприемлемо отклоняющимися от эмпирических распределений. *MD*-оценки обладают большей устойчивостью.

Очевидно, что при очень больших выборках вычисление оценок (1) и (2) связано с серьёзными вычислительными трудностями.

В случае группированной выборки имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины на k непересекающихся интервалов граничными точками

$$x_{(0)} < x_{(1)} < \ldots < x_{(k-1)} < x_{(k)},$$

где $x_{(0)}$ – нижняя грань области определения случайной величины X; $x_{(k)}$ – верхняя грань области определения случайной величины X.

ОМП по группированной выборке [1] вычисляется в результате максимизации функции правдоподобия

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{k} P^{n_i}(\theta), \tag{3}$$

где $P_i(\theta) = \int\limits_{x_{(i-1)}}^{x_{(i)}} f(x, \theta) dx$ – вероятность попадания наблюдения в i-й интервал значений, n_i – количество наблюдений, попавших в i-й интервал, $\sum_{i=1}^{k} n_i = n$.

Оценки по группированным данным можно получать, минимизируя статистику χ^2

$$\hat{\theta} = \arg\min_{\theta} n \sum_{i=1}^{k} \frac{\left(n_i / n - P_i(\theta)\right)^2}{P_i(\theta)}, \tag{4}$$

а также ряд других статистик. В [10] на основании анализа рассмотренной совокупности методов оценивания параметров по группированным данным показано, что все они при соответствующих условиях регулярности дают состоятельные и асимптотически эффективные оценки, но наиболее предпочтительными оценками являются ОМП. Важным достоинством оценок по группированным данным является робастность [11].

При наличии негруппированных данных к оценкам по группированным данным обращаются редко. Связано это с большей трудоёмкостью вычислительного процесса, часто с необходимостью многократного использования численного интегрирования при вычислении $P_i(\theta)$ и требует соответствующей программной поддержки.

В случае больших объёмов выборок ситуация меняется. При фиксированном числе интервалов группирования с ростом объёмов выборок вычислительные затраты не меняются, а возрастают только с увеличением количества интервалов k. Это значит, что в условиях Big Data целесообразно использовать ОМП по группированным выборкам. Это робастные и асимптотически эффективные оценки. При малом k качество оценок можно улучшать, используя асимптотически оптимальное группирование (АОГ) [1, 12, 13], при котором минимизируются потери в информации Фишера, связанные с группированием.

3. О применении критерия χ^2 Пирсона к большим выборкам

Статистику критерия согласия χ^2 Пирсона вычисляют по формуле

$$X_n^2 = n\sum_{i=1}^{k}\frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}. \qquad (5)$$

В случае проверки простой гипотезы при $n \to \infty$ эта статистика подчиняется χ_r^2-распределению с $r = k-1$ степенями свободы, если верна нулевая гипотеза.

При проверке сложной гипотезы и оценивании по выборке m параметров закона статистика (4) в случае справедливости H_0 подчиняется χ_r^2-распределению с $r = k-m-1$ степенями свободы, если оценки получаются минимизацией (4) этой статистики, или используются ОМП (3) (или другие асимптотически эффективные оценки по группированным данным).

При оценивании параметров по негруппированным данным распределение статистики (5) не подчиняется χ_{k-m-1}^2-распределению. При использовании ОМП по негруппированным данным рекомендуется применять критерий Никулина–Рао–Робсона [14, 15].

Принципиальные проблемы, препятствующие применению критерия χ^2 Пирсона для анализа Big Data, отсутствуют: возможны только вычислительные трудности.

Проиллюстрируем результаты применения критерия χ^2 Пирсона на примере достаточно большой выборки, принадлежащёй нормальному закону с плотностью

$$f(x,\theta) = \frac{1}{\theta_1\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta_0)^2}{2\theta_1^2}\right\}.$$

Выборка объёмом $n = 10^7$ смоделирована по стандартному нормальному закону N(0,1) ($\theta_0 = 0$, $\theta_1 = 1$).

В таблице 1 представлены результаты применения критерия при проверке простой гипотезы о принадлежности выборки закону N(0,1) при различном числе интервалов в случае равночастотного группирования (РЧГ) и при $k = 15$ в случае (АОГ). При АОГ максимизируется мощность критерия χ^2 Пирсона относительно близких конкурирующих законов [16, 17, 18]. В таблице приведены значения X_n^{2*} статистики (5), вычисленные по выборке, и соответствующие значения достигнутого уровня значимости $p_{value} = P\{X_n^2 \ge X_n^{2*} | H_0\}$. Как можно видеть, результаты зависят как от способа разбиения, так и от числа интервалов. От этого же зависит и мощность критерия [19].

Таблица 1. Результаты проверки простой гипотезы о согласии с N(0,1)

	АОГ	РЧГ						
	$k=15$	$k=15$	$k=50$	$k=75$	$k=100$	$k=500$	$k=1000$	$k=2000$
X_n^{2*}	7.75162	9.18380	56.8942	79.4904	96.5701	493.995	1044.57	2099.91
p_{value}	0.90186	0.81910	0.20475	0.31026	0.55038	0.55482	0.15403	0.05702

В таблице 2 приведены результаты проверки сложных гипотез. Представлены ОМП $\hat{\theta}_0$ и $\hat{\theta}_1$ по группированным данным, полученные при соответствующем числе интервалов k, значения статистик X_n^{2*} и p_{value}.

ОМП параметров по полной негруппированной выборке $\hat{\theta}_0 = 0.000274$, $\hat{\theta}_1 = 1.000177$. В [20, 21] построены модели распределений статистики (5) для случая проверки сложной гипотезы относительно нормального закона с использованием ОМП по негруппированным дан-

ным и применением АОГ. Вычисленное по выборке значение статистики $X_n^{2*} = 6.600521$ при $k = 15$, а полученная в соответствии с приведенной в [20, 21] моделью предельного распределения оценка $p_{value} = 0.886707$, что свидетельствует о хорошем согласии полной выборки с нормальным законом $\mathrm{N}(0.000274, 1.000177)$.

Таблица 2. Результаты проверки сложной гипотезы

	АОГ	РЧГ						
k	15	15	50	75	100	500	1000	2000
$\hat{\theta}_0$	0.000276	0.000301	0.0002440	0.000270	0.000268	0.000277	0.000273	0.000274
$\hat{\theta}_1$	1.007150	1.002629	1.001730	1.001338	1.001123	1.000399	1.000305	1.000236
X_n^{2*}	927.9202	99.99627	101.7669	104.5111	112.1514	493.7161	1043.471	2098.605
p_{value}	0.0	5.58e-16	6.50e-06	0.007396	0.139377	0.533166	0.149218	0.055723

Обратим внимание, что ОМП по группированной выборке при $k = 2000$ и ОМП по негруппированной очень близки. И в то же время при $k = 2000$ величина p_{value} оказывается много ниже значения 0.886707.

Таким образом, и при проверке сложных гипотез по критерию χ^2 Пирсона результат также существенно зависит от числа интервалов k.

4. Непараметрические критерии согласия и большие выборки

Если опустить рост вычислительных трудностей, то основной причиной возможной некорректности выводов при анализе больших данных с использованием непараметрических критериев согласия является ограниченная точность представления этих данных.

Как правило, объёмы выборок в Big Data (принадлежащих некоторому непрерывному закону распределения) практически неограничены, но сами данные представлены с ограниченной точностью (округлены с некоторым Δ). По сути, “нарушается предположение” о том, что наблюдается непрерывная случайная величина.

Допустим, для проверки простой гипотезы H_0: $F_n(x) = F(x)$, где $F_n(x)$ – эмпирическое распределение, построенное по выборке $x_1, x_2, ..., x_n$ объёма n, применяется критерий согласия со статистикой S, Пусть для данного критерия существует предельное распределение $G(S|H_0)$ статистики. При справедливости H_0 эмпирическое распределение $F_n(x)$, соответствующее выборке непрерывных случайных величин (без округления), при $n \to \infty$ сходится к функции распределения $F(x)$ этой случайной величины. Эмпирическое распределение $G_N(S_n|H_0)$ статистики, строящейся по выборкам непрерывной случайной величины при $n \to \infty$ (и числе имитационных экспериментов $N \to \infty$) сходится к предельному распределению $G(S|H_0)$ этой статистики.

Пусть теперь результаты измерений округляются (фиксируются) с некоторым Δ. Тогда, начиная с некоторого *n*, зависящего от вида $F(x)$, от области определения случайной величины и от Δ, $\max|F_n(x) - F(x)|$ перестанет уменьшаться, а распределение $G_N(S_n|H_0)$ – станет с ростом *n* отклоняться от предельного $G(S|H_0)$ (чем больше Δ, тем при меньшем *n*).

Результаты исследований, демонстрирующих влияние точности регистрации данных на распределения статистик, покажем на 3-х классических критериях согласия.

В критерии Колмогорова рекомендуется использовать статистику с поправкой Большева [9]

$$S_K = \sqrt{n}D_n + \frac{1}{6\sqrt{n}} = \frac{6nD_n + 1}{6\sqrt{n}}, \qquad (6)$$

где $D_n = \max\left(D_n^+, D_n^-\right)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(x_i, \theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n}\left\{F(x_i, \theta) - \frac{i-1}{n}\right\}$; n – объем выборки; $x_1, x_2, \ldots, x_n$ здесь и далее – упорядоченные по возрастанию выборочные значения; $F(x, \theta)$ – функция закона распределения, согласие с которым проверяют. Распределение величины S_K при простой гипотезе в пределе подчиняется закону Колмогорова с функцией распределения $K(S)$ [9].

Статистика критерия Крамера–Мизеса–Смирнова имеет вид

$$S_\omega = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i, \theta) - \frac{2i-1}{2n}\right\}^2, \qquad (7)$$

которая при простой гипотезе в пределе подчиняется закону с функцией распределения $a1(s)$ [9].

Статистика критерия Андерсона–Дарлинга задается выражением [22]

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i, \theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i, \theta))\right\}. \qquad (8)$$

При проверке простой гипотезы эта статистика в пределе подчиняется закону с функцией распределения $a2(s)$ [9].

Распределения статистик (6)–(8) непараметрических критериев согласия исследовались в зависимости от точности регистрации наблюдаемых значений случайных величин. Задавалось число значимых десятичных разрядов, до которых округлялись наблюдаемые величины. Это определяло число уникальных значений, которые могли оказаться в генерируемых выборках. Как правило, число имитационных экспериментов, осуществляемых для моделирования эмпирических распределений статистик, составляло величину $N = 10^6$.

Отклонение реального (эмпирического) распределения статистики от предельного распределения отслеживалось при оценке медианы $\tilde{S}_n$ эмпирического распределения статистики, полученного в результате моделирования. Если реальное распределение статистики при объёмах выборок n не отклоняется от предельного, то вероятность $P\left\{S > \tilde{S}_n\right\}$, вычисляемая по соответствующему предельному распределению равна 0.5. При сдвиге реального распределения статистики в область больших значений (вправо от предельного) оценки $\hat{p}_v = P\left\{S > \tilde{S}_n\right\}$ будут уменьшаться. По величине отклонения оценок $\hat{p}_v$ от 0.5 можно судить о величине погрешности оценки достигнутого уровня значимости p_{value}, вычисляемой по предельному распределению статистики (в случае проверки простых гипотез, соответственно, по $K(S)$, $a1(S)$ и $a2(S)$).

В таблице 3 представлены оценки медиан $\tilde{S}_n$ эмпирических распределений статистик и соответствующие вероятности $\hat{p}_v = P\left\{S > \tilde{S}_n\right\}$, вычисляемые по предельным распределениям статистик критериев при проверке простой гипотезы о принадлежности выборок стандартному нормальному закону в зависимости от объёмов выборок n при регистрации наблюдений с округлением до заданного числа знаков после десятичной точки. В первой колонке таблицы приведены значения $\tilde{S}_n$ и $p_v = P\left\{S > \tilde{S}_n\right\}$ для предельных распределений статистик.

При округлении с точностью до 1 в выборках, принадлежащих $\mathrm{N}(0,1)$, может появляться 9 уникальных значений, при округлении с точностью до $\Delta = 0.1$ – порядка 86 уникальных значений, с точностью $\Delta = 0.01$ – порядка 956, с точностью до $\Delta = 0.001$ – порядка 9830.

Как показали результаты моделирования при округлении наблюдений до целых значений использование предельных распределений статистик критериев **абсолютно** исключено.

При $\Delta = 0.1$ распределения статистики критерия Колмогорова $G(S_n | H_0)$ обладают существенной дискретностью. Для критерия Колмогорова отклонение $G(S_n | H_0)$ от предельного распределения $K(S)$ при $\Delta = 0.1$ следует учитывать уже для $n > 20$, при $\Delta = 0.01$ – для $n > 250$, при $\Delta = 0.001$ величина $n_{\max}$ сдвигается до величины порядка 10^4.

В случае критериев Крамера–Мизеса–Смирнова и Андерсона–Дарлинга отклонение $G(S_n | H_0)$ от предельных $a1(S)$ и $a2(S)$ при $\Delta = 0.1$ надо учитывать для $n > 30$, при $\Delta = 0.01$ – для $n > 1000$, при $\Delta = 0.001$ – величина $n_{\max}$ сдвигается до 5×10^5.

Таблица 3. Оценки медиан эмпирических распределений статистик и вероятностей $\hat{p}_v$

Критерий Колмогорова								
Δ=0.1		$K(S)$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 100$
	$\tilde{S}_n$	0.827574	0.8261	0.8389	0.8480	0.8618	0.8721	0.9149
	$\hat{p}_v$	0.5	0.5023	**0.4897**	0.4663	0.4597	0.4235	0.3724
Δ=0.01		$K(S)$	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 500$	$n = 1000$
	$\tilde{S}_n$	0.827574	0.8289	0.8309	0.8311	0.8348	0.8385	0.85233
	$\hat{p}_v$	0.5	0.4994	0.4962	0.4937	**0.4882**	0.4840	0.4618
Δ=0.001		$K(S)$	$n = 500$	$n = 1000$	$n = 5000$	$n = 10000$	$n = 20000$	$n = 50000$
	$\tilde{S}_n$	0.827574	0.8271	0.8280	0.8301	0.8353	0.8423	0.8538
	$\hat{p}_v$	0.5	0.5007	0.4994	0.4960	0.4879	0.4770	0.4596
Критерий Крамера–Мизеса–Смирнова								
Δ=0.1		$a1(S)$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 100$	$n = 150$
	$\tilde{S}_n$	0.11888	0.1214	0.1218	0.1223	0.1231	0.1267	0.1304
	$\hat{p}_v$	0.5	0.4897	0.4882	0.4861	0.4832	0.4690	0.4551
Δ=0.01		$a1(S)$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 10000$	$n = 20000$
	$\tilde{S}_n$	0.11888	0.1192	0.1193	0.1198	0.1229	0.1263	0.1340
	$\hat{p}_v$	0.5	0.4988	0.4984	0.4962	0.4838	0.4708	0.4423
Δ=0.001		$a1(S)$	$n = 10000$	$n = 5 \times 10^4$	$n = 10^5$	$n = 2 \times 10^5$	$n = 5 \times 10^5$	$n = 10^6$
	$\tilde{S}_n$	0.11888	0.11886	0.11890	0.11887	0.11967	0.1210	0.1250
	$\hat{p}_v$	0.5	0.5001	0.4999	0.5000	0.4968	0.4913	0.4756
Критерий Андерсона–Дарлинга								
Δ=0.1		$a2(S)$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 100$	$n = 150$
	$\tilde{S}_n$	0.774214	0.7798	0.7842	0.7883	0.7931	0.8138	0.8334
	$\hat{p}_v$	0.5	0.4958	0.4926	0.4895	0.4860	0.4712	0.4575
Δ=0.01		$a2(S)$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 10000$	$n = 20000$
	$\tilde{S}_n$	0.774214	0.7744	0.7759	0.7792	0.7956	0.8144	0.8523
	$\hat{p}_v$	0.5	0.5002	0.4987	0.4963	0.4842	0.4708	0.4448
Δ=0.001		$a2(S)$	$n = 10000$	$n = 5 \times 10^4$	$n = 10^5$	$n = 2 \times 10^5$	$n = 5 \times 10^5$	$n = 10^6$
	$\tilde{S}_n$	0.774214	0.7753	0.7762	0.7767	0.7778	0.7922	0.8153
	$\hat{p}_v$	0.5	0.4992	0.4985	0.4982	0.4973	0.4867	0.4701

На рис. 1 показана зависимость распределений статистики (7) критерия Крамера–Мизеса–Смирнова от степени округления Δ при объёмах выборок $n = 1000$ для случая про-

верки простой гипотезы о принадлежности выборки стандартному нормальному закону. На рисунке приведено предельное распределение $a1(S)$, имеющее место в ситуации без округления, а также реальные распределения $G(S_{1000}|H_0)$ статистики при степени округления $\Delta = 0.01, 0.05, 0.1, 0.2, 0.3$. Как можно видеть, при $\Delta = 0.01$ распределение $G(S_{1000}|H_0)$ практически не отличается от $a1(S)$, но с ростом Δ отклонение $G(S_{1000}|H_0)$ от $a1(S)$ быстро увеличивается.

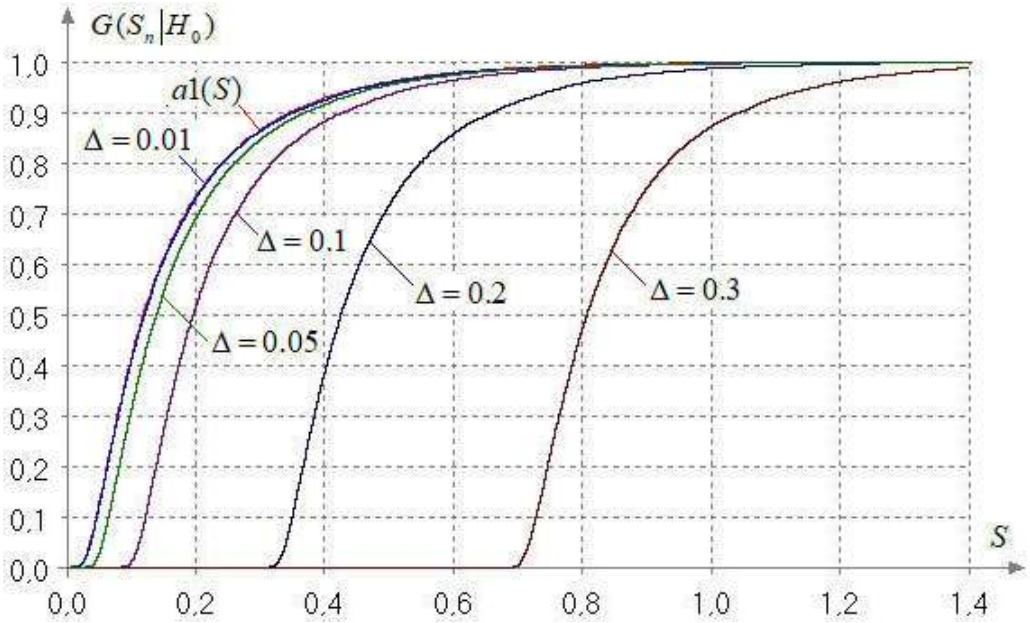

Рис. 1. Распределения статистики $G(S_n|H_0)$ критерия Крамера–Мизеса–Смирнова в зависимости от Δ при $n = 1000$

Следовательно, для того чтобы при анализе больших выборок с использованием соответствующего непараметрического критерия согласия можно было воспользоваться классическими результатами (предельными распределениями), статистика должна вычисляться не по всему большому массиву, а по выборкам, извлекаемым по равномерному закону из “генеральной совокупности”, роль которой в данном случае играет анализируемый большой массив данных. Объём извлекаемой выборки должен учитывать точность фиксируемых данных (количество возможных уникальных значений в выборке) и не превышать некоторой величины $n_{\max}$, при которой (при данной точности) распределение статистики $G(S_{n_{\max}}|H_0)$ критерия при справедливости H_0 ещё реально не отличается от предельного распределения $G(S|H_0)$ статистики этого критерия.

При проверке сложных гипотез проверяемая гипотеза имеет вид H_0: $F(x) \in \{F(x, \theta), \theta \in \Theta\}$, где Θ – область определения параметра θ. Если оценка $\hat{\theta}$ скалярного или векторного параметра закона опирается на ту же самую выборку, по которой проверяется гипотеза, то распределение статистики $G(S|H_0)$ любого непараметрического критерия согласия существенно отличается от предельного, имеющего место при проверке простой гипотезы [23]. При оценивании параметров по этой же выборке на закон распределения статистики $G(S|H_0)$ влияют следующие факторы [24]: вид наблюдаемого закона распределения $F(x, \theta)$, соответствующего истинной гипотезе H_0; тип оцениваемого параметра и число оцениваемых параметров; в некоторых ситуациях конкретное значение параметра (например, в случае гамма-распределения и т.п.); используемый метод оценивания параметров.

Очевидно, что в случае проверки сложных гипотез при анализе Big Data с ограниченной точностью фиксируемых данных мы сталкиваемся с теми же проблемами и должны извлекать из “генеральной совокупности” выборки объёма $n < n_{\max}$, чтобы использовать, например, модели предельных распределений статистик критериев, имеющие место при проверке сложных гипотез [2, 3, 4, 5, 24].

Сделаем важное замечание. Если оценку $\hat{\theta}$ вектора параметров находить одним из рассмотренных выше методов по всему массиву больших данных, а далее критерий применять к выборке объёма $n < n_{\max}$, извлекаемой из этого же массива, то при проверке гипотезы $H_0 : F(x) = F(x, \hat{\theta})$, где $\hat{\theta}$ – полученная ранее оценка, распределение статистики $G(S|H_0)$ будет то же самое, что и при проверке простой гипотезы.

Всё вышесказанное в полной мере относится и к применению к большим выборкам непараметрических критериев согласия Купера [25] и Ватсона [26, 27].

Распределения статистик 3-х критериев согласия Жанга [28], представляющих собой развитие критериев Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга, зависят от объёмов выборок n. Поэтому не может идти речи об использовании предельных распределениях статистик. Но распределения статистик $G(S_n|H_0)$ таким же образом зависят от от степени округления Δ. Следовательно, критические значения статистик, полученные при заданном n для ситуации непрерывных случайных величин, нельзя использовать при тех же n, но при существенной степени округления Δ. Проблема может разрешаться статистическим моделированием (в том числе, в интерактивном режиме [6, 7]) распределений статистик при заданных n и Δ при справедливости проверяемой гипотезы H_0. Построенное в результате N имитационных экспериментов в этих условиях эмпирическое распределение $G_N(S_n|H_0)$ статистики S соответствующего критерия может использоваться для оценки достигнутого уровня значимости p_{value}. Именно так в подобной ситуации эта проблема разрешается в развиваемой программной системе ISW [29].

Отметим, что подобным же образом степень округления регистрируемых данных влияет на свойства множества других критериев, в частности, специальных критериев, ориентированных на проверку гипотезы о принадлежности выборок нормальному закону, о принадлежности выборок равномерному закону, показательному закону и др.

5. Критерии однородности при больших выборках

В критериях однородности одновременно сравнивается 2 и более выборок. В многовыборочных критериях на распределения статистик влияет неравноточность данных, представленных в выборках.

Двухвыборочный критерий однородности Лемана–Розенблатта предложен в работе [30] и исследован в [31]. Статистика, построенная по двум выборкам $x_{11}, x_{12}, ..., x_{1,n_1}$ и $x_{21}, x_{22}, ..., x_{2,n_2}$, используется в форме

$$S_{LR} = \frac{1}{n_1 n_2 (n_1 + n_2)} \left[n_1 \sum_{i=1}^{n_1} \left(r_i - i \right)^2 + n_2 \sum_{j=1}^{n_2} \left(s_j - j \right)^2 \right] - \frac{4 n_1 n_2 - 1}{6(n_1 + n_2)}, \tag{9}$$

где r_i – порядковый номер (ранг) x_{1i}; s_j – порядковый номер (ранг) x_{2j} в объединенном вариационном ряде.

Предельным распределением статистики (9) при справедливости проверяемой гипотезы H_0 является то же самое распределение $a1(s)$ [31], которое является предельным для статистики критерия согласия Крамера–Мизеса–Смирнова.

Рассмотрим, как влияет степень округления на распределения статистик критериев однородности законов в случае справедливости H_0 и принадлежности анализируемых выборок стандартному нормальному закону.

На рис. 2 демонстрируется зависимость распределения статистики $G(S_{LR}|H_0)$ критерия однородности Лемана–Розенблатта от степени округления Δ_2 наблюдений во второй выборке при округлении в первой выборке $\Delta_1 = 0.01$ при объёмах выборок $n_i = 1000$.

Уже при $\Delta_2 = 0.05$ отклонение $G(S_{LR}|H_0)$ от $a1(S)$ оказывается существенным. При фиксированном Δ_2 с ростом объёмов выборок отклонение $G(S_{LR}|H_0)$ от $a1(S)$ быстро увеличивается. Отклонение увеличивается с ростом Δ_2 и фиксированном объёме выборки. Распределения статистики $G(S_{LR}|H_0)$ критерия однородности Лемана–Розенблатта зависят от разности Δ_1 и Δ_2.

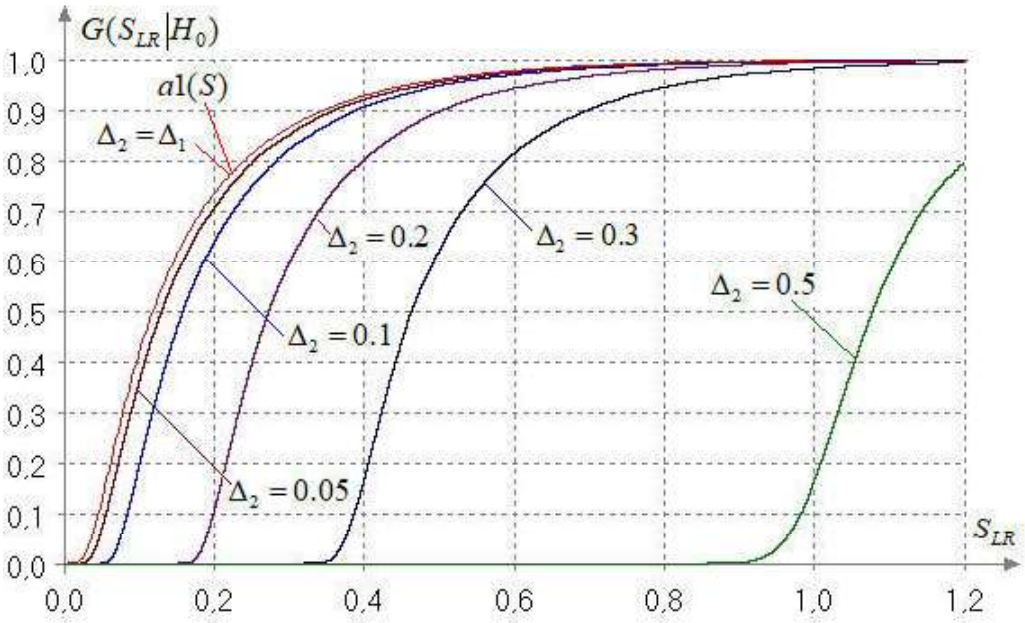

Рис. 2. Распределения статистики $G(S_{LR}|H_0)$ критерия однородности Лемана–Розенблатта при $n_i = 1000$ в зависимости от Δ_2 при $\Delta_1 = 0.01$

Аналогичным образом от разности Δ_1 и Δ_2 зависят распределения других двухвыборочных критериев однородности законов (Смирнова, Андерсона–Дарлинга–Петита). Естественно, что от неравноточности представления данных в анализируемых выборках зависят распределения статистик всех многовыборочных критериев однородности законов, множество которых рассмотрено в [32].

Распределения статистик параметрических критериев однородности средних не страдают такой зависимостью от степени округления результатов наблюдений как рассмотренные выше критерии однородности законов. В то же время следует отметить, что с падением точности регистрируемых данных мощность критериев падает.

Распределения параметрических критериев однородности дисперсий, в отличие от критериев однородности средних, в большей мере зависят от степени округления данных в анализируемых выборках. Некоторым образом, это объясняется большей чувствительностью оценок дисперсии к точности представления результатов измерений.

Можно констатировать, что распределения $G(S|H_0)$ статистик параметрических критериев однородности дисперсий при одинаковой степени округления Δ_i результатов измере-

ний в анализируемых выборках не отличаются от соответствующих распределений в ситуации отсутствия округлений ($\Delta_i = 0,\ \ i = \overline{1,k}$). Но эти же распределения существенно отличаются при неравных Δ_i. При справедливости конкурирующих гипотез степень округления Δ_i (точность регистрации измерений) оказывает существенное влияние и на распределения статистик, и на мощность относительно этих конкурирующих гипотез, в том числе при равных Δ_i. Сказанное справедливо для всего множества параметрических критериев однородности дисперсий, рассмотренных в [32].

6. Заключение

При построении вероятностных моделей по большим выборкам целесообразно использование методов оценивания параметров, предусматривающих группирование данных. В отличие от оценок по негруппированным данным такие оценки робастны, а вычислительные затраты не зависят от объёмов выборок.

Для применения критерия χ^2 Пирсона к анализу больших выборок нет серьёзных возражений: он сохраняет как свои положительные качества, так и свойственные ему недостатки.

Основной проблемой, препятствующей корректному применению непараметрических критериев согласия для анализа больших выборок, является ограниченная точность представления данных. Вследствие ограниченной точности с ростом объёмов выборок реальные распределения статистик отклоняются от предельных, имеющих место в условиях предположения о непрерывности наблюдаемых случайных величин. Поэтому использование классических результатов для соответствующих критериев может приводить к некорректным выводам. В качестве возможного выхода из такой ситуации можно рекомендовать применять критерии к выборкам, извлекаемым из Big data, объём которых ограничивается точностью представления этих данных (количеством возможных уникальных значений в выборке). В качестве другого варианта можно предложить применение методов статистического моделирования для нахождения реальных распределений статистик $G_N(S_n|H_0)$ критериев (соответствующих степени Δ округления данных в анализируемой выборке) с последующим использованием этого $G_N(S_n|H_0)$ для оценки достигнутого уровня значимости p_{value}.

Причиной возможной некорректности выводов при использовании классических результатов, касающихся распределений статистик соответствующих критериев однородности, может оказаться неравноточность измерений в сравниваемых выборках. В качестве выхода из данной ситуации можно предложить статистическое моделирование для нахождения реального распределения статистики $G_N(S_n|H_0)$ применяемого критерия (при соответствующих степенях округления Δ_i и объёмах n_i сравниваемых выборок). Распределение $G_N(S_n|H_0)$ далее можно использовать для оценки достигнутого уровня значимости p_{value}.

Подобная стратегия действий при анализе больших выборок реализуется в программной системе ISW [29].

Литература

1. *Kulldorff G.* Contributions to the theory of estimation from grouped and partially grouped samples. Almqvist & Wiksell. 1961.
2. *Lemeshko B.Yu, Lemeshko S.B.* Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. P. I // Measurement Techniques. 2009. Vol. 52, № 6. – P. 555-565.

3. *Lemeshko B.Yu., Lemeshko S.B.* Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. P. II // Measurement Techniques. 2009. Vol. 52, № 8. – P. 799-812.
4. *Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N.* Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics – Theory and Methods. 2010. Vol. 39, № 3. – P. 460-471.
5. *Lemeshko B.Yu., Lemeshko S.B.* Construction of Statistic Distribution Models for Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses: The Computer Approach // Quality Technology & Quantitative Management. 2011. Vol. 8, No. 4. – P. 359-373.
6. *Lemeshko B.Yu., Lemeshko S.B., Rogozhnikov A.P.* Real-Time Studying of Statistic Distributions of Non-Parametric Goodness-of-Fit Tests when Testing Complex Hypotheses // Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" – AMSA'2011, Novosibirsk, Russia, 20-22 September, 2011. – P. 19-27.
7. *Lemeshko B.Yu., Lemeshko S.B., Rogozhnikov A.P.* Interactive investigation of statistical regularities in testing composite hypotheses of goodness of fit // Statistical Models and Methods for Reliability and Survival Analysis : monograph. – Wiley-ISTE , 2013. – Chap. 5. – P. 61–76.
8. *Лемешко Б.Ю. Лемешко С.Б., Семёнова М.А.* К вопросу статистического анализа больших данных // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2018. № 44. – С. 40-49. DOI: 10.17223/19988605/44/5
9. *Большев Л.Н.,* Смирнов Н.В Таблицы математической статистики. – М. : Наука, 1983. – 416 с.
10. *Рао. С.Р.* Линейные статистические методы и их применения. М.: Наука, 1968. – 548 с.
11. *Лемешко Б.Ю.* Группирование наблюдений как способ получения робастных оценок // Надежность и контроль качества. 1997. № 5. С. 26-35.
12. *Денисов В.И., Лемешко Б.Ю., Цой Е.Б.* Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. В 2-х ч. / Новосиб. гос. техн. ун-т. Новосибирск, 1993. – 347 с.
13. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В. Чимитова. Новосибирск : Изд-во НГТУ, 2011. – 888 с.
14. *Никулин М.С.* О критерии хи-квадрат для непрерывных распределений // Теория вероятностей и ее применение. 1973. Т. XVIII. № 3. – С.75-676.
15. *Rao K.C., Robson D.S.* A chi-squared statistic for goodness-of-fit tests within the exponential family // Commun. Statist. 1974. Vol. 3. – P.1139-1153.
16. *Денисов В.И., Лемешко Б.Ю.* Оптимальное группирование при обработке экспериментальных данных // Измерительные информационные системы. Новосибирск, 1979. – С.5-14.
17. *Лемешко Б.Ю.* Асимптотически оптимальное группирование наблюдений – это обеспечение максимальной мощности критериев // Надежность и контроль качества. 1997. № 8. – С.3-14.
18. *Лемешко Б.Ю.* Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория, 1998. Т. 64. №1. С.56-64.
19. *Лемешко Б.Ю., Чимитова Е.В.* О выборе числа интервалов в критериях согласия типа χ^2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. № 1. С. 61-67.
20. *Лемешко Б.Ю.* Критерии проверки отклонения распределения от нормального закона. Руководство по применению. М.: ИНФРА-М, 2015. – 160 с. DOI: 10.12737/6086
21. *Лемешко Б.Ю.* Критерии согласия типа хи-квадрат при проверке нормальности // Измерительная техника. 2015. № 6. С.3-9.
22. *Anderson T.W., Darling D.A.* A test of goodness of fit // J. Amer. Statist. Assoc. 1954. Vol. 29. P. 765–769.

23. *Kac M., Kiefer J., Wolfowitz J.* On tests of normality and other J. tests of goodness of fit based on distance methods // Ann. Math. Stat. 1955. Vol. 26. – P.189-211.
24. *Лемешко Б.Ю.* Непараметрические критерии согласия: Руководство по применению. М.: ИНФРА-М, 2014. – 163 с. DOI: 10.12737/11873
25. *Kuiper N.H.* Tests concerning random points on a circle // Proc. Konikl. Nederl. Akad. Van Wettenschappen. 1960. Series A. V.63. – P.38-47.
26. *Watson G. S.* Goodness-of-fit tests on a circle. I // Biometrika. 1961. V. 48. No. 1-2. – P.109-114.
27. *Watson G. S.* Goodness-of-fit tests on a circle. II // *Biometrika.* 1962. V. 49. No. 1-2. – P.57-63.
28. *Zhang J.* Powerful goodness-of-fit tests based on the likelihood ratio // Journal of the Royal Statistical Society: Series B. 2002. V.64. № 2. – P.281-294.
29. ISW–Программная система статистического анализа одномерных наблюдений. https://ami.nstu.ru/~headrd/ISW.htm. (дата обр. 03.03.2019)
30. *Lehmann E. L.* Consistency and unbiasedness of certain nonparametric tests // Ann. Math. Statist. – 1951. – Vol. 22, № 1. – P. 165–179.
31. *Rosenblatt M.* Limit theorems associated with variants of the von Mises statistic // Ann. Math. Statist. – 1952. – Vol. 23. – P. 617–623.
32. *Лемешко Б.Ю.* Критерии проверки гипотез об однородности. Руководство по применению. М.: ИНФРА-М, 2017. – 208 с. DOI: 10.12737/22368

Лемешко Борис Юрьевич

Главный научный сотрудник кафедры прикладной и теоретической информатики НГТУ, д.т.н., профессор (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: Lemeshko@ami.nstu.ru, http://www.ami.nstu.ru/~headrd/

Лемешко Станислав Борисович

Старший научный сотрудник кафедры прикладной и теоретической информатики НГТУ, к.т.н., (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: skyer@mail.ru

Веретельникова Ирина Викторовна

Аспирант кафедры прикладной и теоретической информатики НГТУ (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: ira-veterok@mail.ru

Tests for checking statistical hypotheses when analyzing large

B. Yu. Lemeshko, S. B, Lemeshko, I. V. Veretelnikova

Novosibirsk State Technical University

The paper discusses methods for constructing estimates in the analysis of big data (Big Data). The influence on the conclusions by test chi-square Pearson of choosing the number of intervals and the method of grouping is shown. It is shown how the limited accuracy of data representation in large samples affects the distribution of statistics of nonparametric tests of agreement. Recommendations on the application of criteria for analyzing large samples are given.
It is shown that the distribution of statistics of the tests of homogeneity of the laws, as well as of homogeneity of the averages and the homogeneity of the variances, is influenced by the unequal representation of the data in the compared samples.

Keywords: Big Data, parameter estimation, hypothesis testing, goodness of fit tests, homogeneity tests, statistical simulating.