ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2018 Управление, вычислительная техника и информатика № 44

УДК 519.24
DOI: 10.17223/19988605/44/5

Б.Ю. Лемешко, С.Б. Лемешко, М.А. Семенова

К ВОПРОСУ СТАТИСТИЧЕСКОГО АНАЛИЗА БОЛЬШИХ ДАННЫХ

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).

Рассмотрены методы построения оценок при анализе больших данных (Big Data). Демонстрируется влияние на результаты выводов по критерию χ^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.

Ключевые слова: Big Data; оценивание параметров; проверка гипотез; критерии согласия.

Вопросы применения статистических методов к анализу больших массивов данных (Big Data) в последнее время вызывают все больший интерес.

Вполне 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. И тут сталкиваются с тем, что хорошо зарекомендовавшие себя методы и алгоритмы становятся неэффективными из-за «проклятия размерности». Популярные критерии проверки гипотез оказываются неприспособленными для анализа выборок даже порядка тысячи наблюдений. Критерии, которые формально можно использовать при объемах выборок $n \to \infty$, на практике приводят к отклонению даже справедливой проверяемой гипотезы H_0.

В данном случае мы будем касаться только методов и критериев, связанных с анализом одномерных случайных величин, с областью, которая нам наиболее знакома. Можно рассмотреть по крайней мере три ситуации, при которых рост размерности выборок вызывает проблемы в применении методов или критериев.

1. Первая ситуация связана 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 оценок. Естественным выходом является использование методов оценивания, предусматривающих группирование данных.

2. Основная причина, исключающая возможность применения к большим выборкам многих критериев проверки статистических гипотез, заключается в зависимости распределений статистик этих критериев от объемов выборок, в результате чего вся информация о распределениях статистик представлена лишь краткими таблицами критических значений для некоторых объемов n. Отметим, что возможность применения такого рода критерия при «разумных» величинах n легко разрешается интерактивным моделированием распределений статистик при данном n и справедливости проверяемой гипотезы H_0 [1] с последующим использованием построенного эмпирического распределения $G_N(S_n|H_0)$ статистики S для оценки достигнутого уровня значимости p_{value} по значению статистики S^*, вычисленному по анализируемой выборке. Здесь N – количество имитационных экспериментов при статистическом моделировании $G_N(S_n|H_0)$.

3. Существование предельных распределений статистик критериев не гарантирует корректности статистических выводов при использовании последних для анализа больших выборок. Например, применение к выборкам очень большого объема непараметрических критериев согласия, как правило, приводит к отклонению проверяемой гипотезы, даже когда она справедлива. Причина этого кроется в том, что объемы накапливаемых данных практически не ограничены, а исследуемые показатели зафиксированы с ограниченной точностью.

4. Соглашаясь с наличием проблем в применении непараметрических критериев согласия для больших выборок, специалисты рекомендуют использовать критерий χ^2 Пирсона. Однако результаты проверки гипотезы по критериям типа χ^2 бывают неоднозначны, существенно зависят от выбираемого числа интервалов и способа группирования.

В данной работе мы затронем проблемы применения к анализу Big Data некоторых критериев согласия и вопросы, связанные с оцениванием параметров моделей законов распределения.

1. Об оценивании параметров

Для нахождения оценок параметров законов могут использоваться различные методы.

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 ее 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 др. [1]).

При относительно малых объемах 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 серьезными 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 вычисляется в результате максимизации функции правдоподобия

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

где $P_i(\theta) = \int_{x_{(i-1)}}^{x_{(i)}} f(x, \theta)dx$ – 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 в результате минимизации статистики 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 ряда других статистик. В [2] 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 [3].

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

В случае больших объемов выборок ситуация меняется. При фиксированном числе интервалов группирования с ростом объемов 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) [4–6], при котором минимизируются потери в информации Фишера, связанные с группированием.

2. О применении критерия χ^2 Пирсона

Статистику критерия согласия χ^2 Пирсона вычисляют по формуле

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

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 (5) в случае справедливости H_0 подчиняется χ_r^2-распределению с $r = k - m - 1$ степенями свободы, если оценки получаются минимизацией этой статистики (4) или используются ОМП (3) или другие асимптотически эффективные оценки по группированным данным.

При оценивании параметров по негруппированным данным распределение статистики (5) не подчиняется χ^2_{k-m-1}-распределению. При использовании ОМП по негруппированным данным рекомендуется применять критерий Никулина–Рао–Робсона [7, 8].

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, принадлежащей 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 табл. 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 в случае (АОГ) при $k=15$. При АОГ максимизируется мощность критерия χ^2 Пирсона относительно близких конкурирующих законов [9–11]. 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 [12].

В табл. 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}.

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
pvalue	0,90186	0,81910	0,20475	0,31026	0,55038	0,55482	0,15403	0,05702

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

ОМП параметров по полной негруппированной выборке $\hat\theta_0=0,000274$, $\hat\theta_1=1,000177$. В [13, 14] построены модели распределений статистики (5) для случая проверки сложной гипотезы относительно нормального закона с использованием ОМП по негруппированным данным и применением АОГ. Вычисленное по выборке значение статистики $X_n^{2*}=6,600521$ при $k=15$, а полученная в соответствии с приведенной в [13, 14] моделью предельного распределения оценка $p_{value}=0,886707$, что свидетельствует о хорошем согласии полной выборки с нормальным законом $N(0,000274;1,000177)$.

Можно заметить, что и при проверке сложных гипотез результат существенно зависит от числа интервалов k.

3. О применении непараметрических критериев согласия

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

В критерии Колмогорова рекомендуется использовать статистику с поправкой Большева [15]:

$$S_K = \sqrt{n}D_n + \frac{1}{6\sqrt{n}} = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{6}$$

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 $K(S)$ [15].

Статистика критерия Крамера–Мизеса–Смирнова имеет вид:

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

и при простой гипотезе в пределе подчиняется закону с функцией распределения $a1(s)$ [15].

Статистика критерия Андерсона–Дарлинга задается выражением [16]:

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

При проверке простой гипотезы эта статистика в пределе подчиняется закону с функцией распределения $a2(s)$ [15].

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 объемах 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 табл. 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 объемов 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.

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		$K(S)$	$n=50$	$n=100$	$n=200$	$n=300$	$n=500$	$n=1\,000$
	$\tilde{S}_n$	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		$K(S)$	$n=500$	$n=1\,000$	$n=5\,000$	$n=10\,000$	$n=20000$	$n=50\,000$
	$\tilde{S}_n$	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		$a1(S)$	$n=100$	$n=500$	$n=1\,000$	$n=5\,000$	$n=10\,000$	$n=20000$
	$\tilde{S}_n$	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		$a1(S)$	$n=10\,000$	$n=5\times10^4$	$n=10^5$	$n=2\times10^5$	$n=5\times10^5$	$n=10^6$
	$\tilde{S}_n$	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		$a2(S)$	$n=100$	$n=500$	$n=1\,000$	$n=5\,000$	$n=10\,000$	$n=20000$
	$\tilde{S}_n$	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		$a2(S)$	$n=10\,000$	$n=5\times10^4$	$n=10^5$	$n=2\times10^5$	$n=5\times10^5$	$n=10^6$
	$\tilde{S}_n$	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

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 9 830.

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.

Следовательно, при анализе Big Data с использованием соответствующего непараметрического критерия согласия 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. Объем 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 еще реально не отличается от предельного распределения $G(S|H_0)$ 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 [17]. При оценивании параметров по этой же выборке на закон распределения статистики $G(S|H_0)$ влияют следующие факторы [1]: 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 столкнемся с теми же проблемами и должны извлекать из «генеральной совокупности» выборки объема $n < n_{\max}$, чтобы использовать, например, модели предельных распределений статистик критериев, имеющие место при проверке сложных гипотез [1, 18–20].

Если оценку $\hat{\theta}$ вектора параметров находить одним из рассмотренных выше методов по всему массиву больших данных, а далее критерий применять к выборке объема $n < n_{\max}$, 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.

Заключение

В случае больших выборок целесообразно использование методов оценивания параметров, предусматривающих группирование данных. В отличие от оценок по негруппированным данным они робастны, а вычислительные затраты не зависят от объемов выборок.

Нет препятствий для применения к большим выборкам критерия χ^2 Пирсона: он сохраняет как свои положительные качества, так и свойственные ему недостатки.

Ограниченная точность представления данных в больших выборках влияет на распределения статистик непараметрических критериев согласия. Поэтому эти критерии целесообразно применять к выборкам, извлекаемым из Big Data, объем которых ограничивается точностью представления этих данных (количеством возможных уникальных значений в выборке).

ЛИТЕРАТУРА

1. Лемешко Б.Ю. Непараметрические критерии согласия : руководство по применению. М. : ИНФРА-М, 2014. 163 с. DOI: 10.12737/11873.

2. Рао. С.Р. Линейные статистические методы и их применения. М. : Наука, 1968. 548 с.
3. Лемешко Б.Ю. Группирование наблюдений как способ получения робастных оценок // Надежность и контроль качества. 1997. № 5. С. 26–35.
4. Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам. М. : Наука, 1966. 176 с.
5. Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов : в 2 ч. / Новосиб. гос. техн. ун-т. Новосибирск, 1993. 347 с.
6. 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.
7. Никулин М.С. О критерии хи-квадрат для непрерывных распределений // Теория вероятностей и ее применение. 1973. Т. XVIII, № 3. С. 75–676.
8. Rao K.C., Robson D.S. A chi-squared statistic for goodness-of-fit tests within the exponential family // Commun. Statist. 1974. V. 3. P. 1139–1153.
9. Денисов В.И., Лемешко Б.Ю. Оптимальное группирование при обработке экспериментальных данных // Измерительные информационные системы. Новосибирск, 1979. С. 5–14.
10. Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений – это обеспечение максимальной мощности критериев // Надежность и контроль качества. 1997. № 8. С. 3–14.
11. Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория. 1998. Т. 64, № 1. С. 56–64.
12. 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.
13. Лемешко Б.Ю. Критерии проверки отклонения распределения от нормального закона : руководство по применению. М. : ИНФРА-М, 2015. 160 с. DOI: 10.12737/6086.
14. Лемешко Б.Ю. Критерии согласия типа хи-квадрат при проверке нормальности // Измерительная техника. 2015. № 6. С. 3–9.
15. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М. : Наука, 1983. 416 с.
16. Anderson T.W., Darling D.A. A test of goodness of fit // J. Amer. Statist. Assoc. 1954. V. 29. P. 765–769.
17. 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. V. 26. P. 189–211.
18. Лемешко Б.Ю., Лемешко С.Б. Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч. I // Измерительная техника. 2009. № 6. С. 3–11.
19. Лемешко Б.Ю., Лемешко С.Б. Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч. II // Измерительная техника. 2009. № 8. С. 17–26.
20. 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. V. 39, No. 3. P. 460–471.

Поступила в редакцию 15 февраля 2018 г.

Lemeshko B.Yu., Lemeshko S.B., Semenova M.A. (2018) TO QUESTION OF THE STATISTICAL ANALYSIS OF BIG DATA. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 44. pp. 40–49

DOI: 10.17223/19988605/44/5

The search for regularities in Big Data now receives more and more scientific attention. Naturally, the methods for parameter estimation of models and criteria for testing hypotheses from classical mathematical statistic are applied to achieve these aims. At the same time, it is found that well-established methods of evaluation become ineffective because of the "dimensional curse". Most criteria for testing statistical hypotheses are suitable for samples analysis of very limited dimension. The criteria that can be formally used for sizes samples $n \to \infty$ in practice lead to an unjustified rejection of the hypothesis being tested.

In estimation methods that operate with ungrouped data, as the dimension of the analyzed samples increases, the computational costs dramatically increase, the convergence of the iterative algorithms used in the construction of estimates worsens. The non-robustness of the estimates turns out an essential factor.

The reason that excludes the possibility of applying many tests for testing hypotheses to Big Data samples is the dependence of the statistics distribution of these tests on *n* and the availability only short tables of critical values. At reasonable sizes of *n*, this reason can be eliminated by computer technology and statistical simulating methods to find the empirical distribution of statistics necessary for making a decision.

The reason for incorrect conclusions when using tests with known limit distributions of statistics is that sizes *n* in Big Data are "practically unlimited" and these data are presented with limited accuracy.

For a fixed number of intervals with an increase in sample sizes, the computational costs for parameters estimation by grouped data do not change, but increase only with an increase in the number of k intervals. It is recommended that maximum likelihood estimates (MLE) were used for grouped samples. These are robust and asymptotically efficient estimates. For small k, the quality of estimates can be improved using asymptotically optimal grouping, in which the losses in Fisher information associated with grouping are minimized.

Using the example of a Big Data sample, the dependence of the result of applying χ^2 Pearson's criterion for testing a simple and complex hypothesis is shown correspondence on the number of intervals and the method of grouping. It is shown that there are no obstacles to the application of χ^2 Pearson's criterion to large samples, and it retains both its positive qualities and its inherent disadvantages (conclusions are ambiguous, essentially depend on the number of intervals chosen and on the method of grouping).

Statistical distributions of goodness-of-fit tests (Kolmogorov, Cramer–Mises–Smirnov and Anderson–Darling) were studied by statistical simulating methods, depending on the accuracy of the observation record (from the possible number of unique values in the samples).

From obtained results, it follows that when analysis of Big Data using the appropriate non-parametric goodness-of-fit tests, the statistics should not be computed over the entire large array, but on samples extracted from the "general population", whose role in this case is played by the Big Data array being analyzed. The size of the sample to be extracted should take into account the accuracy of the data to be captured (the number of possible unique values in the sample) and not exceed some value of $n_{\max}$, at which (for a given accuracy) the distribution of the statistic $G(S_{n_{\max}}|H_0)$ of the tests for the validity of the hypothesis H_0 does not differ from the limiting distribution $G(S|H_0)$ of this statistic. The presented results allow to estimate $n_{\max}$ values for the considered tests. The estimates of $n_{\max}$ for the Kolmogorov test are substantially lower than for the Cramer–Mises–Smirnov and Anderson–Darling tests. The obtained estimates $n_{\max}$ can apply for using in the goodness-of-fit tests by the Big Data analysis.

Keywords: Big Data; parameter estimation; hypothesis testing; goodness-of-fit test.

LEMESHKO Boris Yurievich (Doctor of Technical Sciences, Professor, Novosibirsk State Technical University, Russian Federation).
E-mail: Lemeshko@ami.nstu.ru

LEMESHKO Stanislav Borisovich (Candidate of Technical Sciences, Associate Professor, Novosibirsk State Technical University, Russian Federation).
E-mail: skyer@mail.ru

SEMENOVA Mariya Alexandrovna (Candidate of Technical Sciences, Associate Professor, Novosibirsk State Technical University, Russian Federation).
E-mail: vedernikova.m.a@gmail.com

REFERENCES

1. Lemeshko, B.Yu. (2014) *Neparametricheskie kriterii soglasiya. Rukovodstvo po primeneniyu* [Nonparametric goodness-of-fit tests]. Moscow: INFRA–M. DOI: 10.12737/11873
2. Rao, S.R. (1968) *Lineynyye statisticheskiye metody i ikh primeneniya* [Linear statistical methods and their applications]. Moscow: Nauka.
3. Lemeshko, B.Yu. (1997) Gruppirovaniye nablyudeniy kak sposob polucheniya robastnykh otsenok [Grouping observations as a way to generate robust estimates]. *Nadezhnost' i kontrol' kachestva.* 5. pp. 26–35.
4. Kulldorf, G. (1966) *Vvedeniye v teoriyu otsenivaniya po gruppirovannym i chastichno gruppirovannym vyborkam* [Introduction to the theory of estimation by grouped and partially grouped samples]. Moscow: Nauka.
5. Denisov, V.I., Lemeshko, B.Yu. & Tsoi, E.B. (1993) *Optimal'noye gruppirovaniye, otsenka parametrov i planirovaniye regressionnykh eksperimentov* [Optimal grouping, parameter estimation, and regression experiment planning]. Novosibirsk: Novosibirsk State Technical University.
6. Lemeshko, B.Yu., Lemeshko, S.B., Postovalov, S.N. & Chimitova, E.V. (2011) *Statisticheskiy analiz dannykh, modelirovanie i issledovanie veroyatnostnykh zakonomernostey. Komp'yuternyy podkhod* [Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach]. Novosibirsk: Novosibirsk State Technical University.
7. Nikulin, M.S. (1973) Chi-square test for continuous distributions with location and scale parameters. *Teoriya veroyatnostey i yeye primeneniya* – Theory of Probability and Its Applications. 18(3). pp. 75–76. (In Russian).
8. Rao, K.C. & Robson, D.S. (1974) A chi-squared statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics.* 3. pp. 1139–1153. DOI: 10.1080/03610927408827216
9. Denisov, V.I., & Lemeshko, B.Yu. (1979) Optimal'noye gruppirovaniye pri obrabotke eksperimental'nykh dannykh [Optimal grouping in the processing of experimental data]. In: *Izmeritel'nyye informatsionnyye sistemy* [Measuring Information Systems]. Novosibirsk: [s.n.]. pp. 5–14.

10. Lemeshko, B.Yu. (1997) Asimptoticheski optimal'noye gruppirovaniye nablyudeniy – eto obespecheniye maksimal'noy moshchnosti kriteriyev [Asymptotically optimal grouping of observations is to ensure the maximum power of the tests]. *Nadezhnost' i kontrol' kachestva*. 8. pp. 3–14.
11. Lemeshko, B.Yu. (1998) Asimptoticheski optimal'noye gruppirovaniye nablyudeniy v kriteriyakh soglasiya [Asymptotically optimum grouping of observations in goodness-of-fit tests]. *Zavodskaya laboratoriya. Diagnostika materialov – Industrial laboratory. Diagnostics of materials*. 64(1). pp. 56–64.
12. Lemeshko, B.Yu. & Chimitova, E.V. (2003) O vybore chisla intervalov v kriteriyakh soglasiya tipa χ^2 [On the choice of the number of intervals in the goodness-of-fit tests of type χ^2]. *Zavodskaya laboratoriya. Diagnostika materialov – Industrial laboratory. Diagnostics of materials*. 69(1). pp. 61–67.
13. Lemeshko, B.Yu. (2015) *Kriterii proverki otkloneniya raspredeleniya ot normal'nogo zakona* [Tests for checking the deviation from normal distribution law]. Moscow: INFRA–M. DOI: 10.12737/6086
14. Lemeshko, B.Yu. (2015) Chi-Square-Type Tests for Verification of Normality. *Measurement Techniques*. 58(6). pp. 581–591. DOI: 10.1007/s11018-015-0759-2
15. Bolshev, L.N. & Smirnov, N. V. (1983) *Tablitsy matematicheskoy statistiki* [Tables for Mathematical Statistics]. Moscow: Nauka.
16. Anderson, T.W. & Darling, D.A. (1954) A test of goodness of fit. *Journal of American Statistics Association*. 29. pp. 765–769.
17. Kac, M., Kiefer, J. & Wolfowitz J. (1955) On tests of normality and other J. tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*. 26. pp. 189–211.
18. Lemeshko, B.Yu. & Lemeshko, S.B. (2009) Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *Measurement Techniques*. 52(6). pp. 555–565. DOI: 10.1007/s11018-009-9330-3
19. Lemeshko, B.Yu. & Lemeshko, S.B. (2009) Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques*. 52(8). pp. 799–812.
20. Lemeshko, B.Yu., Lemeshko, S.B. & Postovalov, S.N. (2010) Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses. *Communications in Statistics – Theory and Methods*. 39(3). pp. 460–471. DOI: 10.1080/03610920903140148