УДК 681.3: 65.012: 004.9 **DOI 10.5072/zenodo.218447**

Джунь Й. В., д.ф.-м.н., профессор (Международный экономико-гуманитарный университет имени академика Степана Демьянчука, г. Ривнэ)

О НЕОБХОДИМОСТИ УЧЕТА НОВЫХ ТЕНДЕНЦИЙ В ОБЛАСТИ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ В ПРАВОВОЙ ИНФОРМАТИКЕ И СТАТИСТИКЕ

Анотація. *Статтю присвячено історії виникнення нових тенденцій в галузі математичної обробки та інтелектуального аналізу статистичних даних. Розкрито, що у статистиці, як науці середніх величин, середнє арифметичне є найкращою оцінкою тільки в разі нормальності вибірки. Показано, що при вибірках обсягом n > 500 приблизно в 85 % випадків їх розподіл не є нормальним і при отриманні статистичних оцінок необхідні інші математичні процедури, описані в «Некласичній теорії похибок вимірів» (НТПВ). Обґрунтовано, що у зв'язку з автоматизацією і комп'ютеризацією вимірювань всі науки вступили в епоху великих вибірок, тому рекомендовано при обробці статистичних даних та розробці програмних продуктів використовувати методи НТПВ при моделюванні, діагностиці і прогностичних обчисленнях в правовій інформатиці та статистиці.*
Ключові слова: *статистичні методи, правова інформатика, некласична теорія помилок.*

Аннотация. *Статья посвящена истории возникновения новых тенденций в области математической обработки и интеллектуального анализа статистических данных. Раскрыто, что в статистике, как науке средних величин, среднее арифметическое является наилучшей оценкой только в случае нормальности выборки. Показано, что при выборках объема n > 500 примерно в 85 % случаев их распределение не является нормальным и при получении статистических оценок необходимы иные математические процедуры, подробно рассмотренные в «Неклассической теории погрешностей измерений» (НТПИ). Обосновано, что в связи с автоматизацией и компьютеризацией измерений все науки вступили в эпоху больших выборок, поэтому рекомендовано при обработке статистических данных и разработке программных продуктов использовать методы НТПИ при моделировании, диагностике и прогностических вычислениях в правовой информатике и статистике.*
Ключевые слова*: статистические методы, правовая информатика, неклассическая теория ошибок.*

Annotation*. The history of genesis of new trends in the field of mathematical processing and the intellectual analysis of statistical data is considered in the article. The arithmetic mean in statistics, as a science of average values, is the best estimate only in the case of sample normality. It is shown that samples' distribution is not normal in about 85 % of cases for samples of volume n > 500 and other mathematical procedures that are necessary for obtaining statistical estimates are discribed in detail in «Nonclassical error theory of measurements» (NETM). In connection with the automation and computerization of measurements, all sciences have entered the era of large samples, therefore, it is recommended using the NETM methods in modeling, its diagnostics and predictive calculations in legal informatics and statistics, when processing statistical data and developing software products.*
Key words*: statistical methods, legal informatics, non-classical error theory.*

Математическая обработка данных в правовой статистике, а также существующие программные продукты, используемые с этой целью, разработаны на основе классических представлений об ошибках наблюдений, изложенных в работах знаменитого немецкого математика К. Ф. Гаусса [1; 2]. Анализ литературных источников свидетельствует, что на основе этих представлений в правовой статистике вычисляются различные средние величины, оценивается их надежность, строятся доверительные интервалы для прогнозных характеристик и т. п. Нужно отметить, что на протяжении более чем 200 лет классические методы прекрасным образом себя зарекомендовали. Так что же произошло? Почему возникла необходимость в иных подходах к математической обработке правовой статистической информации? А произошло то, что осталось до сих пор незамеченным и правильно оцененным ни в европейских, ни в украинских университетах специалистами, обрабатывающими массивы статистической информации: вследствие автоматизации и компьютеризации наблюдений произошло резкое увеличение объемов всевозможной информации, в т. ч. и измерительной. Это произошло во второй половине XX века с началом эры больших выборок.

На первый взгляд кажется, что ничего страшного не произошло: в соответствии с классическими представлениями, выражаемыми законом больших чисел Бернулли, – чем больше наблюдений, тем надежнее средняя оценка наблюдаемой величины. Это написано во всех учебниках по статистике. Однако практика подтверждает совсем иное. Вот что пишет по этому поводу П. Е. Эльясберг, соратник С. П. Королева, осуществлявший математическое обеспечение советской космической программы. В своей работе [3] он сформулировал тезис о том, что свойство состоятельности среднего никогда не осуществляется на практике: «начиная с некоторого момента, дальнейшее увеличение объема измерительной информации не

приводит к повышению точности оценок». Это означает, что в принципе мы никогда не можем полностью исключить систематические ошибки из результатов наблюдений. Но классическая теория математической обработки данных [2] провозглашает в качестве фундаментального постулата следующее: «… нужно с особой силой подчеркнуть, что в последующих исследованиях мы будем говорить о случайных ошибках, не содержащих постоянную часть, ибо по существу, именно они являются предметом нашего исследования; о всех причинах систематических ошибок, мы, по возможности, воздержимся говорить». На практике это означает, что многие статистические методы не работают ввиду наличия систематических ошибок в результатах наблюдений и нерешенным вопросом есть разработка адекватных приемов обработки данных, учитывающие этот фактор.

Не лучше обстояло дело и со вторым фундаментальным постулатом в статистике обработки данных: Гаусс предположил, что случайные ошибки наблюдений следуют нормальному закону, «как наиболее естественной форме» распределения погрешностей [2]. Это было действительно гениальным предположением, вследствие чего этот закон стал использоваться во всех отраслях науки. Но выдающийся немецкий математик и астроном Ф. В. Бессель решил проверить это предположение Гаусса. В 1818 году он исследует эмпирические распределения погрешностей наблюдений 3222 звезд, включенных в его фундаментальный каталог [4], Получился конфуз: исследуемое им распределение ошибок склонений звезд существенно отличаюсь от закона нормального распределения. Мог ли один из крупнейших исследователей и теоретиков ошибок проигнорировать этот факт? По-видимому, он сделал вывод о том, что крупные выборки случайных ошибок нельзя считать нормальными. Подтверждением этого вывода есть то, что опять возвратившись к этому вопросу через 20 лет, он для проверки нормальности случайных ошибок исследовал 4 выборки существенно меньшего объема в 100–470 наблюдений [10], интуитивно чувствуя, что для таких объемов данных закону Гаусса нет альтернативы. В этой работе, Бессель сравнив визуально эмпирические и теоретические частоты 4-х распределений, сделал вывод об «их поразительно близком совпадении» [5]. В то время еще не было каких-либо критериев нормальности, поэтому повторная проверка упомянутого выше вывода, осуществленная Бесселем, показала, что одно из этих четырех исследуемых Бесселем распределений никак нельзя отнести к нормальному, так как оно имеет существенный положительный эксцесс $\varepsilon = +1{,}27 \pm 0{,}28$ [5]. Кроме того, Бессель не только не учел отрицательный результат, полученный им ранее, но и проигнорировал закономерностью, явно присущей всем этим 4 рядам, а именно, – во всех случаях наблюдался избыток ошибок, близких к нулю, над теоретическими гауссовыми частотами. Не смотря на все это

научный мир, благодаря этому исследованию Бесселя, стал полагать, что открыт новый закон природы – и он есть законом ошибок Гаусса. Было забыто замечание самого Гаусса, утверждавшего, что никто не может сказать, каким же на самом деле будет закон распределения случайных ошибок, если наблюдения продолжать до бесконечности. Возникла как бы патовая ситуация – теоретики стали думать, что нормальный закон – это экспериментальный факт, а экспериментаторы – что это математическая теорема. Теоретики стали моделировать этот закон в виде доказательств центральной предельной теории вероятностей. Но практики все чаще и чаще встречались с негауссовыми распределениями ошибок.

Целью нашего исследования является подборка материалов, которые показывают, что закон Гаусса, несмотря на всеобщее его применение, нельзя считать некоторым «всеобщим» законом ошибок.

Наша задача показать, что закон Гаусса не свойственен крупным выборкам. И одним из первых, кто заявил об этом открыто был известный американский математик и астроном С. Ньюком [6], а в [7] он, учитывая негауссов характер ошибок, впервые осуществил неклассический подход к обработке данных, идейно весьма близкий к современным робастным процедурам. Далее в 1900 г. один из основателей математической статистики К. Пирсон в своей статье [8], показал, что некоторые серии наблюдений, опубликованные в поддержку закона Гаусса, показывают такие большие отклонения от него которые, скорее, оправдывают его отбрасывание. Этот кризис, касающийся закона распределения случайных ошибок, успешно разрешил знаменитый кембриджский профессор сэр Гарольд Джеффрис. Он написал: «Действительные распределения ошибок наблюдений обычно следуют нормальному закону достаточно близко и уклонения от него трудно установить, если наблюдений не больше чем 500» [9, § 5.7]. Но если выборка по своему объему $n > 500$, то, как правило, гипотеза нормальности становится несостоятельной. Что же означает этот вывод для правовой статистики, которая, как и всякая другая статистика, является, прежде всего, наукой средних величин? Ответ на этот вопрос дает сам Джеффрис: «Закон Гаусса, конечно, имеет всеобщее применение. В этом случае он допускает три гипотезы:

– что нормальный закон адекватен;

– что среднее есть наилучшее значение оцениваемой величины;

– что средняя квадратическая ошибка есть наилучшей оценкой точности, – взаимно эквивалентны: любая из них заключает в себе две остальные… Если нормальный закон соблюдается, то среднее и стандартное отклонение есть наилучшими оценками, используемыми при поиске истинного значения и их ошибок. Если же нормальный закон неприемлем, то это означает неприемлемость использования арифметического среднего» [10] и, конечно, некорректность оценки его надежности.

Изложим теперь концепцию Джеффриса и результаты ее проверки, полученные нами [15; 16].

Каков же закон ошибок предлагает использовать Джеффрис вместо нормального распределения при наличии больших выборок? Внимательно проанализировав результаты известного эксперимента К. Пирсона [12] и несколько других рядов, он в работе [11] предложил следующее распределение плотности вероятности случайных ошибок:

$$y = \frac{m!}{\left[2\pi(m-0.5)\right]^{0.5}(m-0.5)!\sigma}\left[1+\frac{m^2}{2(m-0.5)^3}\left(\frac{x-a}{\sigma}\right)^2\right]^{-m}, \qquad (1)$$

где a, σ – соответственно параметры положения и рассеяния распределения (1); m – зависящий от эксцесса параметр, который можно считать также мерой уклонения этого распределения от закона Гаусса, для которого $m = \infty$. В [11] Джеффрис также показал, что для полностью независимых случайных ошибок наблюдений, показатель m должен быть в пределах:

$$3 \le m \le 5 \qquad (2)$$

или, что то же самое, распределение (1) должно иметь эксцессы в таких границах:

$$6 \ge \varepsilon \ge 1.2 \qquad (3)$$

Так как (1) является формулой, обобщающей нормальный закон и t – распределение, то для последнего пределы (2) можно представить в виде:

$$5 \le \nu \le 9, \qquad (4)$$

где ν – число степеней свободы распределения Стьюдента.

Здесь следует сделать одно важное замечание. Г. Джеффрис называет (1) распределением Пирсона VII типа. Однако это не совсем отвечает истине. Форму (1) он получил исходя из классической кривой Пирсона VII типа, которая имеет недиагональную информационную матрицу. Классическую кривую Пирсона VII типа он преобразовал к виду (1), который, как и закон Гаусса имеет независимые параметры. Чтобы избежать путаницы, форму (1) правильнее назвать распределением (законом) Пирсона-Джеффриса VII типа, сокращено: *PJVII*–распределением. Обладая необыкновенной научной скромностью, Г. Джеффрис не дал особого названия форме (1), которую он создал. Поэтому некоторые исследователи идентифицируют классическое распределения Пирсона VII типа и даже

обобщенное распределение Коши, с формой (1), хотя этого делать нельзя, так как это различные распределения.

Главная заслуга Джеффриса состоит в том, что вместо закона Гаусса он предложил более глубокую, более обобщенную концепцию идеального вероятностного хаоса в виде формы (1).

Первым, кто по достоинству оценил важность этой концепции Джеффриса, прежде всего в астрометрии и космических исследованиях, был академик АН УССР Е. П. Федоров, известный во всем мире теоретик движения полюсов Земли [13]. Важность изучения этого движения многогранна, она захватывает многие отрасли науки и даже историю человеческой цивилизации. Например, А. Ейнштейн причиной Библейского потопа считал внезапное смещения полюса на 2000 км из района Гренландии к его нынешнему положению. Для изучения движения полюса требовались массовые международные астрономические наблюдения изменяемости широт, что являлось прекрасным полигоном для испытания на практике математических доктрин Г. Джеффриса. Проверка их адекватности, по инициативе Е. П. Федорова, была начата в 1967 г. группой его учеников. Президент НАН Украины Б. Е. Патон так оценил достижения этой группы: «В течении 1959–1973 гг. вокруг Е.П. Федорова, тогда директора Главной астрономической обсерватории АН УССР, сформировался коллектив его учеников, который стал известен как «киевская школа широтников». Вместе с учениками Е. П. Федоров разработал новые методы обработки и оценивания точности широтных наблюдений и применил их при создании уникальной системы координат полюса Земли за 80 лет, которая была названа «киевской». [14, с. 82].

Фундаментальная проверка концепций Джеффриса, выполненная автором по инициативе Е. П. Федорова, подтвердила их правильность [15; 16]. Оказалось, что около 85 % больших выборок с объемами данных n – 500 являются существенно ненормальными и следуют распределению (1). При этом такие же особенности присущи не только астрономическим [17] или космическим [18], но и геодезическим [19], гравиметрическим [20], экономическим [21; 22; 23] и многим другим наблюдениям [24].

Методы обработки таких наблюдений подробно изложены в рамках «Неклассической теории погрешностей измерений» [24]. Имеет ли эта теория значение для правовой статистики, интеллектуального анализа ее данных и прогностических оценках криминогенной, социальной и всяческой иной ситуации? Именно вопросы моделирования таких процессов и составление на основе этого научно-обоснованных прогнозов, требуют применения современных, рафинированных методов анализа данных в условиях, когда все науки, в т. ч. и правоведческие, вступают в эру больших выборок и переизбытка статистической информации.

Таким образом, можно сделать следующий важный вывод: специалисты в области правовой информатики должны учитывать тот факт, что методы статистического анализа данных вступают в эпоху больших выборок. Не нужно забывать о том, что существующие программные продукты для статистического анализа данных, разработаны, в преобладающем своем большинстве, на основе классических представлений об ошибках наблюдений. Эти представления часто становятся совершенно неадекватными в случае больших выборок. Статистическая обработка этих выборок требует других подходов и грамотной диагностики математических моделей, особенно прогностического характера в правовой информатике. Эти подходы, современные адекватные методы обработки больших выборок являются предметом рассмотрения «Неклассической теории погрешностей измерений», которая разработана на факультете кибернетики Международного экономико-гуманитарного университета [24]. Эта теория учитывает современные тенденции в области интеллектуального анализа данных, поэтому ее применение является актуальным.

1. Gauss C. F. Theoria motus corporum in sectionibus conicis Solem ambientium / C. F. Gaus. – 1809. **2.** Gauss C. F. Theoria combinations observationum erroribus minimis obnoxiae / C. F. Gaus. – 1823. **3.** Эльясберг П. Е. Измерительная информация : сколько ее нужно? Как обрабатывать? / П. Е. Эльясберг. – М. : Наука. 1983. – 208 с**. 4**. Bessel F. W. Fundamenta astronomiae / F. W. Bessel. – Konigsberg, 1818. – 325 z. **5.** Bessel F. W. Untersuhungen uber die Wahrscheinlichkeit der Beobachtungs fehler. / F. W. Bessel // Astronomische nachrichten, b. 15, 1838. – 369 z.. **6**. Newcomb S. Generalised Theory of the Combination of Observations so as to obtain the best Result / S. Newcomb, Amer. J. Math. – 1886. – № 1/14, p. 1–249. **7**. Newcomb S. Researches of the Moution of the Moon / S. Newcomb // Astronomical Papers. Published by the US Nautical Office, 1912, vol. 9, p. 1. – 249. **8**. Pearson K. On the Criterion that a given System of Deviation from the Proboble in the Case of a correlated System of Variables is such that it can be reasonably Supposed to have arisen from random Sampling / K. Pearson. – Phil. Mag. – 1900, v, 50, p. 152–160. **9.** Jeffreys H. Theory of Probability / H. Jeffreys. – Oxford. –1939. – 380 p. **10.** Jeffreys H. The Law of Errors and the Combination of Observation / H. Jeffreys. – London : Philos. Trans. Roy. Soc. – 1937, ser, A, № 237, p. 231–271. **11.** Jeffreys H. The Law of Errors in the Greenwich Variation of Latitude observations / H. Jeffreys. – London : Mon. Not. of the RAS, 1939, № 9, p. 703–709. **12.** Pearson K. On the Mathematical Theory of Errors of Judgment with special Reference to the personal Equation / K. Pearson. – Philosophical Transactions of the Royal Socienty of London. Ser. A. 1902, vol. 198, p. 253–296**. 13.** Fedorov E. P. Nutation and forced of the Earths pole / E. P. Fedorov. – London : Pergamon press, 1963. – 152 p. **14.** Федоров Е. П. Избранные труды / Е. П. Федоров, К. : Наукова думка, 2014 – 583 с. 1**5.** Джунь И. В. Анализ параллельных широтных наблюдений, выполненных по общей программе: автореф. дисертации на соискание ученой степени кандидата физико-математических наук : специальность 01.03.01 «Астрометрия и небесная механика» / И. В. Джунь, – Киев : Институт математики АН УССР, 1992 – 12 с. **16**. Джунь И. В. Математическая обработка астрономичес-

кой и космической информации при негауссовых ошибках наблюдений : автореф. диссертации на соискание ученой степени доктора физико-математических наук: специальность 01.03.01 «Астрометрия и небесная механика» / И. В. Джунь, – Киев, ГАО НАН Украины, 1992 – 46 с. **17.** Dzhun J. V. Distribution of Errors in multiple large volume observations / J. V. Dzhun. Springer : Measurement techniques, 2012, vol. 55, № 4, Juli, p. 393–396. **18**. Dzhun J. V. Distribution of Type VII of the Errors of satellite Laser Randing Data / J. V. Dzhun. – Kinematics and Physics of Calestial Bodies. – Allerton Press Inc., New York , 1991, vol. 7, № 3, p.74–84. **19.** Джунь И. В. Метод сравнения точности геодезических приборов, учитывающий эксцесс закона распределения вероятности ошибок / И. В. Джунь. Известие вузов. Геодезия и аэрофотосьемка, 1989, № 3, с. 55–61**. 20.** Джунь И. В. Особенность закона распределения результатов баллистических измерений ускорения силы тяжести / И. В. Джунь, Г. П. Арнаутов, Ю. Ф. Стусь, С. Н. Щеглов. – Издание МГК при Президиуме АН СССР и НПО «Нефтегеофизика». – Повторные гравиметрические наблюдения. Москва : 1984, с. 87–100. **21.** Gazda V. Normal probability Distribution in financial Theory – false Assumption and Consequences / V. Gazda. In : «Business Economics, 1999». Proceding of the International Conference. University of Economics, Faculty of Business Economics, Kosice, 1999, p. 73–75. **22.** Dzhun J. V. The Problems of Probobility Methods in Economics / J. V. Dzhun. – In : Ekonomika firiem, 1998 (zbornik z medzinarodnej Konferencie) /. diel. Bardejovske Kupele 05–06.05.1998, p. 444–448. **23.** Peters E. Fractal Market Analysis. – Applying Chaos Theory to Investment and Economics. / E. Peters. – John Wiley and Sons. Inc., New York, 1981, p.18–53. **24**. Джунь И. В. Неклассическая теория ошибок измерений / И. В. Джунь. – Ровно : Естеро, 2015 – 168 с.

Рецензент: д.геогр.н., професор Калько А. Д.