**УДК 378: 004. 9: 519. 281: 528. 11**

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

## ТЕОРИЯ РАСПРЕДЕЛЕНИЯ БОЛЬШИХ ВЫБОРОК МНОГОКРАТНЫХ НАБЛЮДЕНИЙ И ЕЕ ЗНАЧЕНИЕ В ПРОФЕССИОНАЛЬНОМ ОБРАЗОВАНИИ

***Анотація***. *В статті розглянуто закон розподілу великих вибірок багатократних спостережень, який був запропонований кембриджським професором Г. Джеффрісом. Відмічено, що класичні основи метода найменших квадратів залишались непорушними більше ніж 200 років. Це було обумовлено тим, що виміри велись вручну і як правило, їх обсяги* $n < 500$. *Але в наш час масова автоматизація вимірювальних процесів в усіх галузях знань частіше всього призводить до вибірок обсягом* $n > 500$, *які не вкладаються в рамки нормальності і які вимагають інших засобів математичної обробки. Обґрунтовано, важливість використання цих засобів і нових уявлень про розподіл похибок спостережень у великих вибірках в практиці вищої професійної освіти і педагогічних дослідженнях.*
***Ключові слова:*** *розподіл похибок великих вибірок, практика вищої професійної освіти, розподіл Пірсона-Джеффріса VII типу.*

***Аннотация***. *В статье рассмотрен закон распределения больших выборок многократных наблюдений, который предложен кембриджским профессором Г. Джеффрисом. Отмечено, что классические основы метода наименьших квадратов, оставались незыблемыми более чем 200 лет. Это было обусловлено тем, что измерения велись вручную и их объем был обычно* $n < 500$. *Но в наше время массовая автоматизация измерений во всех областях знаний чаще всего приводит к выборкам объема* $n > 500$, *не укладывающихся в рамки нормальности и требующих иных приемов их математической обработки. Обосновано важность использования этих приемов и новых представлений о распределении ошибок наблюдений в больших выборках в практике высшего профессионального образования и педагогических исследованиях.*
***Ключевые слова:*** *распределение погрешностей больших выборок, практика высшего профессионального образования, распределение Пирсона-Джеффриса VII типа.*

***Annotation.*** *The law of errors distribution of large samples of multiple observations is considered in the article. This law was proposed by Cambridge professor H. Jeffrey. It is noted that the classical foundations of the method of least-squares remained unshakable for more than 200 years. It was due to the fact that the measurements were carried out manually and their volume was*

*usually n < 500. Nowadays, the mass automation of measurements in all demesnes of knowledge leads to samples n > 500 that do not fit into the frames of normality and require other methods of their mathematical processing. The importance of using these methods and new ideas about the errors distribution of observations in large samples in the practice of higher professional education and pedagogical research are noted in the article.*
***Key words:*** *errors distribution of large samples, the practice of higher professional education, Pearson-Jeffreys distribution of type VII.*

**Все в мире** подвержено переменам и, как утверждал основатель математической статистики К. Пирсон, любое научное достижение не является окончательным. Оно – наивероятнейший вывод, который можно получить на основании имеющихся у автора объемов данных. Новая информация или более обширные выборки приводят к новым открытиям и теориям. Но не всегда вовремя и успешно можно проследить за наступившими переменами. Так, например, классический метод наименьших квадратов (МНК) является неотъемлемой частью ряда курсов в высшем профессиональном образовании. В своё время МНК был вершиной математических достижений XVIII века и почти одновременно был создан усилиями трех знаменитых ученых: А. М. Лежандром [1], К. Ф. Гауссом [2] и П. С. Лапласом [3]. Но уже в 30-х годах XX века известный кембриджский математик, геофизик и астроном сэр Г. Джеффрис, исследуя большие выборки (объемом $n > 500$ наблюдений), сделал вывод о несостоятельности для них основополагающей концепции МНК-гипотезы о нормальном распределении ошибок наблюдений.

Фундаментальная проверка вывода Джеффриса, выполненная в НАН Украины под руководством академика Е. П. Федорова подтвердила его правильность [4; 5]. Оказалось, что большие выборки многократных наблюдений в любой отрасли науки, следуют не закону Гаусса, а распределению Пирсона-Джеффриса VII типа (PJVII-распределению). Именно этому распределению следовали большие выборки космических [6], гравиметрических [7], экономических данных [8; 9], ошибки определения фундаментальных физических постоянных: скорости света, постоянной Планка, массы заряда электрона, постоянной сверхтонкой структуры и т.д.

Математическая форма плотности вероятности PJVII-распределения имеет следующий вид:

$$y = c\left[1 + \frac{0.5}{M}\left(\frac{x-\lambda}{\sigma}\right)^2\right]^{-m}, \qquad (1)$$

где $\lambda$, $\sigma$, $m$ – параметры распределения;

постоянная $c = \left[\sqrt{2(m-0{,}5)}\sigma * B\left(m + \frac{1}{2}, \frac{1}{2}\right)\right]^{-1}$;

$B(z,w)$ – бета-функция; $M = \left(m - \frac{1}{2}\right)^3 * m^{-2}$.

**Анализ последних** публикаций и исследований показал, что достижения научной школы Е. П. Федорова в области теории ошибок и теории оценок до настоящего времени все еще не оценены, хотя они явились важнейшим эволюционным этапом в развитии МНК и неклассических процедур в анализе данных. Важнейшим достижением этой школы является создание аналитической теории весовой функции распределения ошибок с определением для неё зон сингулярности, что позволило построить теорию диагностики математического моделирования. Действительный член Метрологической академии РАН, автор основополагающих работ в области теории измерительных систем профессор П. В. Новицкий назвал аналитическую теорию весовой функции научным открытием, превращающим «робастное оценивание» из эвристических попыток в действительную науку [10].

Фундаментальным для анализа данных результатом есть полученный школой Е. П. Федорова результат, подтверждающий универсальность PJVII-распределения для больших выборок наблюдений. Со временем стало очевидно, что эволюция статистических методов анализа данных, пошли именно в том направлении, которое разрабатывали Г. Джеффрис, Е. П. Федоров и его ученики. PJVII-распределение рассматривается сейчас ведущими математиками-статистиками мира как новая вероятностная концепция идеального хаоса [13–15] и работы Г. Джеффриса в этой области признаны пионерскими.

Но по странному стечению обстоятельств основополагающие идеи Г. Джеффриса в области теории методов математической обработки данных, как и работы научной школы Е. П. Федорова, все еще мало известны. В большинстве университетов Украины, да и в европейских университетах о результатах, полученных в Кембридже Г. Джеффрисом и его школой, мало известно. В университетах, как правило, излагают классический МНК без какого-либо учета того, что Джеффрис еще в 1939 г. экспериментально подтвердил его теоретическую и практическую несостоятельность в случае больших выборок [15]. В университетских кругах идеи Джеффриса часто воспринимают не как прорыв к новому, а скорее как занимательный математический курьез, поскольку МНК, как метод моделирования, подтверждал свою эффективность на протяжении более чем 200 лет. Так что же произошло и в чем же сущность наступивших кардинальных изменений? А они такие: на протяжении 200 лет измерения проводились вручную и объемы выборок обычно были малы и не превышали 500 измерений. При таких условиях математические основы МНК остаются еще адекватными. Но произошло то, что было не замечено и все еще должным образом не оценено в высшей школе: во второй половине XX века резко увеличились объемы выборок вследствие автоматизации и компьютеризации измерений. Выборки по объему стали

значительно превышать 500 измерений, вследствие этого основополагающая аксиома МНК-гипотеза нормальности, как правило, становится несостоятельной. К сожалению, до сих пор этот факт большинству университетских исследователей неведом, а изложение курсов анализа данных ведется обычно на уровне представлений начала XVIII века.

Цель настоящего исследования в том, чтобы показать значение тех фундаментальных результатов в развитии теории Data Analysis, которые получены в Кембридже Джеффрисом и Федоровской научной школой в Киеве и Ровно, и которые могут быть определены рамками теории больших выборок (ТБВ). Энциклопедического определения большой выборке нет, но используя классификацию, приведенную в [16] львовским математиком И. Д. Квитом, можно ранжировать выборки следующим образом (табл. 1).

Талица 1

Название выборки в зависимости от её объема

| Название выборки | Объем наблюдений, $n$ |
|---|---|
| Малая выборка | $2 \le n \le 30$ |
| Классическая выборка | $30 < n \le 500$ |
| Большая выборка | $500 < n \le 5000$ |

Приведенная классификация несколько условна, но важна по иной причине: она отражает один из фундаментальных принципов теории познания, согласно которому рост экспериментальной информации об исследуемом явлении есть решающим фактором изменения и уточнения его математической модели. Например, несколько наблюдений дают широкий выбор моделей: при возрастании $n$ создается все более подробная статистическая картина исследуемого явления, отклоняющая предыдущие модели одну за одной. «Всякая теория создается и появляется на свет только для того, чтобы пострадать от фактов» [17]. Иными словами, – время жизни любой теории есть, вообще говоря, функция от количества информации.

Приведенное в табл. 1 разделение выборок по объему измерительной информации важно и в методологическом отношении, так как позволяет классифицировать по этому признаку и методы статистической обработки данных (табл.2).

На то обстоятельство, что большие выборки требуют применения качественно иного метода статистической обработки, впервые обратил внимание Джеффрис [13]. Однако он не сформулировал в законченном виде все основополагающие постулаты этого метода, ограничившись только постулатом о законе (1).

Таблица 2

Классификация методов математико-статистической обработки данных в зависимости от объема выборок

| Малая выборка | Классическая выборка | Большая выборка |
|---|---|---|
| Методы микростатистики, непараметрические критериальные процедуры | Классические методы обработки наблюдений, включая МНК и обычные критериальные процедуры | Методы теории больших выборок на основе использования PJVII-распределения и аналитической теории его весовой функции. |

Попробуем теперь сформулировать полностью основополагающие постулаты ТБВ и дать им краткий комментарий, так как они кардинально меняют традиционные подходы. Например, ныне, опираясь на классические представления, в большинстве университетов учат, что наблюдения, выполненные при постоянных условиях наблюдений имеют равные веса. В ТБВ утверждается, что эти веса могут быть разные. Все зависит от объема выборки. Впервые это показал сам К. Пирсон в своем знаменитом эксперименте [18]. При полностью контролируемых постоянных условиях наблюдений он получал распределения, существенно отличающиеся от нормального закона. Тот факт, что преобладающее большинство больших выборок следует распределению (1), был наглядно продемонстрирован нами в работе [19].

В табл. 3 показано, в чем различие исходных постулатов классической теории ошибок (КТО) и ТБВ.

Таблица 3

Основные постулаты классической теории ошибок (КТО) и теории больших выборок (ТБВ)

| Классическая теория ошибок (КТО) | Теория больших выборок (ТБВ) |
|---|---|
| 1. Ошибки наблюдений при постоянных условиях наблюдений следует закону Гаусса, при котором все наблюдения имеют одинаковый вес (весовая функция нормального распределения $P_i = const$. 2. В результатах наблюдений отсутствуют систематические ошибки. | 1. При большом числе многократных наблюдений их случайные независимые погрешности следуют распределению PJVII- распределению с показателем степени $m$ в пределах $2 \leq m \leq 5$. 2. Индивидуальные веса наблюдений, которые подчиняются PJVII-распределению, характеризует их весовая функция, адаптированная к данному распределению. 3. Влиянием слабых, неисключенных, коррелированных систематических погрешностей можно пренебречь только в том случае, когда весовая функция распределения погрешностей измерений является несингулярной. |

Постулат 1 ТБВ отражает факт универсальности PJVII-распределения, которое фактически является обобщением закона Гаусса и *t*-распределения. Постулат 2 ТБВ отражает наблюдаемый примерно в 75 % случаев факт того, что действительные распределения ошибок или остаточных отклонений имеют существенно отличающийся от нуля положительный эксцесс, а следовательно, а разные веса. Постулат 3 адекватный ответ на несостоятельность постулата 2 КТО. Известный специалист по анализу измерительной информации И. Г. Колчинский писал в [20]: «Источники систематических погрешностей нужно изучать в процессе обработки данных, но никогда нет гарантии, что это можно сделать достаточно хорошо».

Такой вывод означает, что постулат 2 КТО реально выполнить невозможно. Выдающийся специалист, осуществлявший математическое обеспечение советской космической программы, соратник С. П. Королева, П. Е. Эльясберг в [21] высказался еще более определенно: «…опыт решения прикладных задач показывает, что в действительности свойство состоятельности никогда не осуществляется на практике, и, начиная с некоторого момента, дальнейшее увеличении объема используемой измерительной информации не приводит к повышению точности оценок». Этот вывод на деле означает весьма ограниченную дееспособность закона больших чисел (ЗБЧ), согласно которого бесконечное повторение измерений неограниченно приближает нас к «истинному» значению наблюдаемой величины в среднем. С точки зрения ТБВ ЗБЧ несостоятелен, так как после определенного количества измерений начинает проявляться систематическая ошибка метода измерений, которую ни метрологически, ни технически невозможно устранить. В ТБВ найдено приемлемое решение этой проблемы: Джеффрис экспериментально установил какие свойства должно иметь PJVII – распределение при отсутствии заметного влияния систематических ошибок. По Джеффрису, если распределение (1) имеет параметр $m$ в пределах $3 \le m \le 5$, то этим влиянием можно пренебречь. По Хьюберу [II] это требование более жесткое – левая граница в приведенном неравенстве может достигать 2, т.е. объединяя выводы Джеффриса и Хьюбера можно сказать, что если $2 \le m \le 5$, то ошибки наблюдений абсолютно хаотичны и у них нет никакой дополнительной информации. Этот метод ТБВ эффективен при диагностировании качества математической модели (теории) по разностям O–C (Observation–Calculation). Эта новая технология анализа остаточных уклонений O–C на основе параметра $m$ открывает широкий простор для постоянной эволюции и совершенствования математических моделей и теорий в любой отрасли науки. Продвигая с помощью разных усовершенствованный параметр $m$ PJVII-распределения в заветный интервал $2 \le m \le 5$, мы получаем мощный рычаг для непрерывного улучшения модели, а, следовательно, и любой теории.

**Сформулируем теперь главные** выводы нашего рассмотрения – в чем же состоит важность изучения основных идей и подходов ТБВ в высшем профессиональном образовании. Во-первых в том, что современная наука, современное производство стали отраслями больших, если не колоссальных

массивов информации. Большие выборки ошибок измерений, подчиняются обычно не закону Гаусса, а PJVII – распределению. В этом случае средние значения не являются эффективными оценками. Для их получения необходимо использовать весовую функцию. Во-вторых – методы ТБВ, подробно описанные в [10] не только рафинированы, но и просты, и их суть состоит в использовании весовой функции наблюдений $x_i$ которая очень просто находится для распределения (1) по формуле:

$$P_i = \frac{y'}{y(x_i-\lambda)} = \left[\left(\frac{m-0,5}{m}\right)^3 \sigma + \frac{(x_i-\lambda)^2}{2m}\right]^{-1}, \qquad (2)$$

где $y$ – плотность распределения (1);

$x_i - \lambda$ – ошибка наблюдения*;*

$\lambda$ – параметр положения, а $m$ – параметр, характеризующий уклонение PJVII- распределения от нормального закона, для которого $m = \infty$.

Если известен куртозис $\beta_2$ распределения (1), то вес $i$-того наблюдения приблизительно можно определить по формуле [10]:

$$P_i = \frac{5\beta_2 - 9}{2\beta_2 D + (\beta_2 - 3)D}, \qquad (3)$$

где $D$ – дисперсия.

Подчеркнем, что ТБВ не стоит в оппозиции к КТО или к МНК, она является естественной эволюцией классических методов обработки данных, что видно из формулы (3): при $\beta_2 = 3$ (закон Гаусса), мы приходим к классической весовой функции $P_i = D^{-1}$. Кроме того, ТБВ нельзя применить не воспользовавшись сначала классическим МНК. Заметим, что ТБВ может и не потребоваться, если окажется, что куртозис остатков О–С несущественно отличается от нуля. Но если окажется, что $\beta_2 > 3$, то взвешивая весами (2) каждое наблюдение, мы получаем методом приближений эффективные оценки исследуемых величин.

**Несмотря на назревшую** необходимость использования идей ТБВ и неклассических процедур в учебном процессе, об изложенных методах, вследствие наблюдающейся стагнации украинской науки, в университетских кругах практически ничего не известно. Кроме того, в научных журналах сейчас, как показали исследования МГТУ им. Баумана, в научных журналах до 90 % макулатурной информации [22], в которой, можно сказать, утонули блестящие изыскания Г. Джеффриса из Кембриджа и научные открытия школы Е. П. Федорова в НАН Украины по изложенному вопросу. Поэтому, *основная цель нашей работы* есть не только ознакомление с новыми методами, но и призыв к сотрудничеству в столь интригующей области научных исследований.

**1.** Legendre A. M., Nouvelles méthodes pour la détermination des orbites des comètes. / A. M. Legendre, Paris, 1806. **2.** Gauss C. F. Theoria motus corporum coelestium in

sectionibus conicis Solem ambientium / C. F. Gauss, Hamburgi, 1809. **3.** Laplace P. S. Theory analytique des probaboloties / P. S. Laplace, Paris, 1812. **4.** Джунь И. В. Анализ параллельных широтных наблюдений, выполненных по общей программе: автореф. дис. на соискание уч. степени канд. физ.-мат. наук: спец. 01.03.01 «Астрометрия и небесная механика» / И. В. Джунь. – К. : Институт математики АН УССР, 1974. – 14 с. **5.** Джунь И. В. Математическая обработка астрономической и космической информации при негауссовых ошибках наблюдений : автореф. дис. на соис. уч. степени докт. физ.-мат. наук: спец. 01.03.01 «Астрометрия и небесная механика» / И. В. Джунь. – Киев, ГАО НАН Украины, 1992. – 46 с. **6.** Dzhun J. V. Pearson's Distribution of Type VII of the Errors of satellite Laser Ranging Data // J. V. Dzhun, Kinematics and Physics of Calestial Bodies, Allerton Press Inc., New York, 1991, vol. 7, № 3. – P. 74–84. **7.** Джунь И. В. Особенность закона распределения результатов баллистических измерений ускорения силы тяжести / И. В. Джунь, Г. П. Арнаутов, Ю. Ф. Стусь, С. Н. Щеглов // Повторные гравиметрические измерения. – Изд. МГК при Президиуме АН СССР и НПО «Нефтегеофизика». – М.: 1984. – С. 87–100. **8.** Gazda V. Normal probability Distribution in financial Theory-false Assumption and Consequences / V. Gazda. In: «Business Economics 1999». Proceesding of the International Conference. University of Economics, Faculty of Business Economics, Kosice, 1999. – P. 73–75. **9.** Dzhun J. V. The problems of Probability Methods in Economics. In Ekonomika firiem 1998 (Zbornik z medzinarodnej Konferencie) II. diel. Bardejovske Kupele, 5–6. 05. 1998. – P. 444–448. **10.** Джунь И. В. Неклассическая теория погрешностей измерений. / И. В. Джунь, Ровно : Естеро, 2015. – 168 с. **11.** Хампель Ф. Робастность в статистике / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаель. Пер. с англ. М. : Мир. 1989 – 519 с. **12.** Хьюбер П. Робастность в статистике. / П. Хьюбер. Пер. с англ. М. : 1984 – 304 с. **13.** Jeffreys H. The Law of Errors and the Combinations of Observations. // H. Jeffreys. London, Philos. Trans. Roy. Soc., ser. A. – 1937. – № 237. – P. 231–271. 1**4**. Jeffreys H. The Law of Errors in the Greenwich Variation of Latitude observation. // H. Jeffreys. Mon. Not. of the RAS, 1939, vol. 99. № 9. – P. 703–709. **15.** Jeffreys H. Theory of Probability / H. Jeffrey's Sec. Edition. – Oxford, 1940. – 468 p. **16.** Квіт І. Д. Статистична змінна. Ч.1 / І. Д. Квіт. Львів: Вища школи, 1974 – 120 с. **17.** Франс А. Книга Сюзанны / А. Франс – Полн. собр. соч. М. : Гостехиздат, 1957, т. 1, с. 551–610. **18.** Pearson K. On the Mathematical Theory of Errors of Judgment with special Reference to the personal Equation // K. Pearson. Philosophical Transactions of the Royal Society of London. Ser. A., 1902. Vol. 198. – P. 253–296. **19.** Джунь Й. В. Неокласична теорія помилок і її значення для створення нового покоління програмних продуктів в модулі «Data Analysis» // Й. В. Джунь. Психолого – педагогічні основи гуманізації навчально-виховного процесу в школі та ВНЗ. Зб. наук. праць. – Рівне : РВЦ МЕГУ, 2013. – № 2 (10) – С. 353–357. **20.** Колчинский И. Г. Наблюдение и факт в астрономии. / И. Г. Колчинский. – Киев : Наукова думка. – 104 с. **21.** Эльясберг П. Е. Измерительная информация: сколько её нужно? Как обрабатывать? / П. Е. Эльясберг. – М. : Наука. 1983. – 208 с. **22.** Орлов А. И. Высокие статистические технологии // А. И. Орлов. Заводская лаборатория, 2003. – Т. 69. – № 11. – С. 55–60.

Рецензент: д.т.н., профессор Власюк А. П.