К МЕТОДОЛОГИИ ПРОВЕРКИ АДЕКВАТНОСТИ ЦЕНТРАЛЬНОЙ ПРЕДЕЛЬНОЙ ТЕОРЕМЫ*

В.М. Резников

Предлагается методология проверки адекватности центральной предельной теореме на основе связи этой теоремы и теоремы Бернулли. Предложенная методология имеет преимущество перед прямой проверкой адекватности центральной предельной теоремы, так как верификация теоремы Бернулли требует неизмеримо меньше усилий, чем верификация любой предельной теоремы теории вероятностей. Однако центральная предельная теорема намного эффективнее теоремы Бернулли.

Ключевые слова: принцип Курно, теорема Бернулли, центральная предельная теорема, адекватность

Предельными теоремами теории вероятностей называются теоремы, в которых доказывается сходимость распределения суммы независимых или малозависимых случайных величин к некоторым законам распределения. При этом дисперсии (меры разброса) рассматриваемых случайных величин являются примерно одинаковыми. Часто в качестве предельного закона распределения рассматривается закон нормального, или гауссова, распределения. Название теоремы «центральная предельная» предложил Г. Полиа. По мнению многих исследователей, центральная предельная теорема чрезвычайно важна.

Во-первых, нормальное распределение часто является адекватным описанием множества естественно-научных процессов. Например, ошибки измерений часто описываются нормальным распределением. Ошибки измерений возникают под воздействием множества независимых и слабо связанных факторов, при этом вклад каждого фактора в ошибку измерений является ничтожным по сравнению с суммарной ошибкой. Нормальное распределение корректно описывает механизм возникновения ошибок измерений, оно адекватно для описаний многих

* Работа выполнена при финансовой поддержке РФФИ (проект № 11-07-00560а) и Интеграционного проекта СО РАН № 3.

характеристик человека, таких как рост, вес людей в популяции и др. Конечно, нормальное распределение не является универсальным, подходящим для описания любых процессов. Например, для маловероятных событий адекватным является распределение Пуассона, для многих биологических процессов, таких как изменение числа насекомых на листке растения или числа паразитов в органах хозяина, адекватно отрицательное биномиальное распределение. Как однажды заметил статистик Липман, интерес к нормальному распределению связан с тем, что математики думают, что нормальное распределение адекватно для применения во многих науках, а специалисты в конкретных науках широко применяют нормальное распределение, полагая, что его адекватность обоснована математически.

Во-вторых, на основе центральной предельной теоремы выводится теорема Бернулли. Она является следствием теоремы закона больших чисел, которая играет большую роль в становлении и развитии теории вероятностей. Кроме того, получение достоверных результатов на основе центральной предельной теоремы требует существенно меньше данных, чем на их получение на основе теоремы Бернулли. В ряде работ показаны большие трудности корректного применения теоремы Бернулли в практике научных исследований [1].

Настоящая работа посвящена методологическим проблемам применения центральной предельной теоремы, так как в известной литературе этим проблемам уделено немного внимания [2]. Предлагаемая методология основана на связи теоремы Бернулли и центральной предельной теоремы.

Другое название центральной предельной теоремы – «интегральная теорема Муавра – Лапласа». Эта теорема доказывается на основе локальной предельной теоремы Муавра. В теореме Муавра суммы биномиально распределенных случайных величин сходятся не к закону нормального распределения, а к плотности нормального распределения. Поэтому формально теорема Муавра не относится к центральным предельным теоремам. Однако эта теорема играет большую роль в вычислительной статистике, обеспечивая вычисление суммы биномиальных вероятностей путем ее аппроксимации с помощью плотности гауссова распределения. Кроме того, на основе рассуждений А.Н. Ширяева ей можно придать ясную вероятностную интерпретацию [3].

Для суммы Sn биномиально распределенных величин случайных величин ξ_i, $I = 1, n$, (ξ_i принимают два значения, например 1 и 0 с вероятностями p и q) имеет место следующее соотношение:

$$M(Sn/n) = p. \tag{1}$$

Здесь M символ математического ожидания. Из соотношения (1) следует некоторая близость величин Sn/n и p, в явном виде эта близость определяется на основе теоремы Бернулли. Дисперсия $D(Sn/n)$ величины Sn/n определяется следующим образом:

$$D(Sn/n) = M(Sn/n - p)^2 = pq/n. \tag{2}$$

Поэтому по аналогии с основаниями утверждать о близости величин Sn/n и p имеются основания полагать некоторую близость величин $(Sn/n - p)^2$ и pq/n, а также близость величин $Sn/n - p$ и $(pq/n)^{1/2}$. Символически эквивалентность этих величин записывается так:

$$Sn/n - p \sim \sqrt{pq/n}\,. \tag{3}$$

Эквивалентности этих величин придается точный смысл в локальной теореме Муавра. Близость величин (3) определяется с помощью вычисления вероятности следующего вида:

$$P\{|\ Sn - MSn / \sqrt{DSn}\ | \le x\} \tag{4}$$

Пусть $Pn(k)$ – это вероятность реализации k успехов в n испытаниях Бернулли:

$$Pn(k) = C_n^k p^k q^{n-k}. \tag{5}$$

Теорема Муавра обеспечивает асимптотическое вычисление вероятностей в выражении (5).

Локальная теорема Муавра. Если вероятность наступления некоторого события A в n независимых испытаниях постоянна и равна $p(O < p < 1)$, $q = 1 - p$, то вероятность $Pn(m)$ того, что в этих n испытаниях событие A наступит ровно m раз, удовлетворяет при $n \to \infty$ соотношению

$$\sqrt{npq}\, Pn(m) / (1 / \sqrt{2\pi} \times e^{-1/2x^2}) \to 1$$

равномерно для всех m, для которых

$$x = x_{mn} = m - np / \sqrt{npq}$$

находится в каком-либо конечном интервале [4].

Теорема Муавра доказывается путем вычислений на основе формулы Стирлинга.

На основании локальной теоремы доказывается интегральная теорема Муавра – Лапласа [5].

Теорема. Пусть $0 < P < 1$, $Pn(k) = C_n^k p^k q^{n-k}$,

$$P_n(a, b] = \sum_{a < x \le b} p_n(np + x\sqrt{npq}).$$

Здесь $np + x\sqrt{npq}$ целое число. Тогда

$$\sup_{-\infty \le a < b \le \infty} | P_n(a, b] - 1 / \sqrt{2\pi} \int_a^b e^{-x^2/2} dx | \to 0, n \to \infty \cdot \qquad (6)$$

Для прикладных задач окончательный результат формулируется следующим образом.

Для любых A, B, таких, что $\infty- \le A < B \le \infty$ и $n \to \propto$ имеет место

$$P\{A < S_n \le B\} - [\Phi(B - np / \sqrt{npq}) - \Phi(A - np / \sqrt{npq})] \to 0, n \to \infty \qquad (7)$$

Здесь $\Phi(x) = 1 / \sqrt{2\pi} \times \int_{-\infty}^{x} e^{-t^2/2} dt$ (8)

Интегральная теорема Лапласа в прикладной форме [6]. Если μ есть число наступлений изучаемого события A в n независимых испытаниях, причем вероятность A, $0 < P(A) < 1$, то равномерно относительно a и b, $(-\infty \le a \le b \le \infty)$ при $n \to \infty$ имеет место следующее соотношение:

$$P\left\{a \le \mu - np \Big/ \sqrt{npq} < b\right\} \to 1 \Big/ (\sqrt{2\pi}) \int_a^b \mathrm{e}^{-z^2/2} dz.$$

Интерпретация интегральной теоремы становится более ясной, если ее переформулировать с помощью популярных в стохастической математике сумм нормированных случайных величин.

Пусть μ_k – это число появлений событий A в k-м испытании, тогда $\mu = \Sigma\,\mu_k$, $M\Sigma\,\mu_k = np$, $D(\Sigma\,\mu_k) = \sqrt{npq}$.

Выражение $(\Sigma\,\mu_k \; - M\Sigma\,\mu_k)/\,D(\Sigma\,\mu_k)$ называется суммой нормированных случайных величин.

Тогда теорема Лапласа говорит о вероятности того, что сумма нормированных случайных величин находится в пределах от a до b и будет равномерно сходиться к интегралу:

$$1\Big/(\sqrt{2\pi})\int_a^b e^{-z^2/2}dz\ .$$

В контексте проблемы корректного применения центральной предельной теоремы представляет интерес как анализ требований к случайным величинам, на основе которых доказывается теорема, так и оценка трудоемкости этих требований в применении теоремы. В теореме использовались одинаково распределенные независимые случайные величины. Насколько требование независимости и одинаковой распределенности важно для центральной предельной теоремы? В действительности важны оба требования. Так, требование одинаковой распределенности случайных величин обеспечивает одинаковый вклад этих величин в общую дисперсию. Если влияние на суммарную дисперсию некоторых случайных величин будет существенно больше влияния остальных случайных величин, то результирующее распределение будет близким к распределению этих доминирующих величин, а их распределение может быть отличным от нормального распределения.

Существуют различные формализации требования о примерно одинаковом вкладе в итоговое распределение всех слагаемых. Это обеспечивается одинаковой распределенностью всех случайных величин с конечными средними и дисперсиями, причем дисперсии не должны равняться нулю. Одним из условий одинакового вклада случайных величин является то, что в пределе дисперсия суммы случайных величин равняется бесконечности [7]:

$$\lim_{k\to\infty} \mathrm{D}\Sigma\,\mu_k = \infty\ . \qquad (9)$$

Почему бесконечная сумма неотрицательных слагаемых не может быть конечным числом? Если эта сумма равна конечному числу, то подавляющее число слагаемых будут равны нулю. И относительно небольшое число случайных величин определят результирующее распределение: если эти доминирующие случайные величины будут негауссовыми, то и результирующее распределение не будет гауссовым.

Теперь рассмотрим значимость требования независимости случайных величин в центральной предельной теореме. Предположим, что нами нарушается требование независимости и рассматривается сумма n одинаковых случайных величин X_i, $I = 1, n$, $Y = \sum X_i = nX$, где $X = Xi$, $I = 1, n$. В этом случае требование одинакового вклада каждой случайной величины в формирование итогового распределения выполняется. Однако требование независимости будет нарушено, так как в теории вероятностей одинаковые события являются сильно зависимыми. Например, вероятность $P(X/X) = 1$. В рассматриваемом примере распределение Y определяется распределением величины X, последнее в общем случае не является нормальным, поэтому в данном случае центральная предельная теорема неверна. Полученный результат относительно значимости независимости не является неожиданным, так как все теоремы закона больших чисел и другие предельные теоремы верны для независимых и малозависимых случайных величин.

В ряде работ были исследованы проблемы применения теоремы закона больших чисел, в частотности теоремы Бернулли, в практике научных исследований. Оказалось, что корректному применению теоремы Бернулли препятствуют различные трудности. Во-первых, адекватность данных теореме Бернулли базируется на принципе Курно, который не является обоснованным. В силу этого принципа появление маловероятных событий свидетельствует о некорректности проверяемой гипотезы [8]. Однако существует определенные структуры данных, на которых появление маловероятных событий полностью обосновывает проверяемую гипотезу [9]. Во-вторых, последовательная проверка адекватности данных теореме Бернулли приводит к логическому зацикливанию [10]. В-третьих, проверка независимости для больших массивов данных требует огромного объема вычислительной работы. Известно, что центральная предельная теорема является более эффективной, чем теорема Бернулли. Так, на известном примере А.Н. Колмогоровым было показано, что для достижения одной и той же точности вычислений центральная предельная теорема требует в 5 раз меньше данных, чем теорема Бернулли [11]. Поэтому анализ трудностей применения центральной предельной теоремы является актуальной задачей.

Для адекватности данных центральной предельной теореме необходимо проверить условия теоремы на этих данных. Во-первых, в самом простом случае надо проверить, что, вообще говоря, бесконечное множество случайных величин имеют одно и то же определенное распределение с различными параметрами распределения. В случае же теоремы Бернулли имеет место проблема верификации вероятности единственного события с помощью частот. В простейшем случае случайная величина – это переменная, принимающая множество значений с определенными вероятностями. Поэтому верификация адекватности данных одной случайной величине в центральной предельной теореме – более трудная задача, чем верификация единственной вероятности в теореме Бернулли. Во-вторых, как и для теоремы Бернулли, необходима проверка независимости или малой связанности случайных величин. Таким образом, проверка адекватности данных в центральной предельной теореме несоизмеримо сложнее верификации адекватности данных теореме Бернулли.

В связи со сложностями проверки условий центральной предельной теоремы возникает вопрос о редукции проверки адекватности данных теореме к проверке адекватности данных заключению этой теоремы. Действительно, одно дело – проверять адекватность данных бесконечному множеству распределений и совсем другое – проверить адекватность данных единственному нормальному распределению. Конечно, такая задача несоизмеримо проще. Хотя серьезная проверка адекватности данных даже единственному распределению – достаточно серьезная проблема. Однако даже качественное решение проблемы адекватности данных заключению центральной предельной теоремы не гарантирует адекватности данных условиям применения центральной предельной теоремы. Потому что существует множество способов и механизмов порождения нормального распределения, и невозможно обосновать, что в конкретном случае нормальное распределение было порождено с помощью конструкций центральной предельной теоремы.

Необходимы другие подходы к анализу условий применения центральной предельной теоремы. По нашему мнению, в этом плане интересны результаты, касающиеся связи условий применения теоремы Бернулли и центральной предельной теоремы. Впервые обнаружил связь между условиями применения этих теорем А.Я. Хинчин. Существует два основных результата относительно связи этих теорем. Первый результат относится к случайным величинам, которые центрированы медианами.

Определение 1. Случайная величина $\boldsymbol{X}$ центрирована медианой, если выполняется следующее соотношение: $\boldsymbol{P(X} < 0) \leq 1/2 \geq \boldsymbol{P(X} > 0)$.

Определение 2. Последовательность взаимно независимых случайных величин подчиняется *центральному предельному закону,* если существуют такие последовательности чисел $\{a_n\}$ и $\{b_n\}$ (где a_n – положительные числа, b_n – любые числа), что распределение случайной величины

$$\frac{1}{a_n}\sum_{k=1}^{n} x_k - b_n \tag{10}$$

стремится к нормальному закону, при этом отдельные слагаемые

$$\frac{x_k}{a_n} - \frac{b_n}{n}$$

равномерно бесконечно малы по вероятности, т.е. имеет место

$$\lim_{n\to\infty} P\left\{\frac{x_k}{a_n} - \frac{b_n}{n}\right\} = 0 \qquad \{1 \le k \le n\}.$$

Необходимо отметить, что центральный предельный закон является некоторой модификацией центральной предельной теоремы. В частности, в центральном предельном законе используются бесконечно малые по вероятности величины. Такого ограничения нет в центральной предельной теореме, введение этого ограничения оправдано адекватностью для приложений в статистической физике. В свою очередь, теорема закона больших чисел моделируется посредством так называемых устойчивых последовательностей.

Определение 3. Последовательность взаимно независимых неотрицательных случайных величин $\{y_k\}$ называется относительно устойчивой, если существует такая последовательность положительных чисел $\{An\}(n = 1, 2, ...)$, что имеет место следующее предельное соотношение:

$$\lim_{n\to\infty} P\,\frac{1}{A_n}\sum_{k=1}^{n} y_k = 1 \tag{11}$$

При этом, как и в определении центрального предельного закона, отдельные слагаемые равномерно бесконечно малы по вероятности, т.е. имеет место

$$\lim_{n\to\infty} P\left(\frac{y_k}{A_n}\right) = 0 \qquad \{1 \le k \le n\}.$$

Для центральной предельной теоремы, моделируемой посредством центрального предельного закона, и теоремы Бернулли, моделируемой посредством устойчивой последовательности, устанавливается связь с помощью следующей теоремы.

Теорема. Для того чтобы последовательность взаимно независимых случайных величин $\{x_k\}$ центрированных медианами, подчинялась центральному предельному закону, необходимо и достаточно, чтобы последовательность квадратов этих величин $\{x_k^2\}$ была относительно устойчива [12].

Для случая, когда переменные x_k обладают первым и вторым моментами, осуществляется центрирование x_k не с помощью медианы, а посредством среднего. В этом случае коэффициентам a_n приписывают значение B_n:

$$a_n = B_n = \sqrt{\sum_{k=1}^{n} D(\mathrm{x}_k)} = \sqrt{\sum_{k=1}^{n} E(x_k^2)}.$$

При этих условиях определяют сходимость распределения

$$L\left\{ 1\big/B_n \sum_{x=1}^{k} x_k \right\} \to G(x);\, n \to \infty. \tag{12}$$

Здесь $L(x)$ – закон распределения случайной величины X, $G(x)$ – обозначение гауссова распределения. В предельном выражении (11) отдельные слагаемые бесконечно малы по вероятности:

$$\lim_{n\to\infty} P\left(\frac{x_k}{B_n}\right) = 0 \quad \{1 \le k \le n\}.$$

Для рассматриваемого случая не только изменяется формулировка центрального предельного закона, но и по-другому формулируется относительная предельная последовательность. В выражении (11) A_n получает значение

$$A_n = \sum_{k=1}^{n} E(y_k).$$

В таком случае центральный предельный закон формулируется следующим образом:

$$\lim_{n\to\infty} P \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} E(y_k)} = 1.$$

При этом отдельные слагаемые имеют бесконечно малые вероятности:

$$\lim_{n\to\infty} P \frac{y_k}{\sum_{k=1}^{n} E(y_k)} = 0 \qquad \{1 \le k \le n\}. \tag{13}$$

Теорема, связывающая условия применимости теоремы центрального закона и относительной устойчивости последовательностей, имеет следующий вид.

Теорема. Пусть случайные величины $x_к$ (k = 1, 2, ...) имеют конечные моменты второго порядка $E(x_k^2)$ и математические ожидания $E(x_k)$ = 0. Для того чтобы последовательность $\{x_k)$ подчинялась центральному предельному закону, необходимо и достаточно, чтобы последовательность квадратов $\{y_k = x_k^2\}$ подчинялась закону больших чисел [13].

Тема связи условий применимости теорем, относящихся к теореме закона больших чисел, и центральной предельной теоремы является актуальной. Современные результаты, полученные в этой области, представлены в ряде публикаций [14]. Однако в отличие от нашей работы в них не рассматривается проблема применения этих теорем в практике научных исследований. Мы ранее показали чрезвычайную сложность корректного применения теоремы Бернулли в научном исследовании. Трудности применения остальных предельных теорем превышают трудности применения теоремы Бернулли. Поэтому результаты, касающиеся связи центральной предельной теоремы и теоремы Бернулли, скорее имеют теоретический характер. Однако эти результаты могут иметь и определенное практическое значение, так как трудности применения теоремы Бернулли являются минимальными по сравнению с применением всех других известных предельных теорем.

Примечания

1. См.: *Колмогоров А.Н.* Теория вероятностей // Математика, ее содержание, методы и значение. – М.: Изд-во АН СССР. – 1956. – Т. 2; *Алимов Ю.И.* Альтернатива методу математической статистики. – М.: Знание, 1980; *Резников В.М.* Философский и методологический анализ адекватности объективистских стохастических концепций. – Новосибирск, РИЦ Новосиб. гос. ун-та, 2011.
2. Там же.
3. См.: *Ширяев А.Н.* Вероятность. – М.: Наука, 1989.
4. См.: *Гнеденко Б.В.* Курс теории вероятностей. – М.: Наука, 1988.
5. См.: *Ширяев А.Н.* Вероятность.
6. Там же.
7. См.: *Алимов Ю.И.* Элементы теории эксперимента: Опытная проверка утверждений математической статистики. – Свердловск: Урал. пед. ин-т, 1978.
8. См.: *Резников В.М.* Методологические проблемы применения статистических критериев // Вест. Новосиб. гос. ун-та. Сер. Философия. – 2009. – Т. 7, вып. 3. – С. 18–23.
9. См.: *Cohen J.* The earth is round ($p<0.05$) // American Psychologist. – 1994. – V. 49, No. 12. – P. 997–1003.
10. См.: *Алимов Ю.И.* Альтернатива методу математической статистики.
11. См.: *Колмогоров А.Н.* Теория вероятностей.
12. См.: *Райков Д.А.* О связи между центральным предельным законом теории вероятностей и законом больших чисел // Изв. АН СССР. Отд. мат. и естеств. наук, 1938. – С. 324–338.
13. Там же.
14. См., например: *Gut A.* Gnedenko-Raikov's theorem, central limit theory, and the weak law of large numbers // Statist. Probab. Lett – 2006. – No. 76. – С. 1935–1939.

Дата поступления 25.06.2012

Институт философии и права
СО РАН, г. Новосибирск
mathphil1976@gmail.com

***Reznikov, V.M.* On the methodology of verification of adequacy of the central limit theorem**

The paper presents the methodology of verification of adequacy of the central limit theorem based on the connection of the latter with Bernoulli's theorem. The proposed methodology has an advantage over direct verification of adequacy of the central limit theorem because verification of adequacy of Bernoulli's theorem requires considerably less efforts than verification of any limit theorem in probability theory. However, the central limit theorem is much more effective than Bernoulli's theorem.

Keywords: Cournot's principle, Bernoulli's theorem, central limit theorem, adequacy