ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Научная статья

УДК 519.21
DOI: 10.18101/2304-5728-2022-1-26-34

ОЦЕНКА ЧИСЛА СЛАГАЕМЫХ НОРМАЛЬНОЙ АППРОКСИМАЦИИ СУММ НЕЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

© Ганичева Антонина Валериановна
кандидат физико-математических наук, доцент,
Тверская государственная сельскохозяйственная академия
Россия, 170904, г. Тверь, ул. Маршала Василевского (Сахарово), 7
TGAN55@yandex.ru

Аннотация. В работе решается задача определения количества независимых случайных величин с одинаковыми математическими ожиданиями и разными дисперсиями, сумма которых с заданной точностью имеет нормальный закон распределения. Аналогичная задача рассмотрена для средней арифметической выборки из нормального распределения вероятностей. Доказана теорема и получено следствие из нее. Доказательство теоремы основано на разложении характеристических функций в ряд Маклорена. По зависимостям, полученным в теореме, рассчитаны таблицы для определения необходимого числа слагаемых при заданной точности для разных средних квадратических отклонений выборочных наблюдений. Построены графики полученных зависимостей. Зависимость потребного числа слагаемых от точности аппроксимирована полиномом шестой степени. Доказанная в статье теорема и полученные зависимости могут быть использованы в системах тестирования, контроля, наблюдений и диагностики.

Ключевые слова: центральная предельная теорема, нормальное распределение, средняя выборочная, дисперсия, рекуррентный метод, характеристическая функция, ряд Маклорена, точность, относительная погрешность.

Для цитирования
Ганичева А. В. Оценка числа слагаемых нормальной аппроксимации сумм независимых случайных величин // Вестник Бурятского государственного университета. Математика, информатика. 2022. № 1. С. 26–34.

Введение

Проблема оценки числа слагаемых, при котором сумма независимых случайных величин имеет нормальный закон распределения вероятностей, связана с центральной предельной теоремой (ЦПТ).

ЦПТ имеет особое место среди практических применений теории вероятностей. Она позволяет решать многие задачи в различных областях, где рассматриваются суммы случайных величин. Проблема распределения суммы случайных величин по нормальному закону решается в статьях [1; 2; 3] методом моделирования. В статье [1] получено, что при 30 и более слагаемых график плотности распределения близок к нормальному. В работе [2] указывается, что для нормально распределенной генеральной совокупности ЦПТ справедлива даже для выборок с объемом менее 30. Размер выборки 200 слагаемых использовался в работе [3].

Данное исследование является дальнейшим развитием метода из [4] определения числа слагаемых ЦПТ. Новым является распространение этого метода на случай разных значений средних квадратических отклонений [4].

Целью статьи является определение числа слагаемых ЦПТ для получения нормального закона распределения вероятностей с заданной точностью. Для этого нужно получить число n в виде его функциональной зависимости от параметров закона распределения суммы или средней выборочной.

1 Основная теорема и следствие

Сформулируем и докажем следующую теорему.

Теорема. Средняя выборочная ($\overline{x}$) n независимых случайных величин с одинаковыми математическими ожиданиями m и разными дисперсиями, ограниченными сверху значением $\sigma^2_{\max}$, будет иметь нормальный закон распределения (с точностью ε) при

$$n \ge n_0 = \max\left\{ \left\lfloor \sqrt{4,5}\sigma_{\max} \right\rfloor + 2, \left\lfloor \frac{4,5\sigma^2_{\max}}{\sqrt{\varepsilon}} \right\rfloor + 1, \left\lfloor 0,41 + \ln \sigma_{\max} \right\rfloor \right\}.$$

Здесь символом $\lfloor\ \rfloor$ обозначена целая часть числа.

Доказательство.

Определим число n_0 методом характеристических функций. Рассмотрим непрерывные случайные величины.

Пусть $q_{x_j}\left(\frac{1}{n_0}t\right) = \int\limits_{-\infty}^{\infty} e^{\frac{itx_j}{n_0}} \cdot f\left(x_j\right)dx_j$ — характеристическая функция случайной величины $\frac{X_j}{n_0}\left(j = \overline{1, n_0}\right)$. Тогда

$$q_{\overline{x}_j}(t) = \prod_{j=1}^{n_0} q_{x_j}\left(\frac{1}{n_0}t\right) = \prod_{j=1}^{n_0} \int\limits_{-\infty}^{\infty} e^{\frac{itx_j}{n_0}} \cdot f\left(x_j\right)dx_j \qquad (1)$$

характеристическая функция величины $\overline{x}$, т. к. X_j независимы.

Не нарушая общности, считаем, что $m = 0$. Разложим $q_{x_j}\left(\frac{1}{n_0}t\right)$ в ряд Маклорена. Получим: $q_{x_j}(0) = 1;\ q'_{x_j}(0) = \frac{i}{n_0}\int\limits_{-\infty}^{\infty} x_j f\left(x_j\right)dx_j = \frac{i}{n_0}m = 0.$

Вторая производная будет равна:

$$q''_{x_j}(0)=\frac{i^2}{n_0^2}\int_{-\infty}^{\infty}x_j^2 f(x_j)\,dx_j=-\frac{1}{n_0^2}\sigma_j^2.$$

Остаточная сумма ряда $S_{ост}$ определяется выражением:

$$S_{ост}(t)=q_{x_j}(t)-1+\frac{1}{2n_0^2}\cdot t^2.$$

С учетом того, что $q_{x_j}(0)=1$, она будет равна: $S_{ост}(0)=0.$

Тогда

$$q_{x_j}\left(\frac{1}{n_0}t\right)=1-\frac{1}{2n_0^2}\cdot t^2\sigma_j^2,$$

$$q_{\bar{x}}(t)=\prod_{j-1}^{n_0}\left(1-\frac{1}{2n_0^2}\cdot t^2\sigma_j^2\right). \qquad (2)$$

Пусть выполняются следующие условия:

$-3\cdot\sigma_{\max}\le t\le 3\cdot\sigma_{\max}$, при $t^2=9\cdot\sigma_{\max}^2$ $9\cdot\sigma_{\max}^2<2n_0^2$ и скобки в (2) положительные. Заметим, что условие положительности скобок в (2) вытекает из условия $n_0\ge\sqrt{4{,}5}\sigma_{\max}$. Поэтому

$$n_0\ge\sqrt{4{,}5}\cdot\sigma_{\max}. \qquad (3)$$

Для дальнейших выводов будем считать, что

$$n_0\ge\max\left\lfloor\sqrt{4{,}5}\cdot\sigma_{\max}\right\rfloor+2. \qquad (4)$$

Найдем логарифм от (2):

$$\ln q_{\bar{x}}(t)=\sum_{j=1}^{n_0}\ell n\left(1-\frac{1}{2n_0^2}\cdot t^2\cdot\sigma_j^2\right).$$

Ряд Маклорена для данного выражения будет:

$$\ln q_{\bar{x}}(t)=\sum_{j=1}^{n_0}\left(-\frac{1}{2n_0^2}\cdot t^2\cdot\sigma_j^2-\frac{1}{4n_0^4\cdot 2}\cdot t^4\cdot\sigma_j^4-\ldots-\frac{1}{2^k n_0^{2k}k}t^{2k}\cdot\sigma_j^{2k}-\ldots\right).$$

Имеем сумму n_0 рядов. Погрешность (остаточный член) каждого ряда будем оценивать в форме Коши, т. е.

$$r(x)=\frac{x^{k+1}}{k+1}\frac{(1-\xi)^k}{(1-\xi x)^{k+1}},\ 0<\xi<1.$$

В знаменателе стоит разность, т. к. рассматриваем $\ln(1-x)$.

Для каждого ряда суммы имеем: $x=\dfrac{t^2\sigma_j^2}{2n_0^2}$, $0<x<1$;

$$\left(\frac{1-\xi}{1-\xi x}\right)^k<1,\ \text{т. к.}\ \xi x<\xi;\ \ \frac{1}{1-\xi x}<n_0,\ \text{т. к.}\ n_0>2\ \text{и}\ 1>\xi x.$$

Поэтому $r(x) \le \frac{x^{k+1}}{k+1} \cdot n_0$, т. е.

$$R_{\bar{x}}\left(t\right) \le \sum_{j=1}^{n_0} \frac{t^{2k+2} \cdot \sigma_j^{2k+2}}{2^{k+1} \cdot n_0^{2k+1} \cdot \left(k+1\right)}.$$

Потребуем в записанном неравенстве для $k = 1$ выполнения условия:

$$\sum_{j=1}^{n_0} \frac{t^4 \cdot \sigma_j^4}{8n_0^3} \le \frac{\varepsilon}{2}, \tag{5}$$

где ε — сколь угодно малое положительное число.

Так как $t \le 3 \cdot \sigma_{\max}$, то, усилив неравенство (5), получим: $\frac{3^4 \cdot \sigma_{\max}^8}{8n_0^2} \le \frac{\varepsilon}{2}$.

Отсюда:

$$n_0 \ge \left\lfloor \frac{4,5 \cdot \sigma_{\max}^4}{\sqrt{\varepsilon}} \right\rfloor + 1. \tag{6}$$

Можно было бы, заменив t на $3\sigma_{\max}$ в (5), использовать рекуррентный метод. Полагаем сначала $n_0 = 2$ и проверяем выполнение неравенства (5) для двух законов распределения при заданных средних квадратических отклонениях σ_1 и σ_2.Если оно выполняется, то завершаем рекуррентный процесс. В противном случае полагаем $n_0 = n_0 + 1$ и снова проверяем выполнение неравенства (5) для трех законов распределения при заданных σ_1, σ_2 и σ_3 и т. д.

Из (4) и (6) находим

$$n \ge n_0 = \max\left\{ \left\lfloor \sqrt{4,5} \cdot \sigma_{\max} \right\rfloor + 2, \left\lfloor \frac{4,5 \cdot \sigma_{\max}^4}{\sqrt{\varepsilon}} \right\rfloor + 1 \right\}. \tag{7}$$

Таким образом, $\ln q_{\bar{x}}\left(t\right) = -\sum_{j=1}^{n_0} \frac{1}{2n_0^2} \cdot t^2 \cdot \sigma_j^2$ с точностью $\frac{\varepsilon}{2}$.

Положим

$$\sigma = \frac{1}{n_0} \sqrt{\sum_{j=1}^{n_0} \sigma_j^2}.$$

Тогда

$$-\frac{\sigma^2 t^2}{2} - \frac{\varepsilon}{2} = \sum_{j=1}^{n_0} \frac{1}{2n_0^2} t^2 \sigma_j^2 - \frac{\varepsilon}{2} \le \ell n\, q_{\bar{x}}\left(t\right) \le -\sum_{j=1}^{n_0} \frac{1}{2n_0^2} t^2 \sigma_j^2 + \frac{\varepsilon}{2} = -\frac{1}{2}\sigma^2 t^2 + \frac{\varepsilon}{2}. \tag{8}$$

Получаем

$$q_{\bar{x}}\left(t\right) \approx e^{-\sum_{j=1}^{n_0} \frac{1}{2n_0^2} t^2 \cdot \sigma_j^2} = e^{-\frac{1}{2} t^2 \sigma^2} = e^{-\frac{1}{2/\sigma^2} t^2}. \tag{9}$$

Используя дифференциал, оценим погрешность определения функции $q_{\bar{x}}(t)$. А именно:

$$q_{\bar{x}}(t)-e^{-\sum_{j=1}^{n_0}\frac{1}{2n_0^2}t^2\cdot\sigma_j^2}\approx -e^{-\sum_{j=1}^{n_0}\frac{1}{2n_0^2}t^2\cdot\sigma_j^2}\cdot\frac{\varepsilon}{2}<\frac{\varepsilon}{2}. \qquad (10)$$

Очевидно, относительная погрешность также будет меньше ε.
Из определения характеристической функции имеем:

$$q_{\bar{x}}(t)=\int_{-\infty}^{\infty}e^{i\bar{x}t}\cdot f(\bar{x})d\bar{x}.$$

Тогда плотность распределения вероятностей

$$f(\bar{x})=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it\bar{x}}q_{\bar{x}}(t)dt.$$

С учетом формулы, приведенной в книге [5], $\int_{-\infty}^{\infty}e^{-Ax^2\pm 2Bx-C}dx=\sqrt{\frac{\pi}{A}}e^{-\frac{AC-B^2}{A}}$, формул (8)–(10) получаем

$$f(\bar{x})=\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{\bar{x}^2}{2\sigma^2}}\left(1+\frac{\varepsilon}{2}\right).$$

При $\sigma\geq 0{,}399$ $\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{\bar{x}^2}{2\sigma^2}}<1$. Поэтому с точностью $\frac{\varepsilon}{2}$

$$f(\bar{x})=\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{\bar{x}^2}{2\sigma^2}}. \qquad (11)$$

Очевидно, относительная погрешность не превосходит $\frac{\varepsilon}{2}$. Путем соответствующего масштабирования ограничения на σ можно снять.

Значит, $\bar{x}$ имеет нормальное распределение с $m_{\bar{x}}=0$, $\sigma_{\bar{x}}=\sigma$.

Заметим, что при вычислении $f(\bar{x})$ используется несобственный интеграл при $t\in(-\infty,\infty)$. А именно:

$$f(\bar{x})=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it\bar{x}}\cdot e^{-\frac{1}{2}t^2\sigma^2}dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-\left(\frac{1}{2}t^2\sigma^2+it\bar{x}\right)}dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-\frac{1}{2}\sigma^2\left(t+\frac{i\bar{x}}{\sigma^2}\right)^2}dt=$$

$$=\frac{1}{2\pi}e^{-\frac{1}{2}\frac{\bar{x}^2}{\sigma^2}}\cdot\frac{\sqrt{2\pi}}{\sigma}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{\bar{x}^2}{\sigma^2}}. \qquad (12)$$

Однако $t\in[-3\sigma_{\max},3\sigma_{\max}]$. Покажем, что с любой заданной точностью $\frac{\varepsilon}{4}$ при соответствующем n_0 интеграл (12) в пределах от $3\sigma_{\max}$ до ∞ будет сколь угодно мало отличаться от $\frac{\varepsilon}{4}$. Положим $u=\frac{\sigma}{\sqrt{2}}\left(t+\frac{i\overline{x}}{\sigma^2}\right)$, тогда $du=\frac{\sigma}{\sqrt{2}}dt$. Пусть $b_1=\frac{\sigma}{\sqrt{2}}\left(3\sigma_{\max}+\frac{i\overline{x}}{\sigma^2}\right)$.

Имеем: $\int\limits_{3\sigma_{\max}}^{\infty} e^{-\frac{1}{2}\sigma^2\left(t+\frac{i\overline{x}}{\sigma^2}\right)^2}dt=\frac{\sigma}{\sqrt{2}}\int\limits_{b_1}^{\infty}e^{-u^2}du.$

Для вычисления последнего интеграла возведем его в квадрат, затем перейдем к двойному интегралу в полярных координатах:

$\left(\int\limits_{b_1}^{\infty}e^{-u}du\right)^2=\int\limits_{b_1}^{\infty}e^{-u^2}du\cdot\int\limits_{b_1}^{\infty}e^{-v^2}dv=\int\limits_{0}^{\pi}d\varphi\cdot\int\limits_{|b_1|}^{\infty}re^{-r^2}dr$, где $r^2=u^2+v^2$, т. е.

$\frac{\sigma}{\sqrt{2}}\int\limits_{b_1}^{\infty}e^{-u^2}du=\frac{\sqrt{\pi}\sigma}{2}\cdot e^{-\frac{1}{2}|b_1|^2}\le\frac{\varepsilon}{4}$. Отсюда $|b_1|^2\ge-2\ln\frac{\varepsilon}{2\sqrt{\pi}\sigma}$.

Поскольку $\overline{x}\ge m-9\sigma_{\max}$ с вероятностью 0,99, то, усилив последнее неравенство, получим:

$$4,5\sigma_{\min}^2\cdot\sigma_{\max}^2+45,5n_0^2+2n_0\ln\frac{\varepsilon\sqrt{n_0}}{2\sqrt{\pi}\sigma_{\max}}\ge 0,$$

где $\sigma_{\min}^2=\min\sigma_j,\ j=\overline{1,n_0}$, т. е.

$$22,75n_0^2+n_0(\ln\varepsilon+\frac{1}{2}\ln n_0-1,264-\ln\sigma_{\max})+2,25\sigma_{\min}^2\cdot\sigma_{\max}^2\ge 0. \qquad (13)$$

Из этого неравенства при заданных ε, $\sigma_{\min}$ и $\sigma_{\max}$ находится соответствующее значение n_0.

Поскольку $n_0>\frac{1}{2}\ln n_0$, то последнее неравенство можно усилить так:

$$21,75n_0>-\ln\varepsilon+1,264-\ln\sigma_{\max}\text{, т. е.}$$

$$n_0>\frac{7+1,264+\ln\sigma_{\max}}{21,75},$$

т. к. минимальное ε будем рассматривать равным 0,001. Отсюда

$$n_0>0,41+\ln\sigma_{\max}. \qquad (14)$$

Это более простая, но более грубая оценка, чем получаемая из неравенства (13).

Аналогично рассматривается интеграл от $-\infty$ до $-3\sigma_{\max}$.

Следовательно, $f(\overline{x})=\int\limits_{-3\sigma_{\max}}^{3\sigma_{\max}}e^{-it\overline{x}}\cdot q_{\overline{x}}(t)dt$ с точностью $\frac{\varepsilon}{2}$ при n_0, удовлетворяющем (14).

Таким образом, с точностью ε средняя арифметическая количества $n \geq n_0$ суммируемых независимых случайных величин с одинаковыми математическими ожиданиями и разными дисперсиями при выполнении условия

$$n_0 = \max\left\{\left\lfloor \sqrt{4{,}5}\sigma_{\max} \right\rfloor + 2, \left\lfloor \frac{4{,}5\sigma_{\max}^4}{\sqrt{\varepsilon}} \right\rfloor + 1, \left\lfloor \ln \sigma_{\max} + 0{,}41 \right\rfloor\right\} \quad (15)$$

имеет нормальный закон распределения.

Теорема доказана.

2 Исследование зависимости числа слагаемых ЦПТ от точности

В таблице 1 для заданного значения ε приведены соответствующие значения n_0 для $\sigma_{\max} = 1$.

Зависимость n_0 от ε — Таблица 1

ε	0,001	0,002	0,0028	0,003	0,004	0,005	0,01	0,1	0,5	0,9
n_0	102	72	61	59	51	46	33	11	6	6

Зависимость числа слагаемых n_0 от точности ε графически изображена на рис. 1.

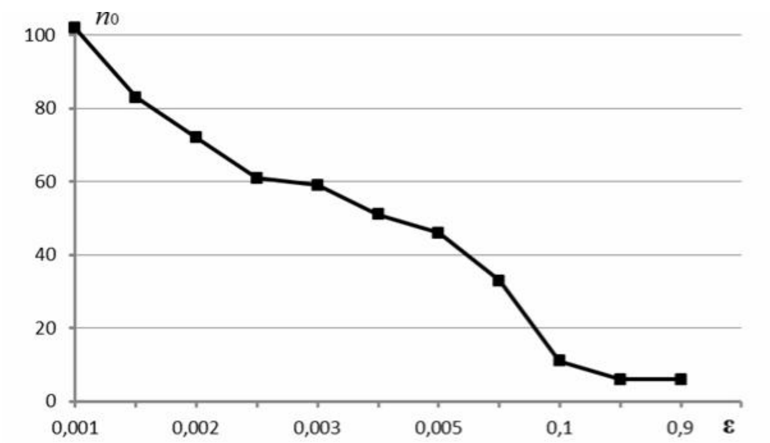

Рис. 1. Зависимость n_0 от ε

С помощью средства MS Excel «Линия тренда» эту зависимость можно аппроксимировать полиномом шестой степени (коэффициент детерминации $R^2 = 0{,}9959$):

$$n_0 = \left\lfloor -0{,}0013\varepsilon^6 + 0{,}0587\varepsilon^5 - 0{,}9298\varepsilon^4 + 6{,}4444\varepsilon^3 - 18{,}656\varepsilon^2 + 5{,}5639\varepsilon + 109{,}25 \right\rfloor.$$

Следствие из теоремы. При числе слагаемых $n \geq n_0$ (n_0 удовлетворяет равенству (14)) сумма n независимых случайных величин (с одинаковыми математическими ожиданиями и дисперсиями, ограниченными сверху значением $\sigma_{\max}$) имеет нормальный закон распределения вероятностей.

При значениях $n \geq n_0$ величина $\overline{x}$ имеет нормальное распределение, поэтому $Y = n\overline{X}$ будет иметь нормальное распределение (при тех же значениях n).

Заключение

На основе доказанной теоремы и ее следствия выведены условия, определяющие количество слагаемых, для которого средняя выборочная и сумма слагаемых (в ЦПТ) имеют нормальный закон распределения (с заданной точностью). Исследована зависимость числа слагаемых ЦПТ от заданной точности и дисперсии выборки.

Литература

1. Пименов С. Ю., Тинаев В. В. Применение центральной предельной теоремы для компьютерного моделирования случайных сигналов // Наука и образование: новое время. 2017. Т. 19, № 2. С. 227–231. Текст: непосредственный.

2. Цурганов А. Г., Макеенко Г. И. Простая иллюстрация центральной предельной теоремы в медицинской статистике // Достижения фундаментальной, клинической медицины и фармации: материалы 71-й научной сессии сотрудников университета. Витебск: ВГМУ, 2016. С. 330–331. Текст: непосредственный.

3. Парахин А. С. Численная проверка центральной предельной теоремы // Математика, информатика, компетентностный подход к обучению в вузе и школе: материалы Всероссийской научно-практической конференции. Курган: Изд-во КГУ, 2015. С. 24–27. Текст: непосредственный.

4. Ганичева А. В. Оценка числа слагаемых центральной предельной теоремы // Прикладная математика и вопросы управления. 2020. № 4. С. 7–19. Текст: непосредственный.

5. Вентцель Е. С. Теория вероятностей. Москва: Изд-во физ.-мат. лит-ры, 2003. 564 с. Текст: непосредственный.

Статья поступила в редакцию 25.02.2022; одобрена после рецензирования 04.03.2022; принята к публикации 15.03.2022.

ESTIMATION OF THE NUMBER OF TERMS OF THE NORMAL APPROXIMATION OF THE SUMS OF INDEPENDENT RANDOM VARIABLES

Antonina V. Ganicheva
Cand. Sci. (Phys. and Math.), A/Prof,
Tver State Agricultural Academy
7 Marshal Vasilevsky Sr. (Sakharovo), Tver, Russia
TGAN55@yandex.ru

Abstract. The paper solves the problem of determining the number of independent random variables with the same mathematical expectations and different variances, the sum of which has a normal distribution law with a given accuracy. A similar problem is considered for an arithmetic mean sample from a normal probability distribution. The theorem is proved and the corollary from it is obtained. The proof of the theorem is based on the decomposition of characteristic functions into a Maclaurin series. Based on the dependencies obtained in the theorem, tables are calculated to determine the required number of terms for a given accuracy for different mean square deviations of sample observations. Graphs of the obtained dependencies are constructed. The dependence of the required number of terms on the accuracy is approximated by a polynomial of the sixth degree. The theorem proved in the article and the obtained dependencies can be used in testing, monitoring, observation and diagnostics systems.

Keywords: central limit theorem, Rayleigh distribution, sample mean, variance, recurrent method, characteristic function, Maclaurin series, accuracy, relative error.

For citation

Ganicheva A. V. Estimation of the Number of Terms of the Normal Approximation of the Sums of Independent Random Variables // Bulletin of Buryat State University. Mathematics, Informatics. 2022. N. 1. Pp. 26–34.

The article was submitted 25.02.2022; approved after reviewing 04.03.2022; accepted for publication 15.03.2022.