ОБЩЕТЕХНИЧЕСКИЕ ЗАДАЧИ И ПУТИ ИХ РЕШЕНИЯ

УДК 378:519.2

Центральная предельная теорема в задачах прогнозирования неуспеваемости студентов вузов

Р. С. Кударов, Р. С. Кударов

Петербургский государственный университет путей сообщения Императора Александра I, Российская Федерация, 190031, Санкт-Петербург, Московский пр., 9

Для цитирования: *Кударов Р. С., Кударов Р. С.* Центральная предельная теорема в задачах прогнозирования неуспеваемости студентов вузов // Бюллетень результатов научных исследований. — 2025. — Вып. 2. — С. 172–189. DOI: 10.20295/2223-9987-2025-2-172-189

Аннотация

Цель: В государственной программе «Научно-технологическое развитие Российской Федерации» важное место отводится подготовке высококвалифицированных инженеров нового поколения, способных обеспечить стране достижение технологического суверенитета. Высшие учебные заведения страны преобразуются в передовые инженерные школы для выпуска специалистов, владеющих современными наукоемкими и мультидисциплинарными технологиями. Повышено внимание к дисциплинам первых лет обучения, без освоения которых дальнейшая учеба становится неполноценной. В настоящей статье приводится способ прогнозирования количества неуспевающих студентов, который призван оказать содействие в планировании мероприятий по обеспечению своевременного выполнения учебного плана. **Методы:** Прогнозирование количества неуспевающих студентов выполняется на основе центральной предельной теоремы. Применимость центральной предельной теоремы устанавливается по условию Ляпунова. Сходимость распределения количества неуспевающих студентов к закону Гаусса изучается с помощью неравенства Эссеена, при этом эмпирическая функция распределения моделируется методом Монте-Карло. **Результаты:** Построен доверительный интервал для оценки количества неуспевающих студентов при известных вероятностях неуспеваемости каждого студента. Введена поправка к надежности доверительного интервала на отклонение эмпирического распределения от закона Гаусса. **Практическая значимость:** Вычислен интервальный прогноз количества неуспевающих первокурсников на конец учебного года.

Ключевые слова: Центральная предельная теорема, закон Гаусса, метод Монте-Карло, Educational Data Mining, анализ образовательных данных, прогнозирование результатов обучения, прогнозирование неуспеваемости студентов.

Введение

Научная работа содержит результаты исследования, начатого введением случайной величины неуспеваемости отдельного студента и представляющего собой последовательное решение следующих задач: суммирование введенных случайных величин, запись предельной формы распределения суммы этих величин, оценка отклонения эмпирического распределения от его предельной формы,

построение доверительного интервала для оценки (т. е. прогнозируемого значения) количества неуспевающих студентов при известных вероятностях неуспеваемости каждого студента, введение поправки к надежности интервальной оценки на отклонение от предельного распределения. На основе реальных данных составляется интервальный прогноз количества неуспевающих студентов.

Публикуемые результаты направлены на достижение целей университета, относящихся к реализации федерального проекта «Передовые инженерные школы».

Центральная предельная теорема

Будем рассматривать неуспеваемость k -го студента как случайную величину:

$$X_k = \begin{cases} 1, \text{если студент имеет задолженности,} \\ 0, \text{если студент задолженностей не имеет.} \end{cases}$$

Тогда количество μ_n фактически неуспевающих студентов (среди общего числа n студентов) по итогам сессии представимо суммой:

$$\mu_n = X_1 + ... + X_k + ... + X_n,$$

в которой каждое слагаемое подчинено распределению Бернулли:

$$\begin{array}{c|c|c} x & 1 & 0 \\ \hline P(X_k = x) & p_k & q_k \end{array}, \quad (1)$$

где p_k — вероятность неуспеваемости k-го студента;

$q_k = 1 - p_k$ — вероятность успеваемости k-го студента.

Случайные величины, входящие в сумму μ_n, положим взаимно независимыми:

$X_1, \ldots, X_k, \ldots, X_n$ — взаимно независимы;

то есть примем тот факт, что по неуспеваемости отдельного студента (или группы студентов) нельзя сделать вывод о неуспеваемости остальных студентов.

Мы исключаем возможность того, что двух или более студентов все экзаменаторы сессии в равной мере аттестуют или не аттестуют. Например, если какой-либо студент успешно осваивает образовательную программу и в состоянии оказать академическую помощь одному или нескольким своим однокурсникам, то успешный результат по всем дисциплинам сессии никто гарантировать этим студентам не может. Обратный случай с побуждением сокурсников к уклонению от плана обучения упреждается постоянным мониторингом учебного процесса и воспитательной системой университета.

Также мы пренебрегаем маловероятными случаями, когда несколько студентов имеют общее, зависящее или независящее от чьей-либо воли, препятствие для

выполнения учебного плана в срок. Например, если в один из дней сессии произошел сбой в работе городского (или пригородного) транспорта, повлекший за собой неявку совокупности студентов на экзаменационное испытание, то у каждого студента этой совокупности без исключения возникнет академическая задолженность. Такие случаи возможны, но слишком редки, чтобы оказать существенное влияние на системный характер изучаемой в настоящей работе закономерности.

Возвращаясь к распределению Бернулли, приступим к суммированию случайных величин X_k. Если для отдельно взятого студента случайная величина $\mu_1 = X_1$ принимает одно из двух возможных значений (1 — «не успевает» или 0 — «успевает») и подчинена закону Бернулли, то для произвольной пары студентов сумма случайных величин $\mu_2 = X_1 + X_2$ имеет три возможных значения (2 — «не успевают оба», 1 — «не успевает один из двух», 0 — «успевают оба») и подчинена закону, отличному от распределения Бернулли:

x	2	1	0
$P(\mu_2 = x)$	$p_1 p_2$	$p_1 q_2 + q_1 p_2$	$q_1 q_2$

.

Распределение Бернулли не является устойчивым относительно суммирования. Однако при увеличении общего числа случайных величин X_k определенная закономерность все же формируется. При больших значениях n в распределении количества μ_n неуспевающих студентов особую роль приобретает закон Гаусса.

Придем к этому математически, используя результаты [1, с. 164–170].

Обозначив математическое ожидание и дисперсию суммы μ_n соответственно через

$$a_n = \sum_{k=1}^{n} M(X_k) = p_1 + ... + p_n, \; B_n = \sum_{k=1}^{n} D(X_k) = p_1 q_1 + ... + p_n q_n,$$

введем нормированную сумму

$$\zeta_n = \frac{\mu_n - a_n}{\sqrt{B_n}}.$$

Теперь убедимся, что выполняется *условие Ляпунова*:

$$\lim_{n\to\infty} L_d = \lim_{n\to\infty} \frac{C_n}{B_n^{1+d/2}} = 0,$$

где $C_n = c_1 + ... + c_n$, $c_k = M\left|X_k - M(X_k)\right|^{2+d}$, $d > 0$.

Из распределения (1) случайных величин X_k вычислим абсолютные центральные моменты порядка $2 + d$:

$$M\left|X_k - M(X_k)\right|^{2+d} = p_k^{2+d} q_k + p_k q_k^{2+d} = p_k q_k \underbrace{(p_k^{1+d} + q_k^{1+d})}_{\le 1} \le p_k q_k,$$

сумму абсолютных случайных моментов:

$$C_n = \sum_{k=1}^{n} p_k q_k (p_k^{1+d} + q_k^{1+d}) \le p_1 q_1 + ... + p_n q_n$$

и предел:

$$\lim_{n\to\infty} L_d \le \lim_{n\to\infty} \frac{p_1 q_1 + ... + p_n q_n}{(p_1 q_1 + ... + p_n q_n)^{1+d/2}} = \lim_{n\to\infty} \frac{1}{(p_1 q_1 + ... + p_n q_n)^{d/2}} = 0.$$

Неотрицательность отношения L_d дает $\lim_{n\to\infty} L_d = 0$.

Выполнимости *условия Ляпунова* достаточно для *выполнения условий Линдеберга*, которые являются необходимыми и достаточными для асимптотической нормальности суммы независимых разнораспределенных случайных величин:

$$F(x) \to G(x),$$

где $F(x)$ — функция распределения нормированной суммы ζ_n;

$G(x)$ — функция распределения стандартного закона Гаусса.

В силу *центральной предельной теоремы* теории вероятностей [2, с. 484] вероятность того, что количество фактически неуспевающих студентов μ_n будет удовлетворять неравенству

$$a_n + x_1\sqrt{B_n} < \mu_n < a_n + x_2\sqrt{B_n}, \quad (2)$$

стремится с ростом числа n студентов к определенному (от x_1 до x_2) интегралу от функции плотности вероятности стандартного закона Гаусса.

На практике предельное значение вероятности выполнения соотношения (2) или *уровень доверия* P находится с помощью функции $G(x)$ распределения стандартного закона Гаусса по формуле:

$$P = G(x_2) - G(x_1). \quad (3)$$

В анализе образовательных данных, при ограниченном числе студентов, отсутствие возможности наращивать n до приемлемой сходимости $F(x)$ к $G(x)$ порождает отклонения, при которых уравнение (3) будет нарушаться (например, вместо уровня доверия P будет иметь место $0{,}95P$ или $0{,}93P$).

В настоящей работе предполагается, что абсолютное отклонение Δ_n функции $F(x)$ от закона $G(x)$ оказывает аддитивное влияние на величину предельной вероятности P. При этом в качестве поправки берется максимально возможная ошибка, равная двум абсолютным отклонениям, поскольку в формуле (3) стандартный закон Гаусса используется дважды (в точках x_1 и x_2). В результате чего вводится понятие скорректированной предельной вероятности P_Δ, или *уровня доверия с поправкой на отклонение от закона Гаусса*, в виде:

$$P_\Delta = P - 2\Delta_n. \quad (4)$$

В качестве оценки величины Δ_n берется *теоретическое отклонение* $\delta_n^{\text{теор}}$, получаемое как $\delta_n^{\text{теор}} = AL_n$ из *неравенства Эссеена* [3, с. 139]:

$$\sup_x |F(x) - \Phi(x)| \le AL_n, \tag{5}$$

где $F(x) = P\left(\overline{B}_n^{-1/2} \sum_{k=1}^{n} \overline{X}_k < x\right)$,

$$L_n = \overline{B}_n^{-3/2} \sum_{k=1}^{n} M\left|\overline{X}_k\right|^3.$$

В неравенстве (5) участвуют центрированные случайные величины $\overline{X}_k = X_k - p_k$, сумма их дисперсий $\overline{B}_n = D(\overline{X}_1) + ... + D(\overline{X}_n) = B_n$ и сумма абсолютных третьих моментов $\sum M\left|\overline{X}_k\right|^3 = \sum p_k q_k (p_k^2 + q_k^2)$.

Применимость неравенства Эссеена требует, сверх принятого выше, существования конечных абсолютных моментов третьего порядка:

$$M\left|X_k - p_k\right|^3 < \infty.$$

В этом несложно убедиться:

$$M\left|X_k - p_k\right|^3 = p_k q_k (p_k^2 + q_k^2) \le 0{,}25 < \infty;$$

что и требовалось показать.

Значение числа A, впервые полученное в 1942 году, до сих пор продолжает улучшаться. В работе [4] наилучшим значением A названа оценка Шевцовой:

$$A = 0{,}469.$$

Теоретически, при неограниченных объемах выборки имеем:

$$\lim_{n\to\infty} L_n = \lim_{n\to\infty} L_d\big|_{d=1} = 0,$$

$$\delta_n^{\text{теор}} \xrightarrow{n\to\infty} 0 \text{ и } P_\Delta \xrightarrow{n\to\infty} P.$$

На деле контингент студентов вуза ограничен конечным числом M и максимально возможный объем выборки не превосходит этого числа. Для имеющегося контингента студентов при максимальном объеме выборки $n_{\max} = M$ поправка $2\Delta_n$ принимает фиксированное значение и регулировать P_Δ возможно лишь путем изменения P, при этом верхней границей P_Δ выступает $1 - 2\Delta_n$:

$$P_\Delta \xrightarrow{P\to 1} 1 - 2\Delta_n.$$

Вычисление отклонения Δ_n позволяет выполнять подгонку P_Δ к ее верхней границе с помощью графика нелинейной[1] зависимости P_Δ от параметра

$$c = \frac{P}{(1-P)},$$

который (параметр c) показывает, во сколько раз вероятность выполнения соотношения (2) превосходит вероятность его невыполнения.

Практический интерес в задачах прогнозирования количества неуспевающих студентов представляет не столько вычисление уровня доверия P при заданных границах интервала (x_1, x_2), сколько построение симметричного относительно ожидаемого количества неуспевающих студентов интервала $(-x, x)$ для заданного уровня доверия P.

Следовательно, вместо соотношения (2) практичнее использовать происходящую из него формулу доверительного интервала для оценки количества μ_n неуспевающих студентов (среди n студентов):

$$P\left(\left|\mu_n - \sum_{k=1}^{n} p_k\right| < x\sqrt{\sum_{k=1}^{n} p_k q_k}\right) = 2\Phi(x) \qquad (6)$$

или

$$P\left(\left|\mu_n - a_n\right| < \varepsilon_\gamma\right) = \gamma, \qquad (7)$$

где a_n — ожидаемое количество неуспевающих студентов;

γ — надежность доверительного интервала;

ε_γ — точность доверительного интервала при надежности γ.

Значение x вычисляется через функцию Лапласа:

$$x = \Phi^{-1}(\gamma / 2),\ \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}dz.$$

В определенных случаях (табл. 1) вместо надежности γ целесообразно использовать *надежность γ_Δ с поправкой на отклонение эмпирической функции распределения от закона Гаусса*: $\gamma_\Delta = \gamma - 2\Delta_n$.

Для использования приведенных выше соотношений необходимо знать вероятности p_k неуспеваемости каждого студента исследуемого контингента. Поскольку истинные p_k узнать невозможно, на практике они замещаются статистическими вероятностями[2], вычисленными прогностической моделью, построенной на данных прошлого опыта. Так, для оценки вероятности p_k неуспеваемости

[1] Графически такая зависимость представляется более удобной для подгонки предельной вероятности, по сравнению с линейным уравнением (4).

[2] Вероятность p_k распределения Бернулли, которая является математическим ожиданием величины X_k, в статистическом моделировании рассматривается как выборочное условное математическое ожидание, зависящее от специально выбранных характеристик (факторов неуспеваемости) студентов.

ТАБЛИЦА 1. Рекомендации к выбору надежности доверительного интервала

Отклонение от закона Гаусса	Рекомендации к выбору надежности	
	Надежность	Примечание
$\Delta_n < 0,005$ (малое)	γ	При малом отклонении поправка не имеет смысла: $\gamma_\Delta \approx \gamma$
$0,005 \le \Delta_n \le 0,05$ (допустимое)	γ_Δ	При допустимом отклонении рекомендуется использовать поправку: $0,9 \cdot \gamma \le \gamma_\Delta < \gamma$
$\Delta_n > 0,05$ (критическое)	$\gamma_\Delta^{\text{крит}}$ или γ^*	При критическом отклонении $\gamma_\Delta^{\text{крит}} < 0,9$ и, возможно, необходим поиск другого правила, взамен уравнения (6), с перерасчетом γ^*

первокурсников в текущем учебном году используется модель, обученная на экзаменационных оценках и характеристиках первокурсников прошлого учебного года.

В настоящей работе расчет теоретического отклонения $\delta_n^{\text{теор}}$, получаемого из неравенства Эссеена, сопровождается вычислением наблюдаемого отклонения $\delta_n^{\text{набл}}$ как наибольшего из абсолютных отклонений в узлах $x_i = 0,\ 1,\ 2,\ ...,\ n,\ n+1$ построения эмпирической функции распределения $F_n(x)$:

$$\delta_n^{\text{набл}} = \max_{x_i} |\, F_n(x_i) - G_n(x_i) \,|,$$

где $G_n(x)$ — закон Гаусса, соответствующий эмпирической функции распределения $F_n(x)$.

Воспроизведение эмпирической функции распределения $F_n(x)$ подразумевает многократную сдачу одной и той же экзаменационной сессии теми же самыми n студентами при сохранении всех сопутствующих условий. Каждая из таких сессий называется *репликой* выборки n студентов.

В каждой *i*-й *реплике* мы наблюдаем за количеством $\mu_n^{(i)}$ неуспевающих студентов (среди n выбранных студентов). По итогам t *реплик* составляется наблюдаемое распределение количества μ неуспевающих студентов:

$$\begin{array}{c|c|c|c|c} \mu_j & 0 & 1 & ... & n \\ \hline w_j & w_0 & w_1 & ... & w_n \end{array}, \qquad (8)$$

где μ_j — возможные значения количества неуспевающих студентов;
w_j — относительная частота появления μ_j в t *репликах*.

Тогда эмпирическая функция распределения $F_n(x)$ строится по уравнению:

$$F_n(x) = \sum_{\mu_j < x} w_j \,.$$

Закон Гаусса $G_n(x)$, соответствующий распределению (8), задается выборочными параметрами:

$$a_{\text{в}} = \sum_{j=0}^{n} w_j \mu_j ,$$

$$s_{\text{в}} = \sqrt{\frac{1}{n}\sum_{j=0}^{n} w_j (\mu_j - a_{\text{в}})^2} .$$

Поскольку осуществлять многочисленные выборки один и тех же n студентов с результатами единственной сессии не представляется возможным, задача *исследования сходимости эмпирической функции распределения количества неуспевающих студентов к закону Гаусса* неразрешима без привлечения методов статистических испытаний.

Метод Монте-Карло

Вообразим, что студент четыре раза проживает одну и ту же сессию. Присвоим студенту индекс $k = 1$ и допустим, что на все время проведения эксперимента вероятность неуспеваемости этого студента оценивается как $p_1 = 0{,}3414$. Тогда случайная величина X_1 неуспеваемости данного студента задается следующим распределением вероятности:

x	1	0
$P(X_1 = x)$	0,3414	0,6586

Воображаемый эксперимент эквивалентен *четырем статистическим испытаниям по разыгрыванию случайной величины* X_1 или *разыгрыванию четырех значений случайной величины* X_1:

$$X_1^{(1)},\ X_1^{(2)},\ X_1^{(3)},\ X_1^{(4)}.$$

Для осуществления статистических испытаний вводится непрерывная случайная величина R, распределенная равномерно в интервале (0; 1). Возможные значения r равномерной случайной величины R называют *случайными числами*[3].

Интервал (0; 1) возможных значений случайной величины R *делится величиной* $p_1 = 0{,}3414$ *на два промежутка:*

$$(0;\ 0{,}3414] \qquad (0{,}3414;\ 1).$$

Затем генерируются[4] четыре *случайных числа* $r_1^{(1)}$, $r_1^{(2)}$, $r_1^{(3)}$ и $r_1^{(4)}$:

0,1009 0,7325 0,3376 0,5201.

[3] Если быть точнее, используется не равномерно распределенная случайная величина, которая принимает значения с бесконечным числом десятичных знаков, а *квазиравномерная* случайная величина, значения которой имеют конечное количество знаков. В настоящей работе применяются *случайные числа* с четырьмя знаками после запятой.

[4] Случайные числа выбраны из книги A Million Random Digits Random Digits with 100 000 Normal Deviates, первоначально опубликованной в 1955 году корпорацией RAND, у которой возникла потребность в генерации случайных чисел для решения задач методом Монте-Карло. В отечественной литературе первые 12 500 из миллиона этих случайных чисел содержатся в «Таблицах математической статистики» Л. Н. Большева и Н. В. Смирнова.

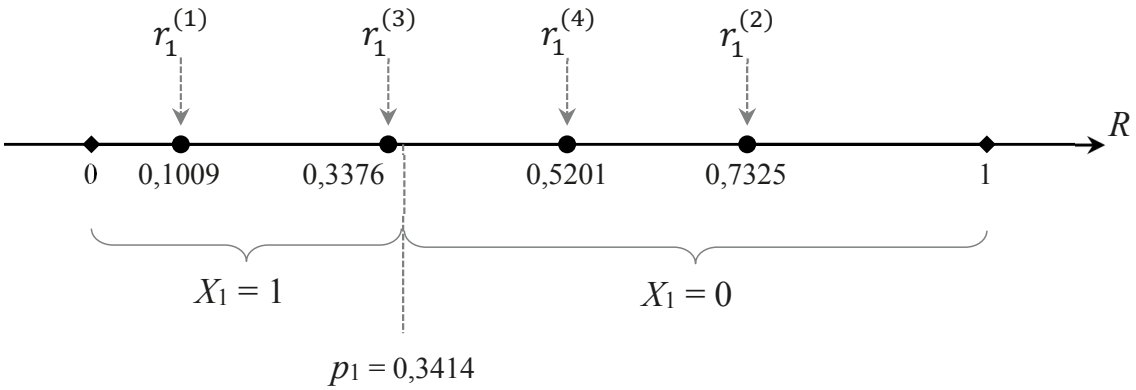

Рис. 1. Разыгрывание значений случайной величины

Поскольку $r_1^{(1)} = 0{,}1009$ попадает в промежуток (0; 0,3414], то $X_1^{(1)} = 1$; аналогично получаем $X_1^{(2)} = 0$, $X_1^{(3)} = 1$ и $X_1^{(4)} = 0$ (рис. 1).

Итог воображаемого эксперимента таков: в первую и третью вымышленные жизни у студента возникли академические долги, а во вторую и четвертую — он сдал сессию без долгов.

Представим теперь, что в том же самом эксперименте вместо одного студента участвуют три студента. Двум новым студентам присвоим индексы $k = 2,\ 3$ и допустим, что оценки вероятности их неуспеваемости равны соответственно $p_2 = 0{,}3031$ и $p_3 = 0{,}7740$.

Для каждого нового студента сгенерируем четыре ($t = 4$) новых случайных числа. Генерацию новых случайных чисел выполним, не изменяя способ генерирования, но изменяя *начальное положение генератора случайных чисел* (*random state*)[5].

Все результаты сведем в таблицу:

	i	1	2	3	4
$p_1 = 0{,}3414$	$r_1^{(i)}$	0,1009	0,7325	0,3376	0,5201
	$X_1^{(i)}$	1	0	1	0
$p_2 = 0{,}3031$	$r_2^{(i)}$	0,3754	0,2048	0,0564	0,8947
	$X_2^{(i)}$	0	1	1	0
$p_3 = 0{,}7740$	$r_3^{(i)}$	0,0842	0,2689	0,5319	0,6450
	$X_3^{(i)}$	1	1	1	1

Здесь i — порядковый номер статистического испытания или номер *реплики*.

[5] Как и в первом эксперименте, здесь случайные числа выбраны из таблицы A Million Random Digits Random Digits with 100 000 Normal Deviates, при этом для каждого следующего студента выбор четырех случайных чисел произведен с начала новой (которая не использовалась до сих пор) строки. В этом случае *начальным положением генератора случайных чисел* служат номера взятых строк: *random state* = 1, 2, 3. Объем *n* выборки и *random state* являются *гиперпараметрами*, изменение которых влечет за собой изменение значений выборочных параметров $a_в$ и $s_в$, вычисленных в статистических испытаниях.

Реплика $i=1$ дала сумму $\mu_3^{(1)} = X_1^{(1)} + X_2^{(1)} + X_3^{(1)} = 2$: в первую вымышленную жизнь среди группы 3 студентов оказалось 2 неуспевающих студента. Другие реплики дали $\mu_3^{(2)} = 2$, $\mu_3^{(3)} = 3$, $\mu_3^{(4)} = 1$.

Таким образом, методом Монте-Карло на основе четырех статистических испытаний ($t=4$) получено эмпирическое распределение случайной величины μ_3:

$$\begin{array}{c|c|c|c|c} \mu_j & 0 & 1 & 2 & 3 \\ \hline w_j & 0 & 1/4 & 1/2 & 1/4 \end{array}, \quad F_3(x) = \begin{cases} 0, & x \le 1 \\ 1/4, & 1 < x \le 2 \\ 3/4, & 2 < x \le 3 \\ 1, & x > 3 \end{cases}.$$

Далее, вычислив выборочные параметры $a_{\text{в}} = 2$ и $s_{\text{в}} \approx 0{,}8165$ закона Гаусса, находим в узлах $x_i = 0,\ 1,\ 2,\ 3,\ 4$ абсолютные отклонения

x_i	0	1	2	3	4
$\lvert F_3(x_i) - G_3(x_i) \rvert$	0,0072	0,1103	0,25	0,1397	0,0072

и получаем наблюдаемое отклонение $\delta^{\text{набл}}$ эмпирической функции распределения $F_3(x)$ от соответствующего закона Гаусса $G_3(x)$:

$$\delta_3^{\text{набл}} = 0{,}25 \ (t=4).$$

При вычислении теоретического отклонения $\delta_3^{\text{теор}}$ проведение статистических испытаний не требуется, используются $p_1 = 0{,}3414$, $p_2 = 0{,}3031$ и $p_3 = 0{,}7740$:

$$\sum_{k=1}^{3} M\left|\overline{X}_k\right|^3 = \sum_{k=1}^{3} p_k q_k (p_k^2 + q_k^2) = 0{,}3595,$$

$$\overline{B}_3^{-3/2} = \left(\sum_{k=1}^{3} p_k q_k \right)^{-3/2} = 2{,}0938,$$

$$\delta_3^{\text{теор}} = 0{,}353.$$

Нам удалось убедиться, что для трех студентов в четырех репликах наблюдаемое отклонение $\delta_3^{\text{набл}}$ не превосходит теоретического отклонения $\delta_3^{\text{теор}}$.

Наперед заметим, что изменение числа реплик t оказывает обратное влияние на величину наблюдаемого отклонения $\delta_n^{\text{набл}}$.

Уменьшив число статистических испытаний до $t=3$, отбрасыванием последней из приведенных выше реплик, получаем завышенное наблюдаемое отклонение:

$$\delta_3^{\text{набл}} = 0{,}2819 \ (t=3).$$

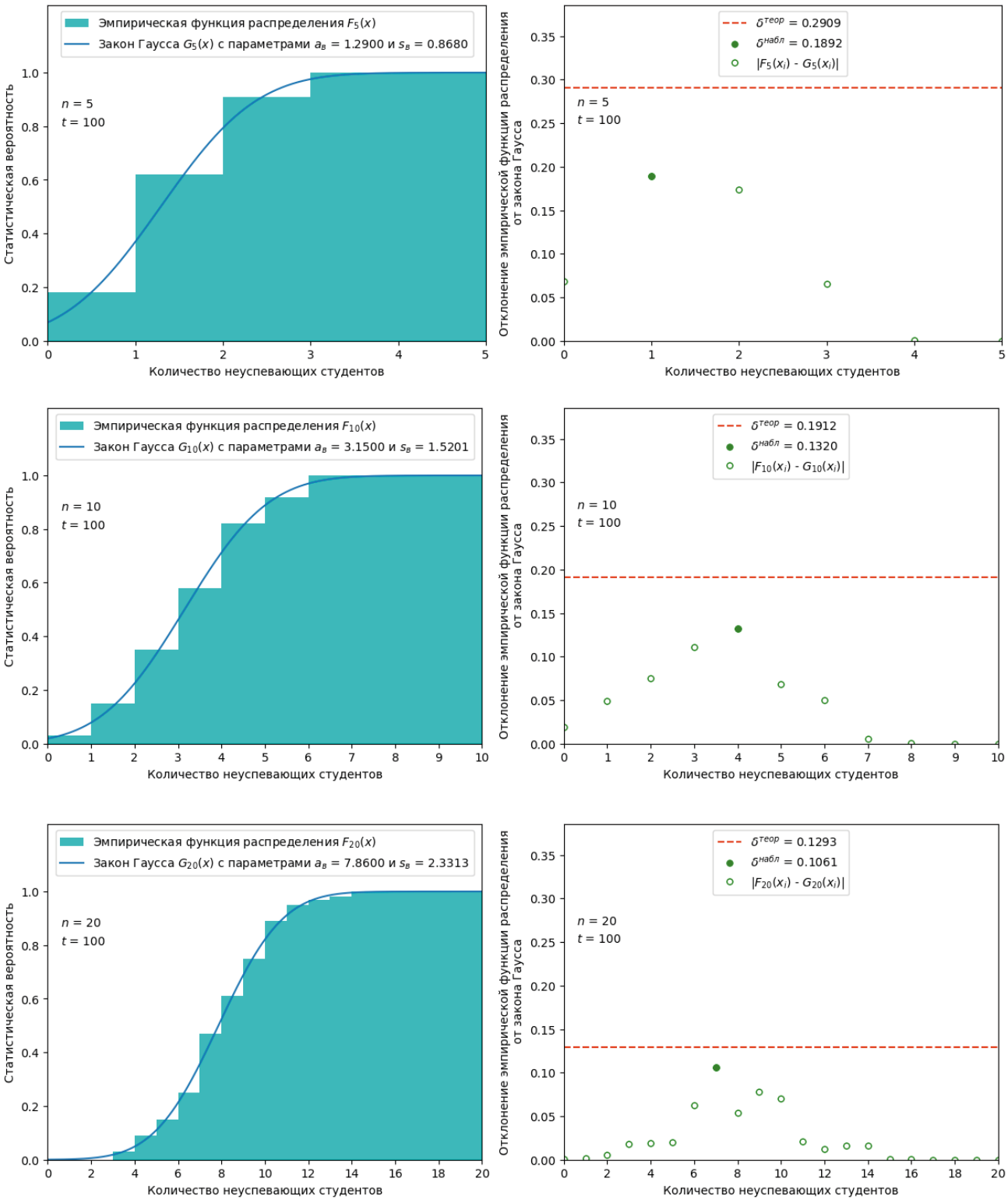

Рис. 2. Сопоставление эмпирического распределения количества неуспевающих студентов с законом Гаусса при наращивании объема выборки студентов $n = 5,\ 10,\ 20$ ($t = 100$ статистических испытаний)

Исследование сходимости распределения количества неуспевающих студентов к закону Гаусса

В начале 2023/24 учебного года вычислены прогнозные вероятности p_k неуспеваемости целого контингента 1232 студентов первого курса очной формы обучения ПГУПС[6].

Выполнен предварительный анализ сходимости эмпирического распределения количества неуспевающих студентов к закону Гаусса при наращивании объема n выборки случайно отобранных студентов: $n = 5,\ 10,\ 20$. На рис. 2 приведено сопоставление эмпирического распределения количества неуспевающих студентов с соответствующим законом Гаусса. Результаты получены в $t = 100$ статистических испытаниях[7].

На графиках рис. 2 видно, как при увеличении количества n студентов эмпирическая функция распределения сближается с законом Гаусса. Наблюдаемые отклонения не превосходят теоретического отклонения $\delta_n^{\text{теор}}$.

Проведено сопоставление эмпирического распределения количества неуспевающих студентов с законом Гаусса для целого контингента студентов $n = 1232$ (рис. 3). На левом графике различие визуально не заметно; на правом — видно, что наблюдаемое отклонение $\delta_{1232}^{\text{набл}} = 0{,}054$ в несколько раз превышает теоретическое отклонение $\delta_{1232}^{\text{теор}} = 0{,}0173$.

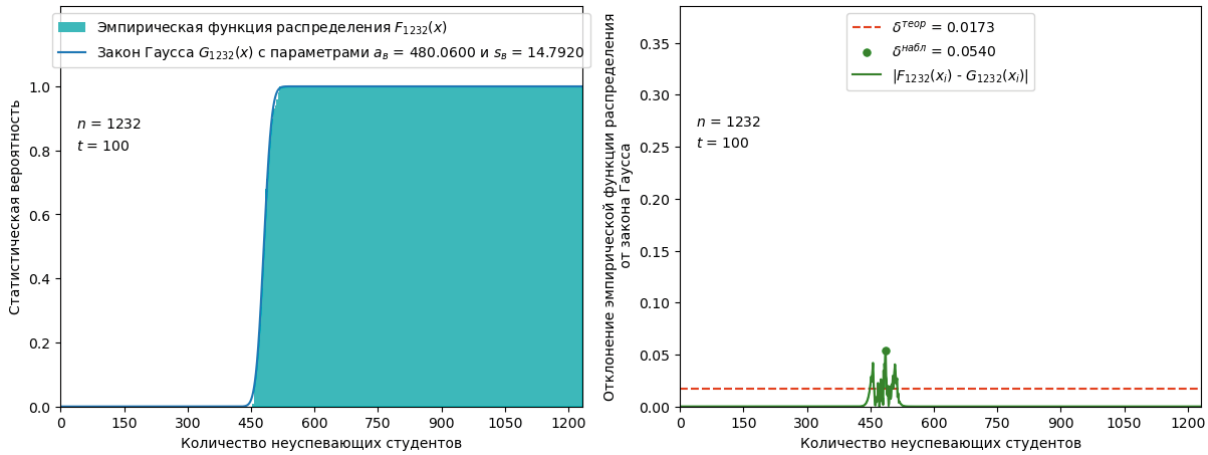

Рис. 3. Сопоставление эмпирического распределения количества неуспевающих студентов с законом Гаусса для целого контингента студентов $n = 1232$ ($t = 100$ статистических испытаний)

[6] По состоянию на 6 сентября 2023 года.

[7] Многократные статистические испытания построены на псевдослучайных числах, сгенерированных алгоритмом *«Вихрь Мерсенна»*. Это наиболее популярный детерминированный и повторяемый алгоритм, разработанный в 1997 году. Последовательность чисел, порожденная им, статистически неотличима от истинно случайной и имеет период, равный числу с шестью тысячами знаков (Matsumoto M., Nishimura T. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator, ACM Transactions on Modeling and Computer Simulation, 1998, Vol. 8, Iss. 1, Pp. 3–30). Использована имплементация этого алгоритма в библиотеке *NumPy* для научных вычислений на языке *Python*.

Динамика наблюдаемого отклонения $\delta_{1232}^{\text{набл}}$ при увеличении числа t статистических испытаний изображена на рис. 4 (левый график).

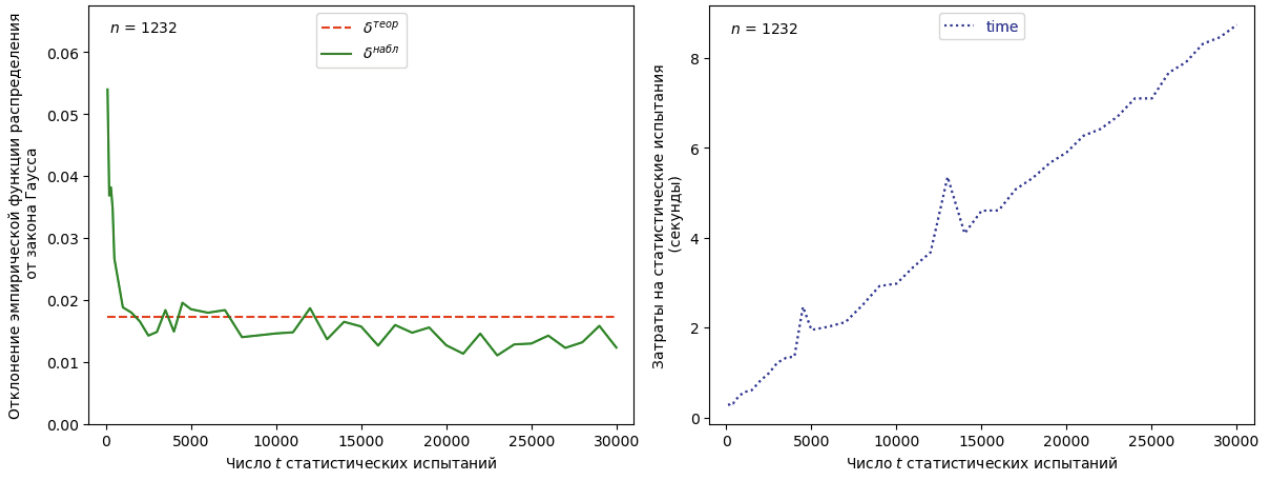

Рис. 4. График *отклонения* эмпирического распределения количества неуспевающих студентов от закона Гаусса и график *затрат* на статистические испытания при различных значениях числа t реплик

Из рис. 4 видно, что при числе $t > 15\ 000$ статистических испытаний наблюдаемые отклонения $\delta_{1232}^{\text{набл}}$ не превосходят теоретического отклонения $\delta_{1232}^{\text{теор}} = 0{,}0173$. Если проведение статистических испытаний не связано с высокими *затратами* на их осуществление, то имеется возможность добиться уровня сходимости, не превышающего $\delta^{\text{теор}}$.

Увеличение числа t статистических испытаний обеспечивает выполнение неравенства $\delta^{\text{набл}} < \delta^{\text{теор}}$ и приближает выборочные параметры $a_{\text{в}}$, $s_{\text{в}}$ к теоретическим параметрам a_n и $b_n = \sqrt{B_n}$ (рис. 5).

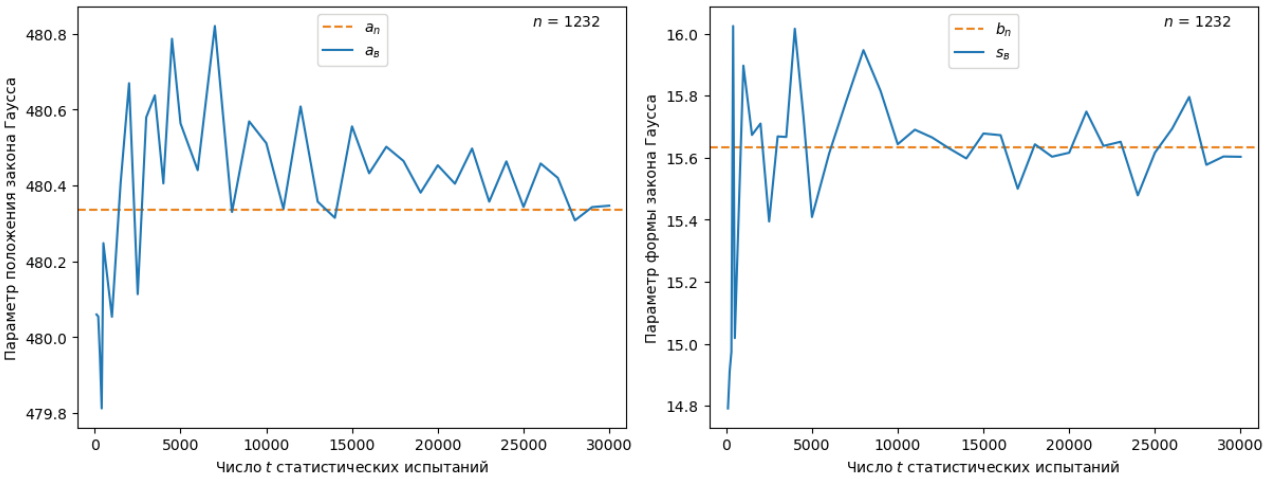

Рис. 5. Графики параметров положения и формы закона Гаусса при различных значениях t числа реплик

После проведения $t = 100\ 000$ статистических испытаний для целого контингента студентов $n = 1232$ имеем практически совпадающие значения параметров $a_{\text{в}} \approx a_n$ и $s_{\text{в}} \approx b_n$:

$a_{\text{в}} = 480{,}3351$ $a_n = 480{,}3346$;

$s_{\text{в}} = 15{,}6116$ $b_n = 15{,}6343$ ($B_n = 244{,}4305$).

Таким образом, установлено, что функция распределения количества неуспевающих студентов (среди целого контингента $n = 1232$ студентов набора 2023 года) сходится к закону Гаусса с параметрами $a_n = 480{,}3346$ и $B_n = 244{,}4305$. Оценка отклонения Δ_n равна 0,0173.

Доверительный интервал для оценки ожидаемого количества неуспевающих студентов

В качестве надежности доверительного интервала выбирается значение $\gamma = 0{,}9995$, которое обеспечивает достаточную близость γ_Δ к верхней границе $1 - 2\Delta_n = 0{,}9654$. Дальнейшее увеличение γ не дает существенного прироста надежности γ_Δ с поправкой (рис. 6).

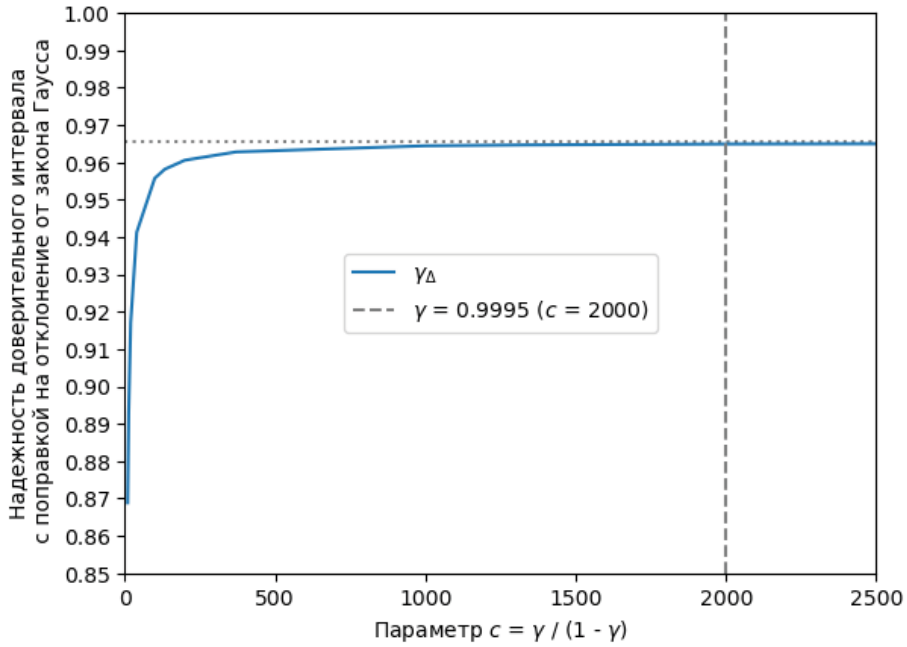

Рис. 6. Подгонка надежности γ_Δ с поправкой к верхней границе $1 - 2\Delta_n = 0{,}9654$

При $\gamma = 0{,}9995$ имеем $x = 3{,}4808$ и по формуле (6) получаем

$$P\left(\left|\mu_{1232} - 480{,}3346\right| < 54{,}4199\right) = 0{,}9995.$$

Таким образом, в начале 2023/24 учебного года получен следующий прогноз с учетом поправки $2\Delta_n = 0{,}0346$: *количество* μ *неуспевающих студентов окажется в интервале* $(426;\ 535)$ *с надежностью* $\gamma_n = 0{,}9649$.

Часть распределения количества неуспевающих студентов, соответствующая интервалу (426; 535), выделена на графиках рис. 7. В этом интервале (с учетом поправки) заключено 96,49 % всех ожидаемых исходов. Исходы за пределами этого интервала относятся к практически невозможным событиям: с вероятностью 0,0351 количество неуспевающих студентов окажется меньше 426 или более 535.

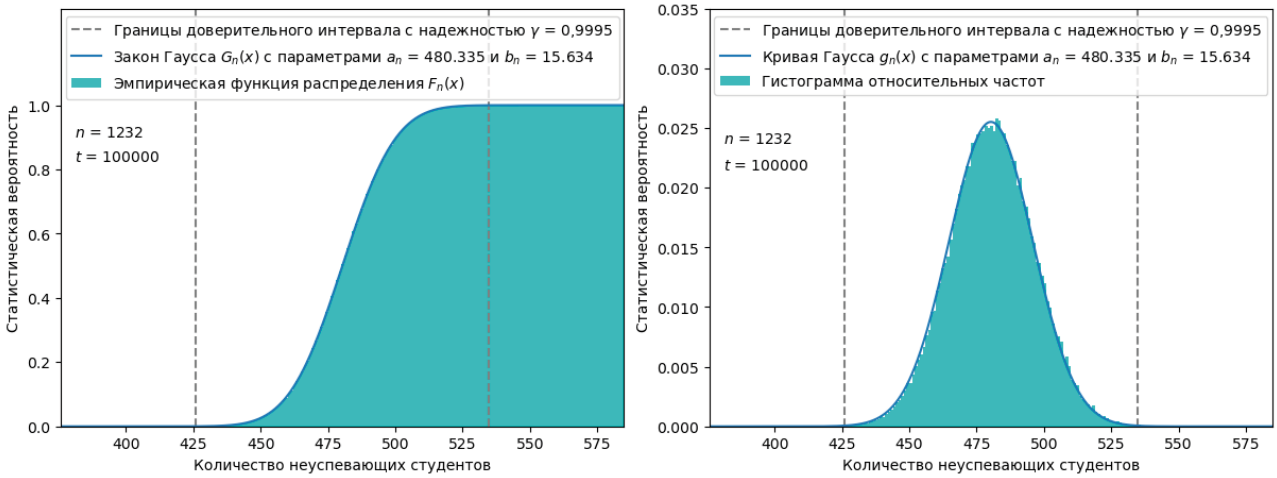

Рис. 7. Границы доверительного интервала

Смысл поправки в том, что с ее учетом в несколько раз возрастают вероятности экстремальных значений:

без поправки	с поправкой
$P(\mu_{1232} < 426) = 0,00025$	$P_\Delta(\mu_{1232} < 426) = 0,01755$
$P(\mu_{1232} > 535) = 0,00025$	$P_\Delta(\mu_{1232} > 535) = 0,01755$

Отклонение распределения количества неуспевающих студентов от соответствующего закона Гаусса на величину $\Delta_n = 0,0173$ приводит к увеличению вероятностей экстремальных значений в 70 раз.

Из рис. 7 видно также, что наряду со сходимостью функций распределений имеет место сходство формы гистограммы относительных частот с формой кривой Гаусса

$$g_n(x) = \frac{1}{\sqrt{2\pi \cdot 244,4305}} e^{-\frac{(z-480,3346)^2}{2 \cdot 244,4305}}.$$

Сравнение с классическим доверительным интервалом

Классический доверительный интервал, получаемый из интегральной теоремы Муавра — Лапласа, использует величину p, которая определяется как доля неуспевающих студентов в завершившемся учебном году. Такой интервал

равносилен построенному в настоящей работе доверительному интервалу (6), в котором все p_k постоянны и равны p.

Известно, что в завершившемся 2022/23 учебном году доля неуспевающих студентов составила 0,38 от целого контингента первокурсников. При $p = 0{,}38$ классический доверительный интервал для оценки количества неуспевающих студентов в 2023/24 учебном году оказывается более смещенным относительно истинного[8] $\mu_{1232} = 496$ по сравнению с построенным в настоящей работе интервалом (рис. 8).

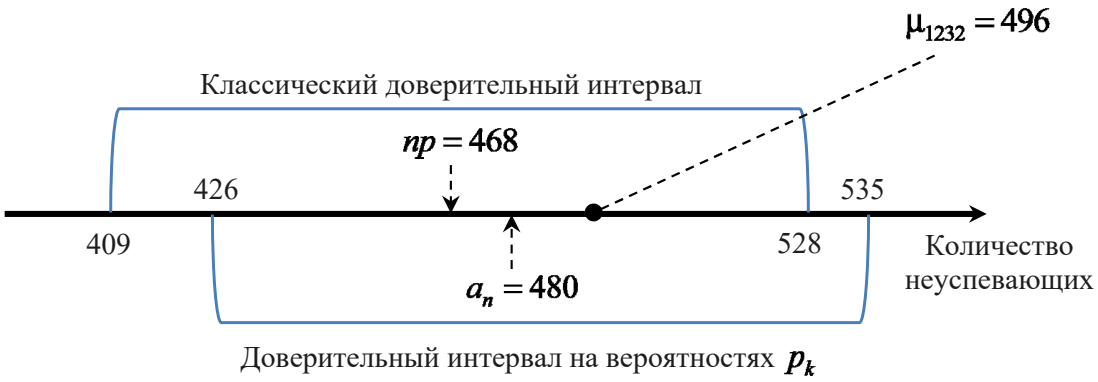

Рис. 8. Границы доверительных интервалов

Параметризация интервального прогноза

Положение и размер доверительного интервала (7) определяется несколькими величинами:

величина	тип	зависимость
a_n	параметр	звисит от p_k
ε_γ	параметр	зависит от p_k и γ
p_k	гиперпараметры	назначаются прогностической моделью
γ, γ_Δ	гиперпараметры	выбираются исследователем

Заключение

В настоящей статье обоснована особая роль закона Гаусса в задаче прогнозирования ожидаемого количества неуспевающих студентов. На основе центральной предельной теоремы построен доверительный интервал для прогнозного количества неуспевающих студентов при известных вероятностях их неуспеваемости. Применимость центральной предельной теоремы установлена по условию Ляпунова. Для оценки отклонения функции распределения количества неуспевающих студентов от закона Гаусса использовано неравенство Эссеена, в котором

[8] Количество неуспевающих первокурсников по состоянию на 1 июня 2024 года.

построение эмпирической функции распределения выполнено методом статистических испытаний Монте-Карло. Предложено использовать вычисленную оценку отклонения для поправки к надежности доверительного интервала.

Теоретические предложения апробированы на реальных данных студентов первого курса ПГУПС. Изучена сходимость эмпирического распределения количества неуспевающих студентов к закону Гаусса. Показано, что отклонения порядка 0,01–0,02 от закона Гаусса приводят к тому, что в десятки раз возрастают вероятности экстремальных значений.

Приведена параметризация интервального прогноза, из которой видно, что положение и размер доверительного интервала определяются оптимальностью прогнозной модели, используемой для вычисления вероятностей неуспеваемости студентов, и заявленным уровнем надежности. Оба обстоятельства требуют дальнейшего рассмотрения. Касательно выбора уровня надежности будет полезно, если руководство установит пределы для практической достоверности (достаточно ли 0,99; 0,999; 0,9995 или требуется этот уровень увеличить). Это обращение к руководству в свое время упоминал Якоб Бернулли в одной из аксиом своего труда «Искусство предположений» [5, с. 31].

Список источников

1. Боровков А. А. Теория вероятностей / А. А. Боровков. — М.: Эдиториал УРСС, 1999. — 472 с.

2. Математический энциклопедический словарь / Гл. ред. Ю. В. Прохоров; ред. кол.: С. И. Адян, Н. С. Бахвалов, В. И. Битюцков, А. П. Ершов и др. — М.: Сов. энциклопедия, 1988. — 847 с.

3. Петров В. В. Суммы независимых случайных величин / В. В. Петров. — М.: Наука. Гл. ред. физ.-мат. лит., 1972. — 416 с.

4. Zolotukhin A. On a bound of the absolute constant in the Berry-Esseen inequality for i.i.d. Bernoulli random variables / A. Zolotukhin, V. Nagaev, V. Chebotarev // Modern Stochastics Theory and Applications. — 2018. — Vol. 5(3). — Pp. 1–26. — DOI: 10.15559/18-VMSTA113.

5. Бернулли Я. О законе больших чисел: Пер. с лат. / Я. Бернулли. — М.: Наука. Гл. ред. физ.-мат. лит., 1986. — 176 с.

Дата поступления: 27.01.2025
Решение о публикации: 13.03.2025

Контактная информация:
КУДАРОВ Руслан Серикович — канд. техн. наук, доц.; r.s.kudarov@gmail.com
КУДАРОВ Рустем Серикович — канд. техн. наук, доц.; kudarovrs@mail.ru

Central Limit Theorem Applied to Predicting Underachievement among University Students

R. S. Kudarov, R. S. Kudarov

Emperor Alexander I Petersburg State Transport University, 9, Moskovsky pr., Saint Petersburg, 190031, Russian Federation

For citation: *Kudarov R. S., Kudarov R. S.* Central Limit Theorem Applied to Predicting Underachievement Among University Students. *Bulletin of scientific research results*, 2025, iss. 2, Pp. 172–189. (In Russian) DOI: 10.20295/2223-9987-2025-2-172-189

Summary

Purpose: The State Programme "Scientific and Technological Development of the Russian Federation" gives an important place to the training of a new generation of highly qualified engineers capable of ensuring the country's technological sovereignty. Higher education institutions are being transformed into "advanced engineering schools" to train specialists in modern knowledge-intensive and multidisciplinary technologies. Greater attention is being paid to the disciplines of the first years of study, without which further study is incomplete. This article presents a method for predicting the number of underachieving students, which should help in planning measures to ensure timely implementation of the curriculum. **Methods:** The prediction of the number of underachieving students is based on the central limit theorem. The applicability of the central limit theorem is established using the Lyapunov condition. The convergence of the distribution of the number of underachieving students to Gauss's law is studied using the Esseen inequality, while the empirical distribution function is modelled using the Monte Carlo method. **Results:** A confidence interval is constructed to estimate the number of underachieving students with known probabilities of underachievement for each student. A modification has been introduced to the reliability of the confidence interval for the deviation of the empirical distribution from Gauss Law. **Practical significance:** The interval prediction of the number of underachieving first-year students at the end of the academic year has been calculated.

Keywords: The central limit theorem, Gauss Law, Monte Carlo method, Educational Data Mining, educational data analysis, forecasting learning outcomes, prediction of student underachievement.

References

1. Borovkov A. A. *Teoriya veroyatnostey* [Probability theory]. Moscow: Editorial URSS Publ., 1999, 472 p. (In Russian)

2. *Matematicheskiy entsiklopedicheskiy slovar'. Gl. red. Yu. V. Prokhorov; red. kol.: S. I. Adyan, N. S. Bakhvalov, V. I. Bityutskov, A. P. Ershov i dr.* [Mathematical encyclopedic dictionary. Ch. ed. Yu. V. Prokhorov; ed. quantity: S. I. Adyan, N. S. Bakhvalov, V. I. Bityutskov, A. P. Ershov et al.]. Moscow: Sov. entsiklopediya Publ., 1988, 847 p. (In Russian)

3. Petrov V. V. *Summy nezavisimykh sluchaynykh velichin* [Sums of independent random variables]. Moscow: Nauka. Gl. red. fiz.-mat. lit. Publ., 1972, 416 p. (In Russian)

4. Zolotukhin A., Nagaev V., Chebotarev V. On a bound of the absolute constant in the Berry-Esseen inequality for i.i.d. Bernoulli random variables. Modern Stochastics Theory and Applications, 2018, vol. 5(3), Pp. 1–26. DOI: 10.15559/18-VMSTA113.

5. Bernulli Ya. *O zakone bol'shikh chisel: Per. s lat.* [On the law of large numbers: Trans. from Latin]. M.: Nauka. Gl. red. fiz.-mat. lit. Publ., 1986, 176 p. (In Russian)

Received: January 27, 2025
Accepted: March 13, 2025

Author's information:
Ruslan S. KUDAROV — PhD in Engineering, Associate Professor; r.s.kudarov@gmail.com
Rustem S. KUDAROV — PhD in Engineering, Associate Professor; kudarovrs@mail.ru