

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (государственный университет)
ФАКУЛЬТЕТ УПРАВЛЕНИЯ И ПРИКЛАДНОЙ МАТЕМАТИКИ
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН
КАФЕДРА «ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ»

Фрей Александр Ильич

**О дискретных аппроксимациях
непрерывных вероятностных распределений**

511656 - Математические и информационные технологии

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Научный руководитель:
н.с. ВЦ РАН, к.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2008

Содержание

1	Введение	4
2	Основные результаты	5
2.1	Определения	5
2.2	Выборка Y_m^L , состоящая из нулей и единиц	6
2.3	Трёхступенчатая выборка Z^L	6
2.4	Вычисление параметров выборок Y_m^L и Z^L	6
2.5	Округление	8
3	Доказательства	9
3.1	Явный вид выражения $Q(Z^L, \varepsilon)$	9
3.2	Метод моментов. Теорема адекватности.	11
3.3	Округление параметров трёхступенчатой выборки.	12
3.4	Контрпримеры.	14
3.4.1	Контрпример для Y_m^L	14
3.4.2	Контрпример для Z^L	15
4	Численные оценки	15
4.1	Классические верхние оценки	15
4.2	Численные результаты для комбинаторных оценок	15
5	Заключение	19

Аннотация

Рассматривается задача оценивания вероятности больших отклонений средних значений в наблюдаемой и скрытой частях выборки. Делается попытка изучения непрерывных распределений чисто комбинаторными методами. Предлагается новый метод построения верхних оценок для вероятности больших отклонений. Суть метода заключается в замене исходной непрерывной выборки определенным дискретным аналогом, для которого вероятность большого отклонения вычисляется в явном виде. Предложен конкретный вид дискретного аналога исходной выборки, для которого в явном виде вычислена вероятность больших отклонений. Сформулирован корректный метод вычисления параметров дискретной выборки. Проведены вычислительные эксперименты. Найдены контрпримеры, при которых предложенные оценки не являются верхними.

1 Введение

В сильной вероятностной аксиоматике известны несколько неравенств, оценивающих вероятность отклонения суммы независимых случайных величин от своего среднего значения. Все они являются уточнениями неравенства Чебышева:

Теорема 1 (Чебышев). Пусть X — случайная величина. Тогда $\forall \varepsilon > 0$ выполнено:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\mathbb{D}(X)}{\varepsilon^2} \quad (1)$$

Введём некоторые обозначения. Пусть X_1, \dots, X_n — независимые случайные величины. Обозначим $S_n = \frac{1}{n} \sum_{i=1}^n X_i$, и

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{D}(X_i).$$

В дальнейшем нас будет интересовать **односторонние** отклонения суммы случайных величин от своего среднего значения:

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq \delta(\varepsilon).$$

Простейшее оценку для $\delta(\varepsilon)$ получаем из неравенства Чебышева-Кантелли:

Теорема 2 (Чебышев-Кантелли). В наших обозначениях $\forall \varepsilon > 0$ выполнено:

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq \frac{\sigma_n^2}{\sigma_n^2 + n\varepsilon^2} \quad (2)$$

Для суммы ограниченных случайных величин выполняется оценка Гефдинга (Hoeffding's inequality). Приведём ее формулировку:

Теорема 3 (Hoeffding). Пусть X_1, \dots, X_n — независимые случайные величины, принимающие значение из отрезка $[0, 1]$ с вероятностью единица. Тогда $\forall \varepsilon > 0$ выполнено:

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq e^{-2n\varepsilon^2} \quad (3)$$

Более строгим является неравенство Бернштейна (Bernstein), уточняющее (3) при известной дисперсии случайных величин.

Теорема 4 (Bernstein). Пусть X_1, \dots, X_n — независимые случайные величины, принимающие значение из отрезка $[0, 1]$ с вероятностью единица. Предположим, что $\mathbb{E}(S_n) = \frac{1}{2}$. Тогда $\forall \varepsilon > 0$ выполнено

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq \exp\left(-\frac{2n\varepsilon^2}{4\sigma_n^2 + \varepsilon}\right) \quad (4)$$

Неравенство МакДиармида (McDiarmid) обобщает (3) на случай произвольных функций от случайной величины (не обязательно суммы, как в неравенстве Гефдинга).

Все эти неравенства применяются, в частности, при оценке обобщающей способности алгоритмов классификации - для оценки отклонения доли ошибок на контрольной выборке от доли ошибок на обучающей выборке. Их недостаток заключается в том, что это "оценки худшего случая". Все эти неравенства существенно завышают величину $P(S_n - E(S_n) \geq \varepsilon)$.

К сожалению, все эти неравенства используют понятие меры. Кроме того, фигурирующие в неравенствах вероятности не поддаются непосредственному измерению. Рассмотрим постановку аналогичной задачи в слабой вероятностной аксиоматике, и постараемся сделать более точные комбинаторные оценки для вероятности больших уклонений.

2 Основные результаты

2.1 Определения

Напомним некоторые определения слабой вероятностной аксиоматики.

Определение 1. Назовём **выборкой** X^L набор из L чисел:

$$X^L = \{x_1, x_2, \dots, x_L\} \in \mathbb{R}^L.$$

Зафиксируем два числа - длину контроля k и длину обучения ℓ (так, что бы выполнялось условие $L = k + \ell$). Рассматриваются всевозможные разбиения выборки на два непересекающихся множества $X_n^L = X_n^k \cup X_n^\ell$, где $n \in \{1, \dots, N\}$, а $N = C_L^k$ - количество различных разбиений.

Гипотеза 1. Все разбиения равновероятны.

Определение 2. Отклонением средних назовём выражение

$$D(X^k, X^\ell) = \frac{1}{k} \sum_{x \in X^k} x - \frac{1}{\ell} \sum_{x \in X^\ell} x \quad (5)$$

Отклонение средних можно также называть "переобученостью". Она показывает, на сколько больше ошибок выдаёт алгоритм на контроле по сравнению с обучением. Требуется оценить сверху эмпирическую функцию распределения переобучености - т.е. долю тех разбиений, в которых переобученость превышает заданный порог ε .

Определение 3. Вероятностью больших уклонений будем называть выражение

$$Q(X^L, \varepsilon) = \frac{1}{N} \sum_{n=1}^N [D(X_n^k, X_n^\ell) \geq \varepsilon], \quad (6)$$

где квадратные скобки переводят истинное условие в единицу, а ложное - в ноль.

Наша задача получить верхнюю оценку $Q(X^L, \varepsilon) \leq \delta(\varepsilon)$, не зависящую от выборки X^L . Применительно к задачам оценки качества обучения будем называть параметр ε точностью, а $\delta(\varepsilon)$ - надёжностью обучения.

2.2 Выборка Y_m^L , состоящая из нулей и единиц

Для вывода оценки типа Гефдинга естественно поместить выборку в единичный гиперкуб. В дальнейшем мы будем предполагать, что $X^L \in [0, 1]^L$. Рассмотрим для начала случай выборки, лежащий непосредственно в вершинах единичного куба: $Y_m^L \in \{0, 1\}^L$ (тут нижний индекс m означает число единиц в выборке: $m = \sum_{i=1}^L y_i$). Для этого случая известна точная комбинаторная оценка, выражаемая через сумму членов гипергеометрического распределения:

$$Q(Y_m^L, \varepsilon) = \sum_{t=s_0}^{s_1} h_{Lm}^{\ell t}, \quad (7)$$

где $s_0 = \max(0, m - k)$, $s_1 = \lfloor (m - \varepsilon k) \frac{\ell}{L} \rfloor$, $h_{Lm}^{\ell t} = \frac{C_{L-m}^{\ell-t} C_m^t}{C_L^\ell}$.

Правда эта оценка надежности все еще зависит от выборки (точнее, от числа m — количества единиц в ней). Что бы избавиться от этого можно взять максимум от правой части выражения (7) по всем $m \in \{0 \dots L\}$. Этот максимум достигается либо при равном, либо при при отличающемся на единицу количестве нулей и единиц. Однако взяв максимум мы, тем самым, сильно завысим оценку (7).

Гипотеза 2 (Бадзян).

$$Q(X^L, \varepsilon) \leq \sup_{m \in \{1, \dots, L\}} Q(Y_m^L, \varepsilon) \quad (8)$$

Андрей Бадзян доказал эту гипотезу для частного случая $k = \ell = \frac{L}{2}$. Численные эксперименты показывают, что она выполняется и для $k \neq \ell$.

2.3 Трёхступенчатая выборка Z^L

Вместо Y_m^L можно рассмотреть другую выборку, более похожую на исходную X^L . Пусть $Z^L(n_q, n_1, q)$ — выборка, в которой n_q чисел равно q , n_1 единиц, а остальные числа равны 0.

Для неё можно вывести следующее выражение для вероятности больших отклонений:

$$Q(Z^L, \varepsilon) = \frac{1}{N} \sum_{\ell_q=0}^{n_q} C_{n_q}^{\ell_q} \sum_{\ell_1=0}^{s_1} C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_1}^{\ell_1}, \quad (9)$$

где

$$s_1 = \left\lfloor \frac{\ell(qn_q + n_1) - k\ell\varepsilon}{L} - \ell_q q \right\rfloor.$$

При $n_q = 0$ это выражение, как и следует ожидать, превращается в (7).

2.4 Вычисление параметров выборок Y_m^L и Z^L

Интересен такой вопрос: как избавиться от взятия максимума в (8)? Другими словами, можно ли каким-то конструктивным способом выбрать число m так, что неравенство по прежнему оставалось в силе. Достаточно очевидно, что $Q(X^L, \varepsilon)$ сильно зависит от дисперсии выборки. Вычисляя ее для Y_m^L , получим:

$$D(Y_m^L) = \frac{m(L-m)}{L^2}.$$

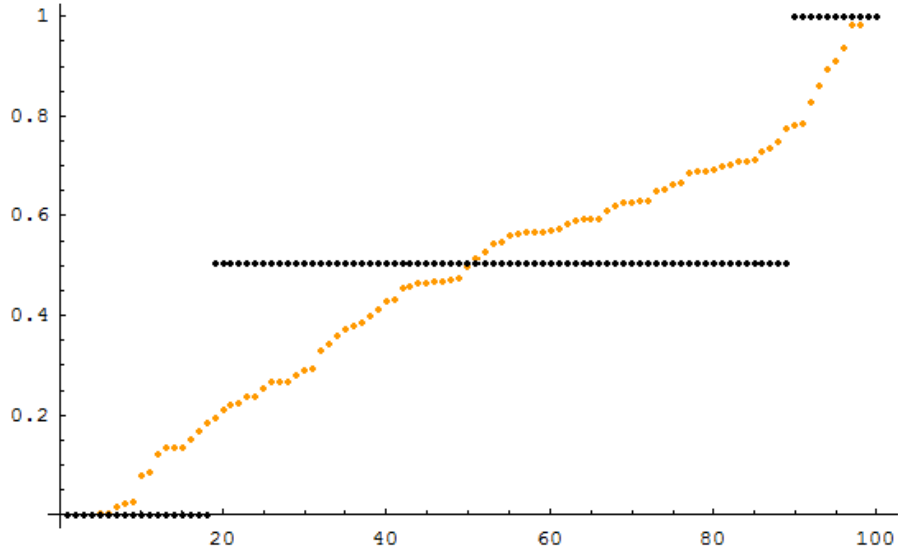


Рис. 1: Вариационные ряды X^L и Z^L

Идея заключается в том, что бы оценить дисперсию исходной выборки $D(X^L)$ и выбрать число m исходя из равенства

$$D(Y_m^L) = D(X^L).$$

Оказывается, такой подход приводит к весьма точной оценке надежности. К сожалению, полученная оценка не является верхней - существуют контрпримеры (см. п. 3.4).

Пусть теперь мы хотим приблизить X^L трёхступенчатой выборкой Z^L . Возникает вопрос: какие параметры исходной выборки нужно знать, что бы с их помощью удачно вычислить параметры выборки Z^L ? Попробуем воспользоваться методом моментов: обозначим 1й, 2й и 3й моменты исходной выборки X^L через $\mu_1 = \sum x_i$, $\mu_2 = \sum x_i^2$ и $\mu_3 = \sum x_i^3$. Приравнивая их к моментам выборки Z^L , получим значения искомым параметров:

$$\begin{cases} q = \frac{\mu_2 - \mu_3}{\mu_1 - \mu_2} \\ n_q = \frac{(\mu_1 - \mu_2)^3}{(\mu_2 - \mu_3)(\mu_1 - 2\mu_2 + \mu_3)} \\ n_1 = \frac{\mu_1\mu_3 - \mu_2^2}{\mu_1 - 2\mu_2 + \mu_3} \end{cases} \quad (10)$$

С помощью неравенства Коши-Буняковского-Шварца можно доказать теорему адекватности (о разумности полученных оценок):

$$0 \leq q \leq 1, n_0 \geq 0, n_q \geq 0, n_1 \geq 0.$$

На Рис. 1 изображены два вариационных ряда - ряд модельной выборки X^L (оранжевый цвет) и ряд выборки Z^L с параметрами, подобранными по формулам (10) (черный цвет). Выборки состоят из $L = 100$ чисел.

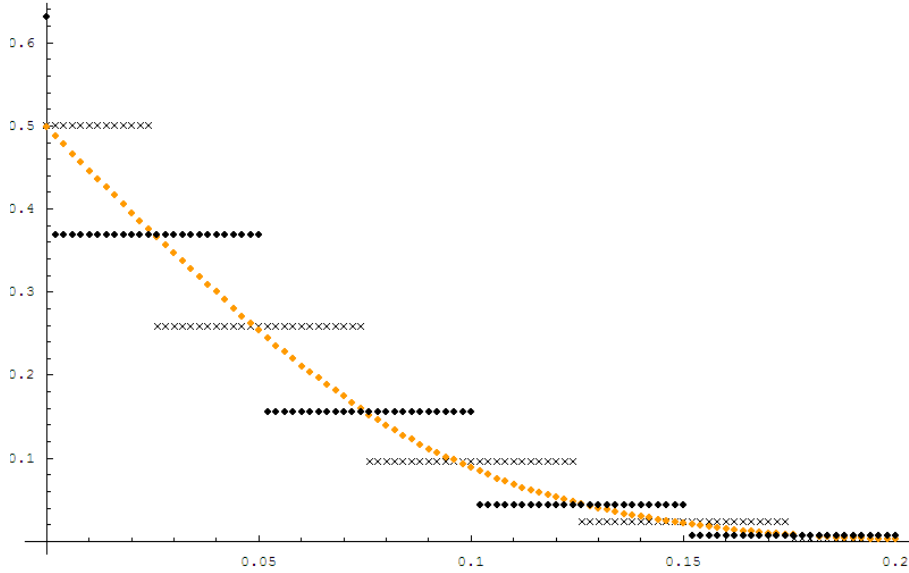


Рис. 2: Ступеньки на графиках - последствия априорного выбора округления.

Рассмотрим еще один, на первый взгляд даже более разумный, вариант выборки-заменителя. Вновь рассмотрим трёхступенчатую выборку и зафиксируем у нее положение ступеньки на уровне $\frac{1}{2}$. Теперь для выбора оставшихся двух параметров достаточно приравнять только первый и второй момент. Прodelывая это, находим

$$\begin{cases} n_{0.5} = 2\mu_2 - \mu_1 \\ n_1 = 4(\mu_1 - \mu_2) \end{cases} \quad (11)$$

Для этой выборки уже нет теоремы адекватности: легко представить себе ситуацию, когда $\mu_2 < \frac{\mu_1}{2}$.

2.5 Округление

Напомним, что число единиц m в выборке Y_m^L предлагается выбирать исходя из равенства дисперсии. Однако вполне очевидно, что при $m \in \mathbb{N}$ можно добиться лишь приближенного равенства дисперсии. Для этого придётся округлить решение уравнения

$$D(Y_m^L) = D(X^L).$$

Приведём численный график, на котором оранжевая кривая соответствует графику $Q(X^L, \varepsilon)$ а два черных графика — функциям $Q(Y_{\lceil m \rceil}^L, \varepsilon)$ и $Q(Y_{\lfloor m \rfloor}^L, \varepsilon)$ (округление происходит либо всегда вверх, либо всегда вниз). Видно, что в зависимости от ε то один, то другой способ округления оказывается более предпочтительным.

Итак, если мы надеемся получить верхнюю оценку, нам придется для каждого ε рассматривать оба способа округления — $Q(Y_{\lceil m \rceil}^L)$, $Q(Y_{\lfloor m \rfloor}^L)$, и выбирать то, которое приводит к большей оценке вероятности большого уклонения. Аналогично предлагается поступать для трёхступенчатой выборки. Однако в этом случае не так просто перечислить все целые тройки чисел (n_0, n_q, n_1) , приближающие истинное решение (10). Все такие точки мы будем называть **окрестностью округления**. Подробнее об округлении параметров трехступенчатой выборки см. в "Доказательствах".

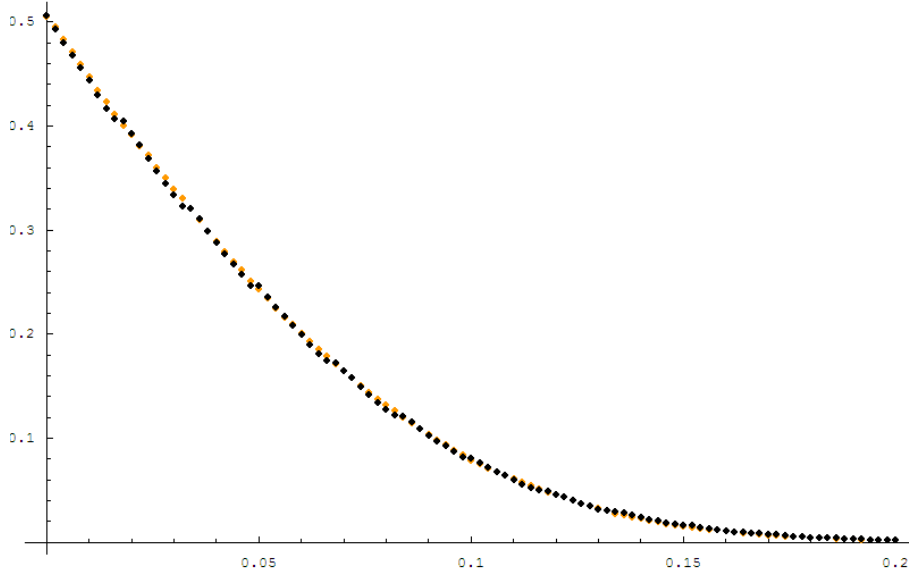


Рис. 3: Точность оценки при усреднении по окрестности округления

К сожалению ни одна из полученных оценок не является верхней. Как для выборки Y_m^L , так и для выборки Z^L существуют контрпримеры, показывающие что даже значительная свобода округления не приводит к строгому результату. В этой ситуации остаётся придумать эвристический метод, дающий как можно более точное приближение. Для этой цели логично не брать максимум, а **усреднять** по всем целочисленным точкам из окрестности округления подходящего размера. Следующий рисунок иллюстрирует точность, с которой данный подход приближает истинное поведение зависимости вероятности большого уклонения от порога ε .

3 Доказательства

3.1 Явный вид выражения $Q(Z^L, \varepsilon)$

Теорема 5 (Вероятность большого уклонения трехступенчатой выборки).

$$Q(Z^L, \varepsilon) = \frac{1}{N} \sum_{\ell_q=0}^{n_q} C_{n_q}^{\ell_q} \sum_{\ell_1=0}^{s_1} C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_1}^{\ell_1}, \quad (12)$$

где

$$s_1 = \left\lfloor \frac{\ell(qn_q + n_1) - k\ell\varepsilon}{L} - \ell_q q \right\rfloor.$$

Доказательство. Мы хотим вычислить выражение для вероятности большого уклонения выборки

$$Q(Z^L, \varepsilon) = \frac{1}{N} \sum_{n=1}^N [D(Z_n^k, Z_n^\ell) \geq \varepsilon] \quad (13)$$

Пусть в тестовой подвыборке оказалось ℓ_q чисел со значением q и ℓ_1 единиц. Распишем условие, при котором выполнено $D(Z_n^k, Z_n^\ell) \geq \varepsilon$.

$$\begin{aligned}
D(Z^\ell, Z^k) &= \frac{1}{k} \sum_{x \in Z^k} x - \frac{1}{\ell} \sum_{x \in Z^\ell} x \\
&= \frac{1}{k} ((n_q - \ell_q)q + (n_1 - \ell_1)) - \frac{1}{\ell} (\ell_q q + \ell_1) \\
&= \frac{L((n_1 - \ell_1) + q(n_q - \ell_q)) - k(n_1 + qn_q)}{k\ell} \geq \varepsilon.
\end{aligned} \tag{14}$$

Левая часть монотонно убывает при увеличении числа единиц в обучающей выборке. Обозначим через s_1 максимальное количество единиц, при котором наше условие выполняется. Выразим s_1 из (14):

$$s_1 = \left\lfloor \frac{\ell(qn_q + n_1) - k\ell\varepsilon}{L} - \ell_q q \right\rfloor$$

Теперь подсчитаем, сколькими способами в обучении могло оказаться ровно ℓ_1 из n_1 единиц, ровно ℓ_q из n_q чисел, равных q , и $\ell - \ell_1 - \ell_q$ из $L - n_1 - n_q$ нулей. Используя биномиальные коэффициенты, искомое выражение приобретает вид

$$C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_q}^{\ell_q} C_{n_1}^{\ell_1}$$

Суммируя по всевозможным значениям ℓ_q и ℓ_1 получаем искомый результат:

$$\begin{aligned}
Q(Z^L, \varepsilon) &= \\
&= \frac{1}{N} \sum_{\ell_q=0}^{n_q} \sum_{\ell_1=0}^{n_1} C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_q}^{\ell_q} C_{n_1}^{\ell_1} \left[\ell_q q + \ell_1 \leq \frac{k\ell\varepsilon + \ell(qn_q + n_1)}{L} \right] = \\
&= \frac{1}{N} \sum_{\ell_q=0}^{n_q} \sum_{\ell_1=0}^{s_1} C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_q}^{\ell_q} C_{n_1}^{\ell_1}.
\end{aligned} \tag{15}$$

Теорема доказана.

Замечание. Подставим в формулы предыдущей леммы $n_q = 0$. Получим

$$\begin{aligned}
s_1 &= \left\lfloor (n_1 - k\varepsilon) \frac{\ell}{L} \right\rfloor \\
Q(Y_{n_1}^L, \varepsilon) &= \frac{1}{N} \sum_{\ell_1=0}^{s_1} C_{L-n_1}^{\ell-\ell_1} C_{n_1}^{\ell_1}
\end{aligned}$$

Теперь заметим, что $n_1 = m$, а начинать суммирование можно с $s_0 = \max(0, m - k)$. Действительно, если единиц в выборке больше, чем длина контроля то есть смысл рассматривать только $\geq (m - k)$ единиц в обучающей выборке (иначе соответствующий биномиальный коэффициент обнулится). Мы получили классическое выражение для суммы членов гипергеометрического распределения:

$$Q(Y_m^L, \varepsilon) = \sum_{\ell_1=s_0}^{s_1} h_{(Lm)}^{\ell \ell_1},$$

где $s_0 = \max(0, m - k)$, $s_1 = \lfloor (m - \varepsilon k) \frac{\ell}{L} \rfloor$, $h_{(Lm)}^{\ell t} = \frac{C_{L-m}^{\ell-t} C_m^t}{C_L^\ell}$.

3.2 Метод моментов. Теорема адекватности.

Теперь рассмотрим подробнее метод моментов для вычисления параметров выборки $Z^L(n_q, n_1, q)$.

Приравнивая моменты исходной выборки X^L и выборки Z^L получаем систему уравнений:

$$\begin{cases} \mu_1 = \sum x_i = n_q q + n_1 \\ \mu_2 = \sum x_i^2 = n_q q^2 + n_1 \\ \mu_3 = \sum x_i^3 = n_q q^3 + n_1 \end{cases}$$

Решая ее относительно n_q, n_1, q получим:

$$\begin{cases} q = \frac{\mu_2 - \mu_3}{\mu_1 - \mu_2} \\ n_q = \frac{(\mu_1 - \mu_2)^3}{(\mu_2 - \mu_3)(\mu_1 - 2\mu_2 + \mu_3)} \\ n_1 = \frac{\mu_1 \mu_3 - \mu_2^2}{\mu_1 - 2\mu_2 + \mu_3} \end{cases} \quad (16)$$

Необходимо убедиться в том, что эти результаты "разумны" — т.е. проверить условия $0 \leq q \leq 1$, а так же $n_0 \geq 0$, $n_q \geq 0$ и $n_1 \geq 0$.

Теорема 6 (Теорема адекватности). *Выражения (16) дают разумные значения параметров выборки Z^L :*

$$0 \leq q \leq 1, n_0 \geq 0, n_q \geq 0, n_1 \geq 0.$$

Доказательство. Напомним, что $x_i \in [0, 1]$, следовательно $x_i^3 \leq x_i^2 \leq x_i$. По определению $\mu_1 = \sum x_i$, $\mu_2 = \sum x_i^2$ и $\mu_3 = \sum x_i^3$, а значит $\mu_3 \leq \mu_2 \leq \mu_1$. Получаем

$$q = \frac{\mu_2 - \mu_3}{\mu_1 - \mu_2} \geq 0.$$

Теперь заметим, что из неравенства $0 \leq (x_i - 1)^2$ следует $2x_i^2 \leq x_i + x_i^3$. Тогда $2\mu_2 \leq \mu_1 + \mu_3$ и $q \leq 1$.

Теперь очевидно, что $n_q \geq 0$ (n_q является произведением положительных множителей). Для доказательства $n_1 \geq 0$ достаточно показать, что $\mu_1 \mu_3 - \mu_2^2 \geq 0$. Рассмотрим вектора

$$\begin{aligned} u &= (\sqrt{x_1}, \sqrt{x_2}, \dots, \sqrt{x_L}) \in \mathbb{R}^L \\ v &= (\sqrt{x_1^3}, \sqrt{x_2^3}, \dots, \sqrt{x_L^3}) \in \mathbb{R}^L \end{aligned}$$

Заметим, что $\|u\|^2 = \mu_1$, $\|v\|^2 = \mu_3$, $(u, v) = \mu_2$. Следовательно $\mu_1 \mu_3 \geq \mu_2^2$ по неравенству Коши-Буняковского-Шварца. Таким образом, $n_1 \geq 0$.

Осталось показать, что $n_0 \geq 0$, т.е. $n_1 + n_q \leq L$. Вычислим:

$$\begin{aligned} n_1 + n_q &= \frac{(\mu_1 - \mu_2)^3}{(\mu_2 - \mu_3)(\mu_1 - 2\mu_2 + \mu_3)} + \frac{\mu_1 \mu_3 - \mu_2^2}{\mu_1 - 2\mu_2 + \mu_3} = \\ &= \frac{\mu_1^2 + \mu_2^2 - \mu_1(\mu_2 + \mu_3)}{\mu_2 - \mu_3} = \frac{(\mu_1 - \mu_2)^2 + \mu_1(\mu_2 - \mu_3)}{\mu_2 - \mu_3} = \\ &= \frac{(\mu_1 - \mu_2)^2}{\mu_2 - \mu_3} + \mu_1. \end{aligned}$$

Мы хотим показать, что это выражение не превосходит длины выборки L . Логично обозначить $L = \sum_{i=1}^L x_i^0 = \mu_0$. Следовательно, мы хотим доказать неравенство

$$(\mu_0 - \mu_1)(\mu_2 - \mu_3) \geq (\mu_1 - \mu_2)^2.$$

Это снова напоминает неравенство Коши-Буняковского-Шварца! Действительно, рассмотрим вектора

$$\begin{aligned} u &= (\sqrt{1 - x_1}, \sqrt{1 - x_2}, \dots, \sqrt{1 - x_L}) \in \mathbb{R}^L \\ v &= (\sqrt{x_1^2 - x_1^3}, \sqrt{x_2^2 - x_2^3}, \dots, \sqrt{x_L^2 - x_L^3}) \in \mathbb{R}^L \end{aligned}$$

Эти вектора специально выбраны так, что квадраты их длин равны соответственно $\|u\|^2 = \mu_0 - \mu_1$ и $\|v\|^2 = \mu_2 - \mu_3$. Тогда их скалярное произведение записывается в виде

$$(u, v) = \sum_{i=1}^L \sqrt{(1 - x_i)(x_i^2 - x_i^3)}.$$

Учитывая то, что $(1 - x_i)(x_i^2 - x_i^3) = (1 - x_i)^2 x_i^2$ получаем следующее выражение для скалярного произведения:

$$(u, v) = \sum_{i=1}^L x_i(1 - x_i) = \mu_1 - \mu_2.$$

Следовательно по неравенству Коши-Буняковского-Шварца получим

$$(\mu_0 - \mu_1)(\mu_2 - \mu_3) = \|u\|^2 \|v\|^2 \geq (u, v)^2 = (\mu_1 - \mu_2)^2.$$

А значит $n_1 + n_q \leq L$, и следовательно $n_0 \geq 0$. Теорема адекватности доказана.

3.3 Округление параметров трехступенчатой выборки.

Вычислим количество способов округления трехступенчатой выборки. В ней есть три целых параметра - n_0 , n_q и n_1 , связанные условием $n_0 + n_q + n_1 = L$. Казалось бы, существует четыре способа округления: можно взять два любых параметра, и для каждого из них есть две возможности: округлять вниз и вверх. Однако один из этих четырех способов обязательно будет нарушать условие на суммарное количество чисел.

Пример. Пусть $L = 4$, а метод моментов дал оценки всех параметров, равные $\frac{4}{3}$. Тогда обязательное условие заключается в том, что два числа нужно округлить вниз, и только одно - вверх. Поэтому существует только три возможности. То же самое имеем и в общем случае.

Определение 4. Окрестностью округления $B_r(x_0, x_q, x_1)$ назовём все целочисленные наборы параметров задачи, получаемые из тройки (x_0, x_q, x_1) изменением каждой координаты не более чем на r , при условии что сохраняется сумма параметров и все параметры остаются неотрицательными.

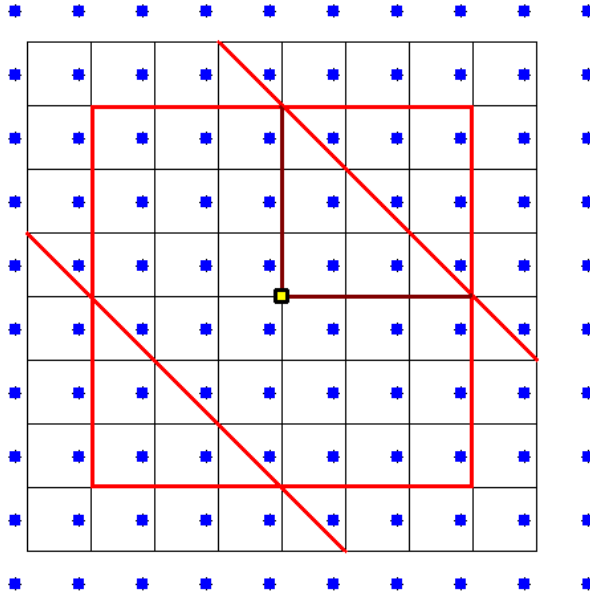


Рис. 4: Иллюстрация к доказательству леммы о размере окрестности.

Параметр r будем называть радиусом округления. Количество точек в окрестности назовем ее размером. Таким образом "округлению до соседнего целого" соответствует $r = 1$.

Очевидно, что размер окрестности радиуса r не превосходит $(2r)^2$. Действительно, в нашем распоряжении только два свободных параметра, и каждый из них может принимать $2r$ различных значений. Можно доказать следующее утверждение.

Теорема 7. Пусть r является целым числом, и все параметры задачи (x_0, x_q, x_1) далеко (т.е. не менее чем на r) отстоят от граничных значений 0 и L и не являются целыми. Тогда размер окрестности простым образом выражается через её радиус:

$$|B_r| = 3r^2.$$

□ Сформулируем для начала следующее вспомогательное утверждение.

Лемма 1. Рассмотрим евклидову плоскость. Назовём узлами все её точки, у которых обе координаты — целочисленные. Пусть на плоскости размещен квадрат со стороной $n \in \mathbb{N}$, причем его стороны параллельны осям координат и не проходят через узлы целочисленной решётки.

Утверждается, что внутри этого квадрата содержится ровно n^2 узлов.

Действительно, спроецируем квадрат на ось абсцисс. Получим отрезок $[a, a + n]$ (тут $a \in \mathbb{R} \setminus \mathbb{N}$ - абсцисса левой стороны квадрата). В этом отрезке лежат следующие целые точки: $[a], \dots, [a+n]$, а их в точности n штук. Аналогично при проецировании на ось ординат получаем n точек, и следовательно внутри квадрата n^2 узлов.

Вернемся к нашей теореме. Обозначим центр окрестности округления через (X, Y, Z) , Пусть (A, B, C) - целочисленная точка из нашей окрестности. Обозначим изменения координат $a = A - X$, $b = B - Y$, $c = C - Z$.

Из того, что сумма параметров при округлении сохраняется, следует условие $a + b + c = 0$. Значит нас интересует количество целочисленных троек (A, B, C) , удовлетворяющих условию $|a| \leq r$, $|b| \leq r$, $|a + b| \leq r$.

Рассмотрим квадрат $|a| \leq r$, $|b| \leq r$. Условия теоремы подобраны так, что ни одна из целочисленных точек (A, B) не попадает на его границу. Следовательно по утверждению доказанной леммы в квадрате содержится ровно $4r^2$ точек. Осталось доказать, что условие $|a + b| \leq r$ отсекает ровно r^2 целочисленных точек (вне зависимости от значения дробных частей исходных параметров).

Это условие отсекает два треугольника площади $\frac{r^2}{2}$ каждый. Количество целочисленных точек, попавших в каждый из этих треугольников, подсчитать трудно. Однако совершенно очевидно, что в сумме в этих треугольниках ровно r^2 точек. Что бы в этом убедиться, достаточно нижний левый треугольник "пририсовать" в другом месте - непосредственно под правым верхним треугольником (от этого количество узлов внутри перемещаемого треугольника не изменится). Теперь можно вновь воспользоваться утверждением леммы для малого квадрата со стороной r .

Иллюстрацией вышесказанного служит Рис. 4. На нём синими квадратами отмечены целочисленные узлы решётки, желтая точка в середине — центр окрестности (округляемая точка), красными линиями показаны условия-ограничения. ■

Для получения оценок предлагается рассматривать **все** точки из окрестности округления (с некоторым фиксированным радиусом и центром, полученным исходя из метода моментов).

3.4 Контрпримеры.

К сожалению, ни одна из предложенных оценок не является верхней. Тому есть несколько контрпримеров.

3.4.1 Контрпример для Y_m^L .

Приведём контрпример, который показывает, что выборка Y_m^L не приводит к верхней оценке, если параметр m выбирать исходя из дисперсии. Рассмотрим выборку $X^L = (0, 0, \dots, 0, 0.25, 0.75, 1, \dots, 1, 1)$. Количество нулей и единиц предполагается одинаковым, и равным $k - 1$. Выборочная дисперсия $D(X^L) = \frac{4k-3}{16k}$. Решая уравнение $D(Y_m^L) = D(X^L)$ относительно m получаем находим

$$\frac{m(2k - m)}{4k^2} = \frac{4k - 3}{16k} \Rightarrow m = k \pm \sqrt{0.75k}$$

Два случая, отличающиеся плюсом и минусом, симметричны, и дают одинаковый результат для вероятности большого отклонения. Выберем знак минус. Вопрос округления этого числа решим в пользу завышения результата: будем рассматривать оба варианта - как $m = \lceil k - \sqrt{0.75k} \rceil$, так и $m = \lfloor k - \sqrt{0.75k} \rfloor$, и выбирать из них максимум. Выберем число ε такое, что бы $Q(Y_m, \varepsilon)$ равнялась нулю, а для исходной выборки была строго больше нуля. Наибольшая отклонение средних для Y_m будет при том разбиении, в котором все единицы попадут в контроль. Т.е необходимо положить

$$\varepsilon > \max_{n=1, \dots, N} D(X_n^k, X_n^\ell) = \frac{m}{k}.$$

Для исходной выборки рассмотрим разбиение, в котором в контроль попали числа, большие $\frac{1}{2}$. Тогда отклонение средних будет равно $\frac{(k-1)+0.75}{k}$, и необходимо положить ε меньше этого числа. Убедимся, что это возможно, т.е. $\lceil k - \sqrt{0.75k} \rceil < k - 0.25$. Это неравенство действительно выполнено при $k \geq 2$.

Проблема не только в том, что наша оценка перестает выполняться лишь при больших значениях ε . Рис. 6 показывает, что для рассмотренной нами выборки "проблемы" начинаются гораздо раньше.

3.4.2 Контрпример для Z^L .

Приведём контрпример, который показывает, что и выборка Z^L не приводит к верхней оценке. Приведенный рисунок соответствует выборке $X^L = (0, 0.2, 0.25, 0.85, 0.8, 1)$. Метод моментов даёт оценки $q \approx 0.50$, $n_q \approx 2.54$, $n_1 \approx 1.84$. Перебирая все 12 точек из B_2 и строя верхнюю огибающую всех полученных графиков для $Q(Z^L, \varepsilon)$ получаем Рис. 7 На нём отчетливо виден участок, где оранжевый график лежит выше черного.

4 Численные оценки

4.1 Классические верхние оценки

Сравним, для начала, поведение верхних оценок классических неравенств. На первый взгляд кажется, что неравенства Гефдинга (3) и Бернштейна (4) являются более строгими, чем простое неравенство Чебышева-Кантелли (2). Однако это далеко не всегда так. Убедимся в этом, построив графики верхних оценок этих неравенств в зависимости от значения порога ε . На Рис. 5 непрерывные графики соответствуют верхним оценкам Чебышева-Кантелли, Гефдинга и Бернштейна.

Для сравнения приведены две точные функции $P(S_n - E(S_n) \geq \varepsilon)$. Гладкий график, построенный по черным точкам, соответствует равномерно-распределенной на $[0, 1]$ случайной величине. График, построенный по красным точкам, соответствует бернулиевской случайной величине, принимающей с вероятностью $\frac{1}{2}$ значения 0 или 1. Для каждой точки графика использовался метод Монте-Карло, оценивающий вероятность по $N = 10^5$ экспериментам. Во всех случаях количество суммируемых случайных величин $n = 7$, дисперсия случайной величины полагалась равной $\sigma^2 = \frac{1}{4}$.

Видно, что все три классических неравенства дают одинаковый по порядку величины результат, существенно завышающий истинную оценку. Более того, значения всех оценок начинаются с единицы, хотя для всех симметричных распределений график должен начинаться с $\frac{1}{2}$.

4.2 Численные результаты для комбинаторных оценок

Теперь рассмотрим несколько модельных выборок X^L (каждый раз $L = 100$) и исследуем зависимость $Q(X^L, \varepsilon)$ от ε . Во всех экспериментах длина контроля полагалась равной длине обучения ($k = \ell = 50$). Величина $Q(X^L, \varepsilon)$ вычислялась численно, по $N = 10^4$ случайным разбиениям. Для численных вычислений и построения графиков использовалась программа Mathematica 5.2.

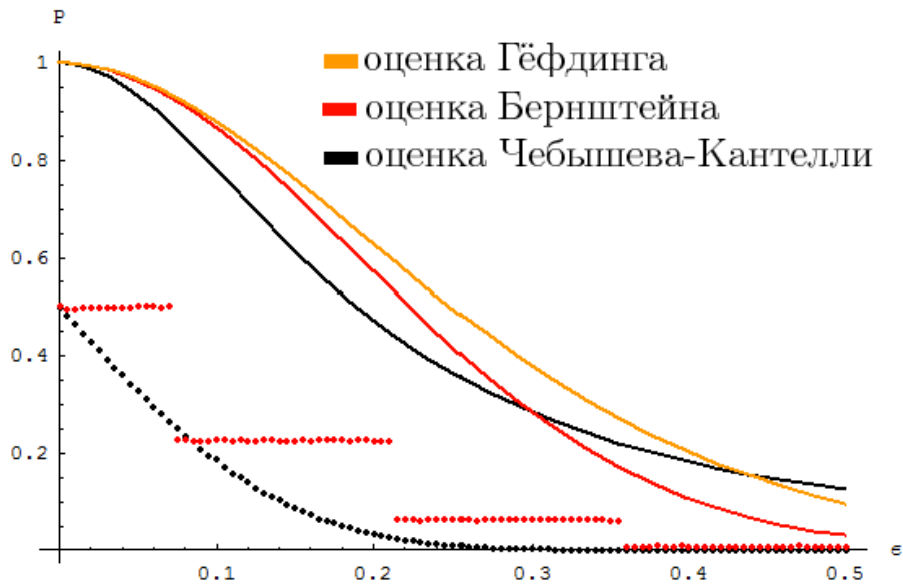


Рис. 5: Верхние оценки вероятности отклонения S_n от $E(S_n)$

На каждом графике красным цветом отмечена оценка худшего случая, полученная из гипергеометрического распределения (т.е. $Q(Y^L, \epsilon)$). Она совпадает с оценкой Андрея Баздяна, и по следствию из доказанной им теоремы всегда лежит выше всех остальных графиков. Оранжевым цветом изображена модельная выборка X^L , черным - выборка Z^L с параметрами, подобранными по методу моментов.

Для сравнения с классическими неравенствами (Чебышева-Кантелли, Гелфдинга, Бернштейна) стоит отождествлять график, построенный по красным точкам (Рис. 4.1) и оценку гипергеометрического распределения.

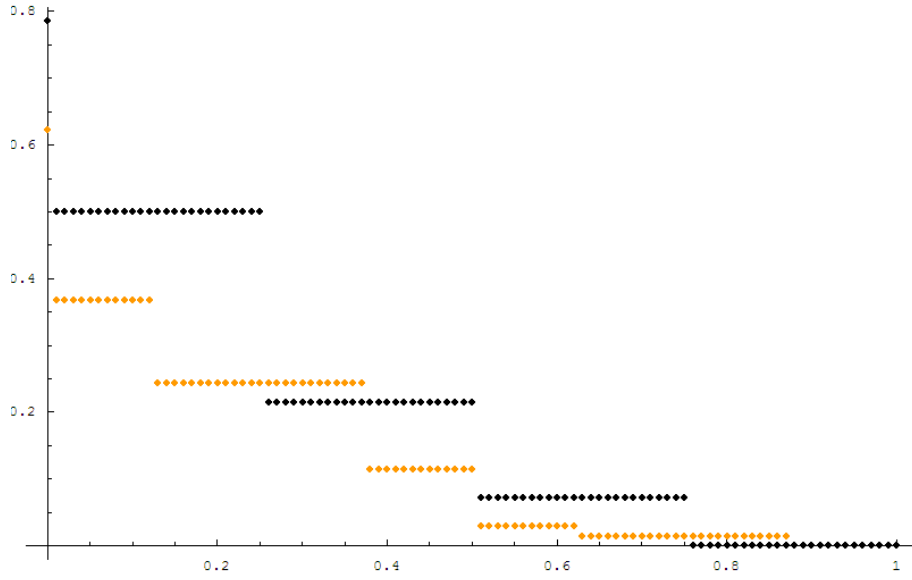


Рис. 6: Контрпример - $Q(Y_m^L, \varepsilon)$ оказывается ниже $Q(X^L, \varepsilon)$. $L=8$.

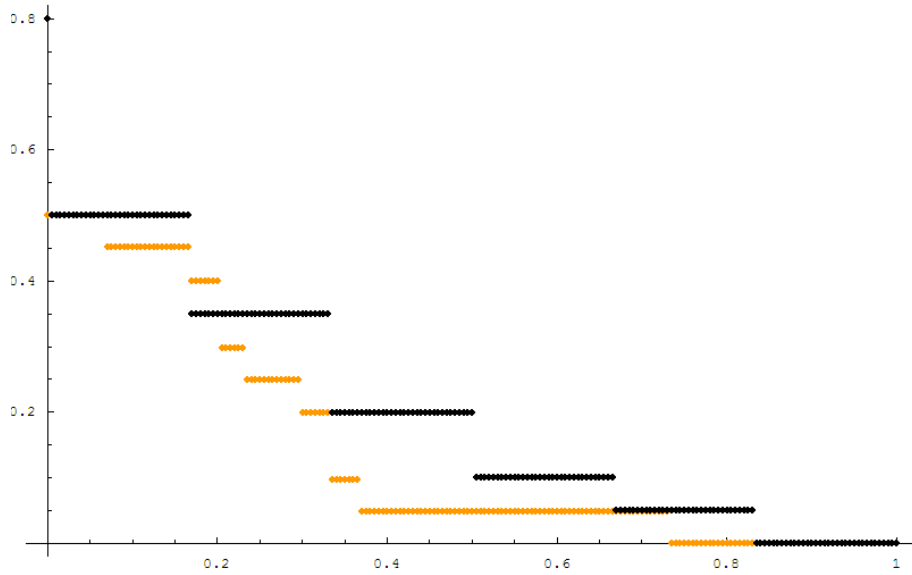
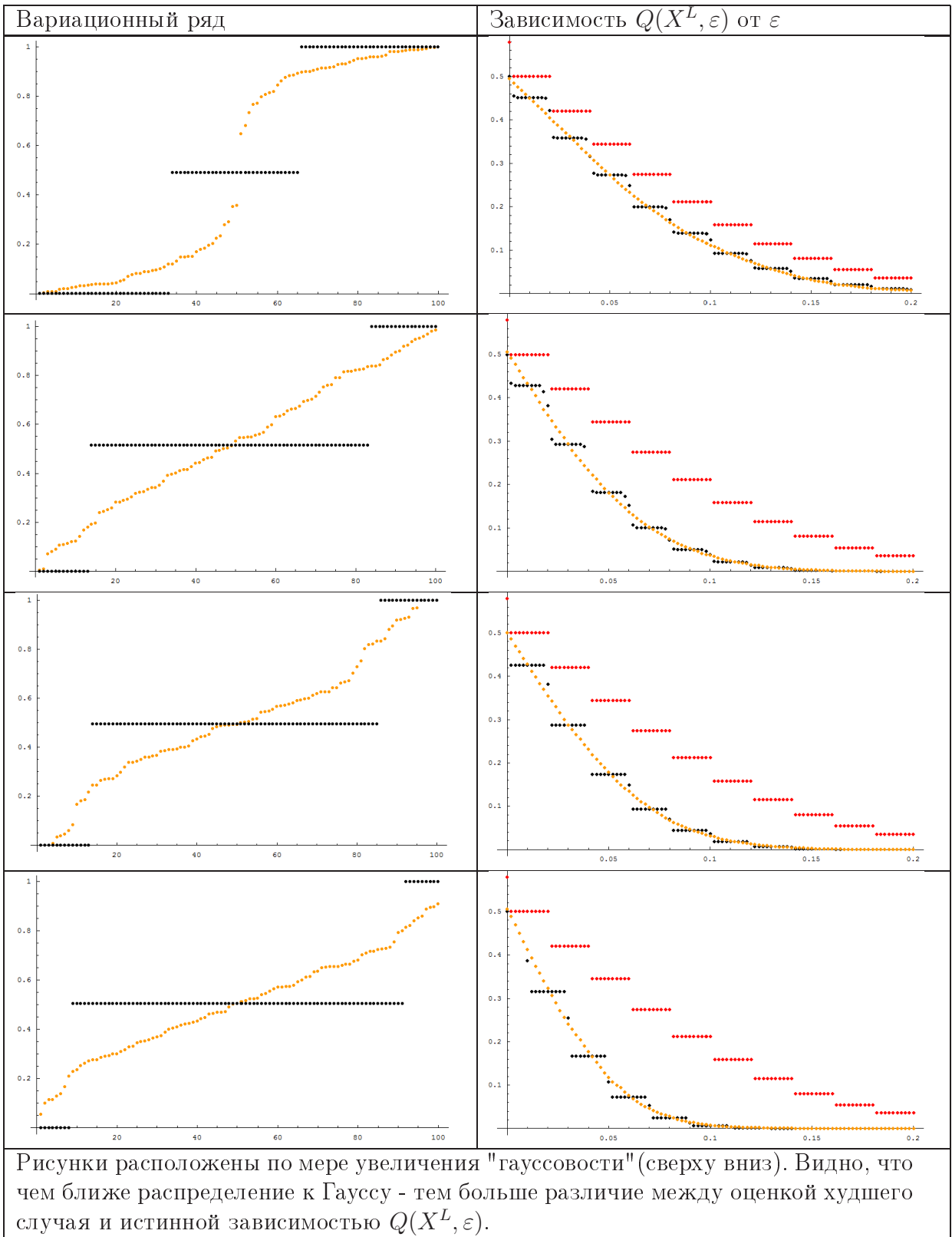


Рис. 7: Контрпример - $Q(Z^L, \varepsilon)$ оказывается ниже $Q(X^L, \varepsilon)$. $L=6$.



5 Заключение

В ходе работы предпринята попытка изучения непрерывных распределений комбинаторными методами. Предложена трехступенчатая выборка в качестве универсального аппроксиматора непрерывных распределений, лежащих на отрезке $[0, 1]$. С помощью удачного выбора параметров этой выборки удается аппроксимировать вероятность больших отклонений для достаточно широкого класса распределений.

В ходе дальнейших исследований предполагается исследовать чувствительность функционала $Q(Z^L, \varepsilon)$ к небольшому изменению параметров исходной выборки. Также предполагается получить асимптотические результаты для больших длин выборок L и оценить скорость сходимости.