

Неравенства Чебышева для статистических характеристик числового набора¹

Г.И.Фалин, д.ф.м.н., проф.
кафедра теории вероятностей
механико-математический факультет
МГУ им М.В.Ломоносова (Москва)

А.И.Фалин, к.ф.м.н., доцент
кафедра общей математики
факультет ВМиК
МГУ им М.В.Ломоносова (Москва)

Неравенства Чебышева – это группа однотипных по внешнему виду и методу доказательства неравенств для вероятностей выхода случайной величины за некоторые границы. Они названы в честь великого русского математика и механика Панфутия Львовича Чебышева (1821-1894), который в 1866 году доказал наиболее важное из этих неравенств и, что самое главное, применил его для простого доказательства одного из наиболее значимых результатов теории вероятностей, так называемого закона больших чисел. Подробнее о неравенстве Чебышева в теории вероятностей можно прочитать в замечательном элементарном курсе теории вероятностей [1], написанном выдающимися советскими математиками Б.В.Гнеденко и А.Я.Хинчиным.

В нашей статье мы расскажем о естественных аналогах некоторых неравенств Чебышева для обычных числовых наборов. Эти неравенства не требуют использования языка теории вероятностей и их доказательства совершенно элементарны. С помощью этих неравенств можно глубже понять изучаемую в школе описательную статистику.

Эта тема представляется нам важной ещё и в связи с патриотическим воспитанием учащихся средних школ на уроках математики. Как отмечал А.Я.Хинчин в своей статье «О воспитательном эффекте уроков математики» (сокращённый вариант этой статьи можно найти в [2]), наилучший способ для этого – рассказывать о понятных школьникам научных открытиях наших великих соотечественников. По мнению Хинчина «этот приём, помимо впечатляющей силы воздействия, особенно ценен ещё тем, что он значительно повышает интерес учащихся к истории математической науки, а

¹ Г.И.Фалин, А.И.Фалин. Неравенства Чебышева для статистических характеристик числового набора. *Математика*, 2011, №17, стр.21-26.

во многих случаях даёт повод и возможность эффективным образом ознакомить учащихся с математическими фактами, выходящими за пределы официальной программы и счастливым образом её дополняющими». Особо выделяя двух великих русских математиков, Н.И.Лобачевского и П.Л.Чебышева, А.Я.Хинчин отмечает, что «значительно менее известны у нас творения... нашего великого учёного П.Л.Чебышева». Сам Хинчин рекомендует знакомить учащихся с деятельностью Чебышева в области теории чисел (просто в то время, когда Хинчин писал свою статью, школьная программа не давала других возможностей). Сейчас, в связи с внедрением в программу средней школы элементов статистики и теории вероятностей уместно рассказать о результатах Чебышева в этой области.

Итак, пусть $x = [x_1, x_2, \dots, x_n]$ – набор из n действительных чисел. Обозначим через $N(x \geq a)$ количество чисел набора, удовлетворяющих неравенству $x \geq a$, а через $\nu(x \geq a) = \frac{N(x \geq a)}{n}$ – долю таких чисел. Чтобы проиллюстрировать это обозначение, рассмотрим набор из 100 чисел

27, 52, 43, 38, 47, 8, 21, 40, 32, 53, 45, 54, 35, 28,
40, 18, 31, 45, 24, 30, 37, 15, 39, 34, 48, 25, 30, 7,
32, 12, 26, 35, 48, 19, 33, 26, 17, 30, 42, 22, 53, 28,
42, 36, 23, 10, 34, 46, 16, 29, 35, 52, 41, 32, 21,
39, 55, 25, 29, 8, 36, 44, 26, 55, 34, 19, 42, 54, 27,
10, 45, 20, 31, 50, 18, 9, 41, 14, 38, 40, 23, 49, 33
15, 24, 46, 36, 28, 32, 37, 51, 20, 29, 47, 33, 27, 41,
22, 39, 40.

Этот набор мы взяли из учебника [3], стр. 251; числа набора – это время (в минутах), которое 100 выбранных наугад учеников гипотетической школы тратят на дорогу в школу.

Если $a = 45$, то, как нетрудно подсчитать, ровно 20 чисел из этого набора удовлетворяют неравенству $x \geq 45$. Поэтому для рассматриваемого набора $N(x \geq 45) = 20$, а $\nu(x \geq 45) = 0,2$. Величина $N(x \geq 45)$ показывает, сколько учеников тратит на дорогу достаточно большое время (по меньшей мере три четверти часа) и намного лучше характеризует набор, чем стандартные характеристики: среднее значение $M_x = \frac{x_1 + x_2 + \dots + x_n}{n}$, медиана μ_x и

дисперсия $D_x = \frac{(x_1 - M_x)^2 + (x_2 - M_x)^2 + \dots + (x_n - M_x)^2}{n}$. Ещё интереснее

величина $\nu(x \geq 45)$, которая говорит, что 20% учеников слишком долго (по меньшей мере три четверти часа) добираются до школы.

Если $a = 6$, то, как нетрудно определить, все числа набора удовлетворяют неравенству $x \geq 6$. Поэтому для рассматриваемого набора $N(x \geq 6) = 100$, а $\nu(x \geq 6) = 1$.

Если $a = 56$, то, как нетрудно определить, ни одно число набора не удовлетворяет неравенству $x \geq 56$. Поэтому для рассматриваемого набора $N(x \geq 56) = 0$, $\nu(x \geq 56) = 0$.

Первое неравенство Чебышева (его иногда называют неравенством Маркова), которое мы сейчас докажем, устанавливает связь между этими новыми характеристиками числового набора и основной мерой положения чисел набора, средним значением M_x .

Теорема 1 (неравенство Чебышева). Для произвольного набора $x = [x_1, x_2, \dots, x_n]$ неотрицательных чисел и любого положительного числа a справедливо неравенство $\nu(x \geq a) \leq \frac{M_x}{a}$.

Доказательство. Разобьём сумму $S = x_1 + x_2 + \dots + x_n$ всех чисел набора на две суммы: $S = S' + S''$. В первую включим те члены набора, которые меньше, чем заданное число a , а во вторую – те члены набора, которые больше или равны a .

Поскольку наш набор состоит из неотрицательных чисел, сумма $S' \geq 0$. Если в наборе нет чисел меньших a , мы положим $S' = 0$, так что в любом случае верно неравенство $S' \geq 0$.

Вторая сумма состоит из $N(x \geq a)$ чисел набора, каждое из которых не меньше a . Поэтому $S'' \geq a \cdot N(x \geq a)$. Если в наборе нет чисел больших или равных, чем a (т.е. $N(x \geq a) = 0$), мы положим $S'' = 0$, так что в любом случае верно неравенство $S'' \geq a \cdot N(x \geq a)$.

Теперь мы имеем:

$$M_x = \frac{x_1 + \dots + x_n}{n} = \frac{S' + S''}{n} \geq \frac{0 + a \cdot N(x \geq a)}{n} = a \cdot \nu(x \geq a),$$

что равносильно доказываемому неравенству Чебышева.

Метод, применённый при доказательстве Теоремы 1 позволяет получить большое число других неравенств. В качестве примера мы докажем следующее утверждение, которое связывает медиану μ_x , среднее значение M_x , наименьшее $x_{(1)} \equiv \min(x_1, \dots, x_n)$ и наибольшее $x_{(n)} \equiv \max(x_1, \dots, x_n)$ числа набора.

Теорема 2. Если набор $x = [x_1, x_2, \dots, x_n]$ состоит из нечётного количества чисел, то

$$\frac{2n}{n+1} \cdot M_x - \frac{n-1}{n+1} \cdot x_{(n)} \leq \mu_x \leq \frac{2n}{n+1} \cdot M_x - \frac{n-1}{n+1} \cdot x_{(1)}. \quad (1)$$

Если же количество чисел в наборе чётно, то

$$\frac{2n}{n+2} \cdot M_x - \frac{n-2}{n+2} \cdot x_{(n)} \leq \mu_x \leq \frac{2n}{n+2} \cdot M_x - \frac{n-2}{n+2} \cdot x_{(1)}. \quad (2)$$

Доказательство. Будем считать, что числа набора упорядочены по возрастанию: $x_1 \leq x_2 \leq \dots \leq x_n$. Тогда $x_{(1)} \equiv \min(x_1, \dots, x_n) = x_1$, $x_{(n)} \equiv \max(x_1, \dots, x_n) = x_n$. Порядок, в котором расположены эти числа не влияет на величину их суммы и, соответственно, на среднее значение набора.

Предположим, что $n = 2k + 1$, т.е. набор состоит из нечётного количества чисел. Тогда медиана μ_x является $(k+1)$ -м в порядке возрастания членом набора: $\mu_x = x_{k+1}$. Запишем сумму S всех чисел набора в виде: $S = (x_1 + \dots + x_k) + x_{k+1} + (x_{k+2} + \dots + x_{2k+1})$.

Все слагаемые в сумме $x_1 + \dots + x_k$ первых k членов набора больше или равны, чем $x_1 = x_{(1)}$ (напомним, что мы упорядочили числа набора по возрастанию), член x_{k+1} в точности равен медиане μ_x , а все слагаемые в сумме $x_{k+2} + \dots + x_{2k+1}$, включающей k последних членов набора больше или равны, чем μ_x . Поэтому $S \geq kx_1 + (k+1)\mu_x$. Поскольку $S = nM_x$, а $k = \frac{n-1}{2}$,

$$\text{мы получим: } \mu_x \leq \frac{2n}{n+1} \cdot M_x - \frac{n-1}{n+1} \cdot x_{(1)}.$$

Для оценки с другой стороны отметим, что все слагаемые в сумме $x_1 + \dots + x_k$ первых k членов набора меньше или равны, чем μ_x , член x_{k+1} в точности равен медиане μ_x , а все слагаемые в сумме $x_{k+2} + \dots + x_{2k+1}$, включающей k последних членов набора меньше или равны, чем $x_n = x_{(n)}$. Поэтому $S \leq kx_n + (k+1)\mu_x$. Поскольку $S = nM_x$, а $k = \frac{n-1}{2}$, мы

$$\text{получим: } \mu_x \geq \frac{2n}{n+1} \cdot M_x - \frac{n-1}{n+1} \cdot x_{(n)}.$$

Аналогичным образом рассматривается и случай, когда $n = 2k$, т.е. набор состоит из чётного количества чисел. Тогда медиана μ_x является средним арифметическим k -го $(k+1)$ -го в порядке возрастания членов набора:

$$\mu_x = \frac{x_k + x_{k+1}}{2}. \quad \text{Запишем сумму } S \text{ всех чисел набора в виде:}$$

$$S = (x_1 + \dots + x_{k-1}) + (x_k + x_{k+1}) + (x_{k+2} + \dots + x_{2k}).$$

Все слагаемые в сумме $x_1 + \dots + x_{k-1}$ первых $k-1$ членов набора больше или равны, чем x_1 , сумма $x_k + x_{k+1}$ в точности равна $2\mu_x$, а все слагаемые в сумме $x_{k+2} + \dots + x_{2k}$, включающей $k-1$ последних членов набора больше или равны, чем μ_x (эти слагаемые не меньше x_{k+1} , которое, в свою очередь,

не меньше, чем среднее арифметическое $\frac{x_k + x_{k+1}}{2} = \mu_x$). Поэтому

$S \geq (k-1)x_1 + (k+1)\mu_x$. Поскольку $S = nM_x$, а $k = \frac{n}{2}$, мы получим:

$$\mu_x \leq \frac{2n}{n+2} \cdot M_x - \frac{n-2}{n+2} \cdot x_{(1)}.$$

Для оценки с другой стороны отметим, что все слагаемые в сумме $x_1 + \dots + x_{k-1}$ первых $k-1$ членов набора меньше или равны, чем μ_x , сумма $x_k + x_{k+1}$ в точности равна $2\mu_x$, а все слагаемые в сумме $x_{k+2} + \dots + x_{2k}$, включающей $k-1$ последних членов набора меньше или равны, чем x_n .

Поэтому $S \leq (k-1)x_n + (k+1)\mu_x$. Поскольку $S = nM_x$, а $k = \frac{n}{2}$, мы получим:

$$\mu_x \geq \frac{2n}{n+2} \cdot M_x - \frac{n-2}{n+2} \cdot x_{(n)}.$$

Замечание 1. Утверждение теоремы 2 неулучшаемо в том смысле, что существуют наборы, для которых оба доказанных двойных неравенства превращаются в точные равенства.

Например, если в случае нечётного $n = 2k + 1$ взять набор из k чисел 0 и $k+1$ чисел 1 (тогда медиана равна 1, среднее значение равно $\frac{k+1}{n} = \frac{n+1}{2n}$, наименьшее значение равно 0, наибольшее значение равно 1), то в точное равенство превратится правая часть первого неравенства. Если же взять набор из $k+1$ чисел 0 и k чисел 1 (тогда медиана равна 0, среднее значение равно $\frac{k}{n} = \frac{n-1}{2n}$, наименьшее значение равно 0, наибольшее значение равно 1), то в точное равенство превратится левая часть первого неравенства.

В случае чётного $n = 2k$ можно взять набор из $k-1$ чисел 0 и $k+1$ чисел 1 (тогда медиана равна 1, среднее значение $\frac{k+1}{n} = \frac{n+2}{2n}$, наименьшее значение равно 0, наибольшее значение равно 1), то в точное равенство превратится правая часть второго неравенства. Если же взять набор из $k+1$ чисел 0 и $k-1$ чисел 1 (тогда медиана равна 0, среднее значение равно $\frac{k-1}{n} = \frac{n-2}{2n}$, наименьшее значение равно 0, наибольшее значение равно 1), то в точное равенство превратится левая часть второго неравенства.

Замечание 2. Если набор $x = [x_1, x_2, \dots, x_n]$ состоит из неотрицательных чисел, то

$$\mu_x \leq 2M_x. \quad (3)$$

Доказательства этого факта основано на простом следствии из неравенств (1) и (2) для наборов из неотрицательных чисел:

$$\mu_x \leq \begin{cases} \frac{2n}{n+1} \cdot M_x, & \text{если } n \text{ нечётно,} \\ \frac{2n}{n+2} \cdot M_x, & \text{если } n \text{ чётно.} \end{cases}$$

Дроби $\frac{n}{n+1}$ и $\frac{n}{n+2}$ строго меньше 1, что влечёт (3). Если исключить вырожденный случай, когда набор состоит из одних нулей, верно строгое неравенство $\mu_x < 2M_x$.

С помощью теоремы 1 можно легко получить ещё одно неравенство. Обычно в учебниках по теории вероятностей именно его (точнее, его вероятностный аналог) называют неравенством Чебышева.

Теорема 3 (неравенство Чебышева для абсолютных отклонений от среднего). Для произвольного числового набора $x = [x_1, x_2, \dots, x_n]$ и любого положительного числа a справедливо неравенство $\nu(|x - M_x| \geq a) \leq \frac{D_x}{a^2}$, т.е. доля тех чисел набора, которые отклоняются от среднего значения на величину a или больше, не превосходит $\frac{D_x}{a^2}$.

Доказательство. Рассмотрим новый набор из n чисел, полученных из чисел исходного набора по формуле $y = (x - M_x)^2$:

$$y_1 = (x_1 - M_x)^2, y_2 = (x_2 - M_x)^2, \dots, y_n = (x_n - M_x)^2.$$

Эти числа, очевидно, неотрицательны и потому к ним применимо первое неравенство Чебышева, доказанное в Теореме 1. При этом в качестве числа a мы возьмём число a^2 :

$$\nu(y \geq a^2) \leq \frac{M_y}{a^2}. \quad (4).$$

Поскольку $y = (x - M_x)^2$, неравенство $y \geq a^2$ равносильно неравенству $(x - M_x)^2 \geq a^2$, или, после извлечения квадратного корня, неравенству $|x - M_x| \geq a$. Соответственно, количество чисел набора $y = [y_1, y_2, \dots, y_n]$, удовлетворяющих неравенству $y \geq a^2$, совпадает с количеством чисел исходного набора $x = [x_1, x_2, \dots, x_n]$, удовлетворяющих неравенству $|x - M_x| \geq a$:

$$N(y \geq a^2) = N(|x - M_x| \geq a) \Leftrightarrow \nu(y \geq a^2) = \nu(|x - M_x| \geq a).$$

Кроме того, среднее значение M_y , равное по определению $\frac{y_1 + y_2 + \dots + y_n}{n}$, может быть переписано как $\frac{(x_1 - M_x)^2 + (x_2 - M_x)^2 + \dots + (x_n - M_x)^2}{n}$, т.е. на самом деле равно дисперсии D_x исходного набора. Заменяя в неравенстве (4) $\nu(y \geq a^2)$ на $\nu(|x - M_x| \geq a)$, а M_y на D_x , мы и получим требуемый результат.

Чтобы проиллюстрировать применение доказанных неравенств Чебышева, рассмотрим набор, приведённый в начале статьи. Нетрудно подсчитать (хотя это довольно утомительно), что $M_x = 32,67$, $D_x = 150,9811$. Поэтому в случае $a=45$ первое неравенство Чебышева примет вид: $0,2 \leq \frac{32,67}{45}$ или, что то же самое, $0,2 \leq 0,726$. Чтобы проиллюстрировать второе неравенство Чебышева возьмём $a=15$. Неравенству $|x - M_x| \geq 15 \Leftrightarrow x \leq 17,67$ или $x \geq 47,67$ удовлетворяет 25 чисел набора. Поскольку $D_x = 150,9811$, второе неравенство Чебышева примет вид $0,25 \leq \frac{150,9811}{15^2}$ или, что то же самое, $0,25 \leq 0,671027$. Конечно, эти числовые неравенства истинны. Но нельзя не отметить, что они довольно «грубые». Истинное значение неравенств Чебышева проявляется в теоретических рассмотрениях теории вероятностей и прежде всего при доказательстве знаменитого «закона больших чисел» (в связи с этим отметим, что $1 - \nu(x \geq a)$ фактически является эмпирической функцией распределения). Сам Чебышев доказал неравенство, называемое его именем, не в терминах описательной статистики, а на языке теории вероятностей, в связи с изучением предельного поведения среднего арифметического большого числа независимых случайных величин.

Сейчас мы рассмотрим одно применение неравенства Чебышева для анализа важной теоретической проблемы описательной статистики.

В школьных учебниках про среднее значение M_x обычно говорят, что числа набора *группируются* вокруг среднего значения, а дисперсия D_x характеризует величину рассеивания данных вокруг среднего арифметического. Однако обычные школьные определения среднего значения и дисперсии не дают никаких оснований для подобных утверждений. В особенности это относится к дисперсии, определение которой содержит смутные рассуждения о необходимости возведения в квадрат отклонений $x_i - M_x$ для того, чтобы «стереть» знаки у этих

отклонений (на самом деле для этого естественно просто взять модули отклонений, но как же тогда объяснить смысл введения дисперсии?).

Неравенство Чебышева позволяет немного прояснить этот вопрос. Прежде всего, отметим следующее соображение: естественно говорить, что числа набора группируются вокруг среднего значения, если лишь небольшая их доля более-менее сильно отклоняется от среднего значения, т.е. если a – предельно допустимое абсолютное отклонение, то число $\nu(|x - M_x| \geq a)$ достаточно мало. Теорема 3 для любого набора оценивает это число дробью $\frac{D_x}{a^2}$. Число a по смыслу рассматриваемой проблемы фиксировано.

Поэтому при малой дисперсии можно говорить, что числа набора группируются вокруг среднего значения. В этом смысле дисперсия может рассматриваться как мера рассеивания.

Ещё один аргумент в пользу этого утверждения может быть получен следующим образом. Пусть предельно допустимое отклонение от среднего (число a) равно $3\sigma_x$, где $\sigma_x = \sqrt{D_x}$ – среднее квадратичное отклонение. Тогда второе неравенство Чебышева примет вид:

$$\nu(|x - M_x| \geq 3\sigma_x) \leq \frac{1}{9},$$

т.е. для *любого* набора количество чисел, отклоняющихся от среднего значения больше (не строго), чем на три средних квадратичных отклонения, не превосходит $\frac{1}{9} \approx 11\%$ от общего количества чисел в наборе, т.е. по меньшей мере 89% всех чисел *любого* набора расположены в интервале $(M_x - 3\sigma_x; M_x + 3\sigma_x)$. Это одна из форм знаменитого правила «трёх сигм». Для конкретных наборов эта доля может быть не 89%, а гораздо выше. Например, для набора, приведённого в начале статьи, как мы уже отмечали, $M_x = 32,67$, $D_x = 150,9811$. Поэтому $\sigma_x \approx 12,29$, так что интервал $(M_x - 3\sigma_x; M_x + 3\sigma_x) = (-4,19; 69,53)$. В него попадает 100% чисел набора.

Если предельно допустимое отклонение от среднего характеризовать не абсолютным значением (числом a), а достаточно малой долей $\varepsilon > 0$ от среднего значения, т.е. считать, что $a = \varepsilon \cdot |M_x|$, то (при $M_x \neq 0$) Теорема 3 примет вид:

$$\nu(|x - M_x| \geq \varepsilon \cdot |M_x|) \leq \frac{1}{\varepsilon^2} \cdot \frac{D_x}{M_x^2}.$$

В этом случае мы можем говорить, что числа набора группируются вокруг среднего значения, если мало число $\frac{D_x}{M_x^2}$ или, что то же самое, мало число

$\frac{\sigma_x}{|M_x|}$. Дробь $\frac{\sigma_x}{|M_x|}$ называется коэффициентом вариации – это ещё одна важная мера разброса чисел набора.

Ещё один подход к определению меры положения числового набора, при котором естественно появляются медиана, среднее арифметическое и дисперсия, а также проясняется связь между средним арифметическим и дисперсией, можно найти в статье авторов [4]. Но истинный смысл введения понятий среднего значения и дисперсии можно объяснить только с помощью предельных теорем теории вероятностей (закона больших чисел и, главным образом, центральной предельной теоремы).

Литература

1. Б.В.Гнеденко, А.Я.Хинчин. Элементарное введение в теорию вероятностей. Изд. 10-е, исправленное. М., УРСС, 2003.
2. Воспитание патриотизма. *Математика*, 2011, №7, стр.4-5.
3. Математика: Алгебра. Функции. Анализ данных: Учебник для 9 класса общеобразовательных учреждений / [Г.В.Дорофеев, С.Б.Суворова, Е.А.Бунимович и др.]; под. ред. Г.В.Дорофеева. – М.: Просвещение, 2005.
4. Г.И.Фалин, А.И.Фалин. Экстремальные свойства среднего значения и медианы. *Математика в школе*, 2011 (в печати).