МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный университет имени М. В. Ломоносова»
МГУ

Факультет космических исследований (ФКИ)

Научно-исследовательская работа по теме
Исследование асимптотических свойств оценок с помощью понятия дефект

Студент: Мартыненко Даниил Юрьевич
Группа: 301
Руководитель: Бенинг Владимир Евгеньевич

Содержание

Введение 3

Определение дефекта 4

Вид дефекта при ассимптотическом разложении среднеквадратичной ошибки 4

Исследование пар оценок с помощью дефекта 5

Цена использования робастной оценки 5

Цена использования несмещенной оценки 6

Цена использования минимаксной оценки 7

Численное вычисление дефекта 8

Численное вычисление дефекта робастной оценки 8

Численное вычисление дефекта несмещенной оценки 9

Численное вычисление дефекта минимаксной оценки 9

Заключение 9

Источники 11

Приложение 12

Введение

Ни для кого не секрет, что статистика и статистические методы на сегодня находят большое число применений в огромном количестве сфер: от таких технических дисциплин, как экономика, социология до гуманитарных, например, история и филология.

И в большинстве этих дисциплин на практике достаточно часто возникают задачи с некоторой выборкой из независимых, одинаково распределенных величин $X_1...X_n$, где n достаточно велико, из какого-то семейства (чаще всего нормального), в котором необходимо оценить некоторые параметры, то есть на основе данных, полученных из наблюдений, получить величину, которая определенным образом (в зависимости от поставленной задачи), приближена к неизвестному параметру. Для этого строят оценки - измеримые функции от выборки.

Но множество возможных оценок достаточно велико, и каждая из них обладает своими особенностями. Некоторые из них более надежны, но менее точны, другие же обладают особыми свойствами, такими как несмещенность, состоятельность, нормальность и т.д. И достаточно часто возникает ситуация, в которой нам необходимо выбрать из двух оценок одну: оценку, более эффективно оценивающую искомый параметр, или оценку, которой требуется большее число наблюдений для достижения эффективности первой оценки, но обладающую нужным нам свойством.

В данной научно-исследовательской работе мы и научимся находить указанную выше "цену" (т.е. необходимое число дополнительных наблюдений) использования оценки с каким-либо асимптотическим свойством.

Цели научно-исследовательской работы:

- Исследовать дефект (та самая цена, указанная выше, более корректное определение мы дадим в дальнейшем) оценок, обладающих некоторыми особыми свойствами. В данной работе исследованы робастные, несмещенные и минимаксные оценки.
- Проверить истинность исследований выше на практических примерах

В самом ходе работы мы:

- Дадим четкое математическое определение дефекта, проанализируем его свойства
- Составим один из видов разложения среднеквадратичной функции риска
- С помощью асимптотического разложения функции риска напишем для некоторых случаев формулу дефекта в явном виде.
- Рассмотрим дефект робастной оценки
- Рассмотрим дефект несмещенной оценки
- Рассмотрим дефект минимаксной оценки
- Для каждого из этих трех случаев построим модели и найдем у них значения дефектов по определению.

Актуальность данной работы достаточно высока, т.к. в большом количестве задач математической статистики ставится вопрос выбора наилучшей в данной ситуации оценки, а дефект достаточно легко позволяет нам понять, насколько одна оценка будет хуже другой с точки зрения числа наблюдений.

Данный раздел в математической статистике уже достаточно активно изучался еще в 20м веке. Одной из первых работ является [1]. Из нее, в частности, взяты основные определения, формула асимптотического разложения функций риска для оценок и кратко описан ее вывод. Кроме того, именно в данной работе разобраны случаи сравнения трех оценок, которые мы также рассмотрим и проверим их практически (сами теоретические сравнения, для лучшего понимания механизма работы дефекта, также будут приведены).

Определение дефекта

Рассмотрим статистическую функцию $g(\theta)$ и две ее оценки(оценки $\hat{\theta}_1$ и $\hat{\theta}_2$). Будем считать, что оценка $\hat{\theta}_2$ "хуже" оценки $\hat{\theta}_1$."Хуже" может определяться по разному в зависимости от того, какую функцию риска мы будем использовать для подсчета ошибок. Например, в задачах статистического оценивания данная качественная мера обычно представляет собой среднеквадратичное отклонение оценки от $g(\theta)$, а при проверках статистических гипотез часто рассматривается мощность критериев. В данной начно-исследовательской работе мы будем использовать именно среднеквадратичное отклонение оценки $\mathrm{MSE} = E_{g(\theta)}[(\hat{\theta} - g(\theta))^2]$ как качественную меру.

Обозначим числом k_n количество наблюдений, которые требуются оценки $\hat{\theta}_2$ для достижения той же эффективности, что и оценке $\hat{\theta}_1$ на n наблюдениях. Тогда определим дефект как разность k_n и n.

$$d = k_n - n \tag{1}$$

Он несет в себе наглядный смысл необходимого количества дополнительных наблюдений, требующихся оценки $\hat{\theta}_2$ для достижения эффективности оценки $\hat{\theta}_1$.

Будем рассматривать асимптотический случай, то есть $n \to \infty$. Если будет существовать предел $\lim_{n\to\infty}(k_n - n)$, то будем называть его ассимптотическим дефектом.

Исторически же рассматривалось отношение k_n к n [Фишер(1925), Пирсон(1950)], так называемая асимптотическая относительная эффективность(в дальнейщем АОЭ)

$$e = \lim_{n\to\infty} \frac{k_n}{n} \tag{2}$$

Например, при $e = 5$ при достаточно большом количестве наблюдений n оценке $\hat{\theta}_2$ требуется приближенно в 5 раз больше наблюдений, чем оценке $\hat{\theta}_1$

Но при исследовании АОЭ возникали ситуации, когда оно было равно одному, но при этом возникала постоянная аддитивная добавка, которую АОЭ не учитывала. Поэтому и было введено понятие дефект для работы именно с такими случаями.

Сразу отметим, что если АОЕ$\neq$1, то дефект будет бесконечно большим, но из того, что АОЕ$=$1 не следует существования конечного ассимптотического дефекта(позже мы рассмотрим такой пример).

Асимптотическое разложение дефекта

Обозначим среднеквадратичную функцию риска оценки $\hat{\theta}_1$ при n наблюдениях как R_n, а среднеквадратичную функцию риска оценки $\hat{\theta}_2$ как R'_n. Рассмотрим k_n, такое что

$\lim_{n\to\infty}(R'_{k_n} - R_n) = 0$

Будем представлять их в виде

$$R_n = \frac{c(\theta)}{n^r} + \frac{a(\theta)}{n^{r+s}} + o\left(\frac{1}{n^{r+s}}\right) \tag{3}$$

$$R'_n = \frac{c(\theta)}{n^r} + \frac{b(\theta)}{n^{r+s}} + o\left(\frac{1}{n^{r+s}}\right), \tag{4}$$

где $a(\theta)$, $b(\theta)$, $c(\theta)$ - величины, не зависящие от n, $r > 0$, $s > 0$ - величины, характеризующие скорость убывания данных последовательностей. Коэффициент при первых членах будет одинаковым, так как иначе отношение среднеквадратичных ошибок не будет равно 1, и ассимптотический дефект существововать не будет.

Теперь, приравняем функции риска оценок R_n и R'_{k_n}:

$$\frac{1}{n^r}\left(1+\frac{a+o(1)}{cn^s}\right)=\frac{1}{{k_n}^r}\left(1+\frac{b+o(1)}{ck_n^s}\right)$$

В силу того, что $\lim_{n\to\infty}\dfrac{n}{k_n}=1$, $k_n=n+d_n$ и $d=\lim_{n\to\infty}d_n$ мы можем выразить d:

$$d=\lim_{n\to\infty}d_n=\frac{b-a}{c\,r}n^{1-s} \qquad (5)$$

Более подробный вывод данной формулы посмотреть в [1].Заметим, что при $s>1$ $d\to\infty$, а при $0<s<1$ $d\to 0$. Таким образом, формула дефекта в общем виде принимает следующий вид:

$$d=\begin{cases}\pm\infty & 0<s<1\\ \text{(b-a)/(cr)} & s=1\\ 0 & s>1\end{cases} \qquad (6)$$

Заметим, что d может принимать любые вещественные значения, хотя изначально мы определили его как целое число. Для этого доопределим k_n на всей вещественной прямой как случайную величину, принимающую значение $[k_n]$ с вероятностью $1-\{k_n\}$ и значение $[k_{n+1}]$ с вероятностью $\{k_n\}$. Тогда саму функцию риска R_n в случае нецелых k_n можно будет записать как

$$R_{k_n}=(1-\{k_n\})R_{[k_n]}+R_{[k_n]+1}\{k_n\} \qquad (7)$$

Таким образом, доопределив k_n на всех вещественных значениях, мы доопределили и d.

Кроме того, из формулы(6) можно получить некоторые свойства для d:

1. **рефлексивность**

   Если d является асимптотическим дефектом оценки $\hat{\theta}_2$ относительно оценки $\hat{\theta}_1$, то $-d$ является асимптотическим дефектом оценки $\hat{\theta}_1$ относительно оценки $\hat{\theta}_2$

2. **транзитивность**

   Если d_1 является асимптотическим дефектом оценки $\hat{\theta}_2$ относительно оценки $\hat{\theta}_1$, d_2 является асимптотическим дефектом оценки $\hat{\theta}_3$ относительно оценки $\hat{\theta}_2$, то асимптотический дефект оценки $\hat{\theta}_3$ относительно оценки $\hat{\theta}_1$ $d_3=d_1+d_2$

Исследование пар оценок с помощью дефекта

Приведенные ниже три исследования дефекта оценок уже были представлены в работе [1]. Подробно рассмотрим механизм их работы, а затем проверим их, найдя в каждом из случаев дефект оценок практически.

Цена использования робастной оценки

Пусть $(X_1...X_N)$ - независимые, одинаково распределенные (в дальнейшем н.о.р) величины из класса распределения $\mathcal{F}$, с математическим ожиданием μ и дисперсией σ^2. Тогда мы можем оценить дисперсию как

$$\hat{\theta}_n=\frac{1}{n}\sum_{i=1}^{N}(X_i-\mu)^2 \qquad (8)$$

Данная оценка несмещенная и состоятельная, значит для определения среднеквадратичной ошибки нам достаточно посчитать ее дисперсию, и она будет равна

$$D(\hat{\theta}_n) = \frac{\gamma\sigma^4}{n} \tag{9}$$

где $\gamma = \frac{\mu_4}{n} - 3$ - коэфффициент эксцессии, μ_4 - четвертый центральный момент распределения $\mathcal{F}$

Данная оценка $\hat{\theta}_n$ конечно же справедлива, только если матожидание μ корректно определено. Но если это не гарантировано, лучше воспользоваться следующей робастной оценкой:

$$\hat{\theta}'_n = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \overline{X})^2, \tag{10}$$

где

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{N} X_i \tag{11}$$

Она также несмещенная и состоятельная, значит среднеквадратичная ошибка снова будет равна дисперсии оценки

$$D(\hat{\theta}'_n) = \sigma^4 \frac{\gamma(n-1)+2}{n(n-1)} \tag{12}$$

Если мы должны сделать выбор из этих двух оценок, то решающим фактором будет цена(дополнительное количество наблюдений), которую мы будем платить за робастную оценку, чтобы достичь эффективности оценки (8)

Четвертый момент μ_4 удовлетворяет соотношению $\mu_4 \geq \sigma^4$ и следовательно, $\gamma+2 \geq 0$ (при этом $\gamma+2 > 0$ кроме одного случая, который мы позднее рассмотрим отдельно). Тогда, зная, что γ неотрицательно, функции риска $R(\hat{\theta}_n)$ и $R(\hat{\theta}'_n)$ удовлетворяют разложению (2) и (3) с $r = 1$, $c = \gamma + 2$, $a = 0$ и $b = 2$. Тогда АОЭ = 1 и асимптотический дефект $d = \frac{2}{\gamma+2}$

В случае принадлежности $(X_1 ... X_N)$ к классу нормальных распределений $\gamma = 0$ и $d = 1$. То есть оценке $\hat{\theta}'_n$ в данном случае потребуется всего на одно наблюдение больше, чтобы достичь той же эффективности, что и оценке $\hat{\theta}_n$.

Несмотря на то, что в случае нормального распределения мы получили только одно дополнительное наблюдение как плату за возможное ошибочное значение μ, дефект в общем случае может быть сколь угодно большим из-за того, что $2 + \gamma$ может сколь угодно близко находиться вблизи нуля. А 0 будет достигаться только в том случае, когда $\mathcal{F}$ таково, что мы с вероятностью $\frac{1}{2}$ попадаем в точки $\mu - \sigma$ и $\mu + \sigma$, и 0 в остальных. Тогда $R_1(\hat{\theta}_n) = 0$, так как $\hat{\theta}_n = \sigma^2$ с вероятностью 1, а $R_2(\hat{\theta}'_n) = \frac{2\sigma^2}{k(k-1)} \to 0$ при $k \to \infty$. Тогда АОЭ и дефект уходят на бесконечность, что говорит нам о том, что в данном случае более надежная оценка никогда не достигнет той же точности, что и другая.

На практике же достаточно часто возникают распределения, схожие с нормальным, но отличающимися от него более тяжелыми хвостами. В таких распределениях значение γ будет больше, чем в случае нормального распределения, т.е. $\gamma > 0$, и тогда дефект будет величиной меньшей единицы, то есть цена за использование робастной оценки будет достаточно мала.

Цена использования несмещенной оценки

Аналогично с первой задачей, рассмотрим н.о.р величины $(X_1 ... X_N)$ из семейства распределения F. Как первую оценку также возьмем

$$\hat{\theta}'_n = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \overline{X})^2, \tag{13}$$

а в качестве второй используем

$$\hat{\theta}_n = \frac{1}{n-c}\sum_{i=1}^{N}(X_i - \overline{X})^2, \tag{14}$$

При $c \neq -1$ она уже будет смещенной, но квадратичная ошибка может быть меньше, чем у первой оценки. Обозначим первую оценку как $\hat{\theta}_{-1}$, а вторую как $\hat{\theta}_{-c}$. Тогда

$$R_n = E(\hat{\theta}_c - \sigma^2)^2 = \frac{\sigma^4}{n(n+c)^2}\{(n-1)((\gamma+2)(n-1)+2) + n(c+1)^2\} \tag{15}$$

$$R_n' = E(\hat{\theta}_c - \sigma^2)^2 = \sigma\{\frac{\gamma+2}{n} + \frac{(c+1)^2 - 2\gamma - 2 - 2c(\gamma+2)}{n^2} + o(\frac{1}{n^2}) \tag{16}$$

В качестве примера рассмотрим $\hat{\theta}_0$. Тогда квадратичные ошибки удовлетворяют(3) и (4) с $r = 1$, $c = \gamma + 2$, $a = 3 - 2\gamma$ и $b = 2$, и тогда

$$d = \frac{(2\gamma+3)}{\gamma+2} \tag{17}$$

Оценка $\hat{\theta}_0$ будет лучше оценки $\hat{\theta}_1$ при $\gamma > \frac{1}{2}$, однако при $\gamma < \frac{1}{2}$, наоборот, $\hat{\theta}_1$ будет лучше $\hat{\theta}_0$

В случае нормальных распределений $\gamma = 0$ и $d = \frac{3}{2}$, то есть для достижения не меньшей точности несмещенной оценке будет нужно на $\frac{3}{2}$ наблюдения больше. Наибольший же выигрыш при различных значениях с достигается при $c = 1$, тогда $d = 2$, и для достижения той же точности нам будут необходимы 2 дополнительных измерения.

Цена использования минимаксной оценки

Пусть X - случайная величина, имеющая вид биномиального распределения с числом наблюдений n и параметром p

Рассмотрим две ее оценки:

$$\hat{\theta}_1 = \frac{X}{n} \tag{18}$$

$$\hat{\theta}_2 = \frac{\sqrt{n}}{1+\sqrt{n}}\frac{X}{n} + \frac{1}{2(1+\sqrt{n})} \tag{19}$$

Первая оценка будет несмещенной, вторая - минимаксной

Среднеквадратичные ошибки этих оценок будут соответственно равны:

$$R_1 = \frac{pq}{n} \tag{20}$$

$$R_2 = \frac{1}{4(1+\sqrt{n})^2} \tag{21}$$

При $p \neq \frac{1}{2}$ АОЭ оценки $\hat{\theta}_2$ относительно оценки $\hat{\theta}_1$ будет число $\frac{pq}{4} < 1$, тогда асимптотический дефект $d \to \infty$. Ситуация же, в которой $p = \frac{1}{2}$, будет кардинально отличаться.

АОЭ $= 1$ и $R_2 < R_1$

При подсчете асимптотического дефекта получим, что $d = 2n^{\frac{1}{2}} + 1 \to \infty$, то есть в данном случае уже для оценки (1) необходимо бесконечно большое число наблюдений для достижения эффективности оценки (2). Кроме того, это показывает нам, что для существования d недостаточно того, что АОЭ $= 1$

Численное вычисление дефекта

Для каждой из задач мы будем вычислять значения дефекта для двух оценок(для краткости в каждом из случаев будем называть их оценка (1) и оценка (2)) по следующему алгоритму:

1. Построим 10000 выборок распределения(в самих задачах укажем, какого именно)
2. Вычислим оценку (1).
3. Для оценки (2) сгенерируем множество наблюдений с числом выборок, полученных из аналитического исследования с окрестностью 5 наблюдений..
4. На каждом из них построим оценку (2).
5. Для каждой из построенных оценок найдем среднеквадратичную ошибку.
6. Сравним среднеквадратичные ошибки, полученные в оценке (1) с каждой из среднеквадратичных ошибок, полученных в оценке (2), посчитаем их разность
7. Найдем среди этих разностей наименьшую по абсолютному значению.
8. Повторим пункты 1-7 еще 20 раз.
9. Увеличим выборку до 40000 наблюдений и повторим для нее пункты 1-8. Этим мы покажем, что при существовании конечного асимптотического дефекта он не будет зависеть от размеров выборки.
10. Графически изобразим зависимость разности значений ошибок оценок от числа наблюдений оценки (2)

Все модели созданы в среде программирования Python 3.9.0.

Численное вычисление дефекта робастной оценки

Найдем дефект оценки $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{N}(X_i - \mu)^2$ относительно оценки $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{N}(X_i - \overline{X})^2$ по описанному выше алгоритму.

Для этого возьмем выборку из 10000 нормальных распределений с математическим ожиданием $\mu = 2$ и дисперсией $\sigma^2 = 1$. Построим график(см. приложение стр. 12)

Модель во всех случаях действительно выдает нам, что оценка (1) стала эффективнее оценки (2) в момент, когда число наблюдений у второй оценки стало равно 10001. Таким образом, дефект $d = 1$, что совпало с результатом аналитического вычисления.

При построении модели из 40000 наблюдений(см приложение стр. 13) мы получаем аналогичный результат, т.е. дефект в данном случае действительно не будет зависеть от размеров выборки.

Численное вычисление дефекта несмещенной оценки

Проведем вычисления для оценок
$\hat{\theta}_{1n} = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$ и $\hat{\theta}_{2n} = \frac{1}{n+1}\sum_{i=1}^{N}(X_i - \overline{X})^2$.

Аналогично первой задаче, проверяем на выборке нормальных распределений с математическим ожиданием $\mu = 2$ и дисперсией $\sigma^2 = 1$ сначала из 10000 наблюдений, а затем из 40000.

Во всех случаях получим, что дефект $d = 2$, что опять-таки говорит нам о верности аналитических вычислений. Графики данных моделей представлены в приложении на страницах 14 и 15 соответственно.

Численное вычисление дефекта минимаксной оценки

$\hat{\theta}_1 = \frac{X}{n}$

$\hat{\theta}_2 = \frac{\sqrt{n}}{1+\sqrt{n}}\frac{X}{n} + \frac{1}{2(1+\sqrt{n})}$

Для данной задачи построим 4 модели, в которых будем рассматривать биномиальные распределения следующего вида:

1. из 10000 наблюдений с параметром p=1/2
2. из 40000 наблюдений с параметром p=1/2
3. из 10000 наблюдений с параметром p=1/3
4. из 40000 наблюдений с параметром p=1/3

При этом для моделей 1-2 (при $p = 1/2$) оценка (1) будет более эффективна, значит необходимо увеличивать размер выборки именно для оценки (2).

В первой модели оценка (2) достигнет оценки (1) лишь при выборке из 10023 наблюдений(см приложение стр. 16).

В модели (2), т.е. при увеличении выборки до 40000 дефект уже будет равен 401(см приложение стр. 17), что говорит нам о том, что он растет с ростом числа наблюдений, т.е. ассимптотический дефект будет бесконечно большим.

Для моделей 3-4 наиболее эффективной оценкой будет оценка (2), значит, необходимо увеличивать число наблюдений уже для первой.

В моделе 3 (см приложение стр. 18) и 4 (см приложение стр. 19) абсолютная разность ошибок достигает минимума, только если оценке (1) дать еще 1040 и 4577 наблюдений соответственно. Это уже числа порядка числа наблюдений, что говорит нам о неединичном значении АОЭ.

Заключение

В ходе данной научно-исследовательской работы мы

- Дали четкое определение дефекта в конечном и асимптотическом случае, ознакомились с его основными свойствами.
- Произвели разложение среднеквадратичной функции риска для дальнейшего его использования в оценивании дефекта.
- Рассмотрели случай с оценками (8) и (10) (цена использования робастной оценки) в общем случае, а затем и для конкретных распределений с конкретными параметрами, расчитали дефект в каждом из них.
- Рассмотрели случай с оценками (13) и (14) (цена использования несмещенной оценки) в общем случае, а затем и для конкретных распределений с конкретными параметрами, расчитали дефект в каждом из них.

- Рассмотрели случай с оценками (18) и (19) для биномиального распределения, в результате которого было выяснено, что при бесконечно больших значениях числа наблюдений дефект будет бесконечно большим.

- Построили математические модели для каждого из предыдущих случаев, на каждом из которых убедились в верности посчитанных в теоретическом случае значений дефектов.

  Все цели работы, представленные во введении, были выполнены.

  Результаты нашей работы носят достаточно актуальный характер, так как в самых различных сферах достаточно часто ставится задача о выборе той или иной оценки для наилучшего приближения статистической функции, мы же привели один из способов решения данной задачи.

Источники

1 Hodges J.L., Lehmann E.L. Deficiency // Annals of Mathematical Statistics. 1970. Vol. 41, № 5. Pp. 783–801.

2 Бенинг В.Е. О дефекте некоторых оценок, основанных на выборках случайного объема // Вестник ТвГУ. Серия: Прикладная математика. 2015. № 1. С. 5–14

Приложение

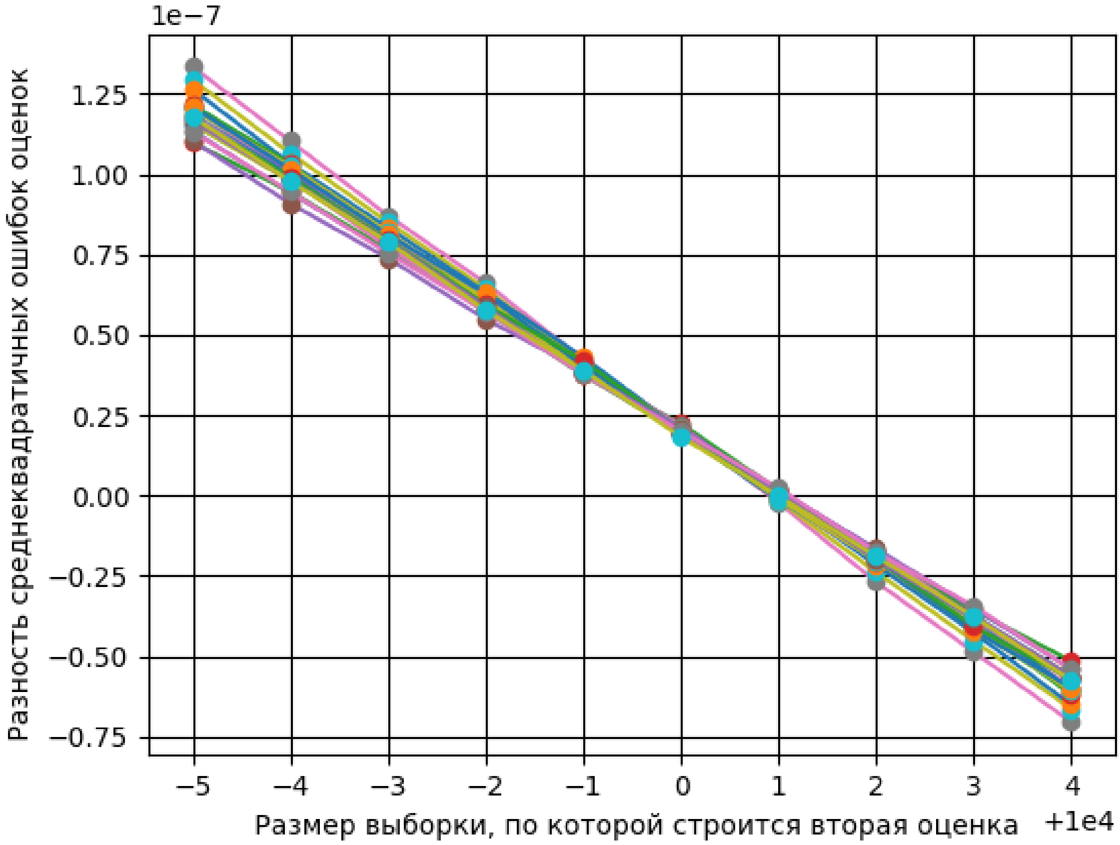

Рис. 0.1. График зависимости разности среднеквадратичных ошибок оценок от размера выборки робастной оценки, 10000 наблюдений

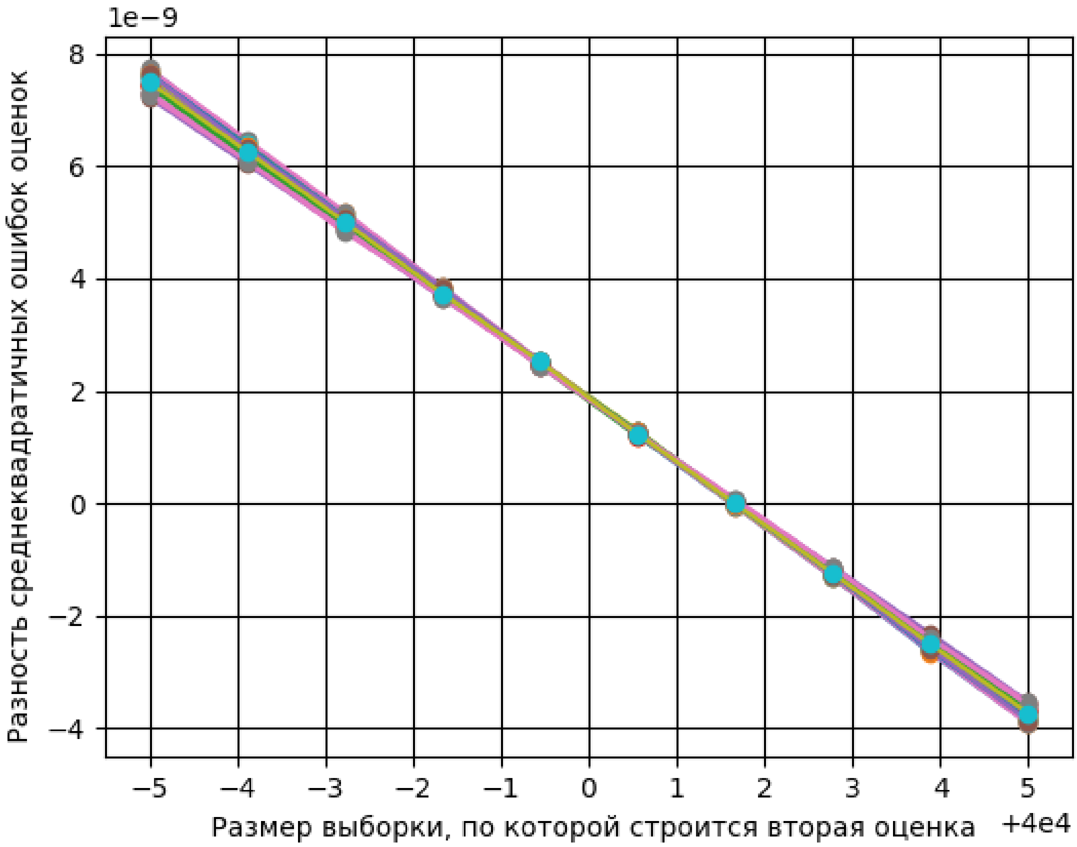

Рис. 0.2. График зависимости разности среднеквадратичных ошибок оценок от размера выборки робастной оценки, 40000 наблюдений

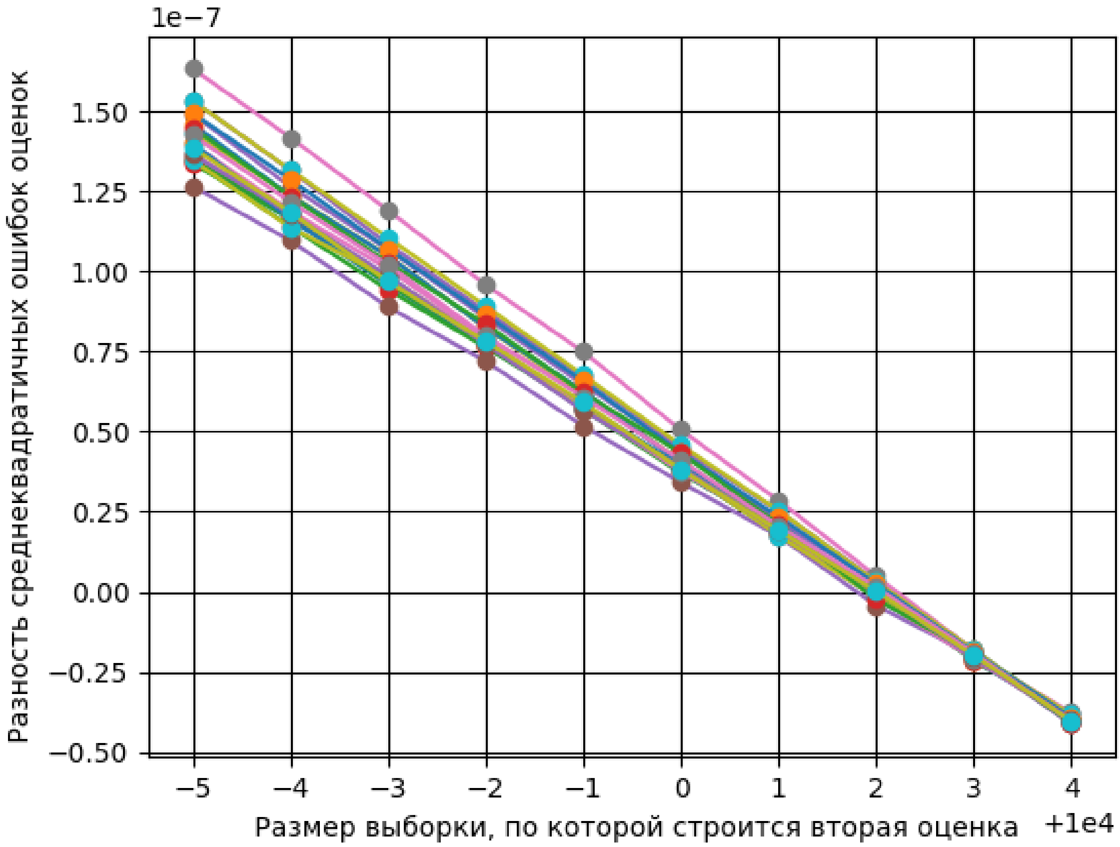

Рис. 0.3. График зависимости разности среднеквадратичных ошибок оценок от размера выборки несмещенной оценки, 10000 наблюдений

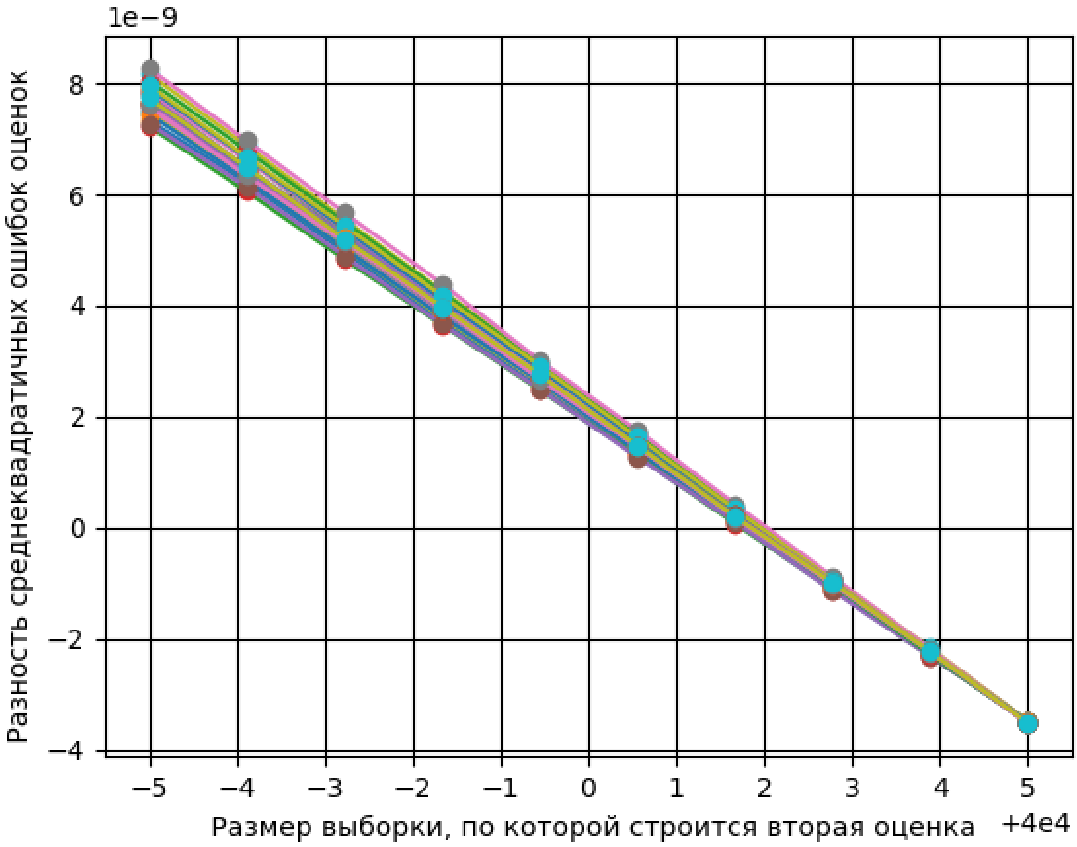

Рис. 0.4. График зависимости разности среднеквадратичных ошибок оценок от размера выборки несмещенной оценки, 40000 наблюдений

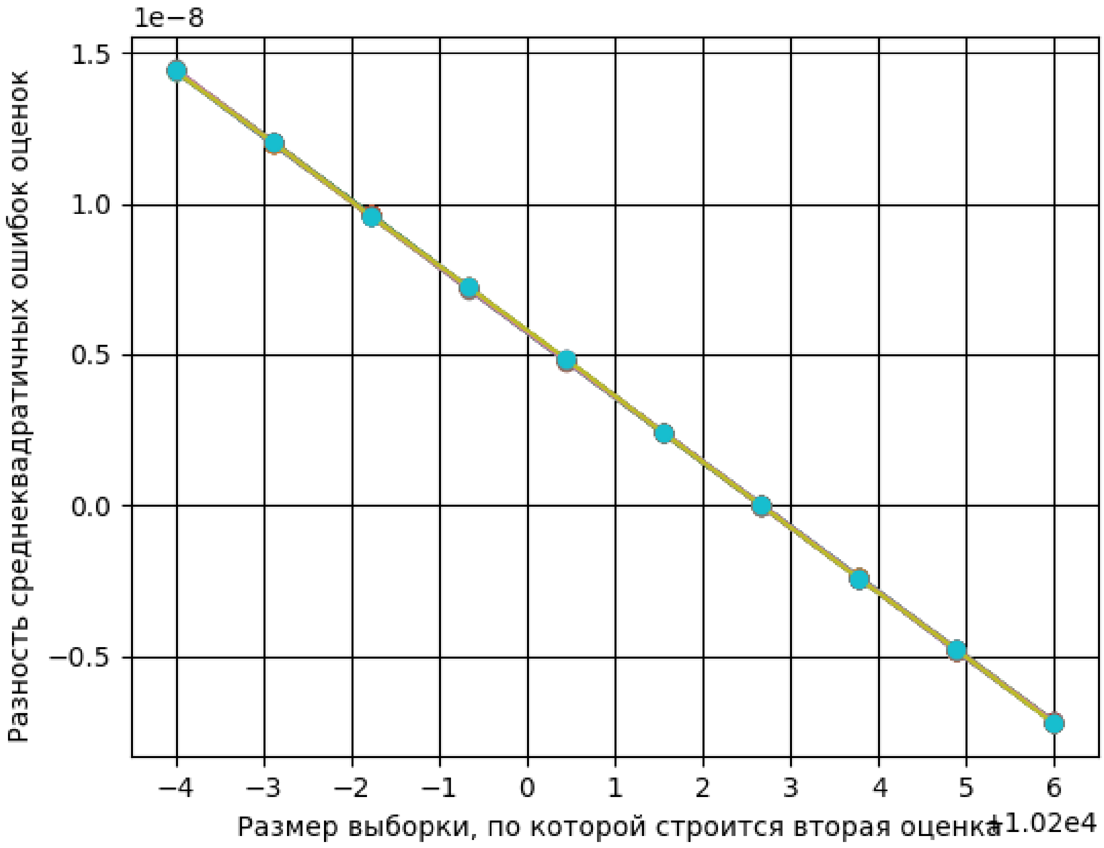

Рис. 0.5. График зависимости разности среднеквадратичных ошибок оценок от размера выборки минимаксной оценки, 10000 наблюдений, p = 1/2

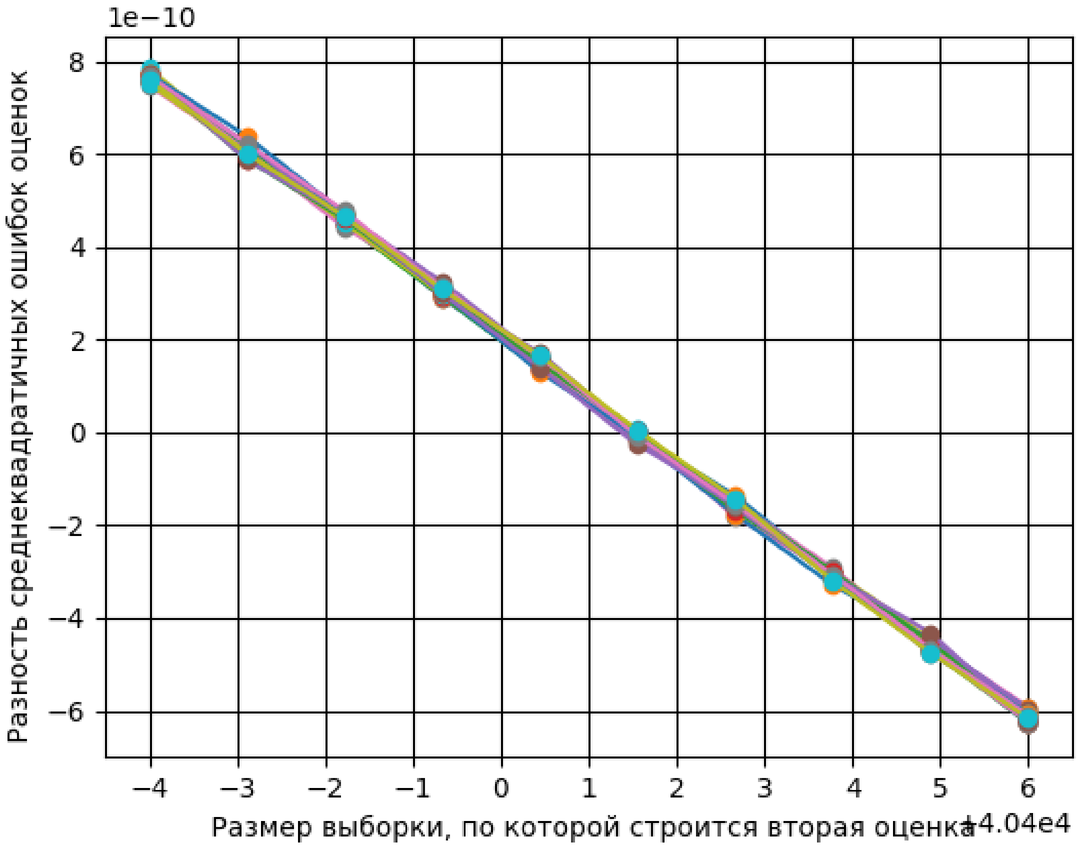

Рис. 0.6. График зависимости разности среднеквадратичных ошибок оценок от размера выборки минимаксной оценки, 40000 наблюдений, p = 1/2

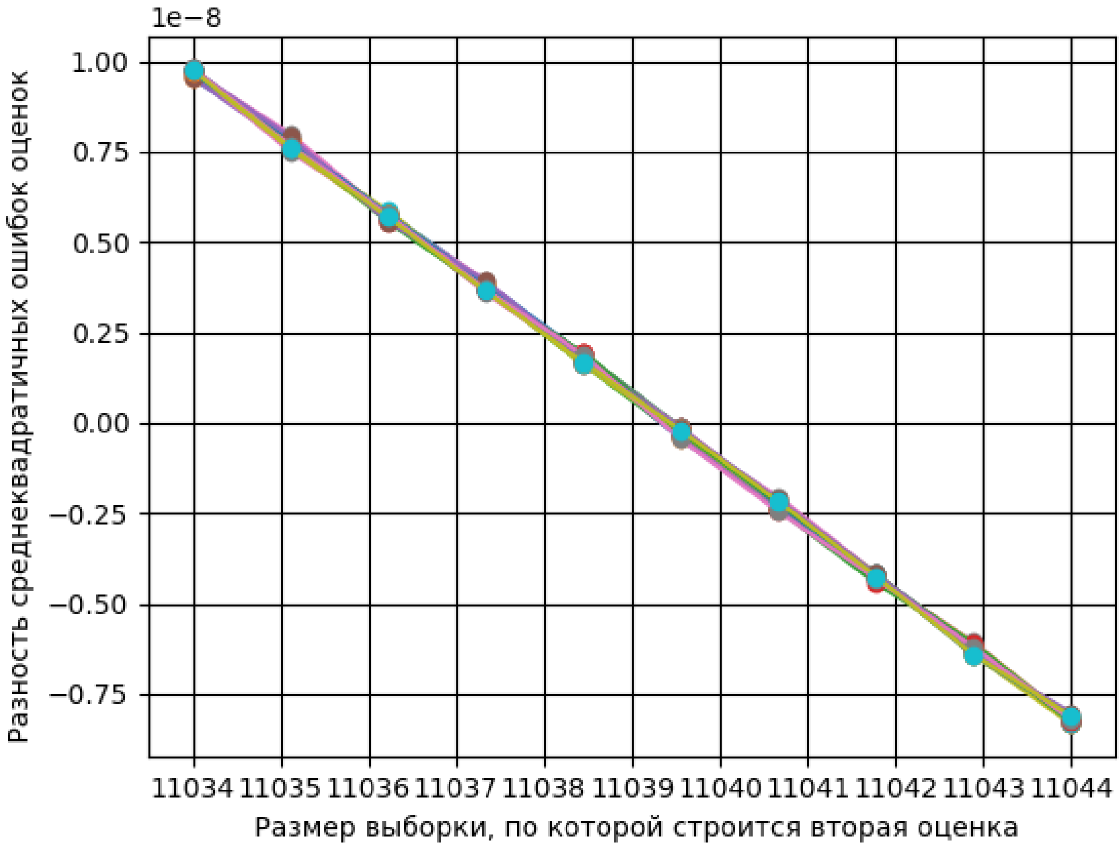

Рис. 0.7. График зависимости разности среднеквадратичных ошибок оценок от размера выборки минимаксной оценки, 10000 наблюдений, p = 1/3

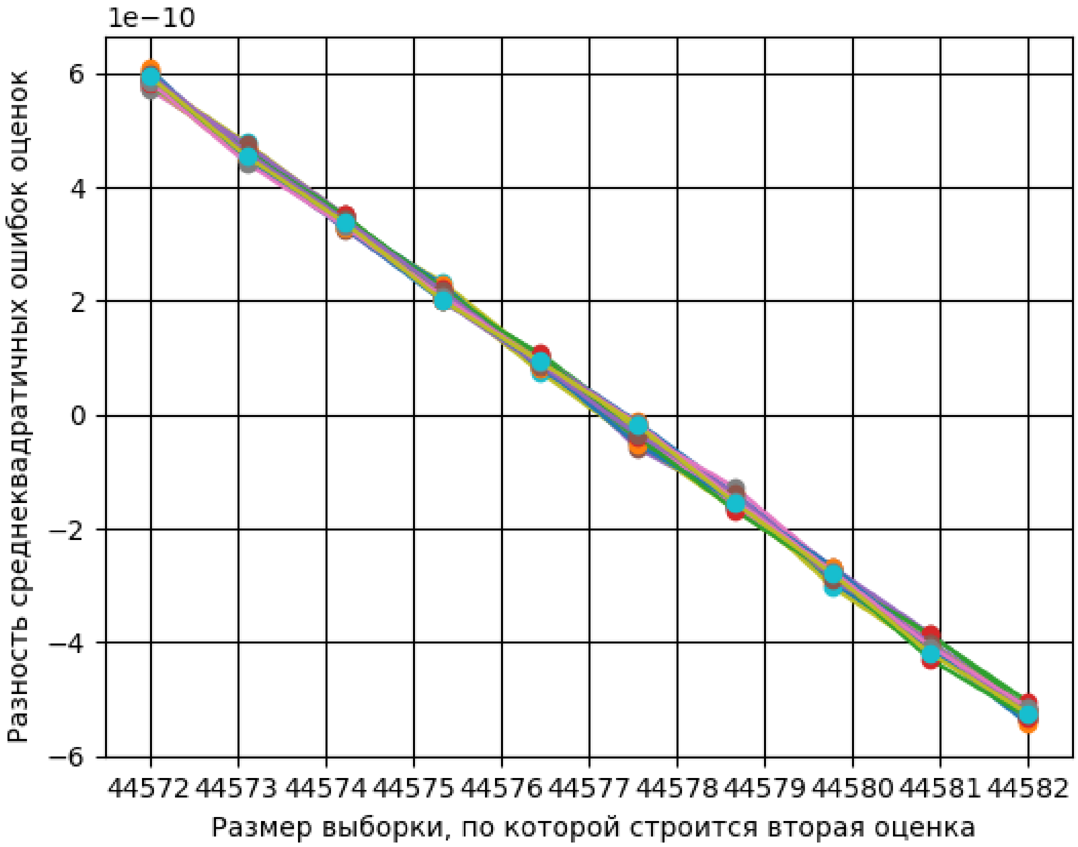

Рис. 0.8. График зависимости разности среднеквадратичных ошибок оценок от размера выборки минимаксной оценки, 40000 наблюдений, p = 1/3