Банк России

ОЦЕНКА ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА И ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ ДЛЯ ПРОИЗВОЛЬНЫХ РАСПРЕДЕЛЕНИЙ

Информационно-аналитический материал

Г. Гамбаров

Москва
2023

ОГЛАВЛЕНИЕ

Материал подготовлен Департаментом статистики.

107016, Москва, ул. Неглинная, 12, к. В
Официальный сайт Банка России: www.cbr.ru

Георгий Гамбаров
Банк России, Департамент статистики
д. э. н., доцент
gambarovgm@cbr.ru

Аннотация

В статье рассматриваются задачи построения доверительного интервала и выявления аномальных значений показателей, распределение которых не подчиняется нормальному распределению. Предложен метод односторонних дисперсий, позволяющий решать данные задачи в случае сильной асимметричности и произвольного эксцесса распределения.

Ключевые слова: аномальные значения показателя, правило трех сигм, метод Тьюки, асимметрия, эксцесс.
Key words: anomalous values, three sigma rule, Tukey method, asymmetry, kurtosis.

Работа посвящена совершенствованию статистических методов обнаружения ошибок в отчетных данных путем исключения аномальных значений, которые возникают в результате ошибок либо в результате сознательного искажения информации. Аномальными значениями некоторого показателя X считаются значения, не принадлежащие распределению значений отчетного показателя, а признаком аномальных значений является их существенное отличие от основной массы значений показателя.

В общем случае отчетность по некоторому показателю X на протяжении *T* периодов сдают *N* подотчетных организаций. Выбор подхода к решению задачи определения аномальных значений показателя X зависит от величин T и N. В первом разделе материала рассматривается случай одного периода (T = 1), во втором – случаи нескольких периодов и различного числа N.

1. Оценка доверительного интервала и выявление аномальных значений в случае одного периода

В данном разделе рассматривается ситуация, когда имеется достаточно большое число значений показателя одного периода (несколько десятков) и требуется определить доверительный интервал значений показателя, а также оценить, случайно ли отличие одного или нескольких значений от остальных значений показателя.

1.1. Оценка доверительного интервала

Наиболее распространенный подход к определению границ доверительного интервала значений показателя – использование среднеквадратического отклонения (σ) и *t*-критерия Стьюдента. Доверительный интервал определяется как:

$$(\overline{X} - t * \sigma,\ \overline{X} + t * \sigma),$$

где $\overline{X}$ – средняя величина;
t – коэффициент доверия;
σ – среднеквадратическое отклонение.

По правилу трех сигм (t = 3) подавляющая доля значений отличается от средней величины $\overline{X}$ менее чем на 3σ. Для нормального распределения вероятность нахождения значений показателя в диапазоне трех сигм – более 0,99 [3].

Достаточно широко распространен также метод Тьюки, в котором нижняя и верхняя границы доверительного интервала определяются как:

$$Q_{\text{нижняя}} = P_{25} - K_T * (P_{75} - P_{25}),$$

$$Q_{\text{верхняя}} = P_{75} + K_T * (P_{75} - P_{25}),$$

где P_{25} и P_{75} – 25-й и 75-й персентили ряда данных соответственно;
K_T – коэффициент Тьюки.

Для нормального распределения доверительный интервал метода Тьюки при $K_T = 1{,}5$ совпадает с доверительным интервалом правила трех сигм. Однако для распределений, отличающихся от нормального, использование правила трех сигм и метода Тьюки приводит к серьезным ошибкам.

Признаки нарушения нормальности распределения

Значения показателя распределены по нормальному закону, если их величина зависит от большого числа независимых факторов. Отклонения от нормального распределения часто обусловлены доминирующим влиянием на величину показателя одного или нескольких связанных факторов, что отражается в первую очередь на таких характеристиках распределения, как его асимметрия и эксцесс [3].

Основные признаки нарушения правила трех сигм – несимметричность распределения, измеряемая коэффициентом асимметрии, и отличие эксцесса распределения от эксцесса нормального распределения. Несимметричность распределения может свидетельствовать о том, что наиболее низкие значения показателя сформировались под влиянием одних факторов, а наиболее высокие значения – под воздействием других факторов.

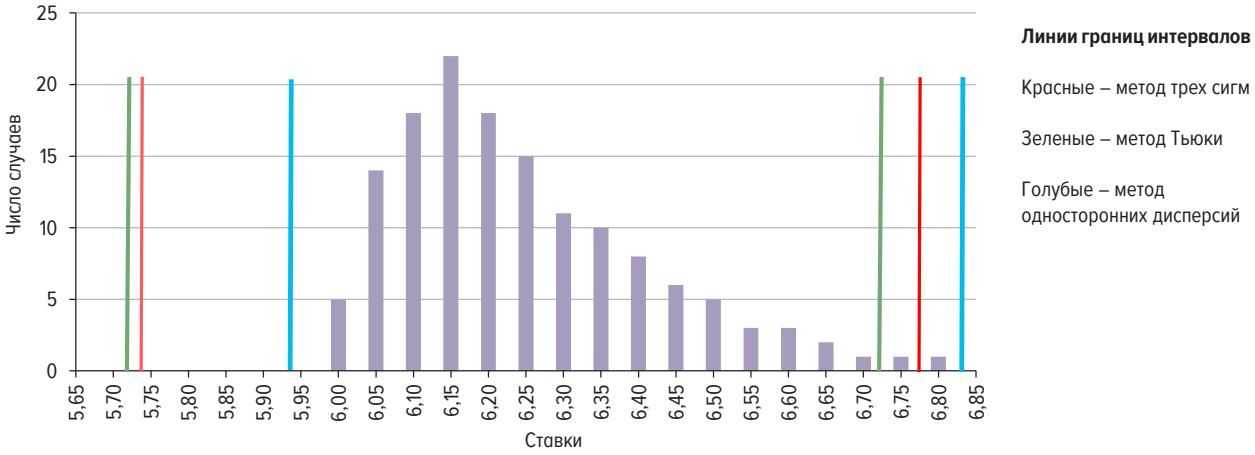
Например, несимметричным будет распределение ставок по выданным банками кредитам в условиях, когда сами банки получают средства по ставкам не ниже 6,0%. В этом случае надежных заемщиков можно кредитовать по ставкам в узком диапазоне (6,05–6,20%), а остальных заемщиков – по ставкам в зависимости от их кредитного рейтинга в более широком диапазоне (6,25–6,80%). Для первой категории заемщиков существенно наличие нижней границы ставок (6,0%), для второй категории наиболее важным фактором является оценка риска невозврата кредита.

Распределение с коротким левым хвостом и длинным правым, соответствующее приведенному примеру, представлено на рисунке 1. При этом эксцесс левой части распределения выше эксцесса нормального распределения, а правой части – ниже. Красные линии расположены по границам диапазона трех сигм: левая граница (3,67) на 3σ меньше средней величины ($\overline{X}$ = 6,25), правая граница (6,78) – на 3σ больше средней величины.

Для значений меньше средней величины диапазон трех сигм избыточно велик, для значений больше средней величины диапазон недостаточно велик, и часть значений, принадлежащих данному распределению, выходит за границы диапазона.

Аналогичная ситуация имеет место при использовании метода Тьюки: границы доверительного интервала, полученного данным методом, изображены на рисунке 1 зелеными линиями. Голубые линии на рисунке 1 отмечают границы интервала, полученного предлагаемым методом – методом односторонних дисперсий (см. описание ниже). Доверительный интервал данного метода включает все значения распределения, они расположены достаточно близко к его крайним значениям.

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ТРЕХ МЕТОДОВ ДЛЯ СКОШЕННОГО РАСПРЕДЕЛЕНИЯ С ДЛИННЫМ ПРАВЫМ ХВОСТОМ *Рис. 1*



Метод односторонних дисперсий

В методе односторонних дисперсий предлагается нижнюю границу доверительного интервала рассчитывать по тем показателям, значения которых ниже средней величины. Верхняя граница доверительного интервала рассчитывается по значениям показателя выше средней величины.

Для расчета нижней границы доверительного интервала вычисляются левая (односторонняя) дисперсия и левый (односторонний) эксцесс. Термин «односторонний» здесь используется потому, что дисперсия и эксцесс вычисляются при условии, что участвующие в расчете значения либо меньше, либо больше средней величины.

В первом случае на основании значений, меньших средней величины, вычисляется левая односторонняя дисперсия:

$$\sigma^2_{\text{левая}} = \sum_{i=1}^{N1}(X_i - \overline{\text{X}})^2) \,/\, \text{N1}, \text{ если } X_i < \overline{\text{X}}, \qquad (1)$$

где $\sigma^2_{\text{левая}}$ – левая односторонняя дисперсия;
$\overline{\text{X}}$ – общая средняя;
X_i – i-ое значение показателя меньше $\overline{\text{X}}$;
N1 – число элементов со значениями показателя меньше $\overline{\text{X}}$.

Аналогично левый односторонний эксцесс вычисляется по формуле:

$$E_{\text{левый}} = \frac{\sum_{i=1}^{N1}(X_i - \overline{\text{X}})^4}{N1 * (\sigma_{\text{левая}})^4} - 3, \text{ если } X_i < \overline{\text{X}}, \qquad (2)$$

где $\sigma_{\text{левая}}$ – квадратный корень из левой односторонней дисперсии (1).

Левая граница доверительного интервала вычисляется по формуле:

$$Q_{\text{левая}} = \overline{\text{X}} - \text{K} * U_{\text{левая}} * \sigma_{\text{левая}}, \qquad (3)$$

где K – коэффициент доверия, равный 3, как и в правиле трех сигм;

$U_{\text{левая}}$ – эмпирическая функция, зависящая от левого одностороннего эксцесса (2), которую предлагается вычислять по формуле:

$$U_{\text{левая}} = (0{,}65 * Ln(3 + E_{\text{левый}}) + 0{,}2)^{0{,}5}. \qquad (4)$$

Правая граница доверительного интервала вычисляется по формуле:

$$Q_{\text{правая}} = \overline{\text{X}} + \text{K} * U_{\text{правая}} * \sigma_{\text{правая}}, \qquad (5)$$

где K – коэффициент доверия, равный 3, как и в (3);
$U_{\text{правая}}$ – функция, зависящая от правого одностороннего эксцесса ($E_{\text{правый}}$), имеющая вид, аналогичный (4):

$$U_{\text{правая}} = (0{,}65 * Ln(3 + E_{\text{правый}}) + 0{,}2)^{0{,}5}, \qquad (6)$$

$\sigma_{\text{правая}}$ – квадратный корень из правой односторонней дисперсии, вычисляемой по формуле:

$$\sigma^2_{\text{правая}} = \sum_{i=1}^{N2}(X_i - \overline{\text{X}})^2) \ / \ \text{N2, если } X_i > \overline{\text{X}}, \qquad (7)$$

где N2 – число элементов со значениями показателя больше $\overline{X}$.

Для симметричных распределений левая и правая односторонние дисперсии равны; для несимметричных – отличаются тем сильнее, чем более несимметрично распределение.

Заметим, что формулы (4) и (6), полученные эмпирическим путем, не являются окончательными и могут быть улучшены.

Выбор в формулах границ доверительных интервалов (3) и (5) величины K, равной 3, соответствует утверждению, что в доверительном интервале находится не меньше 99% значений показателя.

Переход от общих дисперсий к односторонним нивелирует проблемы, связанные с асимметричностью распределений, а использование в формулах (3) и (5) функций $U_{\text{левая}}$ и $U_{\text{правая}}$ нивелирует проблемы, связанные с отличием эксцессов от эксцесса нормальных распределений.

Сравнение метода условных дисперсий с правилом трех сигм и методом Тьюки

Эффективность метода условных дисперсий иллюстрируется на ряде модельных распределений с различными значениями коэффициентов асимметрии и эксцесса. Приложение 1 содержит 20 распределений, два из которых симметричны, а остальные имеют положительную асимметрию, то есть имеют укороченный левый хвост распределения и удлиненный правый. Результаты расчетов дают основание для положительной оценки метода односторонних дисперсий.

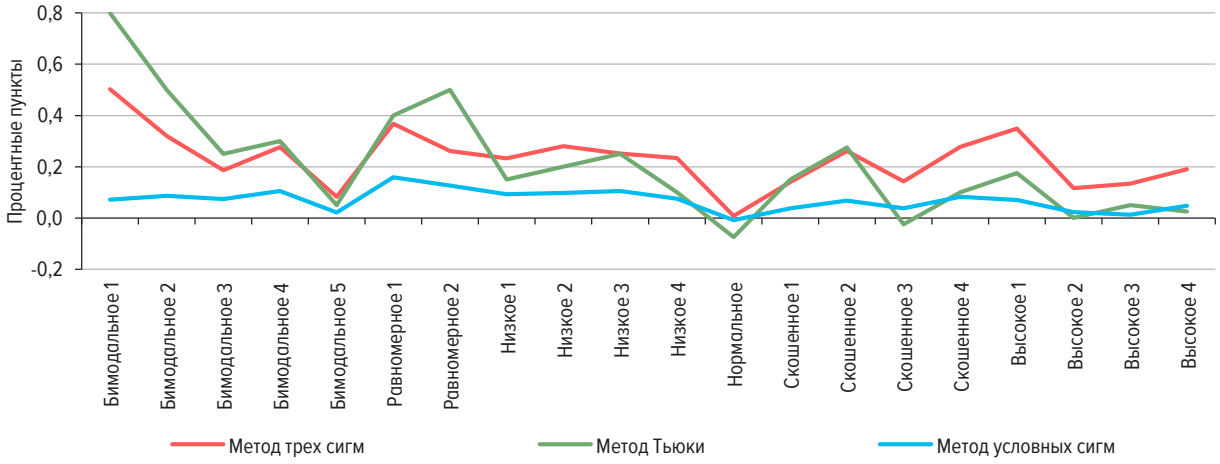
Предполагается, что все распределения в Приложении 1 не содержат аномальных значений, поэтому значения показателя за границами интервалов свидетельствуют о недостаточной ширине рассчитанного интервала, приводящей к ошибке I рода: значения, принадлежащие распределению, отнесены к аномальным.

В ситуации когда левая граница меньше всех значений показателя и находится на значительном расстоянии от минимального значения показателя, существует высокая вероятность ошибки II рода: аномальные значения не будут обнаружены в будущем. Аналогичная вероятность ошибки II рода имеет место, когда правая граница больше всех значений показателя и находится на значительном расстоянии от максимального значения показателя.

На рисунке 2 представлены значения разностей минимального значения распределения и разностей оценок нижних границ, полученных тремя методами: по правилу трех сигм, методом Тьюки [5] и методом односторонних дисперсий. Распределения расположены по мере возрастания их асимметрии и эксцесса. Если значения вне границ рассчитываемого интервала считать аномальными, то отрицательные величины на рисунке 2 свидетельствуют о недостаточной величине рассчитанного интервала, при которой имеет место ошибка I рода. Чем выше значения разностей, тем выше вероятность ошибки II рода.

Как видно на рисунке, два значения, принадлежащие распределению, отнесены к аномальным.

ОТКЛОНЕНИЯ НИЖНИХ ГРАНИЦ ИНТЕРВАЛОВ ЗНАЧЕНИЙ ОТ МИНИМАЛЬНЫХ ЗНАЧЕНИЙ *Рис. 2*



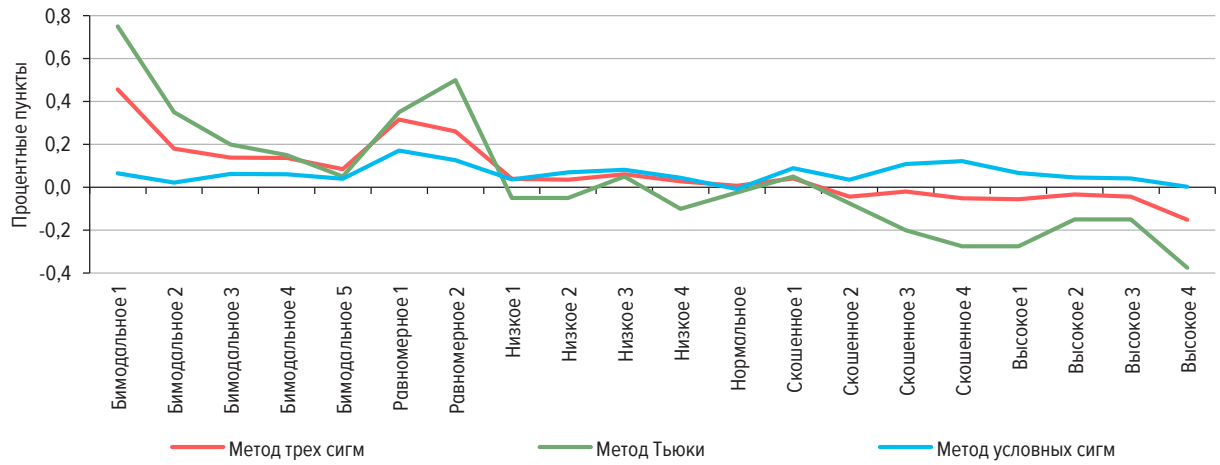
Рассматриваемые методы в случае коротких хвостов практически не допускают ошибок I рода: отсечение значений наблюдалось только два раза по методу Тьюки. Вероятность ошибки II рода заметно меньше у метода односторонних дисперсий. Рассчитанные по данному методу границы интервалов заметно ближе к минимальным значениям распределений.

На рисунке 3 представлены разности максимальных значений распределений и правых границ доверительных интервалов, полученных тремя методами.

По мере увеличения коэффициентов асимметрии и эксцесса метод трех сигм и метод Тьюки приводят к росту ошибки I рода – рассчитанные разности становятся отрицательными, что не наблюдается у границ, рассчитанных методом односторонних дисперсий. Ошибка II рода также меньше у метода условных дисперсий: границы интервалов, полученных данным методом, заметно ближе к границе изменения значений распределений.

Метод односторонних дисперсий дает адекватные результаты для самых разнообразных распределений (Приложение 1).

ОТКЛОНЕНИЯ ВЕРХНИХ ГРАНИЦ ИНТЕРВАЛОВ ЗНАЧЕНИЙ ОТ МАКСИМАЛЬНЫХ ЗНАЧЕНИЙ *Рис. 3*



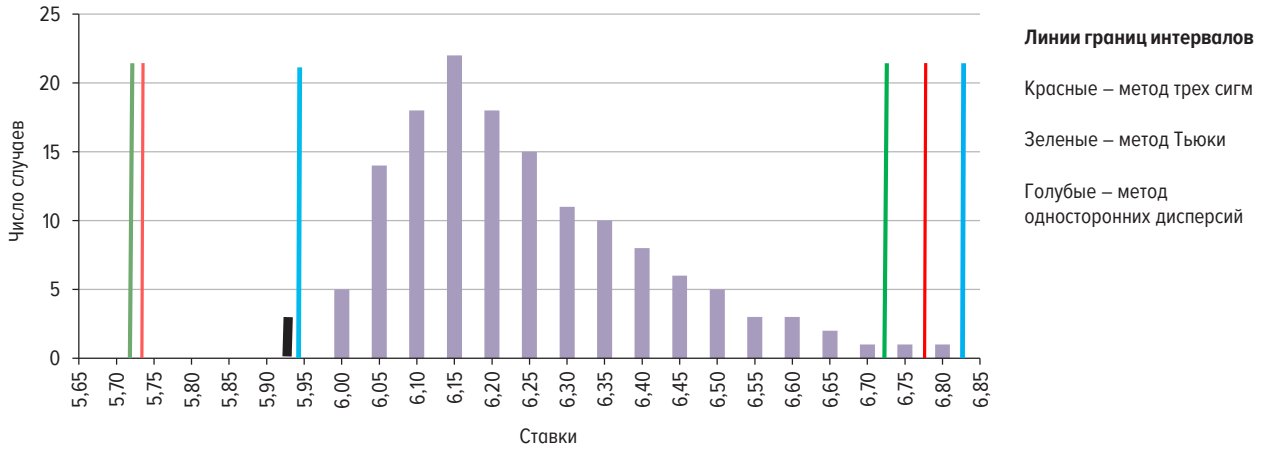
1.2. Определение аномальных значений показателя

Подходы к определению аномальных значений, использующие доверительные интервалы, основываются на предположении об однородности статистической совокупности, то есть сходстве процессов формирования значений показателя. В некоторых случаях распределение величин показателей позволяет с достаточно высокой вероятностью утверждать, что совокупность однородна. Например, в случае гладкого унимодального распределения, которое на практике встречается редко.

При принятии решения об аномальности одного или нескольких значений показателя проверяется гипотеза о его (их) принадлежности к той же генеральной совокупности, что и остальные значения. Для этого применяются критерии сравнения средних величин. В данных критериях используются выборочные дисперсии, которые, как было показано выше, для распределений, отличных от нормальных (особенно для несимметричных), далеко не всегда приводят к адекватным выводам.

Для иллюстрации данного утверждения вернемся к исходному распределению, представленному на рисунке 1. Как и на рисунке 1, красные линии на рисунке 4 отмечают границы доверительного интервала, построенного на основе общей дисперсии (5,73; 6,76) для коэффициента доверия, равного 3. Рассмотрим новое значение 5,92. На рисунке 4 оно изображено черным столбиком. Данное значение заметно отличается от остальных значений, хотя расположено внутри традиционного доверительного интервала.

СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ С ДЛИННЫМ ПРАВЫМ ХВОСТОМ И НОВОЕ ЗНАЧЕНИЕ ПОКАЗАТЕЛЯ *Рис. 4*



Для оценки аномальности нового значения 5,95 проверим гипотезу о его принадлежности к рассматриваемому распределению. Поскольку вторая совокупность представлена одним наблюдением, формула t-статистики для сравнения средних имеет вид:

$$t = \frac{\overline{X} - X}{\sigma} * \sqrt{\frac{N}{N+1}} = 1{,}53, \qquad (8)$$

где $\overline{X}$ – средняя величина исходного распределения, равная 6,25;
X – добавленное значение 5,95;
σ – выборочное среднеквадратическое отклонение исходного распределения, равное 0,170;
N – объем исходной совокупности, равный 143.

Значение t-статистики 1,59 меньше критического значения t-критерия на уровне значимости 0,05, равного 2. Это означает, что с вероятностью 95% нельзя утверждать, что отличие нового значения от старых значений статистически незначимо.

Если же в формуле (8) вместо σ использовать одностороннее среднеквадратическое отклонение, то получаем:

$$t = \frac{\overline{X} - X}{\sigma_{\text{левая}}} * \sqrt{\frac{N}{N+1}} = 3{,}23. \qquad (9)$$

Это означает, что с вероятностью 95% можно утверждать, что отличие нового значения 5,92 от старых значений статистически значимо.

Результат, полученный по формуле (9), представляется более адекватным, чем результат по формуле (8). Это позволяет рекомендовать использование односторонних дисперсий при проверке статистических гипотез сравнения средних.

2. Оценка доверительного интервала и выявление аномальных значений при наличии временных рядов показателя

Наличие отчетности за различные периоды времени, наряду с качественным содержательным анализом однородности совокупности, позволяет использовать статистические методы.

Организации, сдающие отчетность, могут различаться структурой связи своих существенных (внутренних) факторов функционирования, формирующих отчетный показатель. Сходство значений отчетного показателя двух организаций не гарантирует сходства механизмов их формирования. При изменении условий функционирования объектов значения отчетного показателя этих двух организаций в следующем периоде могут заметно отличаться.

В то же время сходная динамика отчетного показателя двух организаций позволяет говорить о сходстве факторов, формирующих данный показатель, то есть об однородности этих организаций в отношении данного показателя. В связи с этим представляется правомерным следующий подход к определению доверительного интервала.

На первом этапе динамические ряды каждой организации очищаются от тренда. Для этого из значений ряда каждой организации вычитаются значения его тренда соответствующего периода.

В случае если длина динамического ряда достаточно велика для построения доверительного интервала, можно использовать метод условных дисперсий.

При недостаточном количестве периодов следует объединять организации со сходной структурой факторов, формирующих отчетный показатель. С этой целью для каждого из вновь полученного временного ряда строится параметрически задаваемая модель, описывающая данный ряд. На практике в качестве моделей используются полиномы второй или третьей степени. При этом важно, чтобы ряды всех организаций описывались моделями с одним и тем же числом параметров.

По полученным наборам параметров проводится классификация организаций. Динамика показателей организаций, попавших в один класс, описывается сходными моделями, что позволяет считать похожими механизмы формирования отчетных показателей и, следовательно, сами организации можно считать однородными в отношении данного отчетного показателя.

Таким образом, все значения организаций, попавших в один класс, образуют совокупность, пригодную для формирования доверительного интервала и оценки аномальности отдельных значений показателя с помощью метода условных дисперсий.

Приложение

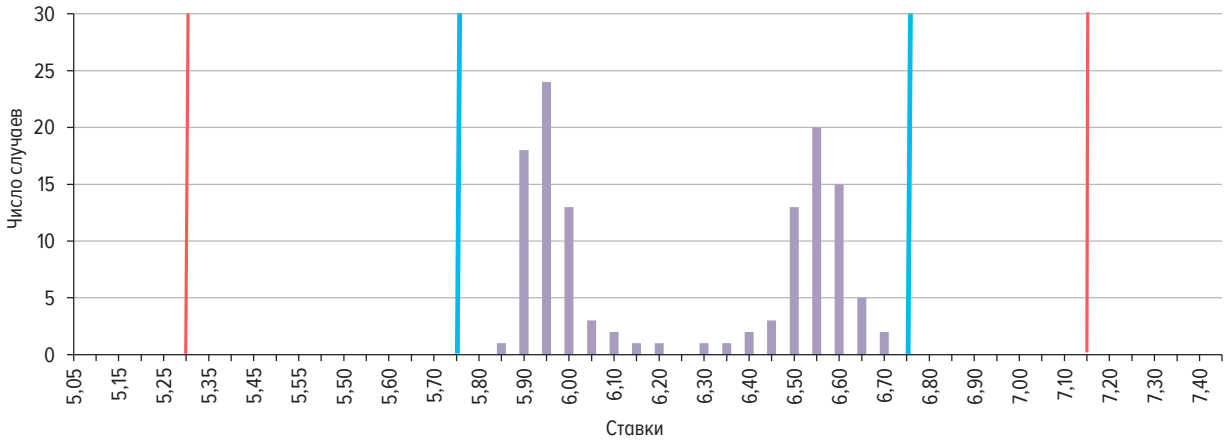
В данном приложении представлены графики 20 распределений с различными значениями коэффициентов асимметрии и эксцесса. На рисунках указаны границы интервалов, полученных методом трех сигм и методом условных дисперсий. Красными линиями отмечены границы интервалов, рассчитанных по методу трех сигм; голубыми – методом условных дисперсий, рассчитанных по формулам (4) и (5). В названии рисунков указаны значения коэффициентов асимметрии (*A*) и эксцесса (*E*), вычисленных соответственно по формулам:

$$A = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^3}{N * \sigma^3},$$

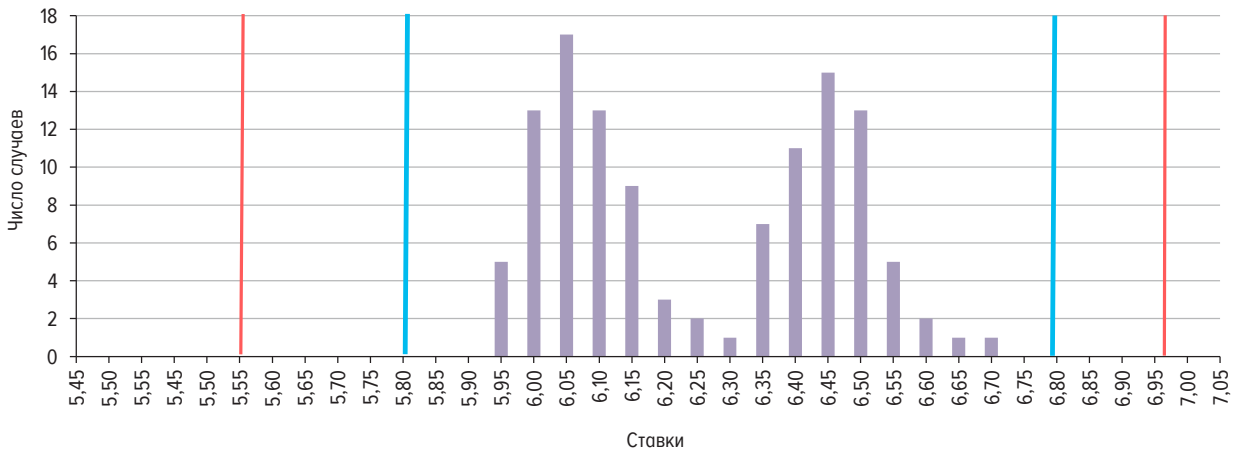
$$E = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^4}{N * \sigma^4},$$

где *N* – численность совокупности.

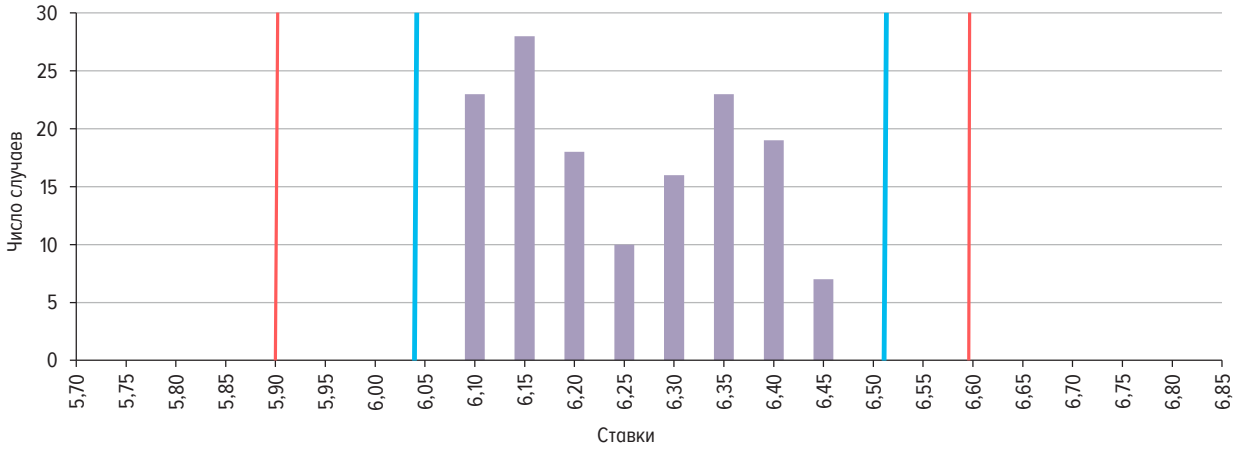
БИМОДАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 1 С АСИММЕТРИЕЙ A = 0,04 И ЭКСЦЕССОМ E = -1,81 *Рис. П-1*



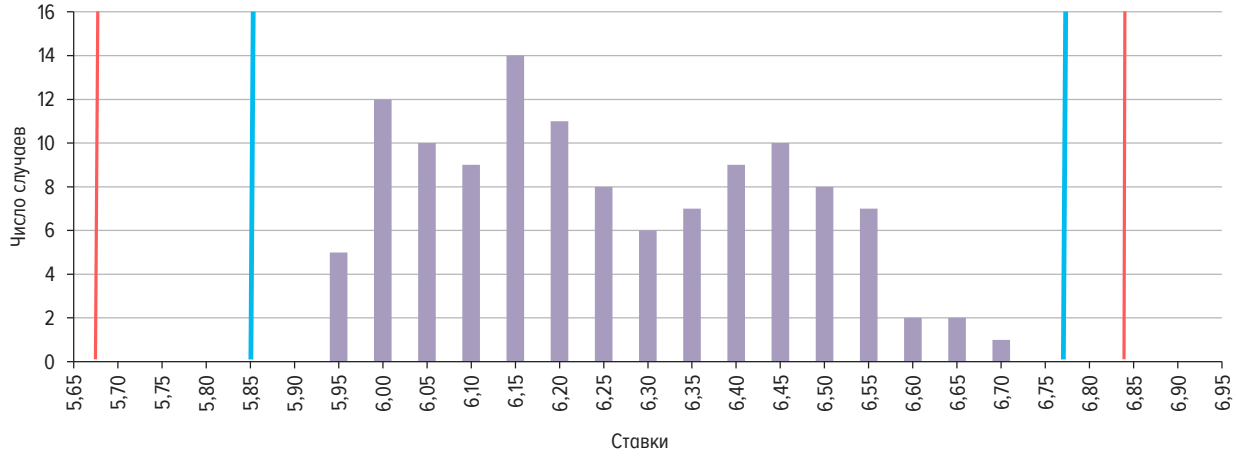
БИМОДАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 2 С АСИММЕТРИЕЙ A = 0,14 И ЭКСЦЕССОМ E = -1,38 *Рис. П-2*



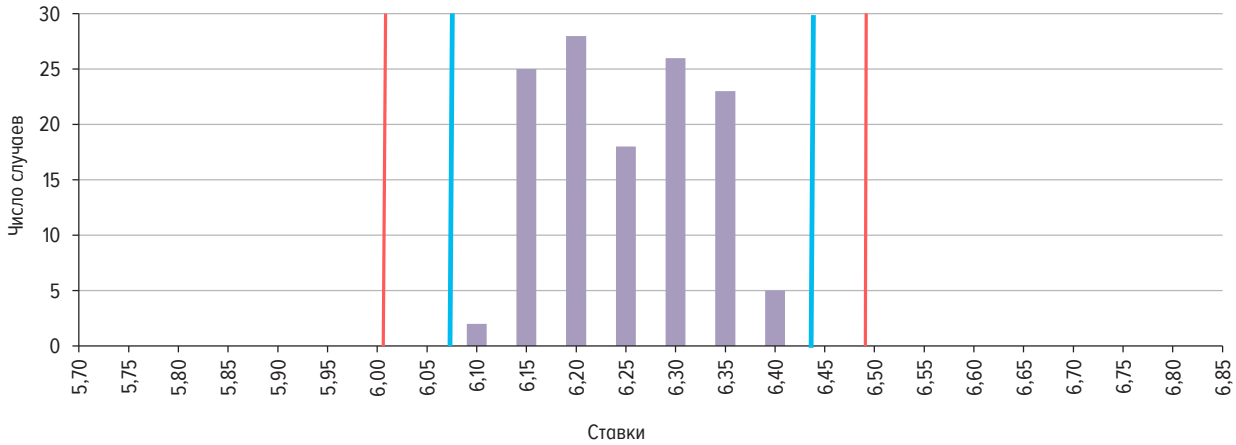
БИМОДАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 3 С АСИММЕТРИЕЙ A = 0,15 И ЭКСЦЕССОМ E = -1,29 *Рис. П-3*



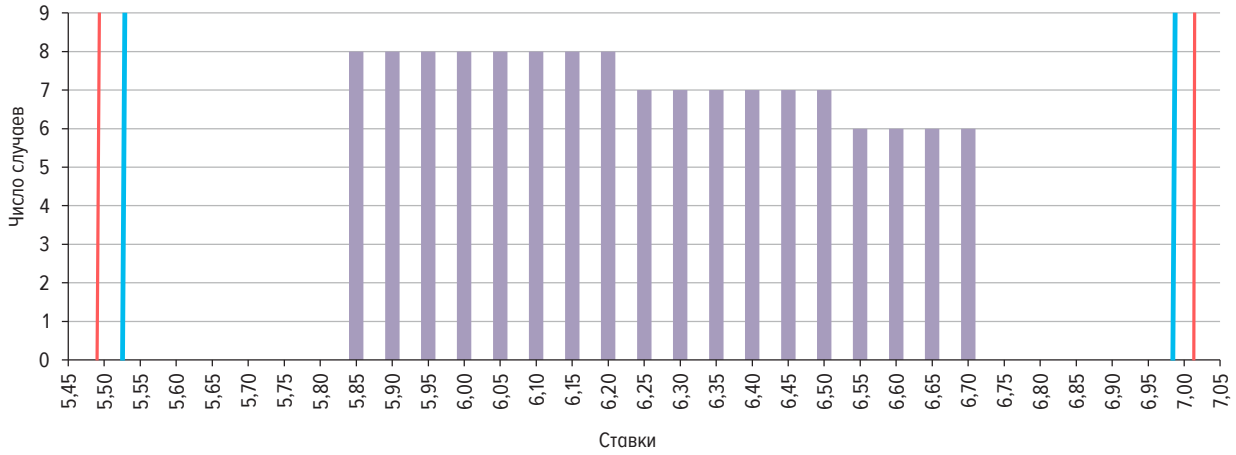
БИМОДАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 4 С АСИММЕТРИЕЙ A = 0,09 И ЭКСЦЕССОМ E = -1,03 *Рис. П-4*



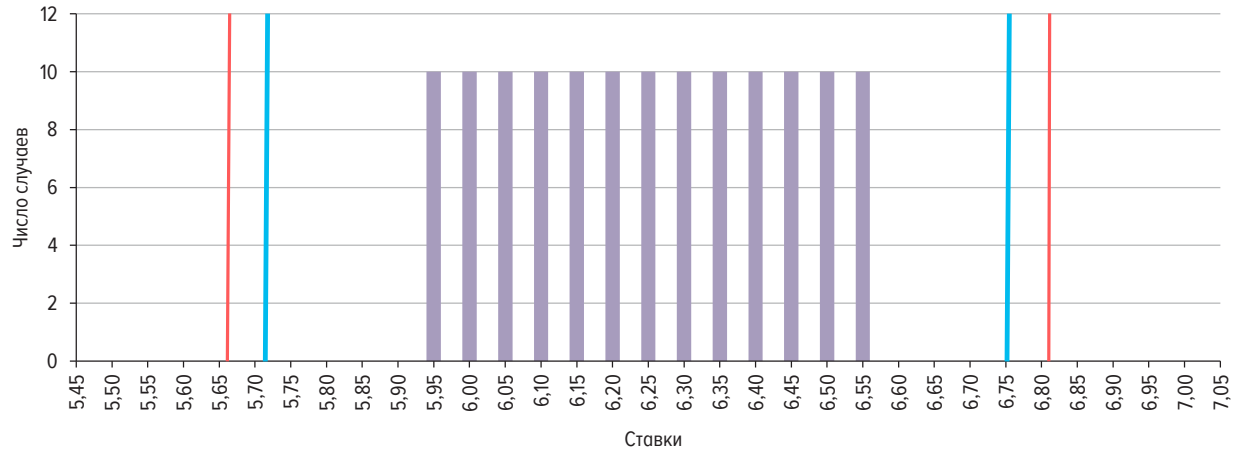
БИМОДАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 5 С АСИММЕТРИЕЙ A = 0,25 И ЭКСЦЕССОМ E = -2,58 *Рис. П-5*



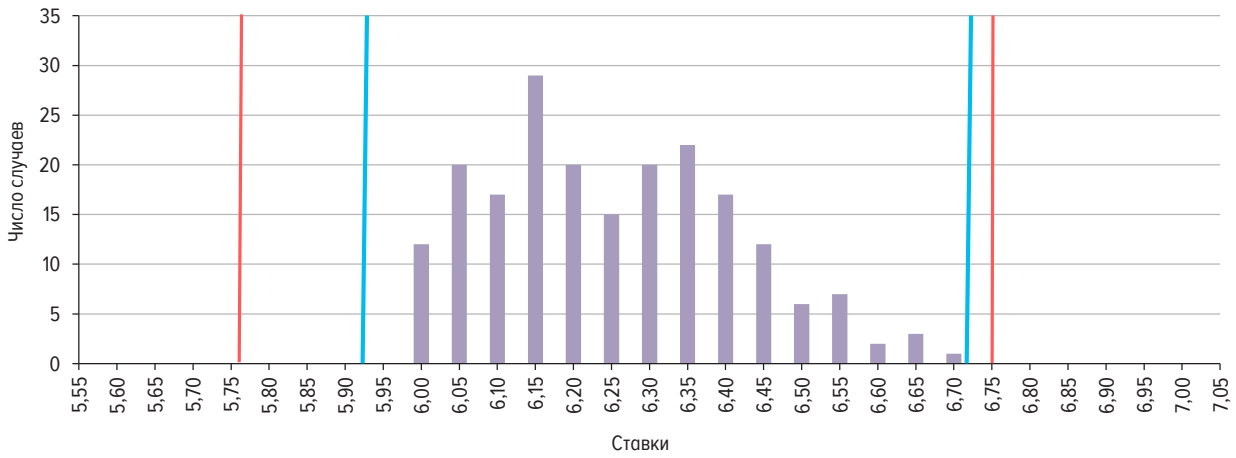
РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ 1 С АСИММЕТРИЕЙ A = 0,13 И ЭКСЦЕССОМ E = -0,90 *Рис. П-6*



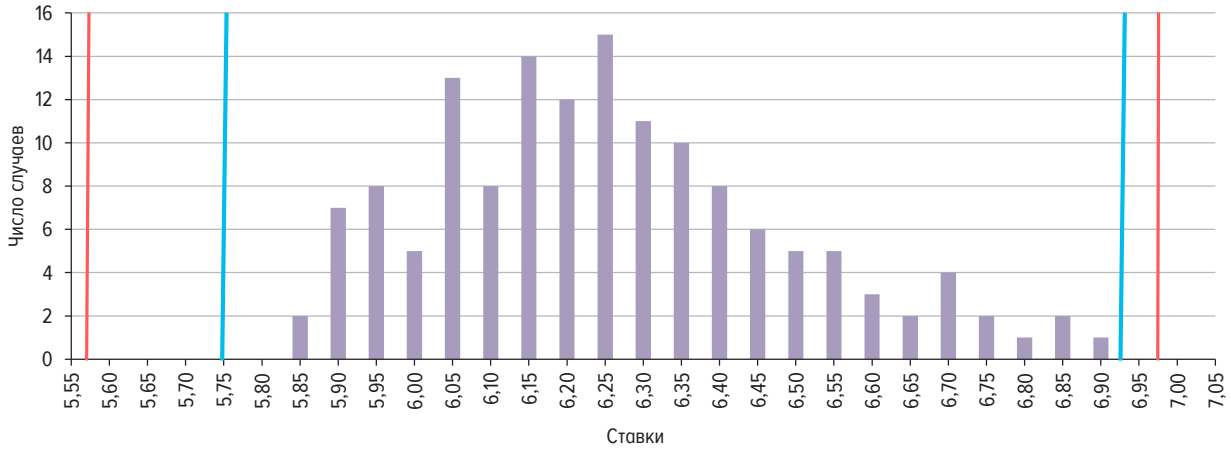
РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ 2 С АСИММЕТРИЕЙ A = 0,00 И ЭКСЦЕССОМ E = -0,84 *Рис. П-7*



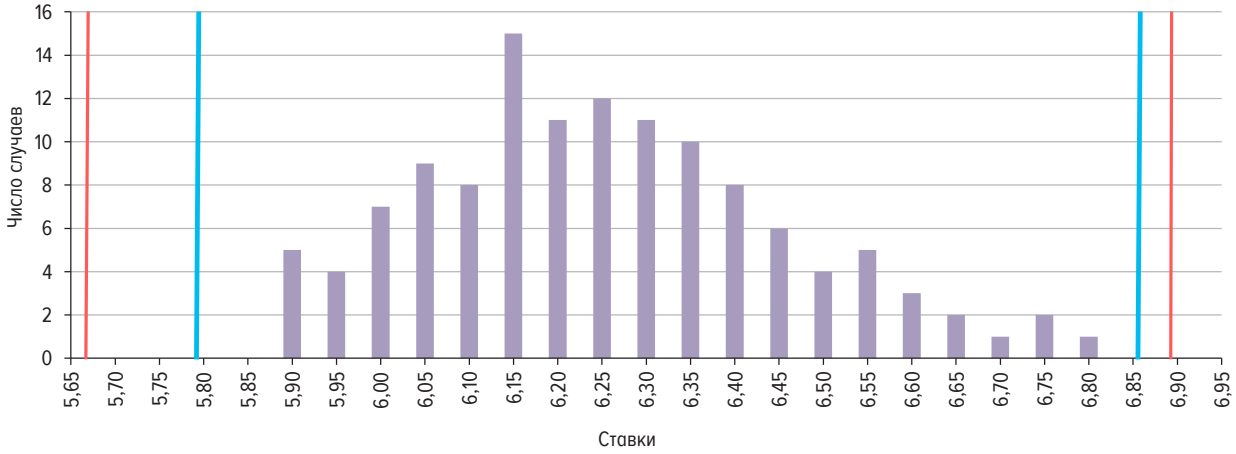
НИЗКОЕ (УНИМОДАЛЬНОЕ) РАСПРЕДЕЛЕНИЕ 1 С АСИММЕТРИЕЙ A = 0,60 И ЭКСЦЕССОМ E = 0,15 *Рис. П-8*



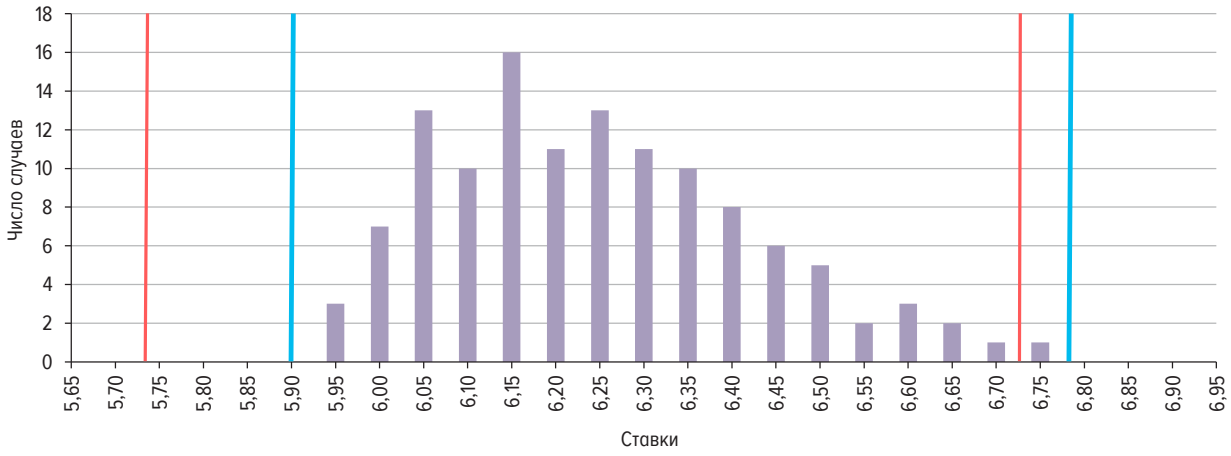
НИЗКОЕ (УНИМОДАЛЬНОЕ) РАСПРЕДЕЛЕНИЕ 2 С АСИММЕТРИЕЙ A = 0,50 И ЭКСЦЕССОМ E = -0,10 *Рис. П-9*



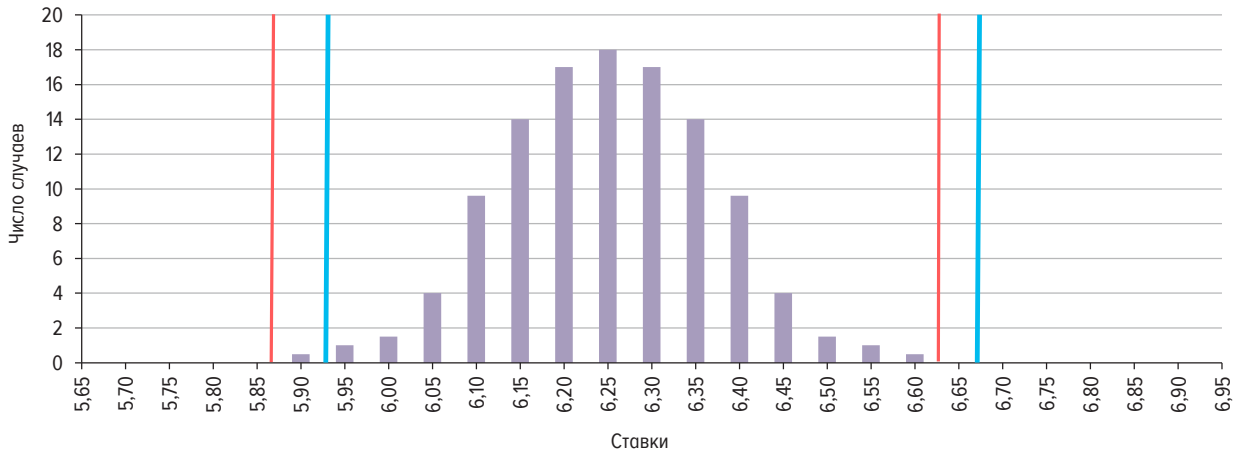
НИЗКОЕ (УНИМОДАЛЬНОЕ) РАСПРЕДЕЛЕНИЕ 3 С АСИММЕТРИЕЙ A = 0,42 И ЭКСЦЕССОМ E = -0,05 *Рис. П-10*



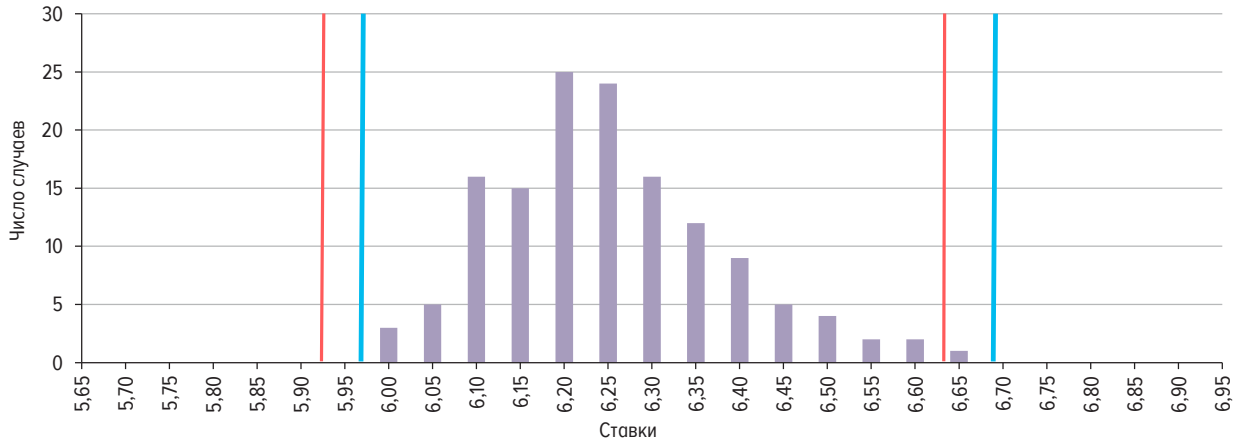
НИЗКОЕ (УНИМОДАЛЬНОЕ) РАСПРЕДЕЛЕНИЕ 4 С АСИММЕТРИЕЙ A = 0,55 И ЭКСЦЕССОМ E = 0,56 *Рис. П-11*



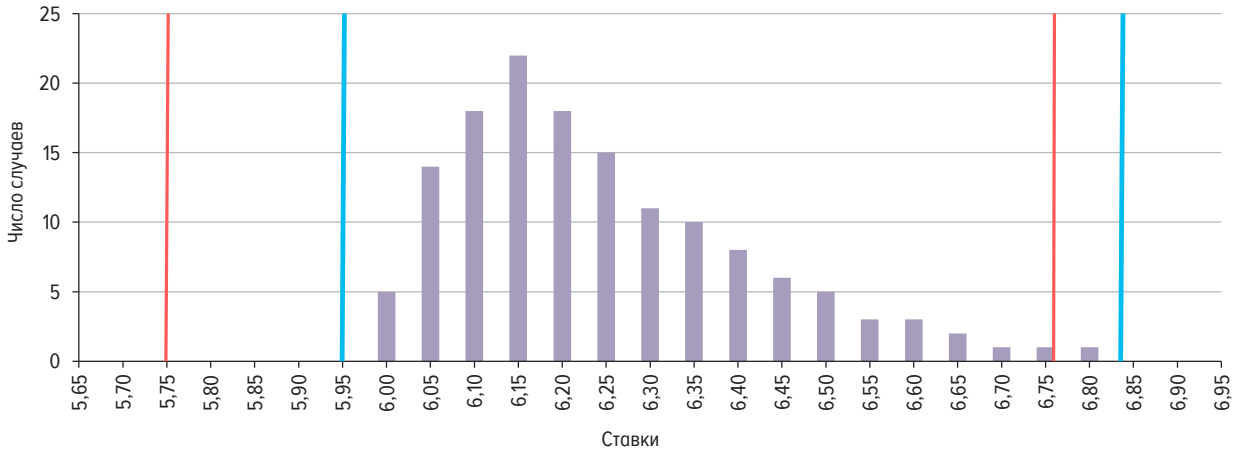
НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ С АСИММЕТРИЕЙ A = 0,00 И ЭКСЦЕССОМ E = 0,00 *Рис. П-12*



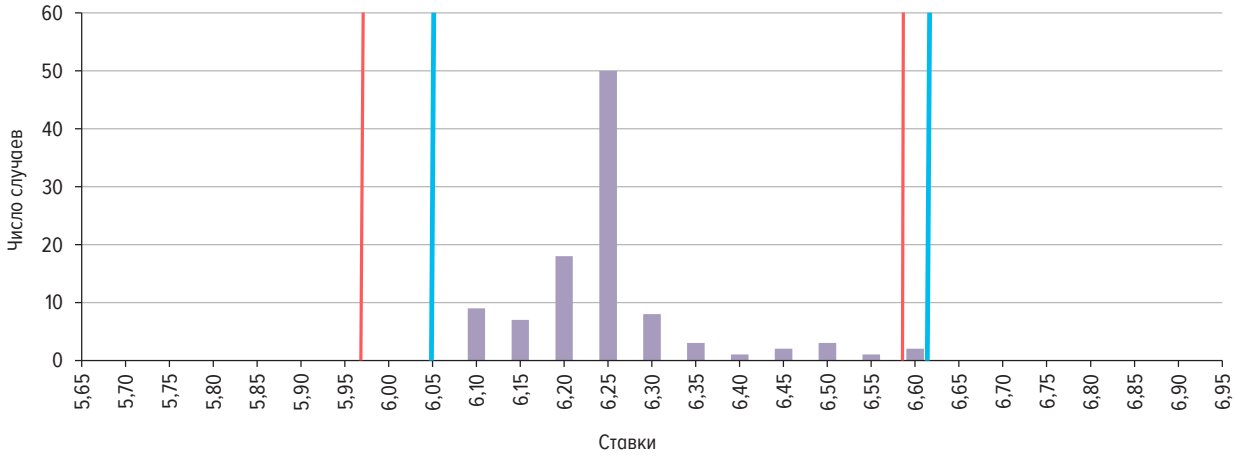
СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ 1 С АСИММЕТРИЕЙ A = 0,60 И ЭКСЦЕССОМ E = 0,15 *Рис. П-13*



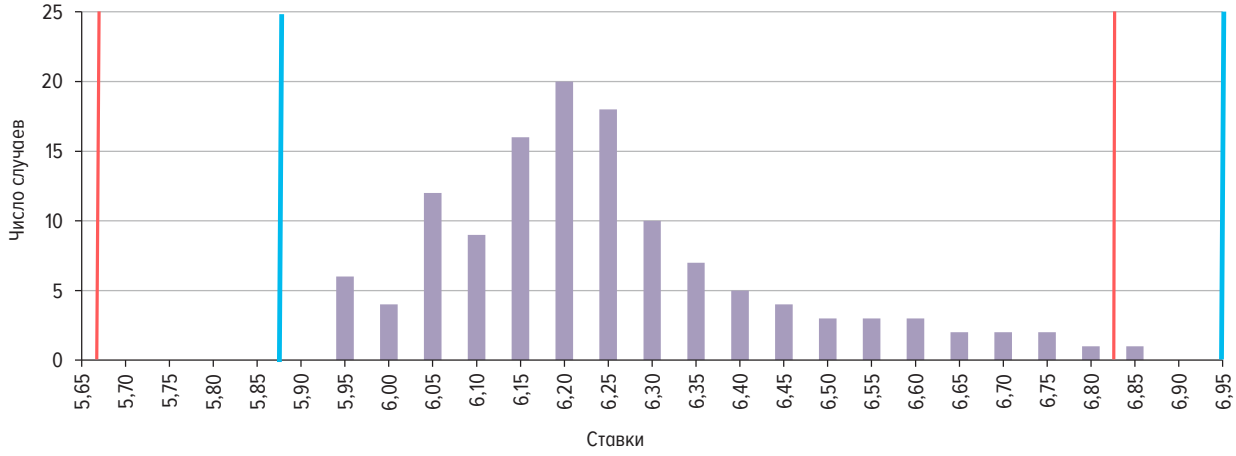
СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ 2 С АСИММЕТРИЕЙ A = 0,92 И ЭКСЦЕССОМ E = 0,84 *Рис. П-14*



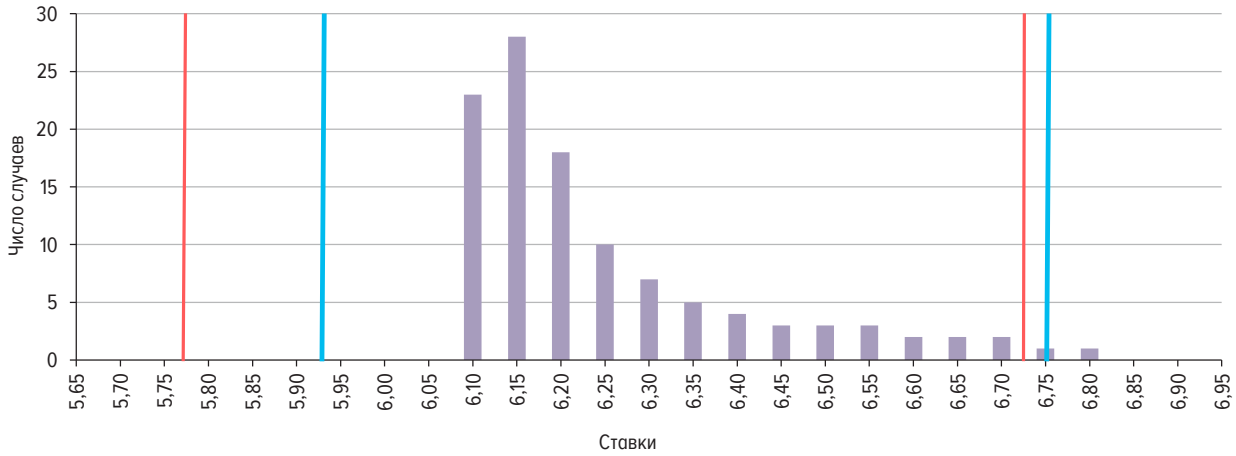
СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ 3 С АСИММЕТРИЕЙ A = 1,46 И ЭКСЦЕССОМ E = 2,33 *Рис. П-15*



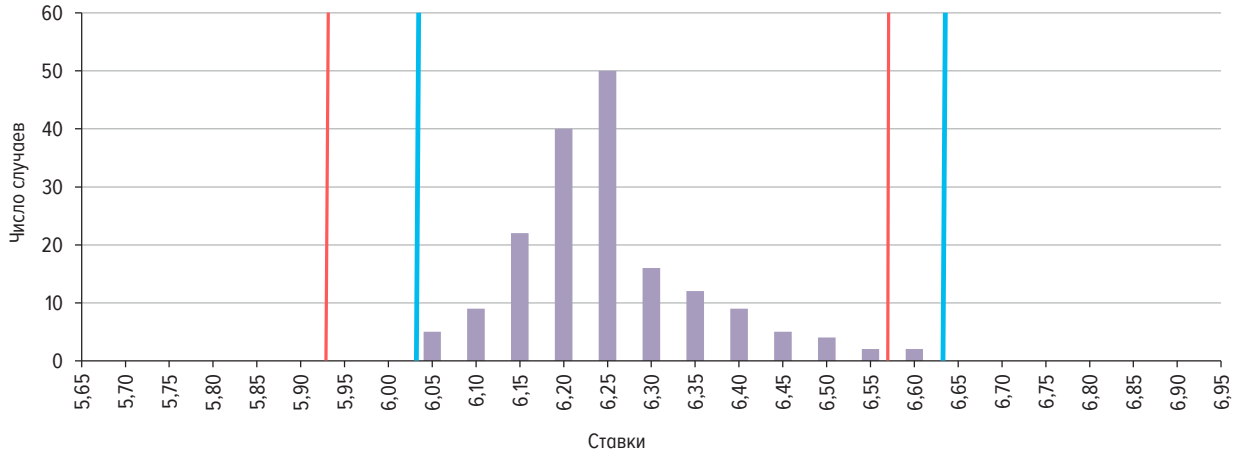
СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ 4 С АСИММЕТРИЕЙ A = 0,95 И ЭКСЦЕССОМ E = 0,14 *Рис. П-16*



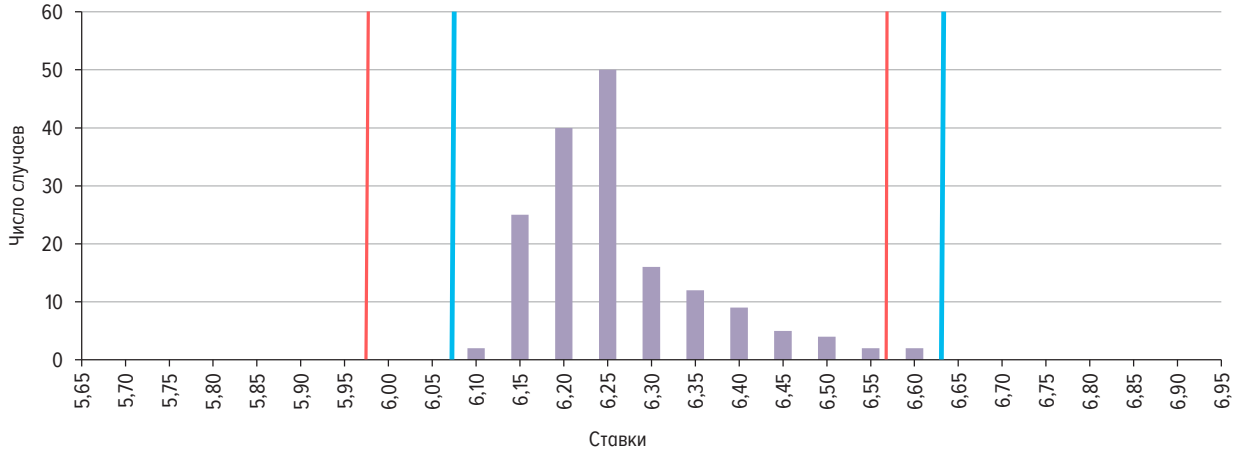
ВЫСОКОЕ РАСПРЕДЕЛЕНИЕ 1 С АСИММЕТРИЕЙ A = 1,46 И ЭКСЦЕССОМ E = -0,30 *Рис. П-17*



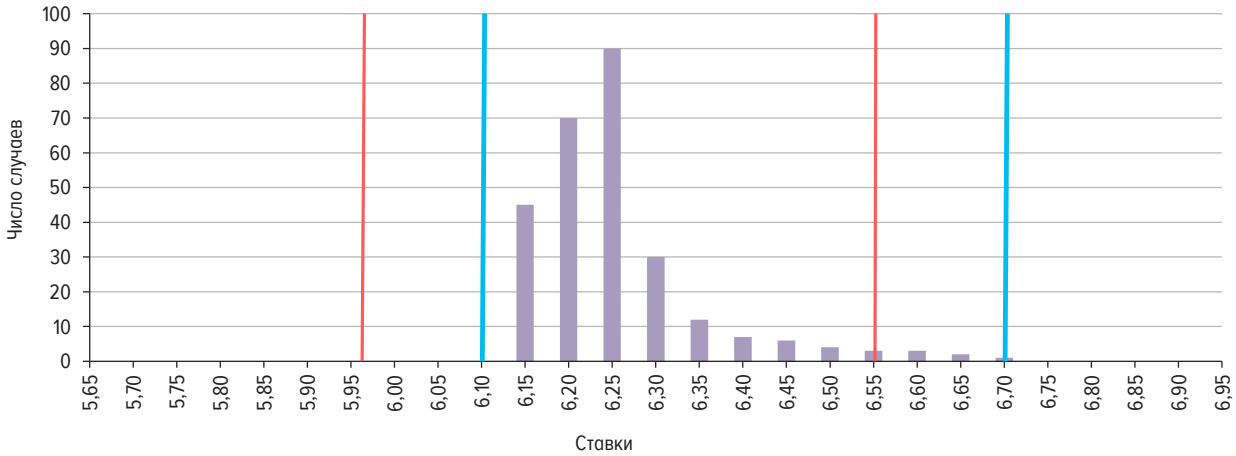
ВЫСОКОЕ РАСПРЕДЕЛЕНИЕ 2 С АСИММЕТРИЕЙ A = 0,90 И ЭКСЦЕССОМ E = 0,72 *Рис. П-18*



ВЫСОКОЕ РАСПРЕДЕЛЕНИЕ 3 С АСИММЕТРИЕЙ A = 1,24 И ЭКСЦЕССОМ E = 0,97 *Рис. П-19*



ВЫСОКОЕ РАСПРЕДЕЛЕНИЕ 4 С АСИММЕТРИЕЙ A = 1,91 И ЭКСЦЕССОМ E = 0,07 *Рис. П-20*



Список литературы

1. Айвазян С., Мхитарян В. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ. 1998.
2. Гамбаров Г.М. Проблемы статистического анализа и оценки стоимости финансовых активов. – М.: МЭСИ. 2010.
3. Общая теория статистики. Учебник для вузов / под ред. И.И. Елисеевой. 5-е изд. перераб. и доп. – М.: Финансы и статистика. 2005.
4. Розанова Н.М., Зарострaтова И.В. Экономический анализ фирм и рынка. М.: ЮНИТИ-ДАНА, 2015. – 280 с.
5. Hoaglin D.C., Mosteller F., Tukey J.W. Understanding Robust and Exploratory Data Analysis. John Wiley @ Sons, Ney York. 2000.
6. Waterson M. Economic theory of the industry. Cambridge University Press. 1984.