Доверительный интервал для среднего по выборке из конечной генеральной совокупности[1]

Зельдин М.А., FRICS
Баринов Н.П., FRICS
Аббасов М.Э.

В оценочной деятельности результат оценки не является точным значением рыночной стоимости – в большинстве случаев имеет место некая неопределенность. В простейшей ситуации сравнительного подхода[2] есть непосредственная возможность расчета доверительного интервала (ДИ) для *среднего по генеральной совокупности значения цены*[3] с надежностью α. Для этого обычно используют формулу[4]:

$$\overline{x} - t_{\alpha}\frac{s}{\sqrt{n}} \leq a \leq \overline{x} + t_{\alpha}\frac{s}{\sqrt{n}}, \qquad (1)$$

где $\overline{x}$ - выборочное среднее, t_{α} - квантиль распределения Стьюдента уровня α, a - среднее по генсовокупности, s - выборочное среднеквадратическое отклонение (СКО), n - число элементов (объем) выборки.

Заметим, что такая оценка границ доверительного интервала базируется на предположении о малой[5] выборке из большой (бесконечной) нормально распределенной генеральной совокупности (ГС). В то же время на практике оценщики обычно имеют дело с относительно небольшими (10-50-100 значений) генеральными совокупностями цен на рынке. Однако, даже имея такую совокупность, из соотношения (1) мы не узнаем точного значения среднего по ней, а лишь получим доверительный интервал с шириной, отличной от нуля. Естественно желание иметь аналогичное (1) выражение, которое бы использовало информацию о конечном объеме генеральной совокупности с тем, чтобы сближение объемов выборки n и генсовокупности N приводило бы к сужению границ доверительного интервала вплоть до нулевой их разности.

Такое выражение известно в литературе, посвященной статистическим методам выборочного исследования [2-5]:

$$\overline{x} - t_{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \leq a \leq \overline{x} + t_{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\ . \qquad (2)$$

где отношение n/N называют долей отбора.

---

[1] опубликовано: ПСМИ Регистр оценщиков, №11, 2012, с.70-75

[2] когда имеется достаточная выборка цен на один товар (или очень близкие товары), не требующая корректировок цен

[3] рассматривается здесь в качестве значения рыночной стоимости

[4] см., например, Закс Л. Статистическое оценивание М.: Статистика, 1976. - 598 с.

[5] в литературе малой считается выборка с числом элементов менее 5% объема генсовокупности

Зная (или ограничивая сверху) объем генсовокупности, можно учитывать долю отбора и получать более точные[6] оценки границ доверительного интервала. Т.е., в оценочной терминологии – можно уменьшать одну из составляющих неопределенности оценки стоимости[7]. Бороться есть за что – при обработке случайной выборки, содержащей половину[8] элементов генсовокупности, доверительный интервал сокращается, при прочих равных, почти на 30%.

При большей доле отбора эффект еще заметнее:

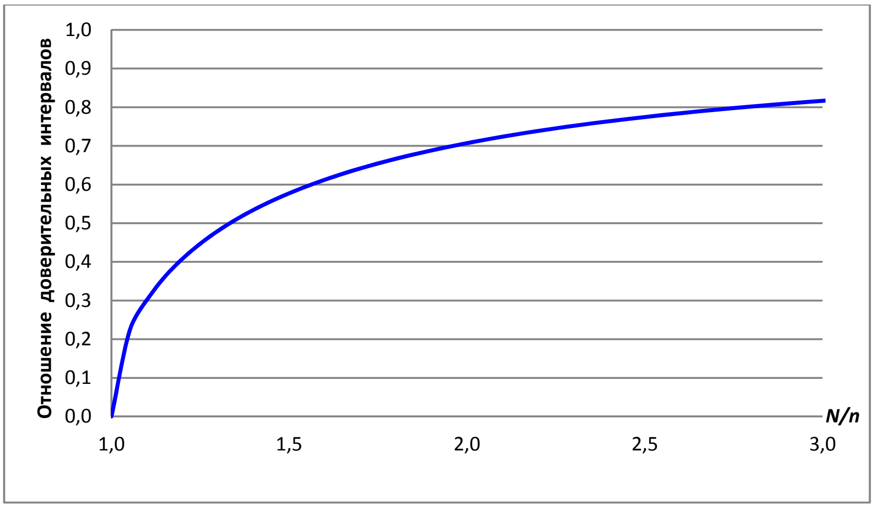

Рис.1 Зависимость отношения доверительных интервалов с учетом доли отбора и «классического» от отношения объемов генеральной совокупности и выборки из нее.

Кроме проблемы «малости» генеральной совокупности, встает вопрос о корректности использования выражений, основанных на нормальности распределения выборочного среднего, для реальных распределениях цен в «малых» ГС[9]. Этот вопрос может быть разрешен путем имитационного моделирования распределений выборочного среднего из совокупностей цен на реальном рынке. Проводя такое моделирование, можно оценить надежность (частость) попадания выборочного среднего в доверительный интервал для него,

---

[6] множитель $\sqrt{1-\dfrac{n}{N}}$, участвующий в (2) — поправка на конечность генеральной совокупности. Предельное значение этого множителя при $N \to \infty$ равно 1, что имеет место в «классическом» подходе.

[7] можно сказать – минимально достижимого значения интервала неопределенности оценки

[8] вполне разумное допущение при оценке на малоактивных рынках после тщательно проведенного поиска аналогов. В ряде случаев можно быть уверенными и в большей доле отбора – 70-90%.

[9] в таких совокупностях распределение цен, как правило, не может быть признано нормальным, см., например, http://www.avg.ru/pressa/methodologicaldevelopments/2012/4/870

рассчитанный с учетом «малости» ГС и в предположении нормальности распределения выборочного среднего[10].

Итак, пусть мы имеем дело с выборкой объема n, в то время как вся генсовокупность имеет объем N. Количество возможных сочетаний (выборок) из N по n равно

$$C_N^n = \frac{N!}{n!(N-n)!}.$$

Известно, что распределение средних, рассчитанных по каждой из этих выборок, при достаточно больших значениях C_N^n и $N>60$ хорошо аппроксимируется нормальным распределением даже тогда, когда распределение самой генсовокупности далеко от такового [3]. В качестве иллюстрации приведем гистограмму распределения населения городов России с числом жителей более 100 тыс. человек[11]:

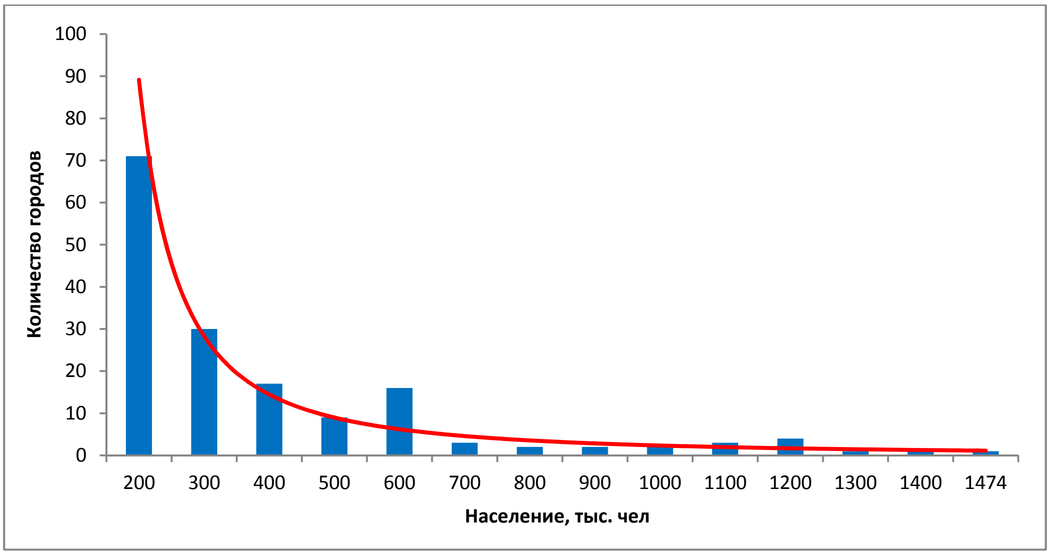

Рис. 2 Гистограмма распределения городов РФ по численности населения >100 тыс. чел.

Выберем из этих 162 городов случайным образом 50 и для них рассчитаем среднюю численность населения. Проделав указанную процедуру 100 тысяч раз, получим такое же количество средних значений, гистограмма распределения которых представлена на рис.3. Для наглядности сравнения на рисунке приведена гистограмма нормального распределения.

[10] Важно! Т.к. рассчитывается доверительный интервал для среднего значения, нам нужно предположение о виде распределения выборочного среднего цен, а не распределения самих цен в генсовокупности или в выборках из нее.

[11] по данным http://www.gks.ru/free_doc/doc_2011/year/year2011.rar , без Москвы и Санкт-Петербурга

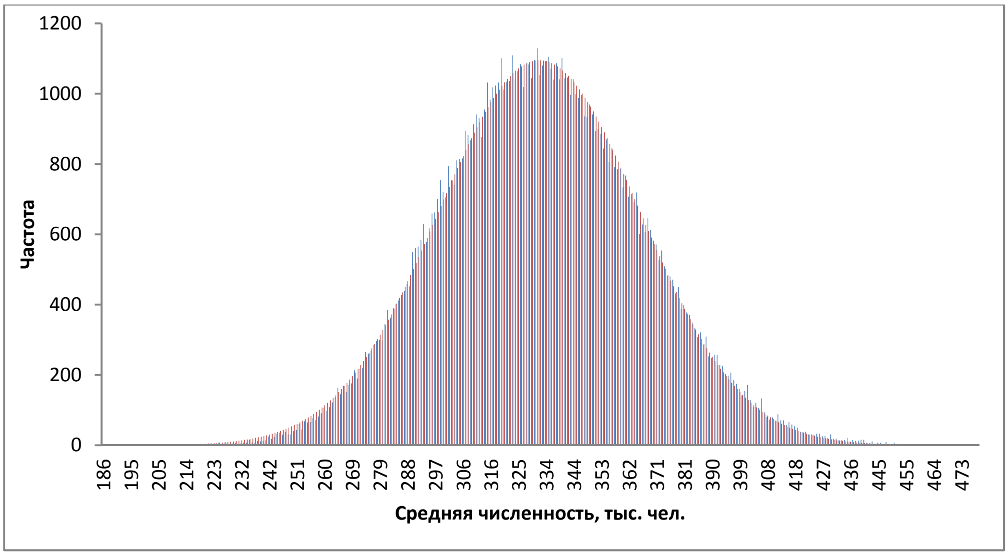

Рис. 3 Гистограммы выборочных средних (синий) и теоретического нормального распределения (красный) численности населения в 50 городах РФ.

Как видим, хотя распределение численности населения в генсовокупности (162 города) не является ни нормальным, ни даже отдаленно симметричным (см. рис.2), распределение средней численности по 50 из них[12] весьма близко к нормальному.

В приведенном примере объем выборки (n=50) достаточно велик по сравнению с выборками цен, которыми обычно оперирует оценщик: 5, в лучшем случае – 10-15 цен. Как обстоит дело в этом случае?

Обратимся к примеру. Примем за генеральную совокупность найденные[13] 94 цены на напольное покрытие (ламинат) с одинаковыми заявленными характеристиками[14]. Гистограмма цен по этой генсовокупности представлена на рис.4. Заметим, распределение цен нельзя признать ни нормальным[15], ни строго симметричным[16]. Выберем также два значения объема выборки n = 5, 10 и проведем аналогичное моделирование. Гистограммы выборочных средних для этих выборок представлены на рис.5-6. Там же приведены гистограммы нормального распределения со средним и дисперсиями[17] для генсовокупности.

[12] доля отбора составляет в данном случае 50/162=31%

[13] интернет-магазины на сайте яндекс-маркет по Москве за 29.04.2011

[14] класс - 32; дизайн - дуб; длина доски 1280 мм; ширина 190 мм; толщина 8 мм, все производители

[15] асимметрия 0.36, эксцесс – 1.33, результат теста по критерию Филлибена на 5% уровне - отрицательный

[16] результат теста по одновыборочному критерию Вилкоксона на 5% уровне - отрицательный

[17] среднеквадратические отклонения рассчитаны с учетом доли отбора, см. выше

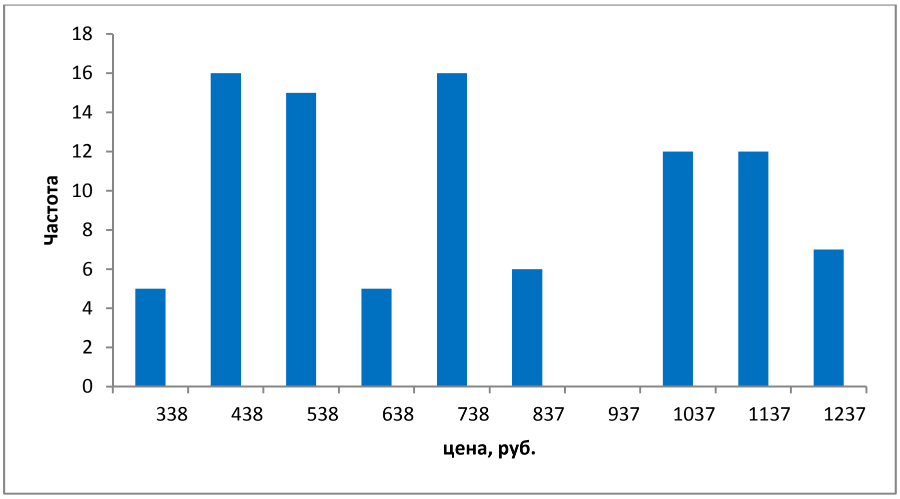

Рис. 4 Распределение цен на ламинат.

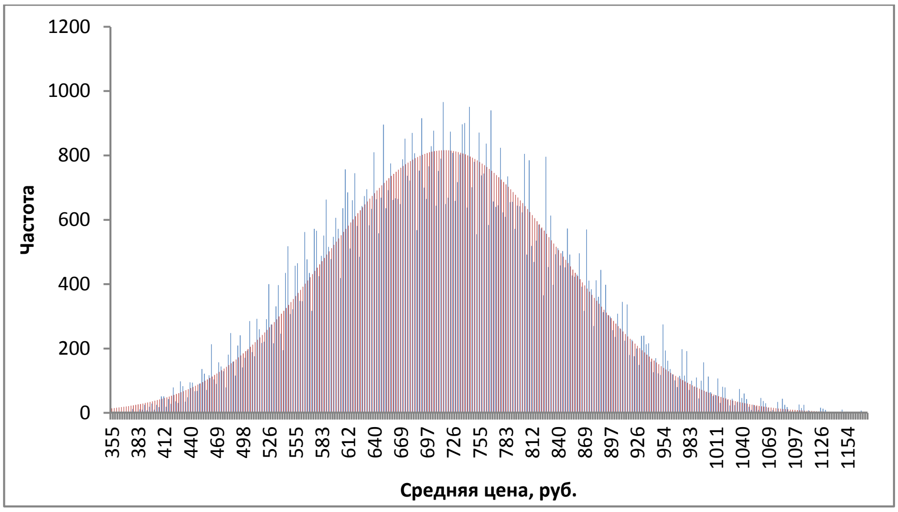

Рис. 5 Гистограмма выборочных средних (синий) и теоретического нормального распределения (красный), 5 цен из 94

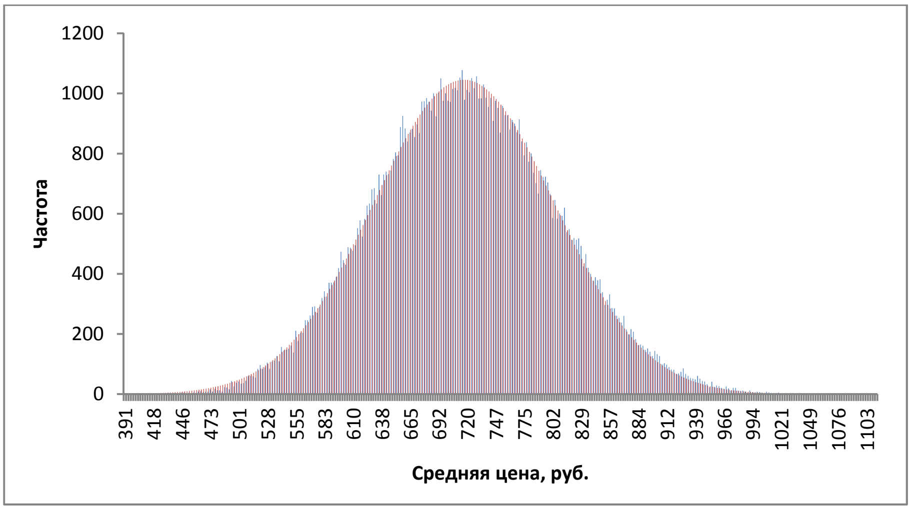

Рис. 6 Гистограмма выборочных средних (синий) и теоретического нормального распределения (красный), 10 цен из 94

Из диаграмм видно, что распределение средней по выборке цены аппроксимируется нормальным законом удовлетворительно для выборок по 5 цен, и очень хорошо - для выборок по 10 цен.

Итак, при относительно больших[18] выборках и генсовокупностях распределение выборочных средних нормализуется даже при явно несимметричных распределениях исходных ГС. При генсовокупностях, близких к симметричным, нормализация распределения выборочных средних происходит и на малых выборках, наблюдаемых на рынке. Однако на совокупностях малого объема невозможно надежно установить вид распределения цен с помощью статистических тестов. В этом случае можно провести непосредственное моделирование расчета доверительного интервала с заданным уровнем надежности, более предпочтительное с точки зрения конечного результата.

Такое моделирования проведено[19] в следующем порядке:

- на сайте Яндекс-маркет отобраны гомогенные товары[20], имеющие одинаковое число цен предложений N = 5, 6, 7, 8, 9, 10, 12, 15, 20, 25.
- Рассматривая выборку цен на отдельный товар как генеральную совокупность, из каждой из них извлечены все возможные[21] сочетания (выборки) меньшего объема n, обеспечивающие долю отбора в пределах n/N = 0,4-0,9.
- Для каждой выборки рассчитаны среднее арифметическое и среднеквадратическое отклонение цен.
- По значениям выборочного СКО цен рассчитаны значения выборочного СКО распределения *средних* цен по формуле[22]: $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$
- По выборочному среднему $\bar{x}$ и СКО $\sigma_{\bar{x}}$ по формуле (2) для среднего по ГС рассчитывался 95%-ый доверительный интервал «Стьюдента-Кокрена»[23], учитывающей долю отбора, а также по формуле (1) - классический 95%-ый доверительный интервал Стьюдента.
- В каждом случае проверялось, попадает ли в построенный доверительный интервал среднее значение, рассчитанное по генеральной совокупности («истинное» среднее).

[18] десятки и сотни единиц
[19] http://www.avg.ru/pressa/methodologicaldevelopments/2012/5/884
[20] 119 товаров, см. http://www.avg.ru/pressa/methodologicaldevelopments/2012/4/870
[21] за исключением сочетаний 13,15, 18 из 25, 10,12 из 20, для которых случайным образом извлечено по 100 тыс. выборок.
[22] см., например, [2-5]
[23] название условно, равно можно назвать «Стьюдента –Йетса»

- По всем сочетаниям, соответствующим выбранному значению N и соответствующим ему набору n, рассчитан процент попадания «истинного» среднего в каждый из двух доверительных интервалов[24].
- Выполнив аналогичные операции для всех 119 товаров, рассчитаны средние частоты попаданий в интервалы по всем генеральным совокупностям одного объема N, а также по всем выборкам одного объема n.

Общие результаты моделирования представлены на следующих рисунках:

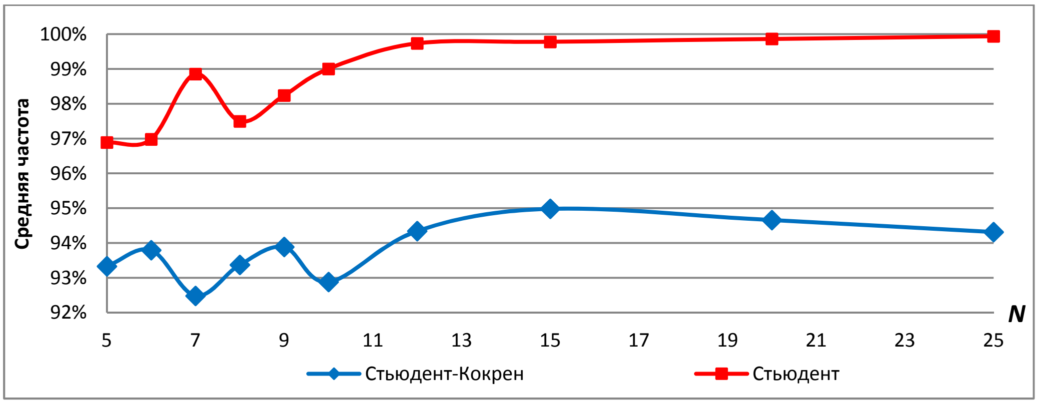

Рис. 7 Зависимость частоты попадания истинного среднего в ДИ от *N* (по всем выборкам)

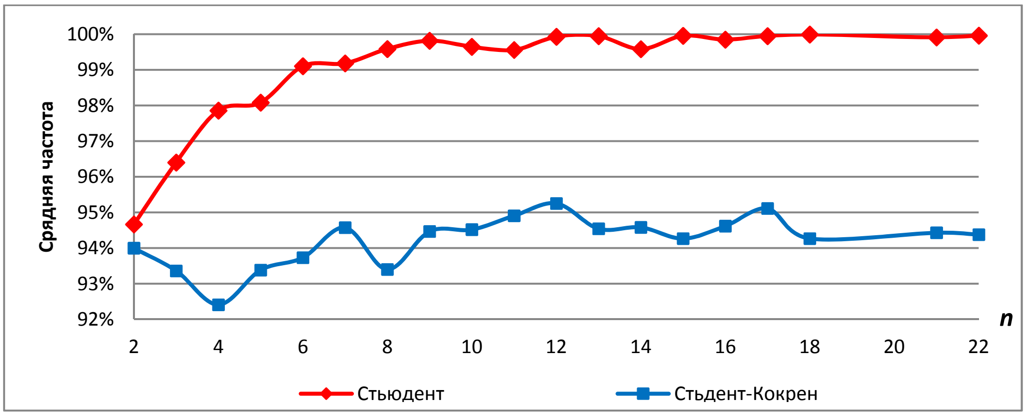

Рис. 8 Зависимость частоты попадания истинного среднего в ДИ от *n* (по всем выборкам)

[24] *Например*, имелись 10 выборок цен объема *N=12* на различные товары, каждая из которых была принята за генеральную совокупность. По каждому товару строились сочетания цен (выборки) объемом *n = 6, 7, 8, 9, 10*, в количестве $C_{12}^{6}=924$, $C_{12}^{7}=792$, $C_{12}^{8}=495$, $C_{12}^{9}=220$, $C_{12}^{10}=66$ соответственно. По каждой такой выборке строились доверительные интервалы Стьюдента-Кокрена и Стьюдента. Таким образом сформировано множество выборок с числом элементов, равным C_N^n, соответствующих данному товару (одной из 10 генеральных совокупностей объема *N=12*) и данному объему выборки *n*. По каждому такому множеству подсчитывалось число попаданий истинного среднего в доверительные интервалы и делилось на общее число элементов множества.

Полученные результаты свидетельствуют о весьма удовлетворительной близости экспериментальных данных ожидаемым: расхождение средних частот попадания истинного среднего в доверительный интервал Стьюдента-Кокрена не превышает 2 процентных пунктов за исключением единичных точек в области особо малых генеральных совокупностей и выборок.

Однако алогичной выглядит понизительная тенденция зависимости частоты попадания от объема ГС в интервале ***N***=15-25 (рис.7, Стьюдент-Кокрен). Заметим, что на практике мы не можем точно знать значения этого объема[25] и для нас важнее зависимость от объема выборки ***n***, но хорошо бы понимать причины подобной тенденции. Картину проясняют аналогичные зависимости, построенные раздельно для выборок из совокупностей с умеренной асимметрией[26]:

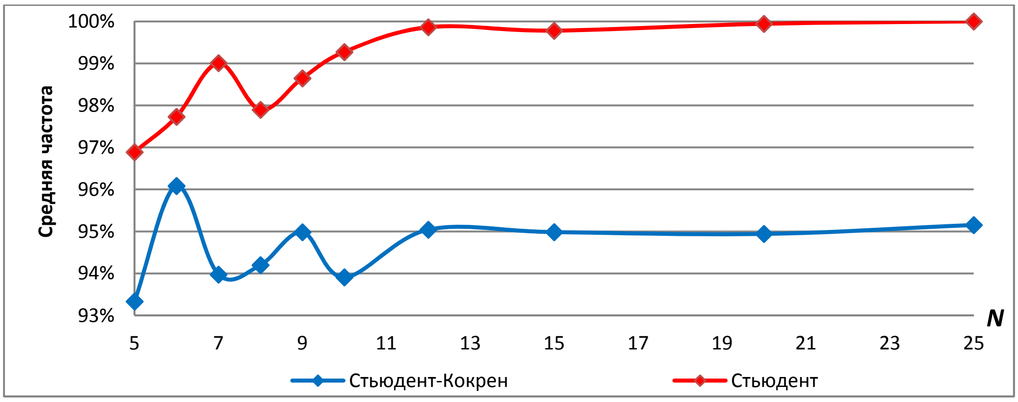

Рис. 9 Зависимость частоты попадания истинного среднего в ДИ от *N* (по «симметричным» совокупностям)

и явно асимметричных[27]:

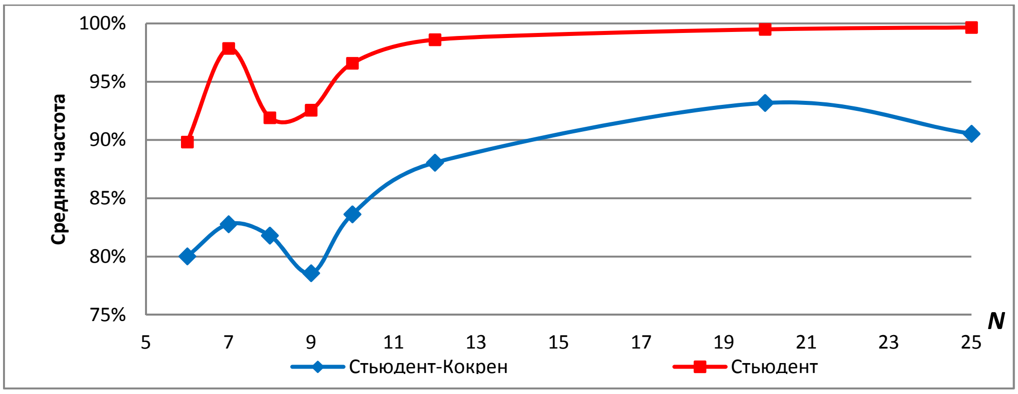

Рис. 10 Зависимость частоты попадания истинного среднего в ДИ от *N* (по асимметричным совокупностям)

[25] всех цен на рассматриваемом рынке

[26] модуль коэффициента асимметрии в пределах 0,01- 1,43, среднее значение – 0,55

[27] модуль коэффициента асимметрии 1,53-2,33, среднее значение - 1,75

Налицо сильное влияние асимметрии ГС на вид рассматриваемых зависимостей. При явной асимметрии распределения цен в ГС небольшого объема (N<25) частота попадания в интервал Стьюдента-Кокрена лишь в отдельных точках превышает 90%. В то же время для совокупностей с умеренной асимметрией отклонение от расчетной частоты (95%) не превышает 1 процентного пункта, начиная уже с N>5 (рис.9).

Аналогичные зависимости от объема выборки ***n*** показывают схожую картину:

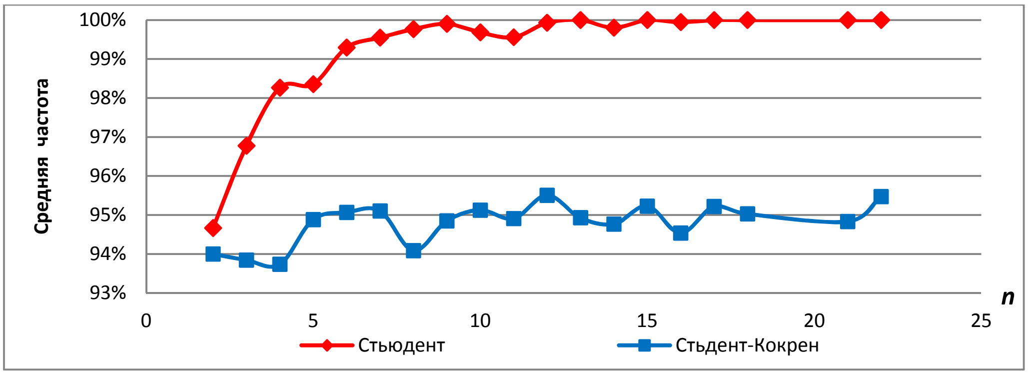

Рис. 11 Зависимость частоты попадания истинного среднего в ДИ от *n* (по «симметричным» совокупностям)

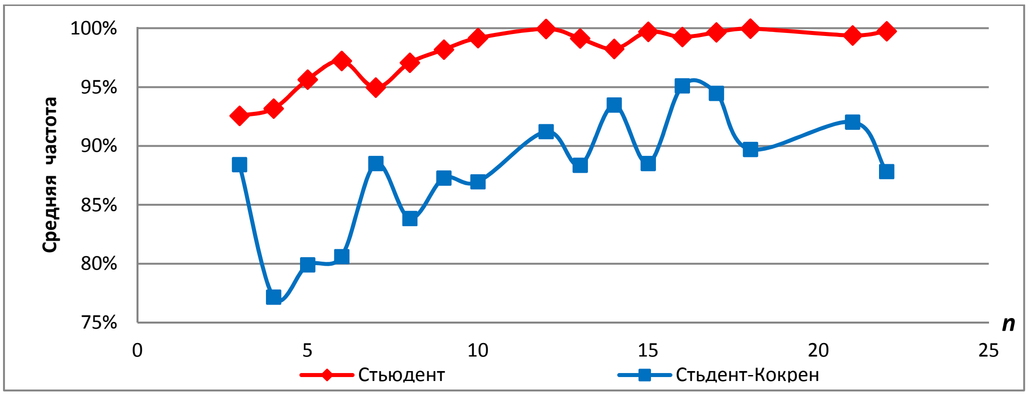

Рис. 12 Зависимость частоты попадания истинного среднего в ДИ от *n* (по асимметричным совокупностям)

Очевидно, что для совокупностей с явно выраженной асимметрией учет доли отбора по соотношению (2) может существенно занижать оценку доверительного интервала с заданной надежностью. В этих случаях, до получения новых соотношений, предпочтительно использовать «классический» интервал Стьюдента.

Наконец, интересно влияние доли отбора на рассматриваемые зависимости. Для его выявления множество выборок было разбито на два подмножества: с долей отбора в пределах n/N=0,4-0,6 и с высокой долей 0,7-0,9. Т.к. практический интерес представляют совокупности с умеренной асимметрией распределения цен, приведем зависимости частоты попадания в доверительный интервал Стьюдента-Кокрена именно для них:

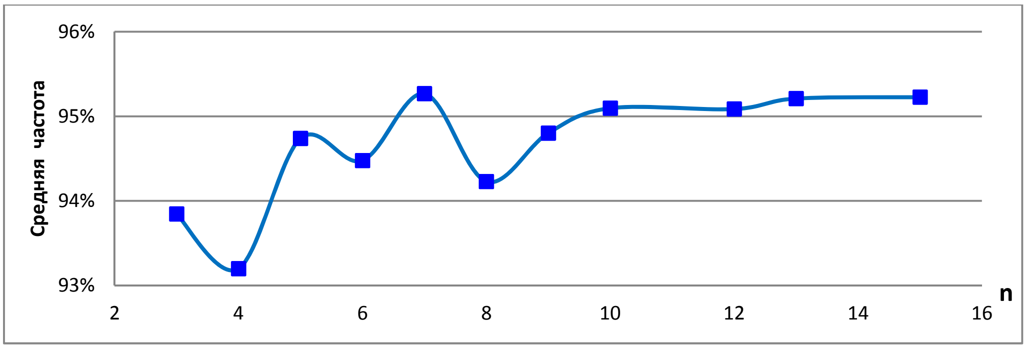

Рис. 13 Зависимость частоты попадания истинного среднего в ДИ Стьюдента-Кокрена от *n* по «симметричным» совокупностям, доли отбора 0,4-06

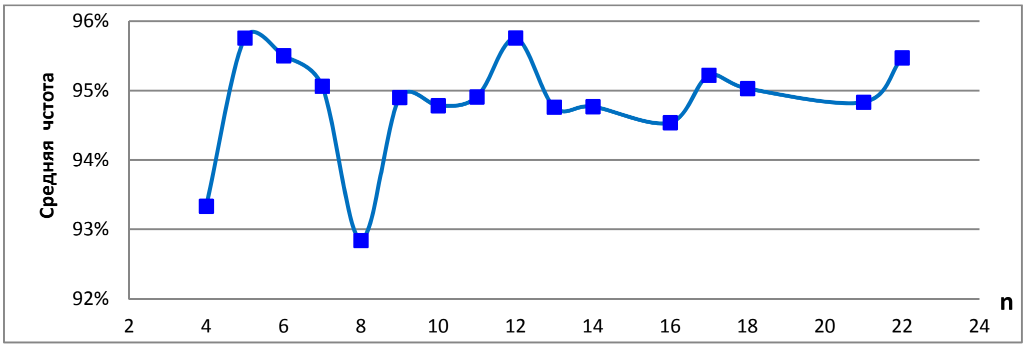

Рис. 14 Зависимость частоты попадания истинного среднего в ДИ Стьюдента-Кокрена от *n* по «симметричным» совокупностям, доли отбора 0,7-0,9

Полученные результаты демонстрируют слабое влияние доли отбора[28] на среднюю частоту попадания «истинного» среднего в расчетный доверительный интервал у ГС с умеренно выраженной асимметрией. А для не самых малых выборок ($n \geq 9$) из таких совокупностей можно утверждать, что это влияние незначимо. Отрадным является небольшие (не более 2 процентных пунктов) отклонения от расчетных величин частоты даже при минимальных объемах выборки – n=3-5 единиц.

Необходимо отметить также, что на «малых» генеральных совокупностях (N=5-25) «классическая» формула (1) существенно завышает доверительный интервал (Стьюдента) практически во всех рассмотренных случаях, за исключением одного - особо малых ($n \leq 5$) выборок из сугубо асимметричных ГС.

Что же можно сказать о практическом применении полученных результатов моделирования?

1. Прежде всего – они открывают возможность более точного[29] определения минимально достижимого интервала неопределенности для оценки рыночной стоимости, т.е. интервала, присущего оценке в условиях, когда никакого искажения в рыночные цены оценщиком не вносится. Образно говоря, теперь мож-

[28] в пределах значений n/N =0,4-0,9

[29] на 30-50%

но уточнить высоту «потолка точности» оценки, выше которого оценщику не прыгнуть.

Заметим, что отказ от модели бесконечной генсовокупности в пользу конечного ее объема N и связанная с этим возможность уменьшения неопределенности не дается даром. Он порождает новую неопределенность – мы не можем знать точного значения N. Выход здесь может быть в допущении доли отбора, «гарантированной снизу» (например, не менее 50%), и получении «оценки сверху» для доверительного интервала[30]. В ряде случаев такая «гарантия» вполне уместна и оценка может быть существенно «точнее» полученной по «классическим» соотношениям.

2. Полезным, по мнению авторов, является подтверждение возможности использования соотношений, полученных в предположении нормально распределенной генсовокупности в условиях реальных рыночных распределений, далеких от нормальных. Ограничителем здесь выступает лишь явно выраженная асимметрия распределения цен.

Необходимо также отметить, что все сказанное предполагает понимание рыночной стоимости как средней по генсовокупности цены. Хотя практика повсеместно реализует именно такое понимание, в последнее время оно стало предметом оживленных дискуссий[31].
С другой стороны, есть данные, позволяющие предположить, что большинство распределений цен сделок с гомогенными товарами на рынке не имеет явно выраженной асимметрии. Если это так, значения арифметического среднего и медианы по генсовокупности весьма близки[32]. И для всех, кто воспринимает рыночную стоимость как отражение центральной тенденции на рынке, почва для споров сужается, а для корректного применения среднего значения - расширяется.

Литература

1. Закс Л. Статистическое оценивание М.: Статистика, 1976. - 598 с.
2. Йетс Ф. Выборочный метод в переписях и обследованиях, пе-ревод с английского, М.: «Статистика», 1965. — 434 с.
3. Кокрен У. Методы выборочного исследования. М.: «Статистика», 1976 – 440с
4. Шварц Г. Выборочный метод. Руководство по применению статистических методов оценивания, перевод с немецкого, М.: «Статистика», 1978. — 213 с.
5. Джессен Р. Методы статистических обследований/Пер, с англ.; Под ред. и с предисл. Е. М. Четыркина. — М.: Финансы и статистика, 1985. — 478 с

[30] и здесь вновь не обойтись без суждения оценщика

[31] В частности, на «оценочных» форумах http://www.appraiser.ru/default.aspx?SectionId=32 и http://www.labrate.ru/cgi-bin/discus/discus.cgi

[32] с модой вопрос сложнее – как определять ее на очень малых совокупностях дискретных величин и что она при этом отражает?