Сергей Пахомов

Оценка погрешностей измерений при тестировании компьютеров

При тестировании компьютеров или их отдельных комплектующих нередко возникают проблемы с корректной оценкой погрешностей измерения, которая в конечном счете позволяет правильно сравнивать производительность тестируемых компьютеров (комплектующих). К сожалению, далеко не всегда этим проблемам уделяется должное внимание. Более того, нередко результаты измерения приводятся как усредненное значение по трем (в лучшем случае — по пяти) прогонам теста без оценки погрешности измерения, что, в свою очередь, может привести к неправильной трактовке результатов измерения.

Рассмотрим простой пример. Допустим, имеются две системы со схожими конфигурациями — A и B. Например, при единичном запуске теста система A продемонстрировала результат X1, а система B — результат X2, причем X1>X2, однако разница результатов невелика. Возникает вопрос: если считать, что большему результату соответствует более высокая производительность, можно ли на основе сравнения результатов X1 и X2 однозначно утверждать, что производительность системы A выше производительности системы B? Оказывается, нет. Более того, при повторном прогоне тестов результаты могут оказаться диаметрально противоположными. Возникает вопрос: сколько раз нужно проводить измерения, чтобы достоверность результатов была гарантирована? В настоящей статье мы постараемся ответить на этот и другие подобные вопросы, связанные с погрешностью измерений при тестировании компьютеров.

Понятия среднего и истинного значений

Итак, допустим, что имеется некий программный тест, который позволяет определять какую-то характеристику ПК. Например, синтетический тест, измеряющий ширину пропускания шины памяти или латентность памяти, либо тест на производительность ПК или его отдельной подсистемы в конкретном приложении или даже в наборе приложений. Мы будем рассматривать только те тесты, которые возвращают результат в числовой форме, то есть предполагается, что измеряемая величина представляется в числовом виде.

Если проводить измерения несколько раз, можно заметить, что постоянно будут получаться различные значения искомой величины. Точное значение искомой величины никогда априори не известно. Можно лишь предположить, что точное (истинное) значение искомой величины может быть определено на основе бесконечного числа измерений с последующим усреднением результата, то есть:

$$x_{ист} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

где x_i — результат, полученный при i-м измерении, n — число измерений.

В реальной же ситуации можно реализовать лишь конечное число измерений и вычислить на основе полученных результатов среднее значение:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Каждое отдельное измерение искомой величины мы будем называть выборкой, а число измерений — объемом выборки. Если речь идет о бесконечном числе измерений (выборка бесконечного объема), то такую выборку принято называть генеральной совокупностью. Соответственно результат, полученный при усреднении n измерений, называют средним по выборке объема n, а истинное значение определяется как среднее по генеральной выборке или — по генеральным средним.

Понятно, что, вообще говоря, истинное значение может и не совпадать со средним по конечной выборке. Также ясно, что чем больше объем выборки, тем меньше разница между истинным значением искомой величины и средним по выборке. Разница же между истинным значением искомой величины и рассчитанным средним по выборке определяет погрешность измерения.

Модуль разницы между истинным значением искомой величины и значением отдельной выборки, то есть одного измерения, характеризует ошибку единичного измерения: $\Delta x_i = |x_i - x_{ист}|$. Но поскольку истинное значение, как мы уже отмечали, априори неизвестно и наилучшим приближением к нему считается среднее по выборке объема n, то под ошибкой (погрешностью) единичного измерения можно понимать модуль разницы между средним по выборке и значением единичного измерения, то есть:

$$\Delta x_i = |x_i - \bar{x}|.$$

Рассмотрим понятие среднего по выборке на примере популярного теста VeriTest Business Winstone 2004 (все приводимые в данной статье примеры реальны). Данный тест мы запустили 50 раз — таким образом, максимальный объем выборки равен 50. Если построить график зависимости среднего по выборке от ее объема (рис. 1), то получится ломаная линия, значения точек которой постепенно сходятся к некоторому истинному значению, определяемому для бесконечного объема выборки.

Результат каждого отдельного измерения (выборка) носит случайный характер. Графически выборка может быть представлена в виде гистограмм и графиков функций распределения.

На рис. 2 приведен пример гистограммы распределения частоты выпадения результатов для рассмотренного выше примера (в тесте VeriTest Business Winstone 2004).

По горизонтальной оси отложены значения результатов теста от минимального до максимального значений для выборки объемом 50, а по вертикальной оси — нормированное число значений из выборки, соответствующих указанным результатам теста. При этом под нормированным значением понимается частота выпадения того или иного результата, то есть если результат X встречается n раз в выборке объемом N, то частота его выпадения будет равна n/N.

Нетрудно заметить, что наибольшее число результатов сосредоточено вокруг некоторого среднего значения, а к краям доля значений плавно уменьшается.

Если количество измерений, то есть объем выборки, устремить к бесконечности, то получится некоторая предельная гистограмма. Плавная кривая, проведенная через вершины столбцов предельной диаграммы, является функцией распределения вероятности $f(x)$.

Конкретная форма функции распределения плотности вероятности зависит от характера случайной величины. Известно большое количество функций распределения, которыми описываются различные физические процессы. В принципе, основываясь на результатах измерений, можно выяснить, какому именно распределению соответствует плотность распределения выпадения случайной величины. Наиболее часто встречаются распределения Гаусса и Пуассона. Однако мы не станем углубляться в дебри статистической математики и лишь заметим, что по форме гистограммы, представленной на рис. 2, можно

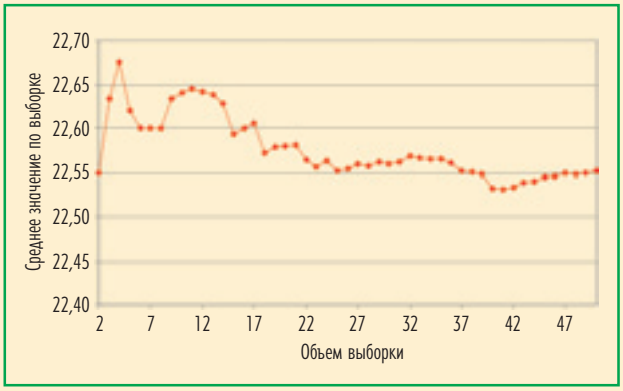

Рис. 1. График зависимости среднего по выборке от ее объема в тесте VeriTest Business Winstone 2004

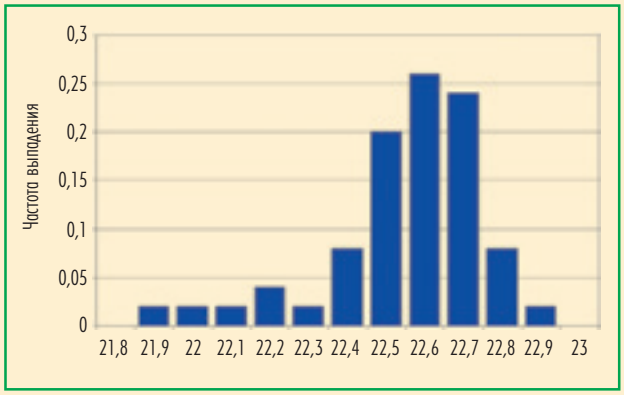

Рис. 2. Гистограмма распределения частоты выпадения результатов в тесте VeriTest Business Winstone 2004

утверждать, что речь идет о гауссовом распределении (предполагается, что результаты теста принимают непрерывный ряд значений, что не совсем точно).

Оценка случайной погрешности измерений

Пусть имеется выборка объема n и x_i — результат отдельного измерения. Как уже отмечалось, наилучшим приближением к истинному значению измеряемой величины (генеральному среднему) является среднее по выборке, то есть:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Для характеристики погрешности измерения вводят понятие среднеквадратичного отклонения (стандартного отклонения) выборки:

$$S_n = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \ .$$

При объеме выборки, стремящемся к бесконечности ($n \to \infty$), стандартное отклонение стремится к некоторому постоянному пределу: $\sigma = lim\ S_n$.

Стандартное отклонение является важным понятием и характеризует рассеивание значений случайной величины в окрестности ее среднего значения.

Зная стандартное отклонение по выборке, можно определить интервал допустимых значений генерального среднего, то есть интервал от $\bar{x} - \Delta x$ до $\bar{x} + \Delta x$, внутри которого с заданной вероятностью p находится истинное значение искомой величины. При этом интервал от $\bar{x} - \Delta x$ до $\bar{x} + \Delta x$ называется доверительным интервалом, а вероятность того, что в этом интервале находится истинное значение искомой величины, — доверительной вероятностью или коэффициентом надежности.

Опуская утомительные математические выкладки, приведем лишь конечный результат. Если производится n измерений искомой величины, то доверительный интервал истинного значения искомой величины определяется по формуле:

$$x_{ист} = \bar{x} \pm t(p,n)\frac{S_n}{\sqrt{n}} \ ,$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ ,$$

$$S_n = \sqrt{\sum_{i=1}^{n} \frac{(\bar{x} - x_i)^2}{n-1}} \ ,$$

поправочные коэффициенты $t\ (n,\ p)$ — коэффициенты Стьюдента. Их зависимость от n и p приведена в табл. 1.

При обработке результатов измерений для вычисления коэффициентов Стьюдента удобно использовать встроенные функции таблиц Excel. В данном случае формула для коэффициента Стьюдента $t\ (n,\ p)$ записывается в виде «=СТЬЮДРАСПОБР((1-p),n-1)».

Как нетрудно заметить, по мере увеличения количества измерений доверительный интервал сужается относительно среднего по выборке при неизменном коэффициенте надежности. Аналогично, уменьшая коэффициент надежности, можно сужать доверительный интервал.

Во многих программных тестах коэффициент надежности выбирается равным 0,90 или 0,95, то есть утверждается, что истинное значение искомой величины с вероятностью 90 или 95% будет находиться внутри доверительного интервала.

Примеры вычисления доверительных интервалов

Рассмотрим типичную ситуацию, когда с помощью программного теста Business Winstone 2004 сравнивается производительность двух процессоров. Допустим, что имеются два процессора с одинаковой тактовой частотой, первый из которых поддерживает технологию Hyper-Threading, а второй — нет. В частности, мы рассмотрим реальные примеры тестирования двухъядерных процессоров Intel Pentium Processor Extreme Edition 840 и Intel Pentium D 840 (напомним, что процессор Intel Pentium Processor Extreme Edition 840 поддерживает технологию Hyper-Threading, а процессор Intel Pentium D 840 — нет). Предполагать априори, как скажется технология Hyper-Threading на производительности ПК при работе с офисными приложениями, довольно трудно, однако понятно, что разница в результатах тестов для этих процессоров будет незначительна.

После первого прогона теста для процессора Intel Pentium Processor Extreme Edition 840 результат оказался равным 22,7, а для процессора Intel Pentium D 840 — 22,5. Понятно, что, основываясь на результатах лишь одного измерения, делать вывод о том, что производительность одного процессора в данном тесте

Таблица 1. Коэффициенты Стьюдента

Количество измерений	p=0,6	p=0,7	p=0,8	p=0,9	p=0,95	p=0,99
2	1,37638	1,96261	3,07768	6,31375	12,70620	63,65674
3	1,06066	1,38621	1,88562	2,91999	4,30265	9,92484
4	0,97847	1,24978	1,63774	2,35336	3,18245	5,84091
5	0,94096	1,18957	1,53321	2,13185	2,77645	4,60409
6	0,91954	1,15577	1,47588	2,01505	2,57058	4,03214
7	0,90570	1,13416	1,43976	1,94318	2,44691	3,70743
8	0,89603	1,11916	1,41492	1,89458	2,36462	3,49948
9	0,88889	1,10815	1,39682	1,85955	2,30600	3,35539
10	0,88340	1,09972	1,38303	1,83311	2,26216	3,24984

Таблица 2. Средние значения и доверительный диапазон, рассчитанные для двух прогонов теста Business Winstone 2004

Процессор	Результаты отдельных прогонов теста		Итоговые интегральные результаты
Intel Pentium Processor Extreme Edition 840	22,7	22,6	22,65±0,64
Intel Pentium D 840	22,5	22,6	22,55±0,4

Таблица 3. Средние значения и доверительный диапазон, рассчитанные для пяти прогонов теста Business Winstone 2004

Процессор	Результаты отдельных прогонов теста		Итоговые интегральные результаты
Intel Pentium Processor Extreme Edition 840	22,7	22,6	22,65±0,64
Intel Pentium D 840	22,5	22,6	22,55±0,4

выше производительности другого процессора (особенно при близких результатах), абсолютно некорректно, поскольку для искомой величины даже не определяется доверительный интервал. Поэтому необходимо продолжить измерение искомой величины. Прогнав тест еще раз, получим два результата, по которым можно рассчитать среднее значение и доверительный диапазон (табл. 2). Несмотря на то что для процессора Intel Pentium Processor Extreme Edition 840 среднее по выборке из двух прогонов теста выше, чем для процессора Intel Pentium D 840, утверждать, что его производительность в тесте Business Winstone 2004 выше, нельзя. Дело в том, что доверительные диапазоны (в дальнейшем все доверительные диапазоны будут рассчитываться для коэффициента надежности 0,95) перекрываются, а поскольку доверительный диапазон определяет тот интервал значений, в котором с вероятностью 95% находится истинное значение искомой величины, сравнивать производительность процессоров на основании полученных данных некорректно.

Если увеличить количество прогонов теста до пяти (табл. 3), то средний результат для процессора Intel Pentium Processor Extreme Edition 840 окажется ниже, чем для процессора Intel Pentium D 840. Однако и в этом случае делать однозначный вывод о том, что производительность процессора Intel Pentium D 840 в данном тесте выше, чем производительность процессора Intel Pentium Processor Extreme Edition 840, было бы некорректно, поскольку доверительные диапазоны искомой величины все еще перекрываются и истинное значение искомой величины для процессора Intel Pentium Processor Extreme Edition 840 может оказаться выше, чем для процессора Intel Pentium D 840. Таким образом, единственный вывод, который можно сделать на основе полученных результатов, заключается в том, что производительность обоих процессоров одинакова в пределах погрешности измерений.

Увеличивая и далее количество прогонов теста Business Winstone 2004, можно добиться того, чтобы доверительные интервалы не перекрывались. В нашем случае для этого потребовалось осуществить 12 прогонов (табл. 4). Как видим, результат процессора Intel Pentium Processor Extreme Edition 840 оказался несколько ниже результата процессора Intel Pentium D 840 (в данном случае технология Hyper-Threading не пошла на пользу), причем корректной будет следующая интерпретация результатов тестирования: истинное значение результата теста Business Winstone 2004 для процессора Intel Pentium Processor Extreme Edition 840 с вероятностью 95% лежит в интервале от 22,49 до 22,63; истинное значение результата теста Business Winstone 2004 для процессора Intel Pentium D 840 с вероятностью 95% лежит в интервале от 22,63 до 22,83.

Таким образом, для того чтобы корректно сравнивать друг с другом компьютеры или отдельные компоненты на основе программных тестов, необходимо задавать столько прогонов тестов, чтобы доверительные интервалы не перекрывали друг друга.

Вычисление погрешности в случае расчета интегральной оценки по совокупности тестов

Нередко в ходе тестирования компьютеров или отдельных подсистем ПК требуется получить интегральный результат тестирования на основе совокупности тестов. К примеру, при тестировании видеокарт могут использоваться разные игровые бенчмарки, а для одного и того же бенчмарка — различные настройки видеодрайвера и разрешения монитора. В итоге получается не один, а целая совокупность результатов, на основе которых видеокарты сравниваются друг с другом. При этом для того, чтобы можно было сравнивать тестируемые системы не по отдельным тестам, а по совокупности всех тестов, нужно свести все результаты тестов к единой интегральной оценке, которая могла бы служить мерой производительности тестируемой системы. То есть если имеется n различных тестов и $x_1, x_2...x_n$ — совокупность результатов этих тестов, то интегральная оценка производительности выражается как некоторая функция:

$$R = f(x_1, x_2...x_n).$$

Тип функции $f(x_1, x_2...x_n)$ зависит от выбранного алгоритма определения интегральной оценки. Это может быть и банальное усреднение всех результатов (среднее арифметическое), и среднее геометрическое, и среднее взвешенное. В конечном счете все определяется тем, насколько выбранный алгоритм нахождения интегральной оценки соответствует действительности.

Точно так же, как для отдельного теста определяются среднее значение и доверительный интервал, для корректного сравнения результатов на основе интегральной оценки необходимо уметь рассчитывать для нее доверительный интервал.

Рассмотрим простой пример. Пусть имеются два теста (Business Winstone 2004 и Multimedia Content Creation Winstone 2004), по совокупным результатам которых требуется получить интегральную оценку производительности ПК. Предположим, что в качестве такой интегральной оценки рассматривается среднее геометрическое результатов обоих тестов, то есть если в первом тесте получен результат x_1, а во втором — результат x_2, то интегральный результат определяется по формуле $\sqrt{x_1 \cdot x_2}$.

Предположим также, что каждый тест запускался пять раз, то есть для каждого теста имеется выборка с объемом, равным пяти (табл. 5).

Наиболее простой и правильный подход заключается в том, чтобы определить интегральную оценку в каждом из пяти прогонов теста и по выборке для интегральной оценки рассчитать среднее значение и доверительный интервал. Тогда окончательный результат будет записан в виде:

$$x = \bar{x} \pm t(p,n)\frac{S_n}{\sqrt{n}},$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$S_n = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n-1}},$$

где $\bar{x}$ — среднее по выборке интегральной оценки, S_n — стандартное отклонение по выборке интегральной оценки, $t(n, p)$ — коэф-

Таблица 4. Средние значения и доверительный диапазон, рассчитанные для десяти прогонов теста Business Winstone 2004

Процессор	Результаты отдельных прогонов теста												Результат
Intel Pentium Processor Extreme Edition 840	22,7	22,6	22,7	22,5	22,3	22,5	22,5	22,6	22,7	22,5	22,6	22,5	22,56±0,07
Intel Pentium D 840	22,5	22,6	22,7	22,5	22,9	22,5	22,5	22,6	22,7	22,5	22,6	22,7	22,73±0,10

Таблица 5. Расчет среднего значения и доверительного интервала для интегральной оценки производительности

Тесты	Результаты					Среднее значение	Доверительный интервал
Business Winstone 2004	22,3	22,6	22,7	22,5	22,7		
Multimedia Content Creation Winstone 2004	31,8	31,9	32	31,9	31,9		
Интегральная оценка производительности	26,63	26,85	26,95	26,79	26,91	26,827	0,156

Таблица 6. Расчет среднего значения и стандартного отклонения для интегральной оценки производительности

Тесты	Результаты					Среднее значение	Стандартное отклонение
Business Winstone 2004	22,3	22,6	22,7	22,5	22,7	22,56	0,167
Multimedia Content Creation Winstone 2004	31,8	31,9	32	31,9	31,9	31,90	0,071
Интегральная оценка производительности						26,827	0,104

Таблица 7. Нахождение стандартного отклонения для интегральной оценки производительности

$f(x_1, x_2 \ldots x_n)$	S_Σ
$\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$	$\frac{\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}}{n} \cdot \sqrt{\sum_{i=1}^{n}\left(\frac{S_i}{x_i}\right)^2}$
$\frac{x_1 + x_2 + \ldots + x_n}{n}$	$\frac{1}{n} \cdot \sqrt{\sum_{i=1}^{n} S_i^{\,2}}$

фициент Стьюдента для доверительной вероятности p и объема выборки n.

Однако при расчете среднего значения и погрешности интегральной оценки производительности может оказаться, что известны лишь среднее значение и доверительный диапазон (или стандартное отклонение) каждого теста в отдельности, но неизвестны результаты прогонов тестов. Или, к примеру, первый тест для уменьшения доверительного интервала прогоняют десять раз, а второй — только пять раз. Понятно, что в данном случае нельзя говорить о выборке интегральной оценки производительности, поэтому погрешность интегральной оценки производительности оценивают иначе.

Если интегральная оценка производительности является функцией $f(x_1, x_2 \ldots x_n)$ от результатов отдельных тестов, а S_i — стандартное отклонение в i-м тесте, то стандартное отклонение интегральной оценки производительности рассчитывается по формуле:

$$S_\Sigma = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial f(x_1, x_2 \ldots x_n)}{\partial x_i} S_i\right)^2},$$

где

$$\frac{\partial f(x_1, x_2 \ldots x_n)}{\partial x_i}$$

частная производная функции $f(x_1, x_2 \ldots x_n)$ по аргументу x_n.

В частности, для приведенного выше примера с двумя тестами и нахождением интегральной оценки производительности как среднего геометрического результатов тестов нетрудно показать, что:

$$S_\Sigma = \frac{1}{2} \cdot \sqrt{x_1 \cdot x_2} \cdot \sqrt{\left(\frac{S_1}{x_1}\right)^2 + \left(\frac{S_2}{x_2}\right)^2}.$$

Рассчитанные по указанным формулам значения стандартного отклонения интегральной оценки производительности приведены в табл. 6. Конечный результат для интегральной оценки производительности в данном случае можно записать в виде:

$$x = \bar{x} \pm S_\Sigma.$$

Как видим, погрешность, определяемая по выборке интегральной оценки производительности, немного отличается от погрешности, которая находится через стандартные отклонения.

В заключение приведем формулы для расчета стандартного отклонения по совокупности тестов для функций $f(x_1, x_2 \ldots x_n)$, определяемых как среднее арифметическое и среднее геометрическое результатов отдельных тестов (табл. 7).