КВАРТИЛЬ: какие формулы расчета использует Excel

Квартиль — одна из статистик, используемая при описании выборок (подробнее о различных статистиках см. Определение среднего значения, вариации и формы распределения. Описательные статистики). В то время как медиана разделяет упорядоченный массив пополам, квартили разбивают набор данных на четыре части. Первый квартиль – это число, разделяющее выборку на две части: 25% элементов меньше, а 75% — больше значения первого квартиля. Третий квартиль — это число, разделяющее выборку также на две части: 75% элементов меньше, а 25% — больше третьего квартиля.

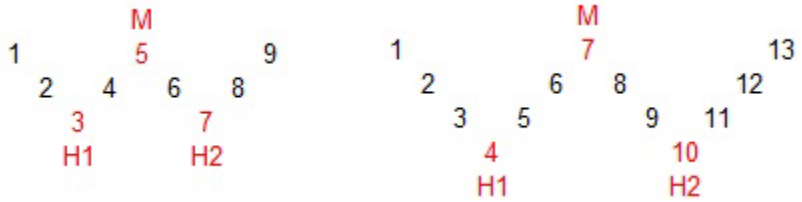

Рис. 1. 5-числовые сводки: M – медиана, H1 и H2 – сгибы (они же квартили)

Для расчета квартилей в Excel2007 и более ранних версиях использовалась функция КВАРТИЛЬ. Начиная с версии Excel2010 применяются две функции: КВАРТИЛЬ.ВКЛ и КВАРТИЛЬ.ИСКЛ (функция КВАРТИЛЬ оставлена для совмещения с более ранними версиями Excel; эта функция возвращает те же значения, что и КВАРТИЛЬ.ВКЛ). Эти две функции возвращают различные значения, но я нигде не нашел, какой алгоритм они используют при расчетах. Замечу, что для корректной работы функций данные можно не упорядочивать.

Изучение литературы показало, что в отличие от большинства других статистик, единодушия в методике расчета квартилей нет)) Я нашел упоминание о девяти различных подходах…

Начнем с **метода Джона Тьюки**, описанного им в, уже ставшем классическом, труде Анализ результатов наблюдений. Разведочный анализ, изданном в 1977 г. Он начинает с введения трех сводок, характеризующих выборку: минимальное, максимальное значения и медиана. Далее он продолжает: «Если мы хотим добавить еще два числа, чтобы образовать 5-числовую сводку, то естественно определять их подсчетом до половины расстояния от каждого из концов к медиане. Процесс нахождения медианы, а затем и этих новых значений можно представить себе, как складывание листа бумаги. Поэтому эти новые значения естественно назвать *сгибами»* (англ. – hinge; рис. 1). Мы их называем квартилями.

Такие рисунки выглядят очень аккуратно, если число элементов выборки N = 4k + 1, например, 9, 13, 17… Но как быть, если в выборке 12 или 19 элементов? Наглядную картину представил Jon Peltier в серии заметок в своем блоге. Упорядочим элементы случайной выборки и разместим их над линейкой (рис. 2; случайная выборка, элементы которой упорядочены называется вариационным рядом). Серые числа под линейкой – индекс ряда (Джон зачем-то в качестве выборки – над линейкой – взял ряд целых чисел; наверное, чтобы запутать нас). Красное число над рядом – значение сводки; если оно дробное, значит полученное значение является интерполяцией между соседними значениями. Мы определяем медиану, как среднее значение набора данных, а первую квартиль – как медиану нижней половины данных.

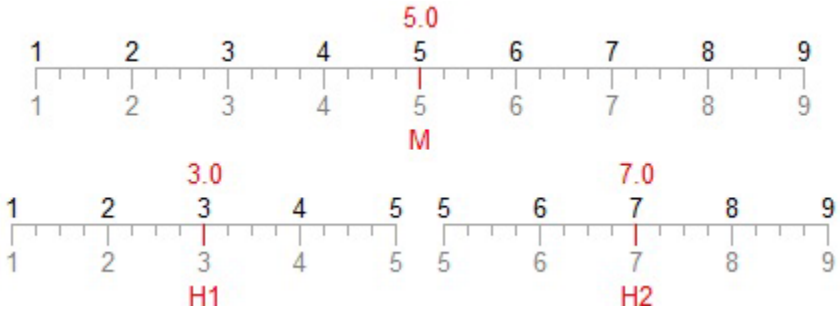

Рис. 2. Инклюзивные квартили

Когда Джон Тьюки впервые предложил такой подход, он решил, что медиана (если число элементов в выборке нечетное) должна быть включена как в нижнюю (левую на рисунке), так и в верхнюю

половинку данных при определении медиан этих половинок, то есть сгибов. Поэтому такой подход и называется инклюзивным (с включением).

Эксклюзивный подход. Некоторым статистикам не нравится, что медиана учитывается дважды. Они решили, что сгибы должны быть определены как медианы верхней и нижней половин набора данных, из которых срединное значение исключено (рис. 3). Такой взгляд отстаивали Moore и McCabe, или кратко M&M. Если набор данных содержит четное количество значений, инклюзивные и эксклюзивные квартили равны, так как нет элемента выборки (соответствующего центральной медиане), который можно было бы включить или исключить из рассмотрения. Для нечетного числа элементов, инклюзивные сгибы всегда ближе к медиане.

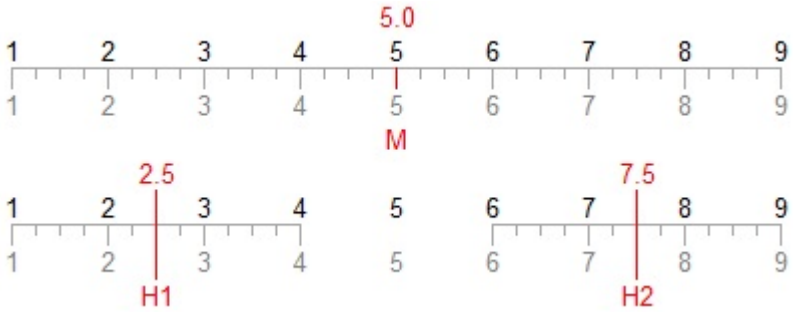

Рис. 3. Эксклюзивные квартили

Третий подход – компромисс между Тьюки и M&M – называется Эмпирическая функции распределения или **Интегральная функция распределения** (английская аббревиатура **CDF**). В случае нечетного числа значений в наборе данных, следует включить или исключить медиану, ориентируясь на то, чтобы оставшиеся половинки содержали нечетное число элементов. Например, если в выборке 9 элементов, медиану следует включить, а при 11 элементах – исключить. В обоих случаях половинки будут содержать по 5 элементов. Преимущество этого компромисса заключается в том, что в качестве значения квартиля всегда получается один из элементов набора данных (а не среднее значение двух соседних элементов). CDF является методом по умолчанию в статистическом пакете SAS.

Все возможные случаи N. Мы не всегда можем изобразить данные в W-образной форме, как на рис. 1, поэтому удобнее пользоваться линейкой. В общем случае возможны четыре варианта по числу элементов в выборке: N = 4k, N = 4k + 1, N = 4k + 2, N = 4k + 3… и три подхода к расчету квартилей: Тьюки, M&M, CDF (рис. 4–7).

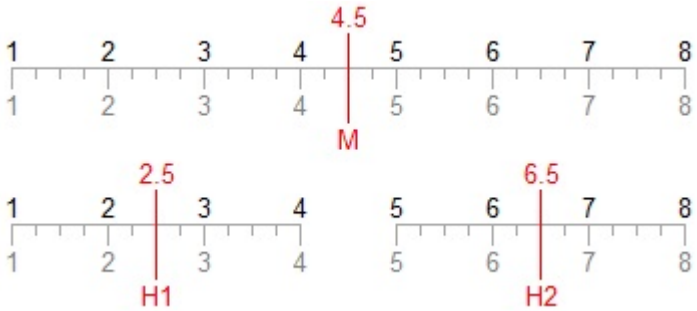

Рис. 4. Число элементов в выборке N = 4k; все три метода дают одинаковые значения квартилей

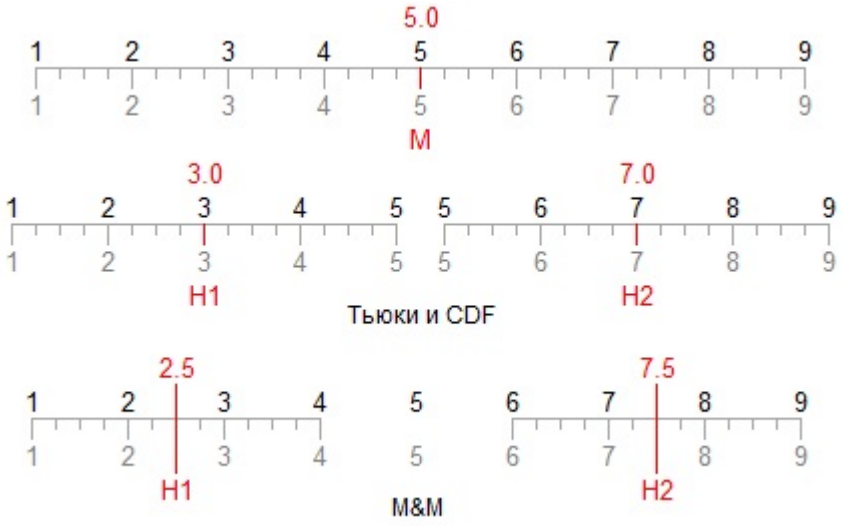

Рис. 5. Число элементов в выборке N = 4k + 1; M&M дает значения, отстоящие дальше от медианы

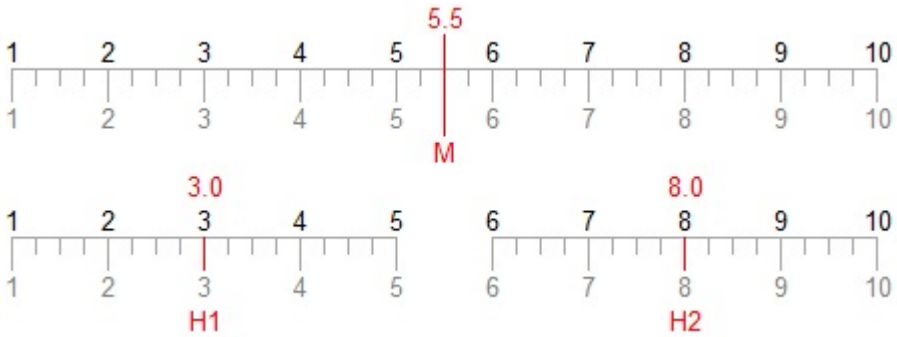

Рис. 6. Число элементов в выборке N = 4k + 2; все три метода дают одинаковые значения квартилей

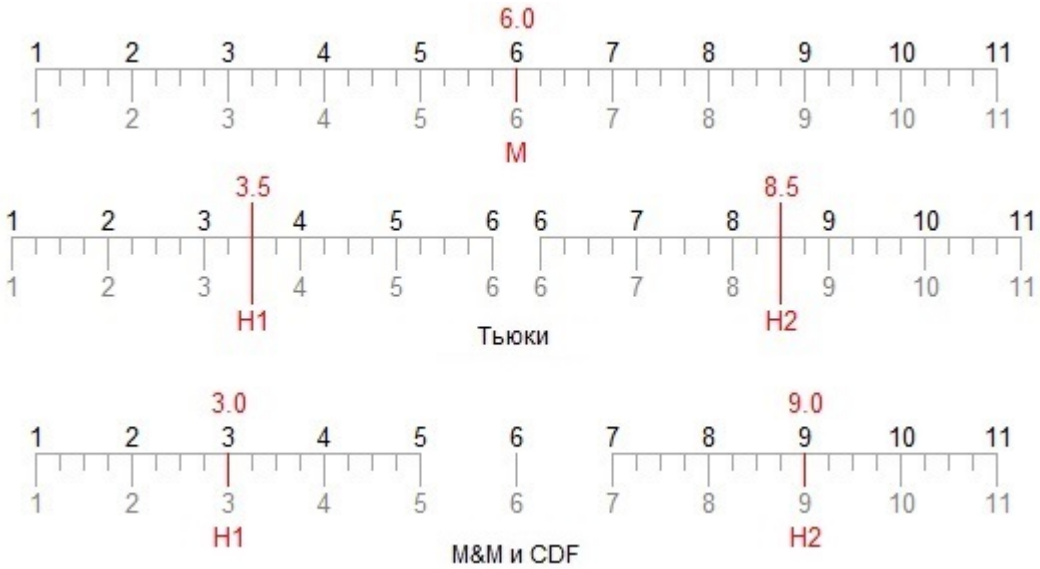

Рис. 7. Число элементов в выборке N = 4k + 3

Методы интерполяции. Помимо трех описанных выше методов, применяют и целый ряд индексных алгоритмов. Мы рассмотрим три из них. Первый индекс во всех методах равен 0, а последний – N–1, N, N + 1. Например, для N=8 индексированные ряды представлены на рис. 8.

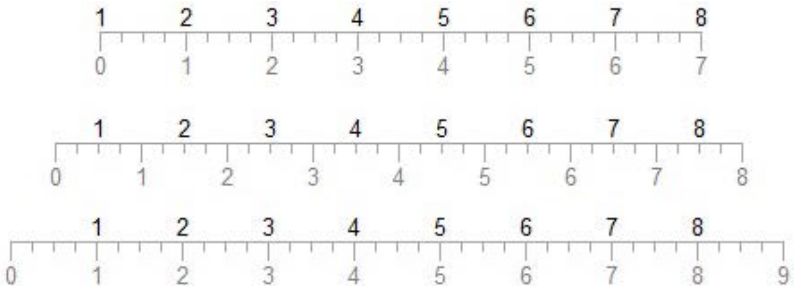

Рис. 8. Индексные ряды на основе N–1, N и N + 1 для N = 8

Положение перцентиля *p* – доля длины индексной линии, или p(N–1), pN, p(N+1), соответственно. p = 0,25 соответствует первому квартилю, а p = 0,75 – третьему. Ниже наглядно представлен расчет квартилей при различном числе элементов в выборке и трех методах интерполяции на основе N–1, N и N + 1 (рис. 9, 11–13). Обратите внимание, что рассчитанные числа (по формулам справа от линеек) являются не значениями квартилей, а значениями индексов квартилей. Над линейками показано значение квартилей для ряда значений {1, 2, 3, 4, 5, 6, 7, 8}.

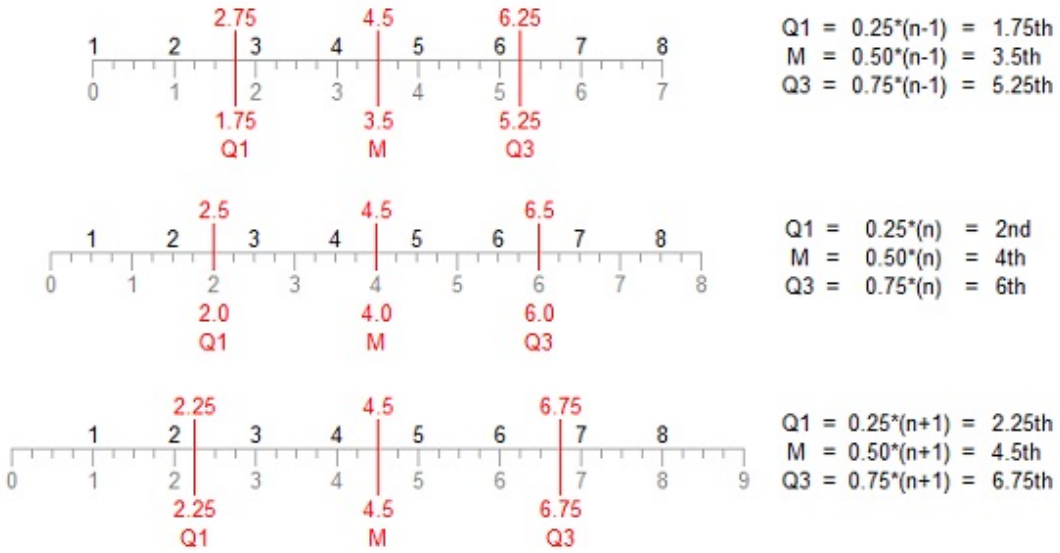

Рис. 9. Число элементов в выборке N = 4k

Если, например, наша выборка {2, 3, 5, 8, 11, 12, 14, 17}, то расчет квартилей на основе N–1-метода даст индексы 1,75, 3,5 и 5,25, и значения квартилей 4,5, 9,5 и 12,5 (рис. 10).

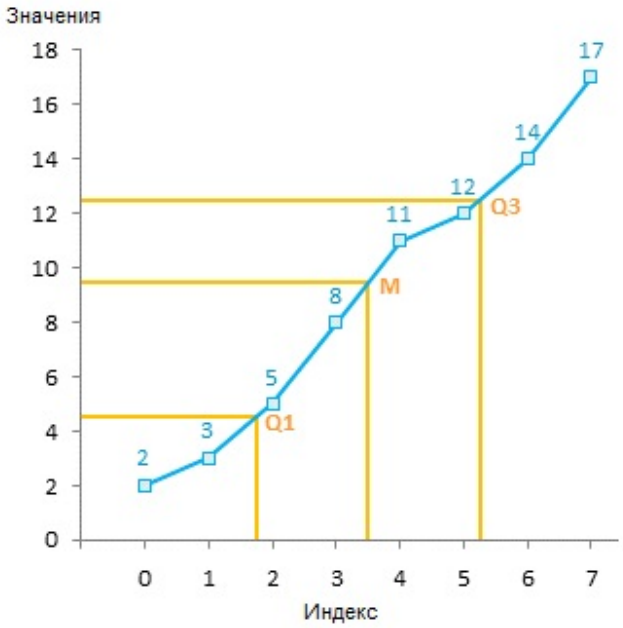

Рис. 10. От индексов к значениям квартилей для N–1-метода и N = 4k

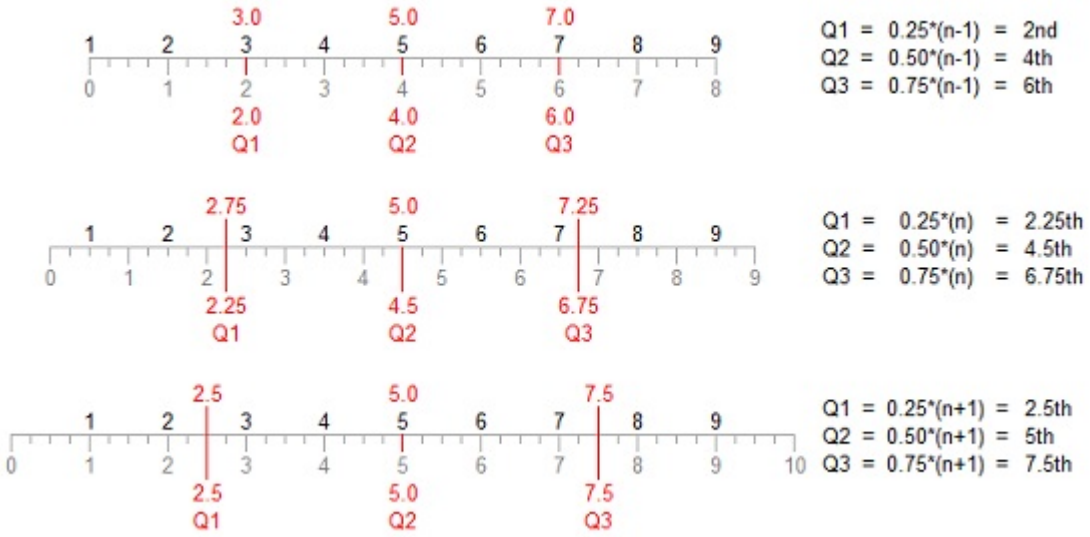

Рис. 11. Число элементов в выборке N = 4k + 1

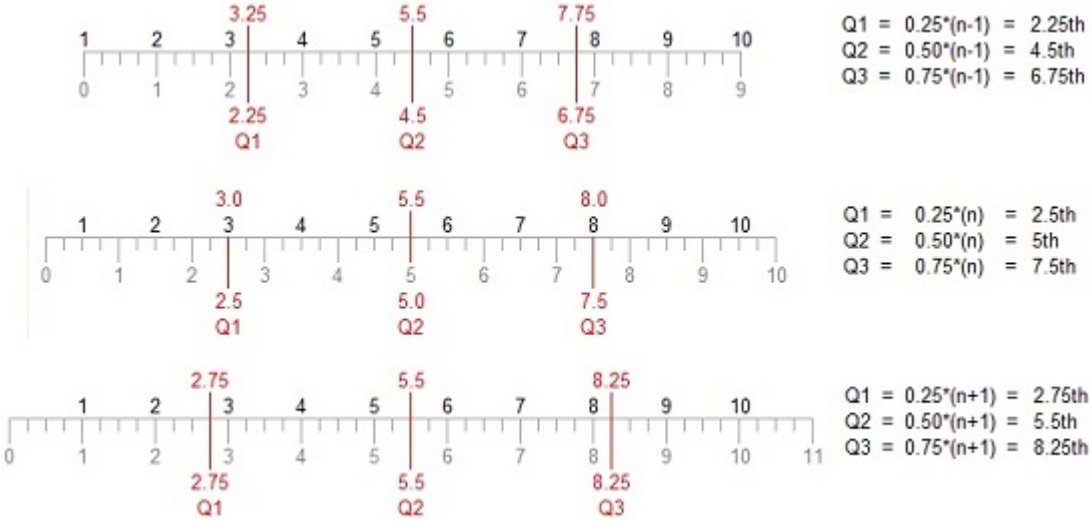

Рис. 12. Число элементов в выборке N = 4k + 2

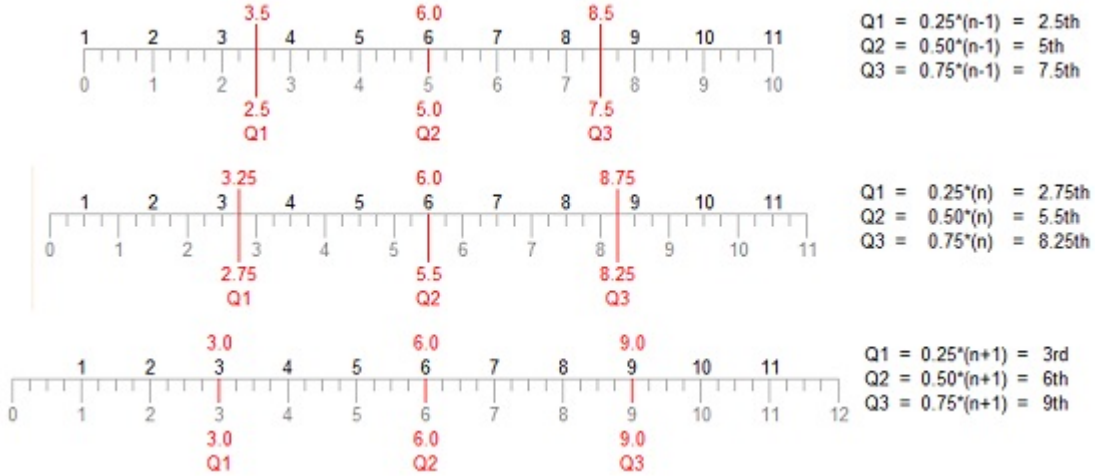

Рис. 13. Число элементов в выборке N = 4k + 3

Какой алгоритм считать стандартным для вычисления квартилей?

В 1996 году Роб Дж. Хиндман и Янан Фан опубликовали статью в American Statistician под названием Квантили выборок в статистических пакетах. В ней они рассматривали различные алгоритмы расчета квантилей (квартили – это частный случай квантилей). Их целью было указать методологию, которая могла бы стать стандартом для поставщиков статистического программного обеспечения, чтобы расчет квартилей не зависел от типа пакета. В статье они описали девять методов для расчета

квантилей. Таблица показывает некоторые статистические пакеты и используемые в них алгоритмы (рис. 14; таблица, этот раздел заметки и код VBA ниже базируются на тексте с сайта Bacon Bits). Обратите внимание, что R и Maple применяют весь спектр алгоритмов.

	Excel	R	SAS	Minitab	Maple	JMP
Method 4 (Linear CDF)		type=4	PCTLDEF=1		method=3	
Method 5 (Piecewise Linear)		type=5			method=4	
Method 6 (N+1)	Квартиль.ИСКЛ	type=6	PCTLDEF=4	Default	method=5	Default
Method 7 (N-1)	Квартиль.ВКЛ Квартиль	type=7			method=6	
Method 8 (N+1/3)		type=8			method=7	
Method 9 (N + 1/4)		type=9			method=8	

Рис. 14. Алгоритмы, используемые в статистических пакетах

Кстати, Хиндман и Фан в завершении своей статьи рекомендовали метод 8 в качестве стандарта для статистических пакетов. По их мнению, этот метод оценки квантиля не зависит от распределения, что делает его наиболее приемлемым для расчета.

Расчет квартилей в Excel

Функция Excel КВАРТИЛЬ.ИСКЛ использует следующую формулу для расчета квартилей:

$$(1)\ Q_p = (1 - (x - i) \cdot A_i + (x - i) \cdot A_{i+1}$$

где Q_p – p-й квантиль: p = 0 – для минимального значения, 0,25 – для первого квартиля, 0,5 – для медианы, 0,75 – для третьего квартиля, 1 – для максимального значения;

x – индекс квантиля (может быть дробным); $x = (n+1)p$, где n – число элементов в выборке; обратите внимание на $(n+1)$, поэтому метод и называется N+1-интерполяция;

i – индекс элемента в упорядоченной выборке; самое большое целое всё еще меньшее, чем x;

$A_1, A_2, \ldots, A_i, A_{i+1}, \ldots, A_n$ – элементы случайной выборки, упорядоченной по возрастанию.

Формула для КВАРТИЛЬ.ВКЛ отличается только методом расчета x: $x = (n-1)p+1$; обратите внимание на $(n-1)$, поэтому метод называется N–1-интерполяция. Подробнее с работой формул можно ознакомиться в приложенном Excel-файле на листе *Формулы*.

Расчет квартилей в R и SAS

Функция *quantile* в R использует все девять алгоритмов расчета квантилей, в соответствии с нумерацией, предложенной Hyndman and Fan в работе 1996 г. (рис. 15; если вы не знакомы с R, рекомендую начать с Алексей Шипунов. Наглядная статистика. Используем R!). Квантиль при i-м методе расчета:

$$(2)\ Q_i(p) = (1 - \gamma)x_j + \gamma x_{j+1}$$

где i – номер метода, 1 ≤ i ≤ 9, (j–m)/n ≤ p < (j–m+1)/n, x_j – j-ый порядковый элемент упорядоченного ряда, n – размер выборки, γ является функцией двух параметров: j = floor(np + m) и g = np + m – j, где floor – функция возвращающая наибольшее целое, но всё еще меньшее, чем аргумент функции (аналог в Excel – ОКРВНИЗ.МАТ), m – константа, определяемая типом алгоритма расчета квантиля. Если вас интересуют подробности, обратитесь к справочной системе R.

SAS предлгает 5 методов расчета квантилей. И вновь за подробностями я отсылаю вас к справочной системе.

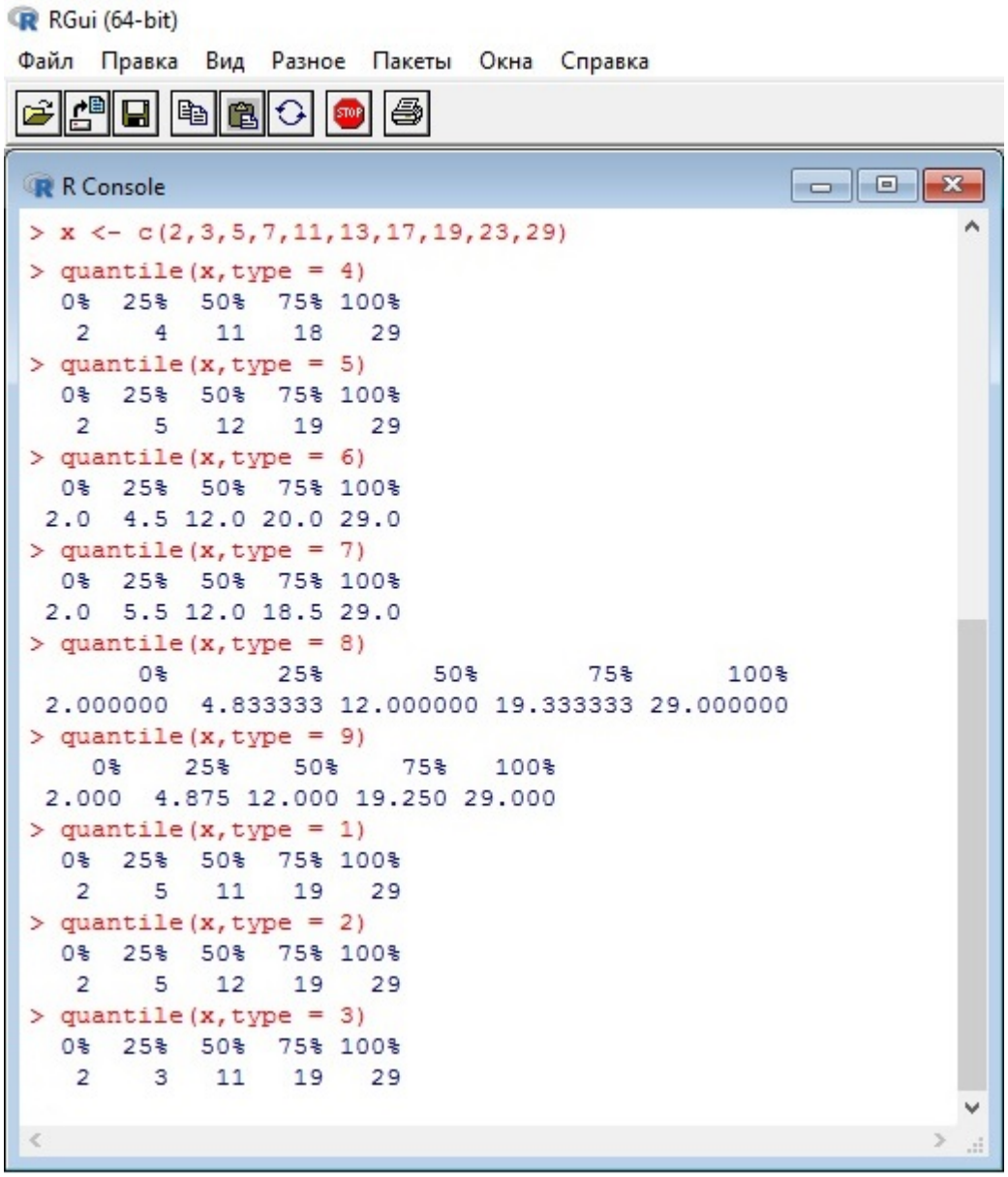

Рис. 15. Расчет квартилей в R девятью способами

Расчет квартилей в Excel любым методом с помощью VBA

Ниже представлен код пользовательской функции, которая позволяет воспроизвести любой из шести методов, перечисленных в таблице на рис. 14. Даже если у вас Excel 2007, и вам недоступна функция КВАРТИЛЬ.ИСКЛ, вы сможете рассчитать квартиль шестым методом с помощью этой функции.

```
Function Quantile(MyRange As Range, p As Double, Optional m As Variant)
'Mike Alexander: www.datapigtechnologies.com
'Based on Code originally posted by Jerry W. Lewis (former Excel MVP)
'**********************************************************************
'This function will replicate various quantile calcuations found in statistical software packages.
'Calculation is determined by the Hyndman-Fan method used.

'Hyndman-Fan Method 4 Replicates: SAS(PCTLDEF=1), R(type=4), Maple(method=3)
'Hyndman-Fan Method 5 Replicates: R(type=5), Maple(method=4)
'Hyndman-Fan Method 6 Replicates: Excel(QUARTILE.EXC), SAS(PCTLDEF=4), R(type=6), Minitab, SPSS, BMDP, JMP, Maple(method=5)
'Hyndman-Fan Method 7 Replicates: Excel (QUARTILE and QUARTILE.INC), R(type=7), S-Plus, Maxima, Maple(method=6)
'Hyndman-Fan Method 8 Replicates: R(type=8), Maple(method=7)
'Hyndman-Fan Method 9 Replicates: R(type=9), Maple(method=8)
```

```
'*****************************************************************************
'Call function from Excel Spreadhseet by entering
'=Quantile(Range, p, m)

'Enter p as the fraction of the population (.25 for quartile 1, .75 for quartile 3, etc....)
'Enter m as the Hyndman-Fan Quantile method number (4, 5, 6, 7, 8 or 9)
'If m is left blank, the function will use method 6 by default
'*****************************************************************************
Dim n As Long
Dim i As Long
Dim QDef As Double
Dim x As Double

'Identify method and set the interpolation basis used
Select Case m
   Case Is = 4
      QDef = 0
   Case Is = 5
      QDef = 0.5
   Case Is = 6
      QDef = p
   Case Is = 7
      QDef = 1 - p
   Case Is = 8
      QDef = (p + 1) / 3
   Case Is = 9
      QDef = (p + 1.5) / 4
   Case Else 'Use Hyndman-Fan 6 by default
      QDef = p
End Select

'Count values within MyRange and calculate the required position index
n = WorksheetFunction.Count(MyRange)
x = n * p + QDef
i = WorksheetFunction.Max(WorksheetFunction.Min(Fix(x), n), 1)

'Perform interpolation and return answer
If (x - i) >= 0 And i < n Then
   Quantile = (1 - (x - i)) * WorksheetFunction.Small(MyRange, i) + (x - i) *
WorksheetFunction.Small(MyRange, i + 1)
Else
   Quantile = WorksheetFunction.Small(MyRange, i)
End If


End Function
```

После того, как вставите код в стандартный модуль книги, вы сможете использовать функцию (рис. 16): =Quantile(MyRange; P; M), где MyRange – диапазон, включающий выборку (можно оставить его неупорядоченным); P – статистика: 0 – минимум, 0,25 – 1-й квартиль, 0,5 – медиана, 0,75 – 3-й квартиль, 1 – максимум; возможно введение иных значений в диапазоне от 0 до 1; M – номер метода из таблицы на рис. 14.

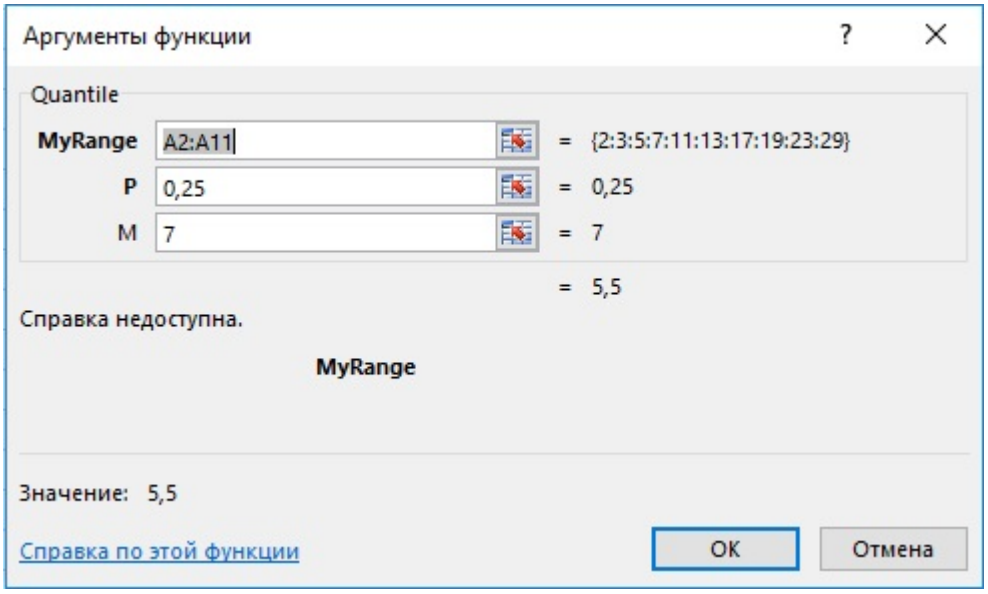

Рис. 16. Синтаксис пользовательской функции *Quantile*

В таблице (рис. 17) приведен расчет квартилей по всем методам. Обратите внимание, как метод 8 (который Хиндман и Фан рекомендуют в качестве стандарта) вычисляет квартили, которые ложатся между значениями, вычисляемыми по методам 6 и 7. Действительно, метод 8 дает наиболее сбалансированный набор квартилей.

Элементы выборки	Номер метода по Хиндману и Фану	Квартили			Межквартильный интервал
		0,25	0,50	0,75	
2	4	4,00	11,00	18,00	14,00
3	5	5,00	12,00	19,00	14,00
5	6	4,50	12,00	20,00	15,50
7	7	5,50	12,00	18,50	13,00
11	8	4,83	12,00	19,33	14,50
13	9	4,88	12,00	19,25	14,38
17					
19					
23					
39					

Рис. 17. Значения квартилей, вычисленные различными методами

Сравнение алгоритмов вычисления квартилей

Стандартом де-факто вычисления квартилей в статистических пакетах и Excel является метод 6 на основе N+1-интерполяции. Если вы хотите, чтобы ваши данные были одинаковыми при использовании различных инструментов, используйте именно этот метод. В Excel он лежит в основе работы функции КВАРТИЛЬ.ИСКЛ. К сожалению, этот метод приводит к увеличению межквартильного интервала. Для нашего примера (рис. 17) с 13,0 до 15,5. Если сравнить все пять методов расчета (рис. 18), то видно, что минимальный межквартильный интервал соответствует методу 7, а максимальный – методу 6. На что это влияет мы рассмотрим в заметке Визуализация статистических данных с помощью диаграммы ящик с усами. Если же вы используете только Excel рекомендую метод 7 на основе N–1-интерполяции. Это позволит вам оперировать с самым узким межквартильным интервалом.

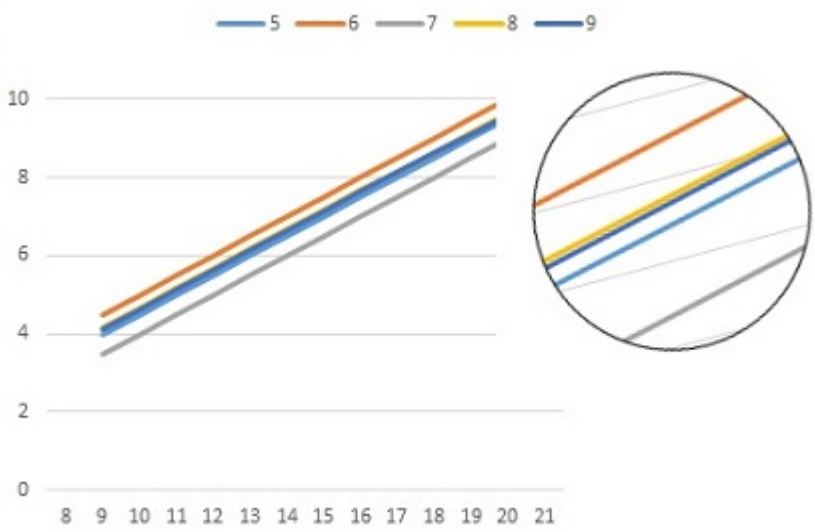

Рис. 18. Влияние алгоритма расчета квартилей на межквартильный интервал; цифры от 5 до 9 – номер метода