РАСЧЕТ ПРОЦЕНТИЛЕЙ И МЕЖКВАРТИЛЬНОГО РАЗМАХА

Квантиль – это значение, которое признак (случайна величина) не превосходит с заданной нами вероятностью.

Для расчета квантилей используются **порядковые статистики**.

Например, если у нас есть вариационный ряд:

$x_{(1)} < x_{(2)} < x_{(3)} < \ldots < x_{(n)}$

то x_i, значение признака для объекта (наблюдения), стоящего на i-м месте, и будет i-й порядковой статистикой.

Выборочные квантили – они же процентили – полезны для более полного понимания разброса в данных.

К примеру, если вы знаете, что ваша оценка за тест = 35 из 50, то само по себе значение не говорит однозначно, хорошо или плохо вы справились. Но если вы знаете также, что 90% проходивших тест решают его не лучше, чем на 35 баллов, то понять, насколько хорошо вы справились, будет проще.

Распространенные статистики, рассчитываемые с использованием процентилей:

Медиана распределения –50-й процентиль, значение, которое делит вариационный ряд пополам: 50% объектов принимают значения ниже медианного, 50% – выше медианного. В примере с тестом это означало бы, что 50% студентов, сдававших тест, сдали его хуже, чем на медианный балл, а другие 50% сдали его лучше, чем на медианный балл.

Межквартильный размах – это границы, в которых находятся 50% «типичных» (распространенных) значений признака, определяемые как разность между 75% и 25% процентилями.

В примере с тестом это диапазон результатов 50% «средних» студентов (не включая 25% «лучших» и 25% «худших»).

Как посчитать процентили?

Нет единственного (общепринятого) способа определения процентилей. Расмотрим четыре основных: два простых и два посложнее. Начнём с простых и рассмотрим их на примере небольшой выборки данных.

Допустим, есть такой ряд данных, уже упорядоченных по возрастанию (если у вас есть данные, не упорядоченные по возрастанию, на первом шаге нужно будет упорядочить, т.е., построить вариационный ряд).

3, 5, 7, 9, 12, 21, 25, 30

Задача: рассчитать на этом ряду данных 25-й процентиль.

Способ 1.

Согласно первому способу расчета 25-й процентиль – это наименьшее из значений, **превышающих** 25% значений признака.

В нашем вариационном ряду 8 наблюдений.

Значит, 25% значений – это (0,25) * 8 = 2 значения. В нашем ряду это значения 3 и 5.

Наименьшее значение, превышающее эти два – **7**. Это и будет значением 25 процентиля.

3, 5, **7**, 9, 12, 21, 25, 30

Способ 2.

Согласно второму способу расчета, 25-й процентиль это наименьшее из значений, **больше либо равное** границе 25% значений признака.

На первом шаге всё так же, как и в первом случае: поскольку в нашем вариационном ряду 8 наблюдений, 25% значений – это (0,25) * 8 = 2 значения.

В нашем ряду это, как мы знаем, значения 3 и 5.

Но поскольку согласно второму способу нам нужно наименьшее из значений, **больше либо равных** 25%, то значение 25-го процентиля будет **5**:

3, **5**, 7, 9, 12, 21, 25, 30

На больших выборках такие способы расчета дают очень похожие результаты, но на небольших выборках, как мы видим, результаты могут отличаться существенно.

Кроме того, есть еще одна проблема: непонятно, как округлять, если порядковый номер статистики получится не целым, а дробным.

Для этого есть третий способ.

Способ 3, взвешенный.

Здесь идея в том, что мы не берем какое-то из значений, уже имеющихся в вариационном ряду, а рассчитываем значение на основе информации, которую содержит вариационный ряд.
Разберём расчет 25 процентиля «взвешенным» способом на нашем примере.

Шаг 1.

Сначала возьмем наш вариационный ряд из 8 объектов и присвоим каждому значению ранг от 1 до 8, где первый ранг будет присвоен наблюдению с наименьшим значением признака, а ранг 8 – объекту с наибольшим значением признака.

Значение	**Ранг**
3	1
5	2
7	3
9	4
12	5
21	6
25	7
30	8

Шаг 2. Считаем ранговую позицию для значения 25-го процентиля. Для этого используется формула:

R = P/100 x (N + 1), где

R – ранговая позиция
P – нужный процентиль
N – объем выборки.

На нашем ряду ранговая позиция для 25-го процентиля считается так:

R = 25/100 x (8 + 1) = 9/4 = 2.25 **(*)**

Если бы ранг был целым числом, как в предыдущем примере, то значение, стоящее на порядковой позиции R, и было бы 25-м процентилем. Но здесь получается, что 25-й процентиль – значение, стоящее на 2,25-й ранговой позиции. В вариационном ряду такого значения нет, значит придется его почитать.

Шаг 3. Если R дробное, то значение 25-го процентиля рассчитывается по следующей схеме:

1. Определяем целую часть от R.
   В случае **(*)** это 2. Обозначим ее IR (от integer portion of R).

2. Определяем дробную часть от R.
   В случае **(*)** это 0,25. Обозначим ее FR (от fractional portion of R).

3. Находим значения, стоящие на ранговых позициях IR и (IR+1).
   В нашем примере это значения, стоящие на позициях 2 и 3 (числа 5 и 7 соответственно).

4. Рассчитываем значение процентиля так:

   4.1. Берем разность значений, стоящих на ранговых позициях IR и (IR+1).
   4.2. Затем умножаем эту разность на FR
   4.3. Прибавляем полученное значение к наименьшему из двух компонентов разности.

   В нашем примере: (7-5) * 0,25 + 5 = 5.5

Отдельный сюжет – о построении межквартильного размаха (МКР).

Можно рассчитывать МКР на основе 25 и 75 процентилей, или 1 и 3 квартилей, каждым из трех описанных способов:

МКР = Q3-Q1.

На наших данных:

Способ 1: МКР = 18

Способ 2: МКР = 16

Способ 3: МКР = 18,5

Любопытно, что SPSS и R имеют разные опции по умолчанию.

В R используется первый способ, в SPSS (если задать автоматический расчет МКР в процедуре EXPLORE [разведочный анализ]) – третий (взвешенный).

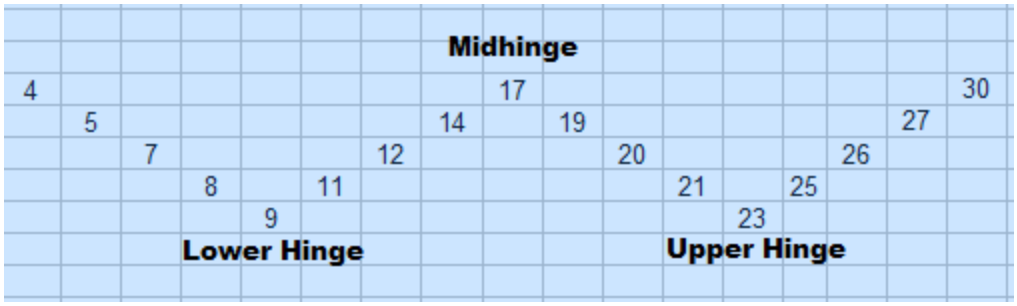
Но для МКР есть и четвертый способ.

Способ 4.

Tukey's hinges (в русскоязычной версии SPSS переведено как «сгибы Тьюки»), также называются «включающими квантилями» (inclusive quantiles)

Идея такая: «сгибы» строятся так, чтобы разделить вариационный рад на 4 части.

Нижний сгиб (lower hinge) – 25-й процентиль,
Средний сгиб (midhinge) – 50-й процентиль,
Верхний сгиб (upper hinge) – 75-й процентиль.



Обычно нижний сгиб соответствует первому квартилю, средний – медиане, верхний – третьему квартиль.
Но есть странное исключение: если при делении объема выборки на 4 в остатке остается 3, то сгибы не будут равны Q1 Q 3. Например, если выборка составляет 35 наблюдений, то 35/4 = 8 и 3 в остатке. Подробнее об исключении и о методе здесь (англ. яз): http://www.statisticshowto.com/upper-hinge-lower-hinge/

Вообще, если объем выборки – величина, кратная 4, то «сгибы» считаются очень просто. Если нет, то нужны дополнительные манипуляции. Как же посчитать сгибы?

1. Если объем выборки – нечетное число.

1. Упорядочиваем выборку

Например: 3, 5, 7, 9, 12, 21, 25, 30, 33

2. Находим медиану (срединное значение; в «нечетном случае» оно одно)

Например: 3, 5, 7, 9, **12**, 21, 25, 30, 33

3. Берем в скобки левую часть вариационного ряда, включая медиану.

(3, 5, 7, 9, **12)**, 21, 25, 30, 33

4. Находим медиану нижней части ряда (того, что в скобках)

(3, 5, **7**, 9, 12), 21, 25, 30, 33

5. Проделываем то же самое для верхней части ряда:

3, 5, 7, 9, (12, 21, **25**, 30, 33)

6. Получаем МКР = 25-7 = 13

2. Если объем выборки – четное число[1]

1. Упорядочиваем выборку

3, 5, 7, 9, 12, 21, 25, 30

2. Находим медианную точку и «режем» ряд при помощи слэша.

3, 5, 7, 9 / 12, 21, 25, 30

3. Считаем среднее арифметическое двух серединных значений нижней части ряда

3, 5, **[(5+7)/2=6]**, 7, 9 / 12, 21**,** 25, 30

[1] Работает, если оно делится на 4; что делать, если не делится – см. ссылку выше.

4. Проделываем то же самое для верхней части ряда: 3, 5, 7, 9 / 12, 21, **(23),** 25, 30

5. Получаем МКР = 23-6 = 17.

P.S.

Забавно, что SPSS при расчете процентилей используются два разных способа: взвешенный (способ 3) и «сгибы Тьюки».

Однако при автоматическом расчете межквартильного размаха (в процедуре EXPLORE - «разведочный анализ») по умолчанию используется только взвешенный способ.

Если не включать опцию расчета процентилей и оставить только статистики «по умолчанию», то мы получим МКР, рассчитанный взвешенным методом:

	Статистика	Стандартная ошибка
Среднее	14	3,55065
Медиана	10,5	
Минимум	3	
Максимум	30	
Размах	27	
Межквартильный размах	**18,50**	

(представлена сокращенная версия таблицы)

Но если включить дополнительную опцию расчета процентилей, то вместе с табличкой статистик будет рассчитана еще одна, и на ее основе можно будет вычислить два разных МКР: взвешенным методом и через сгибы.

	5	10	25	50	75	90	95
Взвешенное среднее (Определение 1)	3	3	5,5	10,5	24	.	.
Сгибы Тьюки			6	10,5	23		

На больших выборках не будет таких драматических различий, но на маленькой выборке мы получили 4 разных значения про подсчете четырьмя разными способами:

МКР1 = 18
МКР2 = 16
МКР3 (взвешенный) = 18,5
МКР4 (сгибы Тьюки) = 17