**Визуализация статистических данных с помощью диаграммы ящик с усами**

**Предыстория.** В связи с холодными днями в мае и июне 2017 г. я изучил климатические данные многолетних наблюдений, и хотел опубликовать заметку о статистическом взгляде на всё это безобразие. В качестве иллюстрации одной из своих идей я построил блочную диаграмму. Во время ее форматирования я обнаружил, что не вполне понимаю, по каким алгоритмам Excel рисует на ней границы квартилей. Углубившись в вопрос, я узнал много нового, о чем захотел поделиться с читателями блога. Но подумал, что описание тонких алгоритмов как бы повисает в воздухе, и поэтому решил начать с небольшого теоретического введения.

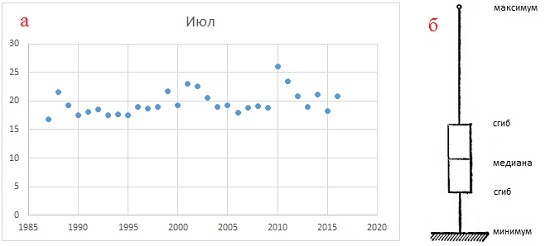


Рис. 1. Визуальное представление выборки: (а) диаграмма разброса показывает все элементы выборки; (б) 5-числовая сводка в виде графика ящик с усами

### 5-числовая сводка данных

Изучая выборку, мы можем визуализировать все ее значения (рис. 1а), а можем научиться выражать в сжатом виде наиболее часто встречающиеся общие особенности. В последнем случае используют сводки данных. Сегодня наиболее популярными сводками (или статистиками) являются среднее значение и среднеквадратичное отклонение (подробнее см. [Определение среднего значения, вариации и формы распределения. Описательные статистики](http://baguzin.ru/wp/?p=5381)). Однако, во-первых, так было не всегда, а во-вторых, не во всех случаях эти статистики лучше всего описывают выборку.

Любопытный пример приводит Нассим Николас Талеб (см. [Черный лебедь. Под знаком непредсказуемости](http://baguzin.ru/wp/?p=1533)). Если вы определяете средний рост 100 человек, то нахождение среди этой сотни самого высокого человека в мире (2 м 72 см) не сильно изменит среднее значение. А вот если вы определяете среднее богатство 100 человек, то наличие среди них Билла Гейтса ($86,8 млрд.) полностью исказит картину. В этих случаях более релевантной сводкой является медиана.

[Медиана](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B4%D0%B8%D0%B0%D0%BD%D0%B0_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)) – такое число, что ровно половина из элементов выборки больше него, а другая половина меньше него. Если случайные данные упорядочить по возрастанию, получится [вариационный ряд](https://ru.wikipedia.org/wiki/%D0%92%D0%B0%D1%80%D0%B8%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D1%80%D1%8F%D0%B4). Если такой ряд содержит нечетное число элементов, то медиана – среднее из них; если четное, то медиана – половина суммы двух средних элементов.

В своей, уже ставшей классической, работе [Анализ результатов наблюдений. Разведочный анализ](http://baguzin.ru/wp/?p=15897) Джон Тьюки предлагает 5-числовые сводки, которые включают: минимальное значение, первый квартиль, медиану, третий квартиль, максимальное значение. Процесс нахождения медианы, а затем квартилей можно представить себе, как складывание листа бумаги. Поэтому эти новые значения Тьюки называет *сгибами* (подробнее см. [КВАРТИЛЬ: какие формулы расчета использует Excel](http://baguzin.ru/wp/?p=17320)). Чтобы представить 5-чиловую сводку визуально, Тьюки предлагает построить график *ящик с усами* (рис. 1б).

### Диаграмма ящик с усами в Excel 2016

В предыдущих версиях Excel приходилось шаманить, чтобы представить статистические данные (см., например, [Биржевая диаграмма, она же блочная, она же ящичная](http://baguzin.ru/wp/?p=4349)). В качестве альтернативы я иногда использовал R (см. [Алексей Шипунов. Наглядная статистика. Используем R!](http://baguzin.ru/wp/?p=14666)). Сейчас же визуализация статистических данных выполняется в несколько кликов. Очень удобно! Вот только Microsoft почему-то сделал весьма краткое описание того, что же собственно отражают элементы диаграммы *ящик с усами* (см. [поддержку MS](https://support.office.com/ru-ru/article/%D0%A1%D0%BE%D0%B7%D0%B4%D0%B0%D0%BD%D0%B8%D0%B5-%D0%B1%D0%BB%D0%BE%D1%87%D0%BD%D0%BE%D0%B9-%D0%B4%D0%B8%D0%B0%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D1%8B-%D1%81-%D0%BE%D0%B3%D1%80%D0%B0%D0%BD%D0%B8%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D1%8F%D0%BC%D0%B8-%D0%B2%D1%8B%D0%B1%D1%80%D0%BE%D1%81%D0%BE%D0%B2-62f4219f-db4b-4754-aca8-4743f6190f0d)). Восполним же этот пробел))

Построим диаграмму *ящик с усами* на основе [данных](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B8%D0%BC%D0%B0%D1%82_%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D1%8B) о средних температурах июля за последние 30 лет (рис. 2).

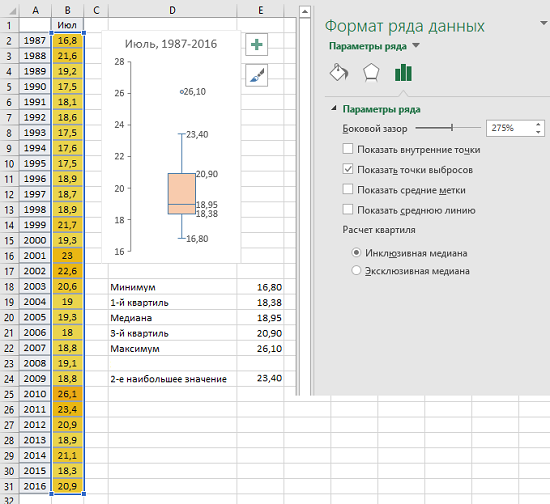


Рис. 2. Основные настройки диаграммы ящик с усами

На рисунке под диаграммой рассчитана 5-числовая сводка. На диаграмме:

* нижний ус соответствует минимальному значению в выборке данных,
* низ ящика – 1-му квартилю,
* линия внутри ящика – медиане,
* верх ящика – 3-му квартилю,
* верхний ус – второму максимальному значению (выбросы не учитываются); если бы данные не содержали выброса, верхний ус соответствовал бы максимальному значению,
* опциально отдельными точками показаны выбросы.

Рассмотрим параметры ряда подробнее (см. *Формат ряда данных* на правой части рис. 2). *Боковой зазор* определяет ширину ящика и ширину горизонтальных границ усов, а также расстояние между ящиками, если их несколько (рис. 3). Параметр *Показать внутренние точки* – говорит сам за себя (рис. 4). *Показать точки выбросов* – любопытно, что Excel просто перестает показывать выбросы, при этом никак не меняет параметры ящика и усов (рис. 5). *Показать средние метки* – наряду с медианной линией отражает в виде крестика среднее арифметическое значение по выборке (рис. 6). Видно, что среднее арифметическое смещено относительно медианы вверх (в сторону выброса). Среднее арифметическое более чувствительно к выбросам по сравнению с медианой. *Показать среднюю линию* – соединяет средние значения разных категорий. У меня не получилось отразить эти линии. А у Дмитрия Езипова [получилось](http://statanaliz.info/excel/diagrammy/110-diagramma-yashchik-s-usami-v-excel-2016).

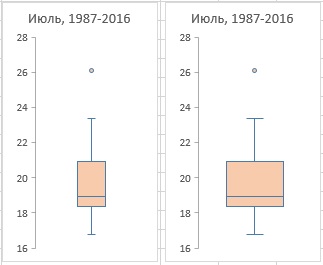


Рис. 3. Боковой зазор: слева – 300%, справа – 100%

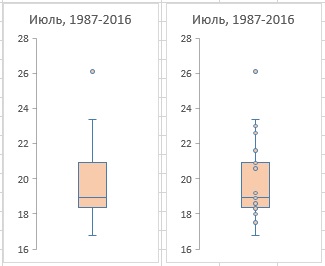


Рис. 4. Показать внутренние точки; слева параметр выключен, справа – включен

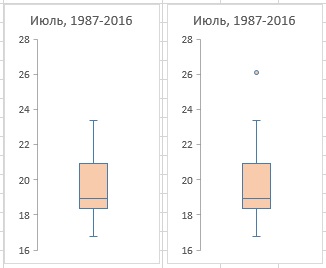


Рис. 5. Показать точки выбросов; слева параметр выключен, справа – включен

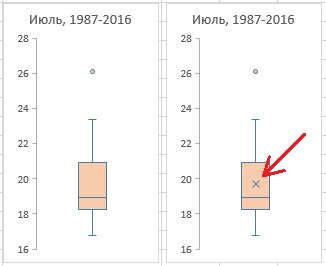


Рис. 6. Показать средние метки; слева параметр выключен, справа – включен

И, наконец, самый непонятный параметр форматирования диаграммы ящик с усами – *Расчет квартиля*. В сообществе статистиков нет единого мнения, по какому алгоритму считать квартили. В Excel используется два алгоритма. Если вы хотите в них досконально разобраться, рекомендую [КВАРТИЛЬ: какие формулы расчета использует Excel](http://baguzin.ru/wp/?p=17320). Если кратко, то в Excel есть две функции КВАРТИЛЬ.ИСКЛ() и КВАРТИЛЬ.ВКЛ(). Первая использует алгоритм эксклюзивной медианы, вторая – инклюзивной. Алгоритм расчета квартилей в первом случае дает значения чуть более далекие от медианы, т.е. ящик немного более вытянутый. Визуально это не разглядеть, и только если добавить подписи данных, то числа покажут этот эффект (рис. 7).

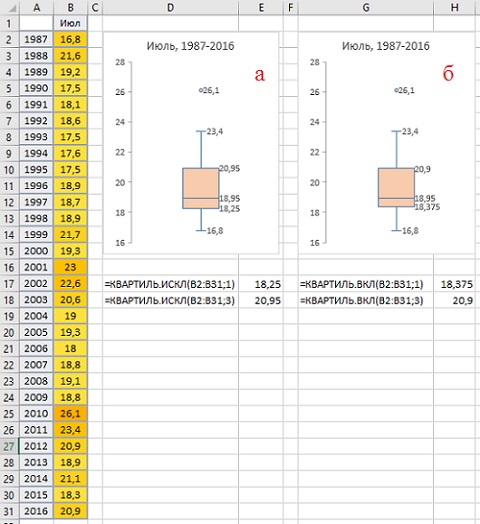


Рис. 7. Расчет квартиля: (а) эксклюзивная медиана, (б) инклюзивная медиана

Если вы обрабатываете данные в разных программах (Excel, R, SAS…), то для совместимости, наверное, лучше использовать квартили, рассчитанные по методу эксклюзивной медианы. Поскольку этот метод используется во всех продуктах, вы получите одинаковые числа в разных программах. Если же вы работаете только в Excel, я бы рекомендовал метод инклюзивной медианы. Он более чувствителен к выбросам, т.е. при прочих равных КВАРТИЛЬ.ВКЛ() определит точку как выброс при меньшем отклонении от медианы (рис. 8).

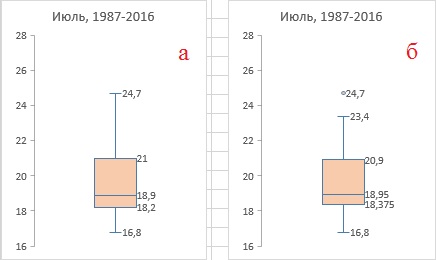


Рис. 8. (б) Точка, которая идентифицируется функцией КВАРТИЛЬ.ВКЛ(), как выброс, в то время, как (а) КВАРТИЛЬ.ИСКЛ() еще не считает эту точку выбросом