**Анализ данных. Пять базовых показателей распределения случайной величины**

[Основные характеристики выборки](http://baguzin.ru/wp/?p=5381) (среднее значение, разброс и форма распределения) позволяют описать свойства данных и перейти к более глубоким исследованиям. Довольно часто для анализа данных применяется подход, основанный на пятерке базовых показателей и построении точечных и/или блочных диаграмм.[[1]](#footnote-1)

Пятерка базовых показателей, обеспечивающих наиболее точную оценку вида распределения, состоит из следующих характеристик:

* Минимальное значение – Хmin,
* Первый квартиль – Q1,
* Медиана,
* Третий квартиль – Q3,
* Максимальное значение – Xmax.

Если данные распределены совершенно симметрично, между пятью базовыми показателями наблюдаются следующие зависимости:

* Расстояние от Хmin до медианы равно расстоянию от медианы до Xmax.
* Расстояние от Хmin до Q1 равно расстоянию от Q3 до Xmax.
* Расстояние от Q1 до медианы равно расстоянию от медианы до Q3.

Когда данные распределены несимметрично, между элементами пятерки показателей возникают следующие зависимости:

* Если распределение имеет положительную асимметрию, расстояние от Хmin до медианы меньше расстояния от медианы до Xmax.
* Если распределение имеет положительную асимметрию, расстояние от Q3 до Xmax больше, чем от Хmin до Q1.
* Если распределение имеет отрицательную асимметрию, расстояние от Хmin до медианы больше расстояния от медианы до Xmax.
* Если распределение имеет отрицательную асимметрию, расстояние от Q3 до Хmin меньше, чем от Xmax до Q1.

Пятерка базовых показателей, характеризующих распределение доходности 15 взаимных фондов с очень высоким уровнем риска представлены на рис. 1.



Рис. 1. Пятерка базовых показателей, характеризующих распределение доходности 15 взаимных фондов с очень высоким уровнем риска

Исследуем на их основе симметричность распределения. Расстояние от медианы до Xmax (18,5 – 6,5 = 12) приблизительно равно расстоянию от Хmin до медианы (6,5 – (–6,1) = 12,6). Однако расстояние от Q3 до Xmax (18,5 – 9,8 = 8,7) превышает расстояние от Хmin до Q1 (–0,7 – (–6,1) = 5,4). Следовательно, распределение пятилетней среднегодовой доходности фондов с очень высоким уровнем риска имеет слабую положительную асимметрию.

**Точечная диаграмма**

Точечная диаграмма позволяет наглядно представить саму выборку, пятерку базовых показателей и интервалы *Х̅ ± S, Х̅ ± 2S*, где *Х̅* – среднее арифметическое выборки, *S* – стандартное отклонение выборки (рис. 2).



Рис. 2. Точечная диаграмма для доходности 15 фондов

В Excel нет стандартной возможности построить точечную масштабированную диаграмму. Наверное, этим и объясняется ее более редкое использование, чем, например, блочной диаграммы. Хотя, как видно из рис. 2, информативность такой диаграммы весьма высока. Кратко опишу шаги построения диаграммы:

1. Для начала постройте обычную точечную диаграмму для диапазона А1:В16 (диапазон В1:В16 добавлен в качестве координаты Y; в нем не просто забиты единицы, а используется формула, позволяющая разместить несколько точек по высоте, если бы в наших данных по доходности встретилось несколько одинаковых чисел, см. Excel-файл).
2. Далее создайте несколько групп данных для значений среднего, первого квартиля, медианы и третьего квартиля; на рис. 2 – это D2:F16.
3. Используя прием со специальной вставкой поместите указатели этих четырех статистик на диаграмму; подробнее см. [Как добавить линию на гистограмму](http://baguzin.ru/wp/?p=1237).
4. Повторите пп. 2 и 3, чтобы отобразить на диаграмме интервалы *Х̅ ± S* и *Х̅ ± 2S.*
5. Пройдите по меню Вставка → Надпись и добавьте текстовое описание дополнительных элементов диаграммы.
6. Отформатируйте диаграмму, чтобы повысить её читаемость; подробнее см. [Принцип Эдварда Тафти минимизации количества элементов диаграммы](http://baguzin.ru/wp/?p=2071) и [Искусство графического представления данных](http://baguzin.ru/wp/?p=5367).

**Блочная диаграмма**

Блочная диаграмма (box-and-whisker diagram) представляет собой удобное средство для изображения пяти базовых показателей (рис. 3).



Рис. 3. Блочная диаграмма, иллюстрирующая показатели среднегодовой доходности 15 фондов с очень высоким уровнем риска; сравните этот рисунок и рис. 2; какое представление вам кажется более информативным?

Вертикальная линия, проведенная внутри прямоугольника, отмечает медиану. Левая сторона прямоугольника соответствует первому квартилю, Q1 а правая сторона — третьему квартилю, Q3. Таким образом, прямоугольник содержит средние 50% элементов выборки. Младшие 25% данных изображаются в виде линии (так называемый ус), соединяющей левую сторону прямоугольника с наименьшим выборочным значением Хmin. Старшим 25% данных соответствует линия, соединяющая правую сторону прямоугольника с наибольшим выборочным значением Хmax. Подробнее о том, как строить блочные диаграммы см. [Excel. Биржевая диаграмма, она же блочная, она же ящичная](http://baguzin.ru/wp/?p=4616).

Предыдущая заметка [Определение среднего значения, вариации и формы распределения. Описательные статистики](http://baguzin.ru/wp/?p=5381)

Следующая заметка

К оглавлению [Статистика для менеджеров с использованием Microsoft Excel](http://baguzin.ru/wp/?p=5285)

1. Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 213–217 [↑](#footnote-ref-1)