Анализ данных. Пять базовых показателей распределения случайной величины

Основные характеристики выборки (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]

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

- Минимальное значение – X_{min},
- Первый квартиль – Q_1,
- Медиана,
- Третий квартиль – Q_3,
- Максимальное значение – X_{max}.

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

- Расстояние от X_{min} до медианы равно расстоянию от медианы до X_{max}.
- Расстояние от X_{min} до Q_1 равно расстоянию от Q_3 до X_{max}.
- Расстояние от Q_1 до медианы равно расстоянию от медианы до Q_3.

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

- Если распределение имеет положительную асимметрию, расстояние от X_{min} до медианы меньше расстояния от медианы до X_{max}.
- Если распределение имеет положительную асимметрию, расстояние от Q_3 до X_{max} больше, чем от X_{min} до Q_1.
- Если распределение имеет отрицательную асимметрию, расстояние от X_{min} до медианы больше расстояния от медианы до X_{max}.
- Если распределение имеет отрицательную асимметрию, расстояние от Q3 до X_{min} меньше, чем от X_{max} до Q_1.

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

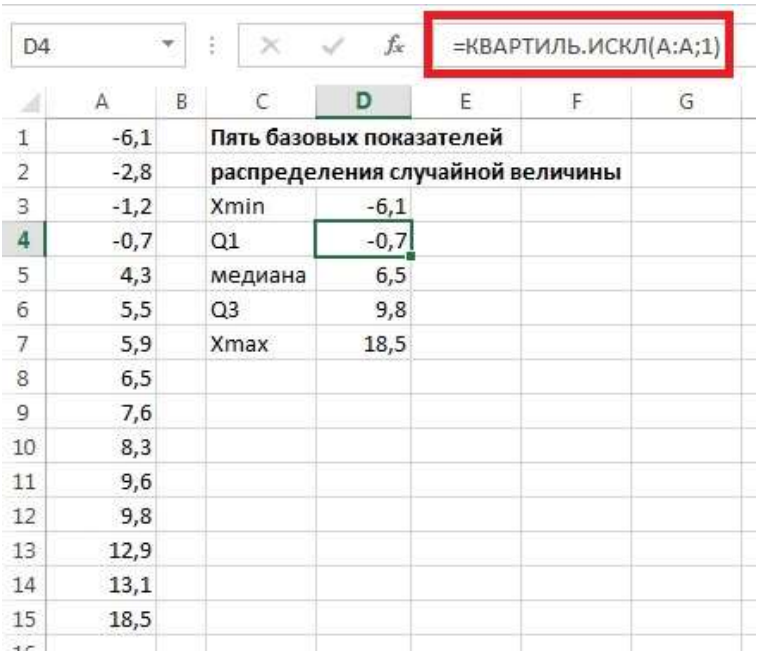

Рис. 1. Пятерка базовых показателей, характеризующих распределение доходности 15 взаимных фондов с очень высоким уровнем риска

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 213–217

Исследуем на их основе симметричность распределения. Расстояние от медианы до X_{max} (18,5 – 6,5 = 12) приблизительно равно расстоянию от X_{min} до медианы (6,5 – (–6,1) = 12,6). Однако расстояние от Q_3 до X_{max} (18,5 – 9,8 = 8,7) превышает расстояние от X_{min} до Q_1 (–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 $\overline{X} \pm S$, $\overline{X} \pm 2S$, где $\overline{X}$ – среднее арифметическое выборки, S – стандартное отклонение выборки (рис. 2).

	A	B	C	D	E	F	G
1	Доходность, %	Количество		Точечная масштабированняя диаграмма			
2	-6,1	1			Доля по горизонтали	Доля по вертикали	
3	-2,8	1		Среднее	6,08	20	
4	-1,2	1			6,08	40	
5	-0,7	1					
6	4,3	1			Доля по горизонтали	Доля по вертикали	
7	5,5	1		1-й квартиль	-0,7	20	
8	5,9	1			-0,7	40	
9	6,5	1					
10	7,6	1			Доля по горизонтали	Доля по вертикали	
11	8,3	1		Медиана	6,5	20	
12	9,6	1			6,5	40	
13	9,8	1					
14	12,9	1			Доля по горизонтали	Доля по вертикали	
15	13,1	1		3-й квартиль	9,8	20	
16	18,5	1			9,8	40	
17							

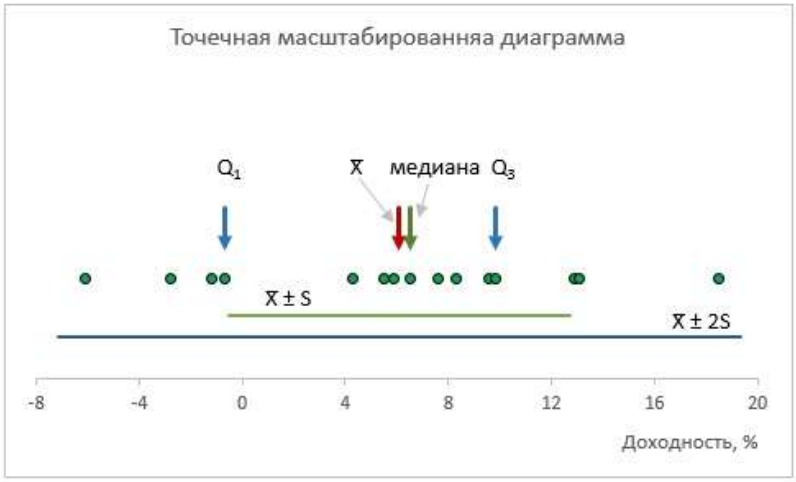

Рис. 2. Точечная диаграмма для доходности 15 фондов

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 A1:B16 (диапазон B1:B16 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.
4. Повторите пп. 2 и 3, чтобы отобразить на диаграмме интервалы $\overline{X} \pm S$ и $\overline{X} \pm 2S$.

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 и Искусство графического представления данных.

Блочная диаграмма

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

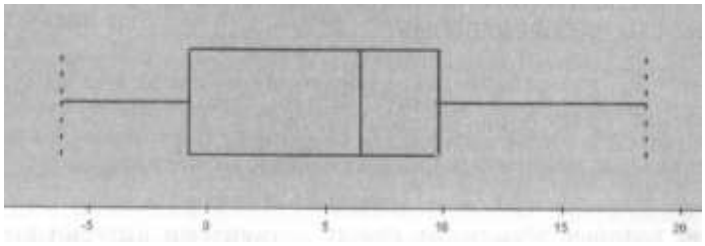

Рис. 3. Блочная диаграмма, иллюстрирующая показатели среднегодовой доходности 15 фондов с очень высоким уровнем риска; сравните этот рисунок и рис. 2; какое представление вам кажется более информативным?

Вертикальная линия, проведенная внутри прямоугольника, отмечает медиану. Левая сторона прямоугольника соответствует первому квартилю, Q_1 а правая сторона — третьему квартилю, Q_3. 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 X_{min}. Старшим 25% данных соответствует линия, соединяющая правую сторону прямоугольника с наибольшим выборочным значением X_{max}. Подробнее о том, как строить блочные диаграммы см. Excel. Биржевая диаграмма, она же блочная, она же ящичная.

Предыдущая заметка Определение среднего значения, вариации и формы распределения. Описательные статистики

Следующая заметка

К оглавлению Статистика для менеджеров с использованием Microsoft Excel