**ОП «Политология», 2019-20**
**Математика и статистика, часть 2**
**Выборки и их описание. (13 мая 2020 г.)**
*А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок*

# Базовые определения

- **Выборка** – последовательность независимых одинаково распределенных случайных величин:

$$x_1, x_2, \ldots, x_i, \ldots, x_n,$$

  где $x_i$ – $i$-тое наблюдение в выборке ($i$-тый элемент), а $n$ – число наблюдений в выборке.

- **Вариационный ряд** – упорядоченная выборка (обычно упорядоченная по возрастанию, от меньшего значения к большему):

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(i)} \le \cdots \le x_{(n)},$$

  где $x_{(1)}$ – наименьшее значение в выборке, а $x_{(n)}$ – наибольшее значение в выборке.

# Выборочные квантили

## 1. Медиана

**Медиана выборки** – это оценка квантиля распределения уровня 0.5, то есть значение, которое 50% значений в выборке не превышают. Другими словами, **медиана** – это центральное значение в вариационном ряду; значение, которое делит упорядоченную выборку на две половины – нижнюю и верхнюю.

Найти значение, которое находится ровно в середине последовательности чисел, просто, но есть проблема: не всегда в центре ряда может оказаться одно число. Возможны два случая: а) число наблюдений в выборке нечётно; б) число наблюдений в выборке чётно.

### Число наблюдений в выборке нечётно

Если в выборке нечётное число наблюдений, медиана – это просто значение, которое находится ровно посередине вариационного ряда.

**Пример 1.** Дана выборка из 7 наблюдений:

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | 10 | 70 | 60 | 80 | 5 | 100 |

Запишем вариационный ряд (упорядочим выборку по возрастанию):

$$5 \qquad 10 \qquad 20 \qquad 60 \qquad 70 \qquad 80 \qquad 100$$

Чтобы найти значение, которое находится посередине, отсчитаем справа и слева одинаковое число наблюдений (в данном случае 3):

$$5 \qquad 10 \qquad 20 \qquad \boxed{60} \qquad 70 \qquad 80 \qquad 100$$

Значение, до которого мы таким образом дошли, 60. Оно и является медианой выборки. Можем записать $\mathrm{med}(x_1 \ldots x_7) = 60$.

Выше было сказано, что медиана делит выборку на две половины. Но нечётное число наблюдений на два не делится. Как быть? Как делить выборку на половины и куда включать медиану? Всё просто: медиану нужно включать в **обе** половины выборки. В нашем примере нижняя половина выборки содержит числа 5, 10, 20, 60, а верхняя половина – 60, 70, 80, 100. В обеих частях одинаковое число наблюдений, значит, они точно являются половинами, мы ничего не перепутали.

### Число наблюдений в выборке чётно

Если число наблюдений в выборке чётно, то для определения медианы понадобится рассчитывать среднее арифметическое двух центральных чисел в вариационном ряду.

**Пример 2.** Дана выборка из 8 наблюдений:

$$20 \qquad 10 \qquad 70 \qquad 60 \qquad 80 \qquad 5 \qquad 100 \qquad 55$$

Запишем вариационный ряд:

$$5 \qquad 10 \qquad 20 \qquad 55 \qquad 60 \qquad 70 \qquad 80 \qquad 100$$

Если мы отсчитаем одинаковое число наблюдений справа и слева (по 3), то дойдем до двух центральных значений в вариационном ряду – 55 и 60:

$$5 \qquad 10 \qquad 20 \qquad \boxed{55 \qquad 60} \qquad 70 \qquad 80 \qquad 100$$

Медианой в таком случае будет среднее арифметическое этих двух чисел. Можем записать:

$$\mathrm{med}(x_1 \ldots x_8) = \frac{55 + 60}{2} = 57.5.$$

Медиану нашли, а как теперь поделить выборку на две половины и куда включить медиану? Всё просто: раз наблюдений в выборке чётное количество, то можем спокойно поделить вариационный ряд на две половины, по $n/2$ наблюдений в каждой. В нашем случае в нижнюю половину выборки входят значения 5, 10, 20, 55, а в верхнюю половину – значения 60, 70, 80, 100. Медиана при этом не входит **ни в одну** половину – она же не принадлежит вариационному ряду (в нем нет значения 57.5), так зачем её тогда куда-то включать?

## 2. Квартили

**Квартили** – значения, которые делят упорядоченную выборку на четыре примерно равные части. В первую часть входят первые 25% наблюдений, во вторую часть входят следующие 25% наблюдений и так далее. Таким образом, первый квартиль отделяет первые 25% значений в вариационном ряду, второй квартиль – первые 50% значений в вариационном ряду, третий квартиль – первые 75% значений, и наконец, четвертый квартиль отделяет 100% значений, то есть все наблюдения в выборке.

Нетрудно заметить, что медиана – это второй квартиль, то есть значение, которое отделяет первую половину значений (0 – 50%) в упорядоченной выборке от второй половины значений (50 – 100%).

**Квартили** – это оценки квантилей распределения уровней 0.25, 0.5, 0.75 и 1 ($x_{0.25}$, $x_{0.5}$, $x_{0.75}$, $x_1$). Для описания выборок нам будут нужны квантили уровней 0.25 и 0.75, первый и третий квартиль или нижний и верхний квартиль. Обозначать их будем следующим образом:

$$\mathrm{Q}_1 = x_{0.25}, \text{ нижний квартиль}$$

$$\mathrm{Q}_3 = x_{0.75}, \text{ верхний квартиль}$$

Как находить нижний и верхний квартили? Просто: нижний квартиль – это медиана нижней половины выборки, а верхний квартиль – это медиана верхней половины выборки. А как находить медиану мы уже разобрали. Рассмотрим следующий пример.

Дана выборка из 9 наблюдений:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 25 | 15 | 7 | 6 | 75 | 15 | 10 | 12 | 18 |

Запишем вариационный ряд:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | 7 | 10 | 12 | 15 | 15 | 18 | 25 | 75 |

Медиана выборки – значение 15. Тогда нижняя половина выборки выглядит следующим образом:

| | | | | |
|---|---|---|---|---|
| 6 | 7 | 10 | 12 | 15 |

Находим медиану нижней половины выборки. Это число 10. Следовательно, $\mathrm{Q}_1 = 10$. Верхняя половина выборки выглядит следующим образом:

| | | | | |
|---|---|---|---|---|
| 15 | 15 | 18 | 25 | 75 |

Находим медиану верхней половины выборки. Это число 18. $\mathrm{Q}_3 = 18$.

С описанием выборок связано ещё одно понятие – **межквартильный размах**. Будем обозначать его IRQ, а определяется он следующим образом:

$$\mathrm{IRQ} = \mathrm{Q}_3 - \mathrm{Q}_1$$

Так, в нашем примере, разобранном выше, $\mathrm{IRQ} = 18 - 10 = 8$. Содержательно межквартильный размах – это одна из мер разброса значений в выборке. Но межквартильный размах очень важен и в «техническом» отношении – именно он используется для поиска нетипичных значений в выборке.

# Поиск нетипичных наблюдений

**Нетипичные наблюдения** в выборке – наблюдения, которые сильно удалены от медианного значения. Иногда нетипичные наблюдения в выборке имеют «естественное» происхождение (существуют объекты, которые сильно отличаются от остальных), а иногда такие наблюдения – просто следствия ошибок (опечатки в данных, неверные единицы измерения и прочее). Нетипичные наблюдения также называют *нехарактерными наблюдениями* или *выбросами (outliers)*.

Вопрос: как определить нетипичные наблюдения в выборке? Ответ: найти границы типичных значений, и все значения, которые выходят за эти границы, считать нетипичными. Границы типичных значений:

$$[\mathrm{Q}_1 - 1.5 \times \mathrm{IRQ}; \mathrm{Q}_3 + 1.5 \times \mathrm{IRQ}]$$

Проверим, есть ли в выборке из нашего примера нетипичные наблюдения. Мы определили, что $\mathrm{Q}_1 = 10$, $\mathrm{Q}_3 = 18$, $\mathrm{IRQ} = 8$. Подставим все значения в формулы:

$$[10 - 1.5 \times 8; 18 + 1.5 \times 8]$$

$$[-2; 30]$$

Видно, что одно наблюдение в этот интервал не входит – это значение 75. Следовательно, в нашей выборке есть одно нетипичное наблюдение – 75.

# Ящик с усами

Для визуализации описательных статистик иногда строят график, который называется *ящик с усами* (*box plot* или *box-and-whiskers plot* на английском). Построение графика, а точнее, его «усов», зависит от того, есть ли в выборке нетипичные наблюдения. Рассмотрим все возможные случаи.

## В выборке нет нетипичных наблюдений

1. Отмечаем горизонтальными линиями нижний квартиль $\mathrm{Q}_1$ и верхний квартиль $\mathrm{Q}_3$, это будут нижняя и верхняя границы «ящика».
2. Достраиваем фигуру до прямоугольника, ширина «ящика» значения не имеет.
3. Внутри «ящика» горизонтальной линией отмечаем медиану. Медиана необязательно должна лежать ровно посередине «ящика», зависит от распределения.
4. Отмечаем минимальное и максимальное значение в выборке, это будут границы «усов» графика. «Дотягиваем» вертикальные «усы» до минимального и максимального значения.

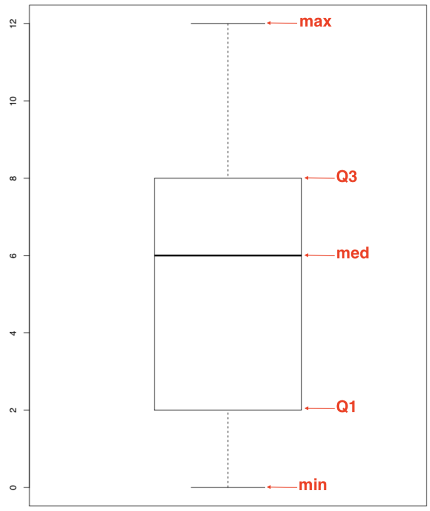


Рис. 1: Нет нетипичных наблюдений

## В выборке есть нетипично маленькие и нетипично большие наблюдения

Повторяем шаги 1-3 из построения графика для выборки без нетипичных наблюдений. Вычисляем границы типичных значений $Q_1 - 1.5 \times IRQ$ и $Q_3 + 1.5 \times IRQ$. Границы «усов» графика – минимальное и максимальное значение в выборке, которые попадают в границы типичных значений. «Дотягиваем» вертикальные «усы» до этих значений. Отмечаем точками все нетипичные значения.

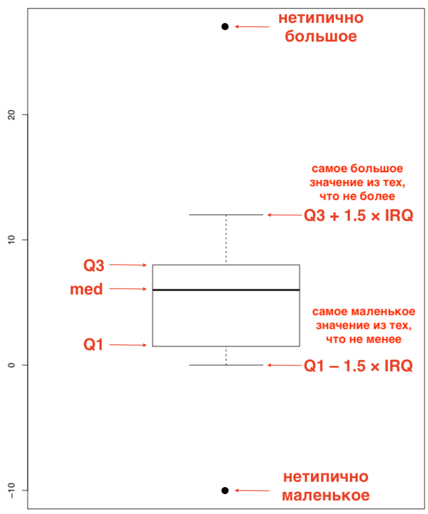


Рис. 2: Есть нетипичные наблюдения

## В выборке есть нетипично маленькие наблюдения

Повторяем шаги 1-3 из построения графика для выборки без нетипичных наблюдений. Вычисляем границы типичных значений $Q_1 - 1.5 \times \text{IRQ}$ и $Q_3 + 1.5 \times \text{IRQ}$. Граница «нижнего» уса графика – минимальное значение в выборке, которое попадает в границы типичных значений. Граница верхнего «уса» – максимальное значение в выборке. «Дотягиваем» вертикальные «усы» до этих значений. Отмечаем точками все нетипичные значения.

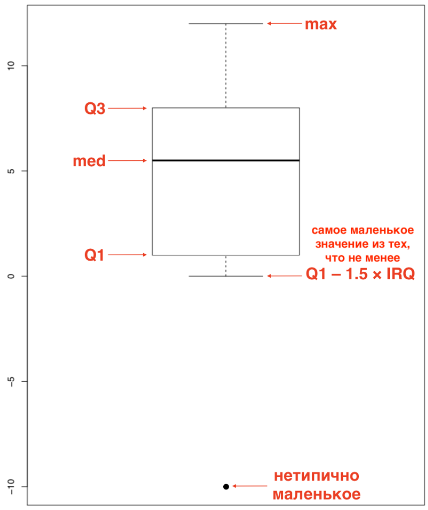


Рис. 3: Нетипично маленькие наблюдения

## В выборке есть нетипично большие наблюдения

Повторяем шаги 1-3 из построения графика для выборки без нетипичных наблюдений. Вычисляем границы типичных значений $Q_1 - 1.5 \times \text{IRQ}$ и $Q_3 + 1.5 \times \text{IRQ}$. Граница «нижнего» уса графика – минимальное значение в выборке. Граница верхнего «уса» – максимальное значение в выборке, которое попадает в границы типичных значений. Дотягиваем» вертикальные «усы» до этих значений. Отмечаем точками все нетипичные значения.

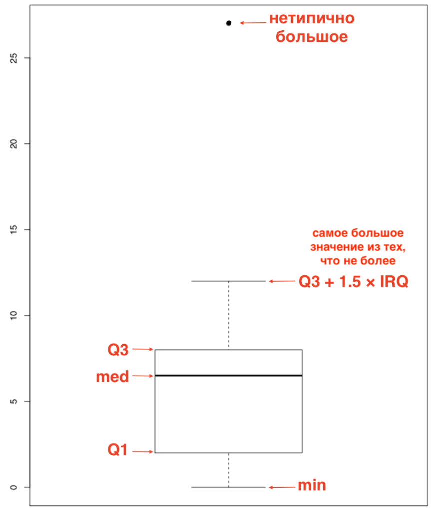


Рис. 4: Нетипично большие наблюдения