ОП «Политология», 2020-21
Введение в ТВиМС
Выборки и их описание. (Краткий конспект)
А. А. Макаров, А. А. Тамбовцева

Базовые определения

Определение 1. Выборка – последовательность независимых одинаково распределённых 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.

Определение 2. Вариационный ряд – упорядоченная выборка (обычно упорядоченная по возрастанию, от меньшего значения к большему):

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

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

Медиана выборки и выборочные квартили

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:

- число наблюдений в выборке нечётно;
- 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

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 по 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 $\text{med}(x_1 \ldots x_7) = 60$.

Выше было сказано, что медиана делит выборку на две половины. Но нечётное число наблюдений не делится на 2 без остатка. Как быть? Всё просто: медиану нужно включить в обе половины выборки. В нашем примере нижняя половина выборки содержит значения 5, 10, 20, 60, а верхняя половина – значения 60, 70, 80, 100.

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:

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

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?

Квартили

Квартили – значения, которые делят упорядоченную выборку на четыре равные части (Рис. 1).

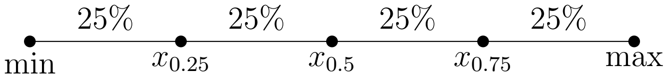

Рис. 1: Квартили

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, четвёртый 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 (обозначим их $x_{0.25}$, $x_{0.5}$, $x_{0.75}$, x_1). Для описания выборок нам будут нужны $x_{0.25}$ и $x_{0.75}$, нижний и верхний квартиль или первый и третий 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 знаем. Рассмотрим следующий пример.

Пример 3. Дана выборка из 9 наблюдений:

$$25 \quad 15 \quad 7 \quad 6 \quad 75 \quad 15 \quad 10 \quad 12 \quad 18$$

Запишем вариационный ряд:

$$6 \quad 7 \quad 10 \quad 12 \quad \boxed{15} \quad 15 \quad 18 \quad 25 \quad 75$$

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

$$6 \quad 7 \quad 10 \quad 12 \quad 15$$

Находим медиану нижней половины выборки. Это число 10. Поэтому $x_{0.25} = 10$. Верхняя половина выборки выглядит следующим образом:

$$15 \quad 15 \quad 18 \quad 25 \quad 75$$

Находим медиану верхней половины выборки. Это число 18. Поэтому $x_{0.75} = 18$.

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

$$\Delta = x_{0.75} - x_{0.25}.$$

Так, в примере 3 межквартильный размах $\Delta = 18 - 10 = 8$. 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 последствия ошибок, например, опечаток в данных или выбор неверных единиц измерения. Нетипичные наблюдения также называют нехарактерными наблюдениями или выбросами (от англ. *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 вычисляются так:

$$[x_{0.25} - 1.5 \times \Delta;\ x_{0.75} + 1.5 \times \Delta].$$

Проверим, есть ли в выборке из примера 3 нетипичные наблюдения. Мы определили, что $x_{0.25} = 10$, $x_{0.75} = 18$, $\Delta = 8$. 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* (от англ. *box plot* или *box-and-whiskers plot*). 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 $x_{0.25}$ и верхний квартиль $x_{0.75}$, 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.

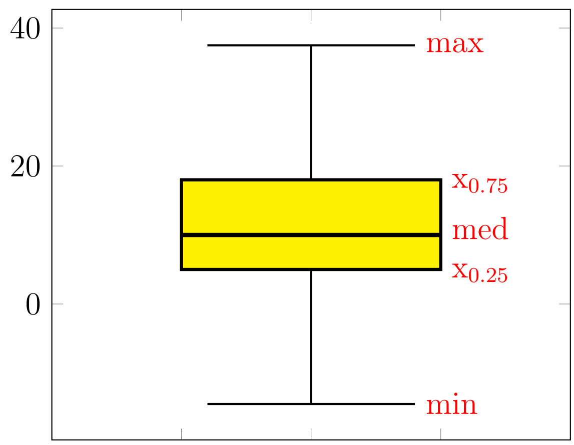

Рис. 2: Нет нетипичных наблюдений

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 $x_{0.25} - 1.5 \times \Delta$ и $x_{0.75} + 1.5 \times \Delta$. Границы «усов» графика – минимальное и максимальное значение в выборке, которые попадают в границы типичных значений (на рисунке x^*_{min} и x^*_{max} соответственно). «Дотягиваем» вертикальные «усы» до этих значений. Отмечаем точками все нетипичные значения.

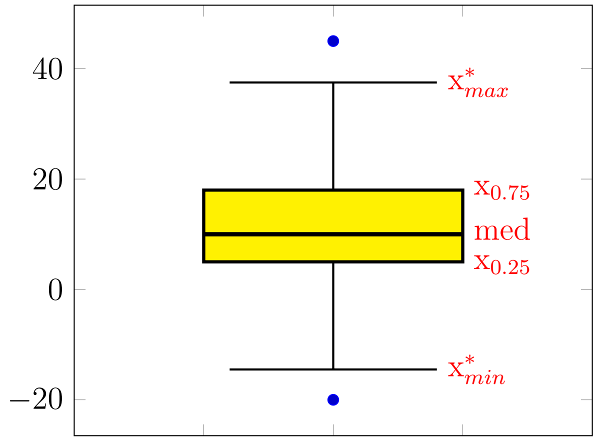

Рис. 3: Нетипичные наблюдения с двух сторон

В выборке есть нетипично маленькие наблюдения

Повторяем шаги 1-3 из построения графика для выборки без нетипичных наблюдений. Вычисляем границы типичных значений $x_{0.25} - 1.5 \times \Delta$ и $x_{0.75} + 1.5 \times \Delta$. 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.

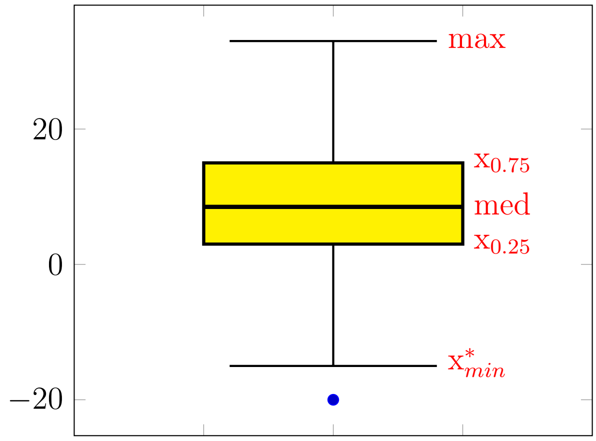

Рис. 4: Нетипично маленькие наблюдения

В выборке есть нетипично большие наблюдения

Повторяем шаги 1-3 из построения графика для выборки без нетипичных наблюдений. Вычисляем границы типичных значений $x_{0.25} - 1.5 \times \Delta$ и $x_{0.75} + 1.5 \times \Delta$. Граница нижнего «уса» графика – минимальное значение в выборке. Граница верхнего «уса»

– максимальное значение в выборке, которое попадает в границы типичных значений (на рисунке x^*_{max}). «Дотягиваем» вертикальные «усы» до этих значений. Отмечаем точками все нетипичные значения.

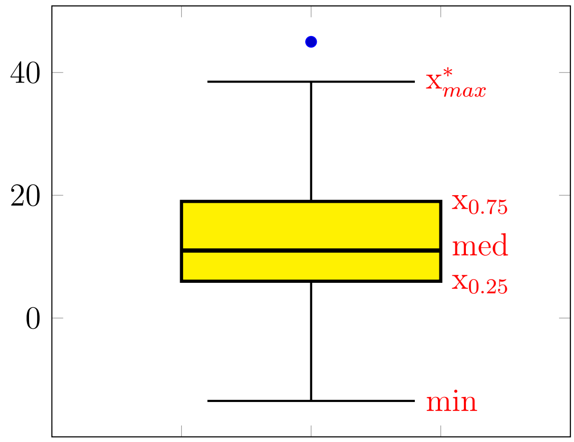

Рис. 5: Нетипично большие наблюдения

Ранги наблюдений

Ранг – порядковый номер наблюдения в вариационном ряду, то есть в выборке, упорядоченной по возрастанию. Будем обозначать ранг буквой R, R_i – ранг i-того наблюдения в выборке.

Возможны два случая:

- выборка не содержит повторяющихся значений;
- выборка содержит повторяющиеся значения.

В выборке нет повторяющихся значений

Если в выборке нет повторящихся значений, ранг наблюдения – просто его порядковый номер в выборке, упорядоченной по возрастанию.

Пример 4. Дана выборка из 7 наблюдений:

$$6 \quad 1 \quad 2 \quad 7 \quad 8 \quad 3 \quad 100$$

Запишем вариационный ряд:

$$1 \quad 2 \quad 3 \quad 6 \quad 7 \quad 8 \quad 100$$

Подпишем номера наблюдений:

1	2	3	6	7	8	100
1	2	3	4	5	6	7

Запишем ранги: $R_1 = 4$, $R_2 = 1$, $R_3 = 2$, $R_4 = 5$, $R_5 = 6$, $R_6 = 3$, $R_7 = 7$.

Внимание: ранги определяются для наблюдений в исходной выборке. Например, здесь R_1 – это ранг первого наблюдения в выборке, то есть порядковый номер «шестёрки» в вариационном ряду, равный 4.

6	1	2	7	8	3	100
1	2	3	6	7	8	100
1	2	3	4	5	6	7

Аналогично для остальных наблюдений.

В выборке есть повторяющиеся значения

Если в выборке есть повторяющиеся значения, то возникает необходимость считать средний ранг.

Пример 5. Дана выборка из 7 наблюдений:

$$6 \qquad 1 \qquad 2 \qquad 7 \qquad 8 \qquad 2 \qquad 100$$

Запишем вариационный ряд:

$$1 \qquad 2 \qquad 2 \qquad 6 \qquad 7 \qquad 8 \qquad 100$$

Для неповторяющихся значений ранги определяются обычным образом (точно так же, как в примере 1):

1	2	2	6	7	8	100
1			4	5	6	7

Для повторяющихся значений считается средний ранг. В данном случае у повторяющихся «двоек» порядковые номера в вариационном ряду (ранги) – это 2 и 3. Посчитаем средний ранг – среднее арифметическое этих чисел:

$$\frac{2+3}{2} = 2.5$$

Следовательно:

1	2	2	6	7	8	100
1	2.5	2.5	4	5	6	7

Запишем ранги: $R_1 = 4$, $R_2 = 1$, $R_3 = 2.5$, $R_4 = 5$, $R_5 = 6$, $R_6 = 2.5$, $R_7 = 7$.

Важно: дробные ранги – это нормально.

Пример 6. Дана выборка из 7 наблюдений:

$$6 \qquad 1 \qquad 7 \qquad 7 \qquad 8 \qquad 7 \qquad 100$$

Запишем вариационный ряд:

$$1 \qquad 6 \qquad 7 \qquad 7 \qquad 7 \qquad 8 \qquad 100$$

Сначала определим ранги неповторяющихся значений:

$$\begin{array}{ccccccc} 1 & 6 & 7 & 7 & 7 & 8 & 100 \\ 1 & 2 & & & & 6 & 7 \end{array}$$

Порядковые номера повторяющихся «семерок» – это 3, 4, 5. Посчитаем средний ранг:

$$\frac{3+4+5}{3} = 4$$

Получаем:

$$\begin{array}{ccccccc} 1 & 6 & 7 & 7 & 7 & 8 & 100 \\ 1 & 2 & 4 & 4 & 4 & 6 & 7 \end{array}$$

Запишем ранги: $R_1 = 2$, $R_2 = 1$, $R_3 = 4$, $R_4 = 4$, $R_5 = 6$, $R_6 = 4$, $R_7 = 7$.

Важно: то, что некоторых «промежуточных» чисел среди рангов нет (например, есть ранги, равные 2 и 4, но нет ранга, равного 3) – это тоже нормально.