ФСН, 2018-19
Математика и статистика
Выборки и их описание. (29 апреля 2019 г.)
А. А. Тамбовцева

Базовые определения

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

$$x_1, x_2, \dots, x_i, \dots, x_n,$$

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):

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(n)},$$

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

Ранги

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

Возможны два случая: 1) выборка не содержит повторяющихся значений; 2) выборка содержит повторяющиеся значения.

1. В выборке нет повторяющихся значений

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

Пример 1. Дана выборка из 7 наблюдений:

6	1	2	7	8	3	100

Запишем вариационный ряд:

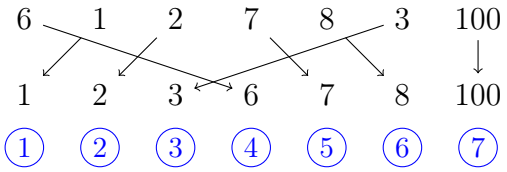
1	2	3	6	7	8	100

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, R_1 – это ранг первого наблюдения в выборке, то есть порядковый номер «шестерки» в вариационном ряду, равный 4.



Аналогично для остальных наблюдений.

2. В выборке есть повторяющиеся значения

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

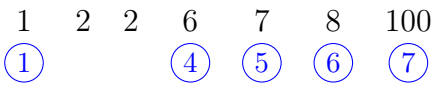
Пример 2. Дана выборка из 7 наблюдений:

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

Запишем вариационный ряд:

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

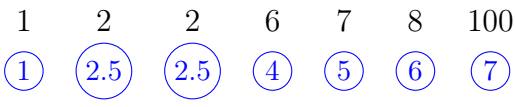
Для неповторяющихся значений ранги определяются обычным образом (точно так же, как в примере 1):



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:



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

Важно: дробные ранги – это нормально. Рассмотрим еще пример.

Пример 3. Дана выборка из 7 наблюдений:

6 1 7 7 8 7 100

Запишем вариационный ряд:

1 6 7 7 7 8 100

Сначала определим ранги неповторяющихся значений:

1	6	7	7	7	8	100
①	②				⑥	⑦

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

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

Получаем:

1	6	7	7	7	8	100
①	②	④	④	④	⑥	⑦

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

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

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 4. Дана выборка из 7 наблюдений:

20 10 70 60 80 5 100

Запишем вариационный ряд (упорядочим выборку по возрастанию):

5 10 20 60 70 80 100

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

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

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

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 5. Дана выборка из 8 наблюдений:

20 10 70 60 80 5 100 55

Запишем вариационный ряд:

5 10 20 55 60 70 80 100

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

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

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

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

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.