Выборки и их описание.

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.

1 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.

1.1 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:

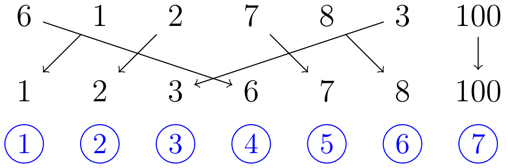
1	2	3	6	7	8	100

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.

1.2 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):

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

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

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

Пример 3. Дана выборка из 7 наблюдений:

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

Запишем вариационный ряд:

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

Сначала определим ранги неповторяющихся значений:

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.

2 Выборочные квантили

Выборочная квантиль определяется аналогично обычным (теоретическим) квантилям, с которыми мы работали в теории вероятностей, только вместо функции распределения рассматривается эмпирическая функция распределения (F_n):

$$F_n(x_p) = p,$$

где p – заданная вероятность (уровень).

3 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: 1) число наблюдений в выборке нечетно; 2) число наблюдений в выборке четно.

3.1 Число наблюдений в выборке нечетно

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

Пример 1. Дана выборка из 7 наблюдений:

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

Запишем вариационный ряд:

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

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 $med(x_1 \ldots x_7) = 60$.

Если формализовать метод нахождения медианы, описанный выше, получим следующее. У нас есть выборка из нечетного числа наблюдений n. Нечетное число всегда можно представить в таком виде: $n = 2k + 1$ (четное плюс 1). Тогда медиану можно определить так:

$$med(x_1 \ldots x_n) = x_{k+1},$$

то есть наблюдение с порядковым номером в вариационном ряду, равным $k + 1$ (здесь $k + 1$ – это именно обычный порядковый номер, не ранг, так как в общем случае значения в выборке могут повторяться, ранги могут быть дробными, и этот способ будет работать несколько иначе).
В нашем примере это выглядит так.

$$n = 7$$

$$7 = 2k + 1$$

$$k = 3$$

Тогда медиана – это $x_{3+1} = x_4$, то есть четвертое наблюдение в вариационном ряду (60).

Выше было сказано, что медиана делит выборку на две половины. Но нечетное число наблюдений на 2 не делится. Как быть? Как делить выборку на половины и куда включать медиану? Все 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.

3.2 Число наблюдений в выборке четно

Если число наблюдений в выборке четно, 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:

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

Если опять описать метод нахождения медианы формально, то получится следующее. У нас есть выборка из четного числа наблюдений, то есть, $n = 2k$ (формула четного числа). Тогда медиану можно определить так:

$$\frac{x_k + x_{k+1}}{2},$$

то есть как среднее арифметическое значений с порядковыми номерами в вариационном ряду k и $k + 1$ (здесь k и $k + 1$ – это именно обычные порядковые номера, не ранги, так как в общем случае значения в выборке могут повторяться, ранги могут быть дробными, и этот способ будет работать несколько иначе).
В нашем примере в выборке 8 наблюдений, $n = 8$.

$$8 = 2k$$

$$k = 4$$

Тогда медиана равна $\frac{x_4 + x_{4+1}}{2}$, то есть среднее арифметическое четвертого и пятого наблюдений в вариационном ряду: $(55 + 60)/2 = 57.5$.

Медиану нашли, а как теперь поделить выборку на две половины и куда включить медиану? Все просто: раз наблюдений в выборке четное 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 ее тогда куда-то включать?

4 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:

$$Q_1 = x_{0.25}, \text{ нижний квартиль}$$

$$Q_3 = x_{0.75}, \text{ верхний квартиль}$$

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, $Q_1 = 10$. Верхняя половина выборки выглядит следующим образом:

15	15	18	25	75

Находим медиану верхней половины выборки. Это число 18. $Q_3 = 18$.

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

$$\Delta = Q_3 - Q_1$$

Так, в нашем примере, разобранном выше, $\Delta = 18 - 10 = 8$ 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.

5 Поиск нетипичных наблюдений

Нетипичные наблюдения в выборке – наблюдения, которые сильно удалены от среднего (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 проч.).

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:

$$[Q_1 - 1.5\Delta; Q_3 + 1.5\Delta]$$

Проверим, есть ли в выборке из нашего примера нетипичные наблюдения. Мы определили, что $Q_1 = 10$, $Q_3 = 18$, $\Delta = 8$. Подставим все значения в формулы:

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

$$[-2; 30]$$

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