ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЮРИДИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ.

ЛЕКЦИЯ 5: «ИНТЕРВАЛЬНЫЙ РЯД. ВЫЧИСЛЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ ПО ИНТЕРВАЛЬНОМУ РЯДУ»

Замечание

При обработке большого числа экспериментальных данных их предварительно группируют и оформляют в виде так называемого **интервального ряда**.

Пример

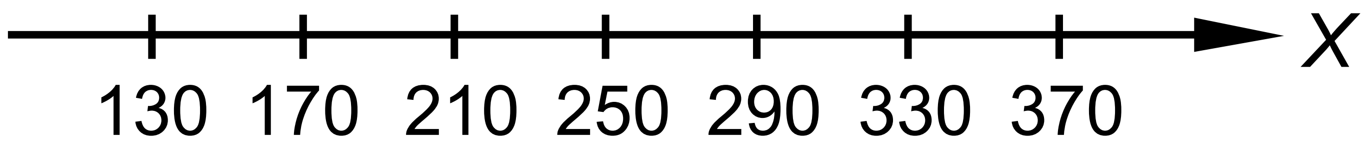
Пусть средняя месячная зарплата за го в некоторых условных единицах каждого из пятидесяти случайно отобранных работников хозяйства такова: 317, 304, 230, 285, 290, 320, 262, 274, 205, 180, 234, 221, 241, 270, 257, 290, 258, 296, 301, 150, 160, 210, 235, 308, 240, 370, 180, 244, 365, 130, 170, 250, 370, 267, 288, 231, 253, 315, 201, 256, 279, 285, 226, 367, 247, 252, 320, 160, 215, 350. Опишите заданную переменную величину с помощью интервального ряда.

Решение (1)

Здесь переменной величиной X является средняя месячная зарплата. Как видно из приведенных данных, наименьшее значение величины X равно 130, а наибольшее — 370. Таким образом, диапазон наблюдений представляет собой интервал $130-370$, длина которого равна $370-130=240$.

Решение (2)

Разобьем диапазон наблюдений на части (**разряды**) так, чтобы каждый разряд содержал несколько экспериментальных данных. Например, разделим интервал $130-370$ на 6 равных частей, тогда длина каждого разряда будет равна 40. Границами разрядов будут числа: 130, 170, 210, 250, 290, 330, 370.



Подсчитаем число значений, попавших в каждый разряд. Например, в первый разряд попадают следующие числа: 150 (1 раз), 160 (2 раза), 130 (1 раз), 170 (1 раз).

Решение (3)

Основная идея интервального ряда: будем считать, что все числа, попавшие в один разряд, приблизительно равны друг другу, и на этом основании заменим их одним числом — серединой этого разряда.

Середина первого разряда — это число 150. Мы будем считать, что оно заменяет собой (приближенно!) число 150 — 1 раз, число 160 — 2 раза, число 130 — 1 раз, но число 170 — только 0,5 раза, поскольку оно находится на границе между первым и вторым разрядами. Во второй разряд мы включим его также с крастностью $1/2$. Сложив кратности, мы получим **абсолютную частоту** первого разряда:

$$m_1 = 1 + 2 + 1 + 0{,}5 = 4{,}5 .$$

Решение (4)

Итак, число m_1 имеет двоякий смысл: с одной стороны, это число попаданий случайной величины X в первый разряд, а с другой — его рассматривают как кратность середины разряда — числа 150. Именно последний факт дает нам основание назвать число m_1 абсолютной частотой.

Разделив абсолютную частоту на число n всех наблюдений, найдем **относительную частоту** $\tilde{p}_1$ первого разряда:

$$\tilde{p}_1 = \frac{m_1}{n} = \frac{4,5}{50} = 0,09\,.$$

Решение (5)

Получив аналогичные вычисления для всех разрядов, мы получим следующую таблицу:

	130–170	170–210	210–250	250–290	290–330	330–370
m_i	4,5	5	12	14,5	9	5
$\tilde{p}_i$	0,09	0,10	0,24	0,29	0,18	0,10

Определение

Эта таблица и называется **интервальным рядом**.

В ней m_i — абсолютные частоты, а $\tilde{p}_i$ — относительные частоты. Сумма всех абсолютных частот равна числу всех приведенных в этой таблице значений переменной величины: $4{,}5+5+12+14{,}5+9+5=50$. Это свойство используется для проверки правильности вычислений. Из него следует, что сумма всех относительных частот равна единице: $0{,}09+0{,}10+0{,}24+0{,}29+0{,}18+0{,}10=1$.

Гистограмма (1)

Интервальный ряд изображают графически в виде **гистограммы**, которая строится так. Сначала вычисляют **плотности частот** h_1, h_2, h_3,... , разделив относительную частоту каждого разряда на его длину:

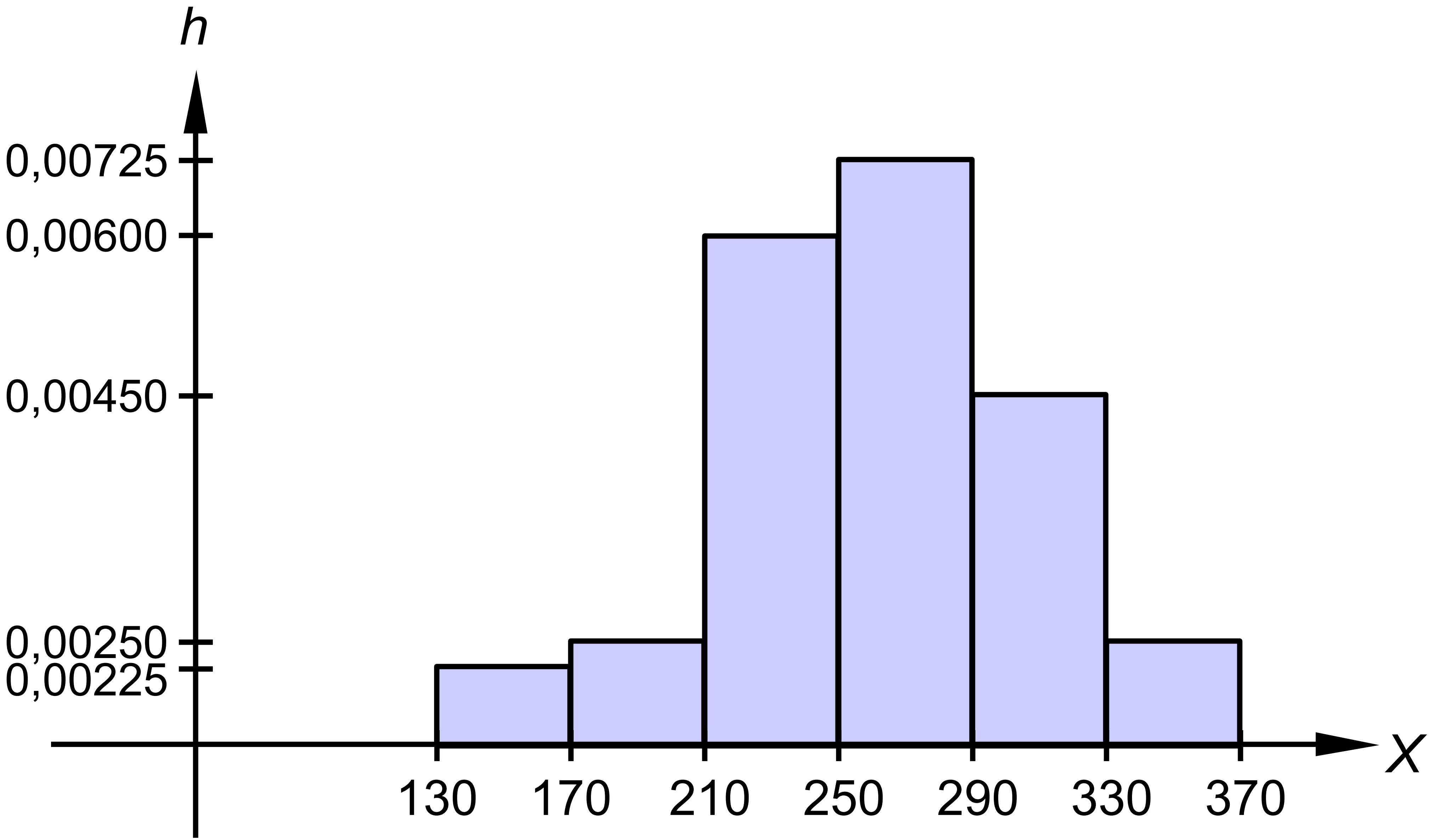
$$h_1 = \frac{0{,}09}{40} = 0{,}00225,\ h_2 = \frac{0{,}10}{40} = 0{,}00250,\ h_3 = \frac{0{,}24}{40} = 0{,}00600,$$

$$h_4 = \frac{0{,}29}{40} = 0{,}00725,\ h_5 = \frac{0{,}18}{40} = 0{,}00450,\ h_6 = \frac{0{,}10}{40} = 0{,}00250.$$

Гистограмма (2)

Затем выбирают на плоскости систему координат и откладывают на оси X значения 130, 170, 210, …, соответствующие границам разрядов. На каждом из отрезков длины 40, как на основании, строят прямоугольник, высота которого равна плотности частоты соответствующего разряда.

Гистограмма (3)



Гистограмма (4)

Заметьте, что высоты h_1, h_2, …, h_6 прямоугольников, образующих гистограмму, выбраны так, что их площади будут $\tilde{p}_1$, $\tilde{p}_2$, …, $\tilde{p}_6$, т.е. Равны соответствующим относительным частотам. Например, площадь первого прямоугольника равна $h_1 \cdot 40 = \frac{0{,}09}{40} \cdot 40 = 0{,}09$, т.е. равна относительной частоте $\tilde{p}_1$.

Гистограмма (5)

Определим, например, долю значений случайной величины X, принадлежащих интервалу $210-300$. Для этого вычислим площадь фигур с основанием $210-300$. Площади первых двух прямоугольников, составляющих фигуру, равны соответственно $\tilde{p}_3 = 0{,}24$ и $\tilde{p}_4 = 0{,}29$; площадь третьего равна $10 \cdot 0{,}0045 = 0{,}045$. Сумма площадей $0{,}24 + 0{,}29 + 0{,}045 = 0{,}575$ и дает нужное число. Иными словами, 57,5% значений величины X находится в границах от 210 до 300.

Замечание

Замечание о выборе масштаба: гистограмма, должна быть наглядной и удобной для проведения по ней расчетов. Поэтому нужно по возможности оптимально использовать площадь чертежа. Чуть позже мы обсудим, как можно автоматизировать процесс построения гистограмм с помощью компьютера.

Задача (1)

В 2010 году по 40 районам области зафиксированы следующие данные о числе преступлений с применением оружия: 19, 32, 26, 3, 8, 26, 2, 14, 6, 18, 3, 5, 10, 35, 4, 7, 4, 12, 23, 8, 9, 2, 14, 6, 2, 6, 7, 13, 6, 9, 17, 4, 3, 2, 6, 9, 27, 1, 7, 2. Составьте по этим данным интервальный ряд, постройте гистограмму и выполните следующее дополнительное задание.

.

Задача (2)

1) В каких границах принимает значения число преступлений в одном районе?

2) Какой из полученных вами разрядов имеет наибольшую частоту и как это можно истолковать?

3) Как вычисляют частоту попадания значений изучаемой величины в тот или иной интервал?

4) Рассчитайте долю тех районов, для которых число преступлений с применением оружия заключено в пределах от 8 до 18.

Решение (1)

Изучаемой величиной X является число преступлений с применением оружия, совершенных в одном районе. Статистические данные приведены по сорока районам. Самое маленькое из данных чисел 1, а самое большое 35. Таким образом, диапазон значений величины X будет интервал $[1;\,35]$. Его длина равна 34 единицам масштаба. Разобьем этот интервал на несколько частей. Это будут разряды интервального ряда.

Решение (2)

Обычно разряды имеют одинаковую длину, но это не очень существенно. Гораздо важнее обеспечить при выборе разрядов выполнение следующих условий:

1) число разрядов не должно быть слишком маленьким (не меньше пяти-шести);

2) число значений величины X, попавших в каждый разряд, тоже не должно быть слишком маленьким (не менее четырех-пяти).

Решение (3)

Эти рекомендации могут меняться в зависимости от конкретного статистического материала и конкретных задач. Проведем прикидку. В нашей задаче дано всего 40 чисел. Полагая в среднем на один разряд 6-7 значений, мы получим 6 разрядов. Поскольку суммарная длина всех разрядов 34, то примерная длина одного разряда равна шести. Возьмем первый разряд длиной, равной 4, тогда все остальные разряды можно взять с одной и той же длиной 6. Теперь составим интервальный ряд, дополнив его строкой, где записаны плотности частот h_i.

Решение (4)

В первый разряд попали 13 чисел: 3, 2, 3, 5, 4, 4, 2, 2, 4, 3, 2, 1, 2. Но соответствующая частота равна $m_1 = 12{,}5$, т.к. Число 5 находится на границе интервала и встречается только один раз. Во второй разряд попадают 15 чисел: 8, 6, 5, 10, 7, 8, 9, 6, 6, 7, 6, 9, 6, 9, 7. Но в таблицу мы запишем число $m_2 = 14{,}5$, т.к. половина пятерки уже отнесена к первому разряду. Действуя далее аналогично, заполняем все строки таблицы.

Решение (5)

Разряды	**1 – 5**	**5 – 11**	**11 – 17**	**17 – 23**	**23 – 29**	**29 – 35**	$\sum$
m_i	12,5	14,5	4,5	3	3,5	2	40
$\tilde{p}_i$	0,3125	0,3625	0,1125	0,0750	0,0875	0,0500	1
h_i	0,0781	0,0604	0,0188	0,0125	0,0146	0,0083	–

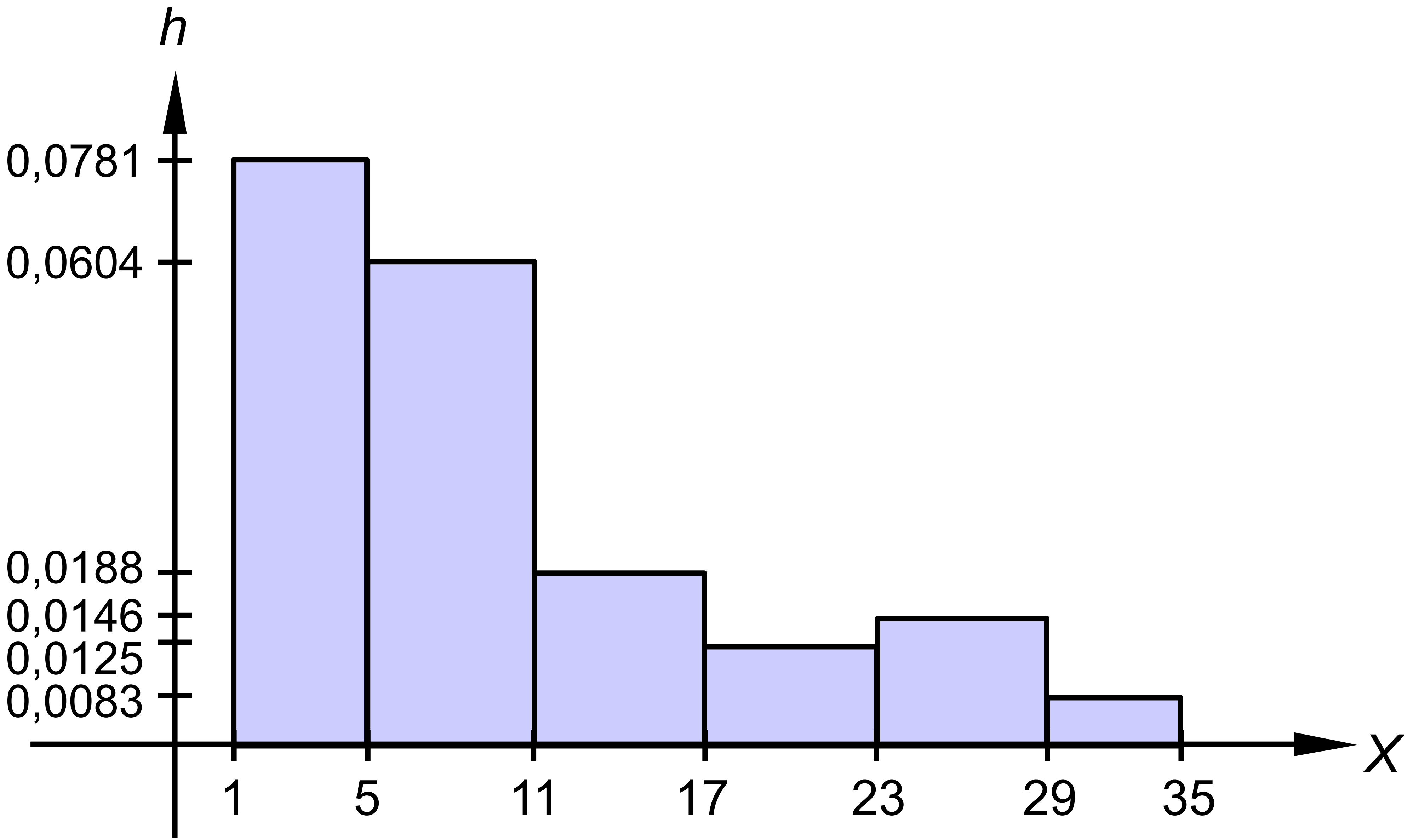
Решение (6)

Несмотря на то, что среди абсолютных частот могут быть дроби, их сумма всегда равна количеству n всех данных чисел.

Если некоторые разряды имеют малую абсолютную частоту, то их можно объединять. Мы могли бы проделать такую операцию, объединив третий разряд с четвертым и пятый с шестым.

Частоты разрядов можно округлять, но так, чтобы и после округления их сумма равнялась единице.

Решение (7)



Решение (8)

Ответим на поставленные вопросы.

Число преступлений в одном районе можем принимать значения от 1 до 35.

Наибольшую частоту 0,3625 имеет разряд $5-11$. Попадание значений величины X в этот разряд по сравнению с другими наиболее вероятно. Иными словами, следует ожидать в среднем по 5-11 преступлений в год.

Решение (9)

Для того чтобы найти частоту попадания величины X в заданный интервал, нужно вычислить площадь участка гистограммы, опирающегося на этот интервал. Например, на интервал $8-18$ опирается прямоугольник с основанием от 8 до 18 и высотой 0,0604, полностью третий столбик гистограммы и прямоугольник с основанием от 17 до 18 и высотой 0,0125. Их суммарная площадь равна 0,3062. Это и будет частотой попадания величины X в промежуток $[8;\,18]$. Следовательно, число районов, в которых совершено от 8 до 18 преступлений, составляет примерно 31% от всех районов области.

Замечание (1)

Как мы уже отмечали, интервальный ряд составляют при обработке больших массивов информации. В таких случаях, как правило, отдельные значения величины X не фиксируют, а подсчитывают абсолютные частоты разрядов, т.е. количество значений величины X, попавших в каждый разряд. Например, статистические данные позволяют точно указать количество малолетних преступников в стране, но указать точный возраст для каждого из них практически невозможно — полученная таблица, если даже и удастся ее составить, будет практически необозримой и крайне неудобной для статистической обработки.

Замечание (2)

Поэтому исследователь, не зная отдельных значений наблюдаемой величины X, не может воспользоваться формулами для вычисления среднего арифметического, дисперсии и среднего квадратического отклонения. Но приближенное значение этих числовых характеристик можно найти с помощью интервального ряда.

Замечание (3)

Для этого вспомним идею интервального ряда: будем считать, что все числа, попавшие в один разряд, приблизительно равны друг другу, и на этом основании заменим их одним числом — серединой разряда.

Пусть всего имеется k разрядов. Обозначим их середины $\tilde{x}_1$, $\tilde{x}_2$, …., $\tilde{x}_k$. Число $\tilde{x}_1$ заменяет собой m_1 чисел, попавших в первый разряд, поэтому будем считать, что число $\tilde{x}_1$ встречается m_1 раз. Иными словами, m_1 — это абсолютная частота значения $\tilde{x}_1$. Точно так же, абсолютную частоту m_2 второго разряда будем считать абсолютной частотой середины $\tilde{x}_2$ второго разряда и т.д.

Замечние (4)

После этого, поделив каждую абсолютную частоту на сумму всех абсолютных частот, мы найдем относительные частоты $\tilde{p}_1$, $\tilde{p}_2$, …., $\tilde{p}_k$, а затем по известным формулам вычислим среднее арифметическое, дисперсию и среднее квадратическое отклонение:

$$\bar{x} = \tilde{x}_1 \tilde{p}_1 + \tilde{x}_2 \tilde{p}_2 + \ldots + \tilde{x}_k \tilde{p}_k,$$

$$D = (\tilde{x}_1 - \bar{x})^2 \tilde{p}_1 + (\tilde{x}_2 - \bar{x})^2 \tilde{p}_2 + \ldots + (\tilde{x}_k - \bar{x})^2 \tilde{p}_k,$$

$$S = \sqrt{(D)}.$$

Задача

Найдите среднее арифметическое, дисперсию и средне квадратическое отклонение по интервальному ряду, представленному следующей таблицей:

	130 – 170	170 – 210	210 – 250	250 – 290	290 – 330	330 – 370
m_i	4,5	5	12	14,5	9	5
$\tilde{p}_i$	0,09	0,10	0,24	0,29	0,18	0,10

Решение (1)

i	$\tilde{x}_i$	$\tilde{p}_i$	$\tilde{x}_i \tilde{p}_i$	$\tilde{x}_i - \bar{x}$	$(\tilde{x}_i - \bar{x})^2$	$(\tilde{x}_i - \bar{x})^2 \tilde{p}_i$
1	150	0,09	13,5	-106,8	11406,24	1026,56
2	190	0,10	19,0	-66,8	4462,24	446,22
3	230	0,24	55,2	-26,8	718,24	172,38
4	270	0,29	78,3	13,2	174,24	90,53
5	310	0,18	55,8	53,2	2830,24	509,44
6	350	0,10	35,0	93,2	8686,24	868,62
$\sum$	**–**	**1**	**256,8**	**–**	**–**	**3113,75**

Решение (2)

В первом столбце записаны номера разрядов, во втором — числа $\tilde{x}_i$ (середины разрядов), в третьем — относительные частоты чисел и т.д. Таблица заполняется по столбцам. Середину разряда вычисляем как полусумму его границ:

$$\tilde{x}_1 = \frac{130+170}{2} = 150,$$

$$\tilde{x}_2 = \frac{170+210}{2} = 190$$

и т.д.

Решение (3)

Сумма чисел четвертого столбца дает среднее арифметическое: $\bar{x} = 256{,}8$. Сумма чисел последнего столбца равна дисперсии: $D = 3113{,}75$. Наконец, среднее квадратическое отклонение: $S = \sqrt{3113{,}75} \approx 55{,}80$.

Поскольку при решении задачи мы заменяли наблюдаемые значения величины X их приближенными значениями — серединами разрядов, в которые они попадали, поэтому среднее арифметическое, дисперсия и среднеквадратическое отклонение также вычислены приближенно, а не точно.

Задача

Управление сельского хозяйства Брюковского района представило сводку по пятидесяти хозяйствам. Согласно этим сводкам урожайность ржи (в центнерах с гектара) в хозяйствах оказалась следующей: 16.8, 15.7, 19.6, 20.1, 16.8, 15.7, 19.0, 19.0, 19.0, 17.0, 17.0, 17.0, 21.3, 18.7, 19.6, 20.0, 15.8, 24.1, 21.0, 18.9, 17.2, 18.7, 20.0, 22.3, 20.1, 21.5, 22.2, 18.5, 19.3, 22.5, 18.1, 19.7, 18.4, 20.1, 19.5, 21.2, 19.7, 18.5, 17.4, 24.1, 22.2, 18.2, 19.7, 18.9, 18.4, 19.4, 17.5, 18.7, 17.7, 16.6. Постройте модель урожайности ржи для одного хозяйства.

Решение (1)

Как мы знаем, в состав модели входят интервальный ряд, гистограмма и числовые характеристики. Начнем с построения интервального ряда.

Самое мальнькое значение урожайности – 15.7, а самое большое – 24.1. Диапазон значений – интервал $[15.7\,;\;24.1]$, его длина равна 8.4.

Прикидка показывает, что можно взять 7 разрядов, тогда длина одного разряда будет равна 1.2.

Решение (2)

Разряды	m_i	$\tilde{p}_i$	h_i
15,7 – 16,9	6	0,12	0,100
16,9 – 18,1	7,5	0,15	0,125
18,1 – 19,3	13,5	0,27	0,225
19,3 – 20,5	14	0,28	0,233
20,5 – 21,7	4	0,08	0,067
21,7 – 22,9	3	0,06	0,050
22,9 – 24,1	2	0,04	0,033
$\sum$	50	1	–

Решение (3)

i	$\tilde{x}_i$	$\tilde{p}_i$	$\tilde{x}_i\tilde{p}_i$	$\tilde{x}_i-\bar{x}$	$(\tilde{x}_i-\bar{x})^2$	$(\tilde{x}_i-\bar{x})^2\tilde{p}_i$
1	16,3	0,12	1,956	-2,868	8,225424	0,987050
2	17,5	0,15	2,625	-1,668	2,849344	0,427401
3	18,7	0,27	5,049	-0,468	0,219024	0,059136
4	19,9	0,28	5,572	0,732	0,535824	0,150030
5	21,1	0,08	1,688	1,932	3,732634	0,298610
6	22,3	0,06	1,338	3,132	9,809424	0,588565
7	23,5	0,04	0,940	4,332	18,766224	0,750648
$\sum$	**–**	**1**	**19,168**	**–**	**–**	**3,2614**
			$\bar{x}$			D

Решение (4)

В последней строке таблицы записаны среднее арифметическое и дисперсия. Найдем среднее квадратическое отклонение

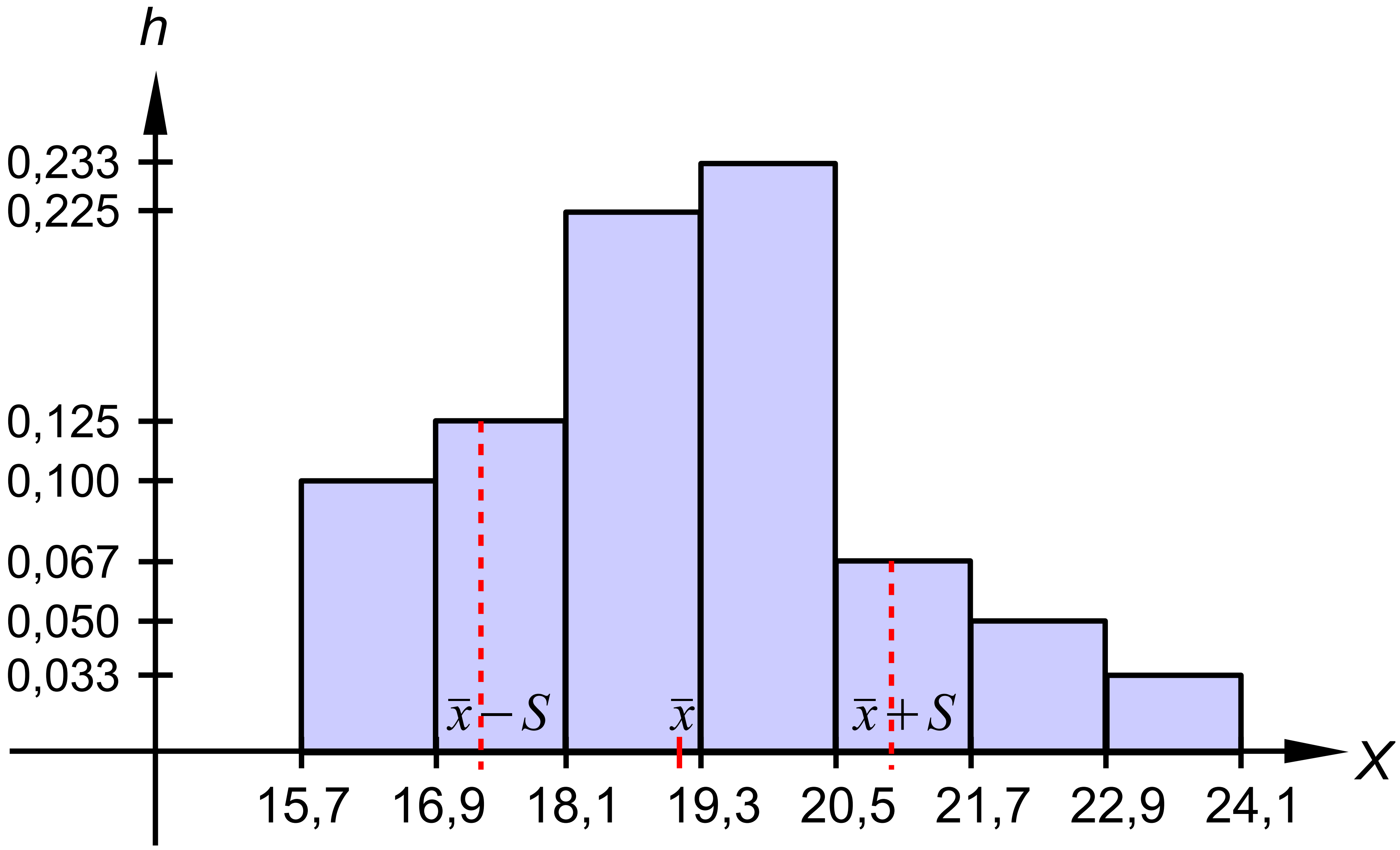
$$S=\sqrt{3{,}2614}=1{,}8059\ldots$$

Поскольку выборочные значения приведены с точностью до десятых, то в записи числовых характеристик следует оставить на один знак после запятой больше. Таким образом,

$$\bar{x}=19{,}17\,,\;\; D=3{,}26\,,\;\; S=1{,}80\,.$$

Построим гистограмму, отметив на ней число $\bar{x}$ и интервал наиболее вероятных значений $\left(\bar{x}-S\,;\;\bar{x}+S\right)=\left(17{,}37\,;\;20{,}97\right)$.

Решение (5)



h
0,233
0,225
0,125
0,100
0,067
0,050
0,033
$\bar{x}-S$
$\bar{x}$
$\bar{x}+S$
X
15,7
16,9
18,1
19,3
20,5
21,7
22,9
24,1

Решение (6)

Урожайность ржи в хозяйствах Брюковского района принимает значения от 15,7 до 21,4 (ц/га). Наиболее вероятные значения урожайности — от 19,3 до 20,5 (14% хозяйств); два процента составляют хозяйства с максимальной урожайностью, от 22,9 до 24,1 (ц/га); 6% хозяйств имеют самую маленькую урожайность — от 15,7 до 16,9 (ц/га). Среднее значение урожайности по всем хозяйствам составляет 19,17 ц/га, среднее отклонение от среднего значения равно 1,80. Хозяйства, в которых урожайность отклоняется от среднего значения не более, чем на 1,80 ц/га, составляет 69% от общего числа хозяйств. Сравнительно малые значения урожайности (меньше 17,37) наблюдается в 16% хозяйств; 15% хозяйств имеют высокую урожайность (от 20,97 до 24,1 ц/га).