§ 19. СТАТИСТИКА — ДИЗАЙН ИНФОРМАЦИИ

Начнем с конкретного примера. Допустим, что в 9-х классах «А» и «Б» измерили рост (в сантиметрах) 50 учеников. Получился набор из 50 чисел. Вряд ли самое маленькое из них будет меньше 140, а самое большое — больше 200. Можно, соблюдая очередность измерений, выписать все данные в строчку через запятую. Можно расположить их в две колонки в соответствии с классными списками. Можно как-то записать их в виде таблицы 5×10 и т. п. В итоге будет собрана полная информация о проведенном измерении. К сожалению, практически при любом способе расположения *абсолютно всех* данных эта информация трудно читается: она не наглядна, занимает много места, никак не упорядочена и т. д.

А представьте результаты, состоящие не из 50 данных, а из 500, 5000 или из миллионов различных чисел! Например, число и размеры вкладов в Сбербанке за текущий год или данные

о производительности труда на предприятиях какой-нибудь отрасли по всей стране, результаты голосования по всем избирательным участкам и т. п. Единственный разумный выход — каким-то образом *преобразовать* первоначальные данные измерения, в первую очередь заметно уменьшив их общее количество. Одна из основных задач статистики как раз и состоит в надлежащей *обработке информации*. Конечно, у статистики есть много других задач: получение и хранение информации, выработка различных прогнозов, оценка их достоверности и т. д. Ни одна из этих целей не достижима без обработки данных.

Итак, в первую очередь займемся статистическими методами обработки информации. Как правило, порядок преобразований первоначально полученной информации таков:

1) сначала данные измерений *упорядочивают* и *группируют*;
2) затем составляют *таблицы распределения данных*;
3) таблицы распределения переводят *в графики распределения*;
4) наконец, получают своего рода *паспорт данных* измерения, в котором собрано небольшое количество основных *числовых характеристик* полученной информации.

Зафиксируем одно конкретное измерение и проследим шаг за шагом, как его данные преобразуются в процессе обработки информации.

И з м е р е н и е (И). У 50 работников городского предприятия попросили оценить время, которое они в среднем тратят на проезд от дома до работы. Получились следующие данные в минутах (с точностью до 10 минут).

20	100	20	30	40	50	30	80	90	40
30	50	20	50	30	30	50	60	60	50
30	40	60	50	100	60	90	10	20	50
90	80	20	40	50	10	50	40	30	40
60	120	30	40	60	20	60	10	50	60

1. Группировка информации. Первое, что следует оценить, — это рамки, в которых вообще могут находиться данные измерения. Менее 10 минут (т. е. 0 минут) никто не заявил (территориально дом и работа — это не одно и то же), а более 180 минут (более трех часов) на проезд по городу в одну сторону вряд ли кто-то будет тратить каждый день. Значит, в принципе в этом

измерении могли получиться числа 10, 20, 30, ..., 160, 170, 180. Мы составили *общий ряд данных*. Данные располагают, как правило, *в порядке возрастания.*

Итак:

Измерение	Общий ряд данных
Время проезда (мин)	10, 20, 30, ..., 170, 180

П р и м е р 1. Выписать общий ряд данных следующих измерений:

а) месяц рождения учеников вашей школы; б) год рождения ваших родственников и знакомых; в) годовой процент начислений по вкладам в банке; г) начальные буквы в первой строке стихотворения.

Р е ш е н и е. а) Всего может получиться 12 месяцев. Если перечислить их не по названиям, а по номерам, то получим общий ряд данных:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

б) Вряд ли у вас есть родственники или знакомые, которым много больше 100 лет, а вот новорожденные вполне могут встретиться. Значит, общий ряд данных выглядит так: 1900, 1901, 1902, ..., 2005, 2006, 2007, 2008.

в) Никакой уважающий себя банк более 15 % годовых не даст. Что касается нижней оценки, то тут менее 0,1 %, которые дает Сбербанк России по вкладам до востребования, невозможно представить. Значит, в этом случае общий ряд данных выглядит так: 0,1; 0,2; ...; 0,9; 1; 2; 3; ...; 14; 15.

г) В первой строке стихотворения в принципе могут встретиться все буквы русского алфавита от А до Я. Следует исключить нереальные случаи (Ь, Ъ, Ы). Оставшиеся буквы можно, например, перенумеровать по порядку и перейти к числовому общему ряду: 1, 2, 3, ..., 29, 30. ▣

Подчеркнем, что определения в статистике не всегда носят столь же точный характер, как, скажем, определения в геометрии или алгебре. Например, в пункте б) примера 1 от добавления 1899 к последовательности 1900, 1901, 1902, ..., 2008 она не перестанет быть общим рядом данных. В пункте в) все годовые проценты можно было измерять с точностью до десятых, и тогда

общий ряд данных составили бы числа 0,1; 0,2; ...; 0,9; 1; 1,1; ...; 14,9; 15.

При проведении конкретного измерения вполне может случиться так, что какие-то данные из общего ряда вообще не встретятся. Значит, надо отличать реально полученные результаты измерения от общего ряда данных. Например, в измерении (И) нам встретились только такие результаты: 10, 20, 30, 40, 50, 60, 80, 90, 100, 120. Каждое из этих чисел называют *вариантой* измерения (несколько непривычно, но в статистике используют слово именно женского рода).

Варианта измерения — *один из результатов этого измерения.*

Если все варианты измерения перечислить по порядку (и без повторений), то получится *ряд данных измерения*. В нашем измерении (И) ряд данных — это 10, 20, 30, 40, 50, 60, 80, 90, 100, 120.

Итак:

Измерение	Общий ряд данных	Ряд данных измерения
Время проезда (мин)	10, 20, 30, ..., 170, 180	10, 20, 30, 40, 50, 60, 80, 90, 100, 120

П р и м е р 2. Выписать ряд данных измерения, состоящего из всех разных букв первых двух строк стихотворений:

а) «Не говори никому / Всё, что ты видел, забудь...»*;

б) «Это дерево сосна, / И судьба сосны ясна...»**.

Р е ш е н и е. а) *а, б, в, г, д, е, ё, з, и, к, л, м, н, о, р, с, т, у, ч, ы, ь*. Тут использована 21 буква на 33 местах и повторений букв очень мало.

б) *а, б, в, д, е, и, н, о, р, с, т, у, ы, ь, э, я*. Здесь использованы 16 букв на 30 местах, повторений больше, и в особенности это относится к букве «с» (6 повторений). ■

Как мы видим, не все варианты конкретного измерения находятся в одинаковом положении. Какие-то встречаются много раз, какие-то реже, а некоторые встречаются по одному разу.

---

* Из стихотворения О. Мандельштама.

** Из стихотворения Ю. Минералова.

Определение. Если среди всех данных конкретного измерения одна из вариант встретилась ровно k раз, то число k называют **кратностью** этой варианты измерения.

Например, в измерении (И) время 60 минут встретилось восемь раз, а время 120 минут встретилось однажды. Значит, кратность варианты 60 равна восьми, а кратность варианты 120 равна единице. Перед дальнейшей обработкой данные измерения удобно *сгруппировать*. Делается это так. Для удобства запишем данные измерения (И) в десятках минут.

~~2~~	~~10~~	~~2~~	3	4	5	3	8	9	4
3	5	2	5	3	3	5	6	6	5
3	4	6	5	10	6	9	1	2	5
9	8	2	4	5	1	5	4	3	4
6	12	3	4	6	2	6	1	5	6

Будем двигаться по строчкам и зачеркивать очередные результаты, а каждое зачеркивание копировать ниже соответствующей варианты в ряду данных. Первые три результата 2, 10, 2 в первой строке зачеркнуты в знак того, что мы их уже учли. Линии, которыми эти результаты перечеркнуты, повторим в выписанном заранее общем ряде данных:

1	2	3	4	5	6	7	8	9	10	11	12
	//								/		

Вот что получится после прохождения первой строки: в ней по два раза встретились варианты 2, 3, 4 и по одному разу — варианты 5, 8, 9, 10.

1	2	3	4	5	6	7	8	9	10	11	12
	//	//	//	/			/	/	/		

Результат после прохождения первых двух строк выглядит так:

1	2	3	4	5	6	7	8	9	10	11	12
	///	~~////~~	//	~~////~~	//		/	/	/		

Для удобства подсчетов вместо каждой пятой черточки проводят линию с другим наклоном, которая перечеркивает четыре

предыдущие черточки. На практике, конечно, все подсчеты производят в одном месте: ведь промежуточные результаты не нужны. Вот как в итоге будет выглядеть результат подсчета кратностей в нашем примере:

1	2	3	4	5	6	8	9	10	12
///	𝍸 /	𝍸 ///	𝍸 //	𝍸 𝍸	𝍸 ///	//	///	//	/

Теперь можно составить *сгруппированный ряд данных*. В нем каждая варианта повторена именно столько раз, сколько она встретилась в измерении, т. е. число повторений каждой варианты равно кратности этой варианты:

$$1, 1, 1, \underbrace{2, \ldots, 2}_{6}, \underbrace{3, \ldots, 3}_{8}, \underbrace{4, \ldots, 4}_{7}, \underbrace{5, \ldots, 5}_{10}, \underbrace{6, \ldots, 6}_{8}, 8, 8, 9, 9, 9, 10, 10, 12$$

На этом заканчивается первый шаг обработки информации — ее упорядочивание и группировка.

2. Табличное представление информации. Внесем в таблицу ряд данных измерения и кратности соответствующих вариант. Получим *таблицу распределения* данных. Вот как это выглядит в измерении (И).

	Варианта										**Сумма**
	1	**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**12**	
Кратность	3	6	8	7	10	8	2	3	2	1	50

Если сложить все кратности, то получится количество всех данных измерения — *объем измерения*. Так как опрашивали 50 работников, то объем измерения (И) равен именно 50. На практике для контроля всегда складывают найденные кратности вариант: сумма должна равняться объему измерения.

Далее, при общей оценке распределения данных не очень важно, что, например, варианта 1 имеет кратность 3 среди всех 50 данных. Так как $\frac{3}{50} = 0{,}06$, то удобнее сказать, что эта варианта занимает шесть сотых общего объема измерения. Так и поступают, т. е. делят кратность варианты на объем измерения и получают **частоту** варианты.

$$\text{Частота варианты} = \frac{\textit{Кратность варианты}}{\textit{Объем измерения}}$$

Частоты всех вариант удобно приписать следующей строкой к уже составленной таблице. Полученную таблицу называют *таблицей распределения частот* измерения. Вот как это выглядит в измерении (И).

	Варианта										Сумма
	1	2	3	4	5	6	8	9	10	12	
Кратность	3	6	8	7	10	8	2	3	2	1	50
Частота	0,06	0,12	0,16	0,14	0,2	0,16	0,04	0,06	0,04	0,02	1

Сумма всех частот всегда равна 1 — ведь это сумма дробей с одинаковым знаменателем, у которых сумма всех числителей как раз и равна этому знаменателю. Для удобства счета и построения графиков частоты переводят в проценты от объема измерения. Тогда таблицу распределения дополняют еще строкой частот в процентах. Она получается из предыдущей строки умножением на 100 %. Итак, для измерения (И) получаем такую таблицу.

	Варианта										Сумма
	1	2	3	4	5	6	8	9	10	12	
Кратность	3	6	8	7	10	8	2	3	2	1	50
Частота	0,06	0,12	0,16	0,14	0,2	0,16	0,04	0,06	0,04	0,02	1
Частота, %	6	12	16	14	20	16	4	6	4	2	100

Сумма всех частот в процентах, конечно же, равна 100.

3. Графическое представление информации. Итак, распределение данных измерения удобно задавать с помощью таблиц. Но мы знаем, что и для функций есть *табличный* способ их задания. Таблицы образуют «мостик», по которому от распределения данных можно перейти к функциям и графикам.

Отложим по оси абсцисс значения из первой строки таблицы распределения, а по оси ординат — значения из ее второй строки. Построим соответствующие точки в координатной плоскости. Получим графическое изображение имеющейся информации — *график распределения выборки*. Построенные точки для наглядности соединяют отрезками. Вот как это выглядит в измерении (И), данные которого мы уже представили в табличном виде.

По оси абсцисс	1	2	3	4	5	6	8	9	10	12
По оси ординат	3	6	8	7	10	8	2	3	2	1

На координатной плоскости мы получили ломаную линию (рис. 134), которая является графиком некоторой кусочно-линейной функции. Эту ломаную называют *многоугольником* или *полигоном распределения* данных. Собственно, *polygon* и переводится как «многоугольник».

Точно так же составленные таблицы распределения частот и распределения частот в процентах позволяют построить *многоугольник частот* и *многоугольник частот в процентах*. Для наглядности в практических приложениях удобнее использовать многоугольники частот в процентах: в них изменения по оси ординат от 1 до 100 более выразительны, чем изменения от 0 до 1.

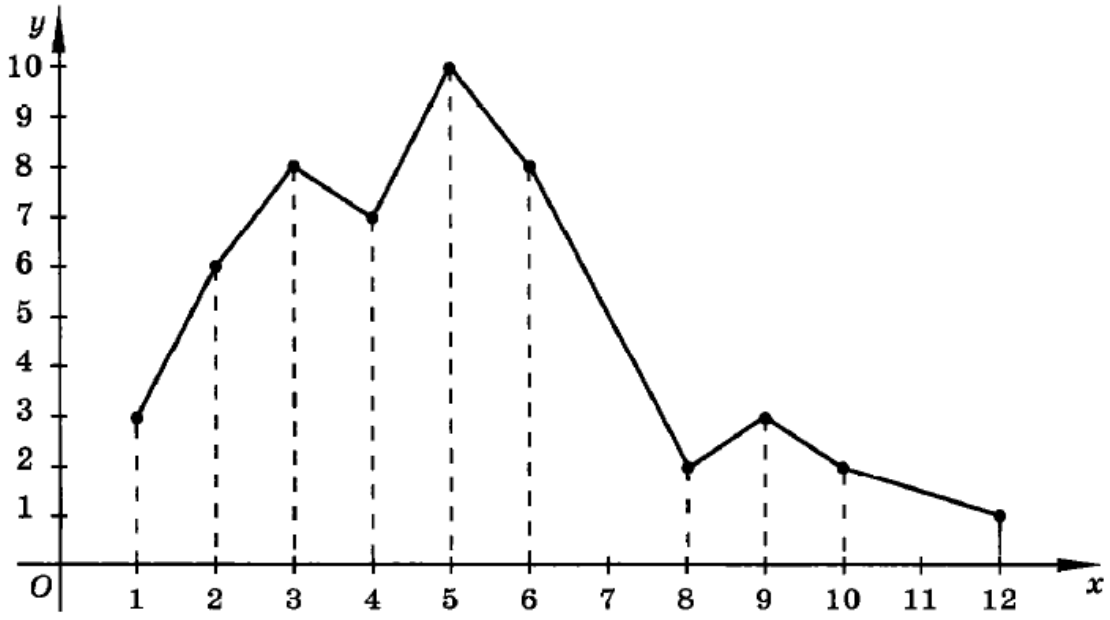

Рис. 134

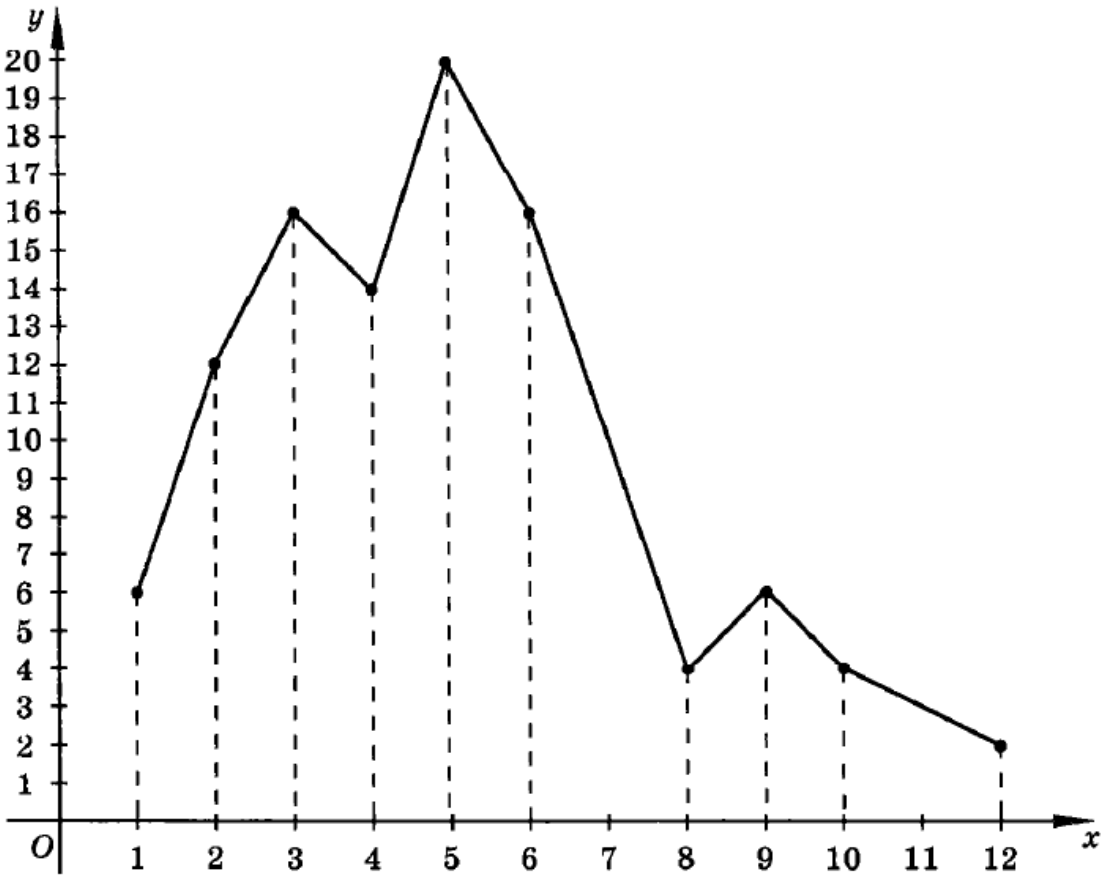

Рис. 135

Построим многоугольник частот в процентах для измерения (И) (рис. 135).

По оси абсцисс	1	2	3	4	5	6	8	9	10	12
По оси ординат	6	12	16	14	20	16	4	6	4	2

Мы видим, что даже для небольшого объема измерений аккуратное «причесывание» информации — довольно кропотливая работа. При оперировании с большими объемами информации используют методы приближенной группировки данных. В таких случаях вариантой измерения является не одно число, а числовой промежуток.

Например, в измерении (И) всех работников предприятия можно разделить на три группы. Во-первых, это те, кто живет близко от работы. Они тратят на дорогу 10, 20 или 30 минут.

Во-вторых, это те, кто живет недалеко. Их путь занимает от 40 до 60 минут. Все остальные живут далеко и проводят в дороге более часа. Тем самым мы разбили промежуток между самой маленькой и самой большой вариантой на промежутки 1—3; 4—6; 8—12 (в десятках минут). Вместо десяти первоначальных вариант получилось всего три новые: близко, недалеко, далеко.

Для каждого промежутка можно найти количество результатов измерения, попавших в этот промежуток. Получим кратности и таблицу распределения новых вариант.

	Варианта			Сумма
	близко	недалеко	далеко	
Кратность	17	25	8	50

Разумеется, можно составить и таблицы распределения частот и процентных частот новых вариант. Вот что получится.

	Варианта			Сумма
	близко	недалеко	далеко	
Кратность	17	25	8	50
Частота, %	34	50	16	100

При такой грубой оценке мы кое-что утеряли из первоначальной информации. Например, теперь неизвестно количество работников, путь которых составляет именно 60 минут. Но мы что-то и приобрели: информация получила более ясное и удобное для объяснения представление. Вот как это выглядит, например, на круговой диаграмме (рис. 136).

Рис. 136

При графическом представлении больших объемов информации многоугольники распределения заменяют *гистограммами*, или *столбчатыми диаграммами*. Вы познакомитесь с ними в старшей школе.

4. Числовые характеристики данных измерения. Каждый человек, кроме своих формальных паспортных данных, обладает

рядом других свойств и качеств. Кто-то лучше всех решает задачи, кто-то брюнет, кто-то замечательно играет на гитаре и т. п. Однако сравнительно небольшая паспортная информация (ФИО, дата рождения, номер и дата выдачи паспорта) позволяет однозначно определить человека, выделить его среди других. У данных измерений тоже есть своего рода краткий паспорт, состоящий из набора основных *числовых характеристик*. Поясним некоторые из них на уже знакомом нам примере измерения (И).

Разность между максимальной и минимальной вариантами называют *размахом измерения*. В измерении (И) размах равен 120 − 10 = 110 минутам. На графике (см. рис. 134, 135) это длина области определения многоугольника распределения данных или распределения частот.

Ту варианту, которая в измерении встретилась чаще других, называют *модой измерения*. Если данные измерения уже собраны в двухстрочную таблицу распределения, то для нахождения моды следует:

— во второй строке (кратность) выбрать наибольшее число;

— от найденного числа подняться на клетку выше: полученное число и будет модой.

Если данные измерения представлены графически в виде многоугольника распределения, то мода — это точка, в которой достигается максимум многоугольника распределения. Например, в измерении (И) мода равна 50 минутам — наибольшее число работников (10) именно так оценивают время своего проезда.

Наиболее важной характеристикой числового ряда данных является *среднее значение (среднее арифметическое,* или просто *среднее)*.

Для нахождения среднего значения следует:

1) просуммировать все данные измерения;

2) полученную сумму разделить на количество данных.

Для подсчета среднего значения удобно использовать сгруппированный ряд данных. Вот как это выглядит в измерении (И).

Сгруппированный ряд данных измерения
$\underbrace{1, 1, 1}_{3}, \underbrace{2, \ldots, 2}_{6}, \underbrace{3, \ldots, 3}_{8}, \underbrace{4, \ldots, 4}_{7}, \underbrace{5, \ldots, 5}_{10}, \underbrace{6, \ldots, 6}_{8}, 8, 8, \underbrace{9, 9, 9}_{3},$ $10, 10, 12$

Найдем среднее значение:

$$\frac{1 \cdot 3 + 2 \cdot 6 + 3 \cdot 8 + 4 \cdot 7 + 5 \cdot 10 + 6 \cdot 8 + 8 \cdot 2 + 9 \cdot 3 + 10 \cdot 2 + 12 \cdot 1}{50} =$$

$$= \frac{3 + 12 + 24 + 28 + 50 + 48 + 16 + 27 + 20 + 12}{50} = 4{,}8$$ (десятков минут). Значит, среднее время проезда для опрошенных работников составляет 48 минут.

Если таблица распределения частот данных уже известна, то среднее значение можно вычислять прямо по ней. Смотрите:

$$\frac{1 \cdot 3 + 2 \cdot 6 + 3 \cdot 8 + \ldots + 12 \cdot 1}{50} = 1 \cdot \frac{3}{50} + 2 \cdot \frac{6}{50} + 3 \cdot \frac{8}{50} + \ldots +$$

$$+ 12 \cdot \frac{1}{50}.$$

Все дроби в последней сумме — это частоты вариант, которые стоят перед этими дробями в качестве множителей. Значит, в таблице распределения частот можно просто перемножить числа в каждом столбце и затем сложить все полученные произведения.

	Варианта										Сумма
	1	2	3	4	5	6	8	9	10	12	
Частота	0,06	0,12	0,16	0,14	0,2	0,16	0,04	0,06	0,04	0,02	1

Проверяйте: $1 \cdot 0{,}06 + 2 \cdot 0{,}12 + 3 \cdot 0{,}16 + 4 \cdot 0{,}14 + 5 \cdot 0{,}2 + 6 \cdot 0{,}16 + 8 \cdot 0{,}04 + 9 \cdot 0{,}06 + 10 \cdot 0{,}04 + 12 \cdot 0{,}02 = 4{,}8$.

Сформулируем общее правило.

Для нахождения среднего значения данных измерения можно:
1) каждую варианту умножить на ее частоту;
2) сложить все полученные произведения.

Мы закончим этот параграф еще одним конкретным примером измерения, кратко повторив для него все шаги 1)—4) обработки данных (см. с. 183).

Пример 3. На вступительном письменном экзамене по математике можно получить от 0 до 10 баллов. Сорок абитуриентов получили такие оценки:

6	7	7	8	9	2	10	6	5	6
7	3	7	9	9	2	3	2	6	6
6	7	8	8	2	6	7	9	7	5
9	8	2	6	6	3	7	7	6	6

а) Составить общий ряд данных; упорядочить и сгруппировать полученные оценки.

б) Составить таблицы распределения данных и распределения частот.

в) Построить графики распределения данных и распределения частот.

г) Найти размах, моду и среднее измерения.

Решение. а) В принципе возможны такие оценки: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Это *общий ряд данных*. В конкретно предложенном измерении встретились только такие оценки: 2, 3, 5, 6, 7, 8, 9, 10. Это *ряд данных*, все числа в нем — *варианты измерения*. Наконец,

$$\underbrace{2, \ldots, 2}_{5}, 3, 3, 3, 5, 5, \underbrace{6, \ldots, 6}_{11}, \underbrace{7, \ldots, 7}_{9}, \underbrace{8, \ldots, 8}_{4}, \underbrace{9, \ldots, 9}_{5}, 10 —$$

это *сгруппированный ряд данных.*

б) Всего выставлено 40 оценок. Значит, 40 — это *объем* данного измерения. Соберем кратности всех восьми вариант в таблицу; подсчитаем и внесем в ту же таблицу все частоты.

	Варианта								Сумма
	2	3	5	6	7	8	9	10	
Кратность	5	3	2	11	9	4	5	1	40
Частота	0,125	0,075	0,05	0,275	0,225	0,1	0,125	0,025	1
Частота, %	12,5	7,5	5	27,5	22,5	10	12,5	2,5	100

Многоугольник распределения данных

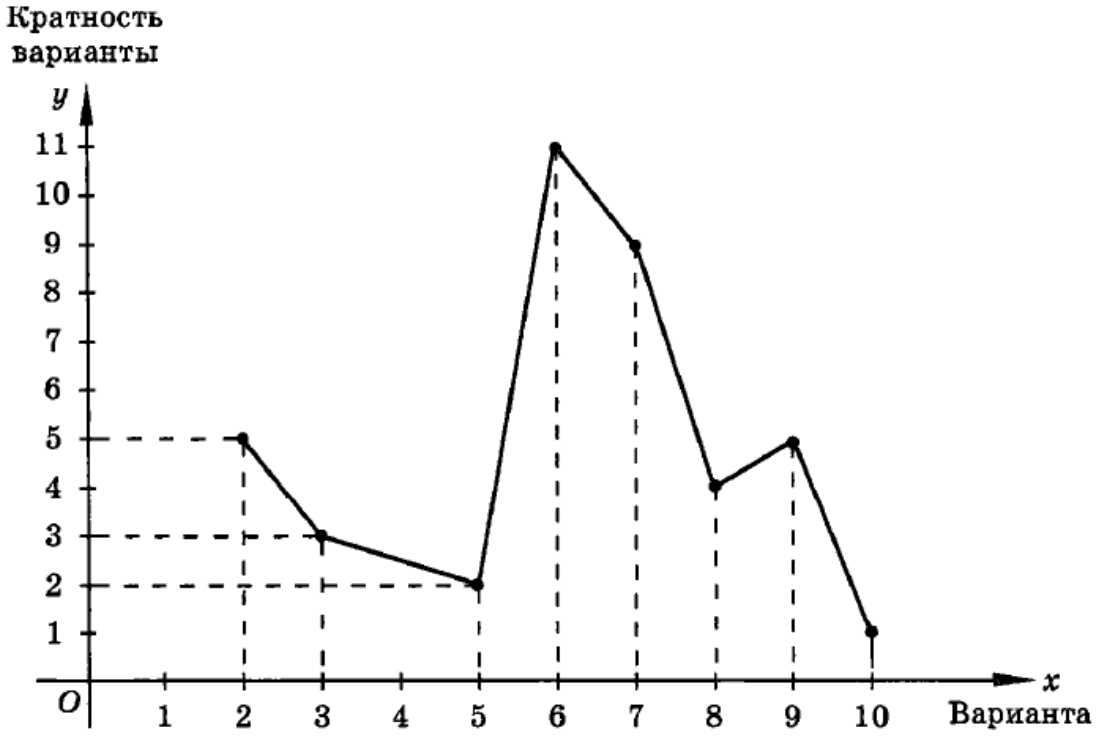

Рис. 137

Многоугольник распределения частот

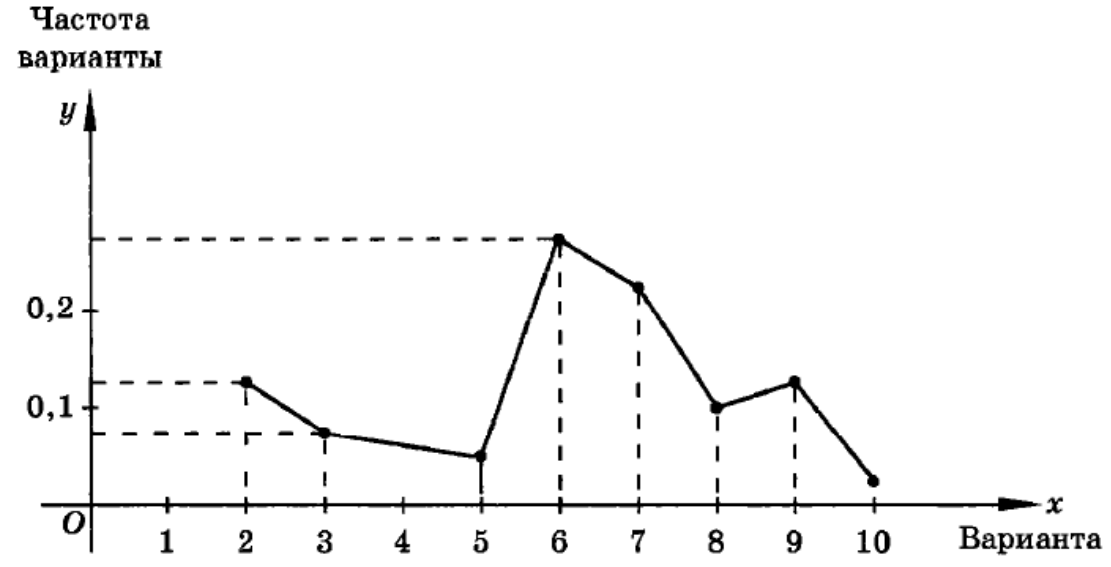

Рис. 138

Рис. 139

в) Полученная таблица распределения позволяет построить три многоугольника распределения: данных, частот и частот в процентах (рис. 137—139).

По существу, различия в этих графиках состоят только в выборе единиц измерения и масштаба по оси ординат.

г) Вернемся к первоначальным данным. *Размах* измерения равен $10 - 2 = 8$. *Мода* равна 6 — эта оценка встретилась чаще других. Наконец, вычислим *среднее значение:*

$$\frac{2 \cdot 5 + 3 \cdot 3 + 5 \cdot 2 + 6 \cdot 11 + 7 \cdot 9 + 8 \cdot 4 + 9 \cdot 5 + 10 \cdot 1}{40} =$$

$$= \frac{245}{40} = 6{,}125. \quad \blacksquare$$