

О ГИСТОГРАММЕ И ЕЁ СВОЙСТВАХ¹

Г.И.Фалин, д.ф.м.н., проф.
кафедра теории вероятностей
механико-математический факультет
МГУ им М.В.Ломоносова (Москва)
<http://mech.math.msu.su/~falin>

А.И.Фалин, к.ф.м.н., доцент
кафедра общей математики
факультет ВМиК
МГУ им М.В.Ломоносова (Москва)

Гистограмма является специфическим способом графического представления набора числовых данных. Мы рассматриваем гистограмму как функцию числового аргумента и устанавливаем её связь с другими способами статистического описания (например, функцией распределения) и основными статистическими характеристиками (скажем, средним).

Ключевые слова: *описательная статистика, гистограмма, среднее значение, функция распределения*

1 Введение

Гистограмма – это широко распространённый в статистике способ графического представления набора значений непрерывной статистической переменной. В школьных учебниках гистограмме уделяется большое внимание, но при этом рассматриваются только простейшие её виды (с интервалами группировки постоянной длины), не показывается её связь с другими способами статистического описания (например, функцией распределения) и основными статистическими характеристиками (скажем, средним), не проводится ясное различие между столбиковой диаграммой и гистограммой. Последнее является следствием недостаточного внимания в

¹ Г.И.Фалин, А.И.Фалин. О гистограмме и её свойствах. *Математика в профильной школе. ФРАКТАЛ*, 2013, №1, стр.44-62.

школе к основным понятиям статистики. Дело в том, что различие между гистограммой и столбиковой диаграммой связано с природой анализируемой переменной: если эта переменная является дискретной, то распределение её значений представляется столбиковой диаграммой, а если непрерывной – то гистограммой. Поэтому в следующем пункте мы вкратце расскажем о непрерывных переменных и (попутно) о некоторых других базовых понятиях статистики.

2 Непрерывные статистические переменные

2.1 Генеральная совокупность. Напомним, что *статистика* – это наука, которая

- изучает проблемы сбора, представления, анализа статистических данных относительно больших групп однородных объектов и явлений,
- разрабатывает методы получения на их основе обоснованных выводов относительно этой группы как целого.

Появившееся в этом определении понятие «статистические данные» нуждается в дальнейшем уточнении. При этом мы попутно дадим определения ещё двух важнейших понятий – «статистического исследования» и «генеральной совокупности».

Статистические данные – это сведения, факты, графические образы и т.п. относительно большой группы однородных объектов, которые используются людьми или организациями в качестве основы для рассуждений, анализа ситуаций, принятия решений и т.п. относительно этой группы как целого. *Статистическое исследование*, анализируя статистические данные, имеет своей целью установление свойств всей группы как целого.

Ниже, говоря о данных, мы всегда будем иметь в виду статистические данные. В математической статистике содержательный смысл данных, как правило, не играет никакой роли; обычно это просто *наборы чисел*. В специальных разделах статистики (медицинская статистика, финансовая статистика и т.д.) для получения практически значимых результатов необходимо в определённой мере принимать в расчёт соответствующую (медицинскую, финансовую и т.д.) информацию, содержащуюся в данных.

Ещё раз подчеркнём (несколько упрощая ситуацию), что статистика изучает только общие свойства группы (или, как обычно говорят в статистике, совокупности) однородных объектов как целого. Если нас интересуют индивидуальные члены этой совокупности сами по себе, статистика бесполезна. Допустим, например, что мы хотим приобрести автомобиль и перед нами в автосалоне стоит два автомобиля – один марки «А», второй – «Б». У них примерно одинаковые технические характеристики и цена. Мы знаем, что в течение трёх лет существенного ремонта требуют примерно 3% автомашин марки «А» и примерно 8% автомашин марки «Б».

Важно понимать, что эта информация совершенно ничего не говорит о качестве тех двух конкретных автомобилей, из которых мы выбираем – она характеризует большие партии машин в целом. В нашей ситуации единственно разумным будет детальная диагностика каждого из двух предложенных автомобилей.

Даже если в результате анализа статистических данных принимается решение относительно индивидуального объекта совокупности, это решение рассматривает этот индивидуальный объект не изолированно, а как элемент этой совокупности, и соотносит свойства этого индивидуального объекта со свойствами всей совокупности объектов как целого.

*Группа объектов, свойства которых изучаются в ходе статистического исследования, называется **генеральной совокупностью** (от латинского generalis – главный).*

При статистическом исследовании чрезвычайно важно понимать, о какой генеральной совокупности идёт речь, т.е. какая группа объектов исследуется. Без этого нельзя правильно организовать сбор и обработку данных, правильно интерпретировать полученные результаты и выработать рекомендации по их практическому применению. В важной методической статье [1] говорится, что «введение термина ... генеральная совокупность... не является обязательным и оставляется на усмотрение учителя и авторов учебных пособий». Однако мы убеждены, что без ясного понимания смысла этого и других базовых понятий (дискретные и непрерывные переменные, числовые и качественные переменные и т.д.) нельзя дать содержательное изложение собственно статистики и, в частности, отчётливо понять, что такое гистограмма. Поэтому соответствующий материал должен обязательно излагаться в школе.

В странах с давней традицией преподавания статистики в школах, например в Великобритании, эти вопросы включены в школьные программы по математике [2] и статистике [3] (статистика в школах Великобритании изучается и в курсе математики, и, по выбору школьника, как отдельный предмет). Соответствующая теория подробно излагается в учебниках (см., например, [4], [5]), а на выпускных школьных экзаменах по математике и статистике (см., например, [6], задача 5) предлагаются задачи по этой теме.

2.2 Статистические переменные. Статистические данные всегда характеризуют какое-то качество, свойство (или набор качеств) объектов, образующих генеральную совокупность. Предположим, например, что генеральная совокупность состоит из трёх учеников, Пети, Маши, Саши, а интересующее нас свойство этих «объектов» – число задач, решённых на контрольной работе по математике. Вообще говоря, это число меняется от ученика к ученику. Поэтому как свойство неопределённого ученика из рассматриваемой группы оно является *переменной величиной*. Поскольку эта переменная появляется в ходе статистического исследования, её называют *статистической переменной*. Мы для краткости будем опускать в

дальнейшем слово «статистическая» и говорить просто о «переменной». В нашем примере переменную обозначим W . Точное значение этой переменной определяется тем, о каком конкретно учащемся идёт речь. Если, например, Петя и Маша решили по 4 задачи, а Саша – 3, то этот факт можно выразить следующим образом: $W(\text{Петя})=4$, $W(\text{Маша})=4$, $W(\text{Саша})=3$. Поскольку значения нашей переменной W являются числами, её называют *числовой* (или *количественной*) переменной, а соответствующий набор значений переменной $[4;4;3]$ – *числовым набором*. Разделять элементы набора друг от друга можно не точкой с запятой «;», но и запятой «,» (конечно, если только нет опасности перепутать её с десятичной запятой).

В статистике обычно работают только с набором значений переменной и фактически отождествляют статистическую переменную и набор её значений. Но важно понимать, что числовые наборы в статистике – это не какие-то произвольные, неизвестно откуда появившиеся группы чисел. Любой числовой набор в статистике всегда является результатом измерения какого-то свойства объектов определённой генеральной совокупности, а каждое число из набора относится к *одному* объекту этой генеральной совокупности. Эти замечания относятся и к нечисловым переменным (их называют *качественными переменными*).

Школьник, знакомый с основами теории множеств, без труда поймёт, что статистическая числовая переменная – это отображение множества изучаемых объектов (генеральной совокупности) в множество действительных чисел.

В школьных учебниках по статистике (см., например, [4]) часто используют термин «данные» вместо термина «переменная»; в более серьёзных пособиях для школьников (см., например, [5]) эти термины употребляют именно так, как принято в современной статистике, т.е. в соответствии с данными выше определениями.

Понятие «набор» нуждается в дополнительных комментариях. В рассматриваемом нами примере для статистического исследования успеваемости в группе не играет роли, кто конкретно из учеников решил данное число задач из набора $[4;4;3]$. Если бы Маша решила 3 задачи, а «Саша» – четыре, то это никак бы не повлияло на общую оценку успеваемости учеников этой группы школьников (хотя, несомненно, обрадовало бы родителей Саши и расстроило родителей Маши). С этой точки зрения наборы $[4;4;3]$ и $[4;3;4]$ следует считать одинаковыми. Неформально можно сказать, что

Числовой (статистический) набор – это конечная неупорядоченная последовательность чисел.

Чтобы не усложнять обозначения, мы будем обозначать набор значений переменной тем же символом, что и саму переменную (как правило, заглавной буквой латинского алфавита). Таким образом, в рассматриваемом примере можно писать: набор $W=[4;4;3]$.

В математике запись $\{4;3\}$ обычно означает *множество* из элементов 4 и 3, а запись $(4;4;3)$ – *последовательность* из трёх элементов 4, 4 и 3 (напомним, что *по определению* всякая последовательность является *упорядоченной*). По этой причине мы решили употреблять для обозначения *неупорядоченного* набора $[4;4;3]$ квадратные скобки.

Ещё раз обратим внимание на то, что $\{4;3\}=\{3;4\}$ (элементы множества не упорядочены), запись $\{4;4;3\}$ не имеет смысла (в множестве каждый элемент указывается только один раз; с некоторой натяжкой её можно было бы отождествить с записью $\{4;3\}$), $(4;4;3) \neq (4;3;4)$, а $[4;4;3]=[4;3;4]$.

Отметим, что в некоторых статистических исследованиях порядок, в котором расположены анализируемые значения, очень важен. В этих случаях мы будем указывать числа, образующие набор, в круглых скобках. Рассмотрим, например, следующую таблицу, в которой указана средняя температура в некотором регионе за 8 лет:

Год	2003	2004	2005	2006	2007	2008	2009	2010
Температура	21,2°	22,1°	23,2°	22,6°	22,9°	23,6°	24,8°	24,6°

Если мы хотим подсчитать среднюю температуру за 8 лет, то порядок чисел в наборе $[21,2; 22,1; 23,2; 22,6; 22,9; 23,6; 24,8; 24,6]$ не играет никакой роли. Но если мы хотим понять, происходит ли изменение климата, то порядок очень важен (в данном регионе явно видна тенденция роста среднемесячной температуры в июле) и чтобы подчеркнуть, что рассматривается упорядоченный набор, мы будем записывать его в круглых скобках: $(21,2; 22,1; 23,2; 22,6; 22,9; 23,6; 24,8; 24,6)$.

Поскольку в статистике любой числовой набор является набором значений определённой статистической переменной, когда мы говорим о среднем значении, дисперсии, функции распределения и т.д. числового набора, мы фактически говорим о среднем значении, дисперсии, функции распределения и т.д. этой переменной.

2.3. Дискретные и непрерывные переменные. Числовые переменные в статистике делятся на два типа: *дискретные* и *непрерывные*.

Числовая переменная называется дискретной, если её значения ясно отделены друг от друга (от латинского *discretus* – разделённый, прерывистый).

С формальной математической точки зрения дискретные переменные – это переменные, значения которых могут быть пронумерованы (что возможно, в частности, если количество значений конечно).

В рассмотренном выше примере переменная W является дискретной, т.к. может принимать только значения: 0, 1, 2, 3 и т.д.

Как мы отмечали, в статистике обычно не играет роли, к какому объекту генеральной совокупности относится то или иное значение числового набора. В таких ситуациях важно лишь общее количество объектов, для которых измеряемая переменная принимает данное значение. Соответственно в таких ситуациях первый шаг в статистической обработке дискретных переменных заключается в вычислении числа повторений каждого значения. Как правило, подсчитывают и (относительную) частоту повторения каждого значения. Результаты представляют в табличной или графической форме (всевозможные диаграммы, включая столбиковую).

На столбиковой диаграмме, представляющей распределение значений дискретной переменной, столбики отделены друг от друга небольшим промежутком. Другие виды диаграмм (особенно популярны круговые диаграммы) обычно дают только приблизительное визуальное впечатление о характере распределения и чаще всего используются только для украшения статьи, доклада и пр.

*Числовая переменная называется **непрерывной**, если её значением может быть любое действительное число из некоторого промежутка (по крайней мере, теоретически).*

Непрерывными являются все переменные, которые выражают физические характеристики объекта: длину, массу, объём, температуру, время и т.д.

При физических измерениях мы всегда определяем значения длины, массы и т.д. приближённо, с точностью, зависящей от вида используемого прибора, так что результат всегда выражается десятичной дробью с конечным числом знаков после запятой. При подходящем выборе единицы измерения результат будет выражен целым числом. Примерно так же обстоит дело и с денежными суммами – они всегда выражаются целым числом копеек. Поэтому с практической точки зрения непрерывных переменных нет вовсе. Однако теоретически измеряемые величины могут выражаться любым действительным числом.

Важно учесть ещё одно соображение. Дискретные переменные характеризуются тем, что их значения ясно отделены друг от друга. Но разница между 456руб. 89 коп. и 457руб. 23коп. настолько мала, что с практической точки зрения вряд ли разумно считать эти величины различными. Если денежные суммы отмечать точками на числовой оси, то даже при выборе единицы масштаба в 1 руб. разница между соседними точками (величиной 1 коп, т.е. 0,01 единицы масштаба) будет настолько мала, что эти дискретные точки сольются в непрерывную линию. Это соображение показывает, что в ряде случаев формально дискретные переменные иногда удобно считать непрерывными.

Для непрерывных переменных любое конкретное значение малоинтересно. Например, при анализе доходов населения совершенно неважно, сколько человек имели в прошлом году суммарный доход в размере 327851 руб. 37 коп. Действительно интересно, например, знать, сколько человек имели годовой доход меньше, чем 150 тыс. руб., т.е. количество малообеспеченных,

сколько человек имели годовой доход от 300 тыс. руб. до 600 тыс. руб., т.е. количество достаточно обеспеченных, и т.д. Таким образом, для непрерывных переменных разумно обращать внимание лишь на классы значений, попадающих в тот или иной промежуток. Поэтому первый шаг в статистической обработке непрерывных переменных заключается в разбиении диапазона возможных значений на несколько промежутков (их называют интервалами группировки) и вычислении количества значений, попадающих в каждый интервал. Интервалы группировки должны быть *непересекающимися* (нельзя, чтобы какое-то значение можно было бы отнести к двум разным интервалам) и *исчерпывающими* (т.е. каждое теоретически возможное значение должно быть отнесено в какой-то интервал).

По поводу интервалов группировки необходимо сделать важное замечание, связанное с процедурой округления значений непрерывных переменных. Предположим, например, что мы измеряем время T на дорогу до школы для 10 учеников и получили следующие значения (в минутах): 18; 23; 29; 18; 37; 25; 25; 21; 18; 27. Если время округлялось до целых минут отбрасыванием секунд, то значение 18 означает, что точное время лежит в промежутке $18 \leq T < 19$. Если же время округлялось до целых минут по обычному правилу (до ближайшего целого), то значение 18 означало бы, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $17,5 \leq T < 18,5$. Если мы собираем в один класс все значения от 10 до 19 включительно, то при первом способе округления этот класс характеризуется неравенством $10 \leq T < 20$, а при втором – неравенством $9,5 \leq T < 19,5$. В каждом случае длина интервала группировки равна 10, но сами интервалы разные. В частности, в первом случае центр интервала (он важен для приближённого расчёта статистических характеристик по сгруппированным значениям) равен 15, а во втором этим центром будет точка 14,5. Чтобы избежать недоразумений, лучше точно описывать классы группировки двойными неравенствами. В школьных учебниках на процедуру округления числовых данных вообще не обращают внимания, что недопустимо при аккуратном статистическом анализе непрерывных переменных.

Общее визуальное представление о характере распределения значений непрерывной переменной дают *гистограмма* и *функция распределения*. Гистограмме посвящена настоящая статья. По поводу функции распределения (она определена также и для дискретных переменных) см. учебное пособие авторов [7] или статью [8].

3 Определение гистограммы

Пусть $X = [x_1, \dots, x_n]$ – набор (вообще говоря, неупорядоченный) значений некоторой непрерывной переменной X . Выберем некоторый промежуток в

качестве множества возможных значений анализируемой переменной. Этот промежуток, конечно, должен содержать все числа x_i , так что он шире отрезка $[\min(x_1, \dots, x_n); \max(x_1, \dots, x_n)]$. Поскольку для непрерывной переменной отдельные значения малоинтересны, совершенно неважно, входят или нет в этот промежуток его граничные точки. Для определённости мы будем рассматривать промежутки вида $[a; b)$.

Предположим, что промежуток возможных значений $[a; b)$ разбит точками $c_0 = a < c_1 < \dots < c_{k-1} < c_k = b$ на k непересекающихся промежутков (классов или интервалов группировки): $[c_0; c_1), [c_1; c_2), \dots, [c_{k-1}; c_k)$. Обратим внимание на то, что мы не предполагаем равенство длин $\Delta_i = c_i - c_{i-1}$ этих интервалов.

Пусть n_i – количество значений, попавших в i -й класс, $f_i = \frac{n_i}{n}$ –

(относительная) частота попадания в i -й класс, $f_i^* = \frac{n_i}{n \cdot \Delta_i}$ – нормированная

частота попадания в i -й класс. Поскольку сумма чисел n_i равна n , сумма относительных частот равна 1.

Определение 1. Функция $h_n(x)$ действительного аргумента x , которая на промежутке $[a; b)$ возможных значений анализируемой переменной определена формулой

$$h_n(x) = f_i^* \equiv \frac{f_i}{\Delta_i}, \text{ если } c_{i-1} < x < c_i,$$

называется *гистограммой*.

В граничных точках интервалов группировки гистограмму можно определять произвольным способом, например, считать, что $h_n(c_{i-1}) = \frac{f_i}{\Delta_i}$, а

можно вообще не определять. Удобно также считать, что вне промежутке $[a; b)$ возможных значений анализируемой переменной гистограмма тождественно равна 0.

График гистограммы получается, если каждый интервал группировки $[c_0; c_1), [c_1; c_2), \dots, [c_{k-1}; c_k)$ поднять на высоту, равную нормированной частоте $f_i^* \equiv \frac{f_i}{\Delta_i}$ попадания переменной в этот интервал. Поэтому график

гистограммы состоит из k конечных промежутков длиной $\Delta_1, \dots, \Delta_k$ и двух бесконечных лучей: левый луч идёт из точки $c_0 \equiv a$ на высоте 0 (совпадая с осью абсцисс) до $-\infty$, правый луч идёт из точки $c_k \equiv b$ на высоте 0 (совпадая с осью абсцисс) до $+\infty$. Функцию с графиком подобного вида называют *кусочно-постоянной*. В граничных точках интервалов группировки гистограмма, как правило, разрывна. Величина скачка зависит от разницы нормированных частот в соседних интервалах группировки (т.е. скачки

могут быть как положительными, так и отрицательными). В зависимости от того, как определять значение гистограммы в граничных точках, функция $h_n(x)$ может быть непрерывна справа или слева. Если нормированные частоты для соседних интервалов группировки равны (эта ситуация встречается крайне редко), то в соответствующей граничной точке гистограмма будет непрерывна.

Гистограмма $h_n(x)$ зависит не только от количества чисел в наборе, но и от самих этих чисел и классов группировки. Поэтому следовало бы как-нибудь отразить это обстоятельство в обозначениях, например, писать $h_n(x | x_1, \dots, x_n; c_0, c_1, \dots, c_k)$ вместо $h_n(x)$. Однако, это сильно усложняло бы обозначения и поэтому мы будем отмечать зависимость функции $h_n(x)$ от чисел набора и классов группировки только указанием общего количества чисел в наборе. По ходу изложения всегда будет понятно, о каком наборе и о каких классах идёт речь, но если возникает опасность путаницы, лучше использовать это более длинное обозначение.

Во многих учебниках, и не только школьных, рассматривают лишь равные интервалы группировки, а под гистограммой понимают ступенчатую *фигуру*, которая образована прямоугольниками, построенными на интервалах группировки как на основаниях; высота прямоугольника, построенного на промежутке $[c_{i-1}; c_i)$, равна количеству значений n_i из соответствующего

класса или частоте $f_i = \frac{n_i}{n}$. В отличие от этого, мы

- считаем, что гистограмма – это *функция*;
- используем нормированные частоты $f_i^* = \frac{n_i}{n \cdot \Delta_i}$ вместо абсолютных количеств n_i или частот $f_i = \frac{n_i}{n}$;
- допускаем интервалы группировки переменной длины.

Проиллюстрируем теперь наше определение гистограммы на конкретном примере. Рассмотрим набор T из 100 чисел

27, 52, 43, 38, 47, 8, 21, 40, 32, 53, 45, 54, 35, 28, 40, 18, 31, 45, 24, 30, 37, 15, 39, 34, 48, 25, 30, 7, 32, 12, 26, 35, 48, 19, 33, 26, 17, 30, 42, 22, 53, 28, 42, 36, 23, 10, 34, 46, 16, 29, 35, 52, 41, 32, 21, 39, 55, 25, 29, 8, 36, 44, 26, 55, 34, 19, 42, 54, 27, 10, 45, 20, 31, 50, 18, 9, 41, 14, 38, 40, 23, 49, 33, 15, 24, 46, 36, 28, 32, 37, 51, 20, 29, 47, 33, 27, 41, 22, 39, 40.

Этот набор мы взяли из учебника [9], стр. 251; числа набора – это время (в минутах), которое 100 выбранных наугад учеников гипотетической школы тратят на дорогу в школу. К сожалению, мы не знаем, как собиралась анализируемые данные – кто-то уже собрал их за нас (в статистике такие данные называют *вторичными*, в отличие от *первичных* данных, которые мы собираем сами). В частности, мы не знаем использованную процедуру

округления. Для определённости будем считать, что время округлялось до целых минут отбрасыванием секунд.

Вначале для этого конкретного числового набора построим гистограмму в соответствии с обычной процедурой.

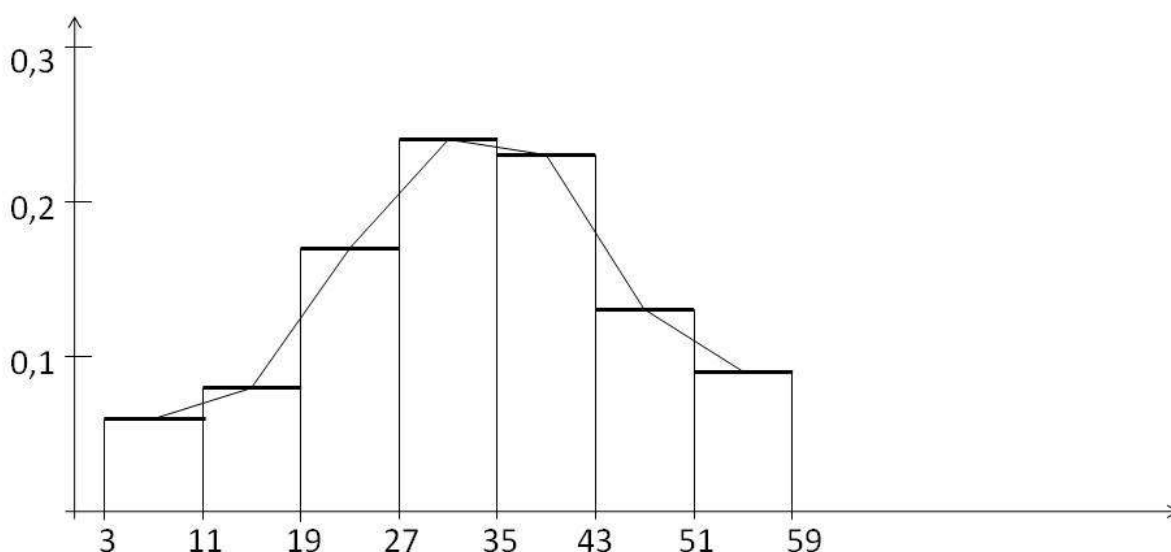
Для построения гистограммы сгруппируем, как это сделано в учебнике [9], эти значения в 7 классов длиной 8 минут каждый: $3 \leq T < 11$, $11 \leq T < 19$, $19 \leq T < 27$, $27 \leq T < 35$, $35 \leq T < 43$, $43 \leq T < 51$, $51 \leq T < 59$.

В учебнике [9] приняты немного другие обозначения для классов, скажем, класс $11 \leq T < 19$ обозначается 11–19, но говорится, что «значение, оказавшееся на границе двух интервалов, будем считать лежащим в правом промежутке». Вместо этой оговорки мы просто точнее описали все классы (сохраняя смысл подхода [9]).

В результате мы получим следующую таблицу (для удобства последующего сравнения двух способов построения гистограммы мы подсчитали и нормированные частоты):

Время (мин)	Количество чисел в классе, n_i	Относительная частота, $f_i = \frac{n_i}{n}$	Нормированная частота, $f_i^* = \frac{n_i}{n\Delta_i}$
3–10	6	0,06	0,00750
11–18	8	0,08	0,01000
19–26	17	0,17	0,02125
27–34	24	0,24	0,03000
35–42	23	0,23	0,02875
43–50	13	0,13	0,01625
51–58	9	0,09	0,01125
Всего	100	1	

Теперь мы можем построить гистограмму. Она изображена на рис.1.

**Рис. 1**

Чтобы нагляднее показать характер гистограммы, на этом же рисунке мы нарисовали ломаную линию, которая соединяет между собой середины горизонтальных отрезков, из которых состоит гистограмма. Эта ломаная в англоязычной литературе по статистике называется *frequency polygon*, что при буквальном переводе на русский язык означает «ломаная линия, образованная частотами».

В школьном учебнике [9] используется термин «полигон относительных частот». В принципе он допустим; в соответствии с Математическим энциклопедическим словарём (Гл.редактор Ю.В.Прохоров. – М., Советская энциклопедия, 1988), «*Полигон* – ломаная линия, составленная из конечного числа прямоугольных отрезков (звеньев). Под полигоном также понимают замкнутую ломаную линию, т.е. многоугольник.» Нельзя, однако, не отметить, что:

(1) в современном русском языке «полигон» обычно означает «участок суши или моря, предназначенный для испытания оружия, боевых средств и техники и для боевой подготовки войск» (Советский энциклопедический словарь, М., Советская энциклопедия, 1980);

(2) в современной математике «полигон» обычно означает «непустое множество с моноидом операторов» (Математическая энциклопедия/ Гл. ред. И.М.Виноградов. – М., Советская энциклопедия, 1984, том 4);

(3) если математики и допускают использование термина «полигон», то лишь как эквивалент терминов «многоугольник» или «ломая линия» *в самом общем их смысле*.

Отметим, кстати, что ломаная частот обычно строится только для гистограмм и не употребляется для столбиковых диаграмм (как рекомендуется в [9]; см., например, британские школьные учебники [4], стр.63, или [5], стр. 16). Иначе говоря, ломаная частот является специфическим графическим представлением распределения *непрерывных* переменных. Как правило, гистограмму и ломаную частот не строят на одном рисунке; мы сделали это лишь для экономии места. Строить ломаную частот можно сразу, без предварительного построения гистограммы. Для этого для каждого класса $[c_{i-1}; c_i)$ нужно вычислить его центр $\frac{c_{i-1} + c_i}{2}$, отметить на координатной плоскости точку $\left(\frac{c_{i-1} + c_i}{2}; f_i\right)$ и соединить все эти точки отрезками.

Следует также отметить, что в теории вероятностей и математической статистике термин «частота» означает отношение числа наблюдений события к общему числу наблюдений. Иначе говоря, вместо относительной частоты следует говорить о частоте. На это обстоятельство обращают внимание и авторы статьи [1].

Теперь применим нашу процедуру построения гистограммы. Прежде всего, в качестве диапазона возможных значений времени на дорогу до школы возьмём более широкий промежуток $0 < T < 60$ (мин). Беглый взгляд на исходные данные показывает, что большинство школьников тратит на дорогу от 20 до 50 минут. Этот промежуток мы разобьём на 6 равных интервалов группировки длиной 5 минут каждый. В каждый из них попадает около 10 чисел набора. В интервалы $0 < T < 20$ и $50 \leq T < 60$ попадает 16 и 10 чисел соответственно. Поэтому мы не будем дробить их на более мелкие промежутки. В результате мы получим следующую таблицу:

Интервал группировки	Длина интервала, Δ_i (мин)	Количество чисел в классе, n_i	Нормированная частота, $f_i^* = \frac{n_i}{n\Delta_i}$
$0 < T < 20$	20	16	0.008
$20 \leq T < 25$	5	10	0.02
$25 \leq T < 30$	5	14	0.028
$30 \leq T < 35$	5	15	0.03
$35 \leq T < 40$	5	13	0.026
$40 \leq T < 45$	5	12	0.024
$45 \leq T < 50$	5	10	0.02
$50 \leq T < 60$	10	10	0.01
Всего	60	100	

Теперь мы можем построить гистограмму. Она изображена на рис.2 (обратим внимание на то, что масштабы по осям не такие, как на рис.1).

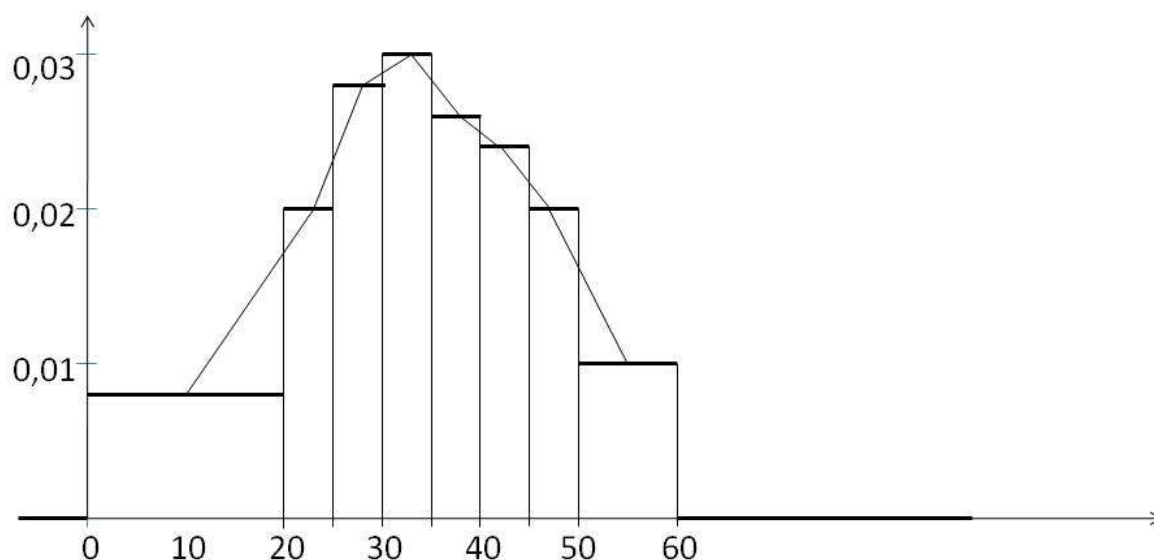


Рис. 2

Как и на рис.1, чтобы нагляднее показать характер распределения, мы изобразили на рис.2 и ломаную частот.

Нетрудно видеть, что в рассматриваемом конкретном случае обе гистограммы дают примерно одно и то же визуальное представление о характере распределения анализируемой переменной. Это связано с большой симметрией набора.

На рис.3 мы изобразили ломаные частот для двух способов группировки в одной системе координат. Непрерывной линией нарисована ломаная нормированных частот для интервалов группировки переменной длины (эта же линия изображена на рис.2). Штрихованной линией нарисована ломаная нормированных частот для интервалов группировки постоянной длины (для удобства сравнения мы использовали нормированные частоты, так что эта линия отличается от линии, изображённой на рис.1, масштабом по оси ординат). Из рис.3 хорошо видно, что ломаные частот практически идентичны.

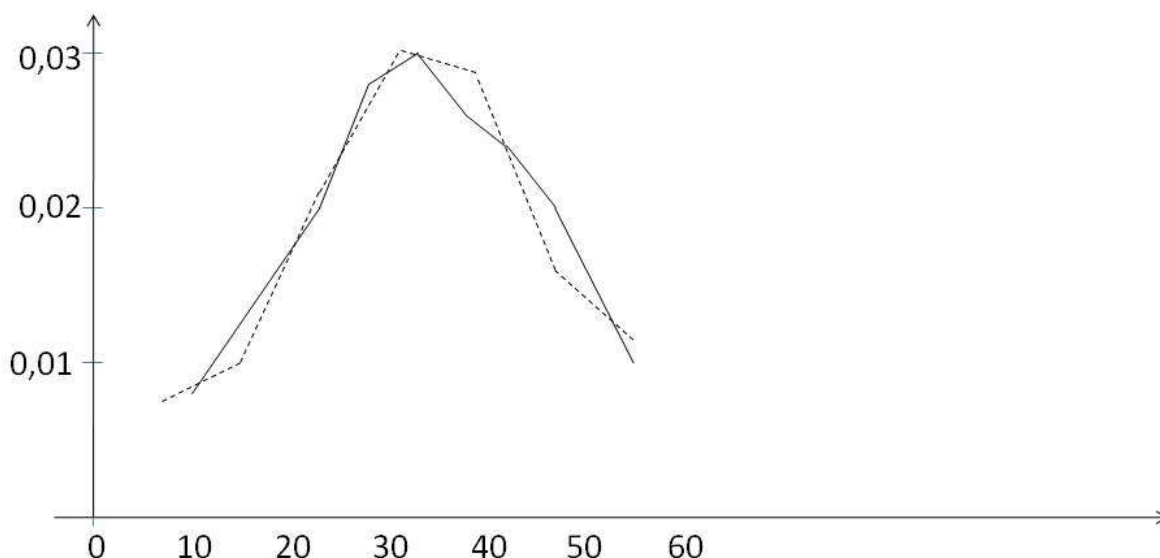


Рис. 3

Для асимметричных распределений с длинными «хвостами» (а на практике часто встречаются именно такие распределения) рассматривать, как это обычно рекомендуется в учебниках по статистике, интервалы группировки равной длины неразумно. Чтобы продемонстрировать это обстоятельство, рассмотрим настоящий, а не «игрушечный» пример (он взят из монографии [10], стр. 64, посвящённой применению статистических методов в страховании).

В 1977 году в США были зафиксированы 40 катастрофических ураганов, каждый из которых привёл к убыткам, превышающим 1,5 миллиона долларов. Величины потерь, округлённые до целого числа миллионов, таковы: 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15, 17, 22, 23, 24, 24, 25, 27, 32, 43. Для удобства последующего анализа значения потерь L упорядочены по возрастанию.

Правильный статистический анализ подобных данных о потерях критически важен для обеспечения финансовой устойчивости страховых компаний. Им занимаются *актуарии* – специалисты по математическим и статистическим расчётам в страховании.

В соответствии с принятой процедурой округления, например, значение 2 означает, что реальные потери L находились в промежутке $1,5 \leq L < 2,5$. Чаще всего в списке встречается число 2, но его нельзя называть модой, т.к. в описательной статистике этот термин употребляется только в случае наборов значений дискретной переменной. В нашей ситуации числу 2 соответствует

целый класс значений (1,5;2,5). Поэтому говорят, что распределение имеет единственный и ярко выраженный *модальный класс* $1,5 \leq L < 2,5$.

Среднее значение потерь равно $M_L = 9,225$, медиана $\mu_L = 5$, среднее квадратичное (стандартное) отклонение $\sigma_L \approx 10,108$. Поэтому коэффициент

асимметрии Пирсона $3 \frac{M_L - \mu_L}{\sigma_L} \equiv 3 \frac{\text{среднее значение} - \text{медиана}}{\text{стандартное отклонение}}$ равен

примерно 1,25. Хорошо известно, что максимальное значение этого коэффициента равно 3 (доказательство см., например, в [11]). Относительно большое значение коэффициента Пирсона для нашего набора свидетельствует о значительной положительной асимметрии набора (это, впрочем, ясно видно и из беглого взгляда на приведённый ряд значений потерь).

Выберем в качестве диапазона возможных потерь промежутков (1,5;49,5) длиной 48 и разобьём его на 4 промежутка одинаковой длины 12: $1,5 < L < 13,5$; $13,5 \leq L < 25,5$; $25,5 \leq L < 37,5$; $37,5 \leq L < 49,5$ с центрами в точках 7,5; 19,5; 31,5; 43,5 соответственно. Количество чисел набора, попадающих в эти классы, равно 30, 7, 2, 1 соответственно. Относительные частоты равны 0,75; 0,175; 0,05; 0,025 соответственно. Нормированные частоты равны 0,0625; 0,0146; 0,0042; 0,0021 соответственно. Построенная по этим данным гистограмма приведена на рис.4. Для удобства сравнения гистограмм мы и в случае равных по длине интервалов группировки строим гистограмму по нормированным частотам.

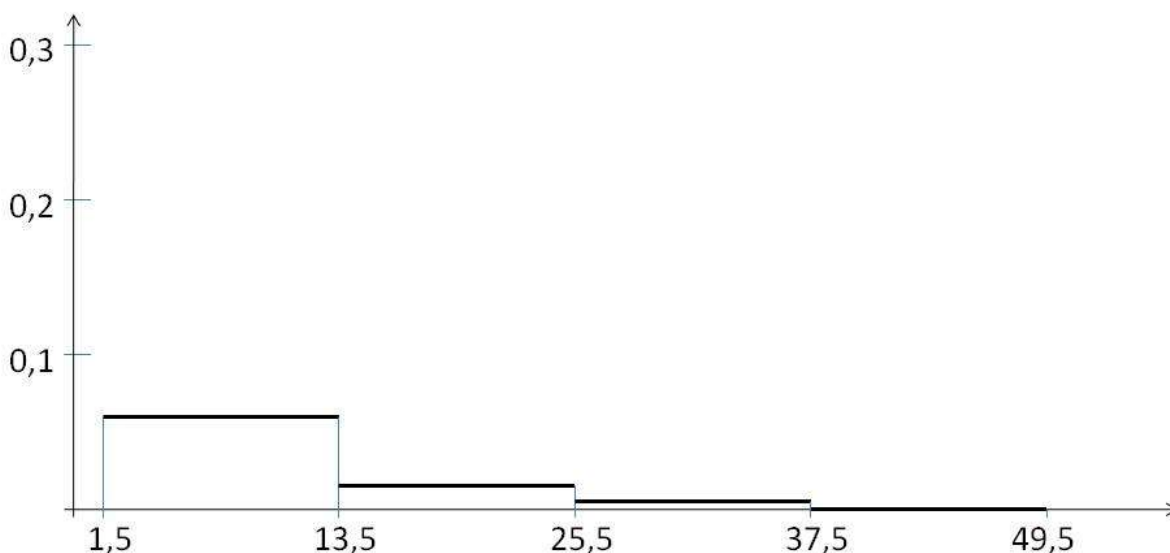


Рис. 4

Теперь разобьём, как сделано в [10], промежуток $(1,5;49,5)$ на 4 промежутка: $1,5 < L < 2,5$; $2,5 \leq L < 6,5$; $6,5 \leq L < 29,5$; $29,5 \leq L < 49,5$ длиной 1, 4, 23, 20 соответственно. Их центры - точки 2; 4,5; 18; 39,5 соответственно. Количество чисел набора, попадающих в эти классы, равно 12, 15, 11, 2 соответственно, так что нормированные частоты равны 0,3000; 0,0938; 0,0120; 0,0025 соответственно. Построенная по этим данным гистограмма приведена на рис.5.

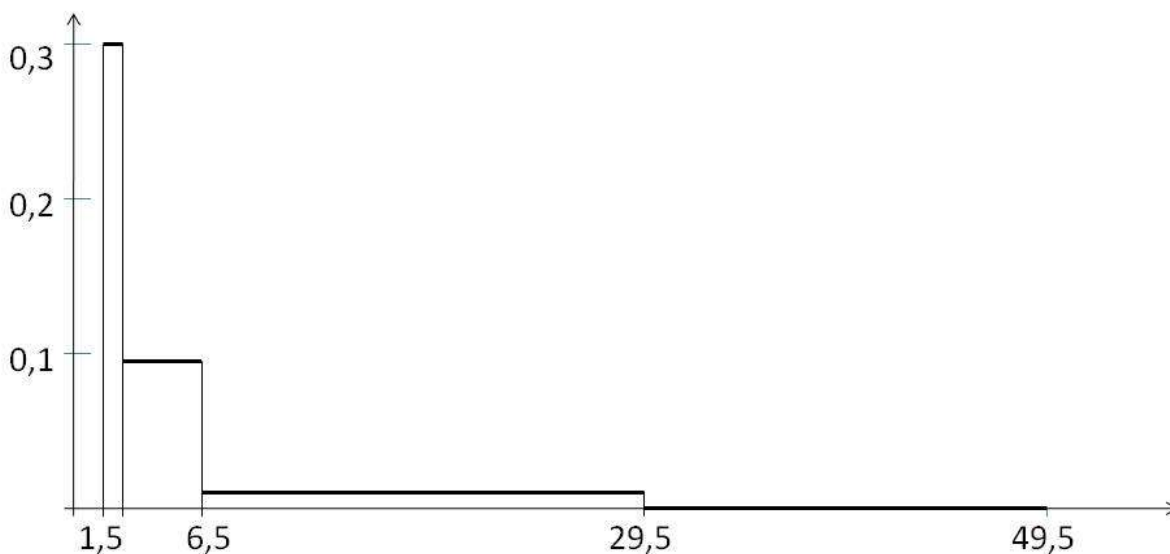


Рис. 5

Как видно из рис.4 и рис.5, гистограммы совершенно различны (на этих рисунках масштаб один и тот же). Особенно хорошо отличие видно на рис.6, где мы изобразили ломаные частот для первого способа группировки (с интервалами группировки равной длины; штрихованная линия) и второго способа (с интервалами группировки разной длины; непрерывная линия). На этом рисунке мы соединили точки, соответствующие крайним интервалам группировки с точками 1,5 и 49,5 на оси абсцисс, чтобы подчеркнуть, что гистограмма равна 0 вне промежутка $(1,5;49,5)$.

Если подсчитать среднее значение набора по сгруппированным данным, заменяя все значения из класса его центром, то первый способ группировки даст значение 11,7, а второй – значение 9,2125. Напомним, что точное среднее значение потерь равно $M_L = 9,225$ (мы игнорируем тот факт, что на самом деле мы работаем с округлёнными значениями потерь). Таким образом, даже это простое соображение показывает, что второй способ группировки лучше. Более тонкий и сложный статистический анализ (детали можно найти в [10], стр. 94-95, 116-118) показывает, что только подход, основанный на интервалах группировки переменной длины, адекватно моделирует реальную ситуацию.

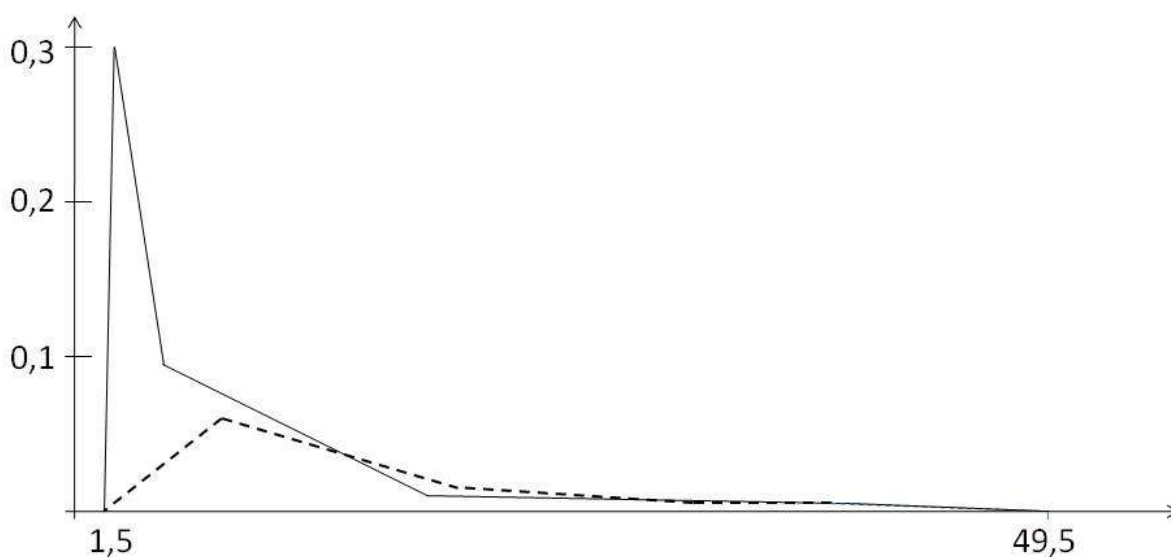


Рис. 6

4 Свойства гистограммы

В этом разделе мы установим несколько важных общих свойств гистограммы $h_n(x)$.

Как и в начале предыдущего раздела, рассмотрим набор $X = [x_1, \dots, x_n]$ значений некоторой непрерывной переменной X и разобьём промежуток возможных значений $[a; b)$ точками $c_0 = a < c_1 < \dots < c_{k-1} < c_k = b$ на k классов: $[c_0; c_1), [c_1; c_2), \dots, [c_{k-1}; c_k)$. Пусть $\Delta_i = c_i - c_{i-1}$ — длина i -го промежутка группировки, n_i — количество значений, попавших в i -й класс,

$f_i = \frac{n_i}{n}$ — (относительная) частота попадания в i -й класс, $f_i^* = \frac{n_i}{n \cdot \Delta_i}$ —

нормированная частота попадания в i -й класс, а гистограмма (функция действительного аргумента x) определена формулой

$$h_n(x) = \begin{cases} f_i^*, & \text{если } c_{i-1} \leq x < c_i, \quad i = 1, \dots, k; \\ 0, & \text{если } x < a \equiv c_0 \text{ или } x \geq b \equiv c_k. \end{cases}$$

Прежде всего отметим, что непосредственно из определения гистограммы следует

Свойство 1. При всех значениях аргумента x верно неравенство

$$h_n(x) \geq 0. \quad (1)$$

Свойство 2. Площадь фигуры, ограниченной снизу осью абсцисс, а сверху гистограммой, равна 1.

Доказательство. Поскольку вне промежутка $[a;b)$ возможных значений переменной X гистограмма тождественно равна 0, фигура, о которой идёт речь, может быть разбита на k прямоугольников; i -й прямоугольник, $i=1, \dots, k$, построен на i -м интервале группировки $[c_{i-1}; c_i)$ как на основании (так что длина основания равна Δ_i), а его высота равна нормированной частоте $f_i^* = \frac{n_i}{n \cdot \Delta_i}$. Поэтому площадь S этой фигуры равна сумме площадей всех k прямоугольников:

$$S = \Delta_1 \cdot f_1^* + \dots + \Delta_k \cdot f_k^* = \frac{n_1}{n} + \dots + \frac{n_k}{n} = \frac{n_1 + \dots + n_k}{n}.$$

Но $n_1 + \dots + n_k$ – это общее количество чисел в наборе, так что $S=1$, как мы и утверждаем.

Замечание. Поскольку площадь фигуры, ограниченной снизу осью абсцисс, а сверху графиком некоторой функции, равна интегралу от этой функции, доказанное свойство можно записать в эквивалентной форме как равенство:

$$\int_{-\infty}^{+\infty} h_n(x) dx = 1. \quad (2)$$

Так как гистограмма тождественно равна 0 вне промежутка $[a;b)$ возможных значений переменной X , пределы интегрирования в интеграле в (2) можно изменить с $-\infty$ и $+\infty$ на a до b соответственно.

Дословное повторение рассуждений, проведённых при доказательстве Свойства 2, позволяет доказать, что площадь фигуры, ограниченной снизу осью абсцисс, сверху гистограммой, а справа вертикальной прямой, проходящей через точку x на оси абсцисс,

- равна 0, если $x \leq a$;
- даётся формулой: $f_1 + \dots + f_{i-1} + f_i \frac{x - c_{i-1}}{c_i - c_{i-1}}$, если $x \in [c_{i-1}; c_i)$;
- равна 1, если $x > b$.

Но по этим же формулам вычисляется непрерывное приближение $F_n^*(x)$ функции распределения набора по сгруппированным значениям (по поводу функции распределения см. учебное пособие авторов [7] или статью [8]). Поэтому верно равенство

$$\int_{-\infty}^x h_n(u) du = F_n^*(x). \quad (3)$$

Теперь мы установим важную формулу, позволяющую подсчитать с помощью гистограммы среднее значение набора. Напомним, что среднее значение M_X набора $X = [x_1, \dots, x_n]$ определяется формулой:

$M_X = \frac{x_1 + \dots + x_n}{n}$. Если числа x_i сгруппированы в классы и $M_X^{(i)}$ – среднее

арифметическое n_i чисел из i -го интервала группировки $[c_{i-1}; c_i)$, то эту формулу можно переписать в виде:

$$M_X = \frac{M_X^{(1)} \cdot n_1 + \dots + M_X^{(k)} \cdot n_k}{n} \equiv M_X^{(1)} \cdot f_1 + \dots + M_X^{(k)} \cdot f_k.$$

Если предположить, что $M_X^{(i)} \approx \frac{c_{i-1} + c_i}{2}$, т.е. что среднее арифметическое чисел из каждого интервала группировки расположено почти посередине этого интервала, то мы получим приближённую формулу:

$$M_X \approx \frac{c_0 + c_1}{2} \cdot f_1 + \dots + \frac{c_{k-1} + c_k}{2} \cdot f_k. \quad (4)$$

Выражение в правой части этого равенства мы назовём средним значением набора, подсчитанным по сгруппированным данным, и обозначим M_X^* . С практической точки зрения можно считать, что $M_X = M_X^*$, т.е. формула (4) является точной.

Свойство 3. Среднее значение набора, подсчитанное по сгруппированным данным, даётся формулой

$$M_X^* = \int_{-\infty}^{+\infty} x h_n(x) dx. \quad (5)$$

Иначе говоря, величина M_X^* равна алгебраической площади (т.е. площади с учётом того, что площадь той части фигуры, которая находится под осью абсцисс, учитывается со знаком «–») фигуры, ограниченной осью абсцисс и графиком функции $x h_n(x)$.

Доказательство. Рассмотрим функцию $x h_n(x)$. Она тождественно равна 0 вне промежутка $[a; b)$ возможных значений переменной X , а на i -м интервале группировки $[c_{i-1}; c_i)$ даётся формулой: $x h_n(x) = f_i^* \cdot x$, так что соответствующий кусок графика функции $x h_n(x)$ является отрезком прямой с угловым коэффициентом f_i^* .

Фигура, о которой идёт речь в формулировке Свойства 3, состоит из k фигур: $\Phi_1, \dots, \Phi_k$. Фигура Φ_i ограничена промежутком $[c_{i-1}; c_i)$ на оси абсцисс, отрезком прямой $y = f_i^* \cdot x$ и двумя вертикальными прямыми, проходящими через точки c_{i-1} , c_i на оси абсцисс. Мы докажем, что

алгебраическая площадь этой фигура равна $\frac{c_{i-1} + c_i}{2} \cdot f_i$, т.е. i -му слагаемому в правой части (4). Отсюда, очевидно, вытекает требуемое утверждение.

Если $f_i^* = 0$, то фигура Φ_i вырождается в отрезок, площадь которого мы считаем равной 0, так что формула $\frac{c_{i-1} + c_i}{2} \cdot f_i$ для её площади верна.

Поэтому ниже будем предполагать, что $f_i^* > 0$.

Возможны 3 случая: (1) $c_{i-1} \geq 0$, (2) $c_i \leq 0$, (3) $c_{i-1} < 0 < c_i$.

Если $c_{i-1} > 0$, т.е. промежуток $[c_{i-1}; c_i)$ расположен на положительной полуоси, то фигура Φ_i лежит выше оси абсцисс и является трапецией с основаниями $c_{i-1} \cdot f_i^*$, $c_i \cdot f_i^*$ и высотой Δ_i (см. рис. 7).

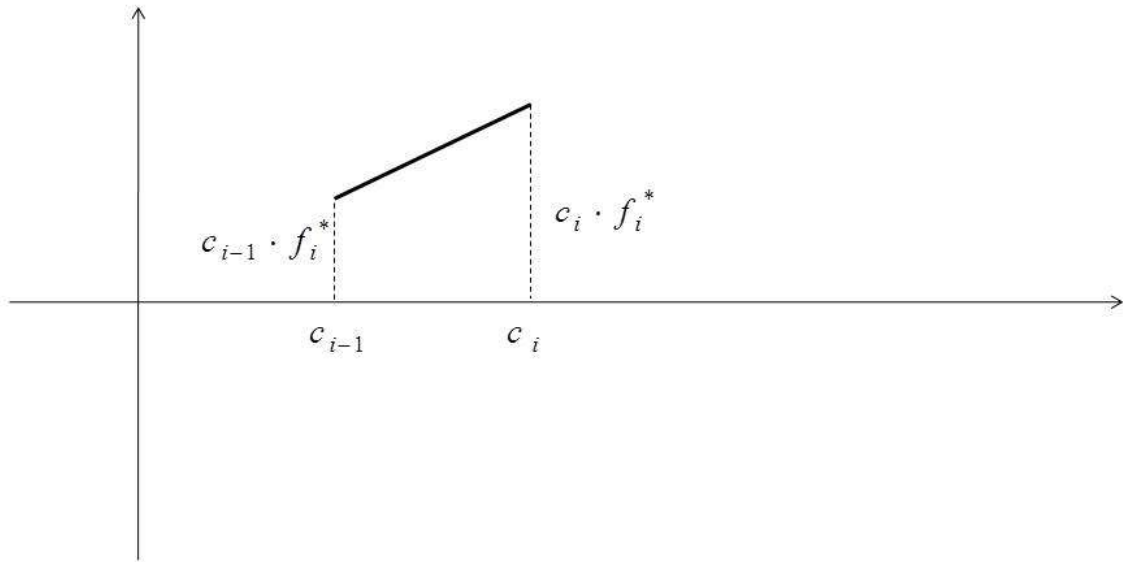


Рис. 7

Поэтому её площадь равна $\frac{c_{i-1} \cdot f_i^* + c_i \cdot f_i^*}{2} \cdot \Delta_i = \frac{c_{i-1} + c_i}{2} \cdot f_i^* \cdot \Delta_i = \frac{c_{i-1} + c_i}{2} \cdot f_i$,

т.е. i -му слагаемому в правой части (4).

Если $c_{i-1} = 0$, то трапеция Φ_i вырождается в прямоугольный треугольник с основанием $c_i \cdot f_i^*$ и высотой Δ_i . Его площадь равна $\frac{1}{2} c_i \cdot f_i^* \cdot \Delta_i = \frac{c_{i-1} + c_i}{2} \cdot f_i$ (с учётом того, что $c_{i-1} = 0$), т.е. опять i -му слагаемому в в правой части (4).

Подобным же образом можно рассмотреть и случай, когда $c_i \leq 0$ или $c_{i-1} < 0 < c_i$.

Доказанные соотношения (1), (2), (3), (5) означают, что гистограмма, определённая в соответствии с нашим определением, обладает всеми свойствами, характерными для плотности случайной величины. Поэтому описанный подход к понятию гистограммы может служить хорошей основой для развития понятной школьникам теории непрерывных случайных величин, включая теорию нормального распределения. Хорошо известно, насколько тяжело изложить эту теорию достаточно строго и в то же время понятно даже студентам математических специальностей.

4 Задачи

Задача 1. В 1982 г. Американский физик Альберт Майкельсон провёл серию опытов по измерению скорости света. Результаты 25 из них представлены в виде гистограммы, которая изображена на рис.8.

Найдите модальный класс. Сколько опытов дали результат в этом диапазоне? Сколько опытов дали результат от 299,80 (тыс.км/сек) до 299,85 (тыс.км/сек)?

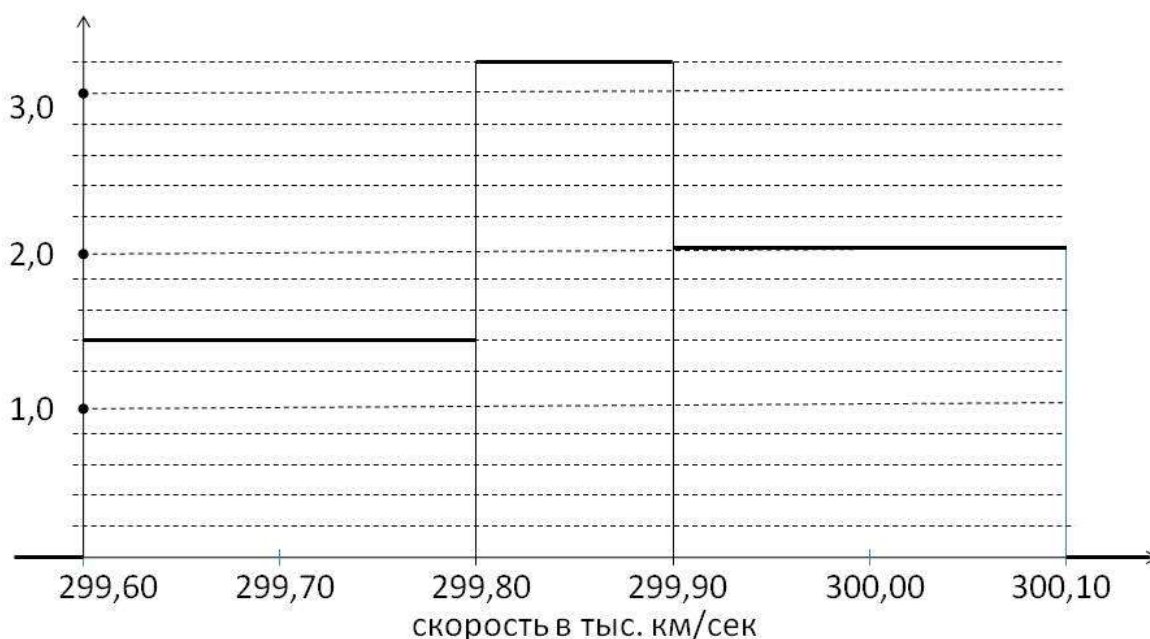


Рис. 8

Решение. Модальный класс определяется интервалом группировки $299,8 \leq v < 299,9$. Соответствующее значение гистограммы равно 3,2, а длина этого интервала равна 0,1. Поэтому количество результатов, попавших в этот класс, равно $25 \cdot 3,2 \cdot 0,1 = 8$.

Если считать, что внутри этого интервала результаты распределены равномерно, то число опытов, давших для скорости света значение от 299,80 до 299,85 равно 4.

Задача 2. Крупная обувная фабрика заказала социологическое исследование, в ходе которого у 200 мужчин была измерена длина стопы. Результаты приведены в следующей таблице.

длина стопы (см)	число случаев	длина стопы (см)	число случаев
$25,0 \leq L < 26,5$	15	$28,0 \leq L < 28,5$	60
$26,5 \leq L < 27,5$	28	$28,5 \leq L < 29,0$	45
$27,5 \leq L < 28,0$	40	$29,0 \leq L < 30,5$	12

Постройте гистограмму по этим данным. Как обувная фабрика может использовать результаты исследования?

Решение. Гистограмма, построенная по этим данным, изображена на рис.9. Фабрика может использовать результаты исследования для того, чтобы определить, какую долю в общем объеме производства должна занимать обувь каждого размера. Разумеется это можно делать только если распределение размера обуви у 200 мужчин, участвовавших в исследовании, примерно такое же, как и у мужчин тех регионов, где фабрика предполагает продавать обувь.

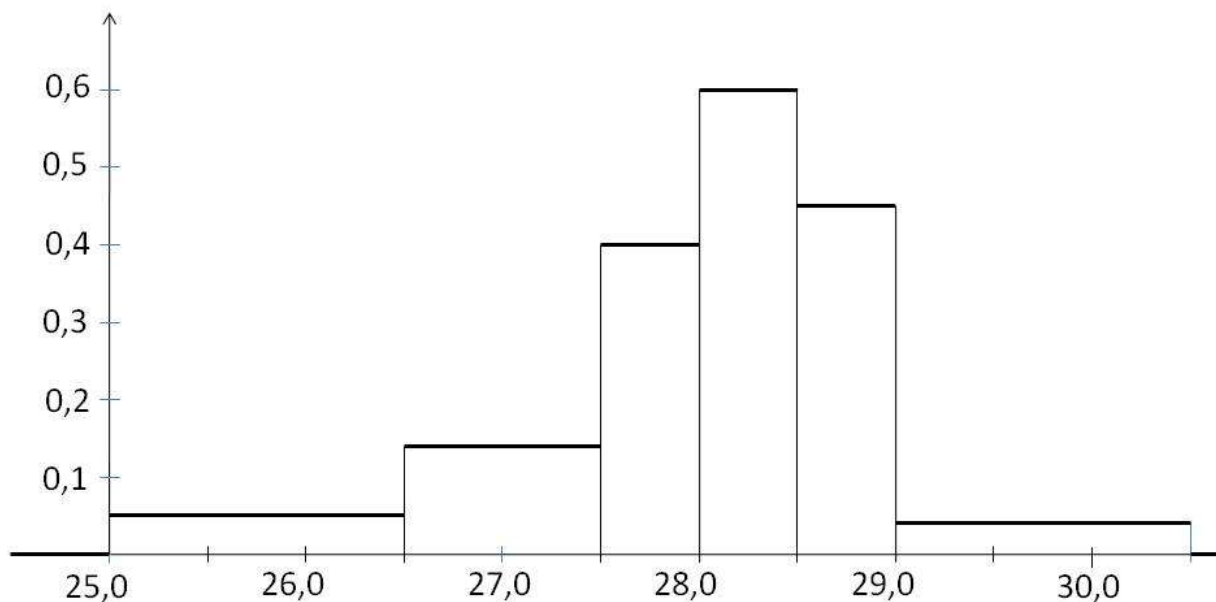


Рис. 9

Задача 3. Школьник рисует гистограмму по сгруппированным значениям непрерывной переменной X . В класс $1,7 \leq x < 2$ попадает 15 значений, а

значение гистограммы для этого класса равно 1,0. В класс $2,0 \leq x < 2,1$ попадает 10 значений. Найдите значение гистограммы для этого класса.

Решение. Используя данные относительно первого класса и формулу $f_i^* = \frac{n_i}{n \cdot \Delta_i}$, мы найдём общее число значений: $n = \frac{n_i}{f_i^* \cdot \Delta_i} = \frac{15}{1,0 \cdot 0,3} = 50$. Теперь из той же формулы можно найти значение гистограммы для второго класса: $f_i^* = \frac{n_i}{n \cdot \Delta_i} = \frac{10}{50 \cdot 0,1} = 2$.

Задача 4. Школьник рисует гистограмму по сгруппированным значениям непрерывной переменной U . Значение гистограммы равно 0,8 для класса $3,3 \leq u < 3,5$ и 0,3 для класса $3,5 \leq u < 3,8$. Затем он решил объединить эти два класса в один. Найдите значение гистограммы для этого класса.

Решение. Используя индекс 1 для величин, относящихся к первому промежутку группировки, и 2 для величин, относящихся ко второму, для искомой величины имеем:

$$f_i^* = \frac{n_1 + n_2}{n \cdot (\Delta_1 + \Delta_2)} = \frac{f_1^* \cdot n \cdot \Delta_1 + f_2^* \cdot n \cdot \Delta_2}{n \cdot (\Delta_1 + \Delta_2)} = \frac{f_1^* \cdot \Delta_1 + f_2^* \cdot \Delta_2}{\Delta_1 + \Delta_2} = 0,5.$$

Литература

1. Е.А.Бунимович, В.А.Булычев и др. О теории вероятностей и статистике в школьном курсе. *Математика в школе*, 2009, №7, стр.3-13.
2. GCSE Specification. Mathematics (Linear) B4365 (for certification 2012 onwards). AQA, 2010.
3. GCSE Specification. Statistics (for certification 2011 onwards). AQA, 2008.
4. A.Ballard, S.Gill, et al. GCSE Statistics. Complete Revision and Practice. Coordination Group Publications Ltd, 2010.
5. S. Dobbs, J. Miller. Statistics 1. Cambridge University Press, 2009.
6. Edexcel GCSE Statistics. Paper 1H. Higher Tier. Wednesday 18 June 2008 – Morning.
7. Г.И.Фалин, А.И.Фалин. Избранные главы описательной статистики: Учебное пособие. – М.: МАКС Пресс, 2011.
8. Г.И.Фалин, А.И.Фалин. Функция распределения числового набора. *Математика в школе*, 2012 (представлено к публикации).
9. Математика: алгебра. Функции. Анализ данных. Учебник для 9 кл. общеобразоват. Учреждений./Г.В.Дорофеев, С.В.Суворов, Е.А.Бунимович и др; под ред. Г.В.Дорофеева. – М.: Просвещение, 2005.
10. R.V.Hogg, S.A.Klugman. Loss Distributions. John Wiley & Sons, 1984.
11. Г.И.Фалин, А.И.Фалин. Одно интересное неравенство для статистических характеристик числового набора. *Математика*, 2012, №1, стр.54-55.