Часть I. Описательная статистика

Глава 1. Особенности применения статистических методов в психолого-педагогических и социальных исследованиях

1.1. Формализация эмпирического базиса исследования. Прежде чем применять математические и статистические методы к изучению тех или иных явлений общественной жизни, необходимо аккуратно описать и непротиворечиво объяснить исследуемую систему фактов, вписав их в рамки некоторой, достаточно формализованной теории.

Многие исследователи, работающие в области психолого-педагогических и социальных наук, имеют смутные представления о сущности математики и используемых в ней методов. По их мнению, математика – это собрание формул, таблиц, графиков, теорем, а её применение сводится к вычислениям, построению кривых, выведению количественных зависимостей и т.п.

Конечно, всё это присутствует в большинстве исследований. Однако не в этом суть математического метода. Более того, если применение математики сводится лишь к бездумному оперированию символами и числами, оно может стать источником грубейших фактических и теоретических ошибок.

Хотя в своей поисковой, творческой деятельности математики, как и исследователи в любой другой области, широко пользуются интуитивными соображениями и идеями, их рассуждениям присущи прежде всего чёткость и однозначность определений, выделение исходных утверждений, строгая дедукция. Именно эти качества имеют в виду, говоря о математизации психолого-педагогических и социальных исследований.

К сожалению, многие исследования в области психолого-педагогических и социальных наук не отвечают этим требованиям. Трактовка одних и тех же понятий в различных научных школах различна, определения расплывчаты и неоднозначны, в основном описательны. Широко используются синонимические определения. Часто одним и тем же термином обозначают разные явления. Редко выделяются исходные предположения, слабо используются средства логического вывода. Много ссылок на результаты недостаточно проверенных экспериментов, цитаты.

Всё это существенно снижает результативность статистических методов. Поэтому, желая применить в своих исследованиях статистические и

вообще математические методы, необходимо, прежде всего, придать соответствующей теории достаточно корректную, логически выдержанную и увязанную с эмпирическими фактами форму. В этом отношении характерно высказывание известного специалиста в области использования статистических методов профессора Ю.И. Алимова: «Научный подход, прежде всего, предполагает создание в любой прикладной теории интуитивно убедительного эмпирического базиса. Сложность, строгость и стоимость математического аппарата – так сказать, математической надстройки – должны сообразовываться с надежностью базиса. Это издавна применяемое прагматическое правило можно назвать принципом равнопрочности всех элементов прикладного исследования».

Привлекая к исследованию статистические методы, надо иметь в виду, что статистика – это средство анализа чисел как таковых, а не как истинных значений некоторого признака. Всякий статистический метод можно применить к любой совокупности чисел, и неизвестен метод, который был бы неэффективным, потому что используемые в нем числа являются «неподходящими». Статистические методы ничего не добавляют и ничего не отнимают от значимости чисел, к которым они применяются.

В конечном счёте, всё дело в той содержательной интерпретации, которую исследователь может придать результатам вычислений. Таким образом, «математика может избавить нас от мучительной необходимости размышлять, но мы должны платить за эту привилегию, испытывая муки раздумий как до того, как математика вступит в действие, так и после».

1.2. Вероятностный характер закономерностей психологии, педагогики и социологии. Большинство утверждений, с которыми учащихся знакомят в школе, носят жесткий, детерминированный характер, например, утверждение, что если треугольник АВС прямоугольный, то квадрат гипотенузы равен сумме квадратов катетов, или, что если масса постоянна, то ускорение пропорционально действующей силе. В рамках аксиоматических теорий это естественно, однако за их пределами это не так. Например, мнение, что сформулированные выше утверждения, или утверждение о том, что сумма углов треугольника равна двум прямым углам, выполняются в окружающем нас мире, является только достаточно вероятным.

Практически все выводы экспериментальных наук, в том числе психологии, педагогики и социологии, носят лишь вероятностный характер.

По своей природе ни один вывод в этих науках не является абсолютно надежным, а связь явлений жестко детерминированной. Поэтому каж-

дый такой вывод, как правило, формулируется достаточно осторожно, со ссылкой на его вероятностный характер. Например:

1. В большинстве случаев наглядность содействует усвоению материала.

2. Мягкое воспитание, как правило, портит ребенка.

3. Девиантное поведение ребёнка чаще всего является следствием плохого семейного воспитания.

4. Если правительство выплатит долги по зарплате, то почти наверняка возрастет инфляция.

5. Маловероятно, чтобы экономические трудности в стране могли сплотить народ.

Обыденная речь, с помощью которой люди относят выводы к «маловероятным» или «почти достоверным», неадекватно служит науке. Эти субъективные оценки варьируют от лица к лицу в зависимости от слов, выбранных для определения правдоподобия, и от их смысла. Последнее обстоятельство создаёт серьёзные трудности при сопоставлении результатов исследований, проведённых разными учёными. В математике уже давно разработаны приёмы, позволяющие при изучении массовых явлений характеризовать числом вероятность осуществления прогнозируемого события. Эти числовые характеристики (числа) также называют *вероятностями*.

Ниже, для тех, кто не изучал теорию вероятностей, расскажем очень кратко и упрощённо, как в простейших случаях вероятность события характеризуется числом.

Рассмотрим урну (ящик), в которой лежат 40 белых и 60 красных шаров. Все шары неразличимы на ощупь. Экспериментатор, тщательно перемешав их, вынимает один шар. Обозначим буквой Б событие, заключающееся в том, что вынутый шар белый, а буквой К – событие, заключающееся в том, что вынутый шар – красный.

Если опыт с извлечением шара повторять многократно, каждый раз возвращая вынутый шар в урну и тщательно перемешивая её содержимое, то естественно ожидать, что красные шары будут встречаться чаще, чем белые, и это подтверждается экспериментом. В этом смысле вероятность события Б меньше вероятности события К. Более того, можно предложить следующий способ числовой оценки вероятности вынуть белый шар. Когда экспериментатор вынимает шар, у него сто возможностей (исходов). И все они равновозможны, ни один шар не имеет преимуществ перед другими. Из них благоприятных (вынут белый шар) всего сорок. Под *вероятно-*

стью события Б (обозначают Р(Б)) понимают отношение числа благоприятных исходов, в нашем случае их 40, к общему числу возможных исходов, т.е. к 100.

Таким образом, $Р(Б) = \frac{40}{100} = 0{,}4$. Аналогично $Р(К) = \frac{60}{100} = 0{,}6$.

В общем случае, если все исходы в эксперименте равновозможны, то вероятность события Н характеризуют числом Р(Н), равным отношению числа исходов, благоприятствующих событию Н, к общему числу возможных исходов, т.е.

$$Р(Н) = \frac{\text{число исходов, благоприятствующих Н}}{\text{общее число равновозможных исходов}}.$$

Это так называемый классический подход к определению вероятности события. Вместе с тем существует и другой, более общий подход к этому понятию, который принято называть статистическим.

Если рассмотренную выше последовательность экспериментов продолжить, фиксируя каждое появление белого шара и вычисляя относительную частоту, то последняя будет приближаться к числу 0,4, т.е. к вероятности появления события Б.

Указанное обстоятельство проверялось многократно и каждый раз, если в урне было m белых и n красных шаров, то в больших последовательностях изымаемых шаров относительная частота появления белого шара оказывалась близкой к числу $\frac{m}{m+n}$, а частота появления красного шара $P(K)$ – близкой к числу $\frac{n}{m+n}$. Эту закономерность можно расценивать как экспериментально доказанный научный факт.

Рассмотрим теперь обратную задачу. Пусть известно, что в урне сто шаров, часть из которых белые, а часть – красные. Надо узнать, сколько тех и других. Для этого разрешается вынимать из урны по одному шару, фиксировать его цвет и возвращать обратно, каждый раз тщательно перемешивая содержимое урны.

Возникает вопрос, как, зная число актов изъятия шара и соответствующую частоту появления белого шара, найти число белых шаров в урне и определить надежность полученного результата.

С подобными задачами то и дело сталкивается экспериментатор, только вместо шаров он имеет дело с людьми, а вместо цвета – с теми или иными личностными качествами.

Пусть, например, исследователь, изучающий уровень религиозности выпускников средних учебных заведений, решил выяснить, какая их часть посещает церковные службы. Для этого из общего их числа (12000 человек) случайным образом было отобрано 1000 респондентов, которым был задан соответствующий вопрос. Предположим, что сорок из них ответили положительно.

Формальный расчет показывает, что среди всех выпускников церковную службу, скорее всего, посещает $12000 \cdot \frac{40}{1000}$, т.е. 480 человек. Но какова надежность этого результата? Может быть, было бы надежнее считать, что их число заключено между 430 и 530? Как при таком расширении границ будет меняться надежность оценки и что вообще означает эта надежность, в каких единицах выражается?

Для того чтобы ответить на этот и многие другие вопросы, надо уточнить многие понятия и узнать много новых фактов из теории вероятностей. Мы в дальнейшем будем обращаться к ним по мере необходимости.

1.3. Особенности измерения психолого-педагогических и социальных явлений. В статистике *измерением* называется процедура, с помощью которой объекты исследования, рассматриваемые как носители определенных отношений между ними, отображаются в некоторую математическую систему с соответствующими отношениями между элементами этой системы. В общественных науках чаще всего используются числовые математические модели.

Как известно, существуют прямые и косвенные процедуры измерения. Прямые – когда откладывается единица измерения и подсчитывается, сколько раз она содержится в объекте. Косвенные – когда измеряют вспомогательные величины, а потом, пользуясь связями, вычисляют требуемую величину.

В психологии, педагогике и социологии тоже применяются как прямые, так и косвенные способы. Прямые – когда измеряют рост, объём грудной клетки, остроту зрения, время выполнения задания, численность населения, объем производимой продукции и т.д. Однако более распространены косвенные методы измерения, например, измерение уровня знаний, умений и навыков, которые проявляются в ответах учеников, качестве и способах решения ими определенных классов задач, успешности выполнения определенных видов деятельности.

Чем сложнее явление, тем труднее провести его измерение. Сравнительно легко измерить физическую силу руки, труднее измерить механическую память, еще труднее – выносливость, настойчивость, волю. С другой стороны, чем глубже проникает психология, педагогика и социология в структуру связей между психической и практической деятельностью человека, тем больше появляется возможностей для оценки внутренних факторов на основе внешней наблюдаемой деятельности.

Приступая к проблеме количественной оценки явления, надо, выделив его характеристические признаки, построить соответствующую теорию. Эталоном для количественной оценки свойств человека, таких, как тонкость восприятия, концентрация внимания, четкость логического мышления, уровень знания и так далее, обычно используется сам человек. Важным средством квантификации служат тесты.

Тест – задание стандартной формы, выполнение которого должно выявить уровень определенных свойств личности, характера интересов, эмоциональных реакций, эстетических и других ориентаций.

Обычно ответы оцениваются в баллах. Сумма баллов характеризует развитость определенных черт личности или социального явления. Диагностическая, прогностическая ценность теста зависит от его научной обоснованности, от того, как именно создавался данный тест. Чем более узким, специфическим является измеряемое качество, тем более оправдано применение метода тестирования, и наоборот.

С точки зрения объекта измерения можно выделить общие тесты интеллекта, дифференциальные тесты способностей, тесты достижений, тесты черт личности и психотехнические тесты.

Тесты интеллекта основаны на предположении, что психическое развитие человека может быть охарактеризовано неким интегральным фактором, который обозначается термином «интеллект» Этот фактор определяет уровень всех разнообразных психических способностей людей и степень успешности решения ими самых разнообразных психологических и практических задач. Измерив относительную величину этого фактора, мы получим характеристику интеллектуального уровня испытуемого. К числу таких тестов относятся: тесты общего интеллекта Лорджа–Торндайка, Оттиса, Бине (в Станфордской переработке), Термена–Мак Намары, Вехлера, Протеуса и SRA (Ассоциация научных исследований); тесты на коэффициент интеллектуальности (JQ), в том числе тесты Айзенка. Заметим, всеми этими тестами нельзя мерить умственное развитие.

Тесты способностей многообразны. Есть тесты, измеряющие способность к концентрации внимания, тесты на определение смысла, на прочность запоминания, на способность к словесному мышлению, арифметическому мышлению, абстрактному мышлению, анализу пространственных отношений, восприятию формы и т.п. Известны тесты способностей, разработанные Беннетом, Симорой и Весманом, тесты первичных духовных способностей Тёрстона.

Тесты достижений выясняют объем и качество усвоенных учащимися знаний, умений и навыков по определенному учебному предмету.

Тесты черт личности служат для количественной оценки рангового уровня развития тех или иных черт личности испытуемого (или одной черты):

1) тесты Пресси для исследования эмоций;

2) тесты Кокса для оценки способностей к моральному различению;

3) тесты Хартгорна для оценки честности и надёжности;

4) тесты Термена и Майлза на установки и интересы.

Все они основаны на рассмотрении личности как простой суммы отдельных функций и черт, которые могут быть изучены и оценены по отдельности.

Многомерные тесты учитывают разные стороны поведения человека при оценке тех или иных сторон личности.

Психотехнические тесты предназначены для оценки уровня пригодности испытуемого для обучения и работы в области той или иной профессии.

Каждый из создаваемых тестов проверяют на валидность, которая отражает степень соответствия измеренного показателя тому, что подлежало измерению. Как правило, оценка валидности осуществляется с помощью коэффициентов связи, других математических или логических средств.

Одной из форм тестирования является анкетирование (анкета). Наиболее простые анкеты состоят из «батареи» высказываний, с которыми респондент может согласиться или не согласиться. Специально разрабатывается система количественной оценки основных показателей.

В социологии с помощью анкет изучаются отношение к труду, использование свободного времени, установки, интересы, мотивация и т.п.

Краткие сведения о требованиях, предъявляемых к тестам и самой процедуре тестирования, приведены в последней главе первой части пособия.

1.4. Типы шкал измерения, применяемых в психолого-педагогических и социологических исследованиях. Шкалой измерения называется алгоритм, с помощью которого каждому наблюдаемому объекту ставится в соответствие некоторое число. Приписываемые же объектам числа называют шкальными значениями этих объектов. Единственное требование, предъявляемое к числам, служащим значениями, состоит в том, что рассматриваемые эмпирические отношения должны переходить в соответствующие им числовые отношения. Существует несколько типов шкал, каждому из которых соответствует своё множество допустимых преобразований.

1. Шкалы наименований (номинальные, классификационные шкалы). При построении шкалы наименований предметы группируются в классы (по некоторому признаку). Каждому классу дается наименование и числовое обозначение. Шкалы наименований применяются, если в качестве моделируемых эмпирических отношений выступают лишь отношения равенства и неравенства между объектами. Например, по полу, месту жительства, национальности и т. п.

При номинальных измерениях и обозначениях шкальным значениям (числам) не приписывается никаких свойств, кроме тождества и различия. Допустимые преобразования – произвольные, взаимно однозначные. Даже при таком простейшем измерении к построению шкалы надо подходить с большой осторожностью. Формируемые классы должны иметь социальную значимость. Исследователь должен решить, что он будет классифицировать, какие категории при этом будут исследоваться.

2. Порядковые шкалы (ранговые шкалы). Порядковое измерение возможно тогда, когда исследователь может обнаружить в предметах различие степеней признака или свойства. Для моделирования используется свойство упорядоченности чисел. Чем сильнее выражено у объекта исследуемое качество, тем большее число ему приписывается. Порядковая шкала не только задает некоторую классификацию на множестве объектов, но и устанавливает определенный порядок между классами. Например, классификация рабочих по разрядам, их классификация по стажу, ранжирование школьников по интеллекту.

Порядковые шкалы допускают произвольные монотонно – возрастающие преобразования, т.е. такие преобразования $y = g(x)$, которые удовлетворяют условию: если $x_1 < x_2$ то $y_1 < y_2$ для любых чисел x_1 и x_2 из области определения. Например, вместо уровней интеллекта JQ: $x_1 = 30$,

$x_2 = 25,\ \ x_3 = 46,\ \ x_3 = 46,\ \ x_4 = 75$ можно ввести уровни: $y_1 = 4,\ y_2 = 3,$ $y_3 = 5,\ y_4 = 6.$

3. Интервальные шкалы. Интервальные шкалы получаются, если в процессе измерения исследователь моделирует не только те отношения, которые моделируются при использовании порядковой шкалы, но и отношение равенства для разностей между изучаемыми объектами. Если группам, характеризуемым некоторым признаком, приписаны числа x_1 и x_2, то имеет смысл рассматривать разности $x_2 - x_1$. Так, измеряя температуру по шкале Цельсия, имеет смысл сравнивать перепады температур. Например, можно утверждать, что падение температуры с 6 до 2° равно падению температуры с 12 до 8°.

Далеко не всякую порядковую шкалу можно превратить в интервальную. Например, классификацию рабочих по разрядам. Естественно, что сопоставлять дистанции между каждой парой разрядов не имеет смысла.

Интервальные шкалы допускают соответствующие положительные линейные преобразования, т.е. преобразования вида: $g(x) = ax + b$, где $a > 0$. Например, значения 24; 21; 15; 12; 15 путём преобразования $g(x) = \frac{1}{3} \cdot x - 3$ можно заменить числами 5; 4; 2; 1; 2.

Главная трудность при построении интервальных шкал состоит в обосновании равенства дистанций между объектами.

Интервальную шкалу с фиксированной единицей измерения (рост в сантиметрах, возраст в годах, доход в рублях, стаж в годах) часто называют шкалой разностей.

4. Шкалы отношений. Отличаются от интервальных шкал тем, что нулевая точка в них не произвольна, а указывает на полное отсутствие измеряемого свойства. Например, рост, вес, температура по шкале Кельвина, время, затрачиваемое на решение задачи.

В шкале отношений имеет смысл рассматривать не только разности, но и отношения чисел (меток). В шкале роста, например, имеет смысл говорить, что ученик А в 2 раза выше ученика В, а в шкале температур по Кельвину, что предмет, нагретый до температуры 430°k в 2 раза теплее предмета, нагретого до температуры в 215°k.

В интервальной же шкале рассмотрение таких отношений не имеет смысла. Если, например, температура воздуха ночью 2°С, а днем 6°С, вряд ли имеет смысл говорить, что днем в три раза теплее, чем ночью. В интер-

вальной шкале метка нуль не означает отсутствия качества (свойства). Шкала отношений допускает положительные преобразования, т.е. преобразования: $g(x) = ax$, где $a > 0$.

1.5. Особенности измерения непрерывных переменных. Настоящего, или точного, значения непрерывной величины никогда нельзя получить, так как процесс измерения, зависящий от чувствительности прибора, на каком-то шаге обрывается. Так, например, если рост, измеряемый мерной линейкой, на которую нанесены деления с интервалом в один сантиметр, зафиксирован в 157 см, то его действительный рост в это время и в этих условиях находится между 156,5 и 157,5 см.

Измерение любой непрерывной величины должно сопровождаться определением точности процесса измерения. Обычно в психолого-педагогических исследованиях время хронометрируется с точностью до 0,1 секунды, рост – с точностью до 1 см, возраст – с точностью до 1 дня. Чувствительность процесса измерения задается минимальной единицей цифровой шкалы, которая фиксируется.

Пределы для точного значения в окрестности любого найденного значения устанавливаются путем прибавления и вычитания половины чувствительности измерительного процесса от найденного значения.

В качестве примера рассмотрим данные, приведённые в таблице 1.1.

Таблица 1.1. Пределы точного измерения в окрестности найденного значения

Переменная	Чувствительность измерения	Результаты измерения	Пределы точного значения
Вес	кг	59 кг	58,5 – 59,5
Возраст	год	25 лет	24,5 – 25,5
Время реакции	0,01 сек	0,53 сек	0,525 – 0,535
Время пробега	0,2 сек	5,6 сек	5,5 – 5,7

Глава 2. Первичная обработка эмпирического материала

2.1. Систематизация эмпирического материала. В различных научных исследованиях и практических вопросах часто приходится рассматривать совокупности вещей или явлений, которые по одним признакам представляют нечто целое и единое, а по другим – разбиваются на группы или классы. Такие совокупности называют *статистическими коллективами* (или статистическими совокупностями); вещи или явления, их составляющие, – *их членами*; признаки, по которым коллектив разбивается на классы или группы, называются *аргументами* коллектива, а число членов в коллективе – *объёмом коллектива*. Статистические коллективы могут состоять из вещей или явлений, относящихся к самым разнообразным областям знания.

Последовательность статистических данных, расположенных в порядке их получения в эксперименте, называют *статистическим рядом.* Если полученные в эксперименте числовые значения расположить в порядке возрастания (убывания), то полученную последовательность называют *вариационным рядом*.

Основным объектом, с которым статистику приходится иметь дело, является *частота* (n_i) – число, которое показывает, сколько раз наблюдалось интересующее исследователя свойство, явление или вещь в рассматриваемой совокупности. С частотой связана *относительная частота, или частость* (ω_i), определяющая долю частот в общей сумме частот, т.е. $\omega_i = n_i / n$, где n – объем совокупности, равный сумме частот ($n = \sum n_i$). Сумма всех относительных частот (частостей) равна единице. Иногда относительную частоту (частость) выражают в процентах, тогда их сумма должна равняться 100%.

В качестве примера рассмотрим результаты тестирования первокурсников математического факультета КГПИ 1991–92 учебного года на предмет измерения их *учебной мобильности* (готовность к усвоению нового учебного материала). Измерения велись по особым тестам, в интервальной шкале, и оценивались в баллах. Оценки фиксировались в порядке их получения. Часть образовавшегося статистического ряда приведёна в таблице 2.1.

Таблица 2.1. Уровни учебной мобильности студентов первого курса математического факультета КГПИ (в баллах от 0 до 150)

(по данным тестирования, проведённого в 1991–92 учебном году)

Номера наблюдений	1	2	3	4	5	6	7	8	9	10
Кол-во набранных баллов	90	66	106	84	105	83	104	82	97	97

11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
59	95	78	70	47	95	100	69	44	80	75	75	51	109	89	58

27	28	29	30	31	32	33	34	35	36	37	38
59	72	74	75	81	71	68	112	62	91	93	84

Однако подобное расположение данных не слишком удобно. Можно лишь с трудом судить, например, о том, будет ли первый в списке ученик с оценкой 90 баллов преуспевающим или только средним по сравнению со своими сокурсниками. Если значения результатов измерения расположить в возрастающем порядке и подсчитать соответствующие частоты, то получим так называемый дискретный вариационный ряд. Он изображён в таблице 2.2.

Таблица 2.2. Уровни учебной мобильности студентов первого курса математического факультета КГПИ (в баллах от 0 до 150)

(по данным тестирования, проведённого в 1991–92 учебном году)

Набрано баллов	44	47	51	58	59	62	66	68	69	70
Частота	1	1	1	1	2	1	1	1	1	1

71	72	74	75	78	80	81	82	83	84	89	90	91	93	95	97
1	1	1	3	1	1	1	1	1	2	1	1	1	1	2	2

100	104	105	106	109	112	Сумма
1	1	1	1	1	1	38

Если каждое значение частоты разделить на количество попавших в выборку студентов, т.е. на 38, то получим распределение относительных частот $\omega_i = n_i / 38$, сумма которых равна единице. В качестве примера приведем несколько первых значений таблицы 2.2 (табл. 2.3).

Таблица 2.3. Уровни учебной мобильности студентов первого курса математического факультета КГПИ (в баллах от 0 до 150)

(по данным тестирования, проведённого в 1991–92 учебном году)

Набрано баллов	44	47	51	58	59	62
Частота (n_i)	1	1	1	1	2	1
Относит. частота (ω_i)	0,0263	0,0263	0,0263	0,0263	0,0526	0,0263

В общем случае дискретный вариационный ряд имеет вид (табл. 2.4):

Таблица 2.4. Вариационный ряд, общий вид

Значения случайной переменной, x_i	x_1	x_2	. . .	x_i	. . .	x_k
Частоты, n_i	n_1	n_2	. . .	n_i	. . .	n_k
Относительная частота, $\omega_i = n_i / n$	ω_1	ω_2	. . .	ω_i	. . .	ω_k

Где $x_1 < x_2 < \ldots < x_k$, а n, равное сумме частот, – объем коллектива.

При большом количестве отдельных значений работа с дискретными вариационными рядами представляет определенные неудобства. В таких случаях ряд преобразуют в так называемый *интервальный вариационный ряд*. В общем случае он имеет вид (табл. 2.5):

Таблица 2.5. Интервальный ряд, общий вид

Интервалы	c_0, c_1	c_1, c_2	. . .	c_{k-1}, c_k
Частоты, n_i	n_1	n_2	. . .	n_k
Относительная частота, $\omega_i = n_i / n$	ω_1	ω_2	. . .	ω_k

При построении интервального вариационного ряда прежде всего необходимо решить вопрос об оптимальном числе интервалов. При этом, как правило, приходится делать выбор между двумя крайностями: слишком крупные интервалы для данного объема выборки скрадывают многие нюансы в описании явления, а слишком мелкие ведут к статически незначимым частотам внутри интервала. Важно, чтобы группировка наиболее полно выявляла существенные свойства распределения.

В статистике существует несколько формальных правил определения оптимального числа интервалов:

а) формула Стэрджеса $m = [1 + 3{,}332 \cdot \lg K] + 1$; б) $m = 5 \cdot [\lg K]$; в) $m = 2 \cdot [\ln K]$; г) $m = [\sqrt{K}]$, где K – объем выборки, [ ] – символ целой части числа.

Однако правила эти при работе с психолого-педагогическими и социологическими данными часто дают либо недостаточное, либо избыточное число интервалов. Опыт показывает, что по каждому из признаков не следует брать меньше пяти и более двадцати группировочных интервалов.

Интервальные ряды распределения могут строиться с равными и неравными интервалами. Неравные интервалы применяются при неравномерном распределении частот значений группировочного признака – для выделения качественно отличных типов явлений. Например, при выборе интервалов группировки по возрасту можно основываться на этапах жизненного цикла.

Если у исследователя нет предварительной информации о характере распределения по тому или иному признаку, то следует задавать равные интервалы. Равные интервалы также наиболее удобны при использовании методов математической статистики.

В этом случае длину интервалов определяют по формуле:

$d = \dfrac{x_{\max} - x_{\min} + 1}{m}$, где $x_{\min}$ – наименьшее, а $x_{\max}$ – наибольшее значение варианты, а m – выбранное число интервалов. Полученное значение d, как правило, округляют до целочисленного значения.

В рассмотренном выше примере формальные методы дают для числа интервалов m значения от 6 до 8. Так как самая высокая оценка $x_{\max} = 112$, а самая низкая $x_{\min} = 44$, то $x_{\max} - x_{\min} + 1 = 69$. Поэтому если количество интервалов взять равным шести, то длина одного интервала будет равна 11,5 ед. Если же количество интервалов взять равным восьми, то длина одного интервала будет равна 8,6 ед. В этих условиях за длину интервала естественно принять число $d = 10$.

Определим теперь нижнюю границу интервалов. Для этого среди целых чисел, кратных $d = 10$, ищем наибольшее, не превышающее $x_{\min} = 44$. В данном случае им будет число 40. В результате получим 8 интервалов: (40;50), (50;60), (60;70), (70;80), (80;90), (90;100), (100;110), (110;120).

На последнем шаге построения интервального вариационного ряда надо подсчитать количество членов коллектива, попадающих в каждый интервал. При этом необходимо условиться, что делать с теми значениями аргумента, которые приходятся на границы промежутков: причислять их к верхней границе интервала, или к нижней, или последовательно то к нижней, то к верхней. В первом случае интервал можно обозначать скобками (…], во втором – […) и, наконец, в третьем случае либо […], либо (…).

Таблица 2.6. Распределение первокурсников математического факультета КГПИ (1991–92 уч. год) по уровням учебной мобильности

Границы интервалов	Частоты	Относительные частоты	Границы интервалов	Частоты	Относительные частоты
[40;50)	2	0,0526	[80;90)	7	0,1842
[50;60)	4	0,1053	[90;100)	7	0,1842
[60;70)	4	0,1053	[100;110)	5	0,1316
[70;80)	8	0,2106	[110;120)	1	0,0263
			Итого	**38**	**1,000**

Выполнив указанную процедуру, получим интервальный вариационный ряд, изображённый в таблице 2.6. Такие ряды часто называют *статистическими интервальными рядами распределения.*

Заметим, что каждую из относительных частот можно рассматривать как *вероятность* того, что студент, случайным образом отобранный из 38-ми проверявшихся, набрал от c_i до c_{i+1} баллов. Например, вероятность того, что он набрал от 70 до 80 баллов, равна 0,2106.

Иногда бывает полезным вновь вернуться к дискретному вариационному ряду, заменив в интервальном ряду каждый интервал его серединой: $x_i^+ = \dfrac{c_{i-1} + c_i}{2}$, где $i = 1,2,\ldots,k$. Полученный центрированный ряд имеет вид (табл.2.7):

Таблица 2.7. Центрированный интервальный ряд, общий вид

Середины интервалов	x_1^+	x_2^+	…	x_i^+	…	x_k^+
Частоты, n_i	n_1	n_2	…	n_i	…	n_k
Относит. частоты, $\omega_i = n_i/n$	ω_1	ω_2	…	ω_i	…	ω_k

В нашем случае такой ряд изображён таблицей 2.8.

Таблица 2.8. Распределение первокурсников математического факультета КГПИ (1991–92 уч. год) по уровням учебной мобильности

Середины интервалов	Частоты	Относительные частоты
45	2	0,053
55	4	0,105
65	4	0,105
75	8	0,211
85	7	0,184
95	7	0,184
105	5	0,132
115	1	0,026
Итого	38	1,0000

Довольно часто исследователь сталкивается с ситуацией, когда необходимо провести перегруппировку материала, задав другие интервалы, но нет возможности при этом обратиться к первоначальным статистическим данным. В этом случае приходится вводить априорное предположение о частотном распределении внутри интервала. Самым простым является предположение о равномерности частотного распределения по отдельным значениям признака. Другие формы распределения требуют довольно громоздких вычислений.

2.2. Представление данных. Статистические таблицы. Графическая интерпретация. Материал, содержащийся в рассмотренных выше основных таблицах, кладётся в основу более сложных таблиц, детализирующих исходные данные. В соответствии с программой исследования и методиками обработки материала составляются более сложные таблицы, позволяющие сопоставлять ряды распределений, и, наконец, комбинационные таблицы, в которых три или более признака перекрещиваются, комбинируются. По таким таблицам устанавливаются, измеряются и анализируются связи между признаками исследуемой совокупности объектов.

Построение таблиц подчинено определенным правилам. Основное содержание таблицы должно быть отражено в названии (круг рассматриваемых вопросов, географические границы статистической совокупности, время, единицы измерения). Таблицы бывают простые, групповые и комбинационные. Простые таблицы представляют собой перечень, список от-

дельных единиц совокупности с количественной (или качественной) характеристикой каждой из них в отдельности.

В групповых таблицах содержится группировка единиц совокупности по одному признаку, например, по полу (табл. 2.9).

Таблица 2.9. Распределение первокурсников КГПУ 1992–93 учебного года по уровню учебной мобильности в зависимости от пола

Середины интервалов	Частоты		Относительные частоты	
	юноши	девушки	юноши	девушки
45	2	-	0,1053	-
55	2	2	0,1053	0,1053
65	2	2	0,1053	0,1053
75	3	5	0,1579	0,2631
85	4	3	0,2105	0,1579
95	3	4	0,1579	0,2105
105	2	3	0,1053	0,1579
115	1	-	0,0525	-
Итого	19	19	1,000	1,000

В комбинационных таблицах содержится группировка единиц совокупности по двум и более признакам, например, по месту жительства и полу (табл. 2.10).

Таблица 2.10. Распределение первокурсников КГПУ 1992–93 учебного года по уровню учебной мобильности в зависимости от места жительства и пола (число испытуемых)

Середины интервалов	Город		Село		Всего
	юноши	девушки	юноши	девушки	
45	1	-	1	-	2
55	1	1	1	1	4
65	2	-	-	2	4
75	1	3	2	2	8
85	2	2	2	1	7
95	2	2	1	2	7
105	1	1	1	2	5
115	-	-	1	-	1
Итого	10	9	9	10	38

Статистические распределения изображаются также в виде диаграмм и графиков. Главным достоинством такого изображения является его наглядность. Наиболее распространёнными среди графических средств изображения статистических распределений являются *гистограммы, полигоны* и *кумуляты* распределения.

Гистограмма – графическое изображение интервального распределения. Строится в декартовой системе координат. При её построении на оси абсцисс откладываются отрезки, изображающие интервалы значений варьирующего признака. На этих отрезках, как на основаниях, строятся прямоугольники, площади которых пропорциональны частотам соответствующих интервалов.

Заметим, что коэффициент пропорциональности k всегда можно выбрать таким образом, чтобы площадь всей ступенчатой фигуры равнялась единице. Для этого достаточно выбрать его таким, чтобы площадь i - го прямоугольника равнялась относительной частоте ω_i, т.е. $h_i \cdot \Delta x_i = \omega_i$ и, следовательно, $h_i = \frac{\omega_i}{\Delta x_i}$. В рассмотренном выше случае (табл. 2.6) гистограмме можно придать вид, приведённый на рисунке 2.1.

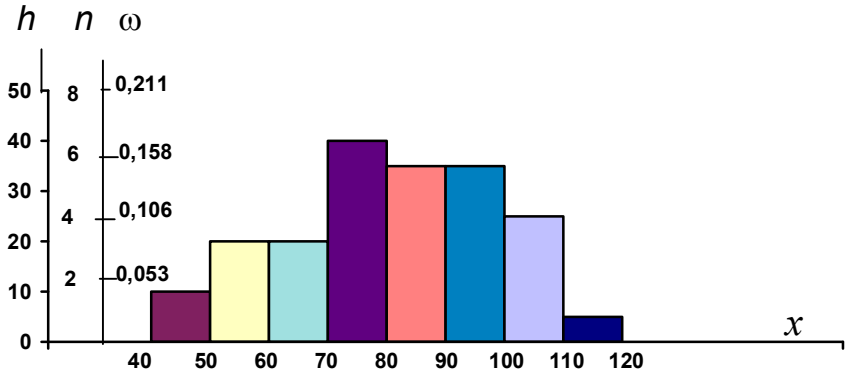

Рис. 2.1. Гистограмма распределения первокурсников КГПУ 1992-93 уч. года по уровню учебной мобильности

Полигон (многоугольник) распределения – одна из форм графического изображения вариационных рядов, как дискретных, так и интервальных. Различают полигоны распределения частот и относительных частот. Если в прямоугольной системе координат построить точки $(x_i; h_i)$, где x_i – значе-

ние признака, а h_i, как и в случае гистограмм, пропорционально либо n_i, либо ω_i, а затем полученные точки соединить отрезками, то полученную ломаную называют *полигоном* распределения. Полигон распределения первокурсников по уровню учебной мобильности приведён на рисунке 2.2.

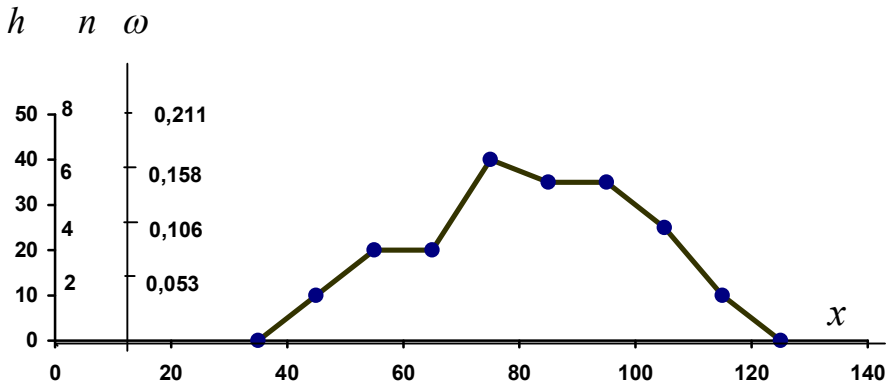

Рис. 2.2. Полигон распределения первокурсников КГПУ 1992–93 уч. года по уровню учебной мобильности

Если распределение уже изображено с помощью гистограммы, то, соединяя середины верхних оснований прямоугольников отрезками, можно получить полигон распределения. Также легко перейти от полигона к гистограмме.

Часто крайние ординаты признака соединяют с серединами примыкающих интервалов. Однако для распределения, где концентрация событий увеличивается на концах полигона, такое изображение может привести к ложным представлениям о сущности явления.

Кумулята – графическое изображение статистического ряда накопленных частот или относительных частот, где по оси абсцисс откладываются значения признака, а по оси ординат – накопленные частоты или накопленные относительные частоты, показывающие, сколько единиц совокупности имеют значения признака, не превосходящие данное значение. В нашем случае (распределение первокурсников по уровням учебной мобильности) расчёт накопленных частот и относительных частот приведён в таблице 2.11, а сама кумулята – на рисунке 2.3.

Таблица 2.11. Расчёт накопленных частот и относительных частот распределения первокурсников по уровням учебной мобильности

Середины интервалов	45	55	65	75	85	95	105	115

Частоты	2	4	4	8	7	7	5	1
Накопленные частоты	2	6	10	18	25	32	37	38
Относит. частоты	0,053	0,105	0,105	0,211	0,184	0,184	0,132	0,026
Накопленные относит. частоты	0,053	0,158	0,263	0,474	0,658	0,842	0,974	1.000

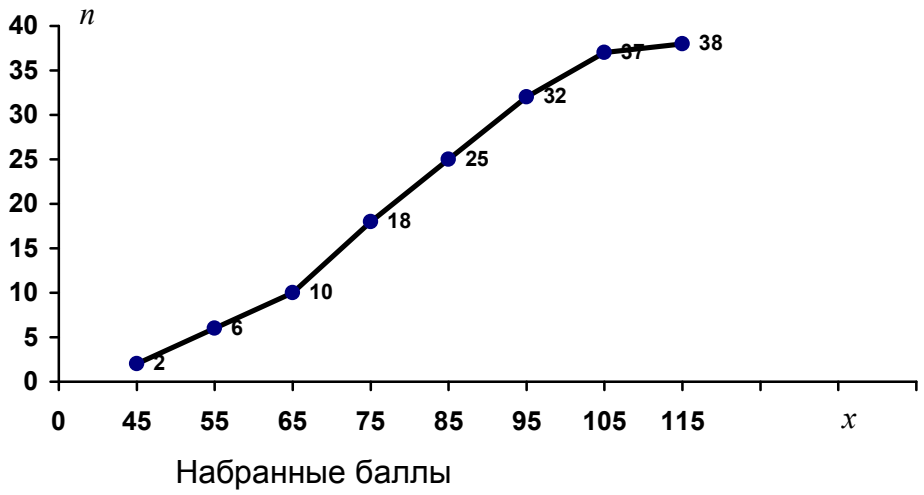

Рис 2.3. Кумулята распределения первокурсников по уровню учебной мобильности

Глава 3. Эмпирические и теоретические распределения

3.1. Кривые эмпирических распределений. Введём теперь понятие *кривой эмпирического распределения*. Рассмотрим полигон относительных частот какого-либо эмпирического (опытного) распределения. Если увеличивать число членов совокупности, уменьшая в то же время интервалы, на которые разбиваются значения аргумента, то звенья полигона будут становиться всё мельче и мельче, а сам полигон всё более похожим на плавную кривую линию.

Действительно, предположим, что, проверяя первокурсников на учебную мобильность, мы протестировали не 38, а 100 человек и получили распределение, задаваемое таблицей 3.1, в которой длина интервалов уменьшена по сравнению с данными таблицы 2.6 в два раза. Соответствующие гистограмма и полигон изображены на рисунке 3.1.

Таким образом, увеличивая объём совокупности и уменьшая промежутки, мы будем получать полигоны, всё более приближающиеся к некоторой плавной кривой, которая будет для них в некотором смысле предельной. Эту кривую называют кривой относительных частот, или *кривой эмпирического распределения* (рисунок 3.2).

Таблица 3.1. Гипотетическое распределение первокурсников по уровню учебной мобильности (в набранных баллах).

Интервалы	40 44	45 49	50 54	55 59	60 64	65 69	70 74	75 79	80 84	85 89	90 94	95 99	100 104	105 109	110 114	115 119
n_i	1	2	4	5	7	10	12	14	13	11	9	5	3	2	1	1

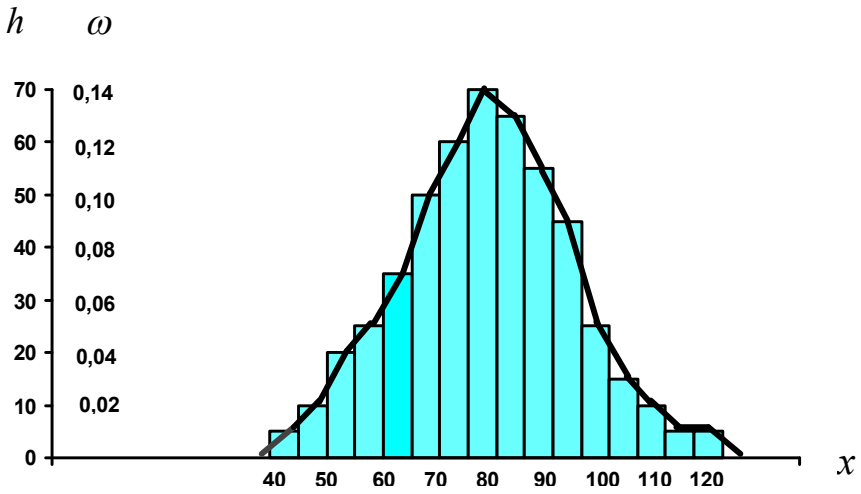

Рис.3.1. Гистограмма и полигон гипотетического распределения 100 первокурсников по уровням учебной мобильности

Кривые эмпирических распределений употребляются для изучения законов распределения относительных частот, их свойств, для сравнения распределений между собой и т.п. Если относительные частоты какой-либо совокупности действительно подчинены некоторому закону, то он проявляется тем отчётливее, чем больше объём совокупности, ибо с увеличением n увеличиваются шансы того, что различные случайные причины, вызывающие отклонение относительной частоты определённых значений аргумента от их истинных значений, уравновесятся. Пользуясь кривой эмпирического распределения, можно найти долю (частоту) или численность членов совокупности, имеющих значения аргумента (в нашем случае количество набранных баллов, характеризующих уровень учебной мобильности), которые заключены между данными пределами, например, между $\alpha = 70$ и $\beta = 90$. С этой целью проведём через точки α и β (рис. 3,2) перпендикуляры к оси ox до пересечения с кривой распределения.

Затем, пользуясь палеткой, найдём площадь заштрихованной части криволинейной трапеции с основанием α, β и площадь всей криволинейной трапеции. Разделив первый результат на второй, получим долю учащихся, набравших от α до β баллов. В нашем случае она равна $0{,}69$.

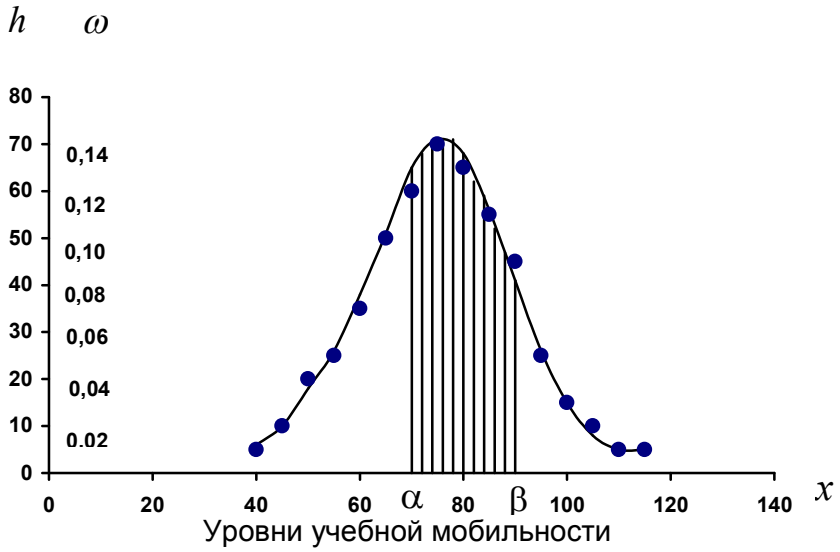

Рис. 3.2. Кривая эмпирического распределения первокурсников КГПУ 1992–93уч. года по уровням учебной мобильности

Формы кривых эмпирических распределений. Формы кривых эмпирических распределений (для краткости часто говорят о форме эмпирических распределений) весьма разнообразны, но среди них можно выделить следующие главные: симметрические, умеренно асимметрические, крайне асимметрические и V-образные. Все эти формы относятся к тем графикам, которые обнаруживают лишь одну наибольшую или наименьшую частоту

(по сравнению с соседними), т.е. к так называемым одновершинным распределениям.

Симметрическими распределениями называют такие распределения, в которых частоты любых двух значений аргумента, равноотстоящих в обе стороны от некоторого его среднего значения, равны между собой (рис. 3.3(а)).

Умеренно асимметрическими называются распределения, в которых частоты аргументов, равноудалённых от некоторого среднего значения, с одной стороны всё время больше или всё время меньше соответствующих частот с другой стороны (рис. 3.3(б)).

Крайнюю асимметрию обнаруживают те распределения, в которых наибольшую частоту имеют наибольшие или наименьшие из значений аргумента, так что все частоты распределения расположены по какую-нибудь сторону от наибольшего или наименьшего значения (рис. 3.3(в)).

Наконец, *V-образное распределение* характеризуется присутствием в нём некоторой средней частоты, которая меньше остальных и по обе стороны от которой частоты возрастают к концам распределения. Такие распределения будет встречаться нам очень редко (рис. 3.3(г)).

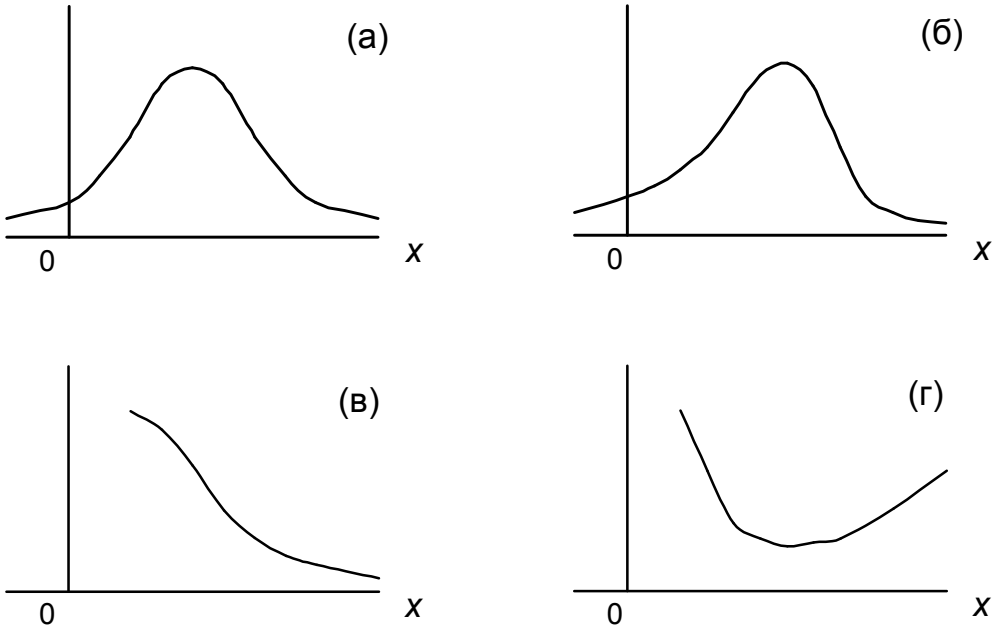

Рис 3.3. Основные формы кривых распределения

Если кривая распределения имеет несколько вершин, то это может указывать на неоднородность исследуемой совокупности и может объясняться наложением двух или более различных совокупностей или наличи-

ем третьей, скрытой, переменной, детерминирующей распределение совокупности на несколько групп, а также рядом других аналогичных причин.

3.2. Квантили. Квантилем называют точку числовой оси (шкалы), которая делит совокупность наблюдений (но не шкалу) на две группы с заданным отношением численности первой из них к численности всей совокупности. Квантиль, делящий совокупность на две равные части, обозначается $K_{1/2}$ и называется *медианой*. Точки числовой оси, делящие данную совокупность на четыре равные части, называют *квартилями*. Левую из этих точек, обозначаемую $K_{1/4}$, называют *нижней квартилью*, правую – $K_{3/4}$ *– верхней квартилью*. Квантили $K_{0,1}, K_{0,2}, \ldots, K_{0,9}$, делящие совокупность (но не шкалу) на десять равных частей, называются *децилями*. Квантили, делящие совокупность на сто равных частей, называются *процентилями*. *Процентиль,* отделяющую 35% совокупности слева, обозначают $K_{0,35}$

Рассмотрим теперь процедуру вычисления квантили $P_{0.25}$ распределения студентов по уровню их учебной мобильности (табл. 3.1). В этих целях дополним таблицу 3.1 строкой, содержащей накопленную частоту.

Таблица 3.2. Гипотетическое распределение первокурсников по уровню учебной мобильности (в набранных баллах).

						M										
Интервалы	40 44	45 49	50 54	55 59	60 64	65 69	70 74	75 79	80 84	85 89	90 94	95 99	100 104	105 109	110 114	115 119
n_i	1	2	4	5	7	10	12	14	13	11	9	5	3	2	1	1
Накопленная частота	1	3	7	12	19	29	41	55	68	79	88	93	96	98	99	100

Шаг 1. Находим число учащихся, попадающих в группу, определяемую заданным квантилем: $pn = 0.25 \cdot 100 = 25$.

Шаг 2. Находим столбец, в который при счете слева попадает 25-й учащийся. Получаем столбец, соответствующий интервалу баллов от 65 до 69 . Обозначаем его буквой М.

Шаг 3. Найдем «фактическую» (см. п. 1.5) нижнюю границу баллов выделенного столбца (М). Обозначим ее буквой L: L=65 - 0,5=64,5.

Шаг 4. Найдем частоту, накопленную к выделенному столбцу. Ее обозначают $cum.f$, и она равна 19.

Шаг 5. Вычтем из np=25 накопленную частоту $cum.f$=19. Получим 6.

Шаг 6. Найдем долю баллов, пропорциональных частоте 6 в выделенном столбце: $x:6=5:10$ откуда $x=\frac{6\cdot 5}{10}=3{,}0$

Шаг 7. $P_{0.25}$=64,5+3,0=67,5.

Общая формула определения квантилей с долей p в группе n оценок имеет вид:

$$P_p = L + \frac{pn-(cum.f)}{f}\cdot(W), \text{ где} \qquad (3.1)$$

np – количество учащихся составляющих p-ю часть совокупности. В нашем случае, когда p=0,25, $np=100\cdot 0{,}25=25$.

L – фактическая нижняя граница интервала, содержащего pn-ю частоту снизу. В нашем случае L=64,5.

$cum.f$ - накопленная к L частота. В нашем случае $cum.f$=19.

f – частота интервала, содержащего частоту pn. В нашем случае $f=10$.

W – ширина любого интервала оценок. В нашем случае W=5.

Пользуясь этим правилом, вычислим еще раз

$$P_{0,25} = L + \frac{pn-cum.f}{f}\cdot W = 64{,}5 + \frac{25-19}{10}\cdot 5 = 67{,}5.$$

Знание квантилей совокупности дает представление о расположении и рассеянии значений случайной величины и, в частности, о функции распределения.

3.3. Распределение вероятностей. Большую роль в статистике играют распределения вероятностей, которые, с одной стороны, противостоят эмпирическим распределениям, а с другой – служат средством их моделирования. В отличие от эмпирических распределений, в основе которых лежат статистические эксперименты и наблюдения, распределения вероятностей основываются на «мысленных» экспериментах, идеализирующих условия эксперимента.

Для более детального знакомства с этим вопросом обратимся к основным понятиям и методам теории вероятностей. Одним из таких понятий является понятие случайной величины и её распределения. В теории вероятностей *под случайной величиной понимается величина, принимающая то или иное числовое значение в зависимости от случая.* Случайные величины делятся на дискретные и непрерывные.

Дискретной случайной величиной называют величину, принимающую случайным образом конечное число или бесконечную последовательность чисел, например, число выстрелов, производимых до первого попадания в цель.

Непрерывной случайной величиной называют случайную величину, принимающую все значения из некоторого интервала, например, расстояние от центра мишени до точки попадания.

Чтобы задать дискретную случайную величину X надо указать множество всех её значений $x_1, x_2, \ldots$ и вероятности, с которыми она принимает каждое из них, т.е. вероятности $p_i = P\left(X = x_i\right) \quad (i = 1..2,...)$, состоящие в том, что случайная величина X принимает значение x_i.

Соотношение, устанавливающее связь между значениями случайной величины и вероятностями этих значений, называют законом распределения случайной величины.

Если речь идет о непрерывной случайной величине (в том смысле, как она была определена выше), то все её значения перечислить невозможно. В этом случае её задают с помощью функции $F(x)$, значение которой в точке x равно вероятности того, что случайная величина X примет значение лежащее левее точки x, т.е. $F(x) = P(X < x)$. *Эту функцию F(x) называют функцией распределения случайной величины или интегральным законом распределения*. Ее также называют *накопленным (кумулятивным) распределением случайной величины.* Из этого определения следует, что для любой пары чисел $(\alpha < \beta)$, вероятность того, что случайная величина X примет значение, лежащее между ними, равна:

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha). \qquad (3.2)$$

Впрочем, такого рода определения бывают полезны не только для непрерывных случайных величин.

Если функция $F(x)$ распределения случайной величины является непрерывной и дифференцируемой*, то функцию производную от неё*

$$\varphi(x) = F'(x)$$

называют плотностью вероятности или дифференциальным законом распределения случайной величины X.

Зная плотность распределения $\varphi(x)$ величины X, легко вычислить вероятность того, что последняя удовлетворяет неравенствам $\alpha < X < \beta$, а именно:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} \varphi(x)dx. \qquad (3.3)$$

Приведенной формуле легко дать наглядное истолкование. Построим график плотности вероятности $y = \varphi(x)$, тогда вероятность того, что случайная величина X примет значение, принадлежащее интервалу (α,β), будет равна площади криволинейной трапеции, построенной на промежутке: (α,β) и ограниченной сверху графиком функции $y = \varphi(x)$.

Учитывая, что вероятность того, что случайная величина X примет значение из интервала $(-\infty,\infty)$ равна единице, т.е. $P(-\infty < X < \infty) = 1$, приведенное выше равенство можно записать в виде $\int_{-\infty}^{\infty} \varphi(x)dx = 1$, откуда следует, что площадь криволинейной трапеции под графиком функции плотности всегда равна единице.

В качестве важного примера непрерывного распределения назовем нормальное распределение, с которым вам придется иметь дело на протяжении всей этой книги и знакомство с которым начнется уже в этой главе. Кроме того, во второй части пособия вы познакомитесь с такими важными для статистики распределениями, как распределение Стьюдента, хи-квадрат распределение, распределение Фишера и некоторыми другими.

Понятие квантиля, использованное нами для анализа экспериментальных распределений, используется также и при изучении теоретических распределений вероятностей, задаваемых интегральной функцией распределения. Если функция распределения $F(x)$ случайной величины X непрерывна и строго монотонна, то для любого p, $0 < p < 1$, *квантиль* порядка p распределения случайной величины X определяется как корень K_p уравнения $F(K_p) = p$. Для применения в математической статистике составлены таблицы квантилей некоторых важных распределений.

Если случайная величина X задана плотностью $\varphi(x)$, то положение точки x на оси аргументов может определяться вероятностью α того, что случайная величина X примет значение, лежащее правее этой точки. В этом случае будем ее обозначать символом $_{\alpha}P$, где P символизирует вид распределения.

3.4. Нормальное распределение. В большинстве эмпирических распределений, с которыми приходится иметь дело исследователю, на изучаемую величину X, помимо одного, главного, фактора, действует большое число слабо действующих случайных факторов (помех). Например,

если рабочий получил задание нарубить из металлического прута стержни длинною 120 мм, то кроме основного фактора – задания, на результат его работы действует множество других второстепенных факторов: возможная неоднородность материала, качество зубила и прочность закрепления прута, твердость руки и периодическое снижение внимания и т.п. В результате длина нарубленных стержней будет несколько отличаться от заданного стандарта то в одну, то в другую стороны. При этом чем больше длина стержня отличается от заданной, тем реже такие стержни встречаются.

Если на основании результатов измерения нарубленных стержней построить полигон распределения их длин, то он примет вид симметричной, колоколообразной кривой. Подобную форму имеют и полигоны, построенные по результатам повторных одинаковых измерений одного и тех же объекта.

В результате полигон распределения оказывается одновершинным и симметричным относительно прямой $x = \mu$, где $\mu = 120$. Чем ближе x_i к 120, тем больше частота и относительная частота. В пределе кривая распределения примет вид, изображённый на рисунке 3.3(а).

Такие распределения так часто встречались в самых различных областях науки и практики, что их стали принимать за норму всякого массового случайного проявления признаков и в соответствии с этим называть *нормальными распределениями.*

В своё время Кетле (1796 – 1874) обнаружил, что рост солдат-ровесников подчиняется нормальному распределению. По его мнению, причиной такого распределения являются ошибки, которые делает природа при воспроизведении среднего идеального человека. Школа Кетле, которая в законе ошибок Муавра (1667 – 1754), Лапласа (1749 – 1827) и Гаусса (1777 – 1855) увидела закон природы, говорила также о «среднем человеке» с его «средней наклонностью к самоубийству», «средней наклонностью к преступлению» и многом тому подобном. В то время как число «лучей» в плавниках камбалы распределено практически нормально, в окружающем нас мире встречается множество распределений, которые лишь с большой натяжкой можно описать нормальным распределением Муавра. Он его открыл и указал на его особое значение.

В настоящее время нормальным называют распределение, плотность которого выражается функцией:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (3.4)$$

Средствами математического анализа легко доказать, что график этой функции имеет вид холма, «склоны» которого полого спускаются к оси абсцисс.

Действительно, приведенная выше функция определена, непрерывна и положительна при любом действительном значении x, и потому её график лежит выше оси абсцисс. Вычисление показывает, что определенный интеграл от этой функции равен единице и поэтому площадь соответствующей криволинейной трапеции при любых μ и σ равна единице.

Так как в точке $x=\mu$ первая производная функции равна нулю, а вторая – отрицательна, то в ней функция плотности имеет максимум, равный, как показывает расчет, $\frac{1}{\sigma\sqrt{2\pi}}$. Если увеличивать значение параметра σ, то вершина «холма» будет опускаться, но зато поднимутся его «склоны» (ведь площадь под кривой не меняется).

Методами математического анализа можно также доказать, что график $\varphi(x)$ имеет две точки перегиба: одна соответствует значению $x=\mu-\sigma$, другая – значению $x=\mu+\sigma$ (рис. 3.4).

Что касается параметра μ, то его значение не влияет на форму графика. С его изменением график только смещается в направлении оси абсцисс. Таким образом, форма кривой нормального распределения зависит только от параметра σ: с его увеличением вершина графика функции опускается, а точки перегиба $\mu-\sigma$ и $\mu+\sigma$ раздвигаются, кривая становится более плосковершинной. С уменьшением значения σ, наоборот, вершина кривой поднимается, а точки перегиба сближаются, кривая становится более островершинной.

При заданных значениях параметров μ и σ можно вычислить долю тех значений X, которые принадлежат промежутку (α,β). Она будет равна площади соответствующей криволинейной трапеции, т.е. $\int_{\alpha}^{\beta}\varphi(x)dx$. Вычисления показывают, что в промежутке $(\mu-\sigma,\mu+\sigma)$ содержится примерно 68% значений нормально распределённой случайной величины. В промежутке $(\mu-2\sigma,\ \mu+2\sigma)$ – 95%, а в промежутке $(\mu-3\sigma,\mu+3\sigma)$ – 99,7%, т.е. практически все значения рассматриваемой случайной величины.

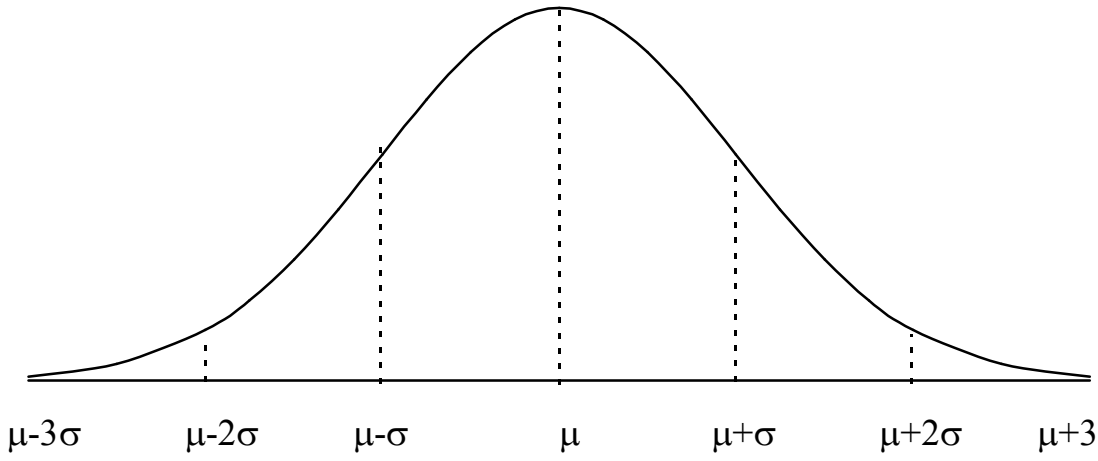

Рис. 3.4. График нормального распределения

Для быстрого построения нормальной кривой, вершина которой имеет абсциссу, равную нулю, можно воспользоваться данными таблицы 3.3, в которой $y_{\max} = \frac{1}{\sigma\sqrt{2\pi}}$.

Таблица 3.3. Опорные точки нормальной кривой

Абсцисса	0	$\pm 0{,}5\sigma$	$\pm 1{,}0\sigma$	$\pm 2{,}0\sigma$	$\pm 3{,}0\sigma$
Ордината	$y_{\max}$	$\frac{7}{8}\, y_{\max}$	$\frac{5}{8}\, y_{\max}$	$\frac{1}{8}\, y_{\max}$	$\frac{1}{80}\, y_{\max}$

В этих условиях особый интерес представляет так называемое стандартное нормальное распределение, при котором $\mu=0$, а $\sigma=1$. В этом случае функция примет вид:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-0{,}5x^2}. \qquad (3.5)$$

При построении графика можно воспользоваться таблицей N Приложения.

Нормальная кривая – это изобретение математиков, довольно хорошо описывающее эмпирический полигон частот величины, находящейся под влиянием многих слабо действующих случайных факторов. Никогда не была, да и не будет получена совокупность данных, которые были бы распределены точно по нормальному закону. Но часто бывает полезным, допуская незначительную ошибку, предположить, что значения случайной величины распределены по нормальному закону.

Переходя от функции плотности нормального распределения с параметрами $\mu = 1,\ \sigma = 1$ $\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$ к интегральной функции рас-

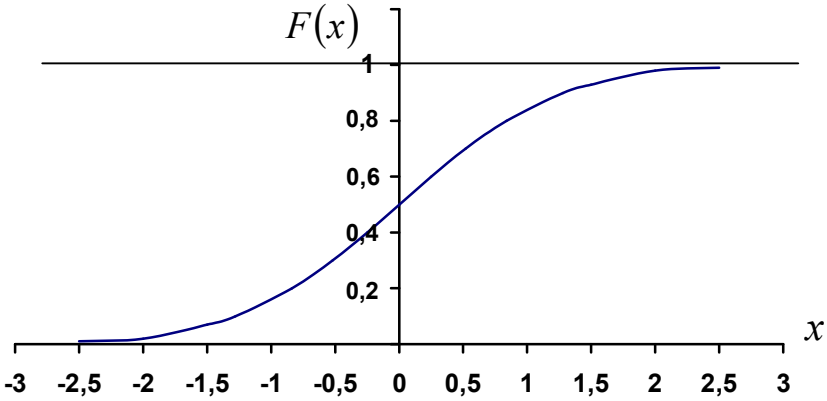

Рис. 3.5. Кумулята стандартного нормального распределения

пределения $F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-0,5t^2}dt$, исследуем эту функцию с помощью децилей.

Для этого удобно воспользоваться функцией Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\cdot\int_{0}^{x} e^{-0,5t^2}dt$, значения которой приведены в таблице N Приложения. Если учесть, что $F(x)=0,5000+\Phi(x)$ и считать функцию $\Phi(x)$ нечетной, то, пользуясь данными таблицы N Приложения, можно построить точки кумуляты стандартного нормального распределения (рис.3.5).

Таблица 3.4. Координаты точек интегральной (кумулятивной) кривой стандартного нормального распределения.

x	-1,28	-0,84	-0,52	-0,25	0	0,25	0,52	0,84	1,28
$\Phi(x)$	-0,4	-0,3	-0,2	-0,1	0	0,1	0,2	0,3	0,4
$F(x)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9

Так как значения $F(x)$ распределены с шагом в 0,1, то децили стандартного нормального распределения равны: $K_{0,1}=-1,28$; $K_{0,2}=-0,84$; $K_{0,3}=-0,52$; $K_{0,4}=-0,25$; $K_{0,5}=0$; $K_{0,6}=0,25$; $K_{0,7}=0,52$; $K_{0,8}=0,84$; $K_{0,9}=1,28$. Квартили этого нормального распределения равны. $K_{1/4}=-0,67$; $K_{3/4}=0,67$.

В дальнейшем мы познакомимся с рядом других теоретических распределения, которыми могут сглаживаться эмпирические полигоны частот, однако «нормальные кривые» имеют серьёзные преимущества, прежде всего потому, что обеспечивают простые и изящные доказательства во многих задачах теории статистического вывода.

Глава 4. Меры центральной тенденции

4.1. Среднее арифметическое. Группировка и построение частотного распределения – лишь первый этап статистического анализа полученных данных. Следующим шагом обработки является получение некоторых обобщающих характеристик, позволяющих глубже понять особенности объекта наблюдения. Сюда прежде всего относятся среднее значение признака, вокруг которого варьируют остальные его значения, и степень разброса этих значений.

В статистике различают несколько видов средних величин: *среднее арифметическое, медиана, мода* и т.д. и несколько показателей разброса значений (мер рассеяния): *дисперсия, среднее квадратичное отклонение, среднее абсолютное отклонение, коэффициент вариации* и т.д.

Из статистических характеристик наиболее важными являются средние величины, представляющие те центры, около которых группируются наблюдаемые значения аргументов и которые определяют в известной мере положение распределений.

Среднее есть абстрактная типическая характеристика всей совокупности. Оно уничтожает, погашает, сглаживает случайные и неслучайные колебания, влияние индивидуальных особенностей и позволяет представить в одной величине некоторую общую характеристику реальной совокупности. Основное условие научного использования средних заключается в том, чтобы каждое среднее характеризовало такую совокупность единиц, которая в существенном отношении, и в первую очередь в отношении усредняемых значений признака, была бы качественно однородной. Среди всего многообразия средних наиболее часто используемой считается *среднее арифметическое*.

Средним арифметическим n чисел a_1, a_2, . . . , a_n называют число: $\bar{x} = \dfrac{a_1 + a_2 + \ldots + a_n}{n}$, или, используя знак суммы, $\bar{x} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} a_i$. Если среди данных чисел встречаются одинаковые, например: x_1 - n_1 раз, x_2 - n_2 - раза, ... , x_k - n_k раз, и $\sum_{i=1}^{k} n_i = n$, то $\bar{x} = \dfrac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \dfrac{\sum_{i=1}^{k} n_i x_i}{n}$.

Таким образом,

$$x = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \qquad (4.1)$$

Например, для восьми чисел 3; 3; 6; 6; 8; 10; 10; 10 среднее арифметическое $\bar{x} = \frac{2\cdot 3 + 2\cdot 6 + 1\cdot 8 + 3\cdot 10}{2+1+3+2} = \frac{56}{8} = 7$.

Среднее арифметическое может быть найдено и по относительным частотам (ω).

Пусть, например, распределение величины X представлено таблицей 4.1.

Таблица 4.1. Данные для расчета средней арифметической

Значения X	x_1	x_2	. . .	x_k
Частоты, n_i	n_1	n_2	. . .	n_k
Относительные частоты, ω_i	ω_1	ω_2	. . .	ω_k

Тогда среднее арифметическое величины X равно:

$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n}$, где $n_1, n_2, ..., n_k$ – веса, а $n = n_1 + n_2 + ... + n_k$.

Разделив слагаемые числителя на n, получим:

$$\bar{x} = \frac{n_1}{n}x_1 + \frac{n_2}{n}x_2 + ... + \frac{n_k}{n}x_k = \omega_1 x_1 + \omega_2 x_2 + ... + \omega_k x_k = \sum_{i=1}^{k}\omega_i x_i$$

$$\bar{x} = \sum_{i=1}^{k}\omega_i x_i \; . \qquad (4.2)$$

В качестве примера рассмотрим распределение первокурсников КГПИ по уровню учебной мобильности (табл. 2.1) и вычислим среднее арифметическое отмеченных в ней баллов. Так как общая сумма балов равна 3050, то, разделив её на количество студентов (38), получим, что $\bar{x} = 80.3$. Несколько иной результат получится, если воспользоваться центрированным интервальным распределением (табл. 2.8):

$$\bar{x} = (2\cdot 45 + 4\cdot 55 + 4\cdot 65 + 8\cdot 75 + 7\cdot 85 + 7\cdot 95 + 5\cdot 105 + 1\cdot 115) : 38 = 80{,}8$$

Различие возникает прежде всего из-за того, что баллы, попавшие в один интервал заменяются серединами интервалов. Естественно, что первый результат более точно соответствует экспериментальным данным.

Рассмотрим теперь основные свойства среднего арифметического.

Свойство 1. Сумма отклонений значений аргумента x от их среднего арифметического, взятых каждое столько раз, сколько раз оно встречается в распределении, равна нулю, т.е. $\sum n_i\left(x_i - \bar{x}\right) = 0$.

Действительно, $\sum n_i\left(x_i - \bar{x}\right) = \sum n_i x_i - \sum n_i \bar{x} = \sum n_i x_i - \bar{x} \cdot \sum n_i =$

$= n \cdot \bar{x} - n \cdot \bar{x} = 0$

Свойство 2. Сумма квадратов отклонений значений аргумента x от их среднего арифметического меньше суммы квадратов отклонений их от любой другой величины, иначе говоря, сумма $\sum n_i\left(x_i - \bar{x}\right)^2$ всегда меньше суммы $\sum n_i\left(x_i - a\right)^2$, какая бы не была взята величина a, отличная от $\bar{x}$.

Это свойство проиллюстрируем на простом примере. Для чисел 3; 3; 6; 6; 8; 10; 10; 10 мы ранее нашли $\bar{x} = 7.$:

Тогда $\sum n_i\left(x_i - \bar{x}\right)^2 = 2 \cdot (3-7)^2 + 2 \cdot (6-7)^2 + 1 \cdot (8-7)^2 + 3 \cdot (10-7)^2 = 62$.

Возьмём теперь какое-либо число $a \neq \bar{x}$, например $a = 6$. Получим:

$$\sum n_i\left(x_i - a\right)^2 = 2 \cdot (3-6)^2 + 2 \cdot (6-6)^2 + 1 \cdot (8-6)^2 + 3 \cdot (10-6)^2 = 70$$

Доказательство этого свойства можно провести следующим образом.

Пусть дана система чисел $x_1, x_2, ..., x_n$ и некоторое число x, тогда $\sum n_i\left(x_i - x\right)^2$ есть функция x. Обозначив её $f(x)$, будем иметь $f(x) = \sum n_i\left(x_i - x\right)^2$.

Чтобы найти, при каком значении x это выражение принимает наименьшее значение (наибольшего значения у него заведомо нет), вычислим производную $f'(x)$ и приравняем ее к нулю:

$$f'(x) = -2 \cdot \sum_{i=1}^{k} n_i\left(x_i - x\right) = -2\left[\sum n_i x_i - \sum n_i x\right] = -2\left[\sum n_i x_i - x \sum n_i\right] =$$

$= -2\left[\sum n_i x_i - nx\right] = 0$. Из последнего равенства получим $x = \dfrac{\sum n_i x_i}{n}$, т.е. $x = \bar{x}$.

Свойство 3. Когда несколько распределений с одним и тем же аргументом x соединяются в одно, тогда среднее арифметическое объединенного распределения будет равно взвешенному среднему средних арифметических соединяемых распределений, причём весами частотных средних будут служить объёмы соединяемых распределений.

Рассмотрим для простоты три распределения (табл. 4.2):

Таблица 4.2. Три распределения для расчета средней арифметической

Распределения	Варианты				Общая численность	Среднее значение
	x_1	x_2	. . .	x_k		
S_1	n_1	n_2	. . .	n_k	N_1	$\bar{x}_1$
S_2	l_1	l_2	. . .	l_k	N_2	$\bar{x}_2$
S_3	g_1	g_2	. . .	g_k	N_3	$\bar{x}_3$

$$N_1=\sum_{i=1}^{k} n_i,\quad N_2=\sum_{i=1}^{k} l_i,\quad N_3=\sum_{i=1}^{k} g_i,\quad \bar{x}_1=\frac{\sum n_i x_i}{N_1},\quad \bar{x}_2=\frac{\sum l_i x_i}{N_2},\quad \bar{x}_3=\frac{\sum g_i x_i}{N_3}.$$

Соединим эти три распределения в одно, тогда $x_1, x_2, \ldots, x_k$ будут иметь частоты $n_1+l_1+g_1;\ n_2+l_2+g_2;\ \ldots\ n_k+l_k+g_k$. Общее среднее $\bar{x}$ будет равно $\bar{x}=\dfrac{(n_1+l_1+g_1)\cdot x_1+(n_2+l_2+g_2)\cdot x_2+\ldots+(n_k+l_k+g_k)\cdot x_k}{(n_1+l_1+g_1)+(n_2+l_2+g_2)+\ldots+(n_k+l_k+g_k)}$, откуда

$\bar{x}=\dfrac{N_1\cdot\bar{x}_1+N_2\cdot\bar{x}_2+N_3\cdot\bar{x}_3}{N_1+N_2+N_3}$. Во всех случаях суммирование по i от 1 до k.

Пример. В трёх классах 8а, 8б и 8в, проведена контрольная работа. В таблице 4.3 приведены ее результаты.

Таблица 4.3. Распределение школьников по качеству выполнения контрольной работы (количество человек)

Классы	Оценки					Итого с/с	Средний балл $\bar{x}$
	1	2	3	4	5		
8а	0	4	14	5	3	26	3,27
8б	0	6	10	8	6	30	3,47
8в	0	3	4	8	7	22	3,86
Итого	0	13	28	21	16	78	3,51

$$\bar{x}=\frac{26\cdot 3{,}27+30\cdot 3{,}47+22\cdot 3{,}86}{26+30+22}=3{,}51.$$

Свойство 4. Если все значения аргумента $x_1, x_2, \ldots, x_n$ подвергнуть одному и тому же положительному линейному преобразованию $y_i=ax_i+b$, где $a>0$, то и их среднее арифметическое окажется подвергнутым тому же преобразованию.

Действительно, пусть среднее арифметическое чисел $x_1, x_2, \ldots, x_n$ равно $\bar{x}$. Заменим x_i на y_i подстановкой $y_i=ax_i+b$, получим ряд $y_1, y_2, \ldots, y_n$, среднее арифметическое которого $\bar{y}=\dfrac{y_1+y_2+\ldots+y_n}{n}=$

$$=\frac{(ax_1+b)+(ax_2+b)+\ldots+(ax_n+b)}{n}=a\cdot\frac{x_1+x_2+\ldots+x_n}{n}+\frac{b+b+\ldots+b}{n}=a\bar{x}+b.$$

Доказанное свойство позволяет упрощать вычисление среднего арифметического. Например, если в приведённом выше расчёте среднего арифметического уровня учебной мобильности (табл. 2,8) от всех значений аргумента отнять 45, после чего полученные разности разделить на десять, т.е. подвергнуть x_i преобразованию $y_i = (x_i - 45)/10$, то получим значения 0; 1; 2; 3; 4; 5; 6; 7 и, соответственно, распределение, приведенное в таблице 4.4.

Таблица 4.4. Распределение, полученное из распределения приведенного в таблице 2.8, путем преобразования $y_i = (x_i - 45)/10$

Аргументы, x_i	0	1	2	3	4	5	6	7	Итого
Частоты, n_i	2	4	4	8	7	7	5	1	38

В этом случае вычислить среднее арифметическое $\bar{y}$ уже значительно проще: $\bar{y} = (2 \cdot 0 + 4 \cdot 1 + 4 \cdot 2 + 8 \cdot 3 + 7 \cdot 4 + 7 \cdot 5 + 5 \cdot 6 + 1 \cdot 7) / 38 = 3{,}6$. Так как по доказанному свойству $\bar{y} = (\bar{x} - 45)/10$, то $\bar{x} = 10 \cdot \bar{y} + 45$, т.е. $\bar{x} = 10 \cdot 3{,}6 + 45 = 81$.

В заключение заметим, что:

1. Процедуру нахождения средней арифметической можно применить к любой конечной совокупности чисел, однако осмысленное, содержательное его истолкование можно получить только в случае работы в интервальной шкале или шкале отношений.

2. Среднее арифметическое представляет величину того же наименования, что и аргумент совокупности, т.к. относительная частота ω_i – отвлеченные числа.

3. Как правило, каждой средней арифметической можно придать содержательный смысл. Например, если $x_1, x_2, ..., x_k$ обозначают различные значения заработной платы, а n_1, n_2,..., n_k - их частоты, то $\bar{x}$, будет той платой, которую получит каждый, если общий заработок распределить поровну между всеми.

4. Среднее арифметическое зависит от величины признака у каждой единицы наблюдения рассматриваемой совокупности.

5. Среднее арифметическое может иметь двоякий смысл: оно может быть только статистической характеристикой совокупности значений переменной величины, не имеющей одного постоянного значения, например,

средняя зарплата, и приближенным значением постоянной величины, если последняя подвергается ряду измерений.

4.2. Медиана. После среднего арифметического наиболее важным средним является *медиана* (Me). Это понятие впервые встретилось нам в пункте 3.2, где трактовалось как квантиль, делящий совокупность на две равные части. Повторяя по существу это определение, назовем *медианой* такое значение Me признака, для которого число объектов наблюдения, имеющих значения признака $x <$ Me, равно числу объектов наблюдения, имеющих значения $x >$ Me.

Предположим сначала, что рассматривается $2m+1$ значений некоторого количественного признака X: x_1, $x_2,...,x_m$, x_{m+1}, $x_{m+2},...,x_{2m+1}$ расположенных в неубывающем порядке, и при этом выполняется строгое неравенство: $x_m < x_{m+1} < x_{m+2}$. Тогда x_{m+1} будет тем значением X, менее которого в нашем ряду будет m членов: $x_1, x_2,...,x_m$ и более которого будет также m членов: $x_{m+2},\ldots,x_{2m+1}$. Значение x_{m+1} будет серединой ряда, или его *медианой*, т.е. $Me = x_{m+1}$.

Предположим теперь, что дано $2m$ неубывающих значений признака X: $x_1,x_2,...,x_m$, $x_{m+1},...,x_{2m}$, и выполняется строгое неравенство $x_m<x_{m+1}$. Тогда в этом ряду нет серединного члена и за середину его можно принять любое значение x, лежащее между x_m и x_{m+1}, так как ниже и выше его опять будет одинаковое число членов. В этом случае за медиану обычно принимают число $Me = \frac{1}{2}\left(x_m + x_{m+1}\right)$.

Замечание: другие случаи нахождения медианы дискретного ряда рассмотрим после того, как познакомимся с правилом нахождения медианы интервального ряда.

В интервальном ряду с различными значениями частот вычисление медианы распадается на три этапа.

Проиллюстрируем их на примере интервального ряда, приведенного в таблице 4.5.

Таблица 4.5. Распределение первокурсников по уровню учебной мобильности (количеству набранных баллов). Частоты – количество человек

Интервалы	40 50	50 60	60 70	70 80	80 90	90 100	100 110	110 120
Частоты	2	4	4	8	7	7	5	1
Накопленная частота (кумулята)	2	6	10	18	25	32	37	38

Следуя методике вычисления квантилей, приведенной в пункте 3.2, на первом этапе строится распределение накопленных частот начиная с меньших значений величины X, т.е. кумулята. Соответствующие значения приведены в последней строке таблицы 4.5.

На втором этапе находят интервал, которому принадлежит первая из накопленных частот, превышающая половину всего объёма совокупности. В нашем случае таким интервалом является интервал [80;90). Его называют медиальным.

На третьем этапе находят значение медианы по формуле:

$$Me = x_0 + \delta \cdot \frac{0{,}5 \cdot n - n_H}{n_{Me}}, \qquad (4.3)$$

где x_0 – нижняя граница медиального интервала, в нашем случае $x_0 = 80$; δ – величина медиального интервала, в нашем случае $\delta = 90 - 80 = 10$; $n = \sum n_i$ – сумма частот интервалов, в нашем случае $n = 38$; n_H – частота, накопленная до медиального интервала, в нашем случае $n_H = 18$; n_{Me} - частота медиального интервала, в нашем случае $n_{Me} = 7$. Пользуясь формулой 4.3, получим: $Me = 80 + 10 \cdot \frac{19 - 18}{7} \approx 81{,}4$.

Теперь мы можем вернуться к вопросу нахождения медианы дискретного ряда, в случае когда признаком является непрерывная величина или хотя бы величина, которую можно интерпретировать как непрерывную.

Рассмотрим, например, таблицу 4.6, в которой приведены итоги письменной контрольной работы по математике на вступительных экзаменах на математическом факультете КГПУ в 1996 году.

Таблица 4.6. Итоги контрольной работы на вступительных экзаменах по математике на математическом факультете КГПУ в 1996 году

Оценки	2	3	4	5	Всего
Частоты	11	66	29	12	118

Для нахождения медианы преобразуем дискретный ряд оценок в интервальный (табл. 4.7).

Таблица 4.7. Итоги письменной контрольной работы на вступительных экзаменах по математике на математическом факультете КГПУ в 1996 году

Интервалы оценок	[1,5-2,5)	[2,5-3,5)	[3,5-4,5)	[4,5-5,5)	Всего
Частоты	11	66	29	12	118
Накопленная частота	11	77	106	118	

Применяя к этому интервальному ряду введённое выше правило, получим: $Me = 2{,}5 + 1 \cdot \dfrac{59 - 11}{66} = 3{,}23.$

Замечание: медиана не изменяется, если любые значения признака, меньшие медианы, изменить как угодно, не превышая, однако, значения Me. Не изменяется она и тогда, когда значения признака, большие медианы, изменять так, чтобы они не стали меньше Me. В то же время подобные изменения сильно влияют на значения средней арифметической. Это свойство медианы делает её более выгодной как средней в тех случаях, когда концы распределения ненадёжны.

В то же время медиана обладает одним существенным недостатком: она не поддаётся так же легко, как средняя арифметическая, аналитическим операциям, например, когда соединяют два распределения с известными медианами, то ничего нельзя сказать о медиане полученного распределения, её надо вычислять заново для объединённой совокупности.

4.3. Мода. Модой в статистике называют наиболее часто встречающееся значение признака, т.е. значение, с которым наиболее вероятно можно встретиться в серии зарегистрированных наблюдений. Обозначают моду символом Mo. В дискретном ряду мода – это значение признака с наибольшей частотой.

Рассмотрим, например, приведённые в таблицах 4.6 и 4.7 итоги письменной работы на вступительных экзаменах по математике на математическом факультете КГПУ в 1996 году. Легко видеть, что наиболее часто встречающееся значение признака (оценки) – 3. Следовательно, Mo = 3. Для сравнения вспомним, что значение медианы Me = 3,23. Среднее же арифметическое, как легко вычислить, равно $\bar{x} = 3{,}36$.

В интервальном ряду (с равными интервалами) модальным называют интервал с наибольшей частотой. Значение моды находится в его пределах и вычисляется по формуле:

$$Mo = x_0 + \delta \frac{n_{Mo} - n^-}{2n_{Mo} - n^- - n^+}\ , \qquad (4.4)$$

где x_0 – нижняя граница модального интервала; δ – величина интервала; n^- – частота интервала, предшествующего модальному; n^+ – частота интервала, следующего за модальным; n_{Mo} – частота модального интервала.

Подсчитаем, например, моду для интервального распределения, приведённого в таблице 4.7.

$Mo = 2{,}5 + 1 \cdot \dfrac{66 - 11}{2 \cdot 66 - 11 - 29} = 3{,}1$. Ближайшей оценкой является 3.

В совокупностях, в которых может быть произведена лишь операция классификации объектов, по какому-либо качественному признаку, т.е. в номинальных шкалах, вычисление моды является единственным способом указать некий «центр тяжести» совокупности.

К недостаткам моды следует отнести следующее: невозможность совершать алгебраические действия; зависимость её величины от интервала группировки; возможность существования в ряду распределения нескольких модальных значений признака.

Замечания:

1. Распределение может не иметь моды, например, если все значения (интервалы) имеют одинаковую или почти одинаковую частоту.

2. Когда два соседних значения (интервала) имеют одинаковую или почти одинаковую частоту и она больше других, то за моду принимают среднее этих значений (общая граница соседних интервалов).

3. Если два несмежных значения (интервала) имеют равные или почти равные наибольшие частоты, то говорят, что распределение имеет две моды. Такое распределение называют бимодальным.

4. Если у распределения одна мода, то его называют унимодальным.

4.4. Сравнение средних (арифметическое среднее, медиана, мода)

1. Между рассмотренными средними ($\bar{x}$, Me, Mo) существует определённая связь, характер которой зависит от вида распределения и, следовательно, от формы полигона или гистограммы.

Если распределение унимодально и симметрично относительно моды, то $\bar{x} = Me = Mo$.

Для унимодальных, умеренно-асимметричных распределений медиана лежит между средним арифметическим и модой, при этом разность между средним арифметическим и модой примерно в три раза больше разности между средним арифметическим и медианой. Иначе говоря, $\bar{x} - Mo = 3(\bar{x} - Me)$, т.е. медиана в три раза ближе к среднему арифметиче-

скому, чем мода. Из приведенной формулы следует, что, зная две средних, можно найти третью. В качестве примера обратимся к распределению первокурсников по результатам вступительной контрольной работы (табл. 4.7). Для него были получены значения $\bar{x} = 3{,}36$; $Me = 3{,}23$; $Mo = 3$, откуда следует, что $Mo < Me < \bar{x}$. Кроме того, $\bar{x} - Mo = 3{,}36 - 3 = 0{,}36$; $\bar{x} - Me = 3{,}36 - 3{,}23 = 0{,}13$, откуда $\bar{x} - Mo$ примерно в три раза больше $\bar{x} - Me$.

2. Целесообразность использования того и иного типа средней величины зависит, по крайней мере, от следующих условий: цели усреднения; типа величины, характера шкалы; вида распределения; уровня измерения признака; вычислительных соображений.

3. Для ряда с открытыми конечными интервалами процедура вычисления среднего арифметического не применима, однако если распределение близко к симметричному, то можно подсчитать тождественную ему в этом случае медиану. В качестве примера рассмотрим распределение, приведённое в таблице 4.8.

Таблица 4.8. Распределение студентов по размеру денежной материальной помощи, оказываемой им родителями в месяц (по данным опроса, проведённого в 1984 году в КГПУ), в тыс. рублей

Размер помощи	Менее 100	от 100 до 150	от 150 до 200	от 200 до 300	более 300
Частота	10	33	82	37	18
Кумулята	10	43	105	162	180

$$Me = 150 + 50 \cdot \frac{90 - 43}{82} = 179. \qquad Mo = 150 + 50 \cdot \frac{82 - 3}{164 - 33 - 37} = 176.$$

$\bar{x}$ вычислить нельзя, но его приближённое значение можно найти, пользуясь соотношением $\bar{x} - Mo = 3 \cdot (\bar{x} - Me)$, откуда

$$\bar{x} = \frac{3 \cdot Me - Mo}{2} = \frac{3 \cdot 179 - 176}{2} = 180{,}5.$$

4**.** На величину среднего арифметического влияет каждое значение варианты. На медиану не влияют величины «больших» и «малых» значений варианты. В малых группах мода может быть очень нестабильной.

Возможно, наиболее простой характеристикой бимодального распределения будет утверждение, что гистограмма бимодальна и имеет V-образную форму с одной модой при $x = x_1$, а другой при $x = x_2$.

4.5. Среднее значение нормально распределённой случайной величины. Если эмпирический полигон частот сглажен некоторой теоретической кривой распределения $\varphi(x)$, то центральную тенденцию характеризуют величиной, которую называют *математическим ожиданием* и обозначают $M[X]$.

Вычисляется $M[X]$ с помощью несобственного интеграла (при условии его абсолютной сходимости): $M[X]=\int\limits_{-\infty}^{+\infty} x\varphi(x)dx$. В частности, математическое ожидание $M\left[X\right]$ нормально распределённой случайной величины, заданной формулой $\varphi\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, равно $\int\limits_{-\infty}^{+\infty} x\varphi(x)dx$.

Выполнив замену переменной $\frac{x-a}{\sigma}=t$, получим:

$$M[X]=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\sigma\cdot t+\mu\right)\cdot e^{-\frac{t^2}{2}}dt=\frac{\sigma}{\sqrt{2\pi}}\cdot\int\limits_{-\infty}^{+\infty} te^{-\frac{t^2}{2}}dt+\frac{\mu}{\sqrt{2\pi}}\cdot\int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}dt.$$

Первый из двух интегралов правой части, ввиду нечётности подынтегральной функции, равен нулю, второй же интеграл подстановкой $t=u\sqrt{2}$ сводится к $\sqrt{2}\cdot\int\limits_{-\infty}^{+\infty} e^{-u^2}du=\sqrt{2\pi}$. В результате получим, что $M[X]=\mu$.

Таким образом, параметр μ в формуле, задающей плотность нормально распределённой случайной величины X, характеризует её центральную тенденцию, по характеру аналогичную среднему арифметическому.

Глава 5. Мера изменчивости

5.1. Размах как мера изменчивости. Меры центральной тенденции ($\bar{x}$, Me, Mo и другие) говорят нам о концентрации группы значений на числовой оси. Каждая из них даёт значение, которое в некотором смысле «представляет», «замещает» все оценки группы. При этом пренебрегают различиями, существующими между отдельными значениями.

Для измерения изменчивости оценок внутри группы требуются другие меры. В этой главе будет введено и обсуждено несколько таких мер, каждая из которых по-своему оценивает вариативность: изменчивость, неоднородность, разброс значений в группе данных. Многие важные функции статистики связаны с процедурами, позволяющими уменьшить, объяснить или интерпретировать изменчивость, которая, в известном смысле, есть неопределённость.

Всякая научная деятельность связана с понятием изменчивости. Когда есть много необъяснимых причин изменчивости, прогнозы не могут быть точными. Зато когда объяснения причин различий представлены в виде некоторой модели, неопределённость можно уменьшить, а часть вариации устранить. Например, если бы было совсем неизвестно, почему люди различаются между собой по умственному развитию, то попытка прогнозировать интеллект наталкивалась бы на большую неопределённость; некоторые люди выглядели бы «смышлёными», а другие – «глупцами», и никто не знал бы, почему. Если же известно, что наследственность и окружающая среда оказывают существенное влияние на *JQ*, то информация о происхождении ребёнка и его воспитании в раннем детстве позволяет дать более точный прогноз его умственного развития. Другими словами, изменчивость JQ у лиц со сходной наследственностью и окружающей средой меньше, чем у людей вообще.

Одной из простейших характеристик изменчивости является *размах*, который измеряет на числовой шкале расстояние, в пределах которого изменяется варианта. Существуют два вида размаха: исключающий и включающий.

Исключающий размах – это разность максимального и минимального значений варианты в группе.

Включающий размах – это разность между *естественной верхней границей* интервала, содержащего максимальное значение, и *естественной нижней границей* интервала, содержащего минимальное значение.

Например, если рост пяти мальчиков, измеренный с точностью до сантиметра, составил 150, 155, 157, 165, 168 сантиметров, то исключаю-

щий размах равен 168-150=18 сантиметрам. Чтобы найти включающий размах, заметим, что фактический рост самого низкого мальчика заключён между 149,5 и 150,5 сантиметрами, а самого высокого – между 167,5 и 168,5 сантиметрами, откуда следует, что включающий размах равен: 168,5 - 149,5 = 19 сантиметрам.

Размах, определяемый только максимальным и минимальным значениями случайной величины, хотя и используется в качестве меры изменчивости, но меры довольно грубой.

Более чувствительной и в то же время стабильной мерой является *размах от 10-го до 90-го процентиля* (D)**,** который исключает 10 процентов наименьших и 10 процентов наибольших значений совокупности и равен : $D = K_{0,90} - K_{0,10}$.

В психологических исследованиях часто используется так называемый *полумеждуквартильный размах*, равный половине размаха совокупности, из которой исключено 25 процентов наименьших и 25 процентов наибольших значений: $Q = \dfrac{K_{0,75} - K_{0,25}}{2}$.

В качестве примера обратимся к дискретному вариационному ряду, приведённому в таблице 2.2. Так как в нем $x_{\max} = 112$, а $x_{\min} = 44$, то исключающий размах $P_{искл} = 112 - 44 = 68$, включающий размах $P_{вкл} = 112{,}5 - 43{,}5 = 69$. Для вычисления размаха от 10 до 90 процентиля исключим из данного вариационного ряда 10 процентов наибольших и 10 процентов наименьших значений. В оставшейся части ряда $x_{\max} = 104, x_{\min} = 59$, следовательно $D = 104 - 59 = 45$. Для вычисления полумеждуквартильного размаха из той же таблицы исключим четверть самых малых и четверть самых больших значений. После этого в образовавшемся усеченном ряду $x_{\max} = 93$, $x_{\min} = 70$ и, следовательно, $Q = \dfrac{93 - 70}{2} = 11{,}5$.

5.2. Дисперсия. При вычислении трёх рассмотренных выше мер изменчивости не учитывается каждое отдельное значение. Теперь введём четвёртую меру, при вычислении которой, как и при вычислении среднего арифметического, используется каждая оценка. В основу её построения кладут значения отклонений, то есть разностей $x_i - \bar{x}$. Однако, как было выяснено в пункте 4.1 (свойство 1), их сумма, распространённая на все члены ряда $x_1, x_2, \ldots, x_n$, равна нулю: $\sum (x_i - \bar{x}) = 0$. Поэтому перед суммированием каждую из этих разностей возводят в квадрат.

Пусть распределение величины X задано таблицей:

Значения X	x_1	x_2	· · ·	x_k
Частоты n_i	n_1	n_2	· · ·	n_k

Тогда, в соответствии со сказанным выше, в основу построения меры изменчивости можно положить сумму $\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$. Естественно, что чем больше n, тем больше эта сумма. Чтобы исключить влияние объёма совокупности, эту сумму следовало бы разделить n. Однако по причинам, речь о которых будет идти ниже, её делят на $n-1$. Полученное число называют *исправленной дисперсией* и обозначают S^2. Таким образом,

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 . \qquad (5.1)$$

В дальнейшем исправленную дисперсию будем называть просто *дисперсией.*

В качестве примера рассмотрим распределение величины X, заданное таблицей 5.1.

Таблица 5.1. Распределение случайной величины X

Значения X	2	3	4	5	6	7	8	9	10	Итого
Частоты n_i	1	2	4	7	12	7	4	2	1	40

Легко подсчитать, что $\bar{x} = 6$. Тогда $\sum_{i=1}^{9} n_i (x_i - \bar{x})^2 = 1(2-6)^2 + 2(3-6)^2 +$

$+4(4-6)^2 + 7(5-6)^2 + 12(6-6)^2 + 7(7-6)^2 + 4(8-6)^2 + 2(9-6)^2 + 1(10-6)^2 = 114,$

откуда $S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \frac{114}{39} = 2{,}923.$

Вычисление дисперсии по формуле (5.1) при больших значениях n становится затруднительным, поэтому естественно эту формулу преобразовать. Раскрыв скобки и просуммировав слагаемые по индексу i от 1 до k, получим: $S_x^2 = \frac{1}{n-1}\left[\sum n_i x_i^2 - 2\bar{x}\sum n_i x_i + \sum n_i \bar{x}^2\right]$. Если учесть, что $\sum n_i x_i = n\bar{x}$, то, сделав замену и приведя подобные члены, получим:

$$S_x^2 = \frac{1}{n-1}\left(\sum n_i x_i^2 - n\bar{x}^2\right) \qquad (5.2)$$

Пользуясь полученной формулой, вычислим дисперсию распределения, приведённого в таблице 5.2.

Таблица 5.2. Распределение случайной величины X

Варианта, X	2	3	4	5	6	7	8	9	10	Итого
Частота, n_i	1	3	4	8	12	7	3	1	1	40

Для этого составим расчётную таблицу 5.3.

Таблица 5.3. Расчет дисперсии распределения, приведенного в таблице 5.2

x_i	n_i	$n_i x_i$	$n_i x_i^2$
2	1	2	4
3	3	9	27
4	4	16	64
5	8	40	200
6	12	72	432

x_i	n_i	$n_i x_i$	$n_i x_i^2$
7	7	49	343
8	3	24	192
9	1	9	81
10	1	10	100
Итого	40	231	1443

$$\bar{x} = \frac{231}{40} = 5{,}775\ ,\quad S_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{10} n_i x_i^2 - n\bar{x}^2\right] = \frac{1443 - 40\cdot(5{,}775)^2}{40-1} = 2{,}794\cdot$$

Сравним полученное значение S_x^2 с дисперсией распределения, приведенного в таблице 5.4.

Таблица 5.4. Распределение случайной величины X

Варианта, X	2	3	4	5	6	7	8	9	10	Итого
Частота, n_i	2	4	5	6	10	5	4	2	2	40

Составив расчетную таблицу, легко подсчитать, что хотя среднее значений оказалось таким же, как в предыдущем ($\bar{x} = 5{,}775$), дисперсия оказалась в полтора раза большей ($S_x^2 = 4{,}179$). При сопоставлении распределений легко увидеть, что разброс значений во втором случае действительно больше, чем в первом.

Рассмотрим таблицу распределения квадратов значений случайной величины X:

X^2	x_1^2	x_2^2	$\cdots$	x_k^2
n_i	n_1	n_2	$\cdots$	n_k

По правилу вычисления средней арифметической, которую обозначим как $\overline{x^2}$, получим, что $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$, откуда $\sum_{i=1}^{k} n_i x_i^2 = n\overline{x^2}$.

Пользуясь этим обозначением, формулу (5.2) можно представить в виде:

$$S_x^2 = \frac{n}{n-1}\left[\overline{x^2} - \bar{x}^2\right]. \qquad (5.3)$$

При достаточно больших значениях n, например $n > 50$, дисперсию можно вычислять по формуле $S_x^2 = \overline{x^2} - \bar{x}^2$.

Замечание: при использовании компьютера проще всего обратиться к программе Excel. Для этого, занеся в столбец соответствующие данные и обратившись к «мастер функции» f_n, вызовите раздел «статистика», а в нем функцию «дисп».

5.3. Основные свойства дисперсии

Свойство 1. Прибавление к каждому значению x_i одного и того же числа C не меняет дисперсию, т.е. $S_{x+c}^2 = S_x^2$.

Пусть дано распределение:

Варианта	$x_1 + c$	$x_2 + c$	…	$x_k + c$
Частота	n_1	n_2	. . .	n_k

в нем среднее арифметическое $\overline{x+c} = \frac{1}{n}\sum n_i\left(x_i + c\right) = \bar{x} + c$, а дисперсия:

$$S_{x+c}^2 = \frac{1}{n-1}\sum n_i\left[\left(x_i + c\right) - \left(\bar{x} + c\right)\right]^2 = \frac{1}{n-1}\sum n_i\left(x_i - \bar{x}\right)^2 = S_x^2 .$$

Свойство 2. При умножении всех значений x_i на постоянную c дисперсия умножается на c^2, т.е. $S_{cx}^2 = c^2 \cdot S_x^2$.

Пусть дано распределение

Варианта	cx_1	cx_2	…	cx_k
Частота	n_1	n_2	…	n_k

$\overline{cx} = c \cdot \bar{x}$. $\quad S_{cx}^2 = \frac{1}{n-1}\sum n_i\left(cx_i - c\bar{x}\right)^2 = c^2\frac{1}{n-1}\sum n_i\left(x_i - \bar{x}\right)^2$.

Откуда $S_{cx}^2 = c^2 \cdot S_x^2$.

Свойство 3. Пусть две группы характеризуются следующими данными: n_a и n_b – их численность; $\bar{x}_a$ и $\bar{x}_b$ – средние значения; S_a^2 и S_b^2 –

дисперсии. Тогда при их объединении получим группу, число элементов в которой равно $n_a + n_b$, среднее арифметическое равно $\bar{x} = \frac{n_a \bar{x}_a + n_b \bar{x}_b}{n_a + n_b}$, а дисперсия $S^2 = \frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2 + n_a(\bar{x}_a - \bar{x})^2 + n_b(\bar{x}_b - \bar{x})^2}{n_a + n_b + 1}$.

5.4. Стандартное отклонение и коэффициент вариации. Мерой изменчивости, тесно связанной с дисперсией, является стандартное, или среднее квадратичное, отклонение.

Стандартным отклонением (S_x) случайной величины X называется положительное значение квадратного корня из дисперсии, т.е. $S_x = \sqrt{S_x^2}$, откуда, учитывая формулу (5.3), получим:

$$S_x = \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)} \ . \qquad (5.4)$$

Стандартное (среднее квадратичное) отклонение является мерой абсолютной колеблемости признака и всегда выражается в тех же единицах измерения, в которых выражен изучаемый признак. Это не позволяет сопоставлять между собой стандартные отклонения различных признаков (в случае разных единиц измерения) в одной и той же совокупности, а так же одного и того же признака в разных совокупностях с различными средними.

Чтобы иметь такую возможность, среднее отклонение часто выражается через соотнесение в процентах к среднему арифметическому. Отношение стандартного отклонения к среднему арифметическому, выраженное в процентах, называется *коэффициентом вариации***:**

$$V = \frac{S_x}{\bar{x}} \cdot 100\%. \qquad (5.5)$$

Этот показатель применим к количественным признакам, измеренным по интервальной шкале или шкале отношений. Его применение при пользовании номинальными или ранговыми шкалами требует тщательной интерпретации.

В качестве примера найдём стандартное отклонение и коэффициент вариации вариационных рядов, приведенных в таблицах 5.2 и 5.4.

Так как в первом случае $S_x = \sqrt{2{,}794} = 1{,}671$, а $\bar{x} = 5{,}775$, то

$$V_x = \frac{1{,}671}{5{,}775} \cdot 100\% = 28{,}9\% .$$

Во втором случае $S_x = \sqrt{4{,}179} = 2{,}004$, а $\bar{x} = 5{,}775$, значит,

$$V_x = \frac{2{,}044}{5{,}775} \cdot 100\% = 35{,}4\% .$$

5.5. Среднее абсолютное отклонение. Ещё одна мера изменчивости – *среднее абсолютное отклонение* – выражается легче, чем стандартное, но используется реже.

Средним абсолютным отклонением M_D называется среднее значение n расстояний оценок x_i от их среднего $\bar{x}$: $M_D = \frac{1}{n} \cdot \sum_{i=1}^{k} n_i \left| x_i - \bar{x} \right|$.

Для среднего абсолютного отклонения нет более простого выражения, которое можно было бы использовать при вычислениях. В качестве примера найдём среднее абсолютное отклонение распределения, приведённого в таблице 5.4. Расчёт приведён в таблице 5.5.

Таблица 5.5. Расчет среднего абсолютного отклонения распределения, приведенного в таблице 5.4

x_i	n_i	$\left\|x_i - \bar{x}\right\|$	$n_i \left\|x_i - \bar{x}\right\|$
2	2	3,775	7,550
3	4	2,775	11,100
4	5	1,775	8,875
5	6	0,775	4,650
6	10	0,225	2,250

x_i	n_i	$\left\|x_i - \bar{x}\right\|$	$n_i \left\|x_i - \bar{x}\right\|$
7	5	1,225	6,125
8	4	2,225	8,900
9	2	3,225	6,450
10	2	4,225	8,450
Итого	40	20,225	64,350

$$M_D = \frac{64{,}35}{40} = 1{,}61 .$$

Среднее абсолютное отклонение, несмотря на логическую простоту, как мера изменчивости используется редко. Одна из причин этого состоит в том, что среднее абсолютное отклонение не имеет теоретического обоснования, в отличие, например, от дисперсии.

Величина стандартного отклонения S_x всегда больше M_D и для достаточно большой совокупности с распределением признака, близкого к *нормальному*, равна $1{,}25 \cdot M_D$.

5.6. Меры изменчивости при работе с номинальными шкалами. При работе с порядковыми и тем более номинальными шкалами нельзя пользоваться такими показателями, как дисперсия и стандартное отклонение. Для таких статистических рядов разработаны иные показатели, харак-

теризующие степень рассеяния, вариативности. В частности, к ним относятся:

1. Индекс качественной вариации (J).

2. Энтропия (H).

Если признак имеет k взаимоисключающих градаций, то для вычисления индекса качественной вариации применяется процедура, которую мы поясним следующим примером.

Пусть, отвечая на некоторый вопрос, 120 студентов распределились на три непересекающиеся группы: A, B и C. В группе A оказалось 30 человек, в группе B – 20 человек, в группе C – 70 человек.

Это распределение сопоставляют с равномерным распределением тех же 120 человек по указанным трём группам A, B и C, при котором в каждую группу попадает по 40 человек. Получим таблицу 5.6.

Таблица 5.6. Расчет индекса качественной вариации

Группы	A	B	C	Итого
Фактическое	30	20	70	120
Равномерное	40	40	40	120

Затем вычисляют величину $J = \frac{30 \cdot 20 + 30 \cdot 70 + 20 \cdot 70}{40 \cdot 40 + 40 \cdot 40 + 40 \cdot 40} = 0{,}85$, или 85%.

Эту величину называют индексом качественной вариации. Она указывает на степень неоднородности полученных ответов. Если бы все ответы попали лишь в одну градацию, то J = 0, что означало бы полное единство в ответах. С аналогичной ситуацией мы встречаемся, используя количественные шкалы: если все $x_i \equiv \bar{x}$, то и S_x^2 и S_x равны нулю. В то же время, если полученное в результате эксперимента распределение окажется равномерным, т.е. если во все градации попадает одно и то же количество ответов ($\frac{n}{k}$, где n – общее число респондентов, а k – число градаций), то $J = 1$, или 100%.

В качестве второго примера рассмотрим мнение школьников 7–10 классов о спортсменах. В частности, их спрашивали, согласны ли они с тем, что спортсмен – это хорошо физически развитый, здоровый человек, но вместе с тем обладающий высоким самомнением и низким уровнем духовной (нравственной, интеллектуальной, эстетической и т.д.) культуры. Школьники, выражая своё мнение о спортсменах, разделились на три группы. В первую из них (группу A) включены школьники, в полной мере разделяющие приведённое выше мнение. В этой группе оказалось 488 че-

ловек. Во вторую группу (группу В) включены школьники, которые в какой-то мере, но не полностью разделяют это мнение. В этой группе оказалось 592 человека. В третью группу (группу С) включены школьники, которые думают иначе (табл. 5.7).

Таблица 5.7. Отношение школьников к высказанному выше утверждению

Группы	А	В	С	Всего
Фактическое распределение	488	592	848	1928
Равномерное распределение	643	643	642	1928

Индекс качественной вариации (J) равен:

$J = \frac{488 \cdot 592 + 488 \cdot 848 + 592 \cdot 848}{643 \cdot 643 + 643 \cdot 642 + 643 \cdot 642} = 0{,}97$, или 97%.

Это сравнительно высокий показатель неоднородности в мнениях школьников по данному вопросу.

Энтропия. Одной из мер вариации признака, не зависящей от уровня измерения (типа применяемой шкалы), служит так называемая *энтропия* - мера неопределённости, вычисляемая по формуле:

$$H = \lg n - \frac{1}{n} \cdot \sum_{i=1}^{k} n_i \lg n_i , \qquad (5.6)$$

где k - число взаимоисключающих градаций, а $n = \sum_{i=1}^{K} n_i$.

Возвращаясь к последнему примеру, где $k = 3$, $n = 1928$, $n_1 = 488$, $n_2 = 592$, $n_3 = 848$, получим:

$H = \lg 1928 - \frac{1}{1928} \cdot (488 \cdot \lg 488 + 592 \cdot \lg 592 + 848 \cdot \lg 848)$, откуда, логарифмируя, получим, что $H = 3.285 - \frac{1}{1928} \cdot (1311.9 + 1641.2 + 2483.3) \approx 0.4654$.

Возникает вопрос, в каких пределах изменяется H, с чем его сравнивать. Велико или мало полученное выше значение?

Из формулы $H = \lg n - \frac{1}{n} \sum_{i=1}^{k} n_i \cdot \lg n_i$, определяющей значение энтропии, следует, что она равна нулю лишь в том случае, когда все респонденты попадают в одну из k групп. Во всех остальных случаях, когда имеется та или иная неопределённость в значениях x_i, энтропия является положи-

тельной величиной. Наибольшее значение энтропия принимает в том случае, когда респонденты равномерно распределены по всем k группам т.е. $n_i = \frac{n}{k}$. Для признака с k градациями $H_{\max} = \lg n - \frac{1}{n} \cdot \sum_{i=1}^{k} \frac{n}{k} \cdot \lg \frac{n}{k}$

Выполнив необходимые преобразования, получим:

$$H_{\max} = \lg n - \frac{1}{k} \cdot k \cdot \lg \frac{n}{k} = \lg n - (\lg n - \lg k) = \lg k .$$

Таким образом, значение H надо сравнивать с $H_{\max}$, которое увеличивается с ростом числа градаций в признаке. Как правило, за меру разброса, вариативности принимают показатель $\eta = \frac{H}{H_{\max}}$.

В рассмотренном выше примере $H_{\max} = \lg 3 = 0{,}4771$ и, следовательно, $\eta = \frac{H}{H_{\max}} = \frac{0{,}4654}{0{,}4771} = 0{,}98$ или 98%. Как видим, он практически совпадает с вычисленным ранее индексом качественной вариации $J = 0{,}97$, или 97%.

Глава 6. Центральные моменты распределения

6.1. Асимметрия распределения. Обратимся теперь к характеристикам других свойств распределений с количественными аргументами. Важнейшими из них являются *асимметрия* и *эксцесс* распределения.

Средние представляют меры положения и дают те точки, около которых концентрируются значения аргумента в распределении. Дисперсия, стандартное отклонение, среднее абсолютное отклонение представляют меры рассеяния значений аргумента. Следующей важной характеристикой распределения является его **асимметрия**, величина, характеризующая степень несимметричности рассеяния значений по одну и по другую сторону от средней. Для её измерения служат различные меры. Рассмотрим одну из них. Пусть дано распределение

Варианта	x_1	x_2	…	x_k	Итого
Частота	n_1	n_2	…	n_k	n

Составим величину $\mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^3$, где $n = \sum_{i=1}^{k} n_i$. Она представляет средний куб отклонений $x_i - \bar{x}$ и обращается в нуль, если распределение симметрично. Если же распределение асимметрично, то значениям аргумента, равноотстоящим по обе стороны от средней, будут соответствовать разные частоты и μ_3 будет либо больше, либо меньше нуля и притом будет отличаться от нуля тем более, чем сильнее выражена асимметрия. Таким образом, μ_3 может служить мерой асимметрии распределения. Чтобы сделать её независимой от характера распределяемой величины, её делят на куб стандартного отклонения. Число $k = \frac{\mu_3}{S_x^3}$ называют **асимметрией**, или **косостью** распределения.

Таким образом, коэффициент асимметрии:

$$k = \frac{1}{nS_ч^3}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^3 . \qquad (6.1)$$

Расчеты выполняются с помощью таблиц, аналогичных тем, которыми мы пользовались при вычислении дисперсии.

Вычислим, например, коэффициент асимметрии K, распределения, заданного таблицей

Варианта	13	14	15	16	17	18	19	20	21	22
Частоты	1	9	14	10	4	3	2	2	1	1

Таблица 6.1. Расчет коэффициента асимметрии, приведенного выше распределения

x_i	n_i	$n_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$n_i(x_i - \bar{x})^3$
13	1	13	-3	9	9	-27	-27
14	9	126	-2	4	36	-8	-72
15	14	210	-1	1	14	-1	-14
16	10	160	0	0	0	0	0
17	4	68	1	1	4	1	4
18	3	54	2	4	12	8	24
19	2	38	3	9	18	27	54
20	2	40	4	16	32	64	128
21	1	21	5	25	25	125	125
22	1	22	6	36	36	216	216
Итого	47	752			186		438
Среднее		16			4,043		

Используя результаты вычислений, получим:

$$n = 47,\ \ \bar{x} = \frac{752}{46} = 16,\ \ S_x^2 = \frac{186}{46} = 4{,}0435,\ S_x = 2{,}01.$$

$$k = \frac{1}{nS_x^3}\sum_{i=1}^{k} n_i(x_i - \bar{x})^3 = \frac{1}{47 \cdot 8.13} \cdot 438 = 1.15\,.$$

Так как $K = 1{,}15 > 0$, то распределение скошено вправо (рис.6.1).

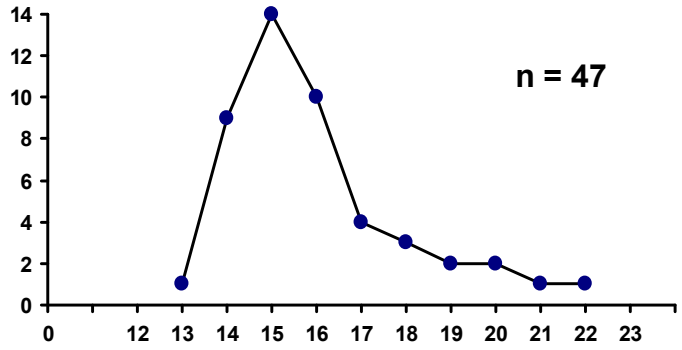

Рис. 6.1

Пример 2. Вычислим асимметрию (косость) распределения, заданного таблицей:

Варианта	11	12	13	14	15	16	17	18	19	20
Частоты	1	1	2	2	3	4	10	14	9	1

Таблица 6.2. Расчет коэффициента асимметрии приведенного выше распределения

x_i	n_i	$n_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^3$	$n_i(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$n_i(x_i - \bar{x})^3$
11	1	11	-6	36	36	-216	-216
12	1	12	-5	25	25	-125	-125
13	2	26	-4	16	32	-64	-128
14	2	28	-3	9	18	-27	-54
15	3	45	-2	4	12	-8	-24
16	4	64	-1	1	4	-1	-4
17	10	170	0	0	0	0	0
18	14	252	1	1	14	1	14
19	9	171	2	4	36	8	72
20	1	20	3	9	9	27	27
Итого	47	799			186		-448
Среднее		17			4,04		

Используя результаты вычислений, получим:

$n = 47, \bar{x} = 17, S^2 = \dfrac{186}{46} = 4{,}04, S_x = 2{,}01$. Выполнив преобразование

$$k = \frac{1}{nS_x^3}\sum_{i=1}^{k} n_i(x_i - \bar{x})^3 = \frac{1}{47 \cdot 8.13} \cdot (-448) = -1.17 < 0.$$

Так как $K = -1.17 < 0$, то распределение скошено влево (рис. 6.2).

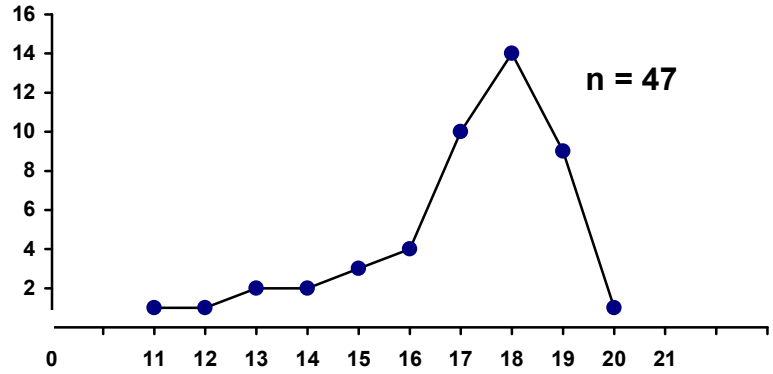

Рис.6.2

Для ускоренного подсчёта асимметрии иногда пользуются формулой

$$K \approx \frac{3(\bar{x} - M_e)}{S_x}, \text{ где } M_e - \text{медиана.} \qquad (6{,}2)$$

6.2. Центральные моменты распределения. Эксцесс распределения. Рассмотренные статистики описывают три свойства распределения: центральную тенденцию, изменчивость, симметрию. Все они в той или

иной мере связаны с понятием *центрального момента*, к которому мы теперь и обратимся. Пусть дано распределение случайной величины X

Варианта	x_1	x_2	…	x_m
Частота	n_1	n_2	…	n_m

Центральным моментом *k-го порядка* случайной дискретной величины X называют величину $\mu_k = \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^k$. Степень k называют порядком момента.

Первый центральный момент ($k=1$) любого распределения, т.е. $\mu_1 = \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^1$, равен нулю, ибо, как было показано в третьей главе, сумма всех отклонений $\left(\sum n_i (x_i - \bar{x})\right)$ всегда равна нулю. Естественно, что первый момент не представляет особого интереса.

Второй центральный момент $\mu_2 = \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^2$ связан с дисперсией.

Действительно $S_x^2 = \frac{\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}{n-1} = \frac{n}{n-1} \cdot \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^2$,

откуда $S_x^2 = \frac{n}{n-1} \cdot \mu_2$ или $\mu_2 = \frac{n-1}{n} \cdot S_x^2$.

Третий центральный момент $\mu_3 = \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^3$ связан с асимметрией, а именно:

действительно $K = \frac{1}{n \cdot S_x^3} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^3 = \frac{1}{S_x^3} \cdot \frac{1}{n} \cdot \sum_{i=1}^{m} n_i (x_i - \bar{x})^3$,

откуда $K = \frac{\mu_3}{S_x^3}$ или $\mu_3 = K \cdot S_x^2$

Рассмотрим *четвёртый центральный момент* $\mu_4 = \frac{1}{n} \cdot \sum_{i=1}^{k} n_i (x_i - \bar{x})^4$.

Если его разделить на S_x^4, то получим величину $\frac{\mu_4}{S_x^4}$, которая в случае унимодального распределения характеризует островершинность кривой. Для

нормального закона распределения $\frac{\mu_4}{S_x^4} = 3$. Для других распределений $\frac{\mu_4}{S_x^4}$ может быть больше или меньше 3. Величину $\frac{\mu_4}{S_x^4} - 3$ называют *эксцессом* распределения. Таким образом,

$$Ek = \frac{1}{n \cdot S_x^4} \cdot \sum_{i=1}^{m} n_i \cdot \left(x_i - \bar{x}\right)^4 - 3. \qquad (6.3)$$

Вычисление эксцесса распределения производится в таблицах, аналогичных тем, по которым мы рассчитывали коэффициент асимметрии.

В качестве примера обратимся к распределению случайной величины, заданной в таблице 6.1.

Таблица 6.3. Расчет эксцесса распределения, приведенного в таблице 6.1

x_i	n_i	$n_i x_i$	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$	$n_i\left(x_i - \bar{x}\right)^2$	$\left(x_i - \bar{x}\right)^4$	$n_i\left(x_i - \bar{x}\right)^4$
13	1	13	-3	9	9	81	81
14	9	126	-2	4	36	16	144
15	14	210	-1	1	14	1	14
16	10	160	0	0	0	0	0
17	4	68	1	1	4	1	4
18	3	54	2	4	12	16	48
19	2	38	3	9	18	81	162
20	2	40	4	16	32	256	512
21	1	21	5	25	25	625	625
22	1	22	6	36	36	1296	1296
Итого	47	752			186		2886
Среднее		16			4,043		

Используя результаты вычислений, получим:

$$n = 47,\ \bar{x} = \frac{752}{46} = 16,\ S_x^2 = \frac{186}{46} = 4{,}0435,\ S_x = 2{,}01.$$

$$Ek = \frac{1}{nS_x^4}\sum_{i=1}^{m} n_i\left(x_i - \bar{x}\right)^4 - 3 = \frac{1}{47 \cdot 16{,}35} \cdot 2886 - 3 = 3{,}756 - 3 = 0{,}756.$$

Если *эксцесс* больше нуля, как в нашем случае, то распределение более островершинное, чем нормальное. Если меньше нуля, то более плосковершинное. На рисунке 6.3 приведены три вида кривых распределения.

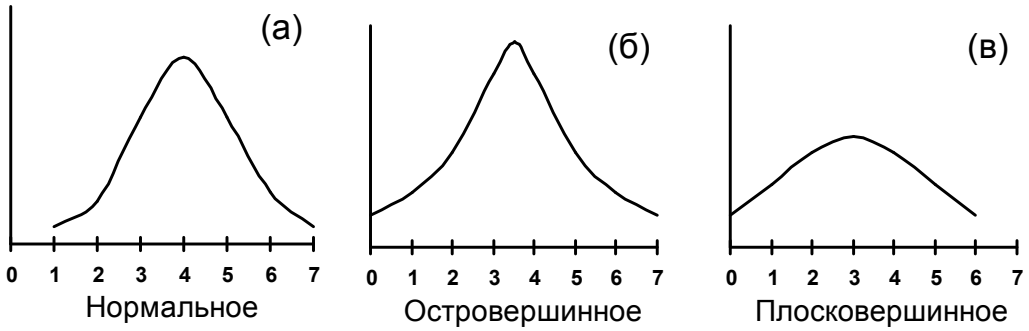

Рис. 6.3

6.3. Центральные моменты нормальнораспределённой случайной величины. Переходя к случаю непрерывной случайной величины, заметим, что *центральный момент* k - го порядка случайной непрерывной величины, заданной своей плотностью $p(x)$, вычисляется по формуле $\mu_k = \int\limits_{-\infty}^{+\infty} (x-m)^k p(x)dx$. Естественно, при условии существования математического ожидания m. Если эмпирический полигон частот сглажен некоторой теоретической кривой распределения с плотностью $p(x)$, то дисперсия как мера изменчивости вычисляется с помощью интеграла $\int\limits_{-\infty}^{+\infty} (x-m)^2 p(x)dx$, где $m-$ математическое ожидание случайной величины X. В частности, если сглаживание полигона частот выполнено нормальной кривой, то дисперсия будет равна. $D[x] = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$.

Произведя замену $\frac{x-\mu}{\sigma} = t$, получим $D[x] = \frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt$. Интегрируя по частям, получим $D[x] = \sigma^2$. При больших значениях n дисперсия нормально распределённой случайной величины $S_x^2 \approx \sigma^2$, а $S_x = \sigma$.

Таким образом, если есть серьёзное основание считать, что рассматриваемое эмпирическое распределение может быть сглажено нормальной кривой с плотностью $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$, то в качестве значений параметров a и σ можно взять $\bar{x}$ и S_x.

В качестве примера обратимся к гипотетическому распределению первокурсников по уровню учебной мобильности, приведённому в таблице 3.1. Соответствующая кривая распределения, приведённая на рисунке 3.2, по форме напоминает нормальную кривую. Пользуясь данными, приведёнными в таблице 3.1, вычислим арифметическое среднее и стандартное отклонение: $\bar{x} \approx 78$, $S_x^2 \approx 225, S_x \approx 15$. Полагая $\mu = 78$ и $\sigma = 15$, получим

$$\varphi(x) = \frac{1}{15\sqrt{2\pi}} \cdot e^{-\frac{(x-78)^2}{2 \cdot 225}} = 0{,}0266 \cdot e^{-\frac{(x-78)^2}{450}}.$$

Учитывая, что при нормальном распределении дискретной случайной величины по шкале интервалов, относительная частота ω_i попадания ее значений в i-ый интервал равен, с одной стороны произведению $\varphi(x_i)$ на длину интервала l (приближенно), а с другой стороны отношению $\tilde{n}_i / n$ где $\tilde{n}_i$ число элементов распределения, попадающих в i- ый интервал, получим

$$\tilde{n}_i = nl \cdot \varphi(x_i) = nl \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \qquad (6.4)$$

где n – общая численность совокупности.

В нашем случае $n = 100$, $l = 5$, $\sigma = 15$, $\mu = 78$, поэтому

$$\tilde{n}_i \approx 13{,}3 \cdot e^{-\frac{(x_i-78)^2}{450}}.$$

Выясним, в какой мере это теоретическое распределение отличается от эмпирического. Подсчитывая значения $\tilde{n}_i$, получим теоретическое распределение, приведённое в таблице 6.4.

Таблица 6.4. Сопоставление эмпирических и расчетно-теоретических данных

Интервалы	40 45	45 50	50 55	55 60	60 65	65 70	70 75	75 80	80 85	85 90	90 95	95 100	100 105	105 110	110 115	115 120
$\tilde{n}_i$	0,8	1,7	3,1	5,2	7,8	10,4	12,5	13,3	12,7	10,9	8,4	5,7	3,5	1,9	0,9	0.4
n_i	1	2	4	5	7	10	12	14	13	11	9	5	3	2	1	1

Как видим, отклонения не очень велики. Во второй части пособия мы сможем оценить значимость этих различий.

6.4. Стандартизация данных. При обработке и анализе данных часто бывает удобным описать место некоторого значения в совокупности, измеряя его отклонение от среднего арифметического в единицах стандартного отклонения. Пусть, например, дана совокупность из двенадцати значений: 32, 34, 37, 38, 38, 40, 40, 40, 42, 42, 47, 50, среднее арифметическое которого $\bar{x} = 40$, а стандартное отклонение $S = 5$. Характеризуя положение значения 34 среди других значений, можно сказать, что оно находится на 6 единиц ниже среднего и что это расстояние составляет $6/5 = 1{,}2$ единиц стандартного отклонения.

Для произвольного x_i в этой совокупности значений его отклонение от среднего, измеренное в единицах стандартного отклонения, может быть вычислено по формуле: $z_i = \dfrac{x_i - 40}{5}$. Применяя эту формулу к каждому значению приведённого выше вариационного ряда, получим ряд:

- 1,6; - 1,2; - 0,6; - 0,4; - 0,4; 0; 0; 0; 0,4; 0,4; 1,4; 2.

Его среднее арифметическое будет равно нулю, $\bar{z} = 0$, а дисперсия $S_z^2 = \dfrac{10{,}96}{11} \approx 1$, а следовательно, $S_z \approx 1$.

Рассмотрим общий случай. Пусть дан вариационный ряд $x_1, x_2, \ldots, x_n$, у которого среднее арифметическое равно $\bar{x}$, а стандартное отклонение S_x. Подвергнем члены данного вариационного ряда преобразованию: $z_i = \dfrac{x_i - \bar{x}}{S_x}$. Получим так называемый стандартизированный вариационный ряд: $z_1, z_2, \ldots, z_n$.

Покажем, что его среднее арифметическое $\bar{z} = 0$, а дисперсия S_z^2, а значит, и стандартное отклонение S_z, равны 1. Действительно, в этом случае $\bar{z} = \dfrac{1}{n}\sum_{i=1}^{n} z_i = \dfrac{1}{n}\dfrac{1}{S_x}\sum\left(x_i - \bar{x}\right) = 0$, т.к. $\sum_{1}^{n}(x_i - \bar{x}) = 0$,

$$S_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2 = \frac{1}{n-1}\sum_{1}^{n} z_i^2 = \frac{1}{n-1}\sum_{1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)^2 = \frac{1}{S_x^2}\cdot\frac{1}{n-1}\sum_{1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{S_x^2}\cdot S_x^2 = 1.$$

Стандартизация данных – не только удобное средство информации о положении отдельных значений коллектива относительно арифметического среднего, измеренного в единицах стандартного отклонения, но и шаг вперёд к преобразованию множества X в произвольную шкалу с удобными характеристиками среднего и стандартного отклонения. Конечно, сами оценки z могут не подходить для некоторых целей. Неудобными могут

оказаться отрицательные или дробные значения z. Однако преобразование z с помощью формулы $t_i = cz_i + d$ при надлежащем выборе коэффициентов c и d позволяет перейти к более удобным значениям $t_1, t_2, ..., t_n$. При этом среднее арифметическое $\bar{t} = d$, дисперсия $S_t^2 = c^2$, а стандартное отклонение $S_t = |c|$.

Применительно к нашему случаю имеет смысл воспользоваться преобразованием $t_i = 10 \cdot z_i + 20$. Тогда получим ряд:

4; 8; 14; 18; 18; 20; 20; 20; 24; 24; 34; 40,

в котором $\bar{t} = 20$, $S_t^2 = 100$, $S_t = 10$.

Существует множество шкал измерения с определёнными средними арифметическими и стандартными отклонениями, которые распространены в педагогике и общественных науках. Например, оценки интеллектуального теста часто преобразуются в шкалу со средним 100 и стандартным отклонением 15 или 16. Широкое применение находят преобразования $t = 10 \cdot z + 50$. Эти и другие распространённые шкалы будут рассматриваться в следующих главах.

Рассмотрим, наконец, какой вид приобретёт нормальное распределение после его стандартизации. Применив преобразование $z = \frac{x - a}{\sigma}$ и учитывая, что после стандартизации $M[z] = 0$, а $D[z] = 1$, т.е. $a = 0$, $\sigma = 1$, получим, что $\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$, и, следовательно, все нормальные распределения после стандартизации приводятся к одному и тому же стандартному виду. В этом случае вероятность того, что нормально распределённая случайная величина X со средним a и дисперсией σ^2 примет значение, принадлежащее промежутку $[\alpha, \beta]$, может быть вычислена по формуле: $\frac{1}{\sqrt{2\pi}} \cdot \int_t^T e^{-\frac{z^2}{2}} dz$, где $t = \frac{\alpha - a}{\sigma}$, $T = \frac{\beta - a}{\sigma}$. Для удобства вычислений составлены таблицы значений функций $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$ и

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{1}{2}t^2} dt$, которые даны в Приложении (табл. N Приложения). Пользуясь этими таблицами, теоретическая частота $\tilde{n}_i$ подсчитывается по формуле $\tilde{n}_i = n \cdot l \cdot \varphi(z_i)$, где n – число элементов коллектива, l – шаг таблицы. Количество элементов, содержащихся между $z = t$ и $z = T$, будет равно $n \cdot \frac{1}{\sqrt{2\pi}} \int_t^T e^{-\frac{z^2}{2}} dz = n \cdot [\Phi(T) - \Phi(t)]$.

В качестве примера обратимся снова к нашему гипотетическому распределению первокурсников по уровню учебной мобильности, приведённому в таблице 3.1. Как мы видели (см. пункт 6.1), $\bar{x} = 78$ $S_x^2 \approx 225$, $S_x = 15$. Нормализуя случайную величину X (центры интервалов) с помощью преобразования $z = \frac{x - 78}{15}$, получим приведённый в таблице 6.1 ряд значений z, для которого $\bar{z} = 0$, $S_z = 1$.

Таблица 6.5. Распределение первокурсников по уровню учебной мобильности

z	-2,37	-2,03	-1,70	-1,37	-1,03	-0,70	-0,37	-0,03	0,30	0,63	0,97	1,30	1,63	1,97	2,30	2,63
$\tilde{n}_i$	0,8	1,7	3,1	5,2	7,9	10,5	12,5	13,4	12,8	11,0	8,3	5,7	3,6	1,9	0,9	0.4

Плотность соответствующего нормального распределения будет $\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}z^2}$, а теоретическая частота $\tilde{n}_i = n \cdot l \cdot \varphi(z_i) =$ $= 100 \cdot 0{,}335 \cdot \varphi(z_i) = 33{,}5 \cdot \varphi(z_i)$. Значения $\varphi(z_i)$ можно находить по таблице N Приложения. Число элементов совокупности, для которых z лежит между $z = -0{,}70$ и $z = 1{,}30$, будет равно $100 \cdot [\Phi(1{,}30) - \Phi(-0{,}70)]$ $= 100 \cdot [\Phi(1{,}30) + \Phi(0{,}70)] = 100 \cdot [0{,}4032 + 0{,}2580] = 66{,}12$.