РАЗДЕЛ II. АГРЕГИРОВАНИЕ СОЦИОЛОГИЧЕСКИХ ДАННЫХ И СООТВЕТСТВУЮЩИЕ МАТЕМАТИЧЕСКИЕ ПРОЦЕДУРЫ

2.1. Частотные распределения

Агрегирование – укрупнение тех или иных показателей посредством их объединения в единую группу, то есть сведение частных показателей к обобщенным. В результате агрегирования социологических данных исследователь получает важные «синтетические измерители», объединяющие в себе множество частных показателей. Агрегирование осуществляется посредством суммирования, группировки или других способов сведения частных показателей в обобщенные. Измеряя характеристики объекта, исследователь собирает первичный статистический материал, который в процессе агрегирования систематизируется и обобщается для выявления характерных черт, свойств, типов социальных явлений, для обнаружения закономерностей изучаемых процессов и проверки исследовательских гипотез. В основе используемых методов обработки первичных данных, их агрегирования, упорядочения лежит, главным образом, статистическая группировка, а также составление статистических таблиц [1, с. 32].

Как говорилось в предыдущем разделе, проводя исследование, социолог изучает признаки, которые характеризуют исследуемую совокупность. На основе анализа этих признаков исследователь делает выводы относительно того или иного социального явления, процесса и т. п.

Признак – случайная величина, которая варьируется, то есть имеет определенное число вариаций, принимает различные значения. Для каждого значения этой величины, по прохождении полевого этапа исследования, будет известна частота встречаемости. Другими словами социологу будет известно одномерное распределение вероятностей, которые задают эту случайную величину.

Результаты измерения, как правило, изначально имеют произвольный и хаотичный порядок. В такой форме полученные данные неудобны для анализа и выявления закономерностей. Первичная обработка статистических данных состоит в упорядочении данных (по возрастанию или убыванию), подсчете некоторых показателей, характеризующих эти значения, в группировании данных.

Ряды распределения − это ряды абсолютных и относительных чисел, которые характеризуют распределение единиц совокупности по качественному (атрибутивному) или количественному признаку. Зарегистрированные в результате наблюдения индивидуальные значения изучаемого варьирующего признака образуют так называемый **первичный ряд.**

Ряд значений признака, или вариант, полученных вследствие массового обследования однородных вещей или явлений, размещенных в порядке возрастания или убывания их величин, вместе с соответствующими частотами

(или относительными частотами) называют **вариационным рядом.**

Если в вариационном ряде значения признака (варианты) заданы в виде отдельных конкретных чисел, то такой ряд называют **дискретным.**

Если в вариационном ряде значения признака заданы в виде интервалов, то такой ряд называют **интервальным**.

Еще есть такое понятие, как «**динамический** (или временной) ряд», показывающее движение признака (изменение его частот) во времени, то есть изменение его в связи с переходом от одного момента или периода времени к следующему. Изучение динамических рядов позволяет установить закономерность в развитии данного явления или признака, определить складывающиеся тенденции и выявить различные колебания, вариации, отклонения.

Число случаев, в которых встречается то или иное значение признака (варианта), называют **абсолютной частотой** этого значения. Результаты социологического исследования фиксируются как статистические наблюдения, которые регистрируются прежде всего в форме первичных абсолютных величин или абсолютных частот.

Например, если мы изучаем распределение респондентов по профессиональному признаку (качественному), для конкретной выборки, объем которой равен 80 человек (n = 80), то ряд распределения абсолютных частот может выглядеть следующим образом:

1. Учитель − 10 человек.
2. Врач − 11 человек.
3. Повар − 9 человек.
4. Водитель − 20 человек.
5. Продавец − 30 человек.

Абсолютные частоты в качестве единицы измерения всегда имеют «штук», «человек», «часов», «килограмм» и т. п.

Но более удобным для восприятия и дальнейшего анализа (например, сравнительного), является представление данных в виде относительных частот. **Относительная частота** − доля или процент объектов, обладающих данным значением признака, по отношению к объему выборки.

Абсолютные частоты, с помощью элементарных математических вычислений из школьной программы, легко переводятся в относительные. В случае с примером, приведенным выше, можем иметь следующую картину:

1. Учитель − 12,5 %.
2. Врач − 13,75%.
3. Повар − 11,25%.
4. Водитель − 25%.
5. Продавец – 37,5%.

Если значения признаков выражены в относительных числах, то эти значения именуются **частостями**. Распределение по признаку профессиональной принадлежности − это распределение по качественному

признаку. Однако эта же совокупность может быть распределена по количественному признаку, например, по возрасту. Диапазон возможных вариант такой переменной, как возраст, очень широк. В предыдущем разделе мы говорили о том, что эту переменную можно представить в виде метрической (числовой) шкалы, а градации еще и выстроить по порядку. Такое представление нам может быть необходимо для того, чтобы расширить спектр возможных математических операций с этой шкалой[10]. Однако вспомним, что представленная таким образом переменная возраета это непрерывная количественная переменная. Она может принимать огромное множество различных значений. Чем больше число этих значений, тем сложнее человеческому мозгу справиться с их обработкой. Например, мы с легкостью может представить трех- или четырехугольник. Однако «двенадцатиугольник» уже представить довольно сложно, а стоугольник – вообще невозможно. Хотя каждый из нас может вполне успешно описать эти фигуры, дать им определение. А для того чтобы мы все-таки смогли представить, вообразить объект исследования, необходимо воспользоваться теми или иными методами агрегирования данных, то есть их «укрупнения» посредством объединения по группам. Агрегирование, как правило, осуществляется посредством группировки, что позволяет информацию об объекте исследования представить в «компактной» форме, такой, которая легко поддавалась бы нашему восприятию и воображению. Для этого осуществляется группировка данных.

Группировка – *разбиение* совокупности на группы, однородные по какому-либо признаку (а с точки зрения отдельных единиц совокупности, наоборот, − *объединение* отдельных единиц в однородные группы). Метод группировки основывается на следующих категориях: группировочный признак, интервал группировки и число групп.

Группировочный признак – это признак, по которому происходит объединение отдельных единиц совокупности в однородные группы (в нашем случае группировочным признаком выступает признак возраста).

Интервал очерчивает количественные границы групп. Как правило, он представляет собой промежуток между максимальными и минимальными значениями признака в группе. Интервалы бывают:

− *равные,* когда разность между максимальным и минимальным значениями в каждом из интервалов одинакова;

− *неравные,* когда, например, ширина интервала постепенно увеличивается, а верхний интервал часто не закрывается вовсе;

− *открытые,* когда имеется только либо верхняя, либо нижняя граница;

− *закрытые,* когда имеются и нижняя, и верхняя границы.

Определение числа групп (числа интервалов) и их границ подчинено ряду условий:

[10] Вспомним, что номинальная шкала – самая ограниченная в плане допустимых операций.

А. Число групп (***k***) детерминируется уровнем однородности/ неоднородности группировочного признака. Чем сильнее неоднородность, чем больше варьирует признак, тем больше должно быть число групп. Например, если мы опросили только школьников 10–11 классов, диапазон изменения значений признака будет небольшим, и, скорее всего, этот диапазон мы разобьем на два возрастных интервала, либо вообще не будем разбивать (все зависит от целей исследования).

Выбор числа интервалов группировки можно осуществить при помощи таблицы 2.1, приведенной ниже:

Таблица 2.1

Таблица определения числа интервалов

Объем выборки, n	***Число интервалов, k***
(25−40)	(5−6)
[40−60)	[6−8)
[60−100)	[8−10)
[100−200)	[10−12)
Больше 200	[12−15)

Б. Число групп должно отражать реальную структуру изучаемой совокупности.

В. Не допускается выделение пустых групп. Если проблема пустых групп все же возникает, при проведении структурных группировок используют неравные интервалы.

Г. Необходимо четко обозначить границы каждого интервала. Для начала нужно найти ширину каждого из интервалов. Шириной интервала **(*h*)** называют разность между верхней и нижней границами интервала.

В случае равных интервалов их ширина находится по следующей формуле:

$$h = \frac{x_{макс} - x_{мин}}{k - 1},$$

где
- h – ширина интервалов;
- $x_{макс}$ и $x_{мин}$ − максимальная и минимальная варианты выборки ($x_{макс}$ и $x_{мин}$ находятся непосредственно по таблице исходных данных);
- k – число интервалов.

Прибавив к этой величине ширину интервала, найдем нижнюю границу второго интервала (x_{n2}):

$$x_{n2} = x_{n1} + h.$$

Это будет одновременно и верхняя граница предыдущего (первого) интервала. В этой связи возникает резонный вопрос: «К какому интервалу относить варианту, находящуюся на стыке двух интервалов?». Однозначного ответа на этот вопрос нет. Авторы многих учебных пособий подчеркивают, что такие варианты могут быть с одинаковыми основаниями отнесены к любому из соседних интервалов, и исследователь сам решает, как поступить в этой ситуации. Тем не менее, мы склонны придерживаться правила, изложенного в следующем абзаце, которое для многих является негласным.

Д). Если в интервальном вариационном ряде в двух последовательных интервалах верхнее предельное значение признака одного интервала равняется нижнему предельному значению второго, условно будем считать, что это число принадлежит *второму интервалу*. В письменной форме это выражается с помощью широко известных условных обозначений – «круглой» и «квадратной» скобки, т. е. $[x_{n2};\ x_{n1+h})$.

Рассмотрим еще два символа, которые нам могут пригодиться при записи интервалов: $-\infty$(минус бесконечность) и $+\infty$(плюс бесконечность). Они не являются числами и вводятся лишь для удобства записи, когда нам неизвестны, либо мы сознательно не хотим задавать самую нижнюю и самую высокую границы всего диапазона значений.

Проиллюстрируем все сказанное примером. Всех респондентов, которых в начале данного параграфа мы распределили по профессиональному признаку, можно распределить и по признаку возрастному. Признак возраста зададим метрической шкалой. При этом представим, что диапазон распределения всех возможных значений признака довольно широк: мы опрашивали все население трудового возраста, начиная с тех, кто только что окончил школу, заканчивая теми, кто в ближайший год собирается выходить на пенсию. Всего нами было опрошено 80 человек (n = 80). Исходя из таблички определения интервалов, мы можем разбить весь диапазон на ~~8~~ 10 интервалов. Самому младшему респонденту, в идеале, должно быть 17 лет, самому старшему – 55. Следовательно, крайними (нижним и верхним) значениями диапазона распределения должны быть, соответственно 17 и 55. Но ведь известно, что некоторые молодые люди оканчивают школу в более раннем возрасте, например, в 15 или 16 лет. Такие же неточности могут быть и в отношении предпенсионного возраста. С учетом этого мы сознательно не ограничиваем крайнюю нижнюю границу первого интервала и крайнюю верхнюю границу последнего. Того человека, которому полных 22 года (27, 32, 37 и т. д.), относим не к первому, а ко второму интервалу. В итоге имеем:

1) моложе 22 лет − 5 человек (6,25%),
2) 22–27 года − 7 человек (8,75%),
3) 27–32 лет − 10 человек (12,5%),
4) 32–37 лет − 11 человек (13,75%),
5) 37–42 лет − 18 человек (22,5%),

6) 42–47 лет− 12 человек (15%),
7) 47–52 лет − 9 человек (11,25%),
8) 52 и старше – 8 человек (10%).

Для удобства занесем все эти цифры в таблицу 2.2, представленную ниже.

Таблица 2.2

Распределение респондентов по возрасту
(абсолютные и относительные для n = 80)

$x_{ni}; x_{ni+h}$ *Возраст*	$(-\infty;22)$	[22; 27)	[27; 32)	[32; 37)	**[37; 42)**	[42; 47)	[47; 52)	$[52;+\infty)$
Абсолютная частота n_i *(человек)*	5	7	10	11	***18***	12	9	8
Относительная частота, частость v_i *(в % ко всем опрош.)*	6,25	8,75	12,5	13,75	***22,5***	15	11,25	10

Из таблицы мы можем увидеть, насколько часто повторяется та или иная варианта, какова частота попадания респондентов в тот или иной интервал. Очевидно, что, например, самая большая частота у интервала [37; 42) и равна она 18 человек или 22,5%.

Помимо этого, рассмотрим так называемые **кумулятивные** вариационные ряды. В таких рядах вместо частот или относительных частот определенных вариант (или интервалов) записаны **накопленные** (кумулятивные) **частоты** или относительные накопленные частоты. Накопленная частота− число, полученное последовательным суммированием частот в направлении от первого интервала к последнему, до того интервала включительно, для которого определяется накопленная частота. Вычисление накопленных частот необходимо для определения некоторых характеристик положения, например, медианы и других квантилей. Более подробно об этом будет сказано в параграфе **2.3** данного учебника. Просчитаем накопленные частоты для нашего примера с возрастом респондентов. Результаты представлены в таблице 2.3.

Таблица 2.3

Распределение респондентов по возрасту

(абсолютные, относительные, кумулятивные частоты для n = 80)

$x_{ni}; x_{ni+h}$ *Возраст*	($-\infty$; 22)	[22; 27)	[27; 32)	[32; 37)	**[37; 42)**	[42; 47)	[47; 52)	[52; $+\infty$)
Абсолютная частота n_i *(человек)*	5	7	10	11	***18***	12	9	8
Относительная частота, частость v_i *(% ко всем опрош.)*	6,25	8,75	12,5	13,75	***22,5***	15	11,25	10
Кумулятивная частота n_i^H *(человек)*	5	12	22	33	51	63	72	80
Кумулятивная частость v_i^H *(% ко всем опрош.)*	6,25	15	27,4	41,15	63,65	78,65	89,9	100

В заключение отметим, что группировка данных, так сказать, предварительная процедура их агрегирования, она формирует основу для последующей сводки и анализа данных. Конечной целью агрегирования является представление множественных частных значений и показателей в одном общем показателе. Агрегированные показатели – обобщенные, синтетические «измерители». В социологии, как правило, используется три основных метода агрегирования данных, заимствованных из описательной статистики: 1) табличное представление; 2) графическое изображение; 3) расчет статистических показателей. К более подробному рассмотрению этих методов мы предлагаем обратиться в последующих параграфах данного раздела.

2.2. Перекрестная классификация и табличное представление социологической информации

В предыдущем параграфе речь шла о группировке как методе обработки социологических данных, позволяющем обеспечить первичное обобщение данных, представление их в более упорядоченном виде. Однако есть еще один, не менее важный метод обработки данных – метод классификации.

Классификация – это систематическое распределение явлений

и объектов по определенным группам, классам, разрядам на основании их сходства и различия [16].

Социологическое исследование редко ограничивается представлением одной переменной в виде ряда интервалов и категорий. Как правило, исследователь заинтересован в обнаружении связи двух и более переменных. Для данной цели более эффективна классификация по нескольким переменным – двухвариантная или поливариантная классификация.

Статистический метод располагает широким разнообразием приемов разной степени сложности для анализа подобных отношений, и одним из простейших приемов является метод перекрестной классификации или «матричный» метод. Согласно этому методу, ряд объектов группируется не по одному признаку, а по двум и более одновременно. Полученные в результате частоты названы совместными, потому что показывают число совместных событий, дают ключи к пониманию того, каким образом можно связать эти переменные, и могут дать больший эффект, чем более квалифицированная, сложная корреляционная техника.

Типы перекрестной классификации и их интерпретация. В таблице 2.4 для 28 объектов, в качестве которых выступают области Украины, Автономная Республика Крым, отдельно г. Севастополь и г. Киев и Украина в целом, построена перекрестная классификация, или «матрица», в которой представлено распределение численности докторов и кандидатов наук, в зависимости от их географического местоположения на карте Украины.

Таблица 2.4

Численность преподавательского состава на 100 студентов дневной формы обучения (распределение по регионам, в %)

Область	*Доктора наук*	*Кандидаты наук*
Автономная республика Крым	0,9	4,5
Севастополь	0,9	4,2
Винницкая	0,8	5,3
Волынская	0,4	4,3
Днепропетровская	0,8	4,1
Донецкая	0,6	3,5
Житомирская	0,4	3,6
Закарпатская	1,1	4,2
Запорожская	0,6	4,5
Ивано-Франковская	1	5
Киев	1,2	5,5
Киевская	0,5	2,7
Кировоградская	0,3	3
Луганская	0,6	4,2
Львовская	0,9	5,5
Николаевская	0,3	3,1
Одесская	1,1	5,5
Полтавская	0,5	4,2
Ровенская	0,4	4,7
Сумская	0,6	4,2
Тернопольская	0,6	3,3
Харьковская	0,9	5,4
Херсонская	0,5	3,3
Хмельницкая	0,5	4,4
Черкасская	0,5	3,8
Черниговская	0,4	3,7
Черновицкая	1	5,8
Украина – всего	0,8	4,6

Данный пример служит наглядной иллюстрацией того, как посредством метода классификации (перекрестной) мы можем осуществить сравнительный анализ распределения преподавательских кадров на востоке, западе, юге, севере и в центре страны.

Таблица *2.4* построена по качественной переменной и двум количественным. Однако аналогичным образом можно построить таблицу по двум качественным признакам (см. Табл. *2.5*).

Таблица 2.5

Распределение ответов респондентов в зависимости от их пола на вопросы анкеты: «На что Вы в наибольшей мере полагаетесь в своей жизни?»

(в % к опрошенным)[11]

Категории	*Мужчины*	*Женщины*
На личные знания, способности и силы	72	61
На поддержку со стороны собственной семьи	20	27
На бога	16	25
На удачу, на судьбу	17	13
На поддержку со стороны родителей (родительской семьи)	13	13
На поддержку со стороны государства	5	6
На поддержку со стороны друзей и знакомых	6	5
Другое	1	1
Трудно ответить	1	2

Как видим из примера, приведенного выше, нет никакой возможности оценить значение цифр, представленных в таблице, осуществить полноценный сравнительный анализ распределений ответов, в зависимости от половой принадлежности. Измерение осуществлялось по шкале с совместимыми альтернативами, следовательно, сумма всех ответов превышает 100%. Поэтому необходимо, чтобы для каждого ряда процентов было указано исходное количество случаев – N с тем, чтобы в распределении нашли свое отражение действительные отношения.

Может быть и такое, что процентные отношения, расположенные в ячейках матрицы (пересечениях строк и столбцов), вычисляются исходя не из полной суммы случаев, а лишь из случаев, относящихся к данной строке или столбцу («маргинальных сумм»). Обычно маргинальные процентные отношения выделяют какую-либо независимую переменную.

Как можно было заметить, результаты группировки и классификации данных социологического исследования довольно часто оформляются в виде статистических таблиц, где излагаются в наглядно-рациональной форме. Однако, следует подчеркнуть, что не всякая таблица может быть названа статистической. Табличные формы календарей, тестовых и опросных листов, таблица умножения не являются статистическими [2, с. 33; 8].

Статистическая таблица − это цифровое выражение итоговой характеристики всей наблюдаемой совокупности или ее составных частей по одному или нескольким существенным признакам. Статистическая таблица содержит два элемента: подлежащее и сказуемое [2, с. 35; 9]. Каждую статистическую таблицу можно рассматривать как «оконченную мысль»

[11] Моніторинг громадської думки населення України. Інформ. бюл. / Укр. ін-т соц. дослідж., Центр «Соц. Моніторинг». – К., 1999. – № 1. – 40 с.

исследователя. И каждая из них имеет свое подлежащее и сказуемое.

Подлежащее статистической таблицы − перечень групп или единиц, составляющих исследуемую совокупность единиц наблюдения.

Сказуемое статистической таблицы − цифровые показатели, с помощью которых дается характеристика выделенных в подлежащем групп и единиц.

Различают *простые, групповые и комбинационные таблицы*. В **простых таблицах**, как правило, содержится справочный материал, где дается перечень групп или единиц, составляющих объект изучения. При этом части подлежащего не являются группами одинакового качества, отсутствует систематизация изучаемых единиц. Сказуемое этих таблиц содержит абсолютные величины, отражающие объемы изучаемых процессов. Групповые и комбинационные таблицы предназначены для научных целей, где, в отличие от простых таблиц, сказуемое составляют средние и относительные величины на основе абсолютных величин.

Групповая таблица − это таблица, где статистическая совокупность разбивается на отдельные группы по какому-либо одному существенному признаку, при этом каждая группа характеризуется рядом показателей. Примером такой группировки может быть разделение семей на группы по месту проживания (сельское и городское), где образуются подгруппы семей по количеству детей. Анализ этих группировок по материалам переписи 1989 года позволил сделать вывод, что большинство семей, независимо от принадлежности к городскому или сельскому населению, имеют только по одному ребенку.

Комбинационная таблица − это таблица, где подлежащее представляет собой группировку единиц совокупности по двум и более признакам, которые распределяются на группы сначала по одному признаку, а затем на подгруппы по другому признаку внутри каждой из уже выделенных групп. Комбинационная таблица устанавливает существенную связь между факторами группировки. Примером комбинационной группировки может быть распределение полиграфических предприятий по трем существенным признакам: степени оснащенности современным полиграфическим оборудованием, степени применения современных технологий и уровню производительности труда. Такого рода статистические таблицы позволяют осуществить всесторонний анализ, но они менее наглядны.

Независимо от количества переменных, для которых строятся таблицы, есть единый перечень общих принципов, которыми при этом следует руководствоваться: количество и величину интервалов необходимо устанавливать таким образом, чтобы это не нарушало и не затуманивало взаимоотношение этих двух переменных. Что касается качественных показателей, признаки должны быть исчерпывающими и взаимоисключающими. Если эти простые принципы будут серьезно приняты во внимание, подготовка таблицы с перекрестной группировкой не составит труда.

В заключении подчеркнем, что при составлении таблиц необходимо соблюдать *общие правила*:

− таблица должна быть легко обозримой;
− общий заголовок должен кратко выражать основное содержание;
− наличие строк «общих итогов»;
− наличие нумерации строк, которые заполняются данными;
− соблюдение правила округления чисел.

2.3. Графическое представление социологической информации

Назначение графика. Весь табличный материал можно представить в графической форме, которая (обычно) более наглядно, чем таблица, выражает картину общего распределения. Таблица частот позволяет с большей легкостью интерпретировать результаты. Главное назначение графика – дать наиболее точное представление о форме частотного распределения – представление, понятное даже совершенно неквалифицированному читателю.

Существует множество видов графического представления, каждое из которых полезно в своем конкретном приложении [4, с. 20−27; 5]. Некоторые из них весьма сложны, но здесь рассматриваются только основные и простейшие: 1) гистограмма; 2) полигон распределения; 3) кумулята; 4) диаграмма полос; 5) статистическая карта; 6) временная диаграмма.

Первые три из указанных шести типов применимы только к количественным данным. Диаграмма полос предназначается, главным образом, для качественных данных; статистическая карта представляет распределение событий по географической площади, а временная диаграмма является графическим вариантом динамического ряда.

Так как графики эквивалентны таблицам, то они должны иметь аналогичные названия и обозначения и подчиняться тем же критериям доступности, простоты и ясности. График нельзя построить до тех пор, пока не будет приготовлена соответствующая таблица.

Построение гистограммы. Гистограмма состоит из ряда соприкасающихся столбцов, высота которых пропорциональна частоте соответствующего класса событий, а ширина пропорциональна величине интервала группировки переменной. Она является не только графической записью абсолютных частот группировок, но и наглядным изображением значения каждой частоты относительно всех других. В качестве примера предлагаем рассмотреть гистограмму, изображающую частотное распределение сети высших учебных заведений III–IV уровня аккредитации Украины (начиная с 1992 г.), которая представлена на рисунке 2.1. Если строить гистограмму вручную, то делается это на обыкновенной бумаге, линованной «в клетку». В качестве первого шага на сетке линий проводятся под прямыми углами две оси приблизительно равной длины, которые пересекаются вблизи нижнего угла с левой стороны страницы. Границы группировок наносятся на горизонтальной

оси X (абсцисс), а частоты группировок наносятся на вертикальной *оси Y* (ординат). Однако, прежде чем располагать интервалы группировок на оси абсцисс, необходимо установить, какое число линейных единиц можно поставить в соответствие каждому интервалу группировки. Для улучшения эстетического вида графика и удобства чтения полезно оставлять свободные отрезки в начале и в конце горизонтальной оси. Учитывая это, подсчитывают полное число единиц длины оси и делят его на число интервалов группировок, которые необходимо нанести на график. В результате получается число линейных единиц, предназначенных для каждого интервала. Теперь, начиная от точки пересечения двух осей, можно откладывать интервалы группировок и помечать их границы острыми вертикальными линиями или толстыми штрихами, размещенными соответственно с внешней стороны оси, так как они могут совпасть со следами первоначальной сетки линий. В связи с тем что эти пометки представляют собой общие точки соприкасающихся интервалов, они же обозначают истинные пределы группировок. Эти истинные пределы неизбежно будут дробными, когда пределы округляются до ближайшего целого.

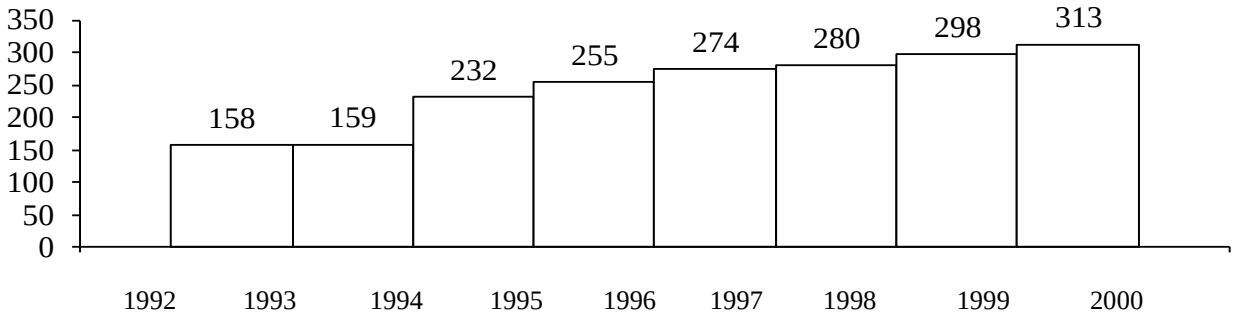


Рис. 2.1. Гистограмма динамики численности высших учебных заведений III−IV уровня аккредитации Украины (в абсолютных частотах)

Аналогичным образом устанавливается шкала частот на вертикальной оси. Наибольшую частоту группировки, которую необходимо разместить на графике, делят на число линейных единиц, имеющихся на предварительно проведенной оси. Тем самым определяется число событий, соответствующих каждой единице вертикальной оси.

Начиная с нулевой точки – точки пересечения осей, обозначения размещаются через равные интервалы и даются такими величинами (5, 10, 15 и т. д.), которые могут быть ясны и понятны.

Построив систему обозначений обеих осей, можно переходить к построению столбцов гистограммы. Вертикальные границы столбцов проводятся из точек истинных границ, а их высоты определяются частотами соответствующих интервалов.

Как правило, отсчет шкалы частот начинается с нуля, в противном случае не соблюдается необходимая пропорциональность площадей столбцов. Практика показывает, что желательно для лучшего восприятия пользоваться осями, равными по длине. Удлиняя ось абсцисс и укорачивая ось ординат, можно сделать гистограмму длинной и плоской, создавая тем самым впечатление большей вариации переменной. С другой стороны, удлинение оси ординат и укорочение оси абсцисс создает высокие узкие фигуры, которые представляют видимость незначительной вариации.

Неравные интервалы группировки. В случае неравных интервалов необходимо добиться возможности сравнения частот и сохранения их пространственных отношений. Поэтому необходимо разбить промежуток, который шире остальных, на два интервала равной ширины; соответственно этому необходимо разделить и его частоту на две равные частоты. Затем эти преобразованные подчастоты вычерчиваются на графике. Если бы начертили непреобразованную частоту, то столбец, соответствующий нестандартному интервалу, имел бы значение в два раза большее по площади, чем он имеет на самом деле, и тем самым создавал бы впечатление, не соответствующее действительности.

Обычно, прежде чем строить график по таблице, в которой существуют неравные интервалы группировок, необходимо представить все большие интервалы в виде кратного числа меньших интервалов и на это число разделить соответствующие частоты. Последний шаг приводит к преобразованным частотам, которые затем и вычерчиваются. Эта процедура находится в согласии с тем принципом, что каждое событие в распределении частот представлено равной площадью графика, так что относительные частоты пропорциональны площадям.

Теперь можно оценить практическую полезность правила о том, что интервалы группировки должны быть равной ширины или, в крайнем случае, должны быть составлены из целого числа меньших интервалов. Поэтому невозможно представить таблицу с открытыми интервалами в виде гистограммы, если не прибегать к произвольному ограничению интервалов – что иногда и приходится делать. Отличительной чертой гистограммы является ее схематическая простота. Столбцы более выразительны, чем числа. Они ясно раскрывают относительную плотность событий в каждом интервале и показывают контур распределения.

Природа и построение полигона распределения. Когда события сосредоточены в относительно небольшом количестве широких интервалов, частоты имеют тенденцию резко, скачком обрываться на границе каждого интервала. Разумнее, однако, предположить, что последовательность частот группировок была бы значительно глаже, если бы применяли большее число относительно небольших интервалов. Полигон распределения предназначен дать такое приближение в виде сглаженной кривой, которая, возможно, возникла бы, если бы размеры интервалов стремились к нулю, а число

наблюдений неограниченно возрастало [4, с. 20−25; 5].

Полигон распределения можно получить из гистограммы, проводя прямые линии через средние точки верхних частей смежных столбцов. Такое преобразование изображено на рисунке 2.2.

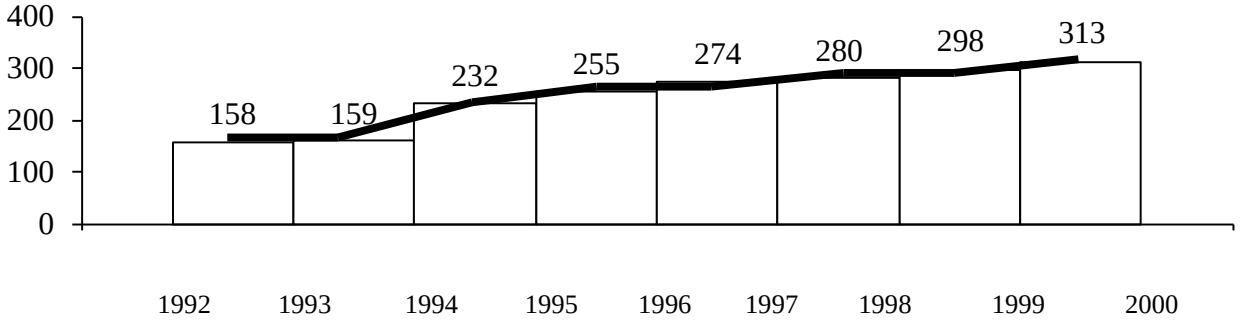


Рис. 2.2. Полигон и гистограмма динамики сети высших учебных заведений III−IV уровня аккредитации Украины (в абсолютных частотах)

На практике строится только один из графиков в зависимости от того, придается ли особое значение «затабулированным» частотам или же тем гипотетическим «точечным» частотам, которые дает полигон распределения. Очевидно, что процедура построения полигона распределения во всем соответствует процедуре построения гистограммы, за исключением конечной стадии. Сначала на линованной бумаге проводятся две оси, затем на базовой линии (оси абсцисс) наносятся интервалы, а шкала частот откладывается вдоль вертикальной оси. Однако на следующей стадии процедуры расходятся. Вместо столбцов наносятся линии, соединяющие точки, расположенные над серединами интервалов на высотах, соответствующих частотам данных интервалов; затем эти точки соединяются отрезками, образуя многоугольник. Нет никакой необходимости продолжать этот многоугольник (полигон) до базовой линии, за пределами области действительных наблюдений. Так, в V-образном распределении, где концентрация событий увеличивается на обоих концах, это было бы даже нелогично. По-видимому, более целесообразно оставить многоугольник открытым на обоих концах и тем самым избежать опасности ложного изображения частот, из которых ни одна фактически не наблюдалась. Иногда все же график доводят до базовой линии, опуская отрезки в средние точки свободных интервалов на обоих концах. Это объясняется стремлением определить площадь, принадлежащую гистограмме, которая графически изображает сумму частот. Хотя это и достаточно разумно, необходимо все же учитывать, что многоугольник неизбежно нарушает (как правило, преуменьшает) отношение между площадями столбцов и частотами группировок. В связи с этим необходимо помнить, что не существует идеальных статистических приемов, всесторонне и точно представляющих

данные. Полигон распределения претендует только на то, чтобы быстро и экономично получить первое приближение теоретической кривой распределения частот величин, сгруппированных в возможно меньшие интервалы. Для целей сравнения можно налагать друг на друга полигоны распределения, представляющие различные распределения одних и тех же переменных, не нарушая их относительных очертаний.

Кумулята. Построение кумуляты. Кумулятивная таблица частот (см. *Табл. 2.6*) дает процент случаев ниже (или выше) каждой данной границы группировки, но не аккумулируют частоты в пределах различных границ, лежащих в пределах вариации переменной. Поэтому она не может давать величины, которые соответствуют всем возможным кумулятивным процентам. Тем не менее часто требуется подобная информация. Так, например, может потребоваться подсчитать процент первоклассников, поступивших в школы г. Харькова с 1998 по 2000 год, или найти такое граничное количество, когда 50% первоклассников поступило в школы в период с 1998 по 2001 год, хотя этот вид информации можно получить из совокупной таблицы с помощью арифметической интерполяции. Эта же информация получается с меньшим усилием и достаточно точно из кумулятивного полигона распределения или кумуляты, как его иногда называют.

Таблица 2.6

Прием в первые классы общеобразовательных школ г. Харькова[12]

Учебный год	*Количество первоклассников: абсолютные частоты* (n_i)	*Кумулятивные частоты абсолютные* (n_i^H)	*Количество первоклассников: относительные частоты* (v_i *в %*)	*Кумулятивные частоты относительные* (v_i^H *%*)
1998/1999	14452	14452	32,64%	32,64%
1999/2000	13601	=14452+13601=28053	30,72%	=32,64%+30,72%=63,36%
2000/2001	16221	44274	36,64%	100,00%
Итого	44274		100,00%	

Построение кумуляты начинается с вычерчивания осей, как и для простого полигона распределения. Интервалы группировок наносятся на ось абсцисс, а частоты – на ось ординат. Так как все частоты теперь кумулятивные, то верхнее деление вертикальной шкалы будет соответствовать полной сумме частот.

Для быстрого понимания и легкости сравнения кумулятивные частоты обычно выражаются в процентах, так что шкала частот простирается от 0 до

[12] Показники роботи закладів освіти та наукових установ області за 2000 рік // Стат. зб. : За заг. ред. О. Л. Сидоренко, Л. О. Бєлової, А. С. Доценка. – Х., 2001 – 87 с.

100. Соответственно двум типам таблиц кумулятивных частот: «меньше чем» и «или более» – существует два типа кумулят. Но эти два графика дают идентичную информацию, так что для всех практических целей строиться будет только один. При построении кумуляты «меньше чем» (см. *Рис. 2.3*) кумулятивные частоты наносятся над истинными верхними пределами интервалов. Эта процедура отличается от процедуры построения простого полигона распределения, где фиксируются средние точки интервалов.

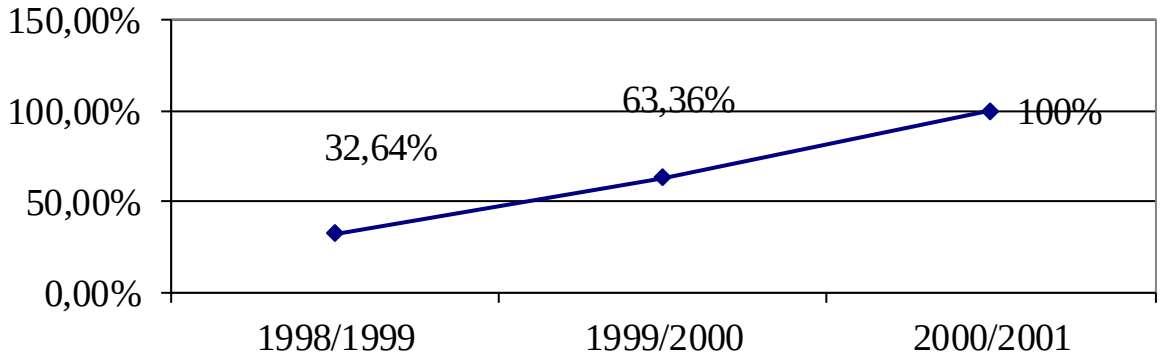


Рис. 2.3. Кумулята (график кумулятивных частот) приема в первые классы общеобразовательных школ г. Харькова

Некоторые авторы пользуются терминологией «меньше чем» и «больше чем», например, «менее чем 8» и «более чем 3». Однако эти термины создают трудности при применении к дискретным данным, если только они не трактуются как непрерывные. Если 33% всех семей состоят менее чем из трех человек, а 45% – более чем из трех, то семья, величиной в три человека, остается неучтенной. Однако когда данные являются непрерывными, то этих трудностей не возникает, так как точка разрыва не занимает никакой части континуума. Обозначение, принятое в этом тексте, является достаточно гибким для того, чтобы удовлетворить требованиям дискретных данных и не сделать насилия над непрерывными данными.

Благодаря такому способу построения шкал, кумулята «менее чем» будет начинаться в нижнем левом углу и проходить по диагонали в верхний правый угол, в то время как кумулята «или более» начинается в верхнем левом углу и движется диагонально к правому нижнему углу. Когда один или оба конца таблицы являются открытыми, кумулята не будет уже распространяться на всю частотную шкалу и, таким образом, будет казаться неполной. В этом случае можно ограничить кумуляту некоторой удобной точкой, не искажая данных.

Кумулята позволяет расчленять частотное распределение в любых точках в зависимости от необходимости, например, мы с легкостью сможем определить медиану (*Me*), найдя точку, четко разделяющую все события исследуемой совокупности на две равные части. Или точку, которая разделяет

нижние 25% и верхние 75% событий, так называемый, первый квартиль (Q_1), как на рисунке 2.4, представленном ниже.

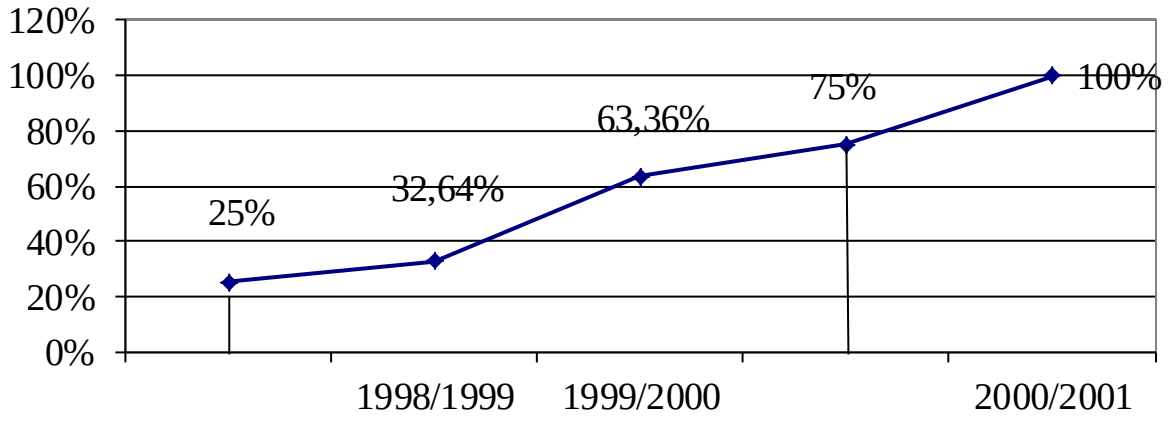


Рис. 2.4. Кумулята приема в первые классы общеобразовательных школ г. Харькова (отображение первого квартиля)

От отметки в 25% на оси частот мы провели линию, параллельную базовой линии, до тех пор, пока она не пересечет кумуляту «меньше чем». Из этого пересечения опускаем на базовую линию перпендикуляр, который фиксирует первый квартиль 1998 года. Эта цифра означает, что 25% первоклассников были приняты в школы г. Харькова в 1998/99 учебном году. Второй и третий квартили находятся аналогично.

Именно легкость получения графического решения дает кумуляте некоторое преимущество перед таблицей кумулятивных частот, при пользовании которой избежать утомительных интерполяций невозможно.

Графики качественных данных. Графическое изображение качественных данных отличается определенным образом от графиков количественных данных. Для графического изображения качественных данных используется длина отрезка, площадь фигуры или интенсивность оттенка цвета. Здесь представляется только три простейших обычно встречаемых типа: 1) *диаграмма полос*; 2) *круговая* (*гартовская) диаграмма*; 3) *статистическая карта*.

Диаграмма полос. Диаграмма полос представляет собой последовательность равноотстоящих друг от друга полос, длины полос пропорциональны соответствующим им частотам (см. *Рис. 2.5*) . Построим диаграмму полос для таблицы 2.5 (см. Стр. 59 данного учебника). Так как табличные признаки имеют только одно измерение, а именно – измерение частоты, то соответствующая диаграмма полос требует только одну шкалу – шкалу частот, которая обычно откладывается вдоль горизонтальной базовой линии. Так как в данном случае отсутствует непрерывная количественная

шкала (как это имеет место в гистограмме), которая определяет размер основания полосы, то полосы могут иметь любую удобную ширину и располагаться в любом возможном порядке. Они вычерчиваются одинаковой ширины, просто для наибольшей наглядности диаграммы; пустое пространство между ними – шириной в половину полосы – служит для подчеркивания дискретного характера качественных данных.

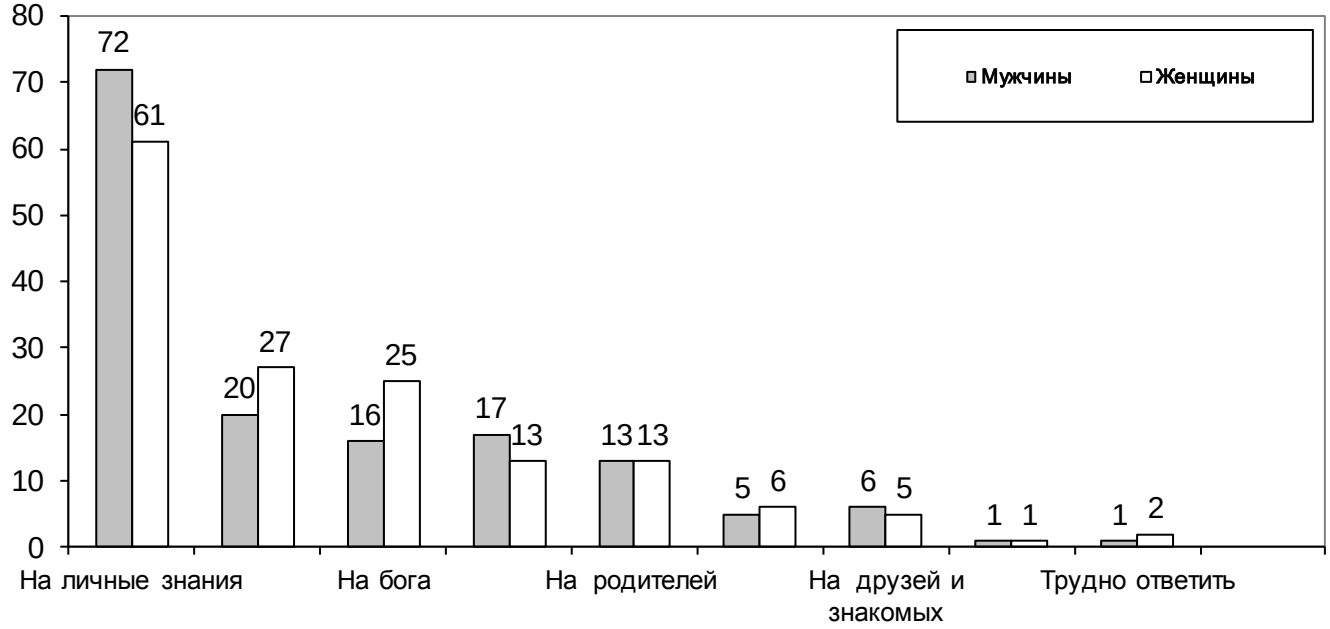


Рис.2.5. Диаграмма полос распределения ответов респондентов в зависимости от их пола на вопрос анкеты: «На кого/что Вы в наибольшей мере полагаетесь в своей жизни?» (% к опрошенным)

Круговая диаграмма – гартовская диаграмма. Как говорит само ее название, круговая диаграмма представляет собой круг, в котором вычерчиваются секторы, площадь которых пропорциональна наблюдаемым частотам (см. *Рис. 2.6*). При построении этой диаграммы частоты, выраженные в процентах, или частости, последовательно вымеряются угломером вдоль окружности в 360^0, или 100%. От этих отметок проводятся радиусы к центру круга, которые разделяют общую площадь на секторы, пропорциональные частостям. Наглядность круговой диаграммы, или, что то же самое, всех качественных диаграмм, часто может быть повышена путем раскрашивания. Построим круговую диаграмму для предыдущего примера (см. *Рис 2.5*.), отразив на ней распределение частот выбора мужчин (см. *Рис*. *2.6*).

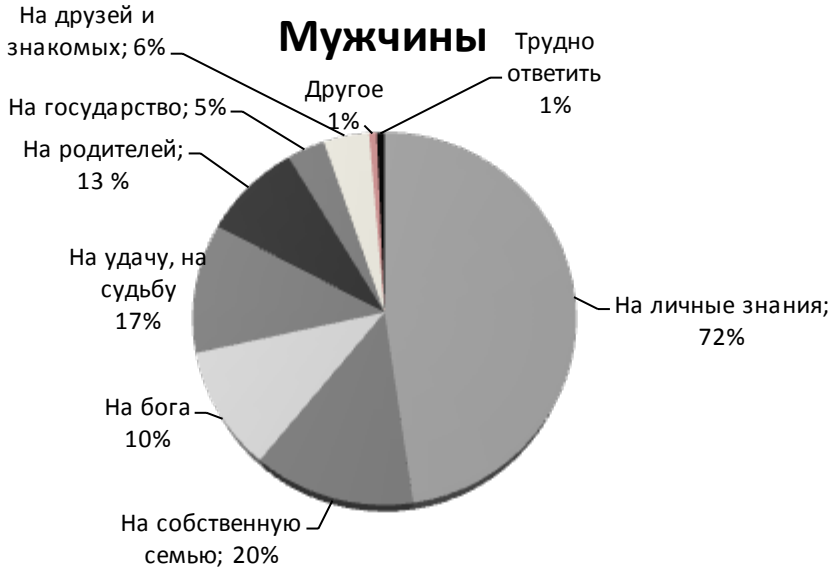


Рис.2.6. Круговая (гартовская) диаграмма распределения ответов мужчин на вопрос анкеты: «На кого/что Вы больше всего полагаетесь в жизни?» (% ко всем опрошенным)

Статистическая карта. Географические изменения в таких вопросах, как состав населения, количество бракосочетаний и разводов, количество преступлений и экономических показателей, – обычный предмет социологического анализа. Систематическое выявление таких измерений может служить целям, как социальной политики, так и научного объяснения. Например, открытие относительно высокой смертности в определенных географических областях или концентрация правонарушений в конкретном городском районе составляет первый шаг в раскрытии причин этих явлений.

Общим принципом составления статистических карт является обозначение различных частот соответствующей плотностью штриховки, которая создается разнообразными поперечными штриховками или различной концентрацией точек. Существуют специальные справочники, в которых детально описываются стандартные способы графического представления, на которых не считаем необходимым останавливаться в данном издании.

Основной проблемой в географическом картографировании является ясное определение пространственной единицы измерения. Такие единицы измерения, как города или государства, в этом отношении обычно не создают трудностей, так как границы этих единиц зафиксированы по закону. Но такие единицы, как городские кварталы, национальные районы и естественные экономические области, ставят множество проблем, которые необходимо разрешить до составления карты. Несмотря на известные сложности, статистическая карта – важный инструмент социологического исследования. В частности, она является существенным элементом в экономической школе

социологии, которая концентрирует внимание на пространственный аспект социального поведения. Статистическая карта состоит из тех контуров, которые считаются существенными для представления и понимания наносимых (на карту) социальных данных.

Временные диаграммы. Временные ряды представляются обыкновенно не в виде таблиц, а в виде графиков по той простой причине, что протяженные временные ряды трудно читать в виде таблиц, а исключительно короткие ряды, во всяком случае, не выразительны. Всякие отклонения от направления, скорости (флуктуации) можно намного легче обнаружить из графика, чем из ряда чисел.

Временные графики существуют *в двух преобладающих формах*: 1) *арифметические диаграммы* и 2) *полулогарифмические диаграммы*. В данном издании подробно остановимся только на арифметической диаграмме.

Арифметические временные диаграммы аналогичны простому графику частот, за исключением того, что строится график количественных наблюдений на интервалах времени, вместо того чтобы относить частоты событий к интервалам переменной. Интервалы времени наносятся вдоль горизонтальной оси, а флуктуации переменных величин наносятся на вертикальной оси.

Вместо ломаной линии временная диаграмма может состоять из ряда отдельных столбцов или полос, чьи высоты пропорциональны относительным величинам временных рядов. Эту временную диаграмму, которая является просто рядом календарных наблюдений, не нужно путать с бинарным распределением таких величин, как возраст бракосочетания и размер семьи, в которых базовая линия представляет параметр времени. Когда время является явной переменной в таких бинарных распределениях, для него существует истинная нулевая точка, как при измерении возраста или в любом воспроизводимом измерении. Во временных рядах, однако, отметки на базовой линии весьма произвольно фиксируют отдельные моменты времени. Тем не менее такие моменты времени часто рассматриваются как интервал, чтобы упорядочить наблюдаемые данные.

Многозначный график. При некоторых обстоятельствах, возможно начертить на одной и той же арифметической шкале два или больше временных ряда (чтобы более ясно выявить соотношения между ними). Такой график вполне понятен, потому что две переменные имеют приблизительно одну и ту же область и общее расположение на шкале. Однако эта процедура должна производиться с некоторой предосторожностью, потому что она может ввести в заблуждение читателя в тех случаях, когда переменные располагаются неодинаково. Ложное впечатление возникает в результате того, что два ряда данных расположены на неравных расстояниях от начала отсчета. Чтобы облегчить сравнение степеней изменения строится полулогарифмический график, который определенным образом преобразует шкалу.

2.4. Характеристики положения (среднее арифметическое, мода, медиана, квантили)

Основные числовые характеристики одномерного распределения: максимум; минимум; средние величины. *Характеристики положения* – величины, определяющие положение центра эмпирического распределения, устанавливающие положение всех единиц распределения вдоль шкалы. Определение этих характеристик может быть весьма результативным и в эмпирической социологии.

По окончании полевого этапа социологического исследования, как правило, социолог сталкивается с огромными массивами данных. Во всей своей полноте эти данные могут быть излишне подробными (дробными), слишком «неудобными» для обращения с ними, громоздкими для анализа и сравнения с другими аналогичными данными. Как мы уже говорили в начале данного раздела, для того чтобы преодолеть громоздкость и излишнюю дробность полученных данных, исследователь обращается к их агрегированию (укрупнению). В связи с этим он апеллирует к математической статистике, которая, в соответствии со своим назначением, должна: а) давать информацию о взаимном расположении фактов и событий; 2) исключать те величины, которые в данный момент не относятся к делу, 3) наглядно представлять общую картину.

Пределом эффективной агрегации, очевидно, было бы сведение множества событий к одной единственной величине, которая некоторым образом представляла бы собой их полную совокупность. Ясно, что никакая отдельная величина не может быть достаточно многосторонней, чтобы отразить каждую характеристику распределения; она может выразить только одно свойство из множества. Эта характерная величина не является точной копией общего. Она лишь приближенно описывает распределение значений.

Любая величина в распределении может описывать всю совокупность, если известно ее относительное положение в распределении. Следовательно, необходимо проанализировать все события в совокупности для того, чтобы оценить представительность любого из них. Но практически, не все величины одинаково полезны в этом отношении. Наиболее удобными и полезными для выделения их из таблиц являются: 1) максимум, 2) минимум и 3) центральные или типичные величины, известные как средние.

Максимум. В некоторых случаях максимальная величина переменной в распределении является единственной представительной величиной. При флуктуациях размера и веса движущегося транспорта такой величиной является наибольший груз, который можно безопасно перевозить по шоссе или мосту – знания «среднего» совершенно недостаточно, так как мост, построенный для средней нагрузки, разрушился бы при максимальной.

Подобно этому вместимость школ, госпиталей и других институтов планируется для ожидаемого максимума, а не для предполагаемого среднего.

Знание средней величины в данном случае бесполезно. Поэтому максимум, как мера положения, представляет собой граничную величину, ниже которой расположены не относящиеся к делу величины.

Минимум. Многие проблемы социологии политики сводятся к выбору минимальной величины распределения в качестве действующей нормы. Очевидно, что минимум – это такая величина, выше которой располагаются все другие величины в данном расположении. Так, для законодательства норм благосостояния населения из распределения доходов «нормальных» семей выводится минимальный доход (прожиточный минимум). Из гипотетического распределения зрелого населения по возрастам устанавливается минимальный возраст для женитьбы, службы в армии, права голосования и права избрания, а также и других социальных обязанностей. Будущий студент вуза может интересоваться только тем, каковы минимальные расходы, необходимые для обучения в нем.

Ни максимум, ни минимум не требуют сколько-нибудь серьезных вычислений как при определении их местоположения, так и в интерпретации. Их смысл весьма прост и, следовательно, не нуждается в более широком обсуждении.

Среднее, особенности выбора среднего. Наиболее общей и распространенной мерой расположения (и, в общем, наиболее полезной) является среднее. Обычно это центральная величина, вокруг которой группируется распределение. Явная тенденция многих статистических совокупностей концентрироваться вокруг центра часто называется «центральной тенденцией», а значение величины в этом центре – «мера центральной тенденции» – обычно называется средней.

Однако этот функциональный центр не обязательно идентичен с серединой области данных наблюдений. Область наибольшей концентрации может находиться как вблизи средней точки области, так и на значительном расстоянии от нее. Распределение оценок в тестах на испытание умственных способностей имеет колоколообразную форму с максимумом в средней точке области. С другой стороны, например, распределение доходов часто представляется V-образными – кривыми, когда большинство единиц распределения сконцентрировано в левой и правой оси шкалы, а не посередине или J-образными – кривыми, центр которых существенно смещен в сторону одной из осей шкалы (см. *Рис 2.10*, с. 81).

Подобно большинству других статистических мер, понятие о среднем имеет свои корни в обычном здравом смысле. Каждый политический деятель, основываясь на результатах исследования «своего» электората, совершенно привычно говорит о «среднем избирателе», «средней семье», «среднем студенте» и т. п. Широкое разнообразие черт, которые сводимы к среднему, видно из популярного описания «среднего» мужчины Украины, «ростом в сто семьдесят девять сантиметров, весом 82 килограмма, который предпочитает брюнеток, футбол, сало, борщ и жаркое и считает, что способность вести

домашнее хозяйство является наиболее важным достоинством жены». Не существует, разумеется, ни одного мужчины, который является средним во всех отношениях; человек среднего роста не обязательно будет человеком среднего ума или средней красоты. Поэтому популярное утверждение, что «среднего человека» не существует, полностью оправдано. Такие фиктивные понятия, как «средняя школа» и «средний украинец», не соответствуют более строгому статистическому понятию, которое применимо только к рядам измерений одной переменной. Тем не менее какой бы смысл не вкладывался в это понятие, политик бессознательно подразумевает то, что все статистики точно признают: среднее – есть разновидность нормы, вокруг которой колеблется переменная. Разница между политическим деятелем и социологом заключается в том, что последний требует большей точности, чем это позволяет неофициальное народное словоупотребление. Поэтому социолог-профессионал разрабатывает терминологию и математические процедуры для измерения среднего и тем самым ограничивается переменными, у которых центральная тенденция допускает некоторые виды квантификации. Различным типам центральных тенденций соответствуют различные средние, каждое из которых отвечает требованиям данной проблемы. Из этих многочисленных типов средних в данном издании подробно обсуждаются только три: среднее арифметическое, мода и медиана.

Среднее арифметическое (*M*[*X*]). Как и в обычном словоупотреблении, на статистическом языке «среднее» означает «типичное», «обычное», «ожидаемое». Среднее арифметическое– такое значение признака, сумма отклонений от которого всех значений признака равна нулю (с учетом знака отклонения). Если речь идет о выборочном исследовании, в основании которого лежит случайная выборка, среднее арифметическое часто называется математическим ожиданием, так как в данном случае имеется в виду среднее значение случайной величины. Следует подчеркнуть, что условное обозначение среднего в данном случае может отличаться. Как правило, если имеется в виду среднее по выборке, то вместо *M*[*X*] используются $\bar{x}$, что необходимо запомнить, во избежание путаницы в формулах. Считается, что любое физическое тело, находясь в неопределенном состоянии, будет стремиться принять состояние равновесия (опоры на свой центр тяжести), точно так же и среднее значение любой случайной величины при достаточно большом количестве испытаний будет стремиться к своему математическому ожиданию. Этот факт доказывается в теории вероятности. Математическим ожиданием случайной величины называется сумма произведений всех возможных значений случайной величины на вероятности этих значений.

В целом вычисление среднего арифметического необходимо для осуществления более сложных математических операций, связанных с анализом социологических данных, в чем можно будет убедиться на последующих занятиях. Именно среднее арифметическое, как математическое ожидание, является одной из основных характеристик выборки, так как с его

помощью можно прогнозировать значения некоторого случайного признака при достаточно долгом периоде испытаний. Например, человек, удрученный затянувшейся «полосой неудач», обычно надеется, что события должны прийти в норму. Он полагает, что существует нечто вроде закона природы, согласно которому полоса неудач должна сбалансироваться удачами.

Процедуры вычисления среднего арифметического для несгруппированных и сгруппированных данных несколько отличаются.

Вычисление среднего арифметического для несгруппированных данных. *Для несгруппированных данных* среднее вычисляется по простой формуле, путем суммирования отдельных величин и последующего деления на общее число событий, что имеет следующий вид:

$$M[X] = \frac{\sum x_i}{N}$$ (формула простого среднего),

где
- M[X]– среднее арифметическое;
- $\sum X_i$ – сумма переменных;
- X_i – значение переменной (ее величина);
- N – число событий.

Вычисление среднего арифметического для сгруппированных данных. В первом подразделе этого раздела мы говорили о том, что сгруппированные данные отличаются от несгруппированных тем, что каждой группе подобных величин приписывается частота или «вес». Поэтому для вычисления среднего сгруппированных событий каждое значение переменной X_i умножается на свою частоту (n_i), затем эти произведения суммируются, и вся сумма делится на сумму всех частот, что имеет следующей вид:

$$M[X] = \frac{\sum x_i \bullet n_i}{N}$$ (формула среднего взвешенного),

где
- X_i – значение переменной (ее величина);
- n_i – частота встречаемого значения переменной;
- N – сумма всех частот, объем исследуемой совокупности.

В этой связи важно понимать, что если применить две разные формулы (простого и взвешенного среднего) к одному и тому же распределению, значения средних, полученные по двум разным формулам, отличаться не будут.

Вычисление среднего арифметического для интервальных рядов.

Процедура вычисления среднего арифметического для интервальных рядов данных, несколько отличается от процедур, описанных выше. Необходимо понять, что для непрерывных интервальных рядов частота интервала совпадает с его средней точкой. Следовательно, перед тем как вычислять среднее арифметическое для интервального ряда, необходимо найти средние точки каждого интервала. Для этого сначала находятся границы

интервалов, а затем вычисляются их средние точки по формуле:

$$x_{ci} = \frac{(x_i + x_{i+1})}{2},$$

где ● x_i – нижняя граница интервала;

● x_{i+1} – верхняя граница интервала.

После чего каждая средняя точка взвешивается (то есть перемножается на частоту интервала) и вычисляется среднее арифметическое для всего интервального ряда по формуле:

$$M[X] = \frac{\sum x_{ci} \bullet n_i}{N},$$

где ● x_{ci} – среднняя точка интервала;

● n_i – частота интервала;

● N – сумма всех частот интервального ряда.

Вычисление среднего ряда средних (комбинированное среднее). Две или более средние величины сами часто усредняются, т. е. можно получать среднее ряда средних. Средние подгрупп до того, как они будут скомбинированы, должны быть взвешены в соответствии со своими N. Недооценка необходимости взвешивать средние может привести к абсурдным результатам. Можно привести пример из игры в баскетбол. Игрок за 75 ударов набрал 25 очков, что дало в среднем 0,33. В этот же день он за пять ударов набрал пять очков, что дало в среднем – 1,00. Каково будет среднее (комбинированное)?

Наивным было бы предположение о том, что если мы сложим две средние величины и разделим затем на 2 (по принципу нахождения среднего арифметического простого), то получим искомый результат. В таком случае, по отношению к рассматриваемому примеру, комбинированное среднее будет равным 0,667 – явно неправильный результат. В таблице 2.7 проиллюстрирована процедура правильного вычисления комбинированного среднего для приведенного выше примера.

Таблица 2.7

Расчетная таблица по нахождению комбинированного среднего

Количество ударов (n_i)	*Общее количество очков* ($\sum xi$)	*Среднее количество очков за каждый удар* ($\bar{x}i$)	$\overline{xi} \times n_i$	*Конечный результат – M [X]*
75	25	25/75=0,33	75×0,33=24,75	
5	5	5/5=1,00	5×1,00=5,00	
ИТОГО: N=80			24,75+5,00=29,75	*M [X]* = =29,75:80= =0,37

Получается, что в среднем за каждый удар в этот день баскетболист получил приблизительно по 0,37 баллов (число, округленное до сотых).

Приведем более близкий социологической практике пример нахождения среднего арифметического для ряда средних. Предположим, нам известны средние показатели численности детей, воспитывающихся в дошкольных учреждениях каждого из районов города. Но перед нами стоит задача выйти на один средний показатель по всему городу. Порядок вычислений в направлении разрешения этой задачи проиллюстрирован в таблице.

Таблица 2.8

Сеть дошкольных учреждений всех ведомств г. Харькова в 1999 г.[13]

№ п/п	*Название района*	*Количество д/у* (n_i)	*В них детей в среднем* ($\bar{x}_i$)	$\overline{Xi} \times n_i$
1	Дзержинский	33	161,06	33×161,06=5315
2	Октябрьский	13	140,62	1828
3	Киевский	33	141,88	4682
4	Коминтерновский	17	182,82	3108
5	Ленинский	23	93,09	2141
6	Московский	35	208,46	7296
7	Орджоникидзевский	24	151,88	3645
8	Червонозаводский	18	126,00	2268
9	Фрунзенский	16	188,56	3017
10	Итого (N)	212		33300

Для определения среднего количества детей (M[X]) в дошкольных учреждениях г. Харькова в 1999 г. необходимо первым делом узнать общее количество детей ($N = \sum \bar{x} \times n$ = 33300), а затем полученный результат разделить на количество дошкольных учреждений – 212. Продолжая этот процесс, найдем, что *M[X]* = 157,07, то есть в среднем на одно дошкольное учреждение г. Харькова в 1999 г. приходится 157,07 ребенка. Дробный характер дискретной величины (в реальной жизни не бывает 1,5 человека и т. п.), обычно, никого не смущает, если речь идет о среднем показателе.

Подводя черту под всем сказанным в отношении среднего арифметического, выделим его *основные свойства:*

[13] Стат. збірник: показники роботи закл. освіти та наук. установ обл. за 1999 рік. [За заг. редакцією О. Л. Сидоренка, А. С. Доценка, П. С. Дементьєва]. – Х., 2000. – 82 с.

1) сумма отклонений различных значений признака от среднего арифметического равна нулю;

2) если от каждой варианты вычесть или к каждой варианте прибавить какое-либо произвольное постоянное число, то среднее увеличится или уменьшится на то же самое число;

3) если каждую варианту умножить (разделить) на какое-либо произвольное постоянное число, то среднее увеличится (уменьшится) во столько же раз;

4) если веса, или частоты, разделить или умножить на какое-либо произвольное постоянное число, то величина среднего не изменится.

Необходимо уяснить и запомнить, что *среднее арифметическое вычисляется и имеет смысл только для метрических и порядковых шкал*. Оно представляет величину каждого события в распределении. В связи с этим оно подвержено влиянию как очень больших, так и крайне малых величин, что особенно заметно в несимметричных распределениях. Для таких распределений более информативными могут быть иные меры усреднения, такие как мода и медиана.

Мода или вероятностное среднее. *Мода представляет собой наиболее часто повторяющуюся величину в упорядоченном распределении; она характеризует то место распределения, где концентрация событий максимальна*. Этимологически она связана с представлением о преобладающей манере одеваться или с этикетом, к которому, как предполагается, приспосабливается большинство той или иной социальной группы. Следовательно, моду (M_o) можно также определить как наиболее вероятную величину, и поэтому ее называют вероятностным средним.

Исследование разговорных выражений наводит на мысль, что под модой в действительности часто подразумевают понятие «среднее». Отчасти это объясняется тем, что признаки, так же как и переменные, могут иметь преобладающие частоты в серии наблюдений. Когда политический деятель говорит о «среднем избирателе», интересы которого он собирается отстаивать, он исходит из того, что большинство избирателей руководствуются своими собственными интересами; или когда официантка замечает, что «средний» клиент не пьет черный кофе, она, вероятно, имеет в виду, что большинство посетителей данного ресторана не пьют черного кофе. Посетители ресторана и избиратели в свою очередь говорят о «средней официантке» и «среднем политическом деятеле».

На языке статистики *мода – это величина, с которой наиболее вероятно можно встретиться в серии зарегистрированных наблюдений*. Таким образом, мода является наиболее вероятной величиной, хотя знание одной лишь моды не позволяет определить степень этой вероятности. Понятно, что если бы частоты всех величин были одинаковыми, то не было бы никакого смысла вводить это понятие.

Вычисление моды. В преобладающем большинстве случаев для

определения моды достаточно просто подсчитать частоту появления каждой величины, так как мода – наиболее часто наблюдающаяся величина. В случае непрерывных данных подразумевается, что эмпирические измерения настолько подробны и кратковременны, что никаких два измерения не дают тождественных результатов. Очевидно, что мода не может выявиться без группировки, так как в противном случае не было бы никаких частот. Само собой разумеется, в случае наличия интервалов они должны быть одинаковыми по ширине; если это правило не соблюдать, можно, выбирая достаточно большие интервалы, получить моду практически любой желаемой величины – явно бессмысленный результат.

Следовательно, необходимо пройти две различные стадии в определении моды: 1) определение модального интервала и места в нем преобладающей или «модальной» частоты и 2) нахождение величины, соответствующей этой частоте. В качестве примера рассмотрим следующую таблицу.

Таблица 2.9

Количество высших учебных заведений Украины III–IV уровня аккредитации, появившихся в определенный учебный год
(известно, что в 1992 году их было 1580)

Годы (X_i; X_{i+1})	Количество учеб. завед. III–IV уровня аккредитации (n_i)	Накопленные частоты (n_i^H)
1993-1994	1	1
1994-1995 *(Модальный интервал)*	**73**	74
1995-1996 *(Медианный интервал)*	**23**	**97**
1996-1997	19	116
1997-1998	6	122
1998-1999	18	140
1999-2000	15	155
Итого	*155*	

В таблице 2.9 наибольшая частота равна 73, следовательно, интервал 1994/1995 гг. является модальным, а модальная величина или приближенная мода равна 1994,5, что является средней точкой модального интервала.

Однако существование «произвола» в выборе ширины интервала вносит некоторую неопределенность в процедуру вычисления моды. Существует два возможных пути избегания этой неопределенности: 1) отказаться от данного типа средней величины и применить другую меру среднего; 2) применить вместо приближенного более точный метод вычисления, который уменьшил бы неопределенность. Следует отметить, что даже при условии неопределенности

в вычислении моды довольно часто в социологической практике бывают такие случаи, когда нельзя отказаться от моды в пользу других мер усреднения. Следовательно, нужно обратить внимание на усовершенствование техники вычисления.

В приближенном методе, рассмотренном выше, находится средняя точка интервала с наибольшей частотой; при этом игнорируются смежные интервалы и их частоты. Однако эти смежные интервалы повлияли бы на величину моды, если бы их границы были иначе расположены. Следовательно, необходимо разработать более «чувствительный» метод, который позволит учесть влияние смежных интервалов.

Метод разностей в вычислении моды. Более совершенный, чем вышеописанный, метод вычисления моды сводится к следующему: 1) вычисляются разности между модальной частотой и частотами смежных интервалов; 2) вычисляется отношение одной из этих разностей (обычно с частотой нижнего интервала) к сумме двух других разностей; 3) это отношение умножается на ширину модального интервала, а затем полученный результат прибавляется к истинной нижней границе модального интервала. В итоге получаем уточненное значение моды. Вычислительная формула в данном случае будет иметь следующий вид:

$$Mo = x_0 + h\frac{n^{mo} - n^{-}}{2n_{mo} - n^{+} - n^{-}},$$

где
- x_o – нижняя граница модального интервала,
- h – ширина модального интервала,
- n_{mo} – частота модального интервала,
- n^{-} – частота интервала, предшествующего модальному интервалу,
- n^{+} – частота последующего интервала.

Применив формулу к таблице 2.9, получим следующие результаты:

x_o=1994 h=1 n_{mo}=73 n^{-}=1 n^{+}=23

$$Mo = 1994 + 1 \times \frac{73 - 1}{2 \cdot 73 - 1 - 23} = 1994{,}59.$$

Бимодальность. Некоторые распределения обнаруживают два максимума и поэтому называются бимодальными, в отличие от унимодальных распределений. Бимодальность распределения может быть следствием наложения двух или более популяций с различными частотными максимумами. Так, например, в полигоне распределения роста взрослого населения, благодаря объединению групп мужчин и женщин, которые характеризуются двумя различными распределениями роста, может возникнуть бимодальность.

Столкнувшись с бимодальностью, исследователь должен попытаться либо разделить распределения, которые ее вызывают, либо (в случае неудачи) принять бимодальность как характеристику, присущую данному распределению.

Медиана. В любой упорядоченной совокупности каждое событие занимает определенное место – первое, второе, десятое или семьдесят пятое или ранг. Очевидно, что любая конкретная численная величина ранга приобретает смысл и значение в зависимости от общего числа рангов. Ранг, равный 10, в ряду из 100, является относительно более высоким, чем ранг 10 в группе из 20.

Точка, которая рассекает упорядоченную (ранжированную) совокупность на две равные части так, что одна половина событий точно находится ниже, а другая половина выше этой точки, называется медианой. Например, в 1950 г. медианный возраст всего населения Соединенных Штатов был равен 30,4 года. Это означает, что одна половина населения была старше, а другая половина населения моложе этого возраста. Так как медиана ясно обозначает положение величины в последовательности, она часто называется «средним положением».

Вычисление медианы. По аналогии со средним арифметическим, существуют разные формулы вычисления медианы для несгруппированных и сгруппированных данных.

Если данные *не сгруппированы,* как правило, придерживаются одной из следующих формул:

$\frac{N}{2}$ или $\frac{N+1}{2}$, *где* N – общее число рангов. При этом, первая формула чаще используется при нечетном количестве рангов, а вторая – при четном количестве.

В случае *если данные сгруппированы,* – вычисления осуществляются по следующей формуле:

$$Me = x_0 + h\frac{\frac{1}{2}N - n_H}{n_{me}},$$

где
- x_o – нижняя граница медианного интервала;
- h – ширина медианного интервала;
- N – объем выборки (соответственно, $\frac{1}{2}N$ - «половинное» событие);
- n_H – частота, накопленная до медианного интервала;
- n_{me} – частота медианного интервала.

Если вкратце описать алгоритм вычисления медианы для данных, сгруппированных в интервалы, то следует начать с (1) ранжирования всех событий или нахождения накопленных частот для каждого интервала, затем (2) найти, так называемое, «половинное событие» (ведь медиана – точка, делящая

пополам), (3) по ряду накопленных частот найти интервал, в который попадает это событие, и (4) осуществив элементарные математические вычисления, найти саму медиану.

Применим данный алгоритм вычисления медианы к таблице 2.9, представленной выше: (1) ранжируем все события путем нахождения накопленных частот для каждого интервала, (2) делим сумму всех частот пополам: $N/2$=155/2=77,5 («половинное» событие), (3) находим медианный интервал, то есть местоположение 77,5-го события в ряду накопленных частот *(очевидно, что «половинное» событие не попадает в первые два интервала, накопленные частоты которых меньше, чем 77,5; частота третьего интервала, равная 97, существенно превосходит отмеченную границу)*, получаем, что 77,5-е событие находится где-то внутри интервала 1995/1996 гг. с частотой в 23, которая, по предположению, однородно распределена по всему интервалу.

Теперь можно определить все неизвестные компоненты формулы вычисления медианы: X_o=1995, h=1, $\frac{1}{2}N$ = 77,5, n_H = 74, n_{me} = 23.

Полученные значения подставим в соответствующую формулу:

$$Me = 1995 + 1 \times \frac{77{,}5 - 24}{23} = 1995{,}15.$$

Интерпретируя эту цифру, скажем, что ровно половина всех учебных заведений III–IV уровня аккредитации появилась в Украине до момента 1995,15 года, и ровно половина – позже.

Медиана дискретных данных. Некоторые авторы ограничивают применение медианы только непрерывными данными по той причине, что дискретные данные по определению не могут быть дробными, как это требуется для медианы. Но такое ограничение не представляется столь необходимым. В распределении размеров семей нет такого размера семьи, чтобы точно 50% семей были больше, а 50% семей – меньше. Можно ли в такой ситуации отказаться от медианы или следует прагматически трактовать данные как непрерывные и принять дробную величину в качестве медианы? Как и ранее, примем последнюю альтернативу. Ведь сущность медианы крайне проста: 50% событий имеют меньшую величину и 50% – большую. Она обладает также еще одним показательным критерием: суммарное расстояние между медианой и каждой из величин распределения всегда меньше, чем подобная величина, вычисленная для любой другой точки. Именно поэтому медиана «ближе» к событиям данного распределения, чем любая другая мера среднего. В этом и состоит смысл того, что медиана занимает центральное положение в распределении.

Другие меры усреднения. Вместо того чтобы вычислять медиану, можно было бы для большей точности локализовать данные в меньшем интервале – верхней четверти, десятой или даже сотой. Для таких дополнительных уточнений требуются меньшие подразделения. Вычисление квартилей, децилей и центилей, которые разделяют множество событий на четвертые, десятые и

сотые части, выполняются совершенно аналогично вычислению медианы, за исключением того, что в формулу для медианы, данную выше, подставляют вместо n_H другую величину, которая соответствует частоте или процентам событий, лежащих ниже рассматриваемого места. Например, для нахождения точки, ниже которой приходится наименьшая четверть случаев, заменяют $N/2$ на $N/4$.

Таким образом, первый квартиль в распределении будет равен:

$$Q_1 = X_o + \delta \frac{N/4 - n_{H1}}{n_{Q1}}.$$

Если потребуется выделить точку, ниже которой находятся 75% событий (или Q_3), то вычисления производятся по следующей формуле:

$$Q_3 = X_o + \delta \frac{3N/4 - n_{H3}}{n_{Q3}}.$$

90-й центиль (или C_{90}) может быть найден по формуле:

$$C_{90} = X_o + \delta \frac{90N/100 - n_{H90}}{n_{C90}}.$$

Квантили в качестве нормирующего критерия. *Медиана, квартили, квантили, децили и центили, которые, согласно своему определению, указывают на долю событий, расположенных ниже или выше данной величины, носят обобщенное название квантили.* Эти меры усреднения можно применять для фиксации относительного положения любой данной величины в своем ряду. Можно локализовать с помощью 90-го центиля вес в 80 килограмм (C_{90}=180). Это будет означать, что 10% населения обладают весом выше, а 90% – меньше данного веса. Точно так же на абстрактной шкале центилей можно локализовать и такие события, как возраст в 62 года, рост в 180 см, уровень интеллекта в 120 баллов и т. д.

Очевидно, что квантили можно рассматривать как стандартизованные меры расположения независимо от метрической системы или типа данных. Таким образом, человек, который находится на 90-м центиле в умственном развитии, может находиться вблизи 90-го центиля и по доходам, что указывает на сходство между этими двумя социальными явлениями. Такой подход позволяет с успехом сопоставлять и сравнивать «несравнимые» величины. Данные характеристики положения не зависят также от вида распределения, то есть от того, является ли оно нормальным, скошенным или прямоугольным. Это обстоятельство еще более повышает ценность квантилей, позволяя с помощью вышеописанных процедур ранжировать любую случайную величину в некотором упорядоченном множестве.

Подводя итог всему сказанному в данном параграфе, представим краткое изложение характеристик средних:

Среднее арифметическое

1. Это такая величина в данной совокупности, которая наблюдалась бы в том случае, если бы все величины были равны.
2. Суммы отклонений от среднего в любую сторону равны; следовательно, алгебраическая сумма отклонений равна нулю.
3. Среднее арифметическое репрезентирует значения каждой величины распределения.
4. Множества имеют одно и только одно среднее.
5. Над средним можно производить алгебраические действия, средние подгрупп можно комбинировать при надлежащем взвешивании.
6. Среднее можно вычислить даже в том случае, когда значения отдельных величин неизвестны при условии, что известна сумма всех величин (N).
7. Для вычисления среднего нет никакой необходимости группировать и упорядочивать величины.
8. Среднее можно вычислить только для закрытых интервалов.
9. Оно фиксировано в том смысле, что процедуры группирования не оказывают сколько-нибудь серьезного влияния на нее.
10. Среднее применимо только к количественным денным.

Мода

1. Наиболее часто наблюдаемая величина в распределении; точка наибольшей плотности.
2. Значение моды определяется преобладающей частотой, а не значениями переменной в распределении.
3. Это наиболее вероятная величина и, следовательно, наиболее типичная.
4. Данное распределение может иметь две или более моды. С другой стороны, в прямоугольном распределении не существует никакой моды.
5. Мода не выражает степени модальности.
6. Над модой нельзя производить алгебраические манипуляции; моды подгрупп нельзя комбинировать.
7. Она неопределенная в том смысле, что зависит от процедуры группировки.
8. Мода определена как для открытых, так и для закрытых распределений.
9. Мода – единственный тип среднего, который может представлять качественные переменные.

Медиана

1. Это величина, расположенная точно в средней точке множества (а не в области изменения переменной); половина событий имеет значения

больше ее, а половина – меньше.

2. Значение медианы определяется ее расположением во множестве данных и не зависит от значения отдельных величин.
3. Сумма расстояний от медианы до других величин множества меньше, чем подобная сумма, вычисленная для любой другой точки распределения.
4. Каждое множество имеет одну и только одну медиану.
5. Над медианой нельзя производить алгебраические действия: медианы подгрупп не могут взвешиваться и комбинироваться.
6. Она фиксирована в том смысле, что процедура группирования не оказывает на нее заметного влияния.
7. Для вычисления медианы все величины должны быть упорядочены и сгруппированы.
8. Медиана вычисляется как для открытых, так и для закрытых интервалов.
9. Качественные данные не позволяют рассчитать медиану.

Подчеркнем, что отнюдь не во всех случаях является целесообразным определение всех характеристик положения. Существует ряд критериев, помогающих решить вопрос о применимости того или иного типа среднего. Об этих критериях и пойдет речь в следующем подразделе.

2.5. Критерии выбора вида усреднения

Сопоставимость средних. В связи с тем что нам часто приходится выбирать между различными видами усреднения, необходимо изложить основные принципы выбора центрального значения. Уже было выяснено, что не существует универсальных типов среднего, которые можно применять во всех случаях. Необходимо всегда помнить, что среднее является единственным представителем распределения, величиной, которая весьма удобна вследствие компактности; однако она в то же время неудобна из-за краткосрочности. В лучшем случае, среднее вскрывает и вычеркивает столько же информации, сколько извлекает из распределения.

Известно три критерия, которые помогают решить вопрос о применимости того или иного типа среднего: (1) цель усреднения, (2) вид распределения данных, (3) ограничения, связанные с «техническими» причинами и типом измерительных шкал.

Цели усреднения. Любое эмпирическое социологическое исследование по существу является попыткой дать ответ на вопрос о природе явления, и статистическая процедура выступает лишь в роли инструмента. Исследователя при вычислении того или иного типа среднего интересуют следующие вопросы: каковы размеры семьи, какова продолжительность жизни, каков возраст населения.

Если размер семьи необходим для целей планирования жилищного строительства, то приближенная мода была бы более подходящей величиной, чем арифметическое среднее или медиана, даже если точная степень модальности неизвестна. Дома строятся не для абстрактных арифметических средних семей, а для реально существующих. Если же необходимо изучить плодовитость, то среднее арифметическое было бы более полезно, так как оно представляет как большие, так и малые семьи.

Вид распределения. Распределения могут иметь самый различный вид, от идеально симметричного до крайне асимметричного. Симметрия означает, что величины распределены идентично по обе стороны от среднего. Степень асимметричности влияет на типичность и представительность средних величин, и, следовательно, ее необходимо принимать во внимание при выборе типа среднего. Иногда утверждают, что если кривая симметрична, то вообще не возникает никаких проблем в выборе способа усреднения, так как среднее, медиана и мода становятся тождественными. Однако это справедливо только в арифметическом смысле, но не в теоретическом. Даже если численные величины средних тождественны каждому типу – среднему арифметическому, моде, медиане, то им при этом соответствуют совершенно различные «образы». Например, «средний» студент считает свою оценку не суммой всех оценок, деленной на N, а скорее просто типичной оценкой. Эти примеры приводят к выводу, что даже при симметричном распределении выбирают такой тип среднего, который, прежде всего, удовлетворяет целям исследования.

По мере того как распределение становится все более и более асимметричным, величины разных типов среднего начинают заметно отличаться, и проблема выбора способа усреднения приобретает серьезное значение. Прежде всего, среднее арифметическое для несимметричного распределения перестает быть типичным. В U-образном распределении среднего вообще практически может и не быть, и поэтому его величина может быть совершенно фиктивной. Многие социологические данные: заработные платы, размеры городов, размеры семей и т. п. – часто несколько асимметричны, что требует особого внимания при выборе типа среднего.

Ограничения, связанные с «техническими» причинами и типом измерительных шкал. Существуют определенные чисто технические особенности вычислений, которые могут вынудить использовать тот или иной тип среднего. Так, например, арифметическое среднее нельзя вычислить для распределений с открытыми интервалами. В этом случае пользуются медианой, если распределение не очень асимметрично, то есть когда среднее и медиана не на много отличаются друг от друга. С другой стороны, когда известны только сумма частот и величин, можно вычислить только среднее арифметическое, хотя другие типы среднего были бы предпочтительнее.

Необходимо всегда принимать во внимание технические возможности вычислительных методов. Так как не существует никакого метода комбинирования или взвешивания медиан и мод, то они обычно являются

завершающим этапом вычислений. Поэтому, когда предвидятся дополнительные вычисления, необходимо выбирать среднее арифметическое и его производные.

Кроме того, выбор той или иной меры усреднения может ограничиваться типом шкалы, с помощью которой измерялся признак. Об этом мы подробно рассказывали в одном из предыдущих параграфов, посвященных рассмотрению различных типов шкал, применяемых в социологическом исследовании. Вспомним, что наибольшие ограничения в этом плане накладывает номинальная шкала. Распределение, полученное по номинальной шкале, мы можем охарактеризовать только с помощью моды.

Минимум, максимум и промежуточные меры. Во многих случаях только средние величины рассматриваются в качестве мер расположения. Однако необходимо подчеркнуть, что данная точка зрения слишком ортодоксальна. Мера расположения не обязательно должна совпадать с мерой типичности. Средние величины, как правило, выражают типичность или представительность; но точно так же максимум, минимум или любая промежуточная величина могут служить мерой расположения, если только они соответствуют всему распределению.

Характеристики средних. Предпосылкой для надлежащего применения каждого типа среднего является знание соответствующих описательных характеристик. Последние часто анализируются с точки их «преимуществ и недостатков». Но такой подход является скорее оценочным, чем описательным, и поэтому не дает строгого изложения сущности каждой процедуры. Преимущества при решении одной проблемы могут оказаться недостатками при решении другой. Поэтому можно формально изложить описательные характеристики, присущие каждому типу среднего, независимо от ситуации, в которой они могут применяться.

Сравнение типов среднего. Подобно любому статистическому показателю средние, прежде всего, применяются для целей сравнения. Наиболее часто средние величины используются для сравнения положений отдельных групп и распределений на одной и той же шкале. Очевидно, однако, что различные типы средних величин принципиально не сравниваемы. Например, среднее арифметическое одного распределения нельзя сравнивать с модой другого распределения.

Различие характеристик средних становится еще более очевидным на примере асимметричных распределений. Будучи подверженным влиянию каждой величины, среднее арифметическое асимметричного распределения будет смещаться в направлении экстремальных величин. На моду края распределения не оказывают никакого влияния, в то время как медиана смещается преимущественно по направлению к "хвосту" асимметричного распределения. Однако это смещение не очень велико, поскольку концентрация событий в «хвосте», как правило, невелика. Если одномодальное распределение скошено вправо, порядок расположения различных типов среднего на базовой

линии будет таков: мода, медиана, среднее арифметическое и интервалы между ними будут изменяться в зависимости от степени искажения (см. *Рис. 2.7*). Если распределение скошено влево, порядок средних будет обратным.

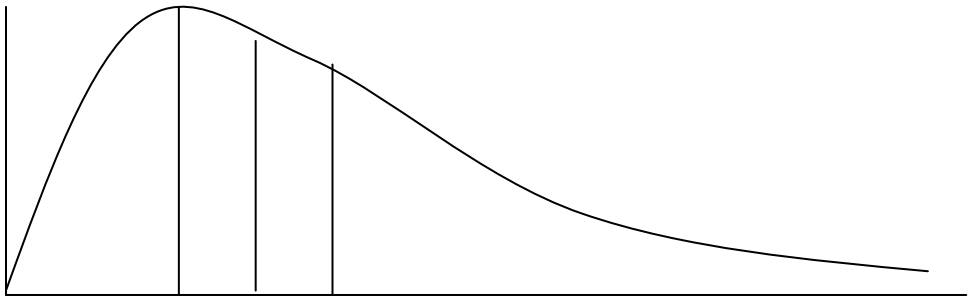


*Рис. 2.7.*Среднее, медиана и мода: правая скошенность

Но даже однотипные средние можно сравнивать только при условии подобия распределений, в том случае, когда «прочие условия равны». Например, сравнение среднего роста мужчин и среднего роста женщин вполне обосновано. Степень различия среднего роста мужчин и женщин надлежащим образом измеряется разницей их средних арифметических. Однако когда распределения заметно отличаются друг от друга, такие сравнения фактически могут ввести в заблуждение; особенно в случае сравнения средних арифметических.

Например, два ряда величин могут иметь равные средние арифметические и совершенно различные типы распределения (как показано на *рис. 2.8*).

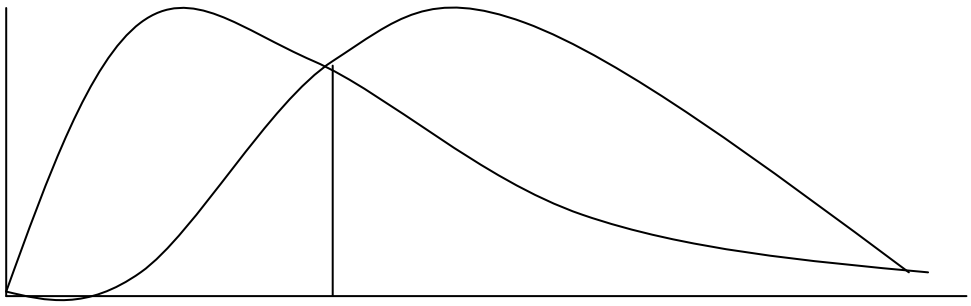


Рис. 2.8. Противоположная скошенность кривых. Одинаковые средние

Эти распределения занимают одинаковую область. Однако их максимумы не совпадают, и поэтому было бы более целесообразно сравнивать данные распределения по их модам.

На рисунке 2.9 максимумы расположены приблизительно в одной и той же области; однако средние арифметические не равны из-за несимметричности (скошенности) одного из распределений.

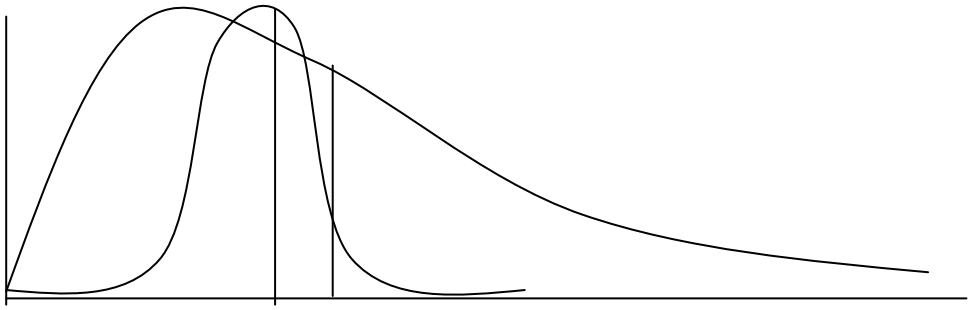


Рис. 2.9. Влияние несимметричности распределения на среднее

Делая общий вывод, следует подчеркнуть, что основные характеристики положения, о которых шла речь в этом параграфе, не представляют полного и всестороннего описания данных. Их следует рассматривать не как автономные величины, а лишь как определенные способы репрезентации конкретных данных. Использование средних ставит важную проблему реконструкции характера исходного распределения по нескольким извлеченным из него величинам. При решении этой проблемы важно знать другие характеристики – рассеяния, к рассмотрению которых мы и предлагаем обратиться в следующем подразделе учебника.

2.6. Характеристики рассеяния (дисперсия, отклонение, коэффициент вариации)

Меры вариации (протяженности) признака. Как уже говорилось, характеристики положения, хоть и являются чрезвычайно важными при изучении варьирующего признака, все же не дают полной информации о нем. Нетрудно представить себе два эмпирических распределения, у которых средние одинаковы, но при этом у одного из них значения признака рассеяны в узком диапазоне вокруг среднего, а у другого – в широком. Поэтому наряду с характеристиками положения нередко определяются и характеристики рассеяния исследуемой совокупности, показывающие, насколько близко/далеко все значения признака отдалены от средних показателей. Характеристики рассеяния выражаются в мерах вариации или мерах протяженности, наиболее употребляемыми из которых являются дисперсия, отклонения (среднее линейное и среднее квадратическое), коэффициенты вариации (для линейного и квадратического отклонений).

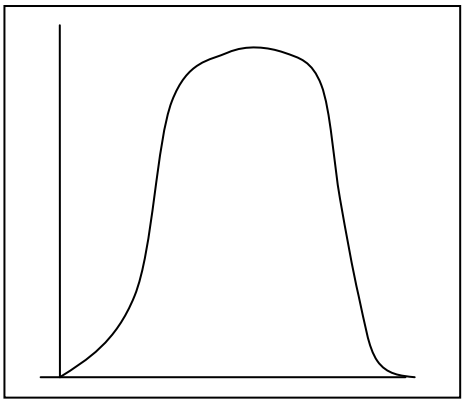
Вообще, понятие вариации лежит в основе всех статистических расчетов.

Многие учебники не устанавливают различия между терминами «варьируемость» и «вариация». Однако заострим внимание на этих различиях. **Варьируемость** − способность изменяться. **Вариация** – проявление такой способности, которую можно описать и измерить.

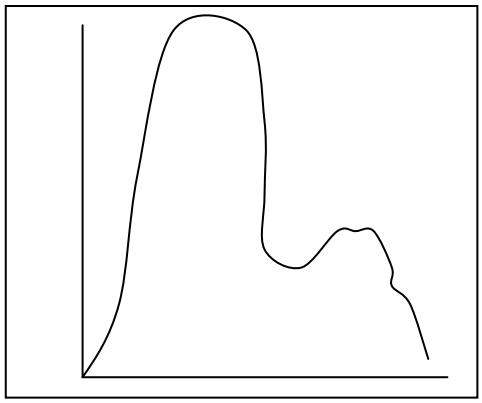
Если бы все величины исследуемого множества были идентичными, то вычисление среднего значения или любой другой статистической величины стало бы излишним. Ведь основная цель усреднения – получение одной величины, которая бы репрезентировала (представляла) целую группу неодинаковых величин. Средние величины и были изобретены для исключения различий между величинами. Однако при определенных обстоятельствах для социолога представляет бóльши й интерес характер отклонений, нежели результат усреднения. Например, оценивая общие способности и успеваемость студента, преподаватель должен принимать во внимание не только его средний балл, но и тенденцию этих баллов. Студент с баллами «5», «4» и «3», в общем, будет оценен иначе, чем студент с баллами «4», «4», «4», хотя средний балл у них одинаковый − «4». Тренеру по баскетболу, который отбирает для университетского первенства одного из двух игроков, имеющих равный средний балл, больше подходит стабильный игрок, который редко отклоняется от среднего, нежели неустойчивый спортсмен, который в среднем показывает низкий уровень игры и лишь иногда эффектно проявляет себя. Аналогично, в двух профессиональных группах, имеющих приблизительно одинаковый среднегодовой заработок, например, профессоров и бизнесменов, могут быть представлены самые различные заработки. Среди профессоров университетов оклады сравнительно мало отклоняются от среднего, тогда как в бизнесе доход является менее устойчивым.

В общем случае знание *картины вариации наблюдаемых значений – того, что статистики называют разбросы, рассеиванием или* **дисперсией**, не менее важно для социолога, чем знание средних величин. Для изучения этой картины в арсенале современной статистики имеется довольно много испытанных приемов и методов. Хотя они различаются в деталях, их можно разделить на три широкие категории в соответствии с процедурой их применения: 1) измерение области, содержащей всю или основную часть распределения; 2) измерение отклонений переменной от центрального значения; и 3) измерение степени однородности качественных переменных.

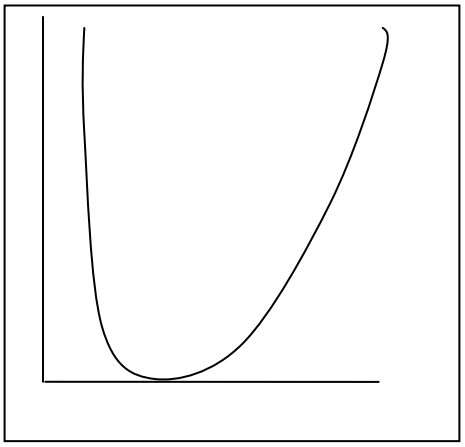
Некоторое общее представление о дисперсии количественных данных можно получить из графиков, представленных на рисунке 2.10. График частот сразу же позволяет прийти к выводу о том, является ли дисперсия симметричной или нет, увеличиваются ли частоты при приближении к среднему арифметическому (унимодальное распределение) или же они возрастают при смещении к концам интервала (U – образное распределение); распределяются ли величины равномерно по всей области вариации переменной или же они концентрируются в двух точках, как, например, в бимодальном распределении.



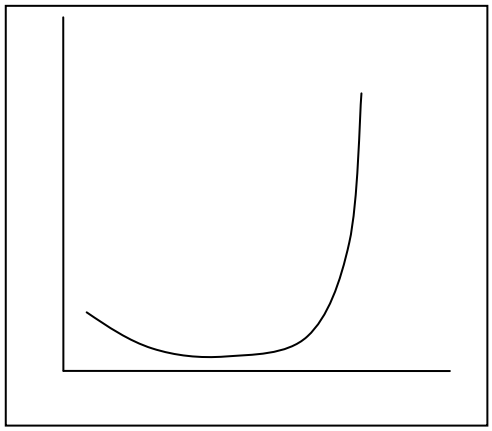
(Кривая унимодального распределения)



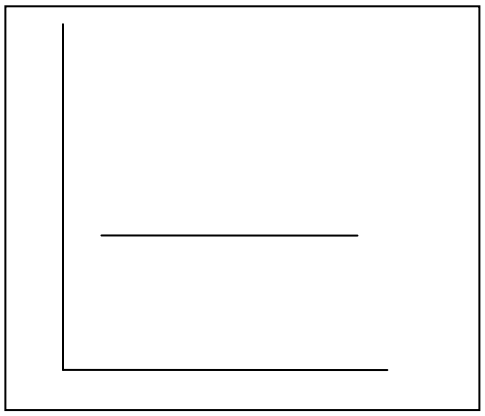
(Кривая бимодального распределения)



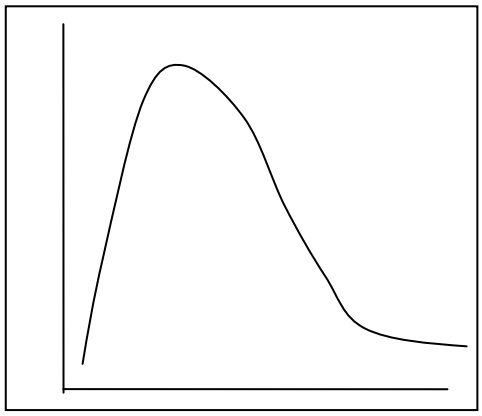
U-образная кривая



J-образная кривая



(Изображение линейного распределения)



(Влево скошенное распределение)

Рис. 2.10. Графики частот, позволяющие судить о дисперсии

Однако такие визуальные впечатления являются индивидуальными и субъективными, а выводы о дисперсии, сделанные на их основе, вряд ли

могут претендовать на научность. В таком случае научная обоснованность информации обеспечивается за счет объективных показателей, отражающих ту или иную меру вариации (протяженности, разброса, рассеяния), которые находятся с помощью стандартных вычислительных процедур

Измерение размаха вариации. Простейшая и самая грубая мера вариации – размах вариации (или «диапазон») *как размер области вариации переменной. По определению,* ***размах вариации*** *– это интервал, заключающий в себе все значения.* Следовательно, он находится точно так же, как и обычный интервал: вычисляется разность между истинными крайними значениями множества переменных, устанавливающими границы размаха вариации:

$$R = (X_{max} - X_{min}).$$

Например, для определения диапазона количества студентов высших учебных заведений III–IV уровня аккредитации нужно найти экстремальные истинные значения, а затем вычитаем одно из другого. Если наименьшее значение 159, а наибольшее – 259, диапазон будет равен разности между этими числами, то есть − 100. Таково расстояние, необходимое для расположения всех наблюдаемых частот. Для другого множества частот, несомненно, можно получить другую величину диапазона, другой размах вариации. Увеличивая количество наблюдений, можно либо расширить этот размах, либо оставить его неизменным, однако сократить − нельзя. По этой причине два или более диапазона сравнимы только в том случае, если они состоят из приблизительно одинакового числа наблюдений. Например, не следует сравнивать диапазоны оценок двух студентов, если каждый диапазон не содержит приблизительно одинаковое число этих оценок.

Для дискретных данных процедура измерения диапазона является точно такой же, как и для непрерывных, за исключением того, что истинные пределы становятся при этом в силу необходимости фиктивными. Так, распределение размеров семей от 2 до 12 человек имеет размах вариации равный 11, что является разностью между 12,5 и 1,5. Это значит, что переменная может принимать 11 и только 11 последовательных значений: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 или 12. В некоторых учебниках эта величина обозначается как «включающий» диапазон и находится по формуле:

$$R = (X_{max} - X_{min}) + 1 = 11.$$

Применив эту формулу в случае рассматриваемого выше примера, получим: (12 – 2) + 1=11.

Значение диапазона определяется весьма просто, подобно любой другой статистической величине, оно даст лишь ограниченную информацию. Поскольку величина его определяется расположением на шкале распределения, необходимо указывать не только его абсолютную величину, но и граничные точки. Повседневная практика подтверждает справедливость этого положения в таких утверждениях, как: «Цена на новые автомобили установится

в диапазоне от 4000 до 5000 долларов» или «Температура завтра будет в диапазоне от 12^0 С до 18^0 С. Следует обратить внимание, что, например, перечень окладов от 500 до 1000 гривен имеет тот же абсолютный размах вариации, что и перечень от 2500 до 3000 гривен, однако будет с совершенно иным содержанием. А при выборе места для отдыха недостаточно знать, что диапазон температур там равен 30^0 С, важно знать абсолютные значения крайних температур.

Характерной особенностью диапазона является то, что он не учитывает структуры вариации в своих пределах, однако иногда именно эта структура представляет наибольший интерес. Диапазон среднегодового дохода семьи в Украине составляет величину свыше 3500 гривен, что не дает никакого ответа на вопрос, сгруппированы ли доходы в середине, концентрируются ли они к концу интервала, ограничивающего диапазон, или же они равномерно распределены по всему диапазону.

Более того, в большинстве наблюдаемых распределений пределы вариации соответствуют весьма малым частотам; следовательно, полный диапазон, определяемый с помощью этих значений, может создать впечатление большей величины вариации, нежели в действительности. Устанавливая возрастной диапазон студентов вуза от 14-летнего вундеркинда до 64-летней бабушки, которая желает присутствовать в вузе вместе со своими внуками, получаем диапазон в 51 год. Но этот результат затмевает тот факт, что большинство студентов отличаются друг от друга лишь несколькими годами, и, следовательно, полный диапазон, как показатель вариации, в данном случае вводит в заблуждение.

Промежуточные диапазоны. Такая зависимость от пределов вариации может быть уменьшена с помощью промежуточных диапазонов, не учитывающих крайние значения переменной. Ограничивая область наибольшей концентрации наблюдений, такой диапазон обеспечивает большую стабильность и надежность. В обычной практике берется разность между 90-м и 10-м центилями, то есть устанавливается диапазон, включающий 80% случаев. Еще более ограниченным диапазоном является промежуточный интервал между первым и третьим квартилями, который содержит в средние 50% случаев. Обычно он называется интерквартильным диапазоном.

Интерквартильный диапазон может быть изображен графически путем нанесения квартилей на базовую линию графика. Подобная процедура обнаруживает степень «сгруппированности» случаев, относящихся к среднему интервалу, вокруг медианы. В рассматриваемом примере четыре интервала, образуемые квартилями, очень сильно различаются по ширине. Хотя каждый из четырех интервалов содержит ровно до 25% общей частоты. Такая классификация наблюдений обратна по отношению к ортодоксальной таблице частот, в которой интервалы выбираются равными, а частоты не равны. Здесь же устанавливаем равные интервалы частот, но допускаем переменную ширину интервала.

Нетрудно убедиться, что построение промежуточных диапазонов, таких, как интервал 10–90% или интерквартильный диапазон, не ограничено никакими рамками. Промежуточные диапазоны в ряде случаев обеспечивают большую ясность в вопросе об относительной концентрации или дисперсии случаев по сравнению с полным диапазоном. Во многих случаях использование стратегически расположенных квартилей (это общий термин мер расчленения) вполне удовлетворительно описывает дисперсию, делает ненужными более усложненные методы.

Отклонение от среднего как мера вариации. Диапазон полный, или промежуточный, позволяет судить об области распределения переменных. Хотя эта мера и имеет некоторое применение, особенно, когда указываются абсолютные границы, она дает информацию лишь о пределах вариации, а не о вариации в этих пределах. Следовательно, можно еще раз повторить, что размах вариации измеряет границы рассеяния, а не полную величину вариаций совокупности. Необходимы некоторые показатели, которые отражали бы степень вариации величин полного распределения.

Попытаемся определить вариацию как отклонение от некоторого значения. Можно, например, получить множество парных разностей для всех величин распределения, а затем «конденсировать» эти разности в некоторый показатель вариации. Однако каждый, кого интересует вариация ряда значений, интуитивно рассматривает их в связи с фиксированным стандартом, который построен на совокупном социальном опыте. «Высокий оклад», «чрезвычайно высокий уровень рождаемости» или «слабые способности» − все эти суждения являются оценкой относительной величины вариации, замеряемой от некоторой установленной основы. Высокий темп рождаемости можно рассматривать как отклонение в положительном направлении, низкий темп рождаемости − как отклонение в отрицательном направлении от нормы. Следовательно, эта норма становится центральной величиной, относительно которой воспринимается и измеряется вариация.

Точно так же можно судить об уровне смертности, об ученических оценках или о преподавательских окладах, особенно, когда они достаточно близки к наблюдаемым экстремумам. Тем не менее более обоснованно выбирать в качестве точки отсчета не субъективные «стандарты», а средние величины. Другими словами, исследователи стремятся организовать наблюдения вокруг среднего значения, взятого в качестве нормы, полагая, что это среднее является характерным значением. Остается решить следующие проблемы: 1) от какого среднего вычислять отклонения; 2) как представить эти отклонения компактным показателем.

Выбор нормы (от которой «отклоняется» отклонение). Вообще, нормой в данном случае считается средний показатель. Однако нам известны, по крайней мере, три разные средние меры (*M[X], Mo, Me*), а выбрать в качестве нормы нужно только одну из них. Очевидно, что для симметричных унимодальных распределений этот выбор не представляет особых трудностей,

поскольку среднее арифметическое, мода и медиана оказываются равными. Однако совершенно симметричные распределения встречаются крайне редко. Следовательно, дилемма возникает при выборе нормы для тех распределений, для которых мода, медиана и среднее арифметическое отличаются друг от друга. Многие исследователи выбирают моду, то есть максимальную частоту в качестве основы для сравнения. Однако более широко в этих целях используются среднее арифметическое или медиана, которые считаются более репрезентативными.

Как уже говорилось, среднее арифметическое – это величина, вариации которой в обе стороны равны. Следовательно, среднее удобно использовать в качестве нормы, однако и медиана столь же успешно может служить началом отсчета вариации. Поскольку медиана делит совокупность распределения на равные части – сумма отклонений от медианы меньше, чем от любой другой точки. Говоря другими словами, медиана – это точка, для которой арифметические «ошибки» минимальны. Приведенное утверждение можно назвать принципом минимального отклонения.

Абсолютные показатели вариации. ***Среднее линейное отклонение ($\overline{d}$)***

Простая сумма арифметических отклонений является бесполезной в качестве показателя вариации, поскольку она непосредственно зависит от количества объектов в распределении. Например, она будет большой для 1000 объектов и малой для 10. Чтобы устранить этот фактор, разделим сумму отклонений на N и, следовательно, будем измерять отклонение, приходящееся на один объект. Этот результат называется средним линейным отклонением ($\overline{d}$) и вычисляется с использованием формул:

1. Для первичного ряда:

$$\overline{d} = \frac{\sum(x_i - \mathrm{M[X]})}{n}.$$

5

2. Для вариационного ряда:

$$\overline{d} = \frac{\sum |x_i - \mathrm{M[X]}| * n_i}{N}.$$

3. Для интервального ряда:

$$\overline{d} = \frac{\sum |x_{ci} - \mathrm{M[X]}| * n_i}{N}.$$

При этом

- X_i – значение переменной (ее величина);
- x_{ci} – средняя точка интервала (в случае интервального ряда);

● M[X] – среднее арифметическое (причем вместо *M[X]* мы можем подставить в формулу *Me,* в том случае, если распределение неравномерно, асимметрично);
● n_i – частота значения переменной (*или частота интервала в случае интервального ряда*);
● *n* – число вариант признака (*в случае первичного ряда*);
● *N* – сумма всех частот.

Квадратичные отклонения. Измерение вариации с помощью простых арифметических (линейных) отклонений от центрального значения является наиболее простой процедурой. Если исследователя интересует лишь наличие или отсутствие дисперсии, то легко вычисляемый $d_{ср}$ вполне подходит для этой цели. Однако вариация, как правило, измеряется через квадратичные отклонения от средних величин. Логика этой операции основывается на принципе минимальных отклонений. Как сумма квадратичных отклонений от среднего арифметического, так и сумма квадратичных отклонений от медианы является минимальной величиной. Это положение получило название «принцип наименьших квадратов» и является одним из наиболее важных принципов статистических расчетов.

Метод квадратичных отклонений может показаться, на первый взгляд, искусственным и излишне усложненным. Если вариация может быть удовлетворительно измерена посредством вычисления линейного отклонения, то какое дополнительное преимущество дает возведение в квадрат? Удовлетворительный ответ на этот вопрос можно получить, лишь углубившись в изучение математической статистики. В рамках этого учебника целесообразность такого углубления представляется сомнительной, потому будем полагаться на практический опыт социологов, которые прибегают к вычислению квадратичного отклонения намного чаще, чем простого среднего отклонения

Квадратичные отклонения могут быть получены несколькими способами, каждый из которых пригоден для определенной цели: 1*) сумма квадратов отклонений (дисперсия); 2) вариация; 3) среднее квадратическое отклонение.*

Сумма квадратов отклонений (дисперсия). Дисперсия (δ^2 или *D*) – величина равная среднему значению отклонений отдельных значений признака от среднего значения. Для того, чтобы найти эту величину, необходимо произвести вычисления по следующим формулам:

1. *Для первичного ряда*:

$$\delta^2 = \frac{\sum(x_i - M[X])^2}{n}.$$

2. Для вариационного ряда:

$$\delta^2 = \frac{\sum(x_i - M[X])^2 * n_i}{N}.$$

3. *Для интервального ряда:*

$$\delta^2 = \frac{\sum(x_{ci} - M[X])^2 * n_i}{N}.$$

При этом

- X_i – значение переменной (ее величина);
- x_{ci} – средняя точка интервала (*в случае интервального ряда*);
- $M[X]$– среднее арифметическое;
- n – число вариант признака;
- n_i – частота значения переменной (*или частота интервала в случае интервального ряда*);
- N – сумма всех частот.

В качестве демонстрации процедуры вычисления дисперсии предлагаем обратиться к примеру, который рассматривался в предыдущем параграфе: «Сеть дошкольных учреждений всех ведомств в г. Харькове в 1999 г.». Мы вычисляли среднее арифметическое для сгруппированных данных, пытаясь таким путем найти среднюю численность детей в харьковских дошкольных учреждениях. По результатам этих вычислений мы получили *M[X] = 157,07*. Этот результат и процедура его нахождения представлены в третьем столбце расчетной (для дисперсии) таблицы, представленной ниже (см. *Табл. 2.10*). Когда среднее нам известно, осуществление последующих действий для нахождения дисперсии не составит большого труда, что также видно из данной таблицы.

Таблица 2.10

Сеть дошкольных учреждений всех ведомств в г. Харькове в 1999 г.[14]

Количество д/у в 1999 г. (n_i)	В них детей в среднем ($\bar{x}_i$)	$n_i \times \bar{x}_i$	$(x_i - M[X])^2$	$(x_i - M[X])^2 \times n_i$
33	161,06	33×161,06= =5315	161,06 - *157,07* = =15,92	525,52
13	140,62	1828	270,75	3519,81
33	141,88	4682	230,77	7615,51
17	182,82	3108	663,24	11275,15
23	93,09	2141	4093,83	94158,09
35	208,46	7296	2640,64	92422,35
24	151,88	3645	26,99	647,71
18	126,00	2268	965,34	17376,21
16	188,56	3017	991,78	15868,44
N=212		33300/212 = *157,07* (M[X])		**243408,78**

Следуя формуле вычисления дисперсии, нам осталось выполнить последнее действие: итоговое табличное значение 243408,78 разделить на сумму всех частот, равную 212. Получаем: σ^2=243408,78 / 212=1148,15.

Среднее квадратическое отклонение (сигма − σ). Поскольку вариация основана на квадратических отклонениях, она не является линейной мерой. Если требуется линейная мера, то необходимо извлечь квадратный корень из обеих частей соотношения. В такой форме эта величина известна как среднее квадратическое отклонение или сигма (σ). Среднее квадратическое отклонение показывает, насколько в среднем каждое значение признака отклоняется от среднего. Геометрически, при нанесении на график, сигма показывает, насколько кривая распределения «размыта» относительно среднего.

Сигма используется как мера вариации, совершенно аналогично $\bar{d}$, от которого отличается главным образом тем, что отклонения квадратичны, а их среднее – линейно. Однако извлечение корня не может полностью уничтожить влияние предшествующего возведения в квадрат; эффект взвешивания частично сохраняется.

Вычисление среднего квадратического отклонения (σ). В принципе среднее квадратическое отклонение является несколько более сложным показателем, нежели среднее линейное отклонение, требуя дополнительного возведения отклонений в квадрат и извлечения квадратного корня из их среднего. Вообще, если нам известна дисперсия, то сигма находится очень

[14] Стат. збірник: показники роботи закл. освіти та наук. установ обл. за 1999 рік. [за заг. редакцією О. Л. Сидоренка, А. С. Доценка, П. С. Дементьєва]. – Х., 2000. – 82 с.

просто, путем извлечения квадратного корня, а именно: $\sigma = \sqrt{\delta^2}$.

Если же дисперсия распределения неизвестна, то среднее квадратическое отклонения для этого распределения находится по следующим формулам:

1. Для первичного ряда:

$$\sigma = \sqrt{\frac{\sum(x_i - M[X])^2}{n}} .$$

2. Для вариационного ряда:

$$\sigma = \sqrt{\frac{\sum(x_i - M[X])^2 * n_i}{N}}.$$

3. Для интервального ряда:

$$\sigma = \sqrt{\frac{\sum(x_{ci} - M[X])^2 * n_i}{N}}.$$

При этом

- X_i – значение переменной (ее величина);
- x_{ci} – средняя точка интервала (*в случае интервального ряда*)
- $M[X]$– среднее арифметическое;
- n – число вариант признака;
- n_i – частота значения переменной (*или частота интервала в случае интервального ряда*);
- N – сумма всех частот.

«Среднее линейное VS среднее квадратическое». Для описательных целей прием отбрасывания знаков при вычислении $\overline{d}$ является совершенно законным. Поскольку $\overline{d}$ измеряет отклонения без возведения в квадрат, то оно по абсолютной величине меньше, нежели σ, которое непропорционально увеличивает большие отклонения в результате возведения в квадрат. Однако пренебрежение знаками делает $\overline{d}$ непригодным для использования в последующих алгебраических вычислениях, независимо от его начала отсчета. Именно поэтому одним из наиболее распространенных средств статистического анализа в течение почти столетия была σ, причем не только как мера дисперсии, но и как составная часть более сложных вычислений. Широкое использование сигмы в какой-то мере объясняется двумя присущими ей достоинствами. *Во-первых,* σ дважды отражает величину каждой переменной распределения: а) точка отсчета, от которой замеряются

отклонения (*M[X]*), сама является репрезентацией всех переменных; б) каждая величина как таковая представлена квадратичным отклонением. *Во-вторых*, возведение отклонений в квадрат автоматически снижает проблему знака отклонения.

Относительные показатели вариации.

Линейный коэффициент вариации. Как уже говорилось, любое отклонение имеет смысл только лишь тогда, когда известно, от чего оно будет «отклоняться», то есть лишь после того, как будет задано начало отсчета или норма. Этот принцип воплощен в коэффициенте вариации, который выражает меру вариации через процентное отклонение от начала отсчета, независимо от того, является ли оно медианой или средним арифметическим. В том случае, если $\bar{d}$ отсчитывался от среднего арифметического, формула коэффициента вариации в знаменателе будет иметь *M[X]*, если же он основан на медиане, формула коэффициента вариации в знаменателе будет иметь *Me.*

$$V_d = \frac{\bar{d}}{\text{M[X]}} \text{ *100 (}\textit{если мы хотим отобразить коэффициент в \%\%}\text{).}$$

При этом

- $\bar{d}$ – среднее линейное отклонение;
- M[X] – среднее арифметическое (в делителе может быть и *Me,* в случае неравномерного, асимметричного распределения).

Подобно среднее квадратическое отклонение также может быть превращено в меру относительной вариации посредством нормирования его по отношению к собственному началу отсчета, то есть среднему арифметическому:

$$V_\sigma = \frac{\sigma}{M[X]} \text{ *100 (}\textit{если мы хотим отобразить коэффициент в \%\%}\text{).}$$

При этом

- δ – среднее квадратическое отклонение;
- $M[X]$ – среднее арифметическое для данного распределения.

Коэффициент вариации является показателем изменчивости признака относительно его средней величины. К примеру, в результате соответствующих вычислений, мы получили V_σ = 0,8. Если выразить его в процентах – имеем 80%. А это значит, что только 20% всего распределения по данному признаку

относительно однородно и приближено к среднему значению. Остальная же часть распределения неоднородна, 80% всех значений очень сильно отличаются от среднего и «далеко» рассеяны по отношению к этому среднему.

Коэффициенты вариации оказываются особенно полезными в процедурах сравнения, поскольку они не зависят от абсолютных значений и от употребляемых единиц измерения. Коэффициент вариации позволяет сравнивать множества малых и больших однородных величин, а также до определенной степени и качественно отличных объектов. Однако *V применим только в тех случаях, когда: (1) наблюдаемые значения имеют нуль; (2) все интервалы равны.* Кроме того, *V* более целесообразно использовать при сравнениях между последовательностями связанных данных. Относительная вариация заработной платы на Востоке Украины может быть меньше, чем на Западе Украины. При измерении симпатии публики к некоторому композитору высокий *V* будет получаться при большом расхождении в мнениях; низкий коэффициент, наоборот, отражал бы тенденцию согласия.

Сопоставление мер средней тенденции и вариации, интерпретация результатов такого сопоставления. Следует подчеркнуть, что малое значение σ при большом среднем указывает на большую однородность данных и в силу этого на типичность среднего, что в некоторых условиях крайне существенно. Среднее, равное 125, при σ=5 и *V*, равным 4%, более репрезентативно, нежели среднее в 125, при σ = 25 и *V* равным 20%. *V*, равное нулю, указывает на отсутствие вариации вообще. Следует отметить, что в той мере, в какой σ увеличивает вариацию относительно среднего арифметического, *V* соответственно увеличивает относительную вариацию.

2.7. Вариация качественных переменных

Вариация качественных переменных. Очевидно, что для номинальных признаков некорректным является использование всех приведенных выше мер разброса. У качественных переменных не существует «нуля» отсчета, и, следовательно, они не имеют величины. Не существует среднего значения диапазона и промежуточных интервалов. Следовательно, не существует и арифметических отклонений.

Это, однако, не означает, что любая группа качественных переменных состоит из совершенно идентичных событий. Попытаемся понять, как можно интерпретировать такой разброс. Два события можно считать различными, если они не обладают различными качествами. Вместо вычисления величин, подсчитываются различия в качествах. Чем больше число различимых пар событий, тем более неоднородна совокупность, и, следовательно, тем больше вариация внутри ее. Аналогично, чем меньше это число, тем больше однородность внутри совокупности и меньше вариация. Поэтому разумно установить показатель качественной вариации по полному числу различных пар событий данного множества. Вопрос теперь лишь в том, *во-первых,* как

подсчитать полное число различий и, *во-вторых*, как превратить это число в компактный показатель.

Чтобы найти полное число различий, суммируются всевозможные различия в группе событий. Например, в множестве из шести мальчиков и шести девочек каждый из шести мальчиков будет отличаться по своим признакам от каждой из шести девочек, давая в итоге 36 различий полов. Если бы имелось девять мальчиков и три девочки, то каждый из девяти мальчиков отличался бы от каждой из трех девочек, что давало бы в итоге 27 различий. В группе 12 мальчиков очевидный результат отсутствия различий был бы получен при умножении 12 на ноль.

Очевидно, что процедура определения полного числа различий сводится к следующему правилу: умножаем частоту каждого признака на частоту каждого отличного от него признака и суммируем эти произведения. Например, в совокупности из четырех католиков, пяти христиан и шести иудеев будем иметь: (4×5) + (4×6) + (5×6) = 74 различия.

Коэффициент качественной вариации (V_q). Число различий, как показатель вариации, сравнимо только с максимально возможным числом различий. Это максимальное число различий будет наблюдаться в том случае, когда все частоты различных признаков равны. Таким образом, максимум вычисляется путем приравнивания частот (т. е. вычисления средней частоты), перемножения частот и суммирования произведений. Другими словами, осуществляются следующие операции: (1) находится средняя частота; (2) этот результат возводится в квадрат; (3) квадрат умножается на число возможных пар признаков. В вышеупомянутом примере девяти мальчиков и трех девочек максимально возможное число различий пола в группе из 12 было бы: 6 (мальчиков)×6 (девочек) = 36, или в этом конкретном случае средняя частота умножается на саму себя.

Относительная величина вариации теперь может быть измерена с помощью отношения между наблюдаемым числом различий его гипотетическим максимумом:

$$Коэффициент\_качественной\_вариации(V_q) = \frac{Полное\_число\_набл.\_различий}{Макс.\_возможн.\_число\_различий}$$

Результат вычислений по соответствующей формуле представляет собой долю всех значений признака, которые сильно неоднородны. Очевидно, что этот показатель принимает значения от 0 до 1. Чем больше коэффициент вариации приближен к нулю, тем меньше вариация значений признака. Как и в случае с вариацией количественных данных, полученное число, которое не может превышать единицы и опускаться ниже нуля, можно представить в %%, умножив полученный результат на 100.

Проиллюстрируем применение данной формулы на предыдущем примере

с 9 мальчиками и 3 девочками: $V_q = \frac{27}{36} = 0{,}75(\times 100) = 75\%$. Этот показатель говорит о том, что вариация значений признака довольно высока, и только 25% (100% минус 75%) относительно однородны.

Среднее число членов каждой из трех упомянутых выше религиозных групп равно пяти. Умножив «5» на «5» и просуммировав эти три произведения, найдем, что максимальное число различий должно быть равно 75. Наблюдаемые различия, как уже было вычислено, равны 75. Следовательно: $V_q = \frac{74}{75} = 0{,}99(\times 100) = 99\%$. Полученная цифра говорит о предельно высокой неоднородности всех значений признака.

Примеры использования коэффициента вариации. Как уже говорилось, рассматриваемый нами коэффициент может быть использован для сравнения тех или иных относительных величин. Например, попытаемся сравнить, насколько вырос/упал уровень «остепененности» системы высшего образования Украины с 1996 по 2000 год (см. *Табл. 2.11*).

Таблица 2.11

Динамика численности основного профессорско-преподавательского состава высших учебных заведений Украины III–IV уровня аккредитации

Годы		Профессора	Доценты
1995	1996	29597	5728
1999	2000	28540	6546

В Украине в 1995/96 учебном году в высших учебных заведениях III–IV уровня аккредитации работало 29597 профессоров и 5728 доцентов, следовательно, максимально возможное число различий будет равно $((29597 + 5728)/2)^2 = 17663^2$, тогда $V_q = \frac{29597 \cdot 5728}{17663^2} = 0{,}54(\times 100) = 54\%$.

В 1999/2000 учебном году ситуация была такова:

$$V_q = \frac{28540 \cdot 6546}{17543^2} = 0{,}61(\times 100) = 61\%.$$

Элементарное нормирование. Необходимость нормировки. Любое событие исследователь рассматривает не изолированно, а в сравнении с конкретной нормой, вытекающей из социальной основы данного события. Например, годовой доход в 3500 гривен воспринимается социологом не как отвлеченное число, а как социальное явление, отнесенное к стандарту, основанному на опыте исследователя; факт рождения ста человек в общности имеет смысл лишь в связи с такими данными, как количество населения, период времени, число рождений в предшествующий год или число рождений в других общностях. Если норма, с которой производится сравнение, не установлена точно, то исследователь невольно установит ее самостоятельно.

Еще большие трудности возникают при сравнении двух и более величин, взятых из различных совокупностей. Например, сравнение умственных способностей мужчин и женщин, проводимое на основе тестов, может быть ошибочным, поскольку известно, что результаты таких тестов зависят от уровня образования, который по полу может быть распределен неравномерно.

Одна из важнейших функций математической статистики по отношению к социологии заключается в предоставлении метода, позволяющего обоснованно сравнивать несколько величин. Существует много способов решения этой задачи; некоторые из них будут рассматриваться в данном подразделе. Совокупность этих процедур можно назвать операциями нормировки, поскольку они устанавливают определенные стандарты наблюдаемых величин. Процесс нормировки уже известен читателю из опыта расчета вариаций, и в настоящей главе остановимся на некоторых других аспектах этой важнейшей статистической процедуры, особенно в применении к социологическим материалам.

Можно осуществлять нормировку приблизительно в следующем порядке сложности: *1) процентные отношения; 2) пропорции; 3) степени; 4) индексы; 5) подклассификация; 6) стандартизация*.

Процентные отношения. Простейшая форма нормировки состоит в приведении рядов абсолютных чисел к стандартной численной основе. Громоздкие абсолютные числа заменяются указанием их отношения к некоторой основе, выраженной в процентах. Вместо того чтобы указывать, что зарегистрированное число юношей и девушек − студентов вуза равно соответственно 9244 и 4622, превращают эти значения в 66,7 и 33,3%. Эта операция настолько привычна, что основной принцип, на котором она основана, не всегда полностью осознается.

Пропорции. Можно сравнивать две величины в форме отношения или выражать одну из них как кратное другой. Отношения бывают различными по составу: можно выделить отношения «часть − часть» частот в пределах одного и того же множества и отношение «целое к целому» между частотами двух взятых переменных. Таким образом, соотношение полов может рассматриваться как отношение «часть − часть», оно сравнивает число мужчин в данной совокупности с числом женщин.

В 1999 году в Украине было 34 млн человек городского населения и 16,1 млн человек − сельского. Соотношение между этими громоздкими числами легче запомнить, если выразить его в следующем виде: 2,1 городского жителя приходится на 1 сельского.

В антропометрии цефалический индекс представляет собой отношение «целого к целому» двух черепных мер − ширины к длине − с целью количественного различия между круглой и овальной формами головы. Отношение умножается на 100, чтобы сделать запись более ясной:

$$Цефалический\_индекс = \frac{ширина}{длина} \times 100\%.$$

Аналогично этому степень умственного развития равна отношению между умственным и хронологическим возрастом испытуемых, что позволяет сравнивать людей различных возрастов по умственному развитию. Таким образом:

$$Iq = \frac{Умственный\_возраст}{Хронологический\_возраст} \times 100\% .$$

Это отношение равно 100 в том случае, когда хронологический возраст и умственный оказываются равными. Другие общепринятые пропорции, используемые в социально-экономическом анализе, – это пропорции: люди – жилье; население – земля; дети – взрослые.

Степени (коэффициенты). Степень является по существу арифметическим средним. Она представляет собой среднее число значений одной переменной, выраженное в единицах другой. Так, степень, равная 20 километрам на 1 литр, есть среднее потребление топлива, при котором расстояние в километрах будет принимать определенное значение при замене одного литра другим. Хотя все степени основаны на прошлых наблюдениях, они обеспечивают предсказание будущего. Поэтому иногда вычисленные степени и средние могут рассматриваться как ожидаемые величины. Будучи в основном результатом большого числа наблюдений, степень часто оказывается эффективным инструментом социологического анализа. Многие степени стали общепринятыми понятиями в области социологии: степень семейности, степень преступности, степени рождаемости и смертности и многие другие видоизменения этих степеней, которые являются рабочими понятиями в социологии.

Статистический смысл степени зависит в основном от двух переменных: от проблемной переменной и от нормированной переменной. При вычислении степеней наиболее важным является выбор нормированной переменной, с которой будет сравниваться проблемная переменная. Например, при вычислении степени рождаемости необходимо выбрать нормирующую совокупность, с которой в дальнейшем будет сравниваться абсолютное число актов рождения. Для этой цели можно было бы использовать либо полную совокупность людей, либо число женщин, достигших зрелого возраста, либо число замужних женщин, достигших зрелого возраста. Наиболее часто используется, хотя и не вполне оправданно, полная совокупность людей: мужчин, женщин, детей.

Следующим этапом построения степени является выбор стандартной численной основы: 10, 100, 1000 или кратное им значение. Назначение числовой основы состоит просто в указании десятичного масштаба для удобства табулирования, для облегчения цитирования и более быстрого понимания. Численный масштаб часто устанавливается по соглашению, особенно когда не существует другого выхода, кроме простого подчинения установившейся традиции. Так, степень рождаемости, равная 24, имеет интернациональный смысл и означает 24 случая рождений на 1000 человек

всего населения в данном году и на данной территории. Такое представление воспринимается более ощутимо, нежели запись: 768 из 32462. Вычисление осуществляется следующим образом: число рождений = 768, полное население = 32462, числовой масштаб = 1000,

$$Степень\_рождаемости = \frac{768}{32462} \times 1000 = 24.$$

Обобщенная формула читалась бы следующим образом:

$$Степень = \frac{Частота\_проблемной\_переменной}{Частота\_нормировочной\_переменной} \times Числовой\_масштаб.$$

В обозначениях:

$$Степень = \frac{r_{nn}}{r_{mn}} \times r_m,$$

где
- r_{nn} – частота задачи;
- r_{mn} – нормирующая частота;
- r_{nn} – числовой масштаб.

Ни нормирующая переменная, ни числовой масштаб не определены твердым соглашением, как, например, для степени рождаемости или смертности. В таких случаях допускается некоторая свобода действий, однако результаты должны снабдиться примечаниями. Степень разводов может быть вычислена как для всего населения, так и для числа супружеских пар в один и тот же год и на одной площади, или даже для числа браков в течение предшествующих десяти лет, как это иногда делается для большинства разводов. Степени преступности могут вычисляться для определенного возраста и конкретных половых групп; степень браков – лишь для возрастной группы от 14 лет и более.

Из предыдущих примеров можно сделать вывод о том, что нормированная переменная должна иметь те же характеристики, что и проблемная переменная – вместе они объединяются общим названием «открытая группа». Если смерть может случиться с каждым человеком, то рождение ребенка, женитьба или замужество и развод – не с каждым. Следовательно, степень, основанная на разумно выбранной открытой совокупности, менее подвержена искажениям из-за внешних факторов. Степени рождаемости, браков и разводов, вычисленные по отношению к полному населению, обычно называются приближенными степенями, а степени, вычисленные для особых групп, обозначаются как «удельные» или «уточненные». Преимущество приближенной степени состоит в ее простоте и удобстве для ориентировочных расчетов. Уточненные степени необходимы для профессионального социологического исследования.

Индексы. В статистике индекс – термин, используемый в разговорном

языке и технике для самых различных типов мер, – обычно относится к более сложным степеням или множествам пропорций. В качестве вторичной меры он обычно предназначается для описания вариации, которая в непосредственном виде могла бы быть совершенно незаметной. В более формализованном варианте он обычно описывает отношение между двумя величинами, одна из которой взята в качестве нормы, или ожидаемой величины, тогда как другая является измеряемой величиной.

Так, индекс стоимости жизни сравнивает цены в конкретном году со средними ценами для «нормального» года. Индекс, равный 139 в 1995 году, при использовании в качестве основного года – 1991 показывает, что стоимость жизни повысилась на 30% по сравнению с основным годом, для которого стоимость принята равной 100. Хотя такой обманчиво простой индекс может бойко цитироваться любым журналистом, его внутреннее содержание, включающее охват, взвешивание, метод усреднения наблюдений, а также выбор основного периода, свидетельствует о его статистической сложности. Аналогично этому V_q сравнивает наблюдаемое число различий признаков заданного множества с максимально возможным числом различий, которое в данном случае служит в качествt нормы.

Таблица 2.12

Вычисление индекса социально-экономической классификации, студентов университета и населения региона

Группа	Университет		Регион %	Разность	Индекс
	кол-во	%%			
Техники	408	19,3	4,7	14,6	411
Инженеры, управленцы	507	24,0	4,7	19,3	511
Бизнесмены	290	13,7	5,0	8,7	274
Клерки	286	13,5	12,8	0,7	105
Фермеры	196	9,3	15,9	–6,6	58
Квалифицированные рабочие	267	12,6	17,0	–4,4	74
Полуквалифициро-ванные рабочие	94	4,5	19,9	–15,4	23
Неквалифицированные рабочие	65	3,1	20,0	–16,9	15
ИТОГО	2,113	100%	100%		

Таблица 2.12 показывает построение и использование индекса для измерения социальной стратификации студентов университета, а также то, насколько представлены в нем различные социальные классы. Логическая основа построенных в этой таблице индексов следующая: дочери состоятельных родителей составляют 19,3% от общего числа девушек

в университете, тогда как в другом регионе самая состоятельная группа составляет 4,7%. Соотношение между такими парами процентных отношений могло бы служить мерой доступности обучения для различных классов. Существует два метода, с помощью которых можно было бы измерять эти различия:

1) простого различия между процентными отношениями;

2) нормированных индексов.

Что касается первой альтернативы, то анализ различий обнаруживает, что при перемещении по социальной шкале различия возрастают. Однако эти различия являются абсолютными, в силу чего их невозможно нормировать по величине или началу отсчета. Чтобы нормировать данные различия, строится индекс, что составляет содержание второй альтернативы.

Таким образом, если бы посещаемость университета распределялась случайным образом между всеми классами населения то можно было бы ожидать, что состоятельная группа, которая составляет 4,7% всего населения, должна дать вклад, равный 4,7% от всего числа всех студентов. Ожидаемая и наблюдаемая доля студентов из состоятельных семей были бы в данном случае идентичными, соотношение между ними было бы равно единице. В действительности, состоятельная часть студентов университета равна 19,3%, что равно $\frac{19,3}{4,7}\cdot 100 = 41,1\%$ соответствующего процента для региона, или в 4,11 раза больше ожидаемого значения. Таким же образом можно нормировать все другие социально-экономические процентные отношения.

Другой подход к определению индекса заключается в нормировке последовательности величин относительно их среднего. В этом случае индекс представляет собой просто отношение между данной величиной и средним значением последовательности. Например, средняя степень смертности в ряде городов равна 10,5. Если все факторы, влияющие на степень смертности, были бы одинаковыми во всех городах, тогда все степени смертности станут идентичными и, следовательно, будут равны среднему значению последовательности. Поскольку при данном предположении среднее является ожидаемым или теоретическим значением, то, чтобы оценить вес факторов, влияющих на различия, индивидуальные степени измеряются по отношению к среднему. Например, если наблюдаемая степень смертности для города равна 7, то мы могли бы вычислить индекс следующим образом:

$$ИНДЕКС = \frac{наблюдамая\_степень}{ожидаемая\_степень}\times 100\% = \frac{7}{10,5}\times 100\% = 67\%.$$

Это значит, что степень смертности в данном городе составляет 67% от средней. С помощью этого метода любой город можно расположить на шкале отношений к среднему. Этот прием нормировки аналогичен V, поскольку он выражает исходные значения через их собственное среднее.

По аналогии находится и, так называемый, сезонный индекс смертности

населения, который вычисляется, как отношение между месячной мерой и среднегодовой мерой. Этот индекс используется для измерения флуктуаций рождаемости, смертности, промышленной продукции и некоторых других экономических показателей.

Нормировка посредством подклассификации. Подобно тому, как отдельное абсолютное значение не имеет смысла до сопоставления с соответствующей нормой, точно так же и степень или процентное отношение практически не имеет социологического смысла, пока они рассматриваются самостоятельно. Они приобретают смысл тогда, когда их сопоставляют с аналогичными степенями, например, сравнивая степень рождаемости двух регионов и степень бракосочетаний католиков и христиан. Однако не следует делать слишком поспешного вывода о существовании причинно-следственного соотношения между такими спаренными переменными. Наблюдаемые вариации степеней могут иногда возникать в результате действия факторов, не учтенных в классификации. Такие факторы можно назвать скрытыми. Во многих случаях сравнение двух или большего числа наблюдаемых величин искажаются в результате воздействия именно таких скрытых факторов. Так, более высокая степень рождаемости в одном регионе может произойти не в силу большей плодовитости его населения, а из-за большего процентного количества женщин, способных к деторождению. Этот случайный фактор не классифицирован в приближенной степени.

Под нормировкой посредством подклассификации обычно подразумевается разделение факторов на «внешние» и «внутренние», причем внешние факторы не должны изменяться в ходе исследования. Из всего сказанного следует, что нельзя распространять на все группы результаты, полученные для соответствующих подгрупп, ведь они отличаются не только весом, но и другими факторами. Поэтому необходимо разработать метод, который позволил бы получить простую, уточненную, но свободную от влияния весов, степень. Соответствующий метод получил название «стандартизация», а получаемая степень – «стандартизованная».

Было обнаружено, например, что повышенную степень преступности населения можно объяснить более высокой долей молодежи в нем. Именно этим, а не избыточной тенденцией к совершению преступлений, объясняется повышенная степень преступности местного населения. В целом же стандартизованные степени преступности для каждой группы населения строятся следующим образом: вычисляют ожидаемое число преступлений для местного населения при условии, что оно имеет такой же возрастной состав, что и другие поселенческие группы. Другими словами, необходимо действовать так, как будто бы все интересующие нас поселенческие группы, имеют одинаковое распределение по возрасту.

Хотя стандартизация кажется полностью формализованной процедурой, нет никакой формулы, которая предписывала бы, насколько подробной и какой именно должна быть подклассификация. Поэтому социолог не освобождается

от содержательного анализа задачи. Например, возраст можно было бы подклассифицировать более чем на три интервала; чем больше число подразделений, тем больше точность сравнений. Однако существуют практические ограничения, за пределы которых распространять подклассификацию нет необходимости. Иногда достаточно самого грубого разделения возраста на три интервала, чтобы установить важность возрастного фактора в степени преступности.

Применение стандартизации не ограничивается степенями и процентами. Любой вид арифметического среднего может быть стандартизован при наличии необходимых данных для подклассификации. Неограниченные возможности стандартизации напоминают еще раз, насколько далека «окончательная истина» от данных, которые лежат перед нами и на которых, тем не менее, часто основываются мнения и действия.

Довольно часто приближенные степени, которые необходимо превратить в нормированные, соответствуют большим территориям и охватывают большие интервалы времени. Трудно сравнивать статистику разных стран, если отсутствует всеобщее соглашение о стандарте. В качестве такого стандарта в 1901 г. часто применялся английский «стандартный миллион». Так как распределение населения по возрасту – один из наиболее искажающих факторов, в интерпретации социальной статистики «стандартный миллион» – есть возрастное распределение одного миллиона британского населения. Такая процедура нормирует степени по возрасту, тем самым превращая их в величины, удобные для сравнения.

Перекрестные таблицы обычно создаются с целью выявления статистических ассоциаций, однако из-за присутствия скрытых факторов, полученные значения ассоциаций не следует рассматривать как безупречно достоверные. Чтобы выявить скрытые факторы, можно подклассифицировать перекрестные таблицы. Для иллюстрации рассмотрим данные (фиктивные) для 52 районов, перекрестно классифицированных близостью к определенному типу производства и по степени правонарушений (см. *Табл. 2.13*).

Таблица 2.13

Приближенные коэффициенты преступности, промышленных городских и сельских районов (возраст населения от 15 до 75 лет)

Тип района	***Коэффициент преступности***					
	Количество			*Процент*		
	Высокий	Низкий	Сумма	***Высокий***	Низкий	Сумма
Промышленный	15	8	23	65	35	100
Сельский	10	19	29	34	66	100
ИТОГО	25	27	52	48	52	100

Эта таблица указывает, что промышленные районы чаще характеризуются высокими степенями правонарушений, чем районы сельских

жителей, указывая, как будто на статистическую связь между местом проживания и преступностью. Эта связь выявляется более четко в процентном распределении, которое показывает, что 65% всех промышленных районов находятся в категории «высокой преступности», тогда как подклассифицированные аналогичным образом сельские районы составляет в этой категории только 34%. С точки зрения приближенной степени, вывод о связи между местом проживания и преступностью кажется убедительным. Однако такой вывод неприемлем для любого специалиста по социальной патологии города. Он указал бы, что жители одних районов сосредоточены в регионах с низким уровнем жизни, тогда как жители других районов чаще проживают в более благоприятно расположенных районах с относительно высоким уровнем жизни. Разумно поэтому спросить, будет ли сохраняться разница между городскими и сельскими районами в отношении преступности, если эти районы нормировать на одинаковый экономический уровень. Чтобы ответить на этот вопрос, необходимо подклассифицировать районы согласно жизненным стандартам и провести сравнения в пределах одинаковых социально-экономических подклассов (см. *Табл. 2.14*).

Таблица 2.14

Коэффициенты преступности, отнесенные к возрасту (15–75 лет) и уровню жизни респондентов, проживаючих в промышленных и сельских районах

Тип района	***Уровень жизни***					
	Высокий уровень			*Низкий уровень*		
	Коэффициент преступности			*Коэффициент преступности*		
	Высокий	Низкий	Итого	Высокий	Низкий	Итого
Промышленный	1	6	7	14	2	16
Сельский	3	18	21	7	1	8
Итого:	4	24	28	21	3	24
Процентное распределение						
Промышленный	14	86	100	88	12	100
Сельский	14	86	100	88	12	100
Итого:	14	86	100	88	12	100

Анализируя эту таблицу, можно увидеть, что связь между местом проживания и преступностью исчезает: из 24 экономически худших районов 21 район или 88% находятся в категории высокой преступности (и это справедливо как для промышленных, так и для сельских районов). Из 28 районов, более развитых экономически, только 4 или 14% находятся в категории высокой преступности независимо от типа района проживания. В результате в этой гипотетической иллюстрации преступность полностью зависит от экономического уровня и совсем не зависит от типа района проживания. Такая связь называется ложной, так как она фактически является результатом действия скрытого социально-экономического фактора, который дает «подлинное» объяснение.

В основном подклассификация – процедура для уточнения сравнений,

она, подобно анатомическому расчленению, является инструментом для более полного и глубокого статистического анализа. Подклассификация отличается от стандартизации тем, что она чисто описательна и имеет столько же показателей, сколько выбрано подклассов. Стандартизованная степень, с другой стороны, является скорее гипотетической, чем описательной; это единый, совокупный показатель, взвешенный по ряду частот подклассов, используемых как стандарт.

Вопросы и задания для самоконтроля

1. Дайте определения следующим терминам: первичный, вариационный и динамический ряд, дискретный и интервальный ряд.

2. Какова разница между частотой и частостью?

3. Дайте определения следующим терминам: абсолютная, относительная, накопленная частота, совместная частота.

4. Распределите показатели из приведенного ниже списка по принципу: абсолютные/относительные частоты: 28 кг; 35%; 246 шт.; 0,45; 55 мин; ⅓.

5. В предыдущем месяце на производственном предприятии была осуществлена аттестация рабочего персонала с применением тестовых методик. В итоге 40 человек из общего числа протестированных работников − 800 аттестацию не прошли. Было принято решение, что этой группе работников необходимо пройти месячный курс интенсивной переподготовки и повышения квалификации. Через месяц с данной группой работников было проведено повторное тестирование и оказалось, что у 3,5% данной группы уровень квалификации все-таки остался на прежнем уровне либо вырос несущественно. Сравните доли работников, не прошедших аттестацию в прошлом месяце и в текущем. Исходя из этого, наметьте кадровые и производственные перспективы предприятия.

6. Дайте определения: перекрестная классификация, перекрестная таблица (матрица), одномерное распределение, двумерное, многомерное распределение, независимые переменные, временной ряд, кумулята и кумулятивный временной ряд.

7. Что такое классификация и группировка? Объясните место и роль метода классификации и группировки в социологическом исследовании.

8. Какие задачи в исследовании совокупностей не могут быть решены с помощью простой группировки? Каковы разновидности сложной группировки?

9. В каких случаях используются неравные интервалы? Какой вид группировки при этом предпочтителен?

10. Какие типы статистических таблиц вам известны? Почему статистическая таблица должна быть легко обозримой и иметь небольшие размеры?

11. Дайте определения: гистограмма, полигон распределения, правило нулевого начала, разрыв шкалы, многозначный график, арифметическая шкала, арифметическая временная диаграмма, график отношений, кумулята, диаграмма полос; круговая диаграмма.

12. В чем заключается необходимость построения графических изображений при обработке социологической информации?

13. Выберите наиболее подходящий тип графика и графически представьте данные следующих таблиц:

а)

Пол студентов	*Мужской*	*Женский*
Доля в общем количестве	*42%*	*58%*

б)

Учебный год	*Количество студентов*
1998−1999	*600*
1999−2000	*640*
2000−2001	*690*
2001−2002	*760*
2002−2003	*850*

с)

Учебный год	*Количество студентов*		
	Муж.	*Жен.*	*Всего*
1998−1999	*270*	*330*	*600*
1999−2000	*300*	*340*	*640*
2000−2001	*340*	*350*	*690*
2001−2002	*380*	*370*	*760*
2002−2003	*440*	*410*	*850*

*Какие выводы можно сделать на основании анализа графиков **b** и **c**? Подумайте и прокомментируйте, как в целом можно проанализировать графики, какие важные выводы можно сделать на их основе?*

14. Дайте определения: характеристики положения, среднее арифметическое (простое и взвешенное), квантили, медиана, медианный интервал, квартиль, дециль, центиль, мода, модальный интервал, модальная частота.

15. Каково будет комбинированное среднее двух групп, если среднее из 100 событий первой будет равно 10, а среднее из 50 событий второй будет равно 15? Каково было бы комбинированное среднее, если бы каждая группа состояла из 50 событий? Из 100 событий?

16. Для приведенного ниже ряда рассчитайте среднее арифметическое (простое и взвешенное), моду и медиану. Есть ли различия в данных показателях? Если есть, то как можно их объяснить?

8 8 7 5 2 8 7 8 9 8 5 8 2 6 8 2 8 4 1 8

17. В течение двух недель в кинотеатре посчитывали количество зрителей и получили следующие данные, представленные в табл. 1

Таблица 1

Данные посещаемости кинотеатра

День недели	*Утренние и дневные сеансы*	*Вечерние сеансы*
1-я неделя		
Понедельник	*376*	*520*
Вторник	*420*	*560*
Среда	*397*	*590*
Четверг	*440*	*558*
Пятница	*432*	*614*
Суббота	*668*	*748*
Воскресенье	*684*	*711*
2-я неделя		
Понедельник	*567*	*712*
Вторник	*611*	*769*
Среда	*581*	*787*
Четверг	*645*	*765*
Пятница	*641*	*837*
Суббота	*892*	*961*
Воскресенье	*878*	*930*

По приведенным в таблице данным рассчитайте следующие показатели: среднее количество зрителей в день и в неделю, в будние и выходные дни, на утренних, дневных и вечерних сеансах, в первую и во вторую неделю. Проанализируйте полученные результаты и сделайте выводы: как изменяется количество зрителей в зависимости от дней недели и времени сеансов, отличается ли количество зрителей в первую и вторую неделю и т. п.

18. По данным Таблицы 1 для каждой недели рассчитайте следующие показатели вариации: дисперсию, среднее квадратическое отклонение, коэффициент вариации. На этой основе сделайте выводы, которые могут заинтересовать администрацию кинотеатра, оказаться полезными в управлении данным заведением.

19. Таблица 2 показывает процент произведений каждого из шести композиторов в репертуаре (например, Киевского оркестра, Харьковского оркестра и т. д.). Вычислите медианный процент для каждого композитора, среднее линейное отклонение, основанное на медиане, и V для каждого композитора. Классифицируйте композиторов в соответствии с их популярностью и интерпретируйте результат.

Таблица 2

Показатели популярности выдающихся композиторов в разных городах Украины

Композитор	*Оркестры городов*							
	Киев	*Харьков*	*Львов*	*Одесса*	*Донецк*	*Ивано-Франковск*	*Днепропетровск*	*Симферополь*
Бах	*0,8*	*4,9*	*2,0*	*1,8*	*2,3*	*2,9*	*3,3*	*1,4*
Бетховен	*9,6*	*10,9*	*8,1*	*11,0*	*11,4*	*10,3*	*10,7*	*9,3*
Берлиоз	*2,5*	*1,6*	*1,1*	*0,7*	*2,6*	*2,3*	*0,6*	*0,1*
Брамс	*9,2*	*9,4*	*7,4*	*11,2*	*11,4*	*9,8*	*10,8*	*11,9*
Прокофьев	*1,2*	*1,2*	*0,5*	*1,1*	*1,1*	*1,1*	*1,8*	*1,0*
Чайковский	*4,5*	*5,9*	*7,3*	*9,6*	*2,8*	*5,4*	*6,5*	*9,3*

20. Объясните, почему среднее квадратическое отклонение, а не вариация, обычно используется как мера дисперсии.

21. Если средний возраст студентов вуза равен 20 годам, а среднее квадратическое отклонение равно 2, то каким будет среднее и сигма (σ) этой группы двадцать лет спустя? Чему будет равен V?

22. Население города состоит из 50% мужчин и 50% женщин; 70% – украинцев, 30% россиян. Можно ли представить переменные одним V_σ? Обоснуйте ответ.

23. Дайте определения: коэффициент качественной вариации, подклассификация; стандартизация, скрытый фактор, причинный фактор, максимум различий, операция нормировки, отношение, процентное отношение, степень, проблемная переменная, переменная нормировки, числовая основа, открытая группа, индекс, ожидаемая величина.

24. Является ли V_q стандартизованной мерой? Аргументируйте ответ.

25. В XVIII веке европейские церкви вели записи смертей и рождений, которые впоследствии были использованы учеными для оценки размеров городов. Так, в Берлине приблизительно в 1700 году произошло около 178 смертей. Оцениваемое отношение числа смертей к общему населению равно 1:35. Вычислите степень смертности на 1000. Оцените размер города Берлина в то время.