Глава I

ИЗМЕРЕНИЕ И АНАЛИЗ РАСПРЕДЕЛЕНИЙ

1. Об измерении в социологии. Классификация социальных признаков по уровням измерения

Количественный анализ применяется при изучении разнообразных форм движения материи, но необходимым условием его эффективности всегда является предварительный качественный, содержательный анализ изучаемых явлений. Как отмечал Гегель, «качество есть непосредственная определенность и с него следует начинать»[1]. Именно качественный анализ определяет постановку задачи, вычленяет предмет исследования, выбирает способы и средства исследования, в частности адекватные задаче количественные методы, использование которых углубляет, делает более конкретным наше знание.

Количественные методы могут быть применены в исследовании лишь после того, как эмпирические данные переведены на язык чисел. Предпосылкой и началом применения количественных методов в социологических исследованиях является измерение. Обычно под измерением понимается «познавательный процесс, в котором определяется отношение одной (измеряемой) величины к другой однородной величине» принимаемой за единицу измерения»[2]. Однако это определение пригодно лишь для измерения количественных[3] (например, стажа, заработной платы и т.п.), а не качественных признаков (например, удовлетворенности, оценки, ориентации и т.п.), так как здесь нет общепризнанных

[8]

эталонов и единиц измерения. Поэтому имеет смысл расширить понятие измерения, понимая под ним процедуру приписывания чисел значениям признака. Цель измерения — получить числовую модель, исследование которой могло бы заменить исследование самого объекта. Это возможно лишь тогда, когда свойства модели соответствуют свойствам объекта, т.е. отношения между числами, образующими числовую модель, соответствуют отношениям между изучаемыми свойствами объекта.

Итак, мы понимаем под *измерением особую процедуру, в результате которой возникает числовая модель объекта* (точнее, изучаемых свойств объекта). При измерении, таким образом, устанавливается соответствие между свойствами объекта и свойствами сопоставленных им чисел. Набор свойств объекта и сопоставляемых им чисел называют шкалой[4] (свойства объекта трактуются здесь очень широко, в частности, под набором свойств понимаются также и различные степени интенсивности одного свойства).

В естественных науках предполагается, что всегда можно пользоваться всеми свойствами чисел. Это обстоятельство настолько привычно, скажем, для физики, что пользуются им обычно автоматически; при этом получаются вполне корректные следствия.

Аксиомы арифметики поэтому так оправданы в физическом мире, что создавались в результате отражения, пусть не всегда осознаваемого (вспомним, например, положение И. Канта об априорности математического знания) свойств и отношений этого мира. Как писал Энгельс, само «понятие числа заимствовано исключительно из внешнего мира, а не возникло

---

[1] Гегель Г. В. Ф. Соч., т. 5, М., 1937, с. 65.

[2] Философская энциклопедия, т. 2. М., 1967, с. 244.

[3] Количественным называется признак, значениями которого служат числа, допускающие сложение; в противном случае признак называется качественным. (Суппес П., Зинес Дж. Основы теории намерений.— В кн.: Психологические измерения. М., 1967, с. 25). После введения понятий уровней измерения различие качественных и количественных признаков станет более ясным.

[4] В теории измерений под шкалой понимают однозначное отображение эмпирической системы с отношениями в числовую систему с соответствующими отношениями. *(Суппес П., Зиме Дж,* Основы теории измерений..., с. 19; *Пфанцагль И.* Теория измерения. М., 1976, с. 23).

в голове из чистого мышления»[5]. Поэтому математические, в частности арифметические, понятия сохраняют следы своего происхождения[6]. Для физика, например, естественно, что масса в 15 кг в 3 раза больше, чем масса в 5 кг, и на 10 кг больше последней. Это кажется столь очевидным, что воспринимается как трюизм. Когда же мы переходим в область психологии или социологии, ситуация значительно усложняется. Здесь исследователь нередко рискует произвести такую арифметическую трактовку своих

[9]

измерений, которая оказалась бы лишенной всякого смысла[7].

Вот почему со всей определенностью нужно подчеркнуть важность изучения базовых эмпирических отношений, которые в конечном счете определяют допустимые операции с числами, приписанными объектам в каждом конкретном случае. Поясним это примером. Предположим, что мы изучаем удовлетворенность работников своей работой (точнее предприятием, на котором они работают).

Обычно в таких случаях вначале выдвигается содержательная модель данной социальной переменной, скажем, из следующих 5 пунктов:

a) вполне удовлетворен работой;

b) скорее удовлетворен, чем не удовлетворен;

c) промежуточная позиция;

d) скорее не удовлетворен, чем удовлетворен;

e) совершенно не удовлетворен.

В качестве эмпирических референтов соотнесения индивидов с позициями модели могут, например, использоваться ответы на вопросы социологической анкеты. Возможные варианты ответов упорядочиваются по схеме так называемого логического квадрата[8]. Рассмотрим построение шкалы с помощью двух вопросов.

Первый — о переходе на другое предприятие и второй — о возврате (в прожективной ситуации: «Допустим, что Вы некоторое время не работали на заводе. Вернулись бы Вы на него?») имеют варианты ответов: «да», «нет», «не знаю».

Схема «логического квадрата» в нашем случае принимает такой вид:

Варианты ответа на вопрос о переходе	Варианты ответа на вопрос о возврате		
	«Да»	«Не знаю»	«Нет»
«Нет»	*a*	*b*	*f*
«Не знаю»	*b*	*c*	*d*
«Да»	*f*	*d*	*e*

Здесь *a, b, c, d, e,* обозначают соответствующие пункты шкалы, *f* — противоречивые ответы.

[10]

Шкалы могут строиться и на большем числе вопросов. Пунктам шкалы и, следовательно, попадающим туда индивидам, приписываются числа *X,* например: 5, 4, 3, 2, 1. Но можно ли считать, что различие в степени удовлетворенности между работниками, попадающими в позиции «*a*» и «*b*», такое же, как между индивидами, попадающими в «*b*» и «*c*», «*c*» и «*d*»? Можно ли утверждать, что индивиды, попадающие в позицию «*b*», вдвое больше удовлетворены, чем те, которые попадают в позицию «*d*»? Ясно, что ответы на эти вопросы должны быть

[5] *Маркс К., Энгельс Ф.* Соч., т. 20, с. 37.

[6] *Реньи А.* Трилогия о математике. М., 1980, с. 44.

[7] *Решлен М.* Измерение в психологии.— В кн.: Экспериментальная психология. М., 1968, с. 197.

[8] Рабочая книга социолога. М., 1976, с. 232.

отрицательными. Мы не имеем права пользоваться свойствами равенства интервалов и отношений, так как данные свойства не обеспечены соответствующими свойствами объектов: между ними установлено лишь отношение порядка.

В принципе можно приписать позициям числа X' = 2, 1, 0, —1, —2 (что означает применение преобразования $X \rightarrow X' = X - 3$). Числа можно возвести в квадрат ($X \rightarrow X' = X^2$) и вообще: любое монотонное преобразование, не изменяющее последовательности чисел, является в данном случае допустимым. Это обстоятельство необходимо учитывать при выборе статистических мер, осуществлении арифметических операций над числами. И так в каждом конкретном случае.

Приписывание чисел пунктам шкалы, как правило, неоднозначно, т.е. числа допускают определенные группы преобразований, не меняющих их (чисел) свойств.

Тип шкалы можно определить допустимыми группами преобразований ее чисел[9] или допустимыми арифметическими операциями над этими числами[10]. При обоих подходах тип шкалы, или уровень измерения, фактически детерминируется эмпирическими свойствами изучаемой системы.

Теоретически существует бесконечное число типов шкал. Но обычно, когда шкалы различают по уровню измерений — от самых «слабых» к самым «сильным», то выделяют 4 уровня. (4 типа шкал): номинальные (ординарные), порядковые (ординальные), интервальные и, наконец, шкалы отношений (релятивные, или пропорциональные).

Такая классификация, как мы увидим, является одновременно классификацией и по допустимым арифметическим операциям, и по допустимым группам преобразований чисел.

[11]

Чем выше уровень шкалы, тем уже круг допустимых преобразований чисел, тем больше арифметических свойств реализуется и, тем самым, шире применяемый статистический аппарат. Для шкал данного уровня можно использовать статистические меры шкал всех предшествующих уровней, но не наоборот.

Познакомимся в общих чертах с основными типами шкал (после изучения статистических мер мы вернемся к шкалам, рассмотрев принципиальный вопрос классификации мер по уровням измерения признаков).

Номинальные шкалы

Для построения этой шкалы необходимо уметь устанавливать отношение равенства (и неравенства) объектов — в смысле рассматриваемого признака — для распределения изучаемой общности на непересекающиеся, дизъюнктивные классы, каждый из которых является отдельным пунктом шкалы. Исследователь должен найти такие эмпирические индикаторы, с помощью которых любой объект можно соотнести с определенным классом, т.е. позицией на шкале. Иногда эта задача решается просто (или сравнительно просто) — установление принадлежности к нации, полу, вероисповеданию и т.д., но зачастую она оказывается далеко не элементарной. Так, длительные поиски предшествовали выделению О.И. Шкаратаном[11] структурных групп, представляющих пункты номинальной шкалы, по которым распределяются члены такой социальной общности, как современное промышленное предприятие. Напомним эти группы:

I — организаторы производственных коллективов;

II — работники высококвалифицированного научно-технического труда;

---

[9] *Stevens S. S.* On the theory of scales of measurement.— Science, 1946, v. 103.

[10] *Coombs C. H.* Theory and methods of social measurement.— In: *Festlnger L., KatzD.* Research methods In behavioral sciences. N. Y., 1953.

[11] *Шкаратан О. И.* Социальная структура советского рабочего класса.— Вопросы философии, 1967, № 1; *Шкаратан 0. И.* Проблемы социальной структуры рабочего класса СССР, историко-социологическое исследование. М., 1970; *Шкаратан О. И., Рукавишников В. О.* Социальные слои в классовой структуре социалистического общества.— Социологические исследования, 1977, № 2.

III — работники квалифицированного умственного труда;

IV — организаторы первичных производственных коллективов;

V — работники высококвалифицированного труда, сочетающие умственные и физические функции при обслуживании сложной техники;

[12]

VI — работники квалифицированного физического ручного труда;

VII — работники квалифицированного, преимущественно физического труда, занятые на машинах и механизмах;

VIII — работники нефизического труда средней квалификации;

IX — работники неквалифицированного физического труда.

В расположении структурных групп интуитивно угадывается известный порядок, но интуиция, «угадывающая» по-рядок, не доказывает его наличия. При детальном рассмотрении мы видим, что «нисходящее» расположение групп не всегда оправдывается; так и творческий характер труда, и престиж, и заработная плата, например, работников V и VI групп могут быть выше, чем у работников I или IV (можно привести и другие примеры несоответствия этому порядку). Следовательно, шкала структурных групп остается неупорядоченной, фактически она номинальная.

Другой пример построения номинальной шкалы — выяснение причин текучести работников. Здесь увеличение числа классов (пунктов), желательное в принципе для более детального изучения проблемы, нередко приводит к увеличению ошибок, уменьшению надежности получаемых результатов за счет нарушения требования дизъюнктивности, т.е. приводит к появлению пересекающихся классов. Например, в одной из работ по текучести выделяется, в частности, такая причина увольнения — «решил перейти к друзьям»[12]. Очевидно, что причиной перехода здесь могут быть и условия труда, и жилищно-бытовые условия («там, говорят, скорее квартиру получить можно») и т.д. Другой источник возможных ошибок — использование слов, допускающих очень широкое толкование, например, «семейные обстоятельства» и др.

Обычно рассматриваемые классы укрупняются в блоки, *содержательно* непересекающиеся. При исследовании текучести, выделяются, например, такие блоки: 1) неудовлетворенность условиями трудовой деятельности; 2) неудовлетворенность заработком; 3) неудовлетворенность жилищно-бытовыми условиями,

При этом итоговые данные оказываются ненадежными, так как закладываются ошибки при распределении недизъ-

[13]

юнктивных (пересекающихся) классов в непересекающиеся блоки (ошибки первой стадии классификации).

Отметим, что для обоснованного построения не «очевидной» шкалы представляется перспективным применение методов таксономии[13].

Итак, хотя номинальная шкала обеспечивает только самый слабый тип измерения, процедура ее построения зачастую не тривиальна. Единственное требование, предъявляемое к числам, приписываемым различным классам в случае номинальных шкал — быть *различными.* Очевидно, эти числа могут быть подвергнуты любому взаимно-однозначному преобразованию, то есть от чисел X всегда можно перейти к $X'=f(X),$ где $f(X)$ — закон взаимно однозначного сопоставления. В дальнейшем мы будем для краткости обозначать это так: $X \rightarrow X'=f(X).$ Здесь числа играют роль символов, «ярлыков», их вполне можно заменить, например, любыми буквами, или какими-либо другими знаками. И то, что обычно выбирают

[12] Социальные проблемы труда и производства. Москва, Варшава, 1969, с. 229.

[13] См. главу VI.

для нумерации позиций натуральные числа 1, 2, 3, ... диктуется лишь соображениями удобства, привычки.

Порядковые шкалы

Для построения такой шкалы необходимо уметь устанавливать не только отношения равенства между объектами (по данному признаку), но и отношения *последовательности* — порядка. Это отношения типа «больше, чем», «лучше, чем» и т.д. Далее, как мы видели, выдвигается содержательная модель признака (см., например, шкалу удовлетворенности работой). Эмпирическим референтом могут быть специальный тест (например, набор проективных ситуаций), вопрос (или, чаще, система вопросов) социологической анкеты, и т.д. С помощью референтов объекты социальной общности соотносятся с пунктами шкалы. Каждому пункту может быть приписано некоторое число. Между этими числами имеют место те же отношения, что и между объектами. Ясно, что и в случае порядковых шкал приписывание чисел неоднозначно.

Этими числами могут быть и 1, 2, 3, 4, ... и 1, 4,9, 16, ... и 1, 3, 5, 7 ... и т.д., т.е. любое преобразование $X \rightarrow X' = \varphi(X)$, где $\varphi(X)$ — монотонно возрастающая функция,

[14]

которая не изменит свойств чисел, приписанных пунктам (свойствам объекта). Известна лишь их последовательность, но не расстояния между ними. Вообще говоря, расстояния между пунктами шкалы не равны (подчеркиваем, что использование рангов может породить иллюзию равенства!), мы не только не можем сказать, *во сколько раз* одно значение признака больше другого, но и на сколько. Следовательно, и числа фактически не несут такой информации.

Понять это помогает простой пример. Рассмотрим такую порядковую шкалу, как итоговое распределение мест в турнирной таблице спортивных состязаний. Ясно, что в общем случае расстояния между этими позициями разные (например, первый «оторвался» от второго больше, чем второй от третьего и т.д.). Конечно, судьи и болельщики знают расстояния (в очках) между различными позициями. В случае порядковой шкалы мы находимся в положении человека, который знает только распределение мест и не может узнать количество очков, набранных разными участниками.

Отметим, что ранги определяют относительную интенсивность качества, но не «абсолютную» величину ее. Ценность шкал этого типа в том, что они устанавливают порядок, а недостаток в том, что этот порядок не является метрическим.

Приведем несколько примеров. Порядковой является шкала ветров Бофорта. Ее пункты: «штиль», «легкий ветер», «свежий», «крепкий», «шторм», «ураган». Каждый из них имеет качественное определение (эмпирический референт). Эти определения основаны на действиях, производимых ветром. Порядок расположения пунктов шкалы фиксируется числом баллов. Так, «легкий ветер», например, 3 балла, «крепкий» — 7, «шторм» — 10 баллов. Сами эти числа фиксируют не абсолютную интенсивность свойства (силы ветра), а лишь отношения последовательности между пунктами. Их нельзя, например, складывать, но можно сравнивать (больше — меньше).

В минералогии существует эталонная шкала твердости из 10 пунктов, каждому из которых приписывается число — от 1 до 10. Пункты расположены в порядке возрастания твердости (шкалируемый признак). Единица соответствует тальку, 10 — алмазу. На этой шкале любому минералу отводится место с помощью такой процедуры: данный минерал располагается между тем, который он царапает, и тем, который царапает его. Так возникает порядковая шкала.

[15]

Педагогическая система балльных оценок — пример порядковой шкалы: мы не можем сказать, что знания студента, получившего 5, на столько больше знаний студента, получившего 4, на сколько знания последнего больше знаний получившего 3. Нельзя также, например, сказать, что знания получившего 4 вдвое больше знаний получившего 2 (очевидна также размытость позиций этой шкалы), хотя можно в идеале утверждать, что знания получившего 5 больше знаний получившего 4 и т.д. Это же относится ко всем балльным шкалам. Поэтому: *шкалы, построенные с помощью балльных оценок, строго можно рассматривать лишь как порядковые, но не метрические.* Число случаев, когда это предается забвению, достаточно велико. Между тем, практически все современные шкалы в социологии и психологии — номинальные и порядковые.

Интервальные шкалы

В основе построения интервальной шкалы лежит эмпирическая процедура, позволяющая определить равенство *дистанций* между *парами* объектов (разумеется, наряду с определением равенства и порядка объектов). Если эта процедура найдена, числа, приписываемые пунктам шкалы, обладают таким свойством: равенство интервалов чисел отвечает равенству эмпирических интервалов, т.е. интервалов между интенсивностями свойств у рассматриваемых пар объектов. Поэтому свойства чисел, приписанных объектам, не изменяются при линейном преобразовании $X \rightarrow X' = aX + b$. Действительно, если для двух пар объектов *A, B* и *C, D* (так мы условно обозначим эти объекты), $X_B - X_A = X_D - X_C$, то и $X'_B - X'_B = X'_D - X'_C$. Но при этом, если $\frac{X_B}{X_A} = \frac{X_D}{X_C}$, то отсюда не следует, что $\frac{X'_B}{X'_A} = \frac{X'_D}{X'_C}$, т.е. нет равенства отношений.

В преобразовании $X \rightarrow X' = aX + b$ есть два неопределенных параметра — *a* и *b,* и поэтому можно сказать, что в шкале интервалов произвольны начало отсчета *(b)* и единица измерения *(a).*

Интервальными являются, например, все температурные (Цельсия, Реомюра, Фаренгейта) шкалы, кроме абсолютной (Кельвина). Как известно, температура по Фаренгейту связана с температурой по Цельсию соотношением $X' = 32 + 1{,}8X$. Выбирая разные значения *X*, можно легко

[16]

убедиться, что в этой шкале нет равенства отношений. У температурных шкал произволен выбор точки отсчета — нуля (в шкале Цельсия, совершенно условно, это температура замерзания воды, например), произволен и масштаб (цена деления разная у шкал Цельсия, Фаренгейта и Реомюра).

Интервальными являются также календарные шкалы. Даты одного и того же события в разных календарях тоже связаны между собой линейным законом.

Подобные шкалы в социологии редки, ими пользуются для измерения пространственных и временных положений объектов. Зато нередки псевдоинтервальные шкалы (шкала Терстоуна, «термометр» общественного мнения и т.д.), т.е. шкалы, по некоторым признакам напоминающие интервальные, но по сути являющиеся порядковыми.

Шкалы отношений

Базовая эмпирическая процедура построения такой шкалы заключается в установлении равенства отношений между *парами* объектов по изучаемому признаку (разумеется, наряду с отношениями равенства, порядка, равенства интервалов между парами объектов). Числа, приписываемые объектам в этом случае, обладают свойствами равенства отношений, т.е. практически удовлетворяют всем арифметическим аксиомам. Допустимые преобразования чисел теперь суть преобразования подобия: $X \rightarrow X' = aX$ $(a > 0)$, т.е. фиксировано начало отсчета, можно лишь менять масштаб, единицу измерения. Следовательно, приписав определенное число какому-нибудь объекту, тем самым фиксируем числа, приписываемые всем другим аналогичным объектам. Классическим примером такой шкалы являются

абсолютная (кельвиновская) температурная шкала, а также обычная числовая шкала счета. Если a=1, то шкалу называют абсолютной. В качестве примера таковой приводят обычно шкалу счета (если считать единицами, а не десятками, сотнями и т.д.).

В социологии такие шкалы используются для измерения «физических» величин — времени (стаж, возраст), счета (заработная плата, доход, премия), когда «экспериментально» определен нуль — начало отсчета. Пример абсолютной шкалы — социометрический статус члена группы (число полученных им выборов).

В зависимости от типа шкалы применяются те или иные методы статистического анализа, после ознакомления с

[17]

которыми мы вернемся к классификации статистических мер по выделенным уровням социологического измерения. Отметим, что различие интервальных шкал и шкал отношений для социологических исследований практически несущественно, эти два типа шкал часто объединяют в один тип и называют *метрическими* шкалами (метр от греческого $\mu\varepsilon\tau\rho\upsilon$ — мера). Особенностью метрических шкал является наличие единицы измерения и допустимость операции сложения. Возвращаясь к определению количественных и качественных признаков, можно сказать, что количественными называются признаки, измеренные с помощью метрических шкал, а качественными — с помощью шкал более низкого уровня (в частности, номинальных и порядковых). Это определение подчеркивает относительность различий качественных и количественных признаков и связь этих различий с уровнем измерения (можно, например, считать, что до изобретения термометра температура была качественным признаком, так как измерялась с помощью порядковой шкалы: горячий, теплый, комнатный, прохладный, холодный, ледяной).

Конкретные шкалы не всегда легко отнести к тому или иному типу. Например, некоторые авторы считают образование (в годах обучения) количественным признаком. Но при строгом подходе в силу разнокачественности одного года обучения в школе, в техникуме и в вузе, этот признак нужно рассматривать как измеренный в порядковой шкале (это следует иметь в виду при выборе статистических мер). То же самое касается квалификации рабочих, измеряемой разрядами. С другой стороны, эти шкалы так же, как, например, балльные оценки знаний в школе, содержат все же больше информации, чем чисто порядковые: между пунктами шкалы существует некоторое, хотя и приближенное равенство. Ведь преподаватель, выставляющий балл, старается использовать шкалу как метрическую, поэтому, например, изменение системы баллов с 2, 3, 4, 5 на 2, 3, 20, 21 рассматривалось бы как некорректное увеличение расстояния между удовлетворительными и хорошими знаниями. Такие шкалы находятся, следовательно, где-то между метрическими и порядковыми (их иногда называют псевдоинтервальными или псевдометрическими), поэтому при строгом подходе корректно применение лишь статистики для порядковых шкал, но в некоторых случаях возможно (при известной осторожности) использование статистики для метрических шкал.

[18]

2. Табулирование. Вариационные ряды. Графики. Приемы наглядного представления социологических данных

Предположим, что мы опросили некоторое множество респондентов с помощью следующей анкеты[14]:

[14] Предлагаемая анкета носит иллюстративный характер, по существу, это фрагмент реальной социологической анкеты, содержащей обычно десятки (или даже сотни) вопросов, в том числе: контактных, функционально-психологических, контрольных и т.п. (См., например, *Ядов В. А.* Социологическое исследование. М., 1972; *Ноэль Э.* Массовые опросы. М., 1978; и др.).

Социологическая анкета

1. Укажите, пожалуйста, Ваш пол:

мужской 1

женский 2

2. Удовлетворены ли Вы своей профессией?

полностью удовлетворен 1

скорее удовлетворен, чем нет 2

затрудняюсь ответить 3

скорее не удовлетворен, чем удовлетворен 4

неудовлетворен 5

3. Укажите, пожалуйста, доход на одного члена Вашей семьи – руб.

Здесь представлены три типа признаков (1-й вопрос порождает номинальную шкалу, 2-й — порядковую и 3-й — метрическую), поэтому на этом примере можно рассмотреть основные специфические для социологии методы представления данных[15]. Прежде всего, сведем информацию к обозримому виду, перенеся данные из анкет в специальную таблицу 1.

Такого рода таблицы называются *матрицами данных.* Дальнейшие преобразования информации направлены на то, чтобы сделать ее более наглядной, представить в более компактной форме. С этой целью подсчитывают, сколько индивидов обладают данным значением признака. Значение признака называют *вариантом,* а число лиц, обладающих данным значением,— его *частотой.* Варианты вместе с частотами образуют *вариационный ряд* данного признака, или *распределение* по данному признаку (в табл. 2 представлены вариационные ряды признаков «пол» и «удовлетворенность профессией», или распределение опрошенных по признакам «пол» и «удовлетворенность профессией»).

[19]

Таблица 1

Условный пример: данные опроса 284 респондентов

Номера индивида (анкеты)	Признак		
	Пол	Удовлетворенность профессией	Доход
1	1	2	80,0
2	2	1	75,3
3	2	5	65,4
.	.	.	…
283	2	4	82,3
284	1	2	95,0

Таблица 2

Распределение опрошенных по признакам «пол» и «удовлетворенность профессией», частости и проценты

Показатель	Признак							Всего
	Пол		Удовлетворенность профессией					
	Номер варианта ответа							
	1	2	1	2	3	4	5	
Частота	104	180	81	83	19	61	40	284

[15] Исключение представляют данные, порождаемые социометрическими вопросами, — методы их анализа рассмотрены в гл. VI.

Относительные частоты, или доли, частости	0,37	0,63	0,28	0,29	0,07	0,21	0,14	1
Процент	37	63	28	29	7	21	14	100

Наряду с вариационными рядами в табл. 2 содержатся также частости и проценты. *Частостями* называют частоты, разделенные на сумму частот по данному признаку, другое название — *относительные частоты,* или *доли частот* (в данном примере сумма частот для признаков «пол» и «удовлетворенность профессией» равна 284), *проценты* представляют собой умноженные на сто частости (доли).

Представить компактно данные, полученные по метрическим шкалам, таким способом, как правило, не удается из-за большого количества вариантов, поэтому для построения распределения диапазон изменения признака разбивают на интервалы и подсчитывают, сколько индивидов имеют значение признака, лежащее в границах каждого интервала (табл. 3).

[20]

Из таблицы ясно, что 2 индивида имеют доход до 65 руб., 32 — от 65 до 74 руб. и т.д. Отметим, однако, что использованные нами значения округлены до целых, т.е. значение 74,3 руб., например, отнесено к интервалу 65—74, а значение 74,6 к интервалу 75—84. Условимся цифру 5 округлять до высшего разряда, т.е. 74,5 до 75 и, следовательно, относить 74,5 к интервалу 75—84. В некоторых работах используются интервалы с совпадающими границами, т.е. в данном примере это были бы границы: до 65, 65—75,

Таблица 3

Распределение опрошенных по признаку «доход»

Показатель	Номер интервала						Всего
	1	2	3	4	5	6	
	Граница интервала, руб.						
	до 65	65-74	75-84	85-94	95-104	105 и выше	
Частота	2	32	50	181	11	8	284

75—85, 85—95, 95—105, 105 и выше. В этом случае 74,5; 74,8; 74,9, например, относятся к интервалу 65 — 75, а значения 75,0; 75,1 и т.д.— к интервалу 75 — 85. (Это важное замечание будет учтено при выводе формул для вычисления медианы и квантилей).

Для описания вариационных рядов введем следующие обозначения. Значения признака *X y* отдельных индивидов, т.е. варианты, обозначим через x_i , i= 1, 2, ..., N, где N — общее число индивидов, или объем совокупности. (Для краткости в дальнейшем мы будем писать $i = \overline{1, N}$). Некоторые варианты могут повториться: например, на предприятии имеется ряд работников с образованием 10 классов и т.д. Пусть различных вариантов k ($k<N$), а обозначение x_i при $i=\overline{1,k}$ соответствует теперь различным вариантам. Общее число индивидов с $X = x_i$ мы будем обозначать $N(X_i)$ или просто N_i (пока рассматривается один признак это возможно). Ясно, что $\sum_{i=1}^{k} N(x_i) = \sum_{i=1}^{k} N_i = N$ .. Величина N_i является частотой, а $v_i=N_i/N$ – частота варианта x_i. Очевидно, что $\sum_{i=1}^{k} v_i = 1$.

[21]

Варианты вместе с частотами образуют вариационный ряд (одномерное распределение признака), который может быть дискретным (в случае номинальных и порядковых признаков, а также для некоторых метрических, например, «число детей в семье», «разряд» для рабочих и т.п.) или непрерывным (для метрических признаков). В случае, если варианты расположены в порядке убывания или возрастания, вариационный ряд называется упорядоченным (ранжированным). Как правило, непрерывные признаки указанным способом преобразуют в дискретные путем введения интервалов. Величина интервала называется интервальной разностью.

Если обозначить левую границу некоторого *l*-го интервала через x'_l, а правую — через x''_l, то ширина интервала, или интервальная разность, равна $I_l = x''_l - x'_l$. Эта формула верна лишь в случае, если границы соседних интервалов совпадают, т.е. $x''_l = x'_{l+1}$. Когда границы интервалов не совпадают (как в табл. 3), то $I_l = x''_l - x'_l + 1$. Например, ширина 3-го интервала равна не 84—75=9, а 10, так как в интервал попадают, как указывалось выше, значения от 74,5 до 84,5 (84,5—74,5=84—75+1=10). Величину $\frac{1}{2}(x''_l + x'_l)$ (для интервалов с совпадающими границами) или $\frac{1}{2}(x''_l + x'_l + 1)$ (для интервалов с несовпадающими границами) назовем серединой или центром интервала. Для нашего примера середина интервала равна $\frac{1}{2}(84+75+1)=80$.

Основным приемом представления и анализа социологических данных является построение одномерных (вариационные ряды) и двумерных распределений признаков (реже 3-мерных и *n*-мерных распределений) или, другими словами, распределений опрошенных по одному, двум, трем и более признакам.

Одномерное распределение

А. **Классификационные и качественные признаки** (номинальные и порядковые шкалы). Допустим, что нам известно одномерное распределение *N* респондентов по некоторому признаку *X,* имеющему *k* градаций (вариантов):

Вариант	x_1	x_2	…	x_k
Частота	N_1	N_2	…	N_k

[22]

Чаще всего одномерное распределение изображается с помощью полигонов и гистограмм распределения. На оси абсцисс откладываются *k* точек, на оси ординат — значения N_i; соединив их ломаной линией, получим *полигон* распределения, если же построить столбики высотой N_i — получим *гистограмму.* Полигоны и гистограммы можно строить не только с использованием частот, но и частостей и процентов.

Таблица 4

Распределение населения СССР по уровню образования в 1979 г. (см. Население СССР. М., 1980 г.)

Уровень образования	Абсолютные показатели, тыс. чел.	Процент
Неполное среднее	52488	37,7
Среднее общее	45099	32,4
Среднее специальное	23439	16,9
Высшее незаконченное	3235	2,3

Высшее оконченное	14826	10,7
Всего	139089	100

Рассмотрим на примере, как строятся указанные виды графиков.

Пример 1. Построим полигон и гистограмму распределения для данных, приведенных в табл. 4.

На рис. 1, отражающем эти данные, изображены две оси ординат — на одной из них отложены абсолютные величины, на другой — проценты. Форма графиков не зависит от вида показателя (частоты, частости или проценты), откладываемого на оси ординат. Полигон (по соглашению) изображают как замкнутую кривую.

Б. **Количественные признаки** (интервальные шкалы и шкалы отношений). Принципиальных различий в построении одномерных распределений количественных признаков по сравнению с изображением качественных признаков нет, но есть некоторые особенности, связанные с тем, что для количественных признаков приобретает смысл понятие ширины интервала. Прежде чем перейти к обсуждению этого вопроса, введем некоторые определения.

Частоту, приходящуюся на единицу интервала (для *l*-го

интервала $\rho_l = \frac{N_l}{I_l}$, назовем *плотностью распределении,* а частость, приходящуюся на единицу интервала — *относительной плотностью распределения.* Особо важную роль играет это понятие в случае неравных интервалов, на чем мы в дальнейшем специально остановимся.

Нам также понадобится понятие *накопленной,* или *кумулятивной, частоты* (частости). Накопленная частота по-

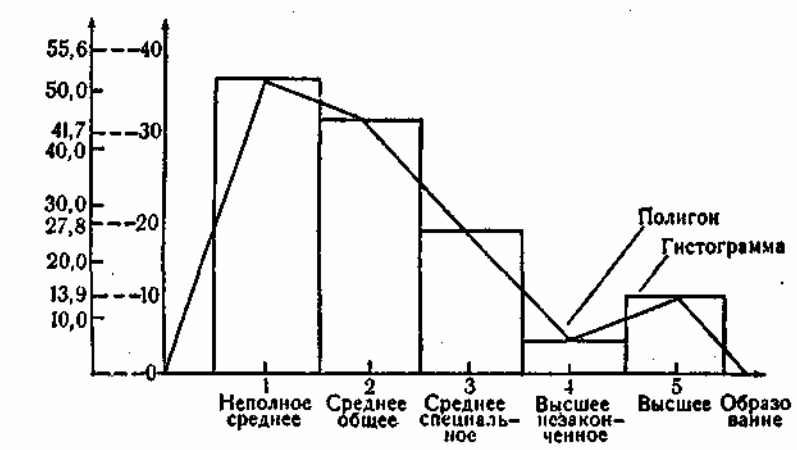

Рис. 1. Уровень образования населения СССР (1979 г.)

казывает число индивидов, у которых варианты не больше (меньше либо равны) данного значения признака.

Скажем, для *l*-ого интервала накопленная частота $F_l = \sum_{i=1}^{l} N_i$ — показывает, у какого числа индивидов $X \le x_l$ или, другими словами: сколько всего индивидов с $X=x_1, X=x_2, ..., X=x_l$. Очевидно $F_k=N$. Кумулятивная частость $f_l = \sum_{i=1}^{l} v_i (l \le k)$ и соответственно $f_k=1$. Тогда ρ_l в процентах равна $\frac{v_l}{I_l}$. 100%.

В конкретных исследованиях нередко используются неравные интервалы. Так как велик диапазон возможных значений, например, возраста работников (свыше 50 лет), то при равных интервалах в случае разумного числа пунктов (10—12) будет слишком большой интервальная разность (около 5 лет), это не позволит достаточно точно изучить по-

[24]

ведение работников разного возраста, особенно молодых (в старших возрастных группах, как показывают исследования, влияние возрастных различий на поведение несколько ниже) Увеличение же дробности, желательное для детального изучения, приводит к очень большому числу пунктов (25—30), существенно затрудняющему анализ материала. Выходом из этого положения является компромиссный вариант: малые интервалы выбираются для групп молодых

Таблица 5

Распределение по возрасту работников Одесского судоремонтного завода им. 50-летия Советской Украины (1971 г.)

Граница интервала, лет	Середина интервала, x_i	v_i, %	f_l, %	I_l	ρ_l, %
16–17	16,5	2,4	2,4	2	1,20
18–19	18,5	5,8	8,2	2	2,90
20–21	20,5	5,1	13,3	2	2,55
22–24	23,0	10,9	24,2	3	3,63
25–30	27,5	15,3	39,5	6	2,55
31–40	35,5	30,2	69,7	10	3,02
41–50	45,5	18,3	88,0	10	1,83
51–60	55,5	8,5	96,5	10	0,85
Свыше 60	65,5	3,5	100,0	10	0,35

работников, а большие — для работников старших возрастных групп.

В настоящее время в социологической литературе обсуждается проблема стандартизации основных измерительных процедур. Дел в том, что данные, получаемые разными исследователями, зачастую несопоставимы (или крайне ограниченно сопоставимы). В значительной мере это результат отсутствия соглашений между исследователями по поводу измерения различных признаков. Практически получается, что число разных градаций одного и того же признака не намного меньше числа исследователей. Осознавая эти трудности, экспертная служба ИСИ АН СССР провела опросы социологов страны, в частности по проблеме «Возраст в конкретных исследованиях». Анализ результатов позволяет дать некоторые рациональные рекомендации для социологов-практиков[16].

[25]

Обратимся к примеру, иллюстрирующему данные выше определения.

Пример 2. В таблице 5 приведено распределение по возрасту работников Одесского судоремонтного завода им. 50-летия Советской Украины (1971 г.). Как видим, при построении распределения использовались неравные интервалы. Рассмотрим, например, интервал 20—21, сюда мы относим индивидов, возраст которых от 19,5 до 21,5, т.е. ширина

[16] *Петренко Е.С., Ярошенко Т.М.* Социально-демографические показатели в социологических исследованиях. М., 1979, с. 40—49.

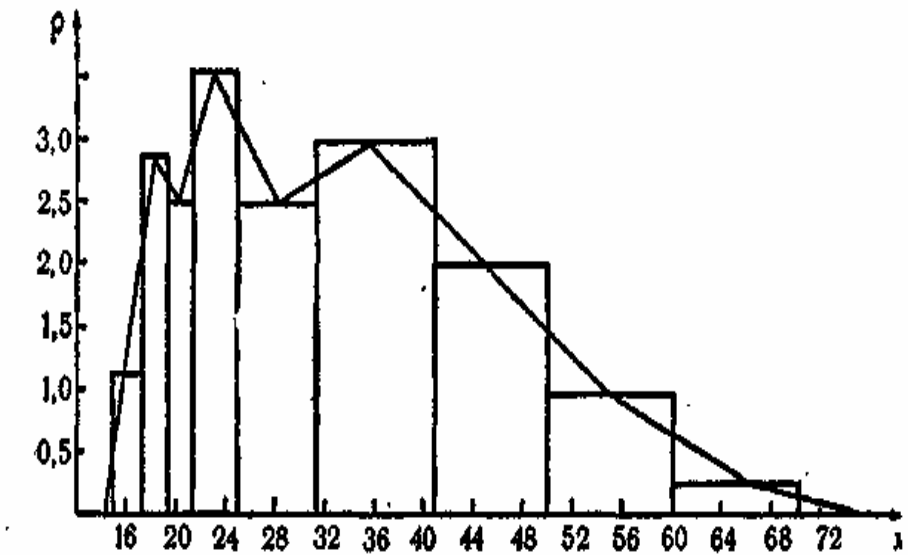

Рис. 2. Полигон и гистограмма распределения при неравных интервалах

интервала 2 года, в интервал 25—30 попадают индивиды, возраст которых от 24,5 до 30,5, т.е. ширина его 6 лет.

Если правая граница предыдущего интервала совпадает с левой последующего (например, в случае интервалов 16— 18, 18—20, 20—22 и т.д.), то следует указать, к какому из них относить граничное значение (в данной книге мы относим его к верхнему интервалу). Отметим, что возникающие трудности, если такое указание не сделано, зачастую преувеличиваются: вероятность того, что мы опрашиваем индивида в день его рождения порядка тысячных долей ($^1/_{365}$)[17].

Из-за наличия неравных интервалов, для построения полигона распределения данных, приведенных в таблице 5, по оси ординат откладывают уже не N_i (или v_i), а плотности ρ_i. Аналогично строится и гистограмма (рис. 2). Отметим, что площадь каждого прямоугольника равна $I_l\rho_l=N_l$, а сумма площадей всех прямоугольников равна N.

Плотность изображается на гистограмме так, как если бы

[26]

она была постоянной внутри интервала. Обычно этого нет, ρ_l — это средняя плотность на интервале. Ясно, что чем меньше интервал, тем ближе полигон к фактическому изменению плотности распределения в зависимости от изменения признака. Для непрерывных признаков в пределе, когда $I_l\to 0$, мы получили бы плавную кривую изменения плотности распределения, которую называют теоретической кривой распределения. Очевидно, площадь, ограниченная кривой распределения, равна 1, если на оси ординат откладывать частости. В дальнейшем мы подробнее остановимся на кривых распределения.

Еще один графический способ изображения вариационного ряда — кумулятивная кривая (ее называют также кумулятой, или кривой накопленных частот). Кумулята строится аналогично полигону, но координаты точек теперь (x_l, F_l) либо (x_l, f_l) т.е. абсциссы те же, а ординаты — накопленные, или кумулятивные, частоты. Ясно, что кумулята — неубывающая кривая.

Упражнение 1. Построить кумуляту по данным табл. № 5.

Кривая, построенная по точкам с координатами (F_l, x_l), называется огивой Гальтона[18].

Упражнение 2. Для нашего примера построить огиву.

---

[17] О понятии вероятности см. Приложение 1.

[18] Кумулята и огива позволяют быстро определить долю лиц, обладающих более высоким (или низким) значением, чем любое фиксированное значение признака. Например, медиана является ординатой такой точки огивы, абсцисса которой равна 0,5.

Форма статистического распределения (вариационного ряда) — вид его графика. Например, полигона. Проанализируем полигон рис. 2. Вначале с увеличением возраста увеличивается плотность распределения. Затем — провал, он связан с уходом молодежи в армию (на обследуемом предприятии работают в основном мужчины). Затем плотность снова возрастает: на предприятие приходят отслужившие. Второй провал связан с историческими условиями жизни страны — эхо войны, следствие низкой рождаемости и выживаемости детей в военные годы (это станет ясно, если сопоставить соответствующие *x c* годом опроса, с течением времени этот провал, естественно, сдвигается вправо. Затем плотность распределения монотонно убывает с увеличением возраста, что естественно.

Полигон — ломаная кривая. Вид полигона зависит от числа различных вариантов. Предел, к которому стремится полигон при увеличении числа вариантов, плавная кривая, которая может быть описана с помощью некоторого аналити-

[27]

ческого выражения: *y=y(x).* Разные распределения описываются с помощью различных функций.

Познакомимся с некоторыми часто встречающимися формами распределений.

Распределение может описываться монотонной — убывающей или возрастающей — функцией типа изображенных на рис. 3 (а и б соответственно).

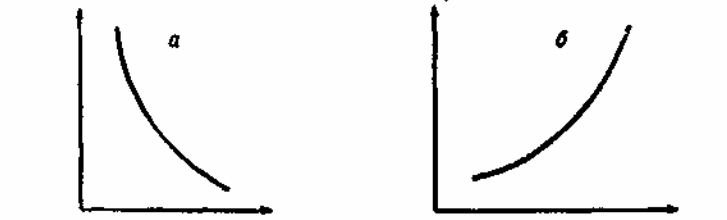

Рис. 3. Монотонно убывающая (а) и монетонно возрастающая (б) функции

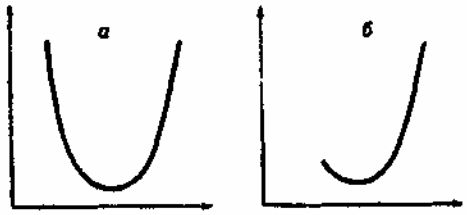

Рис. 4. *U*-образное (а) и *J*-образное (б) распределение

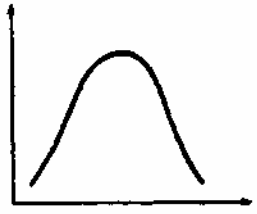

Рис. 5. Колоколообразное распределение

Примером здесь может служить распределение работников по стажу работы на данном предприятии: чем боль-ше стаж, тем меньше работников (это связано с трудовыми перемещениями, с текучестью: уходом «старых» и приходом новых работников).

Распределение может быть *U*-образным (частный случай —*J*-образным, см. рис. 4а и 46 соответственно): например, распределение по удовлетворенности трудовой деятельностью (как правило, часто оказывается меньше всего работников, занимающих на шкале удовлетворенности промежуточную позицию).

Своего рода обратным *U*-образному является так называемое колоколообразное распределение (рис. 5), встречающееся довольно часто в конкретных исследованиях: например, распределение людей по росту, весу, по заработной плате («крайности» встречаются редко). Если частоты вариантов, симметричных относительно центрального, при-

[28]

мерно одинаковы, то распределение называется симметричным, в противном случае — асимметричным. На рис. 6 (а—г) показаны примеры асимметричных распределений:

Одновершинные распределения называются унимодальными, двувершинные — бимодальными и т.д. Многовершинные распределения встречаются реже одновершинных.

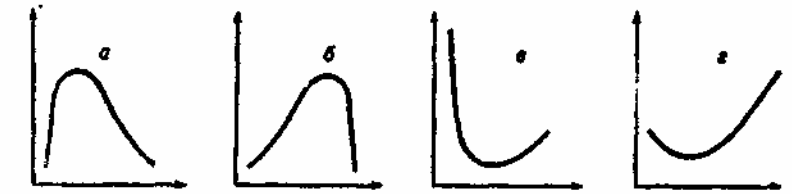

Рис. 6. Асимметричные распределения

Часто встречаются колоколообразные распределения, хотя и не всегда в «чистом» виде: эмпирическое распределение может быть близким к колоколообразному. Особо важную роль в статистике играет распределение, получившее название *нормального* (§ 3 этой главы).

Двухмерные распределения (комбинационные таблицы)

Рассмотрим следующую таблицу, представляющую собой двухмерное распределение[19] по признакам «тип рабочего места» и «удовлетворенность зарплатой» данных выборочного почтового опроса жителей Киева (табл. 6). Такого рода таблицы иногда называют комбинационными, так как в них отражена информация о комбинации двух (в данном случае) или большего числа признаков.

На пересечении *i*-й строки и *j*-го столбца этой таблицы стоит число респондентов, имеющих *i*-е значение первого и одновременно *j*-е значение второго признака, а также процент, который составляет это число от суммы элементов строки. Фактически, таблица представляет собой 4 вариационных ряда (если не считать итогового распределения, которое приведено для удобства пользования таблицей). Поэтому данные этой таблицы можно изобразить на одном графике в виде 4-х полигонов, используя для каждого свой цвет или вид линии (сплошная, пунктирная и т.п.).

[29]

Упражнение 3. Начертить график, представляющий данные двухмерного распределения признаков, приведенные в таблице 6.

Таблица 6

Двухмерное распределение данных почтового опроса жителей г. Киева, абсолютная величина и процент

Признак «тип рабочего места»			Признак «удовлетворенность работой»					Всего
Труд	Квалификация	Номер варианта	Удовлетворен	Скорее да, чем нет	Трудно ответить	Скорее нет, чем да	Неудовлетворен	
			Номер варианта					
			1	2	3	4	5	
Физический	Низкая	1	86 32.3	23 8,6	27 10,2	31 11,6	99 37,2	266 100
	Средняя и высокая	2	393 41,2	110 11,5	117 12,3	114 11,9	221 23,1	955 100

[19] Формализованное описание двухмерных распределений мы приведем ниже при рассмотрении корреляционной таблицы (гл. II, § 1, табл. 15).

Умственный	Не требующая высшего образования	3	182 23,9	64 8,4	89 11,7	121 15,9	306 40,2	762 100
	Требующая высшего образования	4	245 25,8	123 12,9	106 11,1	168 17,7	309 32,5	951 100
	Всего		906 30,9	320 10,9	339 11,6	434 14,8	935 31,9	2934 100

Если один из признаков двухмерного распределения количественный, мы имеем возможность для каждого значения качественного признака рассчитать средние арифметические[20] и таким образом «сжать» информацию, как бы свести ее к одномерному распределению (например, если второй признак не «удовлетворенность», а «доход», то можно было бы рассчитать средний доход для каждого из четырех типов

[30]

Таблица 7

Удовлетворенность респондентов различными сторонами своей работы

Удовлетворенность	Рабочее место				Все группы	Ранг	Среднее квадратическое отклонение	Ранг
	Физического труда		Умственного труда					
	Низкой квалификации	Высокой и средней квалификации	Не требующего высшего образования	Требующего высшего образования				
1	2	3	4	5	6	7	8	9
1. Содержанием труда	0,46	0,61	0,57	0,64	0,60	3	0,079	3
2. Режимом труда	0,54	0,46	0,49	0,55	0,50	4	0,042	6
3. Размером оплаты	-0,06	0,18	-0,20	-0,09	-0,03	7	0,160	1
4. Возможностями повышения квалификации	0,26	0,35	0,22	0,20	0,26	6	0,066	5
5. Отношениями с коллегами	0,92	0,90	0,87	0,84	0,87	1	0,035	7
6. Отношениями с руководителями	0,81	0,73	0,74	0,64	0,72	2	0,070	4
7. Удаленностью работы от места жительства	0,60	0,45	0,54	0,35	0,46	5	0,109	2
8. Возможностями улучшения жилищных условий	-0,17	-0,21	-0,21	-0,14	-0,18	8	0,034	8

[20] Средние арифметические рассматриваются в следующем параграфе.

рабочих мест). На графике в этом случае будет лишь один полигон распределения: на оси абсцисс — качественный признак, по оси ординат откладываются средние значения количественного признака.

Часто так поступают не только для количественных, но и для качественных признаков, измеренных с помощью порядковых шкал: пунктам шкалы приписываются определен-

[31]

ные баллы и находится средний балл[21], или индекс (подробнее этот вопрос будет рассмотрен в § 3). Так, приписав удовлетворенным балл l, тем, кто скорее удовлетворен, чем нет — 0,5, затрудняющимся ответить — 0, тем, кто скорее неудовлетворен, чем удовлетворен — (—0,5) и, наконец, неудовлетворенным — балл (—1), получим для каждого типа рабочих мест следующие средние баллы (индексы) удовлетворенности:

Тип рабочего места	1	2	3	4
Индекс удовлетворенности зарплатой	-0,06	0 ,18	-0,20	-0,09

Таким образом, данные «сжались» до одной строки и могут быть изображены в виде одного полигона (по оси абсцисс — типы рабочих мест, по оси ординат — индексы удовлетворенности). На одном графике можно изобразить данные целого ряда таблиц двухмерных распределений. Так, в проведенном нами опросе работающего населения г. Киева была получена информация об удовлетворенности респондентов различными сторонами работы, или элементами рабочей ситуации (содержанием и режимом труда, зарплатой и т.п.). Индексы удовлетворенности по восьми двухмерным распределениям респондентов для признаков «тип рабочего места» и «удовлетворенность элементом рабочей ситуации» (одно из них было приведено в табл. 6), сведены в таблицу 7. Но прежде, чем перейти к построению графика, сформулируем некоторые общие принципы изображения нескольких полигонов на одном рисунке. Целесообразно рассмотреть отдельно два случая: а) изображаются два полигона; б) три и более.

В первом случае исследователь ставит перед собой цель наглядно представить различия между двумя группами респондентов (или какими-либо двумя другими объектами). При этом на оси абсцисс откладываются значения признака, а полигоны представляют собой распределения каждой из групп по этому признаку или значения некоторого показателя. Значения признаков на оси абсцисс целесообразно откладывать упорядоченными по убыванию разности ординат полигонов. Поясним сказанное примером. В исследова-

[32]

нии межличностных оценок, проведенном В. Шубкиным, Ю. Карповым и Г. Кочетовым[22], каждому из индивидов предлагалось оценить всех членов своего коллектива (в том числе и себя) по семи группам качеств:

I – интеллектуальные качества (одаренность, глубина знаний и т.п.);

II – деловые качества (умение привлечь людей и т.п.);

III – импульсно-волевые свойства (сдержанность, эмоциональность и т.п.);

IV – моральные качества (доброта, скромность и т.п.);

V – качества, характеризующие мотивы поведения (альтруизм, стремление к истине и т.д.);

VI – качества, характеризующие отношения к жизни (оптимизм, юмор и т.п.);

VII – качества, характеризующие физическую привлекательность

---

[21] Как будет показано в гл. III, при строгом подходе эта операция не совсем корректна, так как опирается на некоторые непроверенные предположения. Тем не менее практика применения индексов в социо-логии показывает, что для приближенных оценок их использование час-то правомерно.

[22] *Шубкин В. Н.* Социологические опыты. М., 1970, с. 110—151. [м] Там же, с. 127.

По каждой из групп качеств были найдены коллективная оценка (т.е. оценка данного человека другими) и самооценка (т.е. средняя самооценка членов коллектива). Полученные данные представлены на рисунке, заимствованном из книги В. Шубкина[23] (рис. 7, а). Он дает определенное представление о полученных в результате исследования данных (видно, например, что самооценка выше всего по моральным качествам и качествам, характеризующим отношение к жизни, т.е. п. IV. и VI, что различия оценки и самооценки выше по п. IV, чем по п. III и V и т.д.). Но многие различия оценок и самооценок на графике «не читаются». Например, неясно, по каким пунктам больше различия — по I или по VII, по IV или по VI и т.п.

Чтобы сделать график наглядней и информативней, мы вычли для каждого пункта из коллективной оценки самооценку и расположили качества личности на оси абсцисс по убыванию этой разности (см. рис. 7, б). Интерпретация такого графика существенно облегчается: слева расположены качества личности, для которых оценка других выше, чем самооценка (это интеллектуальные качества и качества, характеризующие физическое совершенство, т.е. п. I и VII, причем по первому из них различия больше), а справа те, по которым индивид оценивает себя выше, чем коллектив (п. IV и VI, а также п. V, III и II, по которым различия приблизительно равны между собой и существенно ниже, чем различия по п. IV и V). Отметим, что при таком способе

[33]

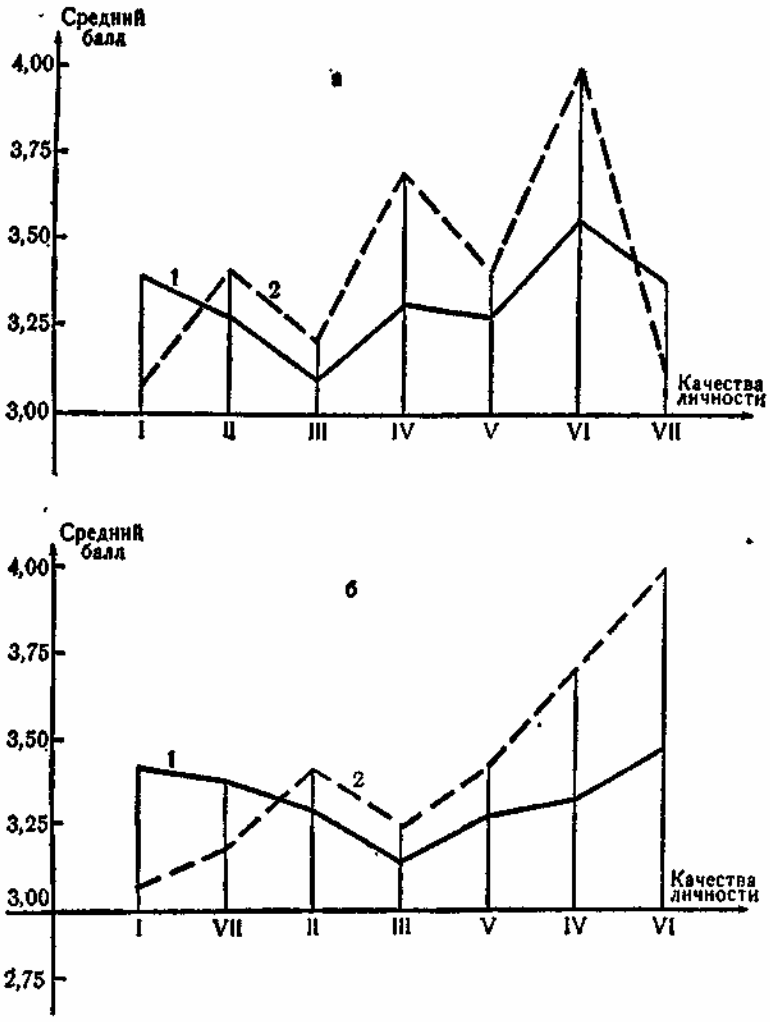

Рис. 7. Коллективная оценка и самооценка качеств личности: график «а» построен неудачно, график «б» — удачно.
Обозначения: *1* — коллективная оценка, *2* — самооценка

[23] Там же, с.127.

построения даже самые запутанные графики с большим количеством пересечений приобретают достаточно простой вид: они содержат *не более одного пересечения,* причем до пересечения один показатель выше другого, а после пересечения наоборот.

[34]

Рассмотрим второй случай — изображение трех и более полигонов на одном графике. Теперь повышение информативности в зависимости от целей анализа осуществляется двумя путями. Первый из них — когда нас интересует преж-

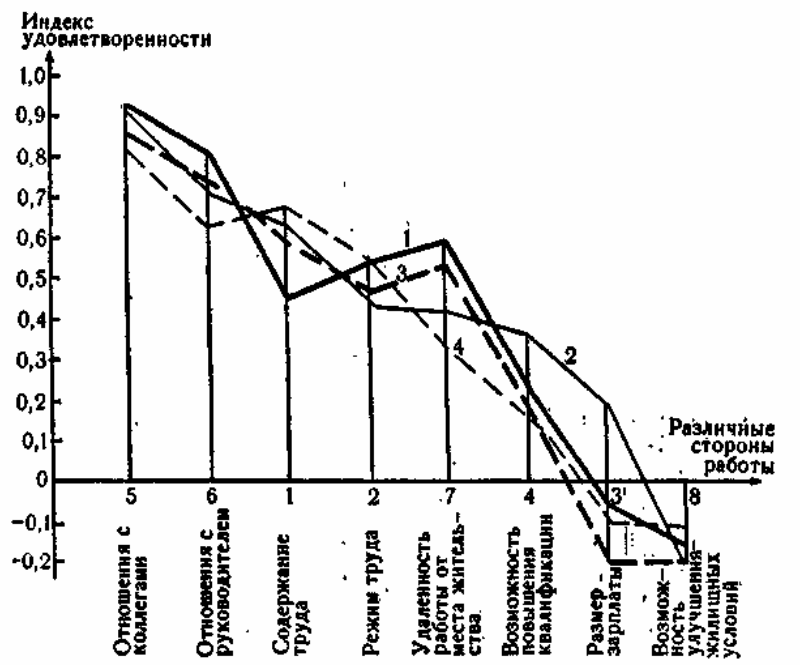

Рис. 8. Удовлетворенность респондентов различными сторонами своей работы (работающее население г. Киева, 1979.). Пункты оси абсцисс расположены по убыванию степени удовлетворенности всех опрошенных.

Обозначения: *1* — лица физического труда низкой квалификации, *2* — лица физического труда средней и высокой квалификации, *3* — лица умственного труда, не требующего высшего образования, *4* — лица умственного труда, требующего высшего образования.

де всего значения изучаемых показателей, а затем уже различия показателей у разных групп респондентов — заключается в расположении пунктов на оси абсцисс по убыванию некоторого усредненного значения изучаемого показателя. Рассмотрим это на примере изображения данных таблицы 7. Предположим, что в первую очередь нас интересует степень удовлетворенности респондентов различными сторонами своей работы, а потом уже различия в удовлетворенности разных групп респондентов. В 6-й колонке таблицы приведены данные об удовлетворенности, рассчитанные для всех

3*

[35]

групп в целом, т.е. для всего массива опрошенных, не расчлененного на группы по характеру труда и уровню квалификации. На рис. 8 на оси абсцисс различные стороны работы представлены в порядке убывания индексов удовлетворенности для всего массива (т.е. в соответствии с рангами,

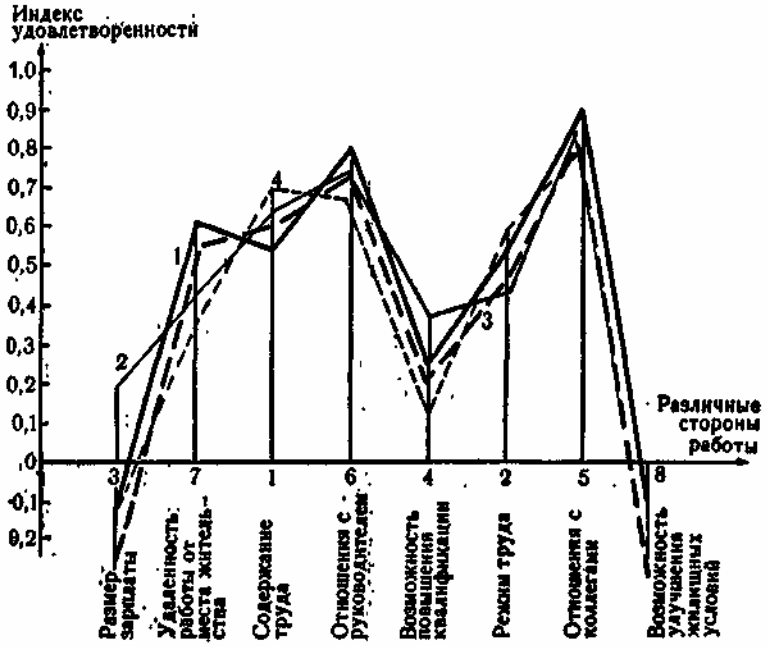

Рис. 9. Удовлетворенность респондентов различными сторонами своей работы (работающее население г. Киева, 1979 г.). Пункты оси абсцисс расположены по убыванию различий между изучаемыми группами.

Обозначения: *1* — лица физического труда низкой квалификации, *2* — лица физического труда средней и высокой квалификации, *3* — лица умственного труда, не требующего высшего образования, *4* — лица умственного труда, требующего высшего образования

приведенными в колонке 7). При этом способе изображения мы имеем возможность при интерпретации обращать внимание прежде всего на наиболее важные, «проблемные» моменты» изучаемых явлений (в данном случае на стороны работы с наиболее низкими показателями удовлетворенности). В тех же случаях, когда нас интересуют прежде всего различия между группами (например, разработка социальных или экономических мероприятий, направленных на уменьшение различий между некоторыми группами респондентов), пункты располагаются по убыванию различий

[36]

между изучаемыми группами. На рис. 9 данные таблицы 7 изображены таким способом. В качестве показателя различий между группами принято среднее квадратическое отклонение[24] (см. колонку 8). На оси абсцисс стороны работы упорядочены по убыванию этого показателя (т.е. в соответствии с рангами колонки 9). В этом случае интерпретация представленных данных проходит иначе, чем в предыдущем. Из рисунка видно, что различие групп по удовлетворенности возможностями улучшения жилищных условий минимально, затем идет удовлетворенность отношениями с коллегами и т.д. Если на рис. 8 было удобно интерпретировать среднюю удовлетворенность и отклонения от нее, то с помощью (рис. 9) удобно интерпретировать различия между группами.

Кроме полигонов и гистограмм, существуют и другие виды графиков, которые используются, однако, значительно реже. В книге Дж.Гласса и Дж.Стенли[25] приводится пример 15-ти различных способов изображения одних и тех же данных. Там же предложены некоторые общие рекомендации для построения графиков[26]. Вместе с тем отметим, что процесс построения графиков плохо формализуется и требует творческого подхода и критического восприятия общих рекомендаций. Нам, в частности, кажется

[24] Для этой цели можно использовать также другие меры вариации (см. § 4 этой главы) и коэффициенты корреляции (например, коэффициент Чупрова между признаком «тип рабочего места» и признаками, характеризующими удовлетворенность сторонами работы).

[25] *Гласс Дж., Стэнли Дж.* Статистические методы в педагогике и психологии. М., 1976, с. 42—43.

[26] Там же, с. 54.

нецелесообразным замыкание полигонов распределения, ухудшающее «чтение» графиков. С другой стороны, неправомерным представляется достаточно распространенное мнение, что на одном графике не следует размещать более трех полигонов, так как целый ряд линий на графике сливается[27]. Сформулированные нами приемы построения графиков вытекают из противоположных соображений. Совпадение полигонов повышает наглядность, облегчает описание сходства и различия: чем больше сливающихся, тем меньше отличающихся точек и тем легче чтение графика (например, из рис. 8 видно, что три группы работников примерно одинаково удовлетворены содержанием труда, у четвертой — работники физического труда низкой квалификации — удовлетворенность этим элементом

[37]

рабочей ситуации значительно ниже; видно также и то, что все группы примерно одинаково оценивают отношения с коллегами, возможность улучшения жилищных условий и т.д.). Думается, что на одном графике вполне можно изображать до 7—8 полигонов распределения.

Завершая изложение способов представления данных, отметим, что построение графиков не только важная часть исследовательской работы, необходимая для повышения наглядности результатов и передачи другим известной автору информации, но и инструмент анализа: продуманный подход к построению графика, стремление сделать его информативным и наглядным позволяют лучше понять структуру полученных данных, глубже проникнуть в сущность изучаемого явления.

3. Меры центральной тенденции

Как мы видели, вариационный ряд может быть описан с помощью набора величин x_i, N_i ($i=\overline{1,k}$). Однако оперирование с полным набором затруднительно. Для удобства изучения необходимо ввести величину, которая, учитывая особенности данного ряда, была бы сводной, итоговой. Такую величину называют *средней.* Средняя не может полностью заменить ряд. Оперируя с нею, мы теряем часть информации, но отражаем типичное для данной совокупности в данных условиях. Средняя характеризует уровень ряда, его центральную тенденцию.

Чтобы средняя величина была действительно обобщающей характеристикой, улавливающей закономерность, она должна применяться к достаточно однородной совокупности. Выведение средних для неоднородной совокупности может привести к бессмысленному результату, например, метко спародированному Г.Успенским усреднению, когда «миллионщик Колотушкин» и «просвирня Кукушкин», имеющий грош, владеют «в среднем по полмиллиону». Такие средние огульны, фиктивны. (Заметим, что в некоторых случаях даже огульная средняя может быть показательной. Например, памятные «четверть лошади» — столько в «среднем» приходилось в царской России на одну ревизскую душу).

Стал классическим пример разоблачения Лениным статистиков народнического толка, выводивших средние для всего крестьянства, не желая видеть, что оно неоднородно, что часть его принадлежит к сельской буржуазии, часть — к

[38]

батракам. Очевидно, «средние», характеризующие крестьянство «в целом», не могли быть научными.

Итак, вычислению средних должно предшествовать обоснованное выделение в изучаемой совокупности достаточно однородных групп.

---

[27] Там же, с. 60.

Говоря о средней, чаще всего имеют в виду среднюю арифметическую $M = \frac{1}{N}\sum_{i=1}^{N}, x_i = \frac{1}{N}\sum_{i=1}^{k} N_i x_i$ (суммируем до N

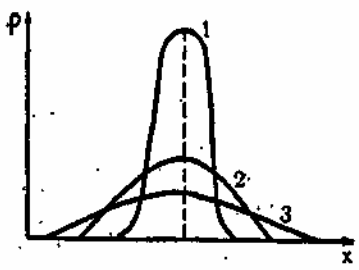

Рис. 10. Распределение с одинаковыми средними

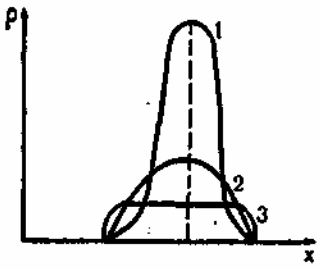

Рис. 11. Распределения с одинаковыми средними и вариационным размахом

или до k). Если все варианты совпадают, то x_i=M, колеблемости (варьирования) нет. Обычно, конечно, $x_i \neq M$. Как же охарактеризовать колеблемость? Простейшей мерой может служить так называемый *вариационный размах* $R=x_{max}-x_{min}$. Для изображенных на рис. 10 распределений такой показатель достаточно эффективен. Все три распределения имеют одинаковые средние $M_1=M_2=M_3$. Ясно, что минимальная колеблемость у распределения 1, максимальная у 3. Как видно из графика, $R_1<R_2<R_3$. Вариационный размах определяется, однако, лишь крайними значениями признака и не отражает колеблемости остальных вариантов. Три распределения, представленных на рис.11, имеют одинаковые R (и M), но явно разные колеблемости. Кроме того, встречаются ситуации, когда вариационный размах в принципе не может быть достаточно достоверно определен (например, доход семей в капиталистических странах — см. пример № 3 этого параграфа). Что же можно еще использовать для описания колеблемости? Величина X_i—M характеризует вклад, вносимый в колеблемость i-ым вариантом.

[39]

Вклад всех вариантов, казалось бы, естественно описать с помощью $\sum_{i=1}^{N}(x_i - M)$. Однако, как легко видеть с учетом определения M, эта величина всегда обращается в нуль, следовательно, она не может быть принята в качестве меры колеблемости. Мы получаем нуль из-за взаимной компенсации отклонений разных знаков, т.е. вправо и влево относительно M. Наверное, целесообразно освободить отклонения от знаков (в самом деле, ведь и отклонения влево, и отклонения вправо — колеблемость, следовательно, они должны равноправно входить в искомый показатель). В простейшем случае это можно осуществить, переходя к величине $(x_i-M)^2$, которая нивелирует различие «правых» и «левых» отклонений вариантов от M, а для полного вклада к $\sum_{i=1}^{N}(x_i - M)^2$. Для сопоставимости различных распределений нужно перейти ко вкладу, приходящемуся на долю одного наблюдения: $\frac{1}{N}\sum_{i=1}^{N}(x_i - M)^2 = D$; эта величина называется *дисперсией,* ее размерность есть квадрат размерности признака. За меру колеблемости естественно принять величину $\sigma = \sqrt{D}$, которая имеет ту же размерность, что и сам признак; она называется *среднеквадратичным (или стандартным) отклонением.* Если колеблемости нет, все $x_i=M$ и $\sigma = 0$. Если а мало, то M хорошо представляет ряд, он достаточно однороден. Чем больше о, тем больше колеблемость.

Итак, а показывает, на сколько в среднем отклоняется каждый вариант от *M.* Допустим, что мы сравниваем признаки, имеющие одинаковую размерность. Например, это могут быть общий трудовой стаж, стаж на данном предприятии и т.д. Если одинаковы *M,* то колеблемость больше у того признака, у которого больше о. Если одинаковы о, то это, вообще говоря, не означает, что одинаковы колеблемости. В этом случае колеблемость там меньше, где больше *M.* Для сопоставлений, очевидно, следует перейти к относительному показателю. Таковыми является *коэффициент вариации,* $C_v = \frac{\sigma}{M} 100\%$. Сравнивая C_v для общего трудового стажа и стажа на данном предприятии, мы можем сопоставить колеблемость данных признаков индивидов изучаемой общности. Пока речь шла о признаках одинаковой размерности; если же сопоставляемые признаки имеют раз-

[40]

личную размерность, то использование коэффициента вариации является единственно возможным способом сравнения колеблемостей. Примерами такого типа являются сопоставления колеблемостей образовательного и квалификационного уровней работников данной профессиональной группы, аналогично для стажа и квалификации, зарплаты и стажа и т.д., в зависимости от стоящей перед исследователем задачи.

Свойства средней арифметической величины.

1. Если все варианты увеличить (или уменьшить) в а раз, то *M* увеличится (или уменьшится) во столько же раз.

Упражнение 4. Показать самостоятельно (для этого нужно использовать свойства сумм — см. Приложение № 2).

2. Если все варианты увеличить на одно и то же число, то и *M* увеличится на то же число.

Упражнение 5. Показать самостоятельно. Указание: для этого нужно сделать переход $x_i \rightarrow x'_i = x_i + a$ и вычислить среднее x'_i с использованием свойств сумм, как и в упражнении № 4.

3. Сумма произведений отклонений вариантов от *M* на частоты равна нулю.

В самом деле, с учетом определения *M* имеем:

$$\sum_{i=1}^{k} N_i (x_i - M) = \sum_{i=1}^{k} N_i x_i - MN = 0 .$$

4. При уменьшении (или увеличении) частот в одно и то же число раз средняя арифметическая не изменяется.

Упражнение 6. Показать справедливость утверждения самостоятельно.

5. Если совокупность *(N)* разбита на *s* непересекающихся классов ($N = \sum_{r=1}^{s} N_r$, здесь N_r – число индивидов в *r*-ом классе), то общая средняя $M = \bar{x}$ равна средней арифметической групповых средних $\overline{x_r}$ ($\overline{x_r} = \frac{1}{N_r} \sum_{i=1}^{k} x_i P_{ri}$, где P_{ri} — число индивидов с $X = x_i$ в *r*-ом классе), взятых с весами N_r.

В самом деле, по определению,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} x_i N(x_i) = \frac{1}{N} \sum_{r=1}^{s} \sum_{i=1}^{k} x_i P_{ri} = \frac{1}{N} \sum_{r=1}^{s} N_r \overline{x_r} ,$$

что и требовалось показать.

[41]

Упражнение 7. Показать, что

$$\bar{x} = \alpha \frac{1}{N} \sum_{i=1}^{k} \frac{x_i - c}{\alpha} N(x_i) + c \quad \text{(I,3,1)}$$

(Для этого нужно воспользоваться свойствами 1 и 2.)

Упражнение 8. Пусть на заводе три цеха: A, B, и C. Допустим, что средний стаж на данном рабочем месте для работников цеха A — 3,8 года, для работников цеха B — 4,0 года, для работников цеха C — 4,2. Чему равен средний стаж на рабочем месте для всего предприятия в целом? Это зависит от того, сколько работников в каждом из цехов. Пусть в цехе A — 100 человек, в цехе B — 400, в цехе C — 500. Тогда средний стаж для всего предприятия равен:

$$3{,}8\frac{100}{1000} + 4\frac{400}{1000} + 4{,}2\frac{500}{1000} = 4{,}1 \text{ (года).}$$

Такое среднее называется *взвешенным.*

Перейдем к изучению других средних.

Медиана Me — значение признака, которое приходится на центральный (средний) член ранжированного ряда.

У одной половины членов ряда значения признака меньше, чем у среднего, у другого — больше. Допустим, что в отделе главного механика работает 9 человек, возраст которых соответственно: 18, 18, 27, 30, 34, 35, 37, 40, 63 (в годах). Тогда, согласно определению, *Me*=34 года: это возраст работника с условным номером 5. Из оставшихся у половины (№ 1 — 4) возраст меньше, у половины (№6 — 9) больше, чем медианный. Допустим, что в отделе главного бухгалтера 6 человек, возраст которых: 19, 23, 38, 42, 54, 67. По определению принимают, что $Me = \frac{38+42}{2} = 40$. Теперь вообще нет работника с медианным возрастом, ко ровно у половины индивидов возраст меньше, чем *Me,* а у другой — больше.

На медиану влияют лишь центральные, срединные значения признака. Если концы распределений — левый или правый — определены ненадежно, то это не исказит *Me,* поможет исказить *M,* которое зависит от *всех* значений признака.

Заметим, что в некоторых ситуациях применение *M* вообще оказывается невозможным, и *Me* выступает в роли средней, репрезентирующей ряд. Это относится к качественным признакам.

Как вычислить медиану в случае интервального ряда?

[42]

Рассмотрим кумулятивный ряд, т.е. ряд накопленных частот. Медианный интервал — тот, на который приходится 0,5*N* наблюдений. Пусть его номер *l*, тогда $N_l = F_l - F_{l-1}$. Все эти варианты заключены между x'_l и x''_l. Мы не знаем точных значений каждого из вариантов, поэтому в простей-шем случае естественно предположить, что внутри интервала все они расположены равномерно, т.е. прирост частоты пропорционален приросту интервала:

$$N_l : I_l = (0{,}5N - F_{l-1}) : (Me - x'_l)$$

Теперь

$$Me = x'_l + I_l \frac{0{,}5N - F_{l-1}}{N_l}. \quad \text{(I,3,2)}$$

Проиллюстрируем это графически:

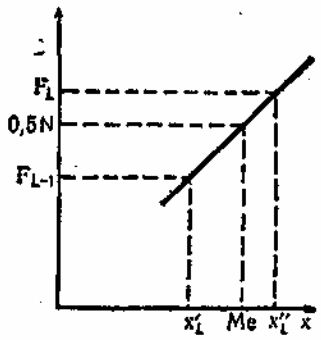

Рис. 12. Определение медианы

Если перейти к частостям, выраженным в %, то:

$$Me = x'_l + I_l \frac{50 - f_{l-1}}{v_l} \quad \text{(I,3,2')}$$

Упражнение 9. По данным примера № 2 вычислить медианный возраст работников.

Мода M_0 — наиболее часто встречающееся в данной совокупности значение признака.

Можно сказать и так: мода — вариант с наибольшей частотой.

Когда продавец говорит о «среднем покупателе», то он, возможно, и не осознавая этого, по существу имеет в виду модального. Мода не отражает степени модальности, сама по себе она не несет информации о том, насколько распространено данное значение признака.

[43]

В отличие от *M* и *Me, Mo* может представлять и классификационные признаки. Можно указать модальную национальность данного государства (например, в СССР это русские), модальную профессию на предприятии или в отрасли и т.д., хотя бессмысленно говорить о средней арифметической или медианной профессии, национальности и т.д.

Как отмечалось, если распределение имеет один максимум, его называют унимодальным (мода — абсцисса макси-

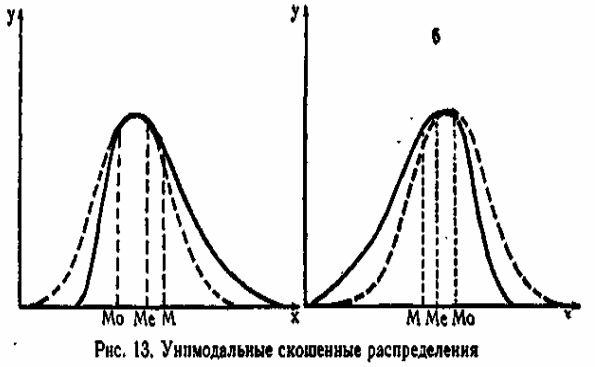

Рис. 13. Унимодальные скошенные распределения

мума), если два, то бимодальным и т.д. Теперь проясняется смысл этих названий. Возрастное распределение населения, например, в отсутствии войн, эпидемий и т.п., обычно имеет колоколообразный вид. У симметричных унимодальных распределений *M=Me—Mo*.

Перейдем к рассмотрению скошенных унимодальных распределений. Сопоставим их с базовым симметричным, которое будем изображать пунктирной кривой. У распределения на рис. 13а «поднят» правый, но «опущен» (по сравнению с симметричным) левый конец. На *Mo* края не влияют (она определяется максимумом, который не изменился, по условию), ее положение не меняется. Как мы видели, на *M* влияют все значения, следовательно, *M* сдвинется, причем в сторону больших значений *X* (поднят правый конец!).

Me тоже сдвигается, но так как она определяется не столько значениями признака, сколько частостей, а в «хвостах» (концах) концентрация события невелика, то и сдвиг *Me*

относительно небольшой. Отсюда становится понятным указанное на рисунке взаимное расположение *Mo, Me, M: Mo<Me<M.*

[44]

Упражнение 10. Показать, что если поднят левый конец, то *M<Me< Mo* (рис. 136).

Итак, если *M, Me, Mo* совпадают (либо близки), то распределение симметрично (либо близкое к симметричному). Если же они значительно разнятся и *Mo > Me,* то имеет место левая асимметрия, если *Mo<Me* — правая.

(Замечание: для контроля вычислений можно использовать то, что *Me* всегда между *Mo* к *M).*

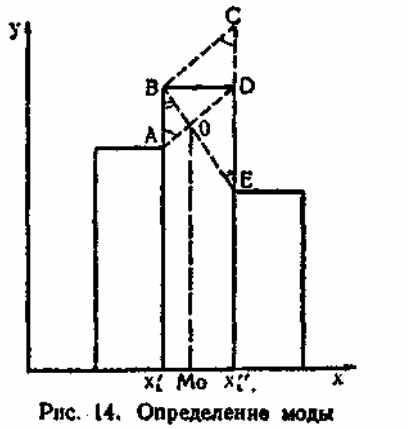

Рис. 14. Определение моды

До сих пор мы предполагали, что умеем вычислять *Mo.* Как же это делать практически в случае наиболее часто встречающихся в социологии интервальных рядов? Прежде всего нужно найти интервал с наибольшим числом наблюдений. Отметим, что при неравных интервалах во избежание ошибок от частоты нужно перейти к плотности. Интервал с наибольшей частотой при равных интервалах (или с наибольшей плотностью при неравных) и есть модальный. Пусть его номер *l.* Естественно предположить, что внутри этого интервала частоты распределены «в согласии» с соседними интервалами: если левый столбик диаграммы (рис. 14) выше, то *Mo* ближе к x_l' если правый, то к x_l''. По определению, в качестве медианы принимается абсцисса точки 0 — пересечения отрезков *BE* и *AD*, удовлетворяющая указанному предположению[28]. Пусть $Mo = x_l' + \Delta x$. Для нахождения Δx проведем (*BC*)II(*AD*) до пересечения с продолжением (*DE*) в точке *C*.

Из подобия треугольников *A0B* и *BCE:*

$$\frac{\Delta x}{I_l} = \frac{|AB|}{|CD| + |DE|};$$

$$|AB| = N_l - N_{l-1}$$

аналогично

$$|CD| + |DE| = |AB| + |DE| = 2N_l - N_{l-1} - N_{l+1}$$

[45]

Следовательно,

$$Mo = x_l' + I_l \frac{N_l - N_{l-1}}{2N_l - N_{l-1} - N_{l+1}}. \quad (I,3,3)$$

[28] Отметим, что мы тем самым доопределили *Mo* (!).

Пример 3. Рассмотрим вычисление средних доходов *(M, Me, Mo)* семей США[29] (1959 г.).

Средняя арифметическая $M \approx 6500$ в данном случае мало показательна, ибо здесь усредняются «тигры и кошки»,

Таблица 8

Распределение годового дохода семей США

Годовой доход в долларах	v_i Частость, %	f_i Кумулятивная астость, %
до 2000	14	14
от 2000 до 4000	21	35
от 4000 до 6000	23	58
от 6000 до 8000	18	76
от 8000 до 10000	10	86
от 10000 до 15000	9	95
свыше 15 000	5	100

что порождает, пользуясь словами В. И. Ленина, «иллюзию благоденствия».

Для уяснения ситуации вычислим *Me* и *Mo.* Медианный и модальный интервалы у нас совпадают, это 4000–6000:

1) именно на этот интервал приходится максимальная частота (23%);

2) на этот же интервал приходится 50% наблюдений.

Из (I,3,2'): *Me* = 5300, т.е. у 50% семей доход на 20% ниже среднего арифметического. Из (I,3,3) *Mo* = 4600, т.е. наиболее часто встречающийся доход примерно на 30% ниже среднего арифметического.

Упражнение 11. Какой процент американских семей в 1959 г. имел доход ниже, чем средний арифметический?

Ответ: 63%

Упражнение 12. Каков процент семей с доходом ниже модального? Ответ: 42%

Пример 4. По данным табл. 9 о распределении роста 1000 взрослых рабочих-мужчин вычислить *M*, *Mo* и *Me*. Полагая C=165,5, a=3, имеем, используя формулу

[46]

(I,3,1): $M = 165{,}5$. Согласно (I, 3,2): $Me = 164 + 3 \times \frac{500 - 403}{201} = 165{,}5$ (см), а по (I, 3,3): $Mo = 165{,}2$ см. Таким образом, *M, Me* и *Mo* практически совпадают.

Начертим гистограмму (рис. 15). Мы видим, что она, как и следовало ожидать, почти симметрична.

Таблица 9

Вычисление среднего арифметического, моды и медианы

X	x_i	N_i	$x_i - C$	$\frac{x_i - C}{\alpha}$	$\frac{x_i - C}{\alpha} N_i$	$\left(\frac{x_i - C}{\alpha}\right)^2 N_i$	F_i
1	2	3	4	5	6	7	8
143–146	144,5	1	–21	–7	–7	49	1
146–149	147,5	2	–18	–6	–12	72	3
149–152	150,5	8	–15	–5	–40	200	11
152–155	153,5	26	–12	–4	–101	416	37
155–158	156,5	65	–9	–3–	–195	585	102
158–161	159,5	120	–6	–2	–240	480	222

[29] Самуэльсон П. Экономика. М., 1964, с.196

161–164	162,5	181	–3	–1	–181	181	403
164–167	165,5	201	0	0	0	0	604
167–170	168,5	170	3	1	170	170	774
170–173	171,5	120	6	2	240	480	894
173–176	174,5	64	9	3	192	576	958
176–179	177,5	28	12	4	112	448	986
179–182	180,5	10	15	5	50	250	996
182–185	183,5	3	18	6	18	108	999
185–188	186,5	1	21	7	7	49	1000

Рассмотренный пример позволяет перейти к очень важному распределению – нормальному. Сперва несколько вводных замечаний. Рассмотрим последовательность $S_n = \left(1+\frac{1}{n}\right)^n$. Легко видеть, что $S_1 = 2; S_2 = 2{,}25; S_3 = 2{,}39, \ldots, S_{100} = 2{,}69.$

Упражнение 13. Используя логарифмирование, вычислить S100.

Предел, к которому стремится S_n при неограниченном увеличении n, оказывается некоторым иррациональным числом, которое обозначается через *e*. Можно показать, что, например, с точностью до 4 знаков после запятой e = 2,7183.

[47]

Говорят, что величина *X* распределена нормально, если теоретическая кривая плотности распределения описывается функцией типа

$y = y_0 e^{-(x-V)^2/2q^2}$ (I, 3, 4)

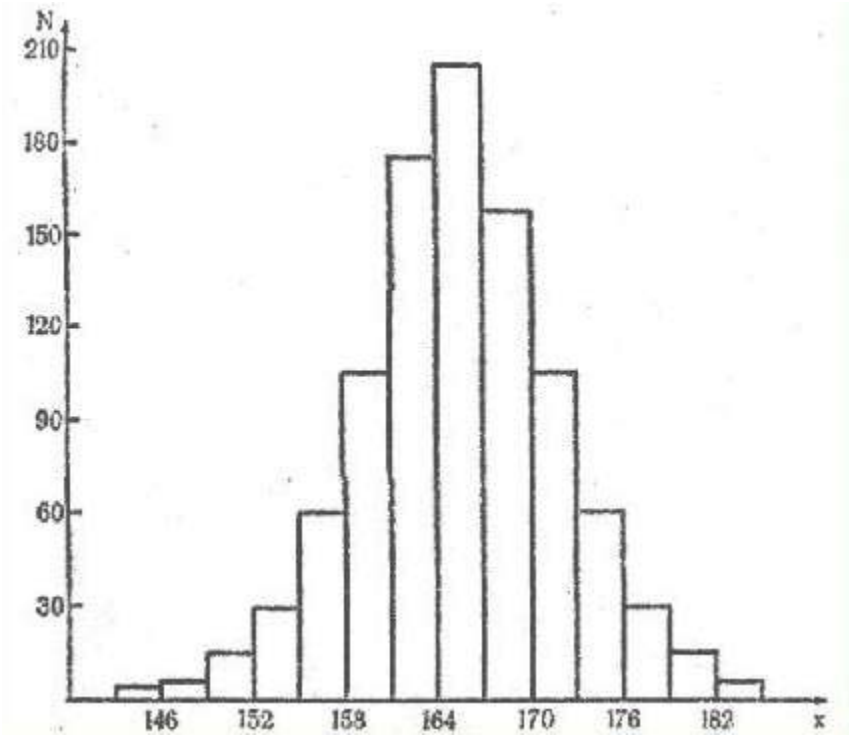

Рис. 15. Гистограмма распределения мужчин по росту

где *M* – среднее арифметическое (абсцисса, относительно которой симметрична кривая), σ – среднее квадратическое, а y_0 – максимальная ордината, равная $\frac{1}{\sigma\sqrt{2\pi}}$. Эта колоколообразная кривая ассимптотически приближается к оси *x*. Нормальное распределение полностью определяется величинами M и σ. Вид кривой не зависит от *M,* которое определяет лишь

положение максимума, его абсциссу (ордината – y_a). Ясно, что $M = M_0 = M_e$. Форма (вид) кривой определяется величиной σ. Вся площадь, ограниченная этой кривой и осью абсцисс, равна **N**, если по оси ординат отложены частоты, или если – частости (именно из этого условия получено значение $y_0 = \frac{1}{\sigma\sqrt{2\pi}}$), или 100% (если –

[48]

проценты). Оказывается, 68,27 % наблюдений заключено между $M - \sigma$ и $M + \sigma$; 95,45% между $M - 2\sigma$ и $M + 2\sigma$; 99,73% – между M-3σ и M+3σ.

Составлены специальные таблицы, в которых для любого z (взятого с определенным шагом) указано, какая площадь, ограниченная кривой нормального распределения, лежит между $M - z\sigma$ и $M + z\sigma$ (см. Приложение 3, табл. А).

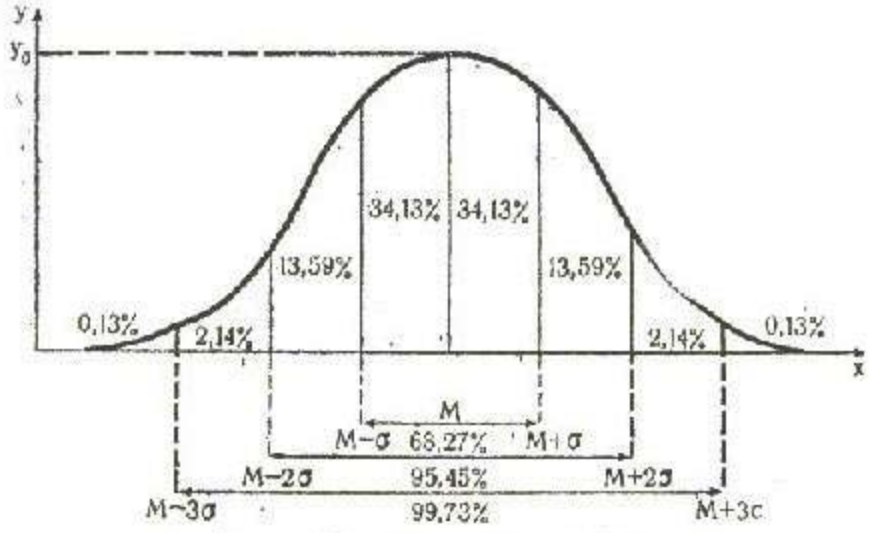

Рис. 16. Нормальное распределение

Так, для $z = 2$ эта площадь равна, как указывалось, 95,45%; для $z = 2{,}5$ – 98,76% и т.п. На рис.16 показано, какие доли площади, ограниченной кривой и осью абсцисс, заключены между соседними ординатами (например, между M и $M + \sigma$ 34,13% общей площади).

Распределение примера 4 очень близко к нормальному, в этом легко убедиться непосредственно. Так как σ = 6, то если бы распределение было точно нормальным, 99,72% наблюдений заключены были бы между 147,5 и 183,5. Легко видеть, что здесь ... 99,8%!

Завершая рассмотрение *M, Me* и *Mo,* сделаем существенное замечание. Как мы видели, среднее арифметическое совокупности, состоящей из нескольких групп, может быть выражено как средневзвешенное групповых средних арифметических. Этим свойством, однако, не обладают ни медиана, ни мода: *Me* и *Mo* для групп, из которых состоит изучаемая совокупность, мы ничего не можем сказать о *Me* и *Mo* этой совокупности: ее параметры нужно

[49]

вычислять заново. Таким образом, *Me* и *Mo* не поддаются арифметическим операциям.

Существуют и другие виды средних величин. Поскольку они не получили широкого применения в социологии, ограничимся кратким знакомством с ними.

Средней геометрической величин $x_1, x_2, \ldots, x_N$ по определению, называется величина $G_N = \sqrt[N]{\prod_{i=1}^{N} x_i}$

Если варианты повторяются, то $G_N = \prod_{i=1}^{R} x_i^{Vi}$

Средней гармонической называется величина $H_N = \dfrac{N}{\sum\limits_{i=1}^{K} \dfrac{1}{x_i}}$.

Если варианты повторяются, то $H_N = \dfrac{1}{\sum\limits_{i=1}^{K} \dfrac{v_i}{x_i}}$. Средняя квадратическая $S_N = \sqrt{\dfrac{1}{N}\sum\limits_{i=1}^{N} x_i^2}$.

Можно доказать, что $H_N \le G_N \le M_N \le S_N$. Оказывается, что H_N, G_N, M_N, S_N могут быть определены с помощью одной формулы: $\alpha_z = \sqrt[z]{\dfrac{1}{N}\sum\limits_{i=1}^{N} x_i^z}$. Полагая, что z = -1, 0, 1, 2, мы получим H_N, G_N, M_N, S_N соответственно. Доказательство этого составляет содержание *упражнения 14*. Примечание. В случае z=0 нужно сперва вычислить $\ln a_z$, а затем перейти к пределу, когда z→0.

Завершая рассмотрение, отметим, что H_N, G_N, M_N, S_N в отличие от *Mo* и *Me* (две последние величины называют структурными средними) зависят от всех значений признака.

4. Меры вариации

В §3 мы уже познакомились с такими мерами колеблемости, как вариационный размах и дисперсия. Ввиду особой значимости для статистики понятия дисперсии остановимся на нем подробнее.

По определению, дисперсия представляет собой среднее арифметическое квадратов отклонений вариантов от среднего арифметического значения признака для данной совокуп-

ности, т.е.

$$D = \frac{1}{N}\sum_{i=1}^{N}(x_i - M)^2 = \frac{1}{N}\sum_{i=1}^{k} N(x_i)(x_i - M)^2 \quad \text{(I, 4,1)}$$

Пример 5. Вычислим *M, D,* σ и C_v для квалификации рабочих Одесского судоремонтного завода образованием 7 классов. Для этой совокупности вариационный ряд имеет следующий вид:

Квалификация (в разрядах)	$x_1=1$	$x_2=2$	$x_3=3$	$x_4=4$	$x_5=5$	$x_6=6$	*Всего*
Частота	*8*	*40*	*42*	*65*	*77*	*53*	*285*

$$M = \bar{x} = \frac{1\cdot 8 + 2\cdot 40 + 3\cdot 42 + 4\cdot 65 + 5\cdot 77 + 6\cdot 53}{285} = 4{,}13 \text{ (разряда)}$$

$$D = \frac{(1-4.13)^2\cdot 8 + (2-4.13)^2\cdot 40 + (3-4.13)^2\cdot 42 + (4-4.13)^2\cdot 65 + (5-4.13)^2\cdot 77 + (6-4.13)^2\cdot 53}{285} =$$

$$= 1.93$$

σ = 1.39 разряда

C_v = 33,8%

Упражнение 15. Найти *M, D,* σ, C_v для рабочих, имеющих общее среднее образование, если распределение имеет вид:

x_i	1	2	3	4	5	6
$N(x_i)$	53	232	212	153	99	34

Ответ: *M* = 3,15; *D* = 1,712,

σ = 1,31 C_v = 41,6%

Так как эти данные получены в одном и том же конкретном социологическом исследовании, целесообразно их сопоставить. Интересно, что у рабочих-судоремонтников с образованием 10-11 классов средний квалификационный разряд на единицу меньше, чем у рабочих с образованием 7 классов. Дело в том, что у рабочих со средним образованием значительно меньше средний стаж (примерно на 9 лет), а для данной специальности стаж в большей мере влияет на квалификацию, чем образование. Подчеркнем, что это локальный вывод, существуют профессии, где решающую роль в квалификации играет образование. Обратим внимание и на то, что группа рабочих с общим средним образованием несколько более разнородна по своему составу в смысле

[51]

квалификации (ср. коэффициенты вариации), а также в смысле стажа и возраста рабочих.

Познакомимся с основными свойствами дисперсии.

1. Если все варианты увеличить (или уменьшить) в одно и то же число, скажем, α раз, то D увеличится (или соответственно уменьшится) в α^2 раз.

Мы совершаем переход $x_i \to x_i' = ax_i$. При этом, очевидно, $M = \bar{x} \to \overline{x'} = ax$, а $D \to D' = \alpha^2 D$.

Заметим, что $\sigma \to \sigma' = \alpha\sigma$.

2. Увеличение (или уменьшение) всех вариант на одну и ту же постоянную величину c не изменит дисперсию. Теперь, $x_i \to x_i' = x_i + c$, очевидно, $\bar{x} \to \overline{x'} = \bar{x} + c$, а $x_i' - \overline{x'} = x_i - \bar{x}$, т.е. $D'' = D$.

3. При увеличении (или уменьшении) всех частот в одно и то же число раз дисперсия не изменится.

4. Дисперсия относительно средней арифметической равна дисперсии относительно произвольной постоянной за вычетом квадрата разности средней арифметической и этой постоянной.

Представляя $(x_i - \overline{x})$ в виде $\left[(x_i - c) - (\bar{x} - c)\right]$, имеем:

$$D = \frac{1}{N}\left[\sum_i N(x_i)(x_i - c)^2 - 2(\bar{x} - c)\sum_i (x_i - c)N(x_i) + (\bar{x} - c)^2 N\right] = \frac{1}{N}\sum_i N(x_i)(x_i - c)^2 - (\bar{x} - c)^2$$

или: $D(\bar{x}) = D(c) - (\bar{x} - c)^2$.

Отсюда: $D(c) - D(\bar{x}) + (\bar{x} - c)^2$ (I,4,2)

т.е $D(c) \geq D(\overline{x})$

Таким образом, дисперсия относительно среднего арифметического (мы ее будем называть собственно дисперсией, или для простоты, просто дисперсией) обладает свойством минимальности: она меньше дисперсии относительно любой другой величины.

5. Дисперсия равна средней арифметической квадратов вариантов, уменьшенной на квадрат средней арифметической.

[52]

В самом деле, полагая в (I,4,2) $c = 0$, получим:

$D = \overline{x^2} - \bar{x}^2$ (I, 4,3)

Следствием свойств 1 и 4 является равенство

$$D = \frac{\alpha^2}{N}\sum_{i=1}^{k} N(x_i)\left(\frac{x_i - c}{\alpha}\right)^2 - (\bar{x} - c)^2 \qquad \text{(I,4,4)}$$

которое может быть использовано для упрощения вычисления дисперсии.

Упражнение 16. Вернемся к примеру 4 §3 и завершим его рассмотрение, вычислив *D,* σ *и* C_v для распределения по росту взрослых мужчин. Для этого могут быть использованы данные таблицы 9, в седьмой колонке которой приведены величины, необходимые для применения формулы (I,4,4). Как и ранее, c = 165,5, а α = 3. Ответ: D = 36,58; σ= 6,05; C_v = 3,8%. Если читатель выполнит вычисление дисперсии и по формуле (I,4,1), то он сумеет оценить преимущество (I,4,4).

Познакомимся с правилом сложения дисперсий. Будем считать, что изучаемая совокупность разбита на s непересекающихся групп.

Пусть в *r*-ой группе x_i встречается P_{r_i} раз, ясно, что $\sum_{i=1}^{k} P_{r_i} = N_r$, т.е. числу индивидов в *r*-ой группе, а $\sum_{r=1}^{s} P_{r_i} = N(x_i)$ – общему числу индивидов с $x = x_i$. Групповое среднее $\overline{x_r} = \frac{1}{N_r}\sum_{i=1}^{k} x_i P_{r_i}$, а групповая дисперсия суть $(r = \overline{1,s})$:

$$\sigma_r^2 = \frac{1}{N_r}\sum_{i=1}^{k}(x_i - \overline{x_r})^2 P_{r_i} = \frac{1}{N_r}\sum_{i=1}^{k} x_i^2 P_{r_i} - \overline{x_r^2} \qquad \text{(I,4,5)}$$

Межгрупповой дисперсией, по определению, называется средняя арифметическая величина квадратов отклонений групповых средних $(\overline{x_r})$ от общей средней $\overline{x}$, т.е.

$$\partial^2 = \frac{1}{N}\sum_{r=1}^{s}(\overline{x_r} - \overline{x})^2 N_r = \frac{1}{N_r}\sum_{r=1}^{s}\overline{x_r^2}N_r - \overline{x^2} \qquad \text{(I,4,6)}$$

Средняя арифметическая групповых дисперсий:

$$\overline{\sigma^2} = \frac{1}{N}\sum_{r=1}^{s} N_r\sigma_r^2 \qquad \text{(I, 4,7)}$$

[53]

Теперь мы получили возможность сформулировать правило сложения дисперсий:

$$\sigma^2 = \overline{\sigma^2} + \delta^2 \qquad \text{(I, 4, 8)}$$

Покажем, что это действительно так.

Из (I, 4, 5):

$$\sum_{i=1}^{k} x_i^2 P_{r_i} = N_r\sigma_r^2 + \overline{x_r^2}N_r \quad (r = \overline{1,s})$$

Запишем s таких равенств и сложим их почленно, тогда:

$$\sum_{i=1}^{k} x_i^2 N(x_i) = \sum_{r=1}^{s} N_r\sigma_r^2 + \sum_{r=1}^{s} N_r\overline{x_r^2} \qquad \text{(I, 4, 9)}$$

Разделив обе части равенства (I, 4, 9) на N и вычтя из них по $\overline{x^2}$, получим с учетом (I, 4, 5-7): $\sigma^2 = \overline{\sigma^2} + \delta^2$, что и требовалось.

Пример 6.

Пусть совокупность из N = 150 индивидов состоит из трех групп (цехов), в первой N_1 = 40, во второй N_2 = 50, в третьей N_3 = 60 человек (здесь s = 3; r = I, 2, 3).

Эмпирические данные сведем в таблицу 10.

Найдем сперва $\overline{x_r} : \overline{x_1} = \frac{65\cdot 7 + 75\cdot 12 + 85\cdot 15 + 55\cdot 6}{40} = 80$ (руб.)

Аналогично: $\overline{x_2} = 95$ руб.; $\overline{x_3} = 105$ руб.

В качестве *упражнения 17* предлагаем вычислить $\bar{x}$ сначала как средневзвешенное $\overline{x_2}$, а затем непосредственно, по определению. В обоих случаях должен получиться один и тот же результат: $\bar{x} = 95$ руб.

Далее. Найдем σ_r (*r*= I, 3, 2). Например,

$$\sigma_1^2 = \frac{(65-80)^2 \cdot 7 + (75-80)^2 \cdot 12 + (85-80)^2 \cdot 15 + (95-80)^2 \cdot 6}{40} = 90$$

Упражнение 18. Найти σ_2^2 и σ_3^2. Ответ: 140; $66\frac{2}{3}$.

Следующий шаг. Вычислим межгрупповую дисперсию:

$$\delta^2 = \frac{1}{N}\sum_{r=1}^{s}(\overline{x_r} - \bar{x})^2 N_r = 100,$$

а также $\overline{\sigma^2} = 97\frac{1}{3}$ и $\sigma^2 = 197\frac{1}{3}$

Таким образом, $\sigma^2 = \delta^2 + \overline{\sigma^2}$.

[54]

Приведем примеры вычислений *M,* σ и C_v для двухмерных распределений.

Пример 7. Пусть первый признак *X* – заработная плата рабочих (в рублях), а второй – *Y* – квалификация (в разрядах). Второй признак дискретный, а первый интервальный (величины соответствующих интервалов представлены в табл. II). 2131 рабочий, подвергнутые обследованию, в частности, по признакам *X* и *Y* распределились так, как

Таблица 10

Пример расчета межгрупповой и внутригрупповой дисперсии

Заработная плата, руб.	x_i, руб.	P_{1i}	P_{2i}	P_{3i}	$N(x_i)$
60—70	65	7	1	0	8
70—80	75	12	5	0	17
80—90	85	15	9	4	28
90—100	95	6	18	8	32
100—110	105	0	12	32	44
110-120	115	0	5	16	21
Всего	—	40	50	60	150

показано в табл. 11. Например, 18 человек имеют первый разряд и получают до 80 руб., 28 – первый разряд и зарплату в интервале от 80 до 100 руб. Всего рабочих с первым разрядом 121,

со вторым 523 и т.д. Всего получающих зарплату до 80 руб. – 107 чел., от 80 до 100 руб. – 216 и т.д.

Таким образом, в колонке $N(x_i)$ по сути представлено распределение рабочих по разрядам, а в строке $N(y_i)$ – по заработной плате. Это вариационные ряды типа ранее рассмотренных. Кроме того, наша табл. 11 содержит специфические ряды типа: распределение рабочих с данной зарплатой по разрядам и распределение рабочих с данным разрядом по величине заработной платы.

В столбцах приведены также средние значения X (например, средняя зарплата рабочих с первым квалификационным разрядом 113,1 руб., а средний разряд рабочих, заработная плата которых от 180 до 200 руб., составляет 4,03 разряда). Далее представлены соответствующие σ и C_v.

Мы приводим эту таблицу не столько из-за ее информационной ценности, сколько для того, чтобы читатель мог

[55]

Таблица 11

Пример расчета M, σ, C_v **(зарплата)**

разряд	Зарплата, руб.												
	до 80	80-100	100-120	120-140	140-160	160-180	180-200	200-220	свыше 220	$N(x_i)$	$\bar{x}$	σ_x	$C_v^{(x)}$
1	18	28	26	26	15	6	1	1	0	121	113,1	30,44	26,9
2	42	77	128	140	67	31	16	13	9	523	125,6	34,93	27,8
3	33	50	84	139	123	20	19	11	8	527	134,4	33,80	25,1
4	7	45	71	66	65	72	44	24	22	416	148,6	42,35	28,4
5	4	12	49	50	57	46	34	38	54	344	167,3	48,51	29,0
6	3	4	23	53	47	29	12	9	20	200	155,7	41,58	26,6
$N(y_i)$	107	216	381	474	374	244	126	96	113	2131			
$\bar{y}$	2,49	2,75	3,15	3,28	3,59	3,85	4,03	4,16	1,60	–			
σ_y	1,13	1,18	1,39	1,41	1,37	1,27	1,17	1,22	1,10	–			
$C_v^{(y)}$	45,3	43,0	44,3	43,0	38,2	33,0	29,0	29,3	23,9	–			

при желании проверить себя и рассчитать показатели, которые были рассмотрены в предыдущих параграфах.

Что же касается содержательной интерпретации данных, кроме достаточно очевидных утверждений типа «с увеличением разряда увеличивается средняя заработная плата», из нее можно почерпнуть менее очевидное: с увеличением заработной платы группы рабочих становятся все более однородными по уровню квалификации (монотонное уменьшение C_v), хотя с увеличением разряда вариация заработной платы не изменяется: разброс примерно один и тот же.

Пример 8.

При изучении связи между признаками квалификация *(X)* и удовлетворенность специальностью (*Y*), в частности, была получена такая таблица (таблица 12).

Как и ранее, X выражается в разрядах $(x_i = i, i = \overline{1,6})$. Признак Y – качественный, его позиции: «удовлетворен», «не знаю, трудно сказать», «не удовлетворен» обозначены в таблице соответственно y_1, y_2, y_3. Удовлетворенность группы рабочих описывается с помощью индекса $J_{спец} = \dfrac{N_+ - N_-}{N_+ + N_0 + N_-}$, где N_+ – число удовлетворенных, N_- – неудовлетворенных, N_0 – не выразивших определенное отношение.

Так как в дальнейшем нам придется неоднократно рассматривать индексы для группы, отметим, что J принимает значения, заключенные между –1 и 1, причем –1 соот-

[56]

ветствует случаю, когда все работники не удовлетворены, 1 означает, что все удовлетворены, а 0 получается в случае, когда число удовлетворенных специальностью равно числу неудовлетворенных. Аналогично конструируются индексы удовлетворенности работой, различными элементами рабочей ситуации и т.д.

Возвратимся к табл. 12. Кроме «очевидных» утверждений типа «с увеличением квалификации увеличивается и удовлетворенность специальностью», из нее следует, что группы индивидов с разной удовлетворенностью специальностью примерно одинаковы по вариации квалификации, а с увеличением квалификации резко возрастает однородность групп по степени удовлетворенности: последняя складывается из все более согласованных индивидуальных оценок.

До сих пор мы рассматривали упорядоченные (количественные и качественные) признаки. Возникает вопрос, что может служить мерой вариации классификационных признаков?

Вариация классификационных признаков. Очевидно, меры, разработанные для признаков, значения которых числа, оказываются теперь непригодными: между объектами разных классов нет упорядочения (все классы равноправны – нельзя выделить континуум, в котором можно было бы упорядочить национальную или расовую принадлежность, членство в различного рода организациях или причины

[57]

увольнения с предприятия и т.д.), нет нуля, нет интервалов, теряют смысл такие понятия, как диапазон, размах, отклонение, столь привычные и удобные, когда значения признаков – числа.

Тем не менее объекты, входящие в разные классы, обладают различными качествами в смысле изучаемого признака: у них разный пол и разная национальность, они принадлежат к разным организациям или указывают разные причины увольнения и т.д.

Таблица 12

Пример расчета M, σ **и** C_v **(удовлетворенность)**

X	Y			$N(x_i)$	$\bar{x}$	σ_x	C_v^x (%)
	y_1	y_2	y_3				
1	66	4	30	100	0,36	1,09	302,8
2	327	16	77	420	0,59	0,86	145,8
3	353	25	61	439	0,66	0,76	115,2
4	295	25	34	354	0,74	0,65	87,8
5	271	16	15	302	0,85	0,49	57,6
6	172	3	6	181	0,92	0,38	41,3
$N(y_i)$	1484	89	223	1796			
$\bar{y}$	3,60	3,47	2,75				
σ_y	1,51	1,22	1,24				
C_v^y (%)	41,9	35,2	45,1				

Попробуем оценивать вариацию с помощью различия в качестве. Чем больше число различных пар объектов, тем, очевидно, больше вариация. Допустим, что у нас всего 2 класса объектов A и B (например, признак «пол»), численность которых N_A и N_B соответственно (объем совокупности $N = N_A + N_B$). В этом случае число различных пар объектов $N_A \cdot N_B$ (скажем,

каждый мужчина, очевидно, отличается от каждой женщины: на одного приходятся N_B женщин, т.е. N_B различий, а на всех N_A мужчин $N_A \cdot N_B$ различий). Для того, чтобы сконструировать нормированную меру, определим, в каком случае число пар максимально.

Как известно, среднее геометрическое двух чисел a и b не превосходит среднего арифметического и равно ему, если $a = b : \sqrt{ab} \le \frac{a+b}{2}$.

Пусть $a = N_A^2, b = N_B^2$, тогда имеем $N_A N_B \le \frac{N_A^2 + N_B^2}{2}$,

[58]

следовательно, $(N_A N_B)$ максимально, когда численности классов одинаковы, т.е. равны $\frac{N}{2}$; $(N_A N_B)_{\max} = \frac{N^2}{4}$, а искомая мера $\frac{4N_A N_B}{N^2} = \left(\frac{G_2}{M_2}\right)^2$. Итак, вариация максимальна, когда классы равнонаполненные, она при этом равна 1. Вариации нет, если, скажем, $N_A = 0$ ($N = N_B$ – все объекты однотипны), мера вариации при этом, очевидно, обращается в нуль.

А как быть, если классов больше чем 2, например, 3? Для двух классов мера равна квадрату отношения среднего геометрического к среднему арифметическому численностей классов. Казалось бы, в случае трех классов A, B, C, мера вариации должна быть $\left(G_3 / M_3\right)^2 = \frac{9N_A N_B N_C}{N^3}$. Легко видеть, что это не так. Допустим, что $N_A = 0$, тогда величина $\left(G_3 / M_3\right)^2$ обращается в нуль, хотя совокупность неоднородна: остались объекты типа B и C. Как же быть?

Составим величину

$$\alpha_3 = \frac{N_A N_B + N_A N_G + N_B N_G}{3\left(\frac{N}{3}\right)^2} \qquad \text{(I, 4,10)}$$

Она обращается в нуль, если по крайней мере два класса пусты (скажем, $N_A = N_B = 0$, т.е. совокупность однородна, состоит только из объектов типа C).

Максимальное значение обсуждаемая величина принимает при $N_A = N_B = N_C$, которое, как легко видеть, равно 1 (при этом различия максимальны). Величину α_3 можно принять в качестве меры вариации.

Упражнение 19. Показать, что $ab + ac + bc \le \frac{(a+b+c)^2}{3}$, причем равенство достигается при a=b=c. Указание: использовать трижды – для всех пар – неравенство между G_2 и M_2. Итак, $0 \le a_3 \le 1$, причем нуль соответствует однородной совокупности (отсутствие вариации), а единица – максимально неоднородной (максимальная вариация, случай равнонаполненных классов).

Упражнение 20. Рассмотреть случай $k = 4 (N = N_1 + N_2 + N_3 + N_4)$

Ответ:

$$\alpha_4 = \frac{8}{3} \frac{N_1 N_2 + N_1 N_3 + N_1 N_4 + N_2 N_3 + N_2 N_4 + N_3 N_4}{N^2} \qquad \text{(I, 4,11)}$$

[59]

Рассмотрим общий случай (произвольное k). Теперь число различий $A = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_i N_j$.

Найдем максимальное А, которое соответствует случаю $N_l = \frac{N}{k} (l = \overline{1,k})$:

$$A_{\max} = \frac{N_2}{k^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 1 = \frac{N^2}{k^2} \left(\sum_{j=2}^{k} 1 + \sum_{j=3}^{k} 1 + \ldots + \sum_{j=k-1}^{k} 1 + \sum_{j=k}^{k} 1 \right) = \frac{N^2}{k^2} \left[(k-1) + (k-2) + \ldots + 2 + 1 \right] =$$

$$= \frac{N^2 (k-1)}{2k},$$

таким образом,

$$\alpha_k = \frac{2k}{k-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} v_i v_j \qquad \text{(I, 4,12)}$$

Для описания вариации можно использовать также и энтропийную меру (см. § 5 главы II).

Квантили. Медиана, как мы видели, это значение признака, которое обладает таким свойством: 50% вариантов меньше, чем Me, 50% – больше. Естественным обобщением медианы является понятие квантиля. Квантиль делит сумму частот на заданное число равных частей. Число частей может быть различным, отсюда и разные квантили – квартили, децили, перцентили.

Квартиль. Квартиль (Q_i) делит сумму частот на четыре равные части. Очевидно, квартилей всего три: Q_1, Q_2, Q_3; Q_1 например, это значение признака, которое обладает таким свойством: 25% вариантов меньше, а 75% – больше его. Q_2 это Me, а Q_3 – значение признака, 75% вариантов меньше которого, а 25% – больше.

Прямые $x = Q_i (i = 1,2,3)$ делят площадь, ограниченную кривой распределения на 4 равные части: $S_1 = S_2 = S_3 = S_4$

На рис. 17а изображено распределение, а на рис. 17б показаны квартили на графике кумулятивной кривой. Подчеркнем, что точки, соответствующие квартилям, вообще говоря, делят отрезок $[x_{\min}, x_{\max}]$ на четыре неравные части. Между Q_1 и Q_3 заключена половина всех вариантов. Чем более плотно распределение, тем отрезок $[Q_1, Q_3]$ меньше. Таким образом, своеобразной мерой «разброса» может служить величина $\Delta Q = Q_3 - Q_1$.

[60]

Дециль. Дециль (D) делит сумму частот на 10 равных частей. Всего децилей, очевидно, девять: $D_1, D_2, \ldots D_9$. Ясно, что $D_5 = Q_2 = Me$. В качестве меры разброса используется также величина $\Delta D = D_9 - D_1$.

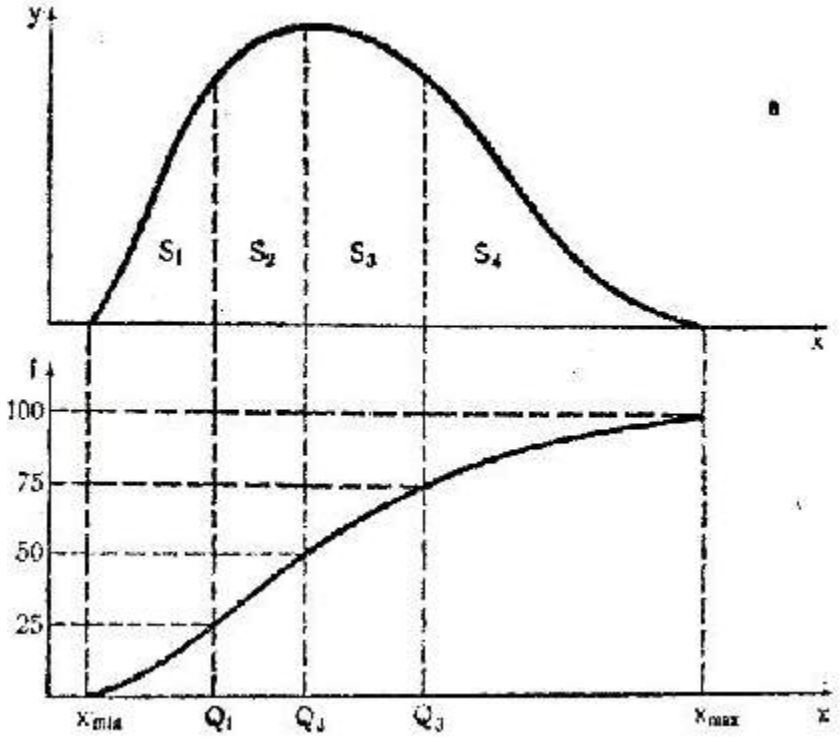

Рис. 17. Квартили на графике распределения (а) и на графике кумулятивной кривой (б)

Перцентиль, по определению, делит сумму частот на 100 равных частей: $C_1, C_2, \ldots, C_{99}$ Легко видеть, что, например, $D_1 = C_{10}$, $Q_1 = C_{25}$, $Me = C_{30}$, $Q_3 = C_{75}$ и т.д.

Как вычислять квантили в случае интервальных рядов? Вспоминая вывод формулы для $Me(Q_2)$, легко понять, что

$$Q_1 = x_i + I_i \frac{0{,}25N - F_{l-1}}{N_l},$$

$$Q_3 = x_i + I_l \frac{0{,}75N - F_{l-1}}{N_l},$$

где l – номер интервала, в который попадает соответствующий квантиль.

[61]

Упражнение 21. Вывести формулы для Q_1 и Q_3. Аналогично, например,

$$D_3 = x_l + I_l \frac{0{,}3N - F_{l-1}}{N_l}$$

$$C_{99} = x_l + I_l \frac{0{,}99N - F_{l-1}}{N_l}$$ и т.д.

Отметим, что квантиль – мера, применимая к самым различным типам упорядоченных данных. При вычислении квантилей вместо частот можно использовать частости.

Пример 9. По данным таблицы № 8 рассчитать Q_1 для годового семейного дохода в США (1959 г.). Нетрудно видеть, что $l = 2$, $x_l = x_2 = 2000$, $f_{l-1} = f_1$ 14, $v_l = v_2 = 21$, теперь $Q_1 \approx 3050$. Таким образом, 25% семей имели доход, меньший 3050 дол.

Упражнение 22. Вычислить Q_3, ΔQ, D_9. Нередко частоты крайних вариантов очень малы, величина вариационного размаха может создать впечатление большей колеблемости (величины вариации), нежели та, которая наиболее характерна для изучаемого распределения. В таких случаях целесообразно вычислять ΔQ или ΔD, в которых отражен диапазон, включающий в себя соответственно 50% и 80% всех наблюдений.

Упражнение 23. Какой процент американских семей имел доход ниже прожиточного минимума (3000 дол.)?

Далее мы рассмотрим применение изученных величин (Me, Q_i, ΔQ) к одной социологической задаче – измерению установки индивидов.

Пример 10. Шкала Терстоуна. С помощью этой шкалы измеряется ориентация (отношение, установка). Терстоун непосредственно изучал отношение к церкви (дальше мы подробно рассмотрим соответствующую процедуру), однако предложенный способ может быть использован для измерения различных установок. Итак, изучаемый признак – отношение.

Пункты шкалы устанавливаются не произвольно, а с помощью отбора суждений, осуществляемого судьями. Сперва при участии представителен обследуемого массива был составлен список, содержавший более ста высказываний, отражающих различное отношение к изучаемому феномену. Затем 300 судьям, представлявшим модель исследуемой аудитории, было предложено разложить карточки с высказываниями на 11 кучек: в первой должны быть суж-

дения наиболее благоприятные для церкви, во-второй – менее и т.д. до 11-ой, куда попадают наименее благоприятные суждения.

После того, как судьи завершили работу, нужно установить цену каждого суждения, меру согласованности судейских решений по каждому суждению и отобрать набор суждений, с помощью которых исследователь может изучать рассматриваемое отношение индивидов данной общности.

Цена суждения определялась как медиана распределения судейских решений, мера согласованности – квантильное отклонение.

Чтобы обработать результаты работы судей, для каждого из суждений первоначального списка составляется такая таблица:

Пункты шкалы	N_i(число судей, поместивших данное суждение в этот пункт)	v_i (% к общему числу судей)	f_i (кумулятивный %)
1	2	3	4
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	12	4%	4%
7	12	4%	8%
8	60	20%	28%
9	66	22%	50%
10	90	30%	80%
11	60	20%	100%
	300	100%	

Затем строится кумулята (рис. 18). При этом предполагается, что отношение изменяется континуально, пункты 1, 2, ..., 11 – отдельные точки, которые выделяют в данном континууме интервалы; ординаты кумуляты соответствуют серединам соответствующих интервалов.

Для представленного на графике суждения, как видно из чертежа, Me = 8,5; Q_1=7,3, Q_3=9,3, ΔQ=2,0. Проделав такую процедуру со всеми суждениями, в итоговую шкалу отбирают те, которые: 1) покрывают более или менее равномерно всю шкалу; 2) имеют наиболее согласованные

[63]

оценки, т.е. из нескольких суждений с близкими Me предпочтение отдается суждению с минимальным квартильным отклонением ΔQ.

Окончательная шкала содержит 10–15 суждений, каждое из которых имеет свой «вес» (цену) – медиану судейских решений. Отобранные суждения предлагаются респон-

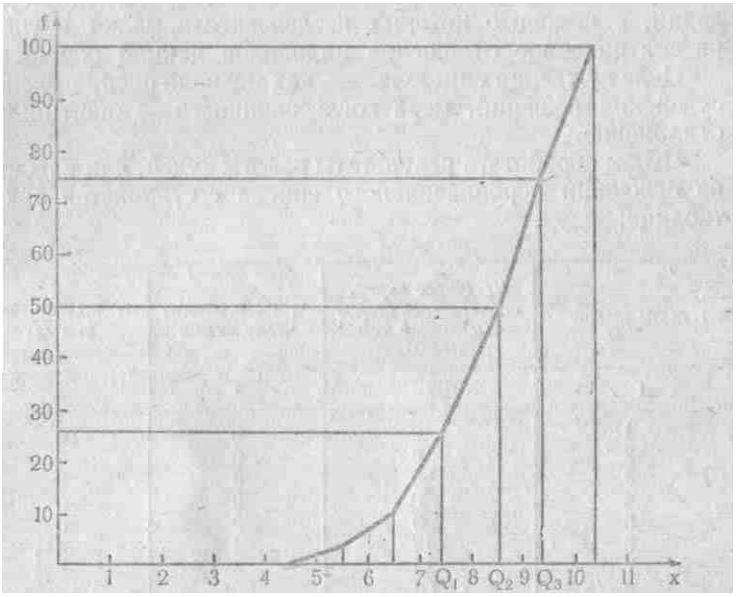

Рис. 18. Кумулята для построения шкалы Терстоуна

денту. Его ранг по данной шкале – медиана «весов» принятых им суждений, т.е. суждений, с которыми он согласен. Если респондент А согласен с такими пятью суждениями, у которых «веса»: 4,4; 4,8; 5,1; 5,6; 6,1, то его ранг 5,1. Если респондент В выбрал четыре суждения с «весами»: 7,6; 8,1; 8,5; 8,7, то его ранг 8,3 (медиана «весов» в случае четного числа суждений, по определению, $\frac{8{,}1 + 8{,}5}{2} = 8{,}3$).

Отметим, что шкала Терстоуна обладает рядом недостатков, устраненных в более совершенных методах[30].

[64]

[30] *Клигер С. А., Косолапов М. С., Толстова Ю. Н.* Шкалирование при сборе и анализе социологической информации. М., 1978, с. 71—81.