СОДЕРЖАНИЕ

1 СТАТИСТИЧЕСКАЯ ОБРАБОТКА БИОМЕДИЦИНСКОЙ ИНФОРМАЦИИ

1.1 Биомедицинская информация и способы ее получения

Биомедицинская информация-это сведения о свойствах биологических объектов и явлениях, являющихся предметами медицинских исследований, а также представления и суждения об этих свойствах и явлениях.

Биомедицинская информация может быть следующих видов:

1) *первичная информация* используемая для получения изображения в медицинской диагностике. Информация получается с использованием сложных диагностических способов, например рентгеновской автоматизированной томографии (АТ), ультразвуковой автоматизированной томографии и других методов. В процессе проведения обследования приёмник излучения (рентгеновского или ультразвукового) диагностических установок накапливает необходимые данные об исследуемом объекте, но для получения изображения с требуемым ракурсом, необходимо производить реорганизацию этих данных. Это требует большего числа вычислений, объём которых зависит от необходимой пространственной и яркостной разрешающей способности. В настоящее время для получения типовой рентгеновской томограммы требуется выполнить несколько сотен миллионов отдельных вычислительных операций. При этом обработка первичных данных должна происходить в реальном масштабе времени, т.е. с минимальной длительности процедуры от облучения пациента до получения результатов анализа.

2) *результаты индивидуального обследования* отдельных пациентов в лечебных учреждениях (поликлиника, клиника и т.д.). Это лабораторные исследования крови, мочи и др., общие рентгеновские обследования, ЭКГ и т.д. Данная информация необходима в комплексе для правильной и своевременной постановки диагноза и выбора метода лечения. Оперативное получение такой информации требует создания специализированных баз данных.

3) *статистическая информация о биологических объектах*, полученная в результате медико статистического исследования.

Например, исследования количества лейкоцитов в крови детей для определения условий, уровень каких-либо вредных веществ в крови для различных физиологических условий и т. д.

Такая информация получается в лечебных учреждениях на фактическом материале обследования пациентов для изучения каких либо закономерностей и тенденций, в НИИ при проведении биологических исследований (например, изучение влияния нового препарата на биологическое существо), в клиниках при проведении клинических исследований и т. д.

Обработка такой информации требует применения математических методов, в частности математической статистики. В приложении к медицине эти методы назвали *медицинской статистикой*, и в приложении к биологии в целом *–биологич еской статистикой*.

При обработке большого объема статистической информации требуется механизация и автоматизация вычислений.

Принципы, аппаратные и программные средства для проведения статистической обработки информации мы и будем изучать в этом курсе.

4) *информация*, *получаемая в области биохимических исследований веществ*, например, при синтезе новых лекарственных препаратов.

С помощью специализированных аппаратных и программных средств становиться возможным детально изучать структуры сложных макромолекул и их химически активные связывающие участки и исследовать как пространственное взаимодействие рецепторов с химически активными участками потенциально полезных лекарств, так и динамику этих молекулярных взаимодействий.

Большинство методов, используемых ранее для изучения этих свойств структуры и функций макромолекул, позволяло лишь косвенно исследовать молекулы. Только рентгеновская кристаллография могла прорисовать атомную структуру биологических молекул, таких как протеины. Основываясь на данных рентгеновской кристаллографии, исследователь был вынужден выполнять трудоемкую работу по созданию трехмерной физической модели молекулы из подручных средств (из палочек и проволоки).

Теперь стало возможным ввести кристаллографические данные любой макромолекулы в ЭВМ, затем изобразить молекулу в цвете и, поворачивая изображение, постепенно строить трехмерный образ структур или любых их частей Кроме того, специальные программы могут заменить одни фрагменты структуры на другие. При обработке в реальном масштабе времени или близком к нему, как хотят биохимики, эти вычисления будут требовать большого быстродействия ЭВМ.

1.2 Организация медико-статистических исследований

При проведении научных исследований в медицине исследователь сталкивается с социально-биологическими явлениями, имеющими случайный, вероятностный характер. Изучение закономерностей, присущих подобным явлениям, производится с помощью методов математической статистики и требует для этого наличие определённого, иногда весьма значительного, числа наблюдений. Организация сбора, количественная характеристика и статистический анализ медицинских наблюдений получили название «*медико-статистическое исследование*».

В содержании медико-статистического исследования выделяют четыре последовательных этапа:

- 1-й –составление плана и программы исследования;
- 2-й – статистическое наблюдение;
- 3-й – статистическую группировку и сводку наблюдений;
- 4-й – статистическую обработку и анализ полученных материалов, оформление результатов исследования.

В процессе реализации 1-го этапа медико-статистического исследования формулируется цель (а в ряде случаев и более частные, детализированные задачи) исследования, составляется его организационный план, кратко излагается содержание последующих этапов статистического исследования (программа исследования).

Хорошо разработанный план медико-статистического исследования, заблаговременное ознакомление его автора с требованиями, предъявляемыми к организации сбора и обработки статистических данных, — одно из главных условий успешности научного исследования.

Рассмотрим основные элементы планирования медико-статистического исследования на тему: «Лечение больных вазомоторным ринитом в амбулаторных условиях».

В качестве цели исследования изберем: выявление наиболее эффективного комплекса мероприятий для лечения больных вазомоторным ринитом в условиях поликлиники.

Далее в процессе планирования медико-статистического исследования разрабатываются программа статистического наблюдения (содержание которой рассматривается ниже при изложении вопросов 2-го этапа статистического исследования), программа статистической группировки и сводки материалов наблюдения (которая рассматривается далее применительно к содержанию 3-го этапа статистического исследования), перечень обобщающих статистических показателей и важнейшие направления анализа полученных данных (что включает в себя 4-й этап статистического исследования).

Познакомимся с важнейшими элементами содержания 2-го этапа *медико-статистического исследования* на основе разбираемого примера.

Объект наблюдения (т. е. совокупность единиц, о которых должны быть собраны статистические сведения, отграничиваемая по территориальному, административному, временному и другим признакам): больные вазомоторным ринитом, лечившиеся в поликлинике №. 8 в 2005 г.

Единица наблюдения (первичный элемент, из которых складывается объект исследования, подлежит чёткому определению в соответствии с целью и задачами исследования): больной вазомоторным ринитом, лечившийся в поликлинике № 8 у отоларинголога и закончивший лечение в 2005 г.

Программа наблюдения (перечень признаков, характеризующих единицу наблюдения с качественной и количественной сторон и подлежащих статистической регистрации): пол, возраст, профессия больного, диагноз, причина болезни характер лечения, срок проявления рецидивов, характеристика носового дыхания и т. п. При составлении программы выбираются только те признаки, которые нужны для ответа на поставленные темой и определенные целью работы вопросы. Практика увеличения числа признаков «на всякий случай» осложняет обработку собранных данных, удлиняет сроки работы, а иногда и скрывает ее смысл.

В соответствии с программой наблюдения разрабатывается учетно-статистический документ, с помощью которого осуществляется регистрация единиц наблюдения: карточка, журнал и т. д.

До утверждения программы наблюдения и ее носителя — учетно-статистического документа, для окончательной отработки всех ее элементов весьма целесообразно провести пробное наблюдение. Оно состоит в том, что на небольшом числе наблюдений проверяется возможность получения ответов на поставленные в учетном документе вопросы. Лишь после проведения пробного наблюдения и необходимой корректировки программы можно приступить к сбору материала, т. е. к *собственно статистическому наблюдению*.

Вид наблюдения по времени его проведения может быть *текущим* или *единовременным*. Большинство явлений медицинского характера наблюдаются непрерывно по мере их возникновения (случаи заболеваний, травм, осложнений, смерти; проведенные лечебно-профилактические мероприятия и т. д.) и потому подлежат текущему учету. К этому виду наблюдения относится и регистрация случаев обращения в поликлинику больных вазомоторным ринитом из нашего примера. В ряде случаев в клинико-статистических и других медицинских исследованиях проводится единовременное наблюдение, т. е. регистрация данных по состоянию на определенный момент времени (на пример, перепись больных).

По степени охвата единиц исследуемого объекта различают *сплошное* и *не сплошное* наблюдение. При сплошном наблюдении регистрируются все единицы составляющие объект наблюдения, как например, все больные вазомоторным ринитом, обратившиеся в поликлинику № 8 в 2005 г. При не сплошном,

частичном наблюдении берется только некоторая часть этих единиц, по которой затем судят о свойствах всего объекта наблюдения. Одним из видов не сплошного наблюдения является специально организуемое *выборочное* наблюдение, которое позволяет по части единиц, составляющих исследуемый объект, получить его наиболее достоверную характеристику.

Основными способами сбора медико-статистических материалов служат *непосредственное наблюдение* (например, регистрация врачами поликлиники больных вазомоторным ринитом), *опрос* и *отчетный способ*.

3-й этап медико-статистического исследования включает *группировку* и *сводку* собранных материалов наблюдения. При этом определяющее значение для раскрытия существа изучаемого явления имеет статистическая группировка, смысл которой заключается в расчленении статистической совокупности на однородные группы по важнейшим признакам. Выделение тех или иных групп не может быть произвольным делом исследователей, а вытекает из сущности и характера явления. Границы групп выбираются в зависимости от темы и цели исследования. Например, значения одного и того же признака — длительности пребывания на койке — могут быть по-разному разделены на группы при изучении лечения больных ангиной, язвенной болезнью желудка или туберкулезом легких.

Группировка может, осуществляться на основе *количественных* или *качественных* (атрибутивных) признаков. При группировке совокупности на равные части по количественному признаку (например, при построении вариационного ряда) важно определить наиболее целесообразное число групп **к**, которое зависит не только от цели исследования, но и от имеющегося числа наблюдений. Для этого можно воспользоваться следующей формулой:

$$\mathbf{к} = 1+3{,}32 \lg n,$$

где n — число наблюдений.

Например, если n = 100, то число групп = $1+3{,}32 \lg 100=7{,}64\approx8$.

Из разных видов группировок в медицине особое место занимает *типологическая* группировка, назначение которой заключается в делении изучаемой совокупности на однородные группы в соответствии с основными типами явления (например, распределение больных по классам, группам и формам болезней).

Результаты группировки статистического материала по отдельным признакам и их различным сочетаниям находят выражение в *статистических таблицах*. Статистическая таблица представляет собой рациональную форму систематизации статистических данных.

Каждая таблица должна иметь общий заголовок, в котором четко, в сжатой форме раскрывается её содержание. В таблице обязательно следует предусмотреть проверочные итоги (как общие, так и групповые) для облегчения необходимых расчетов по данным таблицы. Наименования единиц измерения приведенных в таблице данных указываются в заголовках (обычно в скобках). Нулевые значения признака лучше обозначать знаком тире, чем оставлять клетку

пустой. При отсутствии каких-либо данных необходимо в соответствующей клетке писай «нет сведений».

В медицинской статистике принято деление таблиц на *простые* и *сложные*. Сложные таблицы разделяются на *групповые* и *комбинационные*. Простая таблица содержит в качестве главной группировки перечень наблюдавшихся объектов и их общие числовые значения. В простых таблицах признак главной группировки может быть систематизирован по времени, типам учреждений, территории и т. п. Эти таблицы имеют справочно-информационный характер и могут быть хронологическими, административными или перечневыми. Образец хронологической таблицы представлен в табл. 1.1.

Таблица 1.1 - Число больных вазомоторным ринитом, лечившихся в 1970-1972 г.г.

Годы	Число больных
1970	293
1971	289
1972	317
Всего:	899

Групповая таблица обязательно содержит в главной группировке один группировочный признак (табл.1.2).

Таблица 1.2 - Распределение больных вазомоторным ринитом по возрасту (в абс. цифрах)

Возраст	Число больных
15-19 лет	7
20-29 «	23
30-39 «	47
40-49 «	150
50 лет и старше	90
Всего:	317

Наиболее сложной и ценной в аналитическом отношении является комбинационная таблица, которая в главной группировке (а иногда и в характеризующих признаках) содержит сочетание взаимосвязанных группировочных признаков (табл. 1.3).

По стадии исследования статистические таблицы подразделяют на *разработочные* (рабочие), содержащие только абсолютные данные, и окончательные, *аналитические*, в которых приводятся обобщающие показатели и результаты их статистической оценки. Аналитические таблицы могут быть получены только на последнем этапе медико-статистического исследования.

Содержанию 4-го этапа медико-статистического исследования посвящены последующие главы данного пособия. Важнейшие задачи статистической обра-

ботки и анализа материалов научного медицинского исследования и рекомендуемые для их реализации методы в сводном виде показаны в табл. 1.4.

Таблица 1.3 – Распределение больных по возрасту и полу

Диагноз	До 30 лет			30-49 лет			50 лет и старше			Итого		
	м.	ж.	всего	м.	ж.	всего	м.	ж.	всего	м.	ж.	всего
Гипертоническая болезнь	4	6	10	100	150	250	200	290	490	304	446	750
Кардиосклероз атеросклеротический	1	-	1	65	51	116	185	115	300	251	166	117
Ревматизм	50	80	130	40	65	105	20	30	50	110	175	285
Всего	55	86	141	205	266	471	405	435	840	665	787	1452

Таблица 1.4 – Статистические методы при решении задач научного медицинского исследования

Задачи	Статистические методы для иє следования количественно выраженных признаков	Статистические методы для исследования качественных, атрибутивных признаков, альтернативы
Определение характера распредєления	Метод χ^2, графики	
Определение обобщенных характеристик совокупности	Средние величины (арифметическая, геометрическая, гармоническая), мода, медиана. Графики	Показатели частоты, структуры, соотношения. Графики.
Оценка влияния структуры явления на размеры обобщенных характеристик		Методы стандартизации
Измерение вариабельности явлєний, признаков	Амплитуда вариационного ряда, среднее квадратическое отклонение, дисперсия, дисперсионный анализ -критерий **F,** критерий флюктуаций	Дисперсионный анализ - критерий **F,** критерий флюктуаций

Определение силы влияния различных факторов на вариабельность явлений, признаков	Дисперсионный анализ - сила влияния η^2	Дисперсионный анализ - сила влияния η^2
Оценка достоверности обобщенной характеристики (средней арифметической относительного показателя)	Средняя ошибка средней арифметической, графики	Средняя ошибка относительного показателя, графики
Оценка значимости различий двух совокупностей	а) попарно сопряженные совокупности: средняя ошибка разности средних, критерий t, критерий знаков, максимум-критерий, критерий Вилкоксона б) несопряженные совокупности средняяя ошибка разности средних, критерии t, Уайта, Вилкоксона; критерий χ^2, серийный критерий, критерий Колмогорова -Смирнова	Средняя ошибка разности показателей, критерий Шеллинга – Вольфеля, методы стандартизации, критерий χ^2
Оценка значимости различий 3 и более совокупностей	Дисперсионный анализ, критерий F	Дисперсионный анализ, критерий F, критерий χ^2
Определение связи между явлениями	Коэффициент корреляции, коэффициент регрессии, корреляционное отношение, коэффициент сопряженности, графики	Коэффициент корреляции, критерий Спирмена, критерий Кендела, коэффициент ассоциации, коэффициент сопряженности
Оценка динамики, тенденции	Метод наименьших квадратов, метод расчета скользящей средней, показатели динамического ряда, графики	Показатели динамического ряда, графики

1.3 Относительные величины

Полученные в результате группировки и сводки материалы исследования могут быть представлены в виде статистических таблиц или вариационных рядов, содержащих *абсолютные величины*. Как правило, этих абсолютных данных недостаточно для всестороннего анализа изучаемого явления, и потому исследователи прибегают к вычислению обобщающих статистических показателей: *относительных* и *средних величин*.

Относительные величины — отношения двух чисел, представляющих разные совокупности или части одной совокупности наблюдений. Одна из величин, составляющих отношение, называется основанием (базой) и обычно приравнивается к какому-либо «круглому» числу (100, 1000, 10000 и т. д.) или к единице.

В медицинской статистике применяются относительные величины *распределения, частоты, наглядности, соотношения* и некоторые другие. Для целей научного анализа в медицинских исследованиях чаще всего используются относительные величины *распределения* и *частоты*, которые наиболее полно отражают присущие изучаемым явлениям статистические закономерности.

Относительные величины распределения, или *экстенсивности*, выражают отношение части к целому или распределение целого на его составные части. При вычислении этих показателей за основание принимают величину целого, обычно приравнивая его к 100%, а затем путем отнесения отдельных частей к этому целому определяют их доли в процентах (иногда называемые также *удельным весом*). При полном распределении изучаемого явления на его составные части сумма относительных долей должна составлять ровно 100%.

Пример: Имеются данные о степени активности ревматизма у 2 групп детей. В первой группе 140 детей, больных ревматизмом с поражением сердца и суставов. Ко второй группе отнесены 226 детей, больных ревматизмом с выраженными симптомами поражения нервной системы. Необходимо вычислить (в %) относительные доли детей с разной степенью активности ревматизма в каждой группе и в целом. Принимаем общее число детей в каждой группе и в общем итоге за 100% (табл. 1.5.).

Таблица 1.5 – Распределение детей с ревматизмом по степени активности процесса

Степень активности процесса	1-я группа детей		2-я группа детей		Всего	
	абс. числа	доли в %	абс. числа	доли в %	абс. числа	доли в %
I	15	10,7	145	64,1	160	43,7
II	76	54,3	71	31,4	147	40,1
III	49	35,0	10	4,5	59	16,2
Всего	140	100,0	226	100,0	366	100,0

Полученные относительные величины распределения позволяют сделать вывод, что у больных детей 2- й группы степень активности ревматического процесса ниже, чем у детей 1- й группы: больные с III степенью активности в 1-й группе составляли 35%, а во 2- й — только 4,5%.

Относительными величинами частоты или интенсивности называются относительные числа, показывающие, как часто то или иное явление имеет место в среде, в которой оно происходит за определенный отрезок времени. В зависимости от интенсивности изучаемого явления при вычислении относительных величин частоты размеры основания (среды) приравниваются к 100, 1000, 10000 и т. д. Если вычисление ведётся из расчета на 100, его результат выражают в %, если на 1000 — в ‰ (промилле), если на 10000 — в ‱; (продецимилле) и т. д. Чем реже встречается изучаемое явление, тем большее избирается основание. Так, частота редко встречающихся заболеваний обычно определяется из расчета на 100000 населения.

Пример. В районе проживает 80000 жителей, из них 32 000 мужчин и 48000 женщин. В течение года в поликлинику обратилось 15000 человек по поводу пояснично-крестцового радикулита, из них 7000 мужчин и 8000 женщин. Вычислим относительные величины, характеризующие частоту заболеваемости мужчин и женщин, для чего определим, сколько заболеваний приходилось на 1000 мужчин и 1000 женщин:

$$\text{Частота заболеваемости мужчин} - \frac{7000 * 1000}{32000} = 218{,}7\text{‰}$$

$$\text{Частота заболеваемости женщин} - \frac{8000 * 1000}{48000} = 166{,}7\text{‰}$$

Таким образом, при расчёте относительных величин частоты сопоставляются две различные, но связанные между собой совокупности, например, население и случаи заболевания, в то время как при вычислении относительных величин распределения берётся одна совокупность и определяется, какую долю целого составляет каждая из её частей. Если бы в нашем примере требовалось рассчитать относительные величины распределения больных по полу, то за 100 % следовало принять общее число больных (15000 чел), а затем определить, какие доли (в %) этого числа составляли мужчины и женщины:

$$\frac{7000 \cdot 100}{15\,000} = 46{,}7\%; \quad \frac{8000 \cdot 100}{15\,000} = 53{,}3\%.$$

Сумма полученных показателей равна 100%.

Следует помнить, что, как правило, по относительным величинам распределения нельзя судить о частоте явления.

Пример. Имеются сведения о детях, больных хореей разной степени тяжести и наличии у них изменений со стороны сердца (табл. 1.6).

Из таблицы следует, что у детей с хореей средней тяжести изменения со стороны сердца наблюдаются чаще. Если же вместо необходимых в данном случае показателей частоты воспользоваться для характеристики распространенности поражений сердца относительными величинами распределения, то

может быть сделан неверный вывод о преобладании поражений сердца у больных легкой формой хореи (52,7% против 47,3 %). Между тем большая доля больных легкой формой хореи среди больных с поражениями сердца объясняется не повышенной частотой изменений со стороны сердца у этих больных, а преобладанием легкой формы болезни среди всех больных хореей.

Таблица 1.6 -Час тота поражений сердца у детей, больных хореей

Форма хореи	Число больных	Из них поражением сердца	Относительные величины	
			Частота поражения сердца (%)	Распределение больных с поражением сердца по форме хореи (%)
Легкая	113	39	34,5	52,7
Средней тяжести	77	35	45,4	47,3
Всего	190	74	38,9	100,0

Необходимо указать, что в рассмотренном примере наиболее ценными в аналитическом отношении показателями частоты поражений сердца у больных хореей явились показатели, вычисленные в группах больных с разными формами болезни, позволившие выяснить, что эти поражения чаще встречались у больных хореей средней тяжести. Таким образом, предварительная группировка изучаемой совокупности в соответствии с целью исследования и вычисление дифференцированных относительных величин частоты дают возможность более глубоко проанализировать изучаемые медицинские явления.

Обычно многие относительные величины интенсивности, например заболеваемости, рассчитываются за год. Однако иногда возникает необходимость вычислить эти показатели по данным за более короткий промежуток времени (полугодие, квартал, месяц) Таким образом, чтобы по ним можно было судить о частоте изучаемого явления на протяжении целого года. При этом предполагается, что интенсивность явления в течение всего года останется на уровне, присущем изученному отрезку времени. Определение такого показателя «из расчета на год» производится по формуле:

$$\mathbf{K = n\ 1000\ 12/N\ k} \qquad (1,1)$$

где, **K –**показатель интенсивности за несколько месяцев «из расчета за год»;

n - абсолютное число единиц изучаемого явления (например, заболеваний) за взятое число месяцев;

N - численность среды;

K – число взятых месяцев.

Так если в районе, где проживает 80000 население за 5 месяцев зарегистрировано 4700 заболеваний, то общий показатель заболеваемости за пять месяцев« из расчета на год» составит:

$$\frac{4700 \cdot 1000 \cdot 12}{80\,000 \cdot 5} = 181‰.$$

При сравнении общих показателей частоты какого-либо явления их различия определяются не только разными уровнями распространения этого явления в сравниваемых группах наблюдений, но и неоднородностью состава этих групп. Например различия уровня летальности в лечебных учреждениях определяется не только качеством медицинской помощи, но и составом лечившихся в них больных. Для того, чтобы устранить влияние неоднородности состава сравниваемых групп наблюдений на общие показатели частоты изучаемого явления, прибегают к вычислению так называемых *стандартизованных коэффициентов (показателей)*.

Существует три основных способа расчета стандартизованных показателей: *прямой, косвенный и обратный*.

Рассмотрим методику *прямого метода стандартизации* на примере показателей выявляемости туберкулеза среди обследованных групп населения города Н. в 1960 и 1965 гг. (табл . 1.7).

Таблица 1.7 - Частота выявления туберкулеза среди жителей города Н в 1960 и 1965 гг

Контингенты осмотренных	1960			1965		
	Осмотрено чел.	Выявлено больных		Осмотрено чел	Выявлено больных	
		Абс.число	На 1000 осмотрен.		Абс.число	На 1000 осмотрен.
Взрослые	55670	123	2,2	104186	190	1,8
Подростки	10058	8	0,8	12235	7	0,6
Дети	31550	9	0,3	18969	3	0,16
Всего	97278	140	1,4	135390	200	1,5

Из таблицы видно, что в 1960 г. в городе при профилактическом осмотре 97 278 человек было выявлено 140 больных туберкулезом, или 1,4 на1000 осмотренных, а в 1965г. этот показатель увеличился до 1,5‰. Можно ли считать, что выявляемость туберкулеза выросла? Ответ может быть положительным если говорить об общем показателе выявляемости. Однако из таблицы следует, что в каждой группе осмотренных показатель выявляемости туберкулеза в 1965 г. по сравнению с 1960 г. снизился. Можно предположить, что рост общего показателя выявляемости туберкулеза среди осмотренных в 1965 г. обусловлен изменением их возрастного состава. Действительно, среди осмотренных в 1960 г, доля взрослых составляла 57,2%, а в 1965 г.— 76,9%. Между тем показатель выявляемости туберкулеза у взрослых в несколько раз выше, чем у подростков

и детей В результате абсолютное число выявленных, а вместе с ним и общий показатель выявляемости увеличились. Закономерно возникает вопрос: чему бы равнялись общие показатели выявляемости в сравниваемые годы, если бы состав осмотренного населения был одинаков?

Допустим, что возрастной состав осмотренных в 1965 и 1960 гг был одинаков и равнялся их среднему составу за эти годы. Примем этот средний состав осмотренных за постоянный стандарт для сравнения. Удобнее представить стандарт в относительных величинах приняв общее число обследованных за 1000. В дальнейшем, исходя из избранного стандарта и фактических показателей выявляемости туберкулеза в каждой возрастной группе, вычисляем по гѳ дам так называемые «ожидаемые» числа выявленных больных (табл. 1.8).

Таблица 1.8 - Вычисление стандартизованных показателей выявляемости туберкулеза среди городского населения в 1960 и 1965 гг.

Контингенты осмотренных	Стандарт состава обеих групп		Расчет" ожидаемого" числа выявленных больных, исходя из стандартного состава осмотренных и фактических показателей заболеваемости	
	Абс.число	0/00	1960 г.	1965 г.
Взрослые	79928	687,1	687,1*2,2/1000=1,51	687,1*1,8/1000=1,23
Подростки	11146	95,8	95,8*0,8/1000=0,08	95,8*0,6/1000=0,06
Дети	25259	217,1	217,1*0,3/1000=0,07	217,1*0,16/1000=0,03
Всего	116333	1000,0	1,66	1,32

Стандартизованные общие показатели выявления туберкулёза для каждого года получаются в результате суммирования «ожидаемых» чисел, вычисленных для каждой группы: 1,51 +0,08+0,07 ≈ 1,7‰ и 1,23+0,06+0,03 ≈ 1,3‰. На основании сравнения стандартизованных показателей можно сделать вывод, что при одинаковом возрастном составе осмотренных в 1960 и 1965 гг. общий показатель выявляемости больных туберкулезом в 1965 г. был бы ниже показателя 1960 г. Следовательно, истинная причина повышения общего показателя выявляемости туберкулеза среди осмотренных в 1965 г. заключается в изменении состава осмотренных в сторону увеличения контингента взрослых, что и помог выяснить метод стандартизации.

Стандартизованные показатели служат *только для целей сравнения* и не могут использоваться для измерения действительной частоты анализируемого признака.

В тех случаях, когда частные (например, повозрастные) показатели частоты в одной из сравниваемых групп не известны или образующие их числа малы, прибегают к *косвенному методу стандартизации.*

Пример. В одной клинике апробировался новый метод лечения больных с инсультом. Критерием эффективности являлся показатель летальности, кото-

рый сравнивался с показателем летальности больных, лечившихся старым методом (табл. 1.9).

Таблица 1.9 -Показатели летальности больных с инсультами, лечившихся двумя разными методами

Возраст больных	Число больных, лечившихся старым методом	Из них умерло	Показатель летальности %	Число больных, лечившихся новым методов	Из них умерло	Показатель летальности %
До 20 лет	30	?	?	-	-	-
20-29 лет	85	?	?	65	2	3,8
30-39 лет	90	?	?	55	3	5,4
40-49 лет	103	?	?	120	19	15,8
50 лет и старше	192	?	?	260	40	15,4
Всего	500	61	12,2	500	64	12,8

При сравнении общих показателей летальности можно сделать вывод о неэффективности нового метода лечения, поскольку его применение не привело к снижению показателя летальности, более того, этот показатель оказался выше, чем при лечении старым методом. Однако такой вывод является преждевременным, поскольку состав больных в обеих группах был различным по возрасту. Для устранения (элимирования) влияния на общие показатели летальности неоднородности возрастного состава больных в сравниваемых группах необходимо стандартизовать показатели летальности. В данном случае следует применить косвенный метод стандартизации, так как неизвестны повозрастные показатели летальности в первой группе больных. За стандарт берутся известные показатели летальности (табл. 1.10).

Таблица 1.10 - Вычисление ожидаемых чисел умерших и стандартизованного показателя летальности больных в 1-й группе

Возраст больных	Стандарт - показатель летальности %	“Ожидаемое”число умерших	Вычисление стандартизованного показателя летальности
До 20 лет	-	-	(Действительное число умерших) / (“ Ожидаемое” число умерших)*(Общий показатель летальности стандарта = (61/53,9)*(12,8) = 14,1%
20-29 лет	3,8	3,8*85/100=3,2	
30-39 лет	5,4	5,4*90/100=4,8	
40-49 лет	15,8	15,8*103/100=16,3	
50 лет и старше	15,4	15,4*192/100=29,6	
Всего	12,8	53,9	

Таким образом, элиминировав различия в составе больных по возрасту, мы получили более высокий общий показатель летальности в первой группе (старый метод лечения), что свидетельствует об эффективности лечения больных по новому методу.

Если не известен состав среды (например, населения), в которой происходит изучаемое явление (например, случаи заболевания), то для вычисления стандартизированных показателей применяется *обратный метод Кэрриджа*. Рассмотрим эту методику на примере (табл. 1.11).

Таблица 1.10 – Расчеѓ« ожидаемого» числа осмотренных при выявлении туберкулеза среди разных групп населения в 1965 г.

Контингенты осмотренных	1960 г.			1965 г.			«Ожидаемое» число осмотренных
	Осмотрено чел.	Выявлено больных		Осмотрено чел.	Выявлено больных		
		Абс. число	На 1000 осмотренныхх		Абс. число	На 1000 осмотренных	
Взрослые	55670	123	2,2	?	190	?	X_1=190*1000/2,2=86304
Подростки	10058	8	0,8	?	7	?	X_2=7*1000/0,8=8750
Дети	31550	9	0,3	?	3	?	X_3=3*1000/0,3=10000
Всего	92278	140	1,4 стандарт	135390	200	1,5	105114

В данном примере не известен состав населения в 1965 г., поэтому расчитываем «ожидаемое» число осмотренных, а не больных, как при косвенном методе. За стандарт следует взять показатели выявляемости больных туберкулезом в 1960 г. Для определения числа осмотренных в каждой возрастной группе решаем арифметические пропорции, рассуждая следующим образом: если выявляемость туберкулеза оставалась в 1965 г. такой же, как и в 1960 г., то какое же число населения данной группы необходимо было бы осмотреть, чтобы получить выявленное число больных туберкулезом? Результаты расчетов приведены в табл.1.10. Таким образом, если бы показатели выявляемости остались на на уровне 1960 г., то для выявления фактического числа больных туберкулезом в 1965 г. потребовалось бы осмотреть меньше населения, чем в действительности было осмотрено. Следовательно, увеличение показателя выявляемости в 1965 г. связано с иным (по сравнению с 1960 г.) возрастным составом осмот-

ренных. Стандартизированный показатель выявляемости туберкулеза в 1965 г. вычисляется путем отношения «ожидаемого» числа осмотренных к фактическому их числу и умножением этого отношения на общий показатель стандарта:

$$\eta_{ст} = 105114/135390 \cdot 1,4 = 1,1‰$$

Полученный показатель заметно ниже общего показателя выявляемости туберкулеза в 1960 г. (1,4 ‰). Обратный метод расчета позволил устранить влияние различий возрастного состава осмотренных на общий показатель выявляемости туберкулеза в 1965 г. при отсутствии данных об этом составе.

1.4 Статистическая обработка вариационного ряда

Изучение медицинских явлений, поиск присущих им закономерностей, как правило, связано с повторением (подчас многократным) однородных наблюдений или опытов. При этом исследователя интересуют не отдельные наблюдения, а их обобщенные характеристики, помогающие понять типичные черты изучаемых явлений. Анализируя результаты нескольких серий наблюдений или опытов, исследователь обнаруживает различия в частоте интересующих его признаков, если эти признаки качественные, либо в величине признаков, если их можно оценить количественно.

Во всех случаях обнаружения разброса значений признака исследователю необходимо выяснить, насколько существенен этот разброс, случаен он или нет и каковы факторы, его определяющие. Для решения этих задач необходимо составить *вариационный ряд* и вычислить его *обобщенные характеристики*.

1.4.1 Основные понятия и определения

Всякое множество отдельных объектов, отличающихся друг от друга и в тоже время сходных в некоторых существенных отношениях, составляют так называемую *совокупность*. Например, дети, родившиеся в стране в течение какого-либо периода времени, молекулы вещества в определенном объеме и т.д.

В состав совокупности входят различные *члены* или *единицы совокупности*. Общее число единиц совокупности называется *объемом совокупности*. Каждая единица совокупности характеризуется определенными *признаками*. Например, родившиеся дети – весом, ростом и т.д., молекулы вещества- размером, скоростями хаотичного движения ит.д. Каждый признак принимает различные значения у разных единиц совокупности. Различия в значениях признака между отдельными единицами совокупности называется *вариацией* или *дисперсией*. Понятие «признак варьирует» означает то, что признак принимает различные значения у разных единиц совокупности. Например рост или вес у детей родившихся в стране в течение какого-либо периода времени, размер молекул вещества в определенном объеме и т.д.

Значение признака для той или иной единицы совокупности называется *вариантой* и обозначается $\mathbf{X_i}$ (x_1, x_2, x_3,…x_i…x_n). *Варианта* – это конкретное значение случайной переменной $\mathbf{X_i}$, т.е. величины, изменяющиеся под влиянием многих случайных причин.

Совокупность может состоять из других совокупностей, более частных. Например, совокупность детей, родившихся в стране, можно представить в виде совокупностей по отдельным местностям (область, район, город и т.д.).

Наиболее общую совокупность называют *генеральной*. Это теоретически бесконечно большая совокупность всех единиц. Которые к ней могут быть отнесены.

Совокупность, состоящую из небольшого количества единиц называют *выборочной*. Исследователь, как правило, имеет дело с выборочными совокупностями.

Совокупностью является также объем любых наблюдений или измерений отдельных признаков (вес или рост детей, размер молекул и т.д.). Каждое отдельное наблюдение, при котором устанавливается значение случайной переменной, является единицей совокупности.

Различают вариацию *качественную* и *количественную*. При *качественной* вариации различия между вариантами выражаются каким-либо качеством. В этом случае каждая варианта должна получить качественную характеристику в соответствии с заранее принятыми обозначениями. Например, цвет волос или глаз у родившихся детей При *количественной* вариации сами варианты и различия между ними принимают числовые значения. При этом количественная вариация может *дискретной* и *непрерывной*. При *дискретной* вариации различия между вариантами выражаются целыми числами, между которыми нет и не может быть переходов. Например количество родившихся детей(1, 2, 3, и т. д.). При *непрерывной* вариации значения вариант не обязательно выражаются только целыми числами. Все зависит от степени точности, которая принимается для характеристики данного количественного признака (вес или рост младенца размер вируса или молекулы и т. д.). То есть, между вариантами возможны все переходы. При изучении непрерывной вариации необходимо все единицы совокупности характеризовать с той степенью точности, которая заранее намечена и больше всего подходит в данном конкретном случае.

1.4.2 Методика составления вариационного ряда

Если число наблюдений (**n**) небольшое, то варианты достаточно просто ранжировать, т. е. расположить в порядке возрастания их значений. Например, при измерении размеров вируса орнитоза получены следующие величины (в *мкм*): 0,34; 0,45; 0,20; 0.29; 0,40. Эти варианты нужно записать в такой последовательности: 0,20; 0,29; 0,34; 0,40; 0,45.

При увеличении числа наблюдений обычно отмечаются повторения отдельных вариант. В этом случае для построения вариационного ряда необходимо выписать все значения вариант в порядке возрастания, а затем подсчитать число повторений (частоту – ***f***) каждой варианты и записать их рядом с соответствующими значениями вариант. Например, исследователем произведено 47 измерений мембранного потенциала мышечной клетки в покое (с точностью до 1 мВ). Составленный вариационный ряд показан в табл.1.11.

Таким образом, главными составными элементами вариационного ряда являются

x – варианты -значения варьирующего признака;

f – частоты - число повторений каждой варианты;

n – общее число наблюдений (**n** равно сумме частот, т. е. $n = \sum f$).

Последовательное суммирование частот образует так называемые *накопленные частоты*. Последняя накопленная частота представляет собой общее число наблюдений. Подобным же образом составляется и интервальный вариационный ряд, в котором перечисляются не отдельные варианты, а их группы.

Таблица 1.11 – Результаты измерения потенциала мышечной клетки

Варианта x	Частота f	**Накопленные частоты**
33	1	1
34	2	3
35	4	7
36	5	12
37	8	20
38	10	30
39	7	37
40	6	43
41	3	46
42	1	47
	n=47	

Интервальный вариационный ряд следует составлять в тех случаях, когда исследователь имеет дело с большим разнообразием значений вариант (более 20). Интервалы в таком вариационном ряду целесообразно иметь одинаковыми, т. е. они должны объединять равное число значений вариант. Интервальные вариационные ряды строятся при изучении как дискретных величин (признаков, выражаемых только целым числом, например число посещений, операций, число эритроцитов, частота пульса и т. д.) так и при исследовании непрерывных величин (признаков, регистрируемых дробными числами, например, рост, вес, биохимические показатели т. п.).

Для графического изображения вариационного ряда применяют *полигоны* и *гистограммы* (рис. 1.1.). Полигоны используют для изображения рядов дискретных величин, а гистограммы — непрерывных. При построении полигона на оси абсцисс откладывают значения вариант или их групп, на оси ординат— частоты. Полученные точки соединяют прямыми линиями. При построении гистограммы на оси абсцисс восстанавливают столбики, по высоте соответствующие частотам взятых интервалов, а вся гистограмма приобретает вид суммы прямоугольников.

Графическое изображение вариационного ряда дает ориентировочное представление о законе, которому подчиняется повторяемость вариант, так называемом законе распределения.

Знание закона распределения варьирующих признаков или достаточно достоверное предположение о нем дают возможность исследователю выбрать наиболее правильный и эффективный метод для статистической характеристики имеющихся наблюдений. Если исследуются непрерывные случайные величины и ряд на графике выглядит одновершинной симметричной кривой, то можно предположить, что изучаемые величины подчиняются нормальному закону распределения (см. рис. 1.1.).

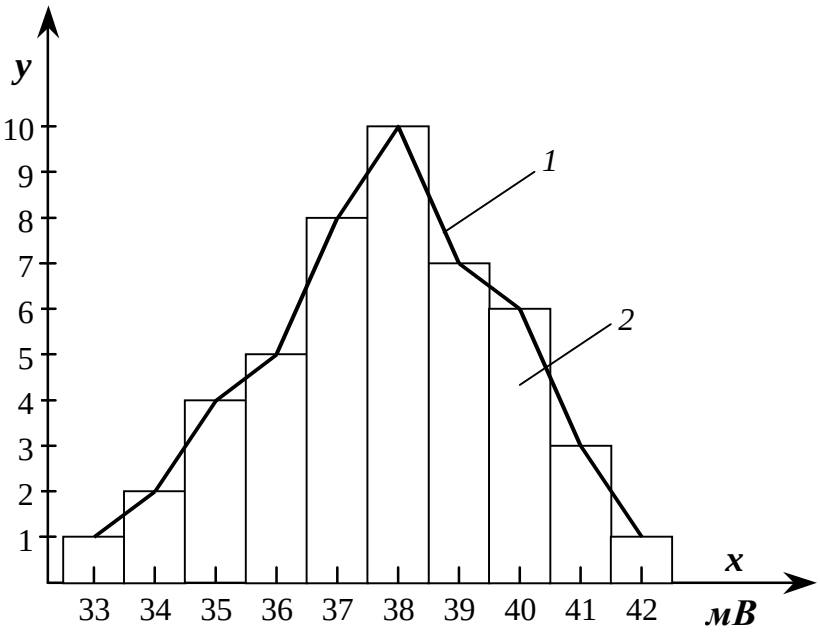

Рис. 1.1. Полигон (1) и гистограмма (2) распределения

1.4.3 Методика статистической обработки вариационного ряда при нормальном законе распределения вариант

Сводными характеристиками значений вариант служат *средняя арифметическая величина, мода, медиана и квартили*. Каждая из этих характеристик своеобразна. Они не могут подменить друг друга и лишь в совокупности достаточно полно и в сжатой форме представляют особенности вариационного ряда.

Наиболее общей характеристикой всех значений вариант является *средняя арифметическая величина*. Различают *среднюю арифметическую простую* и *взвешенную*. Средняя арифметическая простая вычисляется по формуле:

$$\bar{x} = \frac{\sum x}{n} \qquad (1.2)$$

В вариационных рядах, где отдельные варианты встречаются с разной частотой (т.е. имеют разный вес) определяется средняя арифметическая взвешенная по формуле:

$$\bar{x} = \frac{\sum(x \cdot f)}{n} \qquad (1.3)$$

Как видно из формулы, на величине средней арифметической сказывается влияние всех вариант входящих в вариационный ряд, причем это влияние прямо пропорционально числу повторений вариант. Взвешенную среднюю арифметическую величину необходимо вычислять во всех случаях, когда частоты не одинаковы.

В интервальных вариационных рядах при определении средней арифметической величины прежде всего следует определить середины интервалов. Середину интервала при изучении непрерывных величин можно определить как среднюю арифметическую начальных значений двух соседних интервалов. В дискретных рядах середина интервала вычисляется как среднее арифметическое начального и конечного значений данного интервала. Затем значения середин интервалов используют при дальнейших расчетах в качестве вариант ***x***.

Средняя арифметическая величина обладает следующими свойствами:

1) сумма отклонений от средней равна нулю;

2) при умножении (делении) всех вариант на один и тот же множитель (делитель) средняя арифметическая умножается (делится) на тот же множитель (делитель);

3) если прибавить (вычесть) ко всем вариантам одно и то же число, средняя увеличится (уменьшится) на то же число

Эти свойства могут быть использованы для облегчения и упрощения расчета средней арифметической величины.

Первое свойство, например, служит обоснованием расчета средней арифметической по способу моментов:

$$\bar{x} = A + \frac{\sum(x - A) \cdot f}{n} \qquad (1.4)$$

где:

x – середины интервалов вариационного ряда;

A – условная средняя арифметическая, за которую принимают значение середины интервала, имеющего наибольшую частоту.

Особенно удобно способ моментов использовать при вычислении средней арифметической в интервальном вариационном ряду. Для этого необходимо сначала определить середины интервалов. Величину одной из середин интервала следует принять за условную среднюю (А), после чего найти отклонения всех других середин интервалов от этой величины $x—A$. Полученные разности затем необходимо умножить на соответствующие частоты, произведения сум-

мировать и подставить найденную величину $\sum(x - A)\cdot f$ в формулу для вычисления (1.4).

Второе свойство средней арифметической полезно применить при анализе вариационного ряда, состоящего либо из очень больших, либо из очень малых величин. Имеются, например, варианты: 0,0001; 0,0002; 0,0003. Используя это свойство, увеличим их в 10000 раз, получим величины 1, 2, 3. Средняя арифметическая из них равна 2, а искомая средняя арифметическая в 10000 меньше, т. е. 0,0002.

Модой (Мо) называют значение наиболее часто встречающейся варианты. В примере в табл.1.11 это варианта 38 *мВ* В интервальном вариационном ряду мода находится как середина того интервала, которому соответствует наибольшая частота.

Более точно мода определяется по формуле:

$$M_0 = X_{M_0} + \Delta \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})} \qquad (1.5)$$

где:

X_{M_0} – начальное значение интервала, содержащего моду;

Δ – ширина интервала;

f_{M_0} – частота вариант в интервале, содержащем моду;

f_{M_0-1} и f_{M_0+1} – частоты вариант в соседних интервалах.

Как указывалось выше, кривая нормального распределения симметричная и одновершинная. Следовательно, в таком вариационном ряду может быть только одна мода. Если при анализе явления, которое предположительно подчиняется закону нормального распределения, получена, например, несимметричная, двухвершинная (бимодальная) кривая, то следует еще раз проанализировать состав исследуемой группы и, исключив искажающие наблюдения, сделать группу однородной.

Медиана (Ме) — значение варианты, делящей вариационный ряд пополам (с каждой стороны от медианы находится половина вариант).

Квартили (верхний – $Q_в$ и нижний – $Q_н$) — значения вариант, делящих вариационный ряд (вместе с *Ме*) на 4 части. Между $Q_н$ и $Q_в$ находится половина всех вариант. Порядковый номер варианты, являющейся медианой или квартилем, определяется по формулам:

$$Q_н: (n+1) / 4; \quad Me: (n+1) / 2; \quad Q_в: 3\cdot(n+1) / 4; \qquad (1.6)$$

В случае получения дробного значения порядкового номера его округляют до ближайшего целого числа.

Более точный расчет медианы в интервальном вариационном ряду следует производить по формуле:

$$Me = X_{Me} + \Delta \frac{n/2 - S_{Me-1}}{f_{Me}}, \qquad (1.7)$$

где X_{Me} – начальное значение интервала, содержащего медиану;

Δ – ширина интервала;

S_{Me-1} – накопленная частота до интервала, содержащего медиану;

f_{Me} – частота вариант в интервале, содержащем медиану.

Размеры *Мо* и *Ме* не зависят от значений крайних вариант. В симметричном вариационном ряду они равны между собой и совпадают со значением средней арифметической. Мода особенно важна для характеристики несимметричного ряда. Медианой и квартилями обязательно нужно пользоваться при обработке ряда с открытыми крайними интервалами.

После определения обобщенных характеристик вариационного ряда следует установить его *колеблемость*, т.е. размеры варьирования значений изучаемого признака. Приближенно о колеблемости можно судить по *амплитуде* (размаху) вариационного ряда - разности *максимальной* и *минимальной* вариант. Более точно колеблемость ряда характеризует *среднее квадратическое отклонение* (σ), вычисляемое по формуле:

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2 \cdot f}{n}} \qquad (1.8)$$

Квадрат среднего квадратического отклонения (σ^2) называется *дисперсией*.

Небольшая величина среднего квадратического отклонения свидетельствует об однородности исследуемой группы наблюдений. Среднюю арифметическую в таком случае следует признать вполне характерной, типичной для данного вариационного ряда. Однако слишком малая величина σ заставляет думать об искусственном подборе наблюдений. При очень большой σ средняя арифметическая в меньшей степени характеризует весь вариационный ряд, что говорит о значительной вариабельности явления или неоднородности исследуемой группы.

Оценка степени рассеяния вариант около средней может быть произведена с помощью *коэффициента вариации*, вычисляемого по формуле:

$$c = \frac{\sigma}{\overline{x}} \cdot 100\% \qquad (1.9)$$

Значения коэффициента вариации менее10% свидетельствуют о малом рассеянии, от 10 до 20% – о среднем и более 20% – о сильном рассеянии вариант вокруг средней арифметической.

Согласно теории вероятностей в явлениях, подчиняющихся нормальному закону распределения, между значениями средней арифметической, среднего квадратического отклонения и вариантами существует строгая зависимость. Например, 68,3% значений варьирующего признака находятся в пределах

$\bar{x} \pm 1\sigma$; 95,5%—в пределах $\bar{x} \pm 2\sigma$ и 99,7%—в пределах $\bar{x} \pm 3\sigma$. Эти соотношения показаны на рис. 1.2. Указанные взаимоотношения средней арифметической, среднего квадратического отклонения и отдельных вариант иногда называют *правилом трех сигм*. С помощью этого правила, зная $\bar{x}$ и σ (и предполагая нормальным изучаемое распределение), можно получить представление о вероятных размерах варьирующего признака.

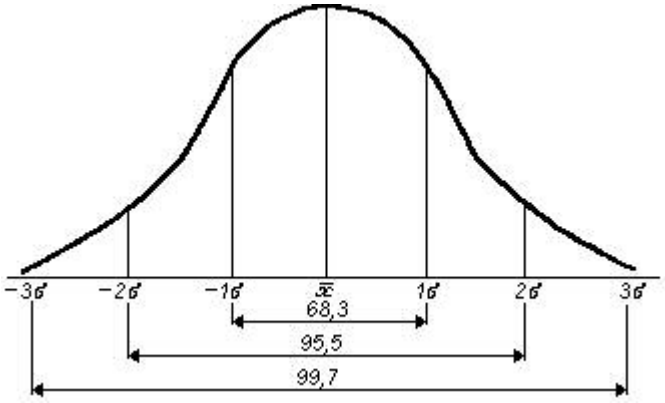

Рис. 1.2. Кривая нормального распределения

Правило трех сигм можно использовать при решении ряда практических задач:

1. Знание значений $\bar{x}$ и σ дает исследователю возможность определить границы средних (нормальных) значений признака. Нормальными обычно рекомендуется считать значения в пределах $\bar{x} \pm 1\sigma$. Иногда пределы нормы определяют с использованием $0{,}5\sigma$; $1{,}34\sigma$ и т.п. Решать этот вопрос должен специалист, знающий существо исследуемого явления.

2. Нормированное отклонение $t = \frac{x - \bar{x}}{\sigma}$., позволяет также решить, относится ли данное наблюдение к интересующей нас совокупности. Ответ будет положительным всегда, когда $t < 3$.

Пусть нам известно, что средняя арифметическая $\bar{x}$ пульса у больных абсцессом мозга равна 50, а σ=±6,5 удара. Требуется определить, может ли относиться больной с частотой пульса x=65 к данной группе больных? $t = \frac{x - \bar{x}}{\sigma} = \frac{65 - 50}{6{,}5} = 2{,}3$ Пульс 65 находится в пределах $2{,}3\sigma$ от $\bar{x}$ = 50, и, следовательно, в рассматриваемом случае ответ должен быть положительным.

3. В некоторых случаях возникает необходимость исключить из наблюдений варианту, почему-то резко отличающуюся от всех остальных («выскакивающую» варианту — $x_в$). Это стремление продиктовано нежеланием получить искаженное представление о средней арифметической. Право исключить эту варианту возникает тогда, когда $t = \frac{x_в - \bar{x}}{\sigma} > 3$, причем $\bar{x}$ и σ рассчитываются без выскакивающей варианты $x_в$.

Описывая, например, небольшую группу больных скарлатиной, врач считал нужным указать средний возраст. Данные больных о возрасте следующие: 1, 3, 3, 5, 7, 11, 12, 32. Средняя арифметическая равна 9 годам. На ее размере, несомненно, сказалось влияние максимальной варианты —32 года Определение средней арифметической без этой варианты дало новую величину

$\bar{x} = \frac{1+3+3+5+7+11+12}{7} = 6$; среднее квадратическое отклонение для нее ±4 года. Подставляем найденные величины в вышеуказанное соотношение:

$t = \frac{32-6}{4} = 6{,}5$, следовательно, варианта 32 года выходит за пределы 3σ. В характеристике среднего возраста больных ею можно пренебречь. При оценке «выскакивающей» варианты можно пользоваться и специальными таблицами.

4. Правило трех сигм используется также для построения теоретического ряда, отвечающего нормальному распределению. Такой ряд, сопоставленный с фактическим, может служить критерием нормальности распределения фактических данных. Построение теоретического ряда, отвечающего нормальному распределению при заданных параметрах x, *n* и σ производится следующим образом:

а) в таблицу вписываются полученные в опыте значения вариант, середины интервалов и соответствующие им частоты;

б) определяются нормированные отклонения каждой середины интервала;

в) по величине нормированного отклонения находятся значения функции нормированного отклонения *f(x)* по справочным таблицам;

г) по формуле $\frac{n \cdot \Delta x}{\sigma} \cdot f(x)$ определить теоретические частоты для каждой середины интервала.

Схема необходимых расчетов приведена в табл. 1.12.

Таблица 1.12 – Расчет теоретических частот, отвечающих нормальному распределению

Содержание цинка в сыворотке крови в физиологических условиях (мкг %)	Частота, *f*	Середина интервала, *x*	Нормированное отклонение значений середины интервала, $t=\left\|\frac{x-\bar{x}}{\sigma}\right\|$	Функция нормированного отклонения, *f(x)*	Теоретическая частота, f_1 $f_1=\frac{n \cdot \Delta x}{\sigma} \cdot f(x)$
75-84	2	80	$\frac{90-120}{15{,}3} = 1{,}96$	0,0132	1,34
85-94	5	90	$\frac{80-120}{15{,}3} = 2{,}61$	0,0584	0,46
95-104	7	100	$\frac{100-120}{15{,}3} = 1{,}31$	0,1691	1,22

105-114	15	110	$\frac{110-120}{15,3}=0,65$	0,3230	1,38
115-124	35	120	$\frac{120-120}{15,3}=0,00$	0,3989	4,00
125-134	17	130	$\frac{130-120}{15,3}=0,65$	0,3230	0,54
135-144	11	140	$\frac{140-120}{15,3}=1,31$	0,1691	0,01
145-154	3	150	$\frac{150-120}{15,3}=1,96$	0,0584	0,13
155-164	1	160	$\frac{160-120}{15,3}=2,61$	0,0132	0,01
Всего	N=96; Δx=10 $\bar{x}$ =120				$\sum f_1$ =96

1.4.4 Расчет статистических характеристик при малом числе наблюдений

Характеристики вариационного ряда *(x, Mo, Me*, квартили), как указывалось, тем точнее отображают явление, чем больше сделано наблюдений. То же относится и к амплитуде вариационного ряда, к среднему квадратическому отклонению. Указанная выше формула (1.8) для расчета среднего квадратического отклонения используется при довольно больших числах наблюдений. Расчет σ при малом числе наблюдений (при n<30— обязательно) проводится по следующей формуле:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \qquad (1.10)$$

Расчет среднего квадратического отклонения значительно упрощается, если пользоваться формулой

$$\sigma = \frac{x_{\max} - x_{\min}}{K} \qquad (1.11)$$

Величину *K* определяют в зависимости от числа наблюдений по табл. 1.13 С. И. Ермолаева (так, при n == 8 *K* == 2,85; при n=12 *K* = 3,26).

Таблица 1.13 -Таблица для определения величины *K*

n	0	1	2	3	4	5	6	7	8	9
0	—	—	1,13	1,69	2,06	2,33	2,53	2,70	2,85	2,97
10	3,08	3,17	3,26	3,34	3,41	3,47	3,53	3,59	3,64	3,69
20	3,73	3,78	3,82	3,86	3,90	3,93	3,96	4,00	4,03	4,06
30	4,09	4,11	4,14	4,16	4,19	4,21	4,24	4,26	4,28	4,30
40	4,32	4,34	4,36	4,38	4,40	4,42	4,43	4,45	4,47	4,48
50	4,50	4,51	4,53	4,54	4,56	4,57	4,59	4,60	4,61	4,63
60	4,64	4,65	4,66	4,68	4,69	4,70	4,71	4,72	4,73	4,74
70	4,75	4,77	4,78	4,79	4,80	4,81	4,82	4,83	4,83	4,84
80	4,85	4,86	4,87	4,88	4,89	4,90	4,91	4,91	4,92	4,93
90	4,94	4,95	4,96	4,96	4,97	4,98	4,99	4,99	5,00	5,01

Расчет среднего квадратического отклонения при помощи этой таблицы рекомендуется производить для ориентировочной оценки.

Приводим расчет средней арифметической и среднего квадратического отклонения при малом числе наблюдений на следующем примере. У 8 больных антракосиликозом I стадии измерялся остаточный объем легких. Были получены следующие значения вариант (в л): 2,05; 2,09; 2,19; 2,20; 2,23; 2,25; 2,28; 2,31. Средняя арифметическая этих значений равна 2,20 л.

Используя способ Ермолаева, определим σ. Амплитуда вариационного ряда составляет 0,26 (2,31— 2,05), число K (соответствующее 8 наблюдениям) равно 2,85. Тогда σ=0,26/2,85=0,091 л.

Точный расчет σ, произведенный по формуле (1.10), дает почти тот же самый результат:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{0,0566}{7}} = 0,0899$$

Значение коэффициента вариации $c = \frac{0,091}{2,20} * 100 = 4,1\%$ позволяет признать рассеяние вариант в этом примере весьма слабым. При увеличении числа подобных наблюдений можно предполагать, что исследователи встретят и другие значения вариант. Чтобы охватить 95% всех ожидаемых значений вариант при большом числе наблюдений, мы должны использовать t = 1,96, а при степени охвата 99% - t = 2,58. В случае малого числа наблюдений величину t при этих расчетах следует брать из таблицы Стьюдента. Эта величина в первом случае при v = n—1 = 7 равна 2,365, во втором — 3,499.

Исходя из этих данных, следует предположить, что при увеличении числа наблюдений 95% значений остаточного объема легких у больных антракосиликозом I стадии не выйдут за пределы $x \pm 2,37\sigma$, т. е. примут значения от 1,99 до

2,41 л, а 99% значений не выйдут за пределы $x \pm 3{,}5\sigma$, т. е. примут значения от 1,88 до 2,52 л.

1.5 Выборочный метод исследований

При выборочном методе исследований, используя обобщенные характеристики выборочной совокупности, исследователь имеет целью распространить полученные выводы на все целое, на всю генеральную совокупность изучаемых явлений и объектов. Это возможно, если выбранная для исследования часть *репрезентативна* целому, т. е. типична и обладает теми же основными чертами, что и все целое. Иными словами, выборка должна «представлять» свою генеральную совокупность. Репрезентативность выборочной совокупности исследуемого материала обеспечивается при выполнении следующих основных условий:

- соблюдение правил формирования выборки;
- использование достаточного числа наблюдений.

1.5.1 Формирование выборочной совокупности

При формировании выборочной совокупности *отбор* из генеральной совокупности единиц для исследования необходимо производить таким образом, чтобы *исключить возможность систематической ошибки* и тем более *любой преднамеренности*.

В выборочном методе различают несколько видов отбора единиц из всей совокупности наблюдений6

- собственно случайный отбор;
- механический отбор;
- типологический;
- серийный.

Собственно случайный отбор заключается в том, что каждая единица исследуемой совокупности имеет равновероятную возможность попасть в выборку, так как отбирается каким либо случайным способом. Для этого обычно применяются:

а) *метод жеребьевки*, при котором выбор единиц наблюдения осуществляется по жребию. Например, на все единицы совокупности заготавливают одинаковые карточки или жетоны, помещают их в ящик и наугад отбирают необходимое число;

б) *метод случайных чисел*, когда отбор необходимого числа ёдиниц наблюдения производится из общего количества пронумерованных единиц совокупности с применением генератора случайных чисел.

Механический отбор предполагает отбор определённой части генеральной совокупности в механическом порядке (каждый второй, пятый, десятый). Например, из 252 больных, лечившихся в больнице по поводу бронхиальной астмы, решено последующее наблюдение во внебольничных условиях вести за пятьюдесятью (пятая часть). Механическому отбору подлежит каждый пятый

больной, например в порядке регистрации в журнале приемного покоя. Следует учитывать, что такой механический отбор не исключает возможности систематической ошибки.

Типологический отбор состоит в том, что случайный или механический отбор единиц наблюдения производится из однородных групп, на которые разбита генеральная совокупность по какому-либо существенному признаку. Типологический отбор имеет целью повышение репрезентативности выборки и поэтому может быть назван направленным отбором.

Серийный отбор отличается от предыдущих тем, что при нем случайным способом отбираются не отдельные единицы, а их целые группы, или серии. Это удобно, например, при определении групп исследования и контроля.

1.5.2 Определение объема выборочной совокупности

Какой бы способ отбора наблюдений ни применил исследователь, всегда сохраняется вопрос: достаточно ли это число для того, чтобы выборка была репрезентативной?

При исследовании явлений, подчиняющихся закону нормального распределения, на поставленный вопрос хорошо отвечает формула *средней ошибки* средней арифметической величины

$$m = \frac{\sigma}{\sqrt{n}} \qquad (1.12)$$

По размерам средней ошибки исследователь может определить, насколько найденная в опыте выборочная средняя величина отличается от средней генеральной совокупности. Малая ошибка указывает на близость этих показателей, большая ошибка такой уверенности не дает. Из формулы видно, что размер средней ошибки прямо пропорционален среднему квадратическому отклонению, т. е. вариабельности явления, и обратно пропорционален корню квадратному из числа наблюдений.

По средней арифметической величине и ее средней ошибке можно представить себе те границы, называемые *доверительными*, в которых с определенной вероятностью может находиться средняя арифметическая величина генеральной совокупности. Предположим, что какое-либо исследование в аналогичных условиях было повторено многими исследователями. Можно ожидать, что полученные при этом выборочные средние будут отличаться друг от друга. Как установлено, распределение этих выборочных средних подчиняется нормальному закону, а средняя арифметическая из них характеризует нам среднюю арифметическую генеральной совокупности ($\bar{x}_{ген}$). Мерой колеблемости выборочных средних является среднее квадратическое отклонение ($\sigma_{ген}$), которое соответствует средней ошибке средней арифметической (m). Как следует из предыдущего изложения, в пределах $\bar{x} \pm \sigma$ находится 68,3% наблюдений (в дан-

ном случае выборочных средних, полученных при повторных испытаниях); в пределах $\bar{x} \pm 2\sigma$ - 95,5% и в пределах $\bar{x} \pm 3\sigma$ - 99,7%. Согласно этому правилу можно представить, что средняя арифметическая генеральной совокупности с вероятностью в 68,3% находится в пределах $\bar{x} \pm m$. Однако степень надежности такого вывода для медицинских исследований считается недостаточной. Увеличить надежность вывода о возможных размерах средней генеральной совокупности можно лишь при расширении доверительных границ до пределов $\bar{x} \pm 2m$, что позволяет повысить вероятность до 95,5%. Интервал в пределах $\bar{x} \pm 3m$ увеличивает надежность вывода до 99,7%.

Величины 1,2 и 3, определяющие заданный интервал доверия (*доверительные границы*), представляют *доверительные коэффициенты* и обозначаются буквой *t* Определяемая этими коэффициентами степень надежности (в % или долях единиц) называется *доверительной вероятностью*. Вычитая из 100 или единицы доверительную вероятность, получаем *уровень значимости*, или *риск ошибки (p)*. Ниже в табл. 1.14 приведены в кратком виде взаимоотношения названных величин.

Таблица 1.14 – Соотношения между доверительным коэффициентом, доверительной вероятностью и уровнем значимости

Доверительный коэффициент, *t*	Доверительная вероятность 1—*p* (в %)	Уровень значимости *p* (в%)
1,0	0,683(68,3)	0,317(31,7)
1,96	***0,950(95,0)***	***0,050(5,0)***
2,0	0,955(95,5)	0,045(4,5)
2,6	0,990(99,0)	0,010(1,0)
3,0	0,997(99,7)	0,003(0,3)
3,3	0,999(99,9)	0,001(0,1)

Значения приведенных величин, выделенные полужирным курсивом, обычно принимаются для достижения минимальной надежности вывода (при большом числе наблюдений).

Итак, указанная выше формула доверительных границ генеральной средней должна включать величину *t*, указывающую на ту степень надежности, с которой исследователь принимает полученный в опыте результат:

$$\bar{x} \pm t \cdot m \qquad (1.13)$$

Таким образом, величина средней ошибки средней арифметической величины позволяет с любой вероятностью установить границы, в пределах которых находится средняя арифметическая величина генеральной совокупности. Она же может указать на достаточность числа наблюдений и репрезентативность выборки.

При планировании исследования необходимое число наблюдений при избранной доверительной вероятности можно рассчитать по формуле:

$$n = \frac{t^2 \cdot \sigma^2}{m_1^2} \qquad (1.14)$$

Размеры t (не менее 1,96 при большом числе наблюдений) и m_1 (допускаемой ошибки отклонения от генеральной средней) исследователь определяет произвольно. Величина σ может быть определена на небольлшом числе предварительных наблюдений.

1.5.3 Сравнение средних арифметических величин двух выборок из совокупности с нормальным распределением вариант

В процессе исследования, как правило, наблюдается не одна, а несколько серий опытов, и среди них выделяется контрольная, или наблюдается несколько групп больных, с одной из которых сравниваются результаты обследования, лечения остальных. Не редко исследователь сопоставляет данные собственного исследования с данными других авторов, полученных в аналогичных условиях. Целью подобных сравнений может быть установление существенности различий между выборочными средними арифметическими, производимые по формуле

$$t = \frac{\overline{x_1} - \overline{x}_2}{\sqrt{m_1^2 + m_2^2}} \qquad (1.15)$$

В числителе формулы – разность средних арифметических, в знаменателе под корнем – суммы квадратов их средних ошибок, или ошибка разности средних ($m_{разн}$). Если t окажется более 1,96, то с уверенностью, превышающей 95%, можно говорить о существенности различия выборочных средних арифметических, т.е. о том, что они представляют разные генеральные совокупности. Чем больше t, тем больше надежность такого вывода. При малых значениях, принимается так называемая *нулевая гипотеза* об *отсутствии существенных различий между средним генеральных совокупностей*.

Пример. У больных хронической пневмонией с легочной недостаточностью I степени было установлено увеличение количества циркулирующей крови. У 47 больных хронической пневмонией I стадии среднее количество циркулирующей крови ($\bar{x}_1$) составило 6,64 л (m_1 = ±0,17 л); у 53 больных хронической пневмонией II стадии $\bar{x}_2$ = 6,83 л (m_2 = ±0,22 л) В контрольной группе больных пневмонией (56 человек) без нарушения функции внешнего дыхания среднее количество циркулирующей крови ($\bar{x}$) равнялось 6,12 л (m = ±0,13).

Разность среднего количества циркулирующей крови у больных хронической пневмонией I стадии контрольной группы ($\bar{x}_1$-$\bar{x}$) оказалась вполне убедительной:

$$t = \frac{6,64 - 6,12}{\sqrt{0,17^2 + 0,13^2}} = \frac{0,52}{\sqrt{0,0458}} = \frac{0,52}{0,21} = 2,5.$$

Так как t>1,96, разность сравниваемых средних не случайная, с вероятностью не меньше 95%.

Еще более убедительной оказалась разность среднего количества циркулирующей крови у больных хронической пневмонией II стадии и контрольной группы ($\bar{x}_2$ и $\bar{x}$):

$$t = \frac{6,83 - 6,12}{\sqrt{0,22^2 + 0,13^2}} = \frac{0,71}{\sqrt{0,0653}} = \frac{0,71}{0,25} = 2,8.$$

Существенность найденных различий установлена с вероятностью не менее 99% (t>2,6)

Сравнение среднего: количества циркулирующая крови у больных хронической пневмонией I и II;стадий ($\bar{x}_1$ и $\bar{x}_2$) приводит к очень небольшой величине t:

$$t = \frac{6,83 - 6,64}{\sqrt{0,22^2 + 0,17^2}} = \frac{0,19}{\sqrt{0,0773}} = \frac{0,19}{0,28} \approx 0,7.$$

Здесь вывод о не случайности различии средних может быть сделан лишь с очень большим риском ошибки — не менее 31,7%. В этом случае правильнее утверждать, что различия средних не доказаны: (предпочтение отдается нулевой гипотезе)

При наличии сомнений в справедливости такого вывода можно рекомендовать увеличение числа наблюдений в обеих группах и повторную статистическую оценку полученных результатов.

Иногда достоверность различий средних арифметических можно доказать простым сравнением их доверительных границ. Если доверительные границы сравниваемых выборочных средних величин полностью или даже частично совпадают, то достоверность различий этих средних не доказана. Если; же сравниваемые доверительные границы не совпадают, то различия средних арифметических следует признать неслучайными.

Пример. Определим, существенны ли различия между средними величинами максимального артериального давления больных острой пневмонией (в лихорадочном периоде), получавших антибиотики (1-я группа— 80 чел) и получавших антибиотики и тауремизин (2-я группа — 97 чел.) (табл. 1.15).

Таблица 1.15 - Оценка существенности различия между средними величинами максимального артериального давления у больных острой пневмонией, лечившихся разными методами

Группы больных	Число больных, n	Средняя величина артериального давления, $\bar{x}$ мм рт.ст.	Средняя ошибка, $\pm m$ мм.рт.ст.	Доверительные границы средней с надежностью в 99,7%, $\bar{x} \pm 3m$ мм рт.ст.
1-я группа	80	95	0,47	93,59 ÷ 96,41
2-я группа	97	116	0,51	113,47÷ 116,53

Полученные доверительные границы сравниваемых средних не совпадают, и потому следует считать различия между средними существенными с весьма высокой надежностью.

Оценка достоверности результатов *малой выборки (n <30)* по ранее рассмотренным формулам может приводить к значительным погрешностям. Поэтому разработаны специальные методы, обеспечивающие репрезентативность данных, полученных в малой выборке. При малом числе наблюдений расчёт средней ошибки средней арифметической производится по формуле (1.12), однако после раскрытия значения σ в малой выборке эта формула получает новый вид:

$$m = \sqrt{\frac{\sum (x - \bar{x})^2}{n(n-1)}} \qquad (1.16)$$

При определении доверительных границ средних арифметических величин из малых выборок значения *t*, соответствующие избранным доверительным вероятностям, следует брать из таблицы Стьюдента. Нахождение табличных *t* требует предварительного определения числа степеней свободы *v*, которое в данном случае равно $v = n\text{-}1$.

Оценка существенности различия средних арифметических величин с помощью t-критерия при *независимых малых выборках* производится по формуле:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} \cdot \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}} \qquad (1.17)$$

В числителе формулы по-прежнему находится разность средних арифметических величин сравниваемых выборок, знаменатель же - ошибка разности определяется так:

$$m_{разн}=\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}\cdot\sigma_{разн}=\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}\cdot\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n_1+n_2-2}} \qquad (1.18)$$

где n_i -численность 1-й-выборки, а n_2 - численность 2-й-выбоки.

Методику расчетов и технику оценки статистической значимости разности средних при малых сериях независимых наблюдений рассмотрим на примере сравнения среднего кардиопортального времени у больных с разными стадиями цирроза печени. Имеется данные о кардиопортальном времени у больных двух групп:

1-я группа — x_1 (больные с активной стадией цирроза)—42, 46, 45, 48, 43, 45,46, 43,45,47, 45; n_1 = 11;

2- группа—x_2(больные с неактивной стадией цирроза) — 38, 40, 36, 35, 37, 38, 39, 38, 40, 37, 38, 37, 36, 37, 38, n_2= 15.

Среднее арифметическое кардиопортальное время больных первой группы ($\bar{x}_1$) равно 45, второй группы ($\bar{x}_2$) - 37,6.

Вычислим среднее квадратическое отклонение разности и подставим его значение в формулу (1.18) для *t:*

$$\sigma_{разн}=\sqrt{\frac{32-27,6}{24}}=1,58$$

$$t=\frac{45-37,6}{\sqrt{\frac{11+15}{11\cdot15}}\cdot1,58}=\frac{7,4}{0,627}=11,8$$

Чтобы оценить существенность различия средних, находим табличные значения *t*. Доверительной вероятности 0,95, и 24 степеням свободы (здесь степени свободы определяются как $n_1 + n_2$—2) соответствует величина t = 2,064, а вероятности 0,99 - t=2,797. Следовательно с вероятностью > 99% можно утверждать, что различие средних величин кардиопортального времени больных сравниваемых групп является статистически значимым.

1.6 Основы дисперсионного анализа

1.6.1 Общие положения

В предыдущей главе была показана методика оценки различий средних арифметических двух выборочных групп наблюдений путем вычисления средней ошибки разности средних и критерия *t*. Применение этого критерия позволяет выяснить, значима ли статистически разность средних величин, велика ли вероятность проявления такой разности случайным образом. При этом подразумевается, что группы исследуемых признаков совершенно однородны и отличаются только по одному какому-либо признаку или методу воздействия на них. Между тем на практике это условие соблюдается далеко не всегда. На размерах явлений и, следовательно, их средних сказывается влияние многочисленных факторов, как постоянных (планируемых или сознательно выделяемых для исследования), так и случайных (многообразных и неопределенных). Например, больные гипертонической болезнью (однородные по полу, возрасту, стадии и длительности заболевания), помимо болезни, подвергаются воздействию других различных факторов, в результате чего у разных больных наблюдается разная высота кровяного давления.

При изучении явлений, сравнении их друг с другом в поисках сходства и различий необходимо обращать внимание не только на размеры средних величин, но и на разнообразие вариант, на варьирование изучаемых признаков. Исследователь может встретить ряды величин не отличающиеся по центральной тенденции (размер средней арифметической), но различные по степени варьирования, одинаковые по величине разброса вариант, но различные по размерам средней арифметической. Установление значимости различий средних арифметических величин, измерение степени влияния факторов и их градаций на варьирующий (результативный) признак наиболее эффективно достигаются путем применения *дисперсионного анализа*. Таким образом сущность дисперсионного анализа заключается в определении значимости влияния отдельных факторов и их относительной роли в общей вариации изучаемого признака.

При решении указанных задач методом дисперсионного анализа используются *вариация* (*S*) и *дисперсия* (σ^2). Под вариацией понимают величину, представляющую собой сумму квадратов отклонений вариант от средней

$$S = \sum (x - \bar{x})^2 \qquad (1.19)$$

Под дисперсией понимают величину

$$\sigma^2 = \frac{S}{\nu} \qquad (1.20)$$

Знаменатель определяет число степеней свободы. Степени свободы обозначают число вариант, которые могут принимать любые значения без измене-

ния их общей суммы. Например, имеются варианты (*x*): 3, 6, 7, 9, 10; *n*=5; Σ*x*=35. Какое число вариант может свободно изменить свои значения без изменения общей их суммы? Очевидно, что только 4 т.е. (*n*-1) варианты могут быть свободно изменены. Значение пятой варианты свободно варьировать не может, так как связано необходимостью обеспечить неизменной сумму Σ*x*=35.

Различают несколько видов вариации и соответствующим им дисперсий. Рассмотрим принципиальную схему вычисления вариации и дисперсий при применении дисперсионного анализа

Пусть проведены 3 группы исследований. В каждой из них использована одна градация какого-либо фактора, например разный уровень содержания кислорода, неодинаковая доза лекарства и т. п.

Обозначим буквами: *a, b* и *c*—варианты каждой группы; $\bar{a}$, $\bar{b}$ и $\bar{c}$– их средние арифметические величины: n_a,n_b,n_c числа наблюдений в каждой группе; $\bar{x}$ — среднюю арифметическую для всех вариант *(a, b* и *c*) и *n*—общее число наблюдений.

Общая вариация (S) вариация — сумма квадратов отклонений всех значений и вариант *(a, b* и *c*) от их общей средней ($\bar{x}$). На ней сказываются влияния как постоянных так и случайных факторов.

Остаточная (внутригрупповая) вариация S_z представляет собой сумму групповых вариаций:

$$S_z = \sum(\sum(a-\bar{a})^2 + \sum(b-\bar{b})^2 + \sum(c-\bar{c})^2) \qquad (1.21)$$

Эта часть общей вариации отражает влияние случайных факторов.

Факториальная (межгрупповая) вариация (S_ϕ) квадрат отклонения групповых средних (*$\bar{a}$, $\bar{b}$* и $\bar{c}$) от общей средней ($\bar{x}$):

$$S_\Phi = (\bar{a}-\bar{x})^2 + (\bar{b}-\bar{x})^2 + (\bar{c}-\bar{x})^2 \qquad (1.22)$$

Именно на её размере сказывается влияние изучаемого фактора.

Общая (σ^2), факториальная ($\sigma_\phi{}^2$) и остаточная ($\sigma_Z{}^2$) дисперсии представляют собой отношение общей, факториальной и остаточной вариации к соответствующему числу степеней свободы *v*. Для общей дисперсии $v = n-1$, где *n*-число наблюдений. Для факториальной дисперсии $v_\Phi = r-1$, где r – число групп. Для остаточной дисперсии $v_Z = n-r$.

Общая сумма квадратов отклонений, или полная вариация, равна сумме отдельных вариаций, т. е. вариации факториальной и остаточной

$$S = S_\phi + S_z \qquad (1.23)$$

Общее число степеней свободы, используемых при расчете факториальной и остаточной дисперсии:

$$v = v_\phi + v_Z = n-r+r-1 = n-1 \qquad (1.24)$$

Для оценки значимости влияния изучаемого фактора определяется отношение факториальной дисперсии к остаточной дисперсии. Размеры полученно-

го критерия $F = (\frac{\sigma_{\Phi}^2}{\sigma_Z^2})$ оцениваются по специальным таблицам. Если эмпирическое значение F превышает табличное, то это свидетельствует о существенности влияния изучаемого фактора. Если же критерий F меньше табличного, то вывод о значимости влияния изучаемого фактора не может считаться доказанным. В зависимости от числа учитываемых организованных факторов различают одно-, двух-, трех- и т. д. факторный дисперсионный анализ. Ниже на примерах показана техника расчетов, необходимых при дисперсионном анализе.

1.6.2 Методика однофакторного дисперсионного анализа

В табл. 1.16 представлены результаты определения противоопухолевого действия креатининсульфата серотонина на прививаемую беспородным мышам опухоль (саркома180). Изучаемый фактор применен в трех разных дозах. Критерием противоопухолевого эффекта являлось снижение среднего веса опухоли (результативный признак) под влиянием различных доз серотонина.

Таблица 1.16 - Результаты определения противоопухолевого действия креатининсульфата серотонина

Доза в *мг/кг*	Вес опухоли (саркома 180) в $\varepsilon - x$. В скобках квадраты вариант (x^2)	Число наблюдений, n	Суммарный вес опухолей, Σx	Средний вес опухолей, $\frac{\Sigma x}{n}$	Сумма квадратов вариант, Σx^2	Квадрат средней величины, $\bar{x}^2$	Произведение квадрата средней величины на число наблюдений, $\bar{x}^2 \cdot n$
2,5	2,0; 2,5; 2,7; 1,7; 1,9; 2,4; (4,00) (6,25) (7,29) (2,89) (3,61) (5,76)	6	13,2	2,2	29,80	4,84	29,04
5,0	1,4; 1,6; 1,7; 2,0; 1,8 (1,96) (2,56) (2,89) (4,00) (3,24)	5	8,5	1,7	14,65	2,89	14,45
25,0	0,9; 1,0; 1,3; 1,6; 1,5; 1,4; 1,4 (0,81) (1,00) (1,69) (2,56) (2,25) (1,96) (1,96)	7	9,1	1,3	12,23	1,69	11,83
Итого		$n = 18$	30,8	—	56,68		55,32
Общая средняя арифметическая ($\bar{x}$) и ее производные				$\bar{x} = 1,7$		$\bar{x}^2 = 2,89$	$\bar{x}^2 \cdot n = \frac{(\Sigma x)^2}{n} = 52,70$

В таблице представлены необходимые промежуточные данные для последующих расчетов по формулам:

1. Полной вариации

$$S = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 56,68 - 52,70 = 3,98;$$

2. Факториальной вариации

$$S_{\Phi} = (n_1\bar{a}^2 + n_2\bar{b}^2 + n_3\bar{c}^2) - \frac{(\Sigma x)^2}{n} = 55{,}32 - 52{,}70 = 2{,}62;$$

3. Остаточной вариации

$$S_Z = S - S_{\Phi} = 3{,}98 - 2{,}62 = 1{,}36$$

4. Факториальной дисперсии

$$\sigma^2_{\Phi} = \frac{S_{\Phi}}{\text{число групп} - 1} = 2{,}62 : 2 = 1{,}31;$$

5. Остаточной дисперсии

$$\sigma^2_Z = \frac{S_Z}{n - \text{число групп}} = 1{,}36 : 15 = 0{,}09:$$

6. Критерия *F*

$$F = \frac{\sigma^2_{\Phi}}{\sigma^2_Z} = 1{,}31 : 0{,}09 = 14{,}5.$$

Табличные значения *F* находим по таблице, учитывая, что число степеней свободы для большей дисперсии указано в графах, а для меньшей дисперсии - в строках: F_{05} = 3,68 и F_{01} = 6,36.

Итоговые данные удобно представлять в следующем виде (табл. 1.17).

Таблица 1.17 Сводная таблица однофакторного дисперсионного анализа

Воздействие на вариабельность	Сумма квадратов отклонений — вариация (*S*)	Число степеней свободы, ν	Дисперсия, σ^2	Отношение дисперсий, *F*	Табличное значение *F* для уровня значимости	
					5%	1%
Исследуемый фактор	2,62	2	1,31	14,5	3,68	6,36
Остальные факторы (случайные причины)	1,36	15	0,09			
Все факторы	3,98	17				

Поскольку вычисленное значение критерия *F* превышает табличные значения, следует заключить, что случайность различий средних значений веса опухолей маловероятна (p<1%). Можно утверждать, что воздействие креатининсульфата серотонина замедляет рост прививаемой опухоли.

Удобство и ценность дисперсионного анализа очевидны. Так, можно оценивать различия в средних величинах и при большем количестве исследуемых групп.

С его помощью можно также определить *силу влияния* (η^2) изучаемого фактора его долю в сумме всех влияний на результативный признак. Эта вели-

чина равна отношению факториальной вариации к общей вариации. В рассмотренном примере $\eta^2 = \frac{S_Ф}{S} = \frac{2,62}{3,98} = 0,658$

Эта величина означает, что воздействие серотонина на вес опухолей значительно и составляет 65,8%, тогда как на долю всех других неучтенных в данном исследовании влияний приходится лишь 34,2%. Средняя ошибка η^2 определяется по формуле:

$$m_{\eta^2} = (1 - \eta^2) \cdot \frac{\text{число групп} - 1}{n - \text{число групп}}, \qquad (1.25)$$

По имеющимся данным, средняя ошибка составляет $m_{\eta^2} = (1 - 0,658) \cdot \frac{2}{15} = 0,342 \cdot 0,133 = 0,045$. Она в 14,6 раза меньше $\eta^2 \left(\frac{\eta^2}{m_{\eta^2}} = \frac{0,658}{0,045} = 14,6 \right)$ что обязывает с полным доверием относиться к показателю силы влияния изучаемого фактора

Доверительные границы силы влияния фактора при уровне значимости p = 0,05 ($F табл.$ = 3,68) могут быть определены по формуле:

$$\eta^2 \pm F_{0,05} \cdot m_{\eta^2} \qquad (1.26)$$

В нашем случае $0,658 \pm 3,68 * 0,045 = 0,658 \pm 0,166$ или 0,492—0,824. Следовательно, при изучении генеральной совокупности доля влияния серотонина может колебаться в пределах от 49,2 до 82,4%. Таким образом, применение дисперсионного анализа с большой достоверностью позволило установить значимость и надежность влияния серотонина на величину опухолей.

Приведенные расчеты представляют пример однофакторного дисперсионного анализа. Из всех многообразных факторов, действующих на размеры опухоли, учтен только один - дозировка серотонина. Варьирование данных внутри каждой группы предполагалось случайным. При использовании однофакторного анализа число наблюдений в отдельных группах может быть либо одинаковым (равномерный статистический комплекс), либо разным (неравномерный статистический комплекс). Техника расчетов при этом не меняется. Если исследователь изучает влияние не одного, а большего числа факторов, требуется использование более сложной методики.

1.6.3 Методика двухфакторного дисперсионного анализа

Изучалась скорость выдоха (л/с) по данным пневмотахометрии у здоровых рабочих угольных шахт, Одним из факторов, учитывавшихся в исследовании, являлись условия производственной деятельности (фактор *A).* Одна группа рабочих не имела контакта с производственной пылью *(A_1);* другую группу соста-

вили рабочие, занятые на подземных работах *(A_2)*. Предполагалось, что работа в запыленных условиях (фактор *A_2)* влияет на скорость выдоха.

В этом же исследовании одновременно изучалось влияние возраста (фактор *B)*. Было отмечено, что с возрастом максимальная скорость выдоха уменьшается. Требовалось установить, связано ли уменьшение скорости выдоха только с возрастом, или же производственные условия также влияют на этот показатель.

Для проведения необходимых расчетов экспериментальные данные сводятся в специальную таблицу (табл. 1.18). Всех исследуемых необходимо разделить на группы по фактору *A* (в примере их 2: *A_1* и *A_2)*, каждую из групп *A* разделить еще на группы по фактору *B* (в примере их также 2: *B_1* и B_2). В каждой из указанных подгрупп следует записать значения вариант — *x* (в нашем примере - величины максимальной скорости выдоха), подсчитать число наблюдений, определить средние арифметические величины.

В аналогичной таблице следует сгруппировать варианты отдельно по фактору *A* и по фактору *B* (табл. 1.19).

Таблица 1.18 - Скорость выдоха у рабочих разного возраста, работающих в разных производственных условиях

	A_1-здоровые рабочие, не имеющие контакта с производственной пылью		A_2-здоровые рабочие, занятые на подземных работах угольных шахт		Итого
	B_1(30-39 л)	B_2(40-49 л)	B_1(30-39 л)	B_2(40-49 л)	
x	4,5; 4,7	4,0; 4,1; 4,2	4,3; 4,5	3,8; 3,9; 4,0	
n	2	3	2	3	10
$\sum x$	9,2	12,3	8,8	11,7	42
$\bar{x}$	4,6	4,1	4,4	3,9	4,2

Таблица 1.19 - Скорость выдоха у работающих в разных производственных условиях лиц разного возраста

	A1-здоровые рабочие, не имеющие контакта с производственной пылью	A2-здоровые рабочие, занятые на подземных работах угольных шахт	B1-рабочие в возрасте 30-39 лет	B2-рабочие в возрасте 40-49 лет
x	4,5; 4,7; 4,0; 4,1; 4,2	4,3; 4,5; 3,8; 3,8; 4,0	4,5; 4,7; 4,3; 4,5	4,0; 4,1; 4,2; 3,8; 3,9; 4,0
n	5	5	4	6
$\sum x$	21,5	20,5	18	24
$\bar{x}$	4,3	4,1	4,5	4

Чтобы определить общую и остаточную вариации по факторам *A* и *B*, а затем соответствующие дисперсии, необходимо найти:

1) квадраты всех вариант—x^2;

2) суммы этих квадратов—Σx^2;

3) суммы вариант, возведенные в квадрат,—$(\Sigma x)^2$;

4) суммы вариант, возведенные в квадрат и поделенные на число наблюдений $-\frac{(\sum x)^2}{n}$.

Последнюю дробь $\frac{(\sum x)^2}{n}$ обозначим *h*, если расчет производится для какой-либо отдельной группы, и *H*, если расчет касается всех наблюдений. Аналогично: *n*—число наблюдений в отдельных группах, а *N*—общее число наблюдений. Через *r* обозначим число групп, на которое делятся наблюдаемые по фактору *A* (r_A) и по фактору *B* (r_B).

Последовательность двухфакторного дисперсионного анализа следующая:

I. Определяем вариацию по фактору *A* (S_A)

$$S_A = h_A — H;$$

II. Определяем вариацию по фактору *B* (S_B)

$$S_B = h_B — H;$$

3. Определяем вариацию по сумме факторов *A* и *B* и взаимодействия этих факторов (S_{A+B+AB})

$$S_{A+B+AB} = \Sigma h — H;$$

4. Определяем вариацию, обусловленную взаимодействием факторов *A* и *B* - так называемый «перекрестный эффект» (S_{AB})

$$S_{AB} = S_{A+B+AB} - S_A - S_B;$$

5. Определяем остаточную вариацию (S_z)

$$S_Z = \Sigma x^2 - \Sigma h;$$

6. Определяем полную вариацию-(*S*)

$$S = \Sigma x^2 - H \quad (S = S_A + S_B + S_{AB} + S_z);$$

7. Определяем число степеней свободы (V_A) и дисперсию (σ_A^2) по фактору *A*

$$V_A = r_A - 1; \qquad \sigma_A^2 = S_A / V_A;$$

8. Определяем число степеней свободы (V_B) и дисперсию (σ_B^2) по фактору *B*

$$v_B = r_B - 1; \quad \sigma_B^2 = \frac{S_B}{v_B}$$

9. Определяем число степеней свободы *(V_{AB})* и дисперсию σ_{AB}^2 для взаимодействия факторов *A* и *B*

$$v_{AB} = v_A \cdot v_B; \quad \sigma_{AB}^2 = \frac{S_{AB}}{v_{AB}}$$

10. Определяем число степеней свободы (ν_Z) и величину остаточной дисперсии (σ_Z)

$$\nu_Z = n - r_A \cdot r_B; \quad \sigma_Z^2 = \frac{S_Z}{\nu_z}$$

11. Определяем критерий *F* — отношение каждой изучаемой дисперсии к остаточной дисперсии .f=σ^2/σ_Z^2

$$\frac{\sigma_A^2}{\sigma_Z^2}; \quad \frac{\sigma_B^2}{\sigma_Z^2}$$

12. Производим оценку критерия *F* по специальной таблице

13. Степень влияния изучаемых факторов определяем по отношению каждой изучаемой вариации к полной вариации

$$\eta_A^2 = \frac{S_A}{S}; \quad \eta_B^2 = \frac{S_B}{S} \text{ и т. д.}$$

Сосредоточим основные расчеты в следующей таблице(табл. 1.20) затем по указанной схеме рассчитаем остальные компоненты двухфакторного дисперсионного анализа.

Таблица 1.20 – Дисперсионный анализ влияния контакта с производственной пылью и возраста на максимальную скорость выдоха

	A_1 –здоровые рабочие, не имеющие контакта с производственной пылью		A_2 –здоровые рабочие, занятые на подземных работах угольных шахтах		r_A=2 r_B=2
	B_1 – 30-39 лет	B_2 – 40-49 лет	B_1 – 30-39 лет	B_2 – 40-49 лет	
x	4,5; 4,7	4,0; 4,1; 4,2	4,3; 4,5	3,8; 3,9; 4,0	$H = \frac{(\sum x)^2}{N} = \frac{(42,0)^2}{10} = 176,40$
n	2	3	2	3	*N*=10
$\sum x$	9,2	12,3	8,8	11,7	$\sum x$=42,0
$\bar{x} = \frac{\sum x}{n}$	4,6	4,1	4,4	3,9	$\bar{x}$ =4,2
$\sum x^2$	42,34	50,45	38,74	45,65	$\sum x^2$=177,18
$(\sum x)^2$	84,64	151,29	77,44	136,89	$(\sum x)^2=(42,0)^2$=1764
$h = \frac{(\sum x)^2}{n}$	42,32	50,43	38,72	45,63	$\sum h$=177,10

h_A	$\frac{(21,5)^2}{5}+\frac{(20,5)^2}{5}=176,5$	
h_B	$\frac{(18,0)^2}{5}+\frac{(24,0)^2}{5}=177,0$	

1. $S_A = h_A — H$ = 176,50 - 176,40 = 0,10.
2. $S_B = h_B — H$ = 177,00 - 176,40 = 0,60.
3. $S_{A+B+AB} = \Sigma h — H$ = 177,10 - 176,40 = 0,70.
4. $S_{AB} = S_{A+B+AB} — S_A - S_B$ = 0,70 -0,10 -0,60=0,00.
5. $S_Z = \Sigma x^2 — \Sigma h$ = 177,18— 177,10 =.0,08.
6. $S = \Sigma x^2 — H$ = 177,18 - 176,40 = 0,78.

7. $\nu_A = r_A - 1 = 2 - 1 = 1;\quad \sigma_A^2 = \frac{S_A}{\nu_A} = \frac{0,10}{1} = 0,10.$

8. $\nu_B = r_B - 1 = 2 - 1 = 1;\quad \sigma_B^2 = \frac{S_B}{\nu_B} = \frac{0,60}{1} = 0,60.$

9. $\nu_{AB} = \nu_A \cdot \nu_B = 1 \cdot 1 = 1;\quad \sigma_{AB}^2 = \frac{S_{AB}}{\nu_{AB}} = \frac{0,00}{1} = 0,00.$

10. $\nu = N - r_A \cdot r_B = 10 - 2 \cdot 2 = 6;\quad \sigma_Z^2 = \frac{S_Z}{\nu_Z} = \frac{0,08}{6} = 0,013.$

11. $F_A = \frac{\sigma_A^2}{\sigma_Z^2} = \frac{0,10}{0,013} = 7,69.$

$F_B = \frac{\sigma_B^2}{\sigma_Z^2} = \frac{0,60}{0,013} = 46,15.$

12. Табличное значение критерия F при числе степеней свободы большей дисперсии $\nu_1 = \nu_A = 1$ и меньшей дисперсии $\nu_2 = \nu_z = 6$ равно 5,99. Так как расчетный критерий F превышает табличное значение при 5% уровне значимости, то это означает, что найденные различия в средних арифметических максимальной скорости выдоха значимы, не случайны, им можно доверять. Риск ошибки этого вывода в отношении фактора A не более 5%, риск ошибки в отношении влияния фактора B—менее 1%. Значимость различий средних в зависимости от взаимодействия фактора A и фактора B по имеющимся данным не установлена.

Степень влияния отдельных изучаемых факторов в общем числе влияний довольно значительна:

$$\eta_A^2 = \frac{S_A}{S} = \frac{0,10}{0,78} = 0,128 \text{ (т. е. 12,8%)},$$

$$\eta_B^2 = \frac{S_B}{S} = \frac{0,60}{0,78} = 0,769 \text{ (т. е. 76,9%)}.$$

Степень влияния условий работы (0,128) у здоровых рабочих оказалась меньше степени влияния возраста (0,769). Суммарное влияние двух выделенных факторов составляет 0,897 (0,128+0,769), т. е. 89,7%. Итоговые данные результатов расчета можно сконцентрировать в таблице (табл. 1.21).

Таблица 1.21 - Сводная таблица двухфакторного дисперсионного анализа

Воздействие на вариабильность	Вариация, S	Степень влияния, η^2, в %	Число степеней свободы, ν	Дисперсия, σ^2	Отношение дисперсий к остаточной дисперсии, F	Табличные значения для уровня вероятности (в %)	
						5	1
Исследуемый фактор A . .	0,10	12,8	1	0,10	7,69	5,99	13,75
Исследуемый фактор B . .	0,60	76,9	1	0,60	46,15	5,99	13,75
Исследуемое взаимодействие факторов AB . . .	0,00	—	1	0,00			
Остальные факторы (случайные причины) . .	0,08	10,3	6	0,013			
Все факторы	0,78	100,0	9				

Рассмотрим еще раз формулу о сумме вариации:

$$S = S_A + S_B + S_{AB} + S_Z$$

Если в эксперименте выделяется действие только одного фактора A, то S_B и S_{AB} входят в S_z. Увеличение S_z за счет S_B и S_{AB} может привести к тому, что значимость влияния фактора A окажется по расчетам недостаточной. Поэтому следует по возможности, кроме фактора A, исследовать какой-либо фактор B, пусть малозначащий, его выделение уменьшит S_z и яснее выявит роль фактора A.

Если исследователя интересует влияние нескольких факторов (более двух), то необходимо провести трехфакторный, четырехфакторный дисперсионный анализ, т. е. многофакторный анализ.

1.6.4 Методика однофакторного дисперсионного анализа альтернативных признаков

При альтернативном варьировании результативного признака (осложнение было или не было; летальный исход наступил или не наступил) необходимо применять следующую методику.

1. Определить факториальную вариацию ($S_Ф$):

$$S_Ф = \Sigma \frac{m_1^2}{n} - \frac{(\Sigma m_1)^2}{N}.$$

2. Определить случайную вариацию (S_z);

$$S_Z = \sum m_1 - \sum \frac{m_1^2}{n}$$

3. Определить общую вариацию (S):

$$S = \Sigma m_1 - \frac{(\Sigma m_1)^2}{N}.$$

4. Определить факториальную дисперсию ($\sigma^2_ф$):

$$\sigma_Ф^2 = \frac{S_Ф}{r-1}.$$

5. Определить случайную дисперсию σ^2_Z:

$$\sigma_z^2 = \frac{S_Z}{N-r}.$$

6. Определить отношение дисперсий (критерий Фишера):

$$F = \frac{\sigma_Ф^2}{\sigma_z^2}.$$

7. Оценить достоверность различия в частоте альтернативного признака по таблице

8. Определить показатель силы влияния изучаемого фактора на результативный признак ($\eta^2_Ф$):

$$\eta_Ф^2 = \frac{S_Ф}{S}.$$

9. Определить ошибку этого показателя:

$$m_{\eta_Ф^2} = \left(1 - \eta_Ф^2\right) \frac{r-1}{N-r}.$$

10. Определить отношение показателя силы влияния изучаемого фактора к его ошибке:

$$\frac{\eta_Ф^2}{m_{\eta_Ф^2}}.$$

11. Оценить достоверность показателя силы влияния и определить доверительные границы этого показателя.

В приведенных формулах использованы обозначения:

r — число изучаемых групп;

m_1 — количество результативных признаков в каждой группе;

n—число наблюдений в каждой группе;

N—общее число наблюдений ($N=\Sigma n$).

Рассмотрим следующий пример. По данным больницы проведено исследование числа послеоперационных осложнений для различных операций. Среднее число осложнений находилось в пределах 0÷16 %. Необходимо определить, оказывает ли влияние характер операции на частоту осложнений. Результаты наблюдений и расчетов представлены в таблице 1.22.

Таблица 1.22 - Дисперсионный анализ послеоперационных осложнений в зависимости от характера операции

	Холецистэктомия	Грыжесечение при неущемленной грыже	Аппендектомия при хроническом аппендиците	Резекция желудка по поводу язвенной бо-	Число групп r=4 $H=\frac{(\sum m_1)^2}{N}=1,7$	Факториальная вариация $S_\Phi=\sum h-H=2,5-1,7=0,8$
n	61	185	48	41	$N=\sum n=335$	$S_z=\sum m_1-\sum h=$ $=24-2,5=21,5$
m_1	10	11	-	3	$\sum m_1=24$	$S=\sum m_1-H=24-1,7=22,3$
$h=\frac{m_1^2}{n}$	1,6	0,7	0,0	0,2	$\sum h=2,5$	$\sigma_\phi^2=\frac{S_\phi}{r-1}=\frac{0.8}{3}=0.3$
$p=\frac{m_1^2}{n}$	0,16	0,07	0,00	0,07	$p=\frac{24}{335}=0.07$	$\sigma_z^2=\frac{S_z}{N-r}=\frac{21.5}{331}=0.07$

$$\eta^2=\frac{S_\Phi}{S}=\frac{0,8}{22,3}=0,036\ (\text{т. е. } 3,6\%)$$

$$m_{\eta^2}=(1-\eta^2)\frac{r-1}{N-r}=(1-0,036)\cdot\frac{3}{331}=0,964\cdot 0,01=0,0096$$

$$\frac{\eta^2}{m_{\eta^2}}=\frac{0,036}{0,0096}=3,7$$

$$\nu_\Phi=r-1=3 \qquad \nu_Z=N-r=331$$

Доверительные границы показателя силы влияния η^2 *в* генеральной совокупности:

$$\eta^2 + F_{0,95} \cdot m_{\eta^2} = 0{,}036 + 2{,}6 \cdot 0{,}0096 = 0\,062$$

$$\eta^2 - F_{0,95} \cdot m_{\eta^2} = 0{,}036 - 2{,}6 \cdot 0{,}0096 = 0{,}010$$

$$F = \frac{\sigma_\Phi^2}{\sigma_Z^2} = \frac{0{,}3}{0{,}07} = 4{,}29$$

Расчеты показывают, что характер оперативного вмешательства оказывает влияние на частоту послеоперационных осложнений, причем влияние этого фактора можно считать значимым с надежностью не менее 99%, а риск ошибки этого вывода составляет менее 1% (4,3 > 3,78). Однако влияние данного фактора невелико и составляет всего 3,6 % в числе других неучтенных факторов. Можно полагать, что в генеральной совокупности процент влияния может оказаться несколько другим, но диапазон предполагаемых его размеров небольшой - не менее 1,0% и не более 6,2%.

Если исследователь не устанавливает большой силы влияния одного изучаемого фактора на результативный признак, следует направить усилия на поиски других факторов, играющих большую роль. Таким образом, очевидный, казалось бы, вывод о том, что частота послеоперационных осложнений в основном обусловлена характером оперативного вмешательства, не подтвердился. Наряду с характером операции, по-видимому, значительную роль играют и другие факторы установление которых представляет большой практический интерес. В рассмотренном примере такими факторами могут быть: состояние больного, его возраст, квалификация хирурга, качество подготовки больного к операции, качество послеоперационного ухода и т. д.

1.7 Определение соответствия эмпирических и теоретических данных

1.7.1 Общие положения

Одной из частых задач научного исследования является определение соответствия (согласия или различия) эмпирического и теоретического распределений или нескольких эмпирических распределений. Кроме методов, указанных выше, в современной статистике для этого широко используется критерий χ^2 (хи-квадрат), предложенный Пирсоном. Расчет χ^2 производится только по абсолютным величинам. В основе метода лежит сопоставление частот, интересующих исследователя распределений. При полном совпадении этих частот (эмпирических величин, полученных в опыте с данными ожидаемыми, теоретическими) χ^2 равен нулю. По мере увеличения различий между сравниваемыми частотами значение χ^2 возрастает. Целью расчетов является доказательство возможности признать или отвергнуть *нулевую гипотезу* т. е. *предположение об отсутствии существенных различий между сравниваемыми данными*. Общая формула для вычисления критерия χ^2

$$\chi^2 = \sum \frac{(\hat{O} - O)^2}{O} \qquad (1.26)$$

где *Ф* — эмпирические данные (частоты полученного распределения);

О — ожидаемые данные (частоты теоретического или другого сравниваемого распределения).

Полученную величину χ^2 необходимо оценить, сравнив ее с табличными значениями.

Табличные значения χ^2 зависят от числа степеней свободы и принятого уровня значимости. Число степеней свободы ν в случаях, когда сопоставляемые данные представлены в виде таблицы определяется по формуле: ν = (число строк — 1)×(число столбцов — 1). Для примера представленного в табл. 1.23 число столбцов равно 2, число строк – 2, а число степеней свободы v=(2-1)*(2-1)=1. Нулевая гипотеза отвергается, если вычисленная величина χ^2 больше табличного, значения χ^2 при уровне значимости 0,01 (риск ошибки меньше 1%), что можно записать как $\chi^2 > \chi^2_{01}$. Нулевая гипотеза принимается, если $\chi^2 \leq \chi^2_{05}$.

Пример. Исследователем изучалась частота побочных явлений при лечении антибиотиками (синтомицин и левомицетин) двух групп больных. В группе больных, получавших кроме антибиотиков витамины, частота побочны явлений оказалась меньше, (9 из 66 больных), чем в группе больных, леченных только антибиотиками (16 из 46). Требуется определить, достаточно ли надёжны эти различия в частоте побочных явлений, можно ли им доверять. Полученные данные следует записать в виде четырехпольной: таблицы (табл.1.23). Примем нулевую гипотезу о том, что витамины не оказывают влияния на частоту побочных явлений, вызываемых антибиотиками, что фактически наблю-

даемые различия в частоте побочных явлений случайны и что при гораздо большем числе наблюдений различий бы не было. Это предположение дает право определить ожидаемую частоту побочных явлений по итоговым данным таблицы и считать ее одинаково возможной в обеих группах больных. Итак, искомый показатель частоты побочных явлений равен $25/111*100=22.5\%$. При этом условии в группе из 45 больных должно иметь побочные явления $22.5*45/100=10$ человек, а в группе из 66 больных - $22.5*66/100=15$ человек.

Таблица 1. 23– Изучение побочных явлений у больных, леченных разными методами

Метод лечения	Побочные явления		Итого
	возникали	не возникали	
Антибиотики	16	29	45
Антибиотики и витамины	9	57	66
Всего:	25	86	111

Следовательно, побочные явления не наблюдались у 35 больных, леченных только антибиотиками (45-10=35) и у 51 больного, леченного кроме антибиотиков и витаминами (66 - 15 = 51).

Рассчитанные таким образом ожидаемые значения запишем в основной четырёхпольной таблице (в скобках). Далее следует вычислить χ^2 и проверить существенны или нет различия между фактическими и теоретически полученными ожидаемыми величинами (табл.1.24).

Таблица 1.24 – Сравнение фактических и ожидаемых чисел побочных явлений у больных, леченных двумя методами

Метод лечения	Побочные явления		Итого
	возникали	не возникали	
Антибиотики	16 (10)	29 (35)	45
Антибиотики и витамины	9 (15)	57 (51)	66
Всего:	25	86	111

$$\chi 2=\sum\frac{(\hat{O}-\hat{I}\,)^2}{\hat{I}}=\frac{(16-10)^2}{10}+\frac{(9-15)^2}{15}+\frac{(29-35)^2}{35}+\frac{(57-51)^2}{51}=7{,}7$$

Сравним полученный результат с табличными значениями χ^2.Одной степени свободы соответствуют: $\chi^2_{0.5}=3.84, \chi^2_{0.1}=6.63$. Вычисленная величина χ^2 (7,7) превышает табличные значение $x_{01}{}^2$ (6,63) что означает возможность отвергнуть нулевую, гипотезу и считать, что различие в частоте побочных явлений у больных сравниваемых групп не случайно. Этот вывод можно сделать с надежностью, превышающей 99%.

Результат применения критерия χ2 позволяет исследователю рекомендовать витамины для предупреждения побочного действия антибиотиков.

1.7..2 Определение соответствия признаков альтернативных явлений

При изучении альтернативных явлений (лечебный эффект от применения метода лечения достигнут или не достигнут, побочные явления при применении нового лекарства наблюдались или не наблюдались и т.д.), когда результаты исследования могут быть представлены в виде четырехпольной таблицы (2 строки × 2 столбца), расчет χ^2 производится по формуле:

$$\chi^2 = \frac{(ad - bc)^2 \cdot n}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)} \qquad (1.\ 27)$$

Буквенные обозначения величин, входящих в формулу, определяются в соответствии со схемой четырехпольной таблицы (табл. 1.25).

Таблица 1.25 – Схема четырехпольной таблицы

	+	–	Итого
+	a	B	a+b
–	c	D	c+d
	a+c	b+d	a+b+c+d=n

Пример. Исследователем изучалась частота побочных явлений при лечении антибиотиками в сочетании с различными витаминами. По результатам исследований выдвинута нулевая гипотеза о том, что вид применяемых витаминов не оказывает влияния на частоту побочных явлений. Результаты исследования представлены в виде четырехпольной таблицы (табл. 1.26).

Таблица 1.26 – Число побочных явлений у больных, получавших разные витамины

Больные получали	Побочные явления		Итого
	возникали	не возникали	
Антибиотики +витам ины С и B_1	5	26	31
Антибиотики +витам ины С , B_1 и РР	4	31	35
Всего	9	57	66

Расчет по этим данным χ2 указанным методом дает следующий результат:

$$\chi^2 = \frac{(5\cdot 31 - 26\cdot 4)^2 \cdot 66}{31\cdot 35\cdot 9\cdot 57} = 0{,}31$$

Табличные значения критерия для данного случая при *ν*=(2-1)*(2-1)=1 равны χ^2_{01}**=6,33** и χ^2_{05}**=3,84**. Малая величина рассчитанного критерия χ2 не дает права отвергнуть нулевую гипотезу. Различия в частоте побочных явлений не доказаны.

Если число наблюдений хотя бы в одной клетке четырехпольной таблицы <4, то рекомендуется использовать формулу:

$$\chi^2 = \frac{(ad - bc - n/2)^2 \cdot n}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)} \qquad (1.28)$$

В тех случаях, когда в четырехпольной таблице общее число наблюдений <30 или число наблюдений хотя бы в одной клетке таблицы <4, правильнее иє пользовать точный критерий Фишера, определяемый по формуле:

$$p = \frac{(a+b)!\cdot(c+d)!\cdot(a+c)!\cdot(b+d)!}{n!\cdot a!\cdot b!\cdot c!\cdot d!} \qquad (1.29)$$

Нулевая гипотеза отвергается, если p < 0,01.Если полученная величина p ≥ 0,05, то принимается нулевая гипотеза.

1.7.3 Определение критерия χ2 по данным, представленным в сложных таблицах

Методика определения χ2 в более сложных таблицах (3×2; 4×2; 5×2 и т. д) принципиально такая же как и в четырехпольной таблице, но требует обязательного расчета ожидаемых величин.

Пример. В таблице 1.27 приведены данные исследователей изучавших частоту кашлевой реакции на вдыхание 1% аэрозоля ацетилхолина у рабочих и работниц углеобогатительной фабрики.

Таблица 1.27 - Частота кашлевой реакции в зависимости от стажа работы

Стаж работы в условиях запыленного воздуха	Всего исследуемых	Число рабочих, у которых наблюдалась кашлевая реакция	
		Абс. число	%
До 1 года	12	1	8,3
1,1-5 лет	20	2	10,0
5,1-10 лет	27	9	33,3
10,1-15 лет	27	8	29,5
15,1 года и более	10	4	40,0
Всего	96	24	25,0

Данные таблицы, казалось бы, дают право сделать вывод о том, что у лиц, более длительно работающих в условиях запыленного воздуха, кашлевая реакция проявляется значительно чаще. Проверим это заключение, с помощью критерия $\chi 2$.

В качестве нулевой гипотезы выскажем предположение о независимости частоты кашлевой реакции от стажа работы в запыленном воздухе, и, следовательно, о случайности различий этих показателей у рабочих разных по стажу групп. Исходя из этого, будем считать, что при большом числе наблюдений частота кашлевой реакций во всех группах должна быть одинакова и равняться 25% т.е. величине показателя, полученного по итоговым данным.

Основанный на таком предположении расчет ожидаемых чисел показан в табл. 1.28.

Таблица 1.28 - Сравнение фактических и ожидаемых частот кашлевой

Стаж работы в условиях запыленного воздуха	Всего исследуемых	Фактические величины				Ожидаемые величины	
		Кашлевая реакция				Кашлевая реакция	
		Наблюдалась		Не наблюдалась		Наблюдалась	Не наблюдалась
		%	Абс. число	%	Абс. число		
До 1 года	12	8,3	1	91,7	11	$\frac{25,0 \cdot 12}{100} = 3,0$	12-3,0=9,0
1,1-5 лет	20	10,0	2	90,0	18	$\frac{25,0 \cdot 20}{100} = 5,0$	20-5,0=15,0
5,1-10 лет	27	33,3	9	66,7	18	$\frac{25,0 \cdot 27}{100} = 6,8$	27-6,8=20,2
10,1-15 лет	27	29,5	8	70,5	19	$\frac{25,0 \cdot 27}{100} = 6,8$	27-6,8=20,2
15,1 года и более	10	40,0	4	60,0	6	$\frac{25,0 \cdot 10}{100} = 2,5$	10-2,5=7,5
Всего	96	25,0	24	75,0	72	≈ 24	≈ 72

Вычисленные таким образом ожидаемые величины дают возможность определить $\chi 2$ по формуле (1. 26).

Результаты всех расчетов, необходимых для определения $\chi 2$, с указанием ожидаемых величин в скобках рядом с фактическими, показаны в табл. 1.29.

Для оценки полученной величины $\chi 2$ определим число степеней свободы согласно приведённой выше формуле: $v = (5-1)*(2-1) = 4$. Этому числу степеней свободы соответствуют критические значения $\chi 2$:

$$\chi 2_{05} = 9,49 \text{ и } \chi 2_{01} = 13,28$$

Величина χ2 =6,6 намного меньше табличных значений, и это не дает права отвергнуть нулевую гипотезу. Различия частоты кашлевой реакции у рабочих, отличающихся по стажу работы в условиях запыленного воздуха, остаются не доказанными.

Таблица 1.29 – Определение соответствия фактических и ожидаемых частот кашлевой реакции

Стаж работы в условиях запыленного воздуха	Кашлевая реакция		Разности между фактическими и ожидаемыми значениями (Ф—О)		$(Ф-О)^2$		$\frac{(Ф-О)^2}{О}$	
	наблюдалась	не наблюдалась						
До 1 года	1(3,0)	11(9,0)	2	2	4	4	1,33	0,44
1, 1=5 лет	2(5,0)	18(15,0)	3	3	9	9	1,80	0,60
5, 1=10 лет	9(6,8)	18(20,2)	2,2	2,2	4,84	4,84	0,71	0,24
10, 1=15 лет	8(6,8)	19(20,2)	1,2	1,2	1,44	1,44	0,21	0,07
15, 1 года и более .	4(2,5)	6(7,5)	1,5	1,5	2,25	2,25	0,90	0,3
Всего	24	72					$\chi^2 = \Sigma \frac{(Ф-О)^2}{О} = 6{,}60$	

Если метод χ2 используется при сопоставлении рядов измерений длины, веса, времени и т. д., то расчеты χ2 следует вести по формуле:

$$\chi^2 = \frac{1}{n_1 \cdot n_2} \Sigma \frac{(f_1 n_2 - f_2 n_1)^2}{f_1 + f_2} \qquad (1.30)$$

где f_1 и f_2 — частоты сравниваемых вариационных рядов;

n_1 и n_2 –с оответствующие числа наблюдений.

Пример. Сопоставим два ряда данных, полученных у онкологических больных, оперированных по поводу опухоли брюшной полости в горизонтальном положении и положении Тренделенбурга (табл. 1.30). Исследователь изучал влияние искусственной вентиляции легких на размеры венозного давления. Для этого больным измеряли венозное давление (в мм вод. ст.) перед операцией (до наркоза) и через 15 мин после операции. Было установлено, что венозное давление до операции у больных обеих групп было в среднем одинаковым. Через 15 мин после операции венозное давление заметно изменилось: более значительно у больных, оперированных в положении Тренделенбурга, и меньше у больных, оперированных в горизонтальном положении. Для доказательства

этих различий применим метод χ2. Последовательность расчёта критерия χ2 представлена в табл. 1.30.

Таблица 1.30 – Оценка соответствия разности венозного давления у больных, оперированных в разном положении

Разности венозного давления в *мм вод. ст.* (середины интервала – *X)*	Число больных оперированных		f_1+f_2	f_1*n_2	f_2*n_1	$f_1*n_2 - f_2*n_1$	$(f_1*n_2 - f_2*n_1)^2$	$\frac{(f_1*n_2 - f_2*n_1)^2}{f_1+f_2}$
	в горизонтальном положении, f_1	в положении Тренделенбурга, f_2						
5	7	4	11	455	240	215	46225	4202,3
15	12	2	14	780	120	660	435600	31114,3
25	15	4	19	975	240	735	540225	28432,9
35	20	5	22	1300	300	1000	1000000	40000,0
45	6	5	11	390	300	90	-8100	736,4
55		19	19		1140	-1140	1299600	68400,0
65		15	15		900	-900	810000	54000,0
75		8	8		480	-480	230000	28800,0
85		3	3		180	-180	324000	10800,0
	n_1=60	n_2=65						Сумма=266485,9

Значение критерия $\chi 2$ при числе степеней свободы v=9-1=8 будет равно

$$\chi^2 = \frac{1}{60 \cdot 65} \cdot 266485{,}9 = \frac{266485{,}9}{3900} = 68{,}33;$$

Так как табличные значения $\chi 2$ ($\chi 2_{05}$= 15,51; $\chi 2_{01}$= 20,09) много меньше найденной величины (68,33), то нулевая гипотеза предполагавшая отсутствие различий в повышении венозного давления у больных, оперированных в разных положениях, отвергается с большой степенью надежности: риск ошибки не превышает 1%; (p< 0,01).

1.7.4 Проверка соответствия фактических частот вариационного ряда теоретическому распределению

Метод χ^2 может использоваться также для проверки соответствия фактических, полученных в опыте, частот вариационного ряда теоретическому распределению. Это важно для распознавания характера распределения (нормальное, биномиальное, Пуассона) значений изучаемого признака и выбора методов последующей статистической обработки.

Число степеней свободы во всех случаях, когда предполагается нормальное распределение, равняется *v = число строк – 3*, для биномиального распределения и распределения Пуассона *v = число строк – 2*.

Методику проверки соответствия фактических частот вариационного ряда теоретическому распределению рассмотрим на примере проверки соответствия содержания цинка в сыворотке крови нормальному закону распределения, приведенных в табл.1.31 .

Вначале выпишем фактические и теоретические частоты содержания цинка в сыворотке крови и определим χ^2. Результаты расчетов представлены в табл.1.31.

Таблица 1.31 - Определение соответствия частот, полученных в опыте, нормальному распределению

Содержание цинка в сыворотке крови в физиологических условиях (мкг %)	Фактические эмпирические частоты, *Ф*	Теоретические (ожидаемые) частоты, *О*	*Ф-О*	$(Ф-О)^2$	$\frac{(Ф-О)^2}{О}$
75-84	2	0,9	1,1	1,21	1,34
85-94	5	3,7	1,3	1,69	0,46
95-104	7	10,6	3,6	12,96	1,22
105-114	15	20,3	5,3	28,09	1,38
115-124	35	25,0	10,0	100,00	4,00
125-134	17	20,3	3,3	10,89	0,54
135-144	11	10,6	0,4	0,16	0,01

145-154 155-164	3 1	3,7 0,9	0,7 0,1	0,49 0,01	0,13 0,01
	n=96	n=96			χ^2=9,09

Число степеней свободы в нашем примере ν = 9 - 3 = 6. Этому числу степеней свободы соответствуют значения: $\chi 2_{05}$ = 2,59 и $\chi 2_{01}$= 16,81. Найденная величин $\chi 2$ (9,09) меньше табличных значений, следовательно принимается нулевая гипотеза. Существенные различия между фактическими и теоретическими частотами отсутствуют. Эмпирически полученный вариационный ряд соответствует нормальному распределению.

При определении соответствия эмпирического распределения теоретическому следует обращать внимание на крайние частоты теоретического ряда. Минимально допустимые размеры этих частот зависят от числа степеней свободы (табл. 1.32).

Таблица 1.32 – Минимально допустимые значения теоретических частот

Число степеней свободы, ν	1	2	3-6	> 6
Минимально допустимые размеры теоретических частот	4	2	1	0,5

1.8 Корреляционный анализ

Одной из важных задач исследовательской работы является выявление и измерение связи между признаками, характеризующими изучаемые явления или процессы. Различают *функциональную* и *корреляционную* связи.

При наличии *функциональной* связи изменение величины одного признака неизбежно вызывает совершенно определенные изменения величины другого признака. Примером такой связи может служить зависимость площади круга от его радиуса. Функциональная связь между явлениями присуща неживой природе В биологических науках чаще приходится иметь дело с иной связью между явлениями, когда одной и той же величине одного признака соответствует ряд варьирующих значений другого признака, что обусловлено чрезвычайным многообразием взаимодействия различных явлений живой природы. Такого рода связь носит название *корреляционной* (correlation—соответствие, соотносительность). В то время как функциональная связь имеет место в каждом отдельном наблюдении, корреляционная связь проявляется только при многочисленном сопоставлении признаков.

Рассмотрим, например, связь между возрастом детей дошкольников и их ростом (табл. 1.33). Из приведенных данных видно, что с возрастом рост детей увеличивается, и поэтому можно предположить наличие связи между указанными признаками.

Таблица 1.33 - Рост детей дошкольников разного возраста

Возраст	3 года	4 года	5 лет	6 лет	7 лет
Рост в *см* . .	100,3	102,9	108,1	113,7	118,3
	92,6	100,1	106,8	113,8	119,2
	93,8	101,6	107,8	113,3	119,4
	93,7	98,4	104,6	111,8	116,1
	94,2	99,4	107,4	112,1	

Вместе с тем следует отметить, что одному и тому же возрасту соответствует различный рост детей Это происходит потому, что рост детей определяется не только возрастом: на него влияют многие другие факторы, в том числе условия жизни, питание, занятия физкультурой и др. Таким образом, можно прийти к выводу, что связь между возрастом и ростом детей является *корреляционной*.

Исследователю следует помнить, что обнаружение корреляции между сопоставляемыми явлениями не говорит еще о существовании причинной связи между ними. Для установления последней необходим всесторонний логический

и специальный анализ существа изучаемых процессов. Статистический же метод позволяет обосновать полученные в результате научного исследования выводы о наличии тех или иных связей между явлениями, выделить самые главные из них.

1.8.1 Способы выявления корреляционной связи

Наиболее простым способом выявления корреляционной связи является графический.

Например, в эксперименте на 13 кошках получены следующие данные об интрасклеральном и внутриглазном давлении. Уровень интрасклерального давления (*x*)—19,8 7,8 12,7 13,4 10,3 13,7 16,2 15,4 21,5 8,1 11,7 7,6 6,1. Уровень внутриглазного давления (*y*)— 32,5 16,1 21,3 26,8 23,4 19,7 22,9 22,2 22,6 17,6 14,3 18,6 21,4. Необходимо установить, имеется ли корреляционная связь между этими признаками.

На листе бумаги начертим под прямым углом две оси координат, из которых одна - ось абсцисс - будет соответствовать интрасклеральному давлению (*x*), а другая - ось ординат - внутриглазному давлению (*y*). Тогда каждой паре значащий *x и y* на диаграмме будет соответствовать определенная точка (рис. 8.1).

Полученное на диаграмме скопление точек может быть очерчено *эллипсоидальной замкнутой кривой*, длинная ось которой образует острый угол с осью абсцисс (*x*). При этом наглядно видны взаимоотношения между сопоставляемыми признаками. Преобладающая часть точек располагается вблизи длинной оси эллипса, так как большим значениям признака (*y*) обычно соответствуют большие значения признака (*x*), и наоборот, меньшие - меньшим. Такого рода график носит название графика *корреляционного поля*. *Вытянутый характер кривой*, охватывающей точки корреляционного поля, *и угол с осями графика, близкий к 45°*, указывает на *наличие корреляционной связи* между интрасклеральным и внутриглазным давлением. В том случае, если в результате построения графика окажется, что *длинная ось эллипса параллельна одной из осей координат или скопление точек образует круг*, то можно полагать, что между исследуемыми признаками *связь отсутствует*. В ряде случаев *корреляционное поле* может принимать *дугообразную форму* и тем самым свидетельствовать о возможности *криволинейной связи* между признаками.

При наличии большого числа измерений (несколько десятков и более) для выявления связи между двумя признака ми целесообразно данные *сгруппировать и занести в специальную таблицу*, которую иногда называют *корреляционной решеткой*. Допустим, что для изучения физического развития у 100 школьников были измерены рост и вес. Для того, чтобы на основании этих данных построить корреляционную таблицу, сгруппируем данные о росте ребят (*x*) и запишем их в заголовок горизонтальных строк таблицы, а группировку ве-

са- в заголовок вертикальных столбцов (граф). Затем в каждую клетку на пересечении строк и столбцов запишем число детей, имевших соответствующие величины роста и веса. Например, в клетке на пересечении строки 117,5 - 122,4 *см* и столбца 22,5—25,4 *кг* в табл. 1.34 указаны два школьника, имевших рост и вес в этих пределах. Итоговые строка и столбец покажут распределение обследованных по каждому из признаков отдельно.

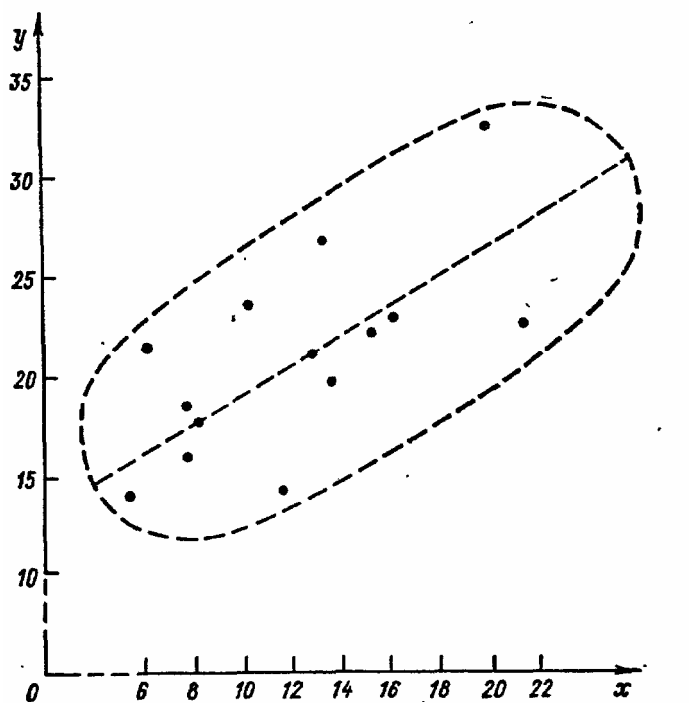

Рис. 8.1. Корреляционное поле

Таблица 1.34 - Распределение школьников по росту и весу

Рост в см. (x)	Вес в кг (y)					Итого
	22,5-25,5	**25,5-28,4**	**28,5-31,4**	**31,5-34,4**	**34,5-37,4**	
117,5-122,4	2	6	-	-	-	8
122,5-127,4	-	4	12	2	-	18
127,5-132,4	-	2	10	10	-	22
132,5-137,4	-	2	12	14	4	32
137,5-142,4	-	-	2	8	4	14
142,5-147,4	-	-	-	2	2	4
147,5-152,4	-	-	-	-	2	2
Всего	2	14	36	36	12	100

По характеру расположения данных, сконцентрированных по диагонали таблицы, можно предположить наличие корреляции между ростом и весом детей

1.8.2 Виды и теснота корреляционной связи

Корреляционная связь может быть *прямолинейной (линейной)* и *криволинейной*. При *прямолинейной корреляции изменение значений одного признака сопровождается равнонаправленным (в сторону увеличения или уменьшения) изменением значений другого признака*. Если же *изменение одного признака приводит к неодинаковым изменениям другого, например, вначале к увеличению, а затем — к уменьшению величин зависимого признака, то такая связь носит название криволинейной*. По форме *линейная* связь между явлениями может быть *прямой (положительной)*, когда с увеличением значений одного признака увеличиваются значения другого, и *обратной (отрицательной)*, когда с увеличением значений одного признака значения другого уменьшаются. Для измерения и оценки связи при прямолинейной корреляции применяется *коэффициент корреляции* (r), при криволинейной корреляции - *корреляционное отношение* (η).

Степень связи между явлениями, ее теснота определяется величиной коэффициента корреляции, который колеблется в пределах от 0 до ±1. При $r = 0$ связь отсутствует, при $r = \pm 1$ — связь полная, функциональная (табл. 1.35).

Таблица 1.35 - Схема оценки тесноты корреляционной связи по коэффициенту корреляции

Теснота связи	Величина коэффициента корреляции при наличии	
	Прямой связи (+)	Обратной связи (-)
Связь отсутствует	0	0
Связь слабая	От 0до +0,3	От0до -0,3
Связь умеренная	От+0,31 до +0,7	От -0,31 до -0,7
Связь сильная	От +0,7 до +1	От -0,7 до -1
Связь полная (функциональная)	+1.0	-1.0

1.8.2 Определение коэффициент корреляции при малом числе наблюдений

При малом числе наблюдений и линейной зависимости между признаками коэффициент корреляции целесообразно рассчитывать, пользуясь следующими формулами:

$$r_{xy} = \frac{\sum (x - \bar{x}) \times (y - \bar{y})}{n \times \sigma_x \times \sigma_y} \quad (1.31)$$

$$r_{xy} = \frac{\sum xy - n \times \bar{x} \times \bar{y}}{n \times \sigma_x \times \sigma_y} \quad (1.32)$$

где r_{xy} — коэффициент линейной корреляции;

x и y — коррелируемые (сопоставляемые) величины;

$\bar{x}$ и $\bar{x}$ — средние арифметические ряда *x* и ряда *y;*

σ_x, σ_y — средние квадратическое отклонения сопоставляемых рядов;

n — число сравниваемых пар.

Использование формулы (1.32) предпочтительнее, так как не требует определения отклонений вариант от средних. В этом случае среднее квадратическое отклонение в каждом ряду следует вычислять по формуле:

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} \quad (1.33)$$

Допустим, необходимо определить корреляционную связь между интрасклеральным и внутриглазным давлением на основании проведенных измерений у 13 кошек (табл. 1.36).

Полученный коэффициент корреляции свидетельствует о наличии между интрасклеральным и внутриглазным давлением у кошек исследованной группы прямой умеренной связи.

1.8.3 Определение коэффициент корреляции при большом числе наблюдений

Приведенные формулы удобны для расчета коэффициентов корреляции при небольшом числе наблюдений (обычно меньше 30—50). Если число наблюдений велико то для вычисления коэффициента корреляции целесообразно сначала построить корреляционную таблицу. При этом данные наблюдений, размещенные в таблице, должны быть сгруппированы.

Например, необходимо установить, имеется ли связь между количеством нейтрофилов и общим числом лейкоцитов у 142 обследованных детей. Распределение их по числу лейкоцитов и нейтрофилов представлено в табл. 1.37.

В графе 12 таблицы указаны частоты *(f_y)* признака *y* (нейтрофилез), в строке 16—частоты *(f_x)* второго признака *x* (число лейкоцитов). В клетке на пересечении графы f_y и строки f_x приведена общая сумма наблюдений, равная в нашем

Таблица 1.36 - Вычисление коэффициента корреляции по не сгруппированным данным

Номер опыта	Интрасклеральное венозное давление(x)	Внутриглазное давление (y)	x^2	y^2	xy
1	19,8	32,5	392,04	1056,25	643,50
2	7,8	16,1	60,84	259,21	125,58
3	12,7	21,3	161,26	453,69	270,51
4	13,4	26,8	179,56	718,24	359,12
5	10,3	23,4	106,09	547,66	241,02
6	13,7	19,7	187,69	388,09	269,89
7	16,2	22,9	262,44	524,41	370,98
8	15,4	22,2	237,16	492,84	341,88
9	21,5	22,6	462,25	510,76	485,90
10	8,1	17,6	65,61	309,76	142,56
11	11,7	14,3	136,89	204,49	167,31
12	7,6	18,6	57,76	345,96	141,36
13	6,1	21,4	37,21	457,96	130,54
n=13	$\sum x$ = 164,3	$\sum y$ = 279,4	$\sum x^2$=2346,83	$\sum y^2$=6269,22	$\sum xy$=3690,15
	$\bar{x}$= 12,64	$\bar{y}$=21,49	σ_x = 4,74	σ_y = 4,70	

$$r_{xy} = \frac{\sum xy - n \times \bar{x} \times \bar{y}}{n \times \sigma_x \times \sigma_y} = \frac{3690{,}15 - 13 \cdot 12{,}64 \cdot 21{,}49}{13 \cdot 4{,}74 \cdot 4{,}70} = +0{,}55$$

примере 142. Расчет коэффициента корреляции производится по несколько видоизмененной основной формуле:

$$r_{xy} = \frac{\sum(\sum f_{xy} a_x) a_y - n \frac{\sum f_x a_x}{n} \times \frac{\sum f_y a_y}{n}}{n \times \sigma_x^1 \times \sigma_y^1} \quad (1.34)$$

$$\sigma_x^1 = \sqrt{\frac{\sum f_x a_x^2}{n} - \left(\frac{\sum f_x a_x}{n}\right)^2} \quad (1.35)$$

$$\sigma_y^1 = \sqrt{\frac{\sum f_y a_y^2}{n} - \left(\frac{\sum f_y a_y}{n}\right)^2} \quad (1.36)$$

где: f_x— итоговые числа наблюдений в отдельных графах ряда x;

f_y—итоговые числа наблюдений в отдельных строках ряда *y;*

f_{xy} ~ числа наблюдений в клетках таблицы на пересечении граф *x* и строк *y;*

a_x и a_y — отклонения от условных средних рядов *x* и *y* в единицах их интервалов;

n — общее число наблюдений;

σ_x^1 и σ_y^1 — средние квадратическия отклонения рядов *x* и *y* в единицах интервалов.

Таблица 1.37 - Определение коэффициента корреляции между нейтрофилезом и лейкоцитозом у детей

Нейрофилы (в тыс.), *y*	Строки	Лейкоциты (в тыс.) *x*											t_y	a_y	$f_y a_y$	$f_y a^2{}_y$	$\Sigma f_{xy} a_y$	$(\Sigma f_{xy} a_x) a_y$
		4—5.9	6—7.9	8—9.9	**10—11.9**	12—13.9	14—15.9	16—17.9	18—19.9	20—21.9	22—23.9	24—25.9						
		графы																
		1	2	3	**4**	5	6	7	8	9	10	11	12	13	14	15	16	17
0 – 0,9	1	3	5	2	**7**	1	0	1					19	-3	-57	171	-17	+51
1,0 – 1,9	2	6	9	7	**4**	1	0						27	-2	-54	108	-42	+84
2 – 2,9	3		3	10	**14**	8	3	3					41	-1	-41	41	+7	-7
3 – 3,9	**4**	**1**	**3**	**2**	**2**	**5**	**1**						**18**	**0**	**0**	**0**	**+5**	**0**
4 -- 4,9	5		1	2	**2**	441	4						13	1	13	13	+8	8
5 – 5,9	6			1	**1**		1	2	1				10	2	20	40	+15	30
6 –6,9	7					1	1	11	2	1			6	3	18	54	+19	57
7 – 7,9	8				**1**		1						3	4	12	48	+5	20
8 – 8,9	9									1			2	5	10	50	+11	55
9 – 9,9	10										1		2	6	12	72	+7	42
10 –10,9	11										1		0	7	0	0	0	0
11 – 11,9	12												0	8	0	0	0	0
12 – 12,9	13												0	9	0	0	0	0
13 – 13,9	14												0	10	0	0	0	0
14 –14,9	15											1	1	11	11	121	7	77
f_x	16	10	21	24	**31**	25	14	9	3	2	2	1	142		Σ=-56	Σ=718		Σ=417
a_x	17	-3	-2	-1	**0**	1	2	3	4	5	6	7	-					
$f_x a_x$	18	-30	-42	-24	**0**	25	28	27	12	10	12	7	Σ=25					
$a^2{}_x$	19	9	4	1	**0**	1	4	9	11	25	36	49						
f_x x $a^2{}_x$	20	90	84	24	**0**	25	56	81	48	50	72	49	Σ=579					

Пользуясь приведенной формулой, рассчитываем коэффициент корреляции для нашего примера. Прежде всего необходимо определить отклонения каждой варианты в своем ряду от условного среднего значения в единицах интервалов. Для этого произвольно за условную среднюю ряда *x* примем варианту 10—11,9, а в ряду *y* - варианту 3—3,9. Для большей наглядности выделим соответствующие графу и строку полужирным шрифтом. Далее для каждого ряда запишем отклонения (a_x и a_y) вариант от условных средних арифметических величин в единицах интервалов (при этом все разности между соседними групповыми вариантами условно принимаем равными единице). В таком случае отклонения в ряду *x* (строка 17) будут следующими: —3, —2, —1, 0, 1, 2, 3, 4, 5, 6, 7; а в ряду *y* (графа 13): —3, —2, —1, 0, 1, 2,3,4,5,6, 7,8,9, 10, 11.

Для того, чтобы найти средние квадратические отклонения в единицах интервалов по приведенной выше формуле, необходимо вычислить сумму произведений отклонений на соответствующие частоты—$\Sigma f_x a_x$ и $\Sigma f_y a_y$:

$\Sigma f_x a_x$=10 (-3)+21 (-2)+24(-l)+31*0+ ... ...+1·7=25;

$\Sigma f_y a_y$ ==19·(—3)+27·(—2)+41 (—1) +...+11 l == —56.

Далее определяем сумму произведений квадратов отклонений на соответствующие частоты:

$\Sigma f_x a_x^2$ =-9·10+4·21 +1·24+ ... +49·1=579;

$\Sigma f_y a_y^2$ = 9·19+4·27+1·41+0·18+ ... +121·1=718

Промежуточные и окончательные результаты подсчета по этим формулам записаны в графе 15 и строке 20 таблицы.

Вычисляем значения σ_x^1 и σ_y^1:

$$\sigma_x^1 = \sqrt{\frac{\Sigma f_x \cdot a_x^2}{n} - \left(\frac{\Sigma f_x \cdot a_x}{n}\right)^2} = \sqrt{\frac{579}{142} - \left(\frac{25}{142}\right)^2} = 2{,}0$$

$$\sigma_y^1 = \sqrt{\frac{\Sigma f_y \cdot a_y^2}{n} - \left(\frac{\Sigma f_y \cdot a_y}{n}\right)^2} = \sqrt{\frac{718}{142} - \left(\frac{-56}{142}\right)^2} = 2{,}21$$

Остается определить величину $\Sigma(\Sigma f_{xy} a_x) a_y$ для числителя формулы коэффициента корреляции. Вычисляем сумму произведений отклонений ряда *x* и ряда *y* на соответствующие частоты, т. е. $\Sigma f_{xy} a_x a_y$. Вначале получаем сумму произведений $\Sigma f_{xy} a_x$ путем перемножения чисел, стоящих в клетках таблицы и представляющих из себя частоты совмещенных значений *x* и *y* (f_{xy}) на величины отклонений вариант ряда *x* от условной средней, т. е. на a_x. Эти произведения суммируем для каждой горизонтальной строки и заносим в графу 16. Например, для первой строки

$\Sigma f_{xy} a_x$ =3·(-3)+5·(-2)+2·(-l)+7·(0) + 1-1+0·2+1·3=(—9)+(—10)+(-2)+1 +3 = -17

Затем каждую из величин $\Sigma f_{xy}a_x$ перемножаем на условные отклонения ряда *y* для данной строки (a_y) и результаты записываем в графу 17. Суммируя полученные произведения, получаем величину $\Sigma(\Sigma f_{xy}a_x)a_y$=417.

Теперь в нашем распоряжении имеются все необходимые величины для вычисления *r* по формуле:

$$r_{xy} = \frac{\Sigma(\Sigma f_{xy} \cdot a_x)a_y - n\frac{\Sigma f_x a_x}{n} \cdot \frac{\Sigma f_y \cdot a_y}{n}}{n\sigma'_x\sigma'_y} =$$

$$= \frac{417 - 142 \cdot \frac{25}{142} \cdot \frac{(-56)}{142}}{142 \cdot 2{,}01 \cdot 2{,}21} = \frac{417 + 9{,}86}{630{,}78} = +0{,}68.$$

Величина полученного коэффициента корреляции говорит об умеренной тесноте связи исследованных признаков, а знак свидетельствует о прямом характере этой связи.

Иногда при наличии линейной связи можно, используя коэффициент корреляции, оценить влияние признака-фактора на результативный признак. Для этого применяется квадрат коэффициента корреляции, называемый *коэффициентом детерминации* (r^2). Если выразить коэффициент детерминации в процентах, то он покажет долю влияния данного факториального признака на результативный. Например, коэффициент корреляции между ростом и весом детей равен +0,75, тогда коэффициент детерминации будет: r_{xy}^2=0,75^2=0,56. Если принять все факторы, влияющие на вес тела, за 100%, то на долю роста приходится 56%.

1.8.4 Средняя ошибка коэффициента корреляции

Поскольку коэффициент корреляции в клинических исследованиях рассчитывается обычно для ограниченного числа наблюдений, нередко возникает вопрос о надежности полученного коэффициента С этой целью определяют *среднюю ошибку коэффициента корреляции*. При достаточно большом числе наблюдений (больше 100) средняя ошибка коэффициента корреляции *(m_r)* вычисляется по формуле:

$$m_r = \frac{1 - r_{xy}^2}{\sqrt{n}} \qquad (1.37)$$

где *n* — число парных наблюдений.

В том случае, если число наблюдений меньше 100, но больше 30, точнее определять среднюю ошибку коэффициента корреляции, пользуясь формулой:

$$m_r = \frac{1 - r_{xy}^2}{\sqrt{n-1}} \qquad (1.38)$$

С достаточной для медицинских исследований надежностью о наличии той или иной степени связи можно утверждать только тогда, когда величина коэффициента корреляции превышает или равняется величине трех своих ошибок *(r_{xy}>=3m_r)*. Обычно это отношение коэффициента корреляции *(r_{xy})* к его средней ошибке *(m_r)* обозначают буквой *t* и называют критерием достоверности:

$$t_r = \frac{r_{xy}}{m_r} \tag{1.39}$$

Если t_r >= 3, то коэффициент корреляции достоверен. В рассмотренном выше примере число наблюдений 142, а коэффициент корреляции 0,68. Тогда

$$m_r = \frac{1 - r_{xy}^2}{\sqrt{n}} = \frac{1 - (0{,}68)^2}{\sqrt{142}} = 0{,}045$$

$$t_r = \frac{r}{m_r} = \frac{0{,}68}{0{,}045} = 15,$$

т. е. коэффициент корреляции вполне достоверен.

В случае малой выборки (число наблюдений меньше 30) для оценки достоверности коэффициента корреляции, т. е. для определения соответствия коэффициента корреляции, вычисленного по выборочным данным, действительным размерам связи в генеральной совокупности, средняя ошибка коэффициента корреляции *(m_r)* определяется по формуле:

$$m_r = \frac{\sqrt{1 - r_{xy}^2}}{\sqrt{n-2}} \tag{1.40}$$

Значения критерия t_r оцениваются по таблице *t* Стьюдента при числе степеней свободы $v = n - 2$. Если величина t_r больше табличного значения t_{05}, то коэффициент корреляции признается надежным с доверительной вероятностью больше 95%. Например, имеется коэффициент корреляции, равный +0,72 при числе наблюдений 28. Тогда

$$m_r = \sqrt{\frac{1 - (0{,}72)^2}{28 - 2}} = \pm 0{,}019$$

$$t_r = \frac{0{,}72}{0{,}019} = 35{,}9$$

Полученное значение t_r = 35,9 значительно больше табличного t_{01} = 2,779, следовательно, полученному коэффициенту корреляции можно доверять с высокой степенью вероятности (>99%).

В медицинской практике нередко возникает необходимость сравнить между собой два выборочных коэффициента корреляции и определить, существенна ли разница между ними. Ввиду того, что распределение коэффициента корреляции отличается от нормального, для оценки значимости различия между двумя коэффициентами корреляции рекомендуется использовать величину Z,

предложенную Р. Фишером. Величины Z, соответствующие различным значениям коэффициента корреляции, представлены в табл. 1.38.

Например, при исследовании тесноты связи между ростом и весом девочек и мальчиков было установлено, что у мальчиков коэффициент корреляции равен 0,5, а у девочек — 0,7. При этом обследовано 20 мальчиков и 30 девочек. Можно ли считать, что у девочек сильнее выражена связь между ростом и весом, чем у мальчиков? Для решения этого вопроса переведем значение наших коэффициентов корреляции (r) в величины Z. Находим по таблице, что r = 0,5 соответствует Z = 0,5493, а для r = 0,7 соответствует Z = 0,8673. Ошибка разности вычисляется по формуле:

$$m_z = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} = \sqrt{\frac{1}{20 - 3} + \frac{1}{30 - 3}} = \sqrt{0{,}10} = 0{,}316$$

Вычисляем критерий значимости различий:

$$t_z = \frac{z_1 - z_2}{m_z} = \frac{0{,}8673 - 0{,}5493}{0{,}316} = \frac{0{,}3188}{0{,}316} = 1{,}005$$

Разность признается значимой, если $t_z \geq 3$. В нашем примере $t_z < 3$; следовательно, на основании полученных коэффициентов корреляции нельзя делать вывод о более выраженной связи между ростом и весом у девочек.

Таблица 1.38 - Таблица величин Z

r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,90	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467
0,80	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3101	1,3758	1,4219
0,70	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,60	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7623	0,8107	0,8291	0,8480
0,50	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,40	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,30	0,3045	0,3205	0,3316	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,20	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,10	0,1003	0,1104	0,1205	0,1307	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,00	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0600	0,0701	0,0802	0,0902

1.8.5 Определение тесноты связи между качественными признаками

При изучении зависимости *качественных признаков* используется *коэффициент сопряженности*. Для определения тесноты связи в случае альтернативной изменчивости двух сопоставляемых признаков имеющиеся данные сводятся в четырехпольную таблицу, и коэффициент сопряженности вычисляется по формуле:

$$C_1 = \frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}} \qquad (1.41)$$

Если ранее по данным этой таблицы был вычислен критерий χ^2, то коэффициент сопряженности вычисляется по формуле:

$$C_1 = \sqrt{\frac{\chi^2}{a+b+c+d}} = \sqrt{\frac{\chi^2}{n}} \qquad (1.42)$$

Например, требуется установить, имеется ли связь между степенью тяжести ревматизма и эффективностью тонзиллэктомии (табл. 1.39).

Таблица 1.39 - Эффективность тонзиллэктомии в зависимости от симптоматики ревматизма

Симптоматика ревматизма	Результат лечения		Итого
	успешное	неэффективное	
Больные, имевшие изменения со стороны сердца и суставов.		9	26
Больные, имевшие изменения только со стороны сердца . .		8	16
Всего	25	17	42

$$C_1 = \frac{(17 \cdot 8) - (8 \cdot 9)}{25 \cdot 17 \cdot 26 \cdot 16} = \frac{64}{420{,}2} = 0{,}15.$$

Коэффициент сопряженности изменяется в пределах от +1 до -1 и оценивается аналогично коэффициенту корреляции.

При сопоставлении качественных признаков, имеющих три и больше групп, для определения тесноты связи, пользуются *коэффициентом средней квадратичной сопряженности Пирсона*:

$$C_1 = \sqrt{\frac{\phi^2}{1+\phi^2}} \qquad (1.43)$$

и *коэффициентом взаимной сопряженности Чупрова*:

$$K = \sqrt{\frac{\phi^2}{\sqrt{(k_1-1)\cdot(k_2-1)}}} \qquad (1.44)$$

где k_1 — число групп по столбцам;

k_2 — число групп по строкам таблицы;

$\varphi^2 + 1$ — равно о сумме отношений квадратов частот каждой клетки таблицы к произведению итогов строк и соответствующих итогов столбцов

$$\varphi^2 + 1 = \left(\sum \frac{m_{xy}^2}{m_x \cdot m_y} \right) \qquad (1.45)$$

Пример. Вычислим коэффициент средней квадратической сопряженности Пирсона между гистологической структурой и типом роста опухоли по данным таблицы 1.40.

Находим значение $\varphi^2 + 1$:

$$\phi^2 + 1 = \sum \frac{m_{xy}^2}{m_x \cdot m_y} = \frac{11^2}{20 \cdot 21} + \frac{6^2}{33 \cdot 21} + \frac{2^2}{14 \cdot 21} + \frac{2^2}{6 \cdot 21} + \frac{3^2}{20 \cdot 15} + \frac{10^2}{33 \cdot 15} + \frac{1^2}{14 \cdot 15} + \frac{1^2}{6 \cdot 15} + \frac{3^2}{20 \cdot 12} +$$

$$+ \frac{5^2}{33 \cdot 12} + \frac{3^2}{14 \cdot 12} + \frac{1^2}{6 \cdot 12} + \frac{1^2}{20 \cdot 12} + \frac{7^2}{33 \cdot 11} + \frac{3^2}{14 \cdot 11} + \frac{1^2}{33 \cdot 6} + \frac{5^2}{14 \cdot 6} + \frac{2^2}{20 \cdot 8} + \frac{4^2}{33 \cdot 6} + \frac{2^2}{6 \cdot 8} = 1{,}47$$

Отсюда находим $\varphi^2 = 1{,}47 - 1 = 0{,}47$.

Коэффициент средней квадратичной сопряженности Пирсона:

$$C_1 = \sqrt{\frac{0{,}47}{1{,}47}} = 0{,}565$$

Коэффициент взаимной сопряженности Чупрова равняется

$$K = \sqrt{\frac{0{,}47}{\sqrt{(4-1)\cdot(6-1)}}} = \sqrt{\frac{0{,}47}{\sqrt{3 \cdot 5}}} = \sqrt{\frac{0{,}47}{\sqrt{15}}} = \sqrt{0{,}12} = 0{,}348$$

Полученный коэффициент *К* также свидетельствует о наличии связи между рассматриваемыми признаками.

Таблица 1.40 - Зависимость между гистологической структурой опухоли и типом ее роста

Гистологическая структура (*y*)	Тип роста опухоли (*x*)				Итого (m_y)
	экзофитный	язвенно-инфильтративный	диффузно-инфильтративный	переходный	
Аденокарцинома	11(m_{xy})	6	2	2	21
Cr. simplex……	3	10	1	1	15
Солидный рак……	3	5	3	1	12
Слизистый……	1	7	3	—	11
Фиброзный рак…	—	1	5	—	6
Смешанные формы	2	4	—	2	8
Всего (m_x)……	20	33	14	6	73

При применении коэффициента сопряженности C_1 следует учитывать, что он всегда меньше 1 и теоретическая его величина зависит от числа строк и столбцов таблицы. Поэтому вычисление коэффициента C_1 правомочно только тогда, когда каждый из сопоставляемых признаков имеет не менее 5 градаций (таблица 5×5 групп). Коэффициент Чупрова, который всегда меньше коэффициента C_1 не имеет этого ограничения.

Достоверность выборочного коэффициента взаимной сопряженности оценивается с помощью критерия χ^2. Полученная величина $\chi^2 = n\varphi^2$ сопоставляется с табличными значениями χ^2 при числе степеней свободы $v=(k-1)(k_2-1)$ и $p = 0{,}05$.

1.8.6 Множественная корреляция

Приведенные выше коэффициенты корреляции характеризуют связь между двумя признаками. В медицинской практике часто наблюдаются процессы, в которых взаимно связаны не два, а большее число варьирующих признаков. В ряде случаев необходимо установить тесноту связи между двумя признаками при условии, что третий признак не меняется, т. е. при исключении влияния третьего признака. Такая связь оценивается с помощью парциальных (частичных) коэффициентов корреляции при условном допущении постоянства одного из трех коррелируемых признаков по следующим формулам:

$$r_{ab}^{c} = \frac{r_{ab} - r_{ac} \cdot r_{bc}}{\sqrt{(1 - r_{ac}^2) \cdot (1 - r_{bc}^2)}}, \quad (1.46)$$

$$r_{bc}^{a} = \frac{r_{bc} - r_{ab} \cdot r_{ac}}{\sqrt{(1 - r_{ab}^2) \cdot (1 - r_{ac}^2)}}, \quad (1.47)$$

$$r_{ac}^{b} = \frac{r_{ac} - r_{ab} \cdot r_{bc}}{\sqrt{(1 - r_{ab}^2) \cdot (1 - r_{bc}^2)}} \quad (1.48)$$

где r_{ab}^{c} - коэффициент корреляции между признаками *a* и *b*, при исключении влияния признака *c*;

r_{ac}^{b} - коэффициент корреляции между признаками *a* и *c*, при исключении влияния признака *b;*

r_{bc}^{a} - коэффициент корреляции между признаками *b* и *c*, при исключении влияния признака *a*.

Условное обозначение признака, влияние которого элиминируется, выставлено в виде верхнего индекса.

Например, известно, что между возрастом (*a*), ростом *(b)* и весом (*c*) детей существует сильная корреляционная связь. Парные коэффициенты корреляции равны:

$$r_{ab} = +0{,}73,\ r_{bc} = +0{,}82 \text{ и } r_{ac} = +0{,}87$$

Вычислим парциальные коэффициенты корреляции для определения тесноты связи между возрастом и весом при устранении влияния роста:

$$r_{ac}^{b} = \frac{r_{ac} - r_{ab} \cdot r_{bc}}{\sqrt{(1 - r_{ab}^2) \cdot (1 - r_{ac}^2)}} = \frac{0{,}87 - (0{,}73 \cdot 0{,}82)}{\sqrt{(1 - 0{,}73^2) \cdot (1 - 0{,}82^2)}} = \frac{0{,}27}{\sqrt{0{,}31}} = 0{,}48.$$

Парциальный коэффициент корреляции между ростом и весом при элиминировании влияния возраста:

$$r_{bc}^{a} = \frac{r_{bc} - r_{ab} \cdot r_{ac}}{\sqrt{(1 - r_{ab}^2) \cdot (1 - r_{ac}^2)}} = \frac{0{,}82 - 0{,}73 \cdot 0{,}87}{\sqrt{(1 - 0{,}73^2) \cdot (1 - 0{,}87^2)}} = 0{,}59.$$

Вычисленные парциальные коэффициенты корреляции оказались меньше обычных. Проведенный анализ позволил полнее выявить влияние каждого отдельного признака.

1.8.7 Понятие о корреляционном отношении

В случаях нелинейной (криволинейной) корреляции для измерения тесноты связи применяется *корреляционное отношение* (η), которое более точно дает сведения о степени зависимости признаков, причем следует помнить, что при нелинейной связи коэффициент корреляции и корреляционное отношение численно не совпадают друг с другом, при линейной связи они примерно равны между собой. Поэтому нередко, если вид связи не определяется достаточно четко, наряду с коэффициентом корреляции, вычисляют и корреляционное отношение.

В табл. 1.41 представлены результаты одновременного определения содержания сиаловой кислоты и 17-кетостероидов в суточной моче больных. Коэффициент корреляции, вычисленный по этим данным, равен -0,46. Однако характер расположения данных в корреляционной таблице позволяет усомниться в наличии линейной связи между признаками, так как частоты P_{xy} расположены преимущественно в левом верхнем углу таблицы. В этом случае следует вычислить корреляционное отношение.

Таблица 1.41 - Содержание сиаловой кислоты и 17-кетостероидов в суточной моче больных

Содержание 17-кетостероидов в суточной моче в *мг* (*y*)	Середина интервалов	Уровень сиаловой кислоты (*x*)						P_y	$\bar{x}_y$
		90–100	101–140	141–180	181–199	200–210	211–250		
		95	120,5	160,5	190	205	230,5		
2,00–5,00	3,5	–	1	–	–	–	1	2	175,5
5,01–9,99	7,5	1	2	2	1	7	3	16	185,5
10,00–15,00	12,5	3	11	6	2	–	–	22	134,2
15,01–20,00	17,5	2	8	4	–	–	–	14	128,3
20,01–25,99	23,0	2	–	1	–	–	–	3	116,8
26,00–31,99	29,0	1	1	1	–	–	–	3	125,3
P_x		9	23	14	3	7	4	60	146,7
$\bar{y}_x$		17,2	14,1	15,1	10,8	7,5	6,5	13,3	

Корреляционное отношение *y* на *x* определяется по формуле:

$$\eta_{yx} = \sqrt{\frac{\sigma^2_{\bar{y}_x}}{\sigma^2_y}} \tag{1.49}$$

где $\bar{y}_x$ — частная средняя каждой группы;

$\sigma^2_{\bar{y}_x}$ —межгрупповая дисперсия;

σ^2_y — общая дисперсия *y*;

P_x, P_y — частота значений каждой группы.

Межгрупповая дисперсия и общая дисперсия *y* рассчитываются по следующим формулам:

$$\sigma^2_{\bar{y}_ч} = \frac{\sum(\bar{y}_x - \bar{y})^2 P_x}{n}$$
$$\sigma^2_y = \frac{\sum(y - \bar{y})^2 P_y}{n} \tag{1.50}$$

Расчеты указанных дисперсий представлены в таблицах 1.42 и 1.43. Ввиду неодинаковых группировочных интервалов средние величины $\bar{x}$ и $\bar{y}$ вычислялись по формулам:

$$\bar{x} = \frac{\sum(P_x \cdot x_i)}{n}, \quad \bar{y} = \frac{\sum(P_y \cdot y_i)}{n} \tag{1.51}$$

где x_i, y_i — середины интервалов, соответствующие частотам *Px* и *Py*. Например, $\Sigma P_x x_i$=9·95+23·120,5 + 14·160,5 + 3·190 +7·205 + 4·230,5 = 8801,5.

Таблица 1.42-Расчет общей дисперсии признака *y*

y	P_y	$y - \bar{y}$	$(y - \bar{y})^2$	$P_y (y - \bar{y})^2$	
3,5	2	−9,8	96,04	198,08	$\sigma^2_y = \frac{\Sigma_{Py}(y-\bar{y})^2}{n} = \frac{1799,40}{60} = 29,9$
7,5	16	−5,8	33,64	538,24	
12,5	22	−0,8	0,64	14,08	
17,5	14	+4,0	16,00	224,00	
23,0	3	+9,0	81,00	243,00	
29,0	3	+14,0	196,00	588,00	
$\bar{y} = 13,3$	$n = 60$			$\Sigma = 1799,40$	

Таблица 1.43 - Расчет межгрупповой дисперсии признака *y*

$\bar{y}_x$	P_x	$\bar{y}_x - \bar{y}$	$(\bar{y}_x - \bar{y})^2$	$P_x(\bar{y}_x - \bar{y})^2$	
17,2	9	+3,9	15,21	136,89	$\sigma^2_{\bar{y}_x} = \frac{\Sigma P_x(\bar{y}_x - \bar{y})^2}{n} = \frac{636,16}{60} = 10,6$
14,1	23	+0,8	0,61	14,72	
15,1	14	+1,8	3,24	45,36	
10,8	3	−2,5	6,25	18,75	
7,5	7	−5,8	33,64	235,48	
6,5	4	−6,8	46,24	184,96	
$\bar{y} = 13,3$	$n = 60$			$\Sigma = 636,16$	

Корреляционное отношение в этом случае будет равно

$$\eta_{yx} = \sqrt{\frac{\sigma^2_{\bar{y}_x}}{\sigma^2_y}} = \sqrt{\frac{10,6}{29,9}} = \sqrt{0,36} = 0,60.$$

Корреляционное отношение *y* на *x* не совпадает с корреляционным отношением *x* на *y*.

Поэтому следует вычислить и корреляционное отношение *x* на *y*:

$$\eta_{xy} = \sqrt{\frac{\sigma^2_{\bar{x}_y}}{\sigma^2_x}}, \qquad (1.52)$$

где $\bar{x}_y$ - частная средняя каждой группы;

$\sigma^2_{\bar{x}_y}$ -межгрупповая дисперсия;

σ^2_x - общая дисперсия признака *x*.

Межгрупповая дисперсия и общая дисперсия *x* рассчитываются по следующим формулам:

$$\sigma^2_{\bar{x}_y} = \frac{\sum(\bar{x}_y - \bar{x})^2 P_y}{n}$$
$$\sigma^2_x = \frac{\sum(x - \bar{x})^2 P_x}{n} \qquad (1.53)$$

Расчеты указанных дисперсий представлены в таблицах 1.44 и 1.45.

Корреляционное отношение η_{xy} в этом случае будет равно

$$\eta_{xy} = \sqrt{\frac{768,8}{1669,56}} = \sqrt{0,47} = 0,69$$

Полученные корреляционные отношения ($\eta_{yx} = 0,60$, а $\eta_{xy} = 0,69$) не совпадают друг с другом и отличаются от $r = -0,46$, что указывает на нелинейность связи. Вместе с тем величина корреляционного отношения свидетельствует о выраженной связи между изучаемыми признаками.

Таблица 1.44 - Расчет общей дисперсии признака x

x	P_x	$x-\bar{x}$	$(x-\bar{x})^2$	$P_x(x-\bar{x})^2$	
95	9	−51,7	2672,89	24056,01	$\sigma_x^2 = \frac{100173,75}{60} = 1669,56$
120,5	23	−26,2	686,44	15788,12	
160,5	14	+14,2	201,64	2822,96	
190,0	3	+43,3	1874,89	5624,67	
205,0	7	+58,3	3398,89	23792,23	
230,5	4	+83,8	7022,44	28089,76	
$\bar{x} = 146,7$	$n=60$			$\Sigma = 100173,75$	

Таблица 1.45 - Расчет межгрупповой дисперсии признака x

$\bar{x}_y$	P_y	$\bar{x}_y-\bar{x}$	$(\bar{x}_y-\bar{x})^2$	$P_y(\bar{x}_y-\bar{x})^2$	
175,5	2	+28,8	829,44	1658,88	$\sigma^2_{\bar{x}_y} = \frac{45530,05}{60} = 758,8$
185,8	16	+39,1	1977,37	31637,22	
134,2	22	−12,5	156,25	3437,50	
128,3	14	−18,4	338,56	4739,84	
116,8	3	−29,9	894,01	2682,03	
125,3	3	−21,4	457,96	1373,88	
$\bar{x} = 146,7$	$n=60$			$\Sigma = 45530,05$	

Средняя ошибка выборочного коэффициента корреляционного отношения определяется по формуле:

$$m_\eta = \frac{1-\eta^2}{\sqrt{n}}; \qquad (1.54)$$

Оценка этой ошибки производится так же, как и коэффициента корреляции.

1.9 Основы регрессионного анализа

В медицинских исследованиях подчас требуется вычислить не только меру связи между двумя явлениями, но определить характер изменения одной величины от изменений другой. Для этого определяют *коэффициент регрессии*, который рассчитывают при линейной корреляции по следующим формулам:

$$R_{x/y} = r_{xy} \frac{\sigma_x}{\sigma_y} \tag{1.55}$$

$$R_{y/x} = r_{xy} \frac{\sigma_y}{\sigma_x} \tag{1.56}$$

Первое уравнение и полученный из него коэффициент регрессии показывает, насколько изменяется в среднем признак x при изменении признака y на какую-либо величину; второе уравнение показывает обратные взаимоотношения, т. е. как изменяется в среднем признак y при изменении признака x.

Например, имеются следующие данные: коэффициент корреляции (r_{xy}) между интрасклеральным венозным давлением (x) и внутриглазным давлением (y) равен 0,719; x = 9,97 и σ_x = ±4,52; y =18,4 и σ_y = ±5,7. Отсюда:

$$R_{x/y} = 0{,}719 \cdot \frac{4{,}52}{5{,}7} = 0{,}57$$

$$R_{y/x} = 0{,}719 \cdot \frac{5{,}7}{4{,}52} = 0{,}94$$

Следовательно, с увеличением интрасклерального венозного давления на единицу внутриглазное давление в среднем увеличивается на 0,94, а с увеличением внутриглазного давления на единицу интрасклеральное в среднем увеличивается на 0,57.

Зная коэффициент регрессии, можно рассчитать любые значения y при разных значениях x и определить x при любых значениях y. Этим целям служат уравнения регрессии при прямолинейной связи:

$$y_i - \bar{y} = R_{y/x}(x_i - \bar{x}) \tag{1.57}$$

$$x_i - \bar{x} = R_{x/y}(y_i - \bar{y}) \tag{1.58}$$

Уравнение (1.57) представляет уравнение регрессии y по ***x***.

Уравнение (1.58) представляет уравнение регрессии x по y.

Рассчитаем в качестве примера ожидаемые числа внутриглазного давления (y) при заданных значениях (x) — интрасклерального венозного давления и сравним полученные теоретические величины внутриглазного давления с фактическими.

Подставим в формулу (1.57) числовые значения и получим:

$$y_i - 18{,}4 = 0{,}94\,(x_i - 9{,}97);$$

$$y_i - 18{,}4 = 0{,}94x_i - 9{,}36;$$

$$y_i = 0{,}94x_i + 9{,}04.$$

Решаем наше уравнение регрессии, подставляя различные значения x (табл. 1.46).

Таблица 1.46 – Результаты вычислений по уравнению регрессии

Интрасклеральное венозное давление (x)	Внутриглазное давление (y)
x_1=19,8	y_1= 0,94·19,8 +9,04 =27,65
x_2=7,8	y_2 =0,94·7,8+9,04= 16,37
x_3=12,7	y_3=0,94·12,7+9,04=20,98
x_4=13,4	y_4 =0,94·13,4 +9,04 =21,63

Сопоставим полученные данные с фактическими (табл. 1.47).

Таблица 1.47 – Внутриглазное давление при разных уровнях интрасклерального венозного давления

Интрасклеральное венозное давление в опыте	Внутриглазное давление в опыте	Внутриглазное давление, вычисленное по формуле	Разность между опытными и вычисленными по формуле величинами
1	2	3	4
19,8	34,9	27,65	+7,25
13,4	26,8	21,63	+5,17
12,7	21,3	20,98	+0,32
7,8	16,1	16,37	-0,27

Полученный ряд (столбец 3) носит название *ряда регрессии*. Для оценки величин, составляющих ряд регрессии, применяется *среднеквадратическое отклонение регрессии*

$$\sigma_{R_{y/x}} = \sigma_y \sqrt{1 - r_{xy}^2} \qquad (1.59)$$

где σ_y — среднеквадратическое отклонение изучаемого признака;

r_{xy} — коэффициент корреляции.

В нашем примере σ_y =±5,7; r_{xy} =± 0,719

$$\sigma_{R_{y/x}} = 5{,}7 \cdot \sqrt{1 - 0{,}719^2} = 5{,}7 \cdot 0{,}7 = \pm 3{,}99$$

Зная средние значения внутриглазного давления, мы можем определить и их колебания, в пределах которых могут находиться индивидуальные значения внутриглазного давления.

Средняя ошибка коэффициентов регрессии определяется по формулам:

$$m_{R_{x/y}} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n}} \quad \text{или} \quad m_{R_{y/x}} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n}} \qquad (1.60)$$

Полученные выборочные коэффициенты регрессии можно оценить с помощью критерия *t:*

$$t_{R_{x/y}} = \frac{R_{x/y}}{m_{R_{x/y}}} \quad \text{или} \quad t_{R_{y/x}} = \frac{R_{y/x}}{m_{R_{y/x}}} \tag{1.61}$$

1.10 Непараметрические критерии в медицинских исследованиях

Рассмотренные в предыдущих главах методы статистической обработки клинических и лабораторных данных имеют некоторые ограничения. Так, критерий Стьюдента и критерий Фишера применяются при нормальном распределении вариант изучаемых признаков, а критерий χ^2 требует определенного числа частот в сравниваемых рядах. Поэтому, наряду с *параметрическими критериями*, в практике статистической обработки данных стали применять *непараметрические критерии*, не зависящие от формы распределения. Непараметрическими они называются потому, что при их использовании не требуется вычисления параметров распределения: средних величин, дисперсии и др. Более простая техника вычисления этих критериев по сравнению с параметрическими, высокая статистическая мощность (чувствительность) некоторых из них способствует их широкому распространению. Кроме того, отдельные непараметрические критерии могут применяться даже тогда, когда изучаемые признаки выражены качественными признаками. Рассмотрим несколько непараметрических критериев.

1.10.1 Критерии для характеристики одной совокупности

Критерий проверки гипотезы о случайном характере флюктуации. Допустим необходимо установить, случайны ли колебаниями микроскопических измерений одного и того же штриха (в *мкм),* или наблюдается систематическое смещение результатов измерений: 3,68 3,11 4,76 2,75 4,15 5,08 2,96, 6,35 3,78 4,49 2,81 4,65 3,27 4,08 4,51 4,43 3,43 4,26 2,48 4,84

Решим эту задачу с помощью критерия проверка гипотезы о случайном характере флюктуации (колебаний). С этой целью расположим варианты нашего ряда в возрастающем порядке:

2,48 2,75 2,81 2,96 3,11 3,27 3,43 3,68 4,08 4,15 4,26 4,43 4,49 4,51 4,65 4,49 4,76 4,84 5,08 6,35

Находим медиану ряда, которая равна полусумме 10-й и 11-й вариант (так как n= 20) или

$$Me=(4,08+4,15)/2=4,115$$

Далее сравниваем медиану с исходными данными, расположенными в первоначальном порядке, при этом варианты большие, чем медиана, заменяем знаком (+), меньшие —знаком (—).

3,68 3,11 4,76 2,75 4,15 5,08 2,96, 6,35 3,78 4,49 2,81 4,65 3,27 4,08 4,51 4,43 3,43 4,26 2,48 4,84

— — + — + + — + — + — + — — + + — + — +

Подсчитываем число серий (R) одинаковых знаков, которых в нашем примере оказалось R = 16. Полученное число серий (R) сравниваем с табличными

значениями $R_{0,025}$ и $R_{0,975}$. Если $R < R_{0,25}$ или $> R_{0,975}$, то нулевая гипотеза о случайном характере колебания отвергаются.

При n=20 число серий, необходимых для признания колебаний случайными, находится в пределах от 6 до 15, следовательно, в нашем примере колебания результатов измерений следует признать неслучайными, систематическими. Следует проверить правильность методики измерений, изучаемых микроскопических штрихов.

Обязательным условием применения данного критерия является наличие результатов исследования одного и того же материала, выраженного в непрерывной числовой форме.

1.10.2 Критерии различия для двух сопряженных совокупностей

Критерий знаков. При применении критерия знаков определяется направленность изменений в каждой паре наблюдений, т. е. знак разнос между значениями *x* (до опыта) и *y* (после опыта). Затем подсчитывается число парных наблюдений, давших положительные (+) и отрицательные (-) разности. Наблюдения, имеющие нулевые разности из расчета исключаются. Число пар с менее часто встречающимся алгебраическим знаком обозначают буквой *Z*. Полученную величину *Z* сравнивают с табличными критическими числами менее часто встречающихся знаков Z_{05} и Z_{01} при разных числах наблюдений. Нулевая гипотеза т. е. предположение о том, что полученная в опыте разница случайна, принимается при $Z > Z_{05}$ и отвергается при $Z \leq Z_{05}$ ($p < 0,05$) или $Z \leq Z_{01}$ ($p<0,01$), когда обнаруженные различия признаются существенными

Например, сравнивается содержание лецитина у больных сахарным диабетом до лечения и после приема разных доз витамина B_6 (табл. 1.48). Из таблицы видно, что при приеме витамина B_6 в дозах 350 до 700 *мг* содержание лецитина увеличилось у всех больных. Следовательно, минимальное число наблюдений с одинаковым знаком (—) равно 0. Таким образом, $Z=0$. Критические значения Z при 14 наблюдениях равны $Z_{05}=3$, а $Z_{01}=2$. Итак, $Z < Z_{01}$. Различия в содержании лецитина до и после лёчения витамином B_6 в дозах 350—700 *мг* можно признать существенными ($p<0,01$).

После приема витамина в дозах 1300— 2000 мг *Z* оказалось равным трем. Критические значения *Z* для полученного результата при 13 наблюдениях, так как пара наблюдений, не давшая изменения (0), не принимается во внимание, равны $Z_{05}=3$, а $Z_{01}=1$. Нулевая гипотеза в этом случае может быть отвергнута, так как $Z= Z_{05}$. Различия в содержании лецитина у больных после приёма больших доз витамина B_6 по сравнению с исходным уровнем существенны.

Критерий знаков учитывает только знак разностей сравниваемых наблюдений, не отражая величины разности. Поэтому критерий знаков имеет ограниченную мощность, и если он не уловил различий, можно применить максимум-критерий или самый мощный из этой группы — парный критерий Вилкок-

сона, которые учитывают не только направленность, но и величину сдвига наблюдении.

Таблица 1.48 - Применение критерия знаков для оценки различий в содержании лецитина у больных сахарным диабетом до и после лечения

Больной	Лецитин (в мг%)				
	до лечения (контроль)	после приема от 350–700 *мг* витамина B_6	разница с контролем	после приема 1300–2000 мг витамина B_6	разница с контролем
А	180	245	+	322	+
Б	114	201	+	163	+
В	234	273	+	546	+
Г	382	385	+	382	0
З	209	221	+	93	–
Ж	252	392	+	150	–
Е	135	190	+	168	+
К	130	161	+	47	–
Л	56	168	+	168	+
М	197	202	+	546	+
Н	201	308	+	317	+
Р	98	240	+	327	+
Т	155	278	+	267	+
П	201	257	+	349	+

Максимум-критерий. Для его определения необходимо найти разности сравниваемых пар. Полученные разности располагают в убывающем порядке их абсолютной величины не учитывая их алгебраических знаков. После этого подсчитывают число первых (наибольших по абсолютной величине) разностей с одинаковым знаком. Если первые шесть наибольших разностей имеют одинаковый знак, то нулевая гипотеза об отсутствии различий сравниваемых рядов отвергается с вероятностью 95% (p=0,05). При наличии 8 наибольших разностей с одинаковым знаком нулевая гипотеза отвергается с вероятностью 99%, а при 11 — 99,9%.

В табл. 1.49 представлены данные о пороге безусловных реакций до и пѳ сле введения аминазина. В примере содержится 14 однозначных наибольших разностей. Следовательно, различия в величине порога безусловных реакций до и после введения аминазина существенны, с вероятностью ошибки менее 0,1 % (P < 0,001).

Парный критерий Вилкоксона. При вычислении этого критерия определяем разности между сравниваемыми парами наблюдений. Каждой полученной разности присваивают номер (ранг) в зависимости от её абсолютной величины. Наименьшей разности присваиваются первый номер. В том случае, если разно-

сти двух и более пар одинаковы, им присваивают ранг, равный средней арифметической величине из порядковых номеров.

Пары наблюдений, имеющие нулевые разности, из разработки исключаются. Далее вычисляют суммы рангов разностей, имеющих одинаковые знаки. Меньшую сумму *(T)* сравнивают с критическими значениями таких сумм при разном числе парных наблюдений.

Таблица 1.49 - Применение максимум-критерия для оценки влияния аминазина на порог, безусловных реакций

Больной	Порог безусловных реакций (в В)		Разность	Разности, расположенные по абсолютной величине
	до введения	после введения аминазина		
А.	80	70	+10	+30
Б.	90	75	+5	+20
В.	80	80	0	+20
Г.	130	110	+20	+20
Д.	80	60	+20	+20
Е.	100	100	0	+20
Ж.	100	90	+10	+20
З.	90	70	+20	+20
И.	100	80	+20	+20
К.	80	50	+30	+15
Л.	80	60	+20	+15
М.	90	70	+20	+10
Н.	120	110	+10	+10
О.	115	120	–5	+10
П.	90	100	–10	–10
Р.	75	60	+15	+5
С.	90	70	+20	–5
Т.	90	75	+15	0
У.	100	80	+20	0

Нулевая гипотеза принимается при $T = T_{05,}$ нулевая гипотеза отвергается и различия признаются существенными при $T<T_{05}$ или $T< T_{01}$.

Например, требуется установить, достоверны ли различия в уровне содержания гемоглобина у больных с хронической уремией до и после лечения их специальным рационом (табл. 1.50).

Сумма рангов отрицательных разностей равна $T(-) = 3+5,5+3=11,5$. Сумма рангов положительных разностей равна $T(+) = 1+6+8+2=17$. За величину критерия принимаем меньшую сумму $T = 11,5$. При $n=7$ (одно наблюдение с нулевой

разностью из расчета изъято) критическое значение критерия равно $T_{0,5}$=3. Поскольку $T>T_{0,5}$, то принимается нулевая гипотеза: различия в опыте не существенны.

Таблица 1.50 - Применение критерия Вилкоксона для оценки различий в содержании гемоглобина у больных с хронической уремией до и после лечения

Больной	Гемоглобин в %		Разность	Ранговый номер разности
	до лечения	после лечения		
А	60	59	1	1
Б	74	68	6	5,5
В	78	70	8	7
Г	47	49	–2	3
Д	40	40	0	–
К	74	80	–6	5,5
И	62	60	2	3
П	65	67	–2	3

1.10.3 Критерии различия для двух несопряженных совокупностей

Критерий Уайта. Критерий применяется для ориентировочной оценки различий двух независимых рядов наблюдений. (например, опытная группа сравнивается с контрольной).Для расчета критерия необходимо сопоставляемые ряды расположить в один ранжированный ряд в порядке возрастания либо убывания полученных величин. Каждому значению объединенного ряда присваивается порядковый номер. Далее подсчитываются суммы номеров для каждого из сопоставляемых рядов. Меньшая из полученных сумм *(К) с*равнивается с критическими значениями сумм, приведённых. Нулевая гипотеза принимается при $K \geq K_{05}$ и отвергается при $K<K_{05}$ или $K<K_{01}$.

Пример. Требуется установить, существенны ли различия в содержании свободного сульфаниламида у кроликов контрольной группы и кроликов, которым водился кортизон. Для получения сравнимых результатов вычислялось отношение содержания свободного сульфаниламида в экссудате к содержанию его в плазме.

Данные контрольной группы *(x):* 0,63 0,60 0,67 0,94 0,62 1,09 0,88 0,96 1,12 1,00 1,03 1,00 0,64 0,81 0,76; n_x =15.

Данные опытной группы *(y):* 1,96 1,64 0,90 1,61 0,44 1,24 1,23 1,

20 2,00 1,56 :1,67 1,76 1,23 1,46 1,50 n_y =15.

Составляем общий ранжированный ряд, располагая в отдельных строках данные групп *x* и *y*, и присваиваем каждому значению порядковый номер.

Значения	***x***	**0,6**	**0,62**	**0,63**	**064**	**0,67**	**0,76**	**0,81**	**0,88**		**0,94**	**0,96**	**1**	**1**	**1,03**
	y									**0,9**					
Ранги	***x***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**		**10**	**11**	**12**	**13**	**14**
	y									**9**					

Значения	***x***	**1,09**	**1,12**												
	y	**15**	**16**												
Ранги	***x***			**1,2**	**1,23**	**1,23**	**1,25**	**1,44**	**1,46**	**1,5**	**1,56**	**1,61**	**1,64**	**1,67**	**1,76**	**1,96**	**2,0**
	y			**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**

Подсчитываем сумму рангов (K) значений x и y: K_x=127 и Ky ==338. Меньшая сумма K=K_x= 127. При n_x=15 и n_y=15 критические значения критерия равны K_{05}=185 и K_{01}=171. Поскольку полученное $K<K_{01}$, в содержании свободного сульфаниламида у кроликов сравниваемых групп выявлены существенные различая (P<0,01).

Если число наблюдений n_x или n_y выходит за пределы, указанные в таблице критических значений критерия, т. е. больше 15 или 28, то для оценки различий можно использовать величину W

$$W = \frac{n_x(n_x + n_y + 1) - 2K}{\sqrt{n_x \cdot n_y(n_x + n_y + 1)}} \qquad (1.62)$$

тде —сумма рангов. Критические значения W: W_{05} = 1,13 и W_{01} = 1,49. Нулевая гипотеза отвергается при $W > W_{05}$ или $W > W_{01}$.

Критерий *U* (Вилкоксона—Манна—Уитни). Для расчета критерия U необходимо расположить данные первой (x) и второй группы (y) в один ряд по их величине, но так, чтобы было видно, к какой группе принадлежит каждая величина. Затем подсчитывается число инверсий, т. е. число величин второй группы (y), предшествующих каждой величине первой группы (x). Полученное число инверсий сравниваем с их критическими значениями. Если полученное число инверсий $U > U_{05}$, то нулевая гипотеза принимается и различия между группами признаются не значимыми. Если $U < U_{05}$ или U_{01} то различия считаются существенными с соответствующими уровнями значимости.

Например, требуется установить, существенна ли разница в содержании ацетилированного сульфаниламида у кроликов, которым вводился физиологический раствор, и теми животными, которым вводился кортизон (табл. 1.51). Из таблицы видно, что величинам группы x: 8,8 и 9,6 предшествует одна величина группы y-7,1, т.е. они имеют по одной инверсии. Величинам 14,3 и 14,7 предшествуют по две величины другой группы (7,1 и 13,4). Полученное число инверсий суммируем и находим U=114. По таблице критических значений критерия находим, что при n_x =n_y=15 U_{05} =72. Следовательно, $U > U_{05}$, и различия между группами не существенны.

Серийный критерий S (В а л ь д а— В о л ь ф о в и ц а). В основе критерия лежит подсчет числа серий, т. е. чередований вариант двух сравниваемых групп наблюдений в общем ранжированном ряду. Например, в ряду, составленном из вариант групп x и y: $\frac{xxx}{1}\frac{yy}{2}\frac{xxxx}{3}\frac{y}{4}\frac{xxx}{5}$ имеется пять серий. Полученное число серий сравнивается с критическим числом серий, при котором уже принимается нулевая гипотеза (предположение о принадлежности сравниваемых групп к одной генеральной совокупности). При меньшем числе серий ($S<S_{05}$) нулевая гипотеза должна быть отвергнута. Значение S_{01} приблизительно соответствует значению $S_{05}-2$.

Таблица 1.51 - Применение критерия U для оценки различий в содержании ацетилированного сульфаниламида у двух групп кроликов

Ранжированные данные: % ацетилированного сульфаниламида у кроликов, которым вводился		Последовательность величин	Число инверсий
физиологический раствор (x)	кортизон (y)		
	7,1	y	
8,8		x	1
9,6		x	1
	13,4	y	
14,3		x	2
14,7		x	2
	16,3	y	
20,4		x	3
	20,4	y	
	25,4	y	
25,6		x	5
	26	y	
	32,8	y	
33,3		x	7
35		x	7
	36,7	y	
37,5		x	8
	40	y	
43,6		x	9
	46,9	y	
	48,3	y	
	50	y	
	50	y	
52,3		x	13
	52,9	y	
55,3		x	14

57,9		*x*	14
64,6		*x*	14
64,8		*x*	14
	65,2	*y*	
nx=15	ny=15	U=114	

Если число наблюдений в одной или обеих группах >20, то число серий *S* оценивается с помощью случайной переменной величины U_s по формуле:

$$U_s = \frac{\bar{S} - S - 0{,}5}{\sigma_x} \qquad (1.63)$$

где

$$\bar{S} = \frac{a}{b} + 1, \quad \sigma_s = \sqrt{\frac{a(a-b)}{b^2(b-1)}}, \quad a = 2n_x \cdot n_y, \quad b = n_x + n_y$$

Нулевая гипотеза принимается при $U_s \leq 1{,}96$ и отвергается при $U_s > 1{,}96$ (p< 0,05) или $U_s > 2{,}58$ (p<.0.01).

Следует подчеркнуть, что серийный критерий обнаруживает различия не только по центральной тенденции, но и по рассеянию вариант. Рассмотрим вычисление его на примере оценки различий в уровне механической резистентности эритроцитов (в %) у больных шизофренией (*x*) и больных, не страдающих данным заболеванием (*y*)

X – 3,8 0,5 1,7 1,0 5,4 4,9 3,1 4,5 n_x=8

Y – 10,5 13,8 9,2 6,2 7,6 3,0 5,3 8,2 3,9 7,0 3,5 5,0 2,2 6,2 n_y=14

Построим ранжированный ряд, сохраняя варианты x и *y* в отдельных строках и подсчитаем число серий:

Величина	***x***	**0,5 1,0 1,7**		**3,1**		**3,8**		**4,5 4,9**		**5,4**	
	Число серий	**1**		**3**		**5**		**7**		**9**	
Величина	***y***		**2,2 3,0**		**3,5**		**3,9**		**5,0 5,3**		**6,2 6,2 7,0 7,6**
	Число серий		**2**		**4**		**6**		**8**		**10**

В полученном ряду содержится 10 серий (*S*=10). По таблице находим, что при n_x=8 и n_y=14 критическое значение S_{05}=7. Так как $S > S_{05}$, то различия между группами больных по уровню механической резистентности эритроцитов признать существенными нельзя (p>0.05).

Критерий Колмагорова-Смирнова. Этот критерий является более мощным чем серийный критерий, особенно при большом числе наблюдений. Он основан на сравнении рядов накопленных частот. Схему вычисления критерия рассмотрим на примере определения значимости различий в росте однослой-

ных культур фибробластов под влиянием кортизон-ацетата *(x)* и в контроле (типичная среда без стероидов —*y*) (табл. 1.52).

Методика расчета критерия Колмогорова — Смирнова будет следующая:

1. Располагаем варианты обеих групп в один возрастающий ряд (графа 1);

2. В графах 2 и 3, записываем раздельно частоты вариант каждой группы (*fx* и f_y);

3. Последовательно, суммируем частоты *fx* и f_y составляя ряды накопленных частот S_x и S_y, которые записываем в графах 4 и 5;

4. Путем деления накопленных частот на число наблюдений в каждой группе получаем ряды накопленных частостей (графы 6 и 7);

Таблица 1.52 - Применение критерия Колмогорова — Смирнова для оценки различий в росте двух культур фибробластов

Варианты обоих рядов в возрастающем порядке	Частоты вариант по группам		Накопленные частоты по группам		Накопленные частости по группам		Разности *sx/nx-sy/ny* (без учета знаков)
	f_x	f_y	S_x	S_y	S_x/n_x	S_y/n_y	
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
198	0	1	0	1	0	0,091	0,091
242	0	1	0	2	0	0,182	0,182
253	0	1	0	3	0	0,273	0,273
264	0	1	0	4	0	0,364	0,364
286	0	1	0	5	0	0,455	0,455
297	0	1	0	6	0	0,545	0,545
319	1	2	1	8	0,091	0,727	0,636
341	0	1	1	9	0,091	0,818	0,727
352	0	1	1	10	0,091	0,909	0,818
385	1	0	2	10	0,182	0,909	0,827
429	1	1	3	11	0,273	1	0,717
440	1	0	4	11	0,364	1	0,636
473	2	0	6	11	0,515	1	0,455
517	1	0	7	11	0,636	1	0,364
539	1	0	8	11	0,727	1	0,273
594	1	0	9	11	0,818	1	0,182
638	1	0	10	11	0,909	1	0,091
660	1	0	11	11	1	1	0

5. Определяем разности между накопленными частностями (без учета их алгебраических знаков) и находим максимальную разность *D* (в нашем примере 0,818).

6. Вычисляем критерий Колмогорова.—Смирнова λ^2 по формуле:

$$\lambda^2 = D^2 \cdot \frac{n_x \cdot n_y}{n_x + n_y} = 0{,}818^2 \frac{11 \cdot 11}{11+11} = 3{,}76 \qquad (1.64)$$

7. Сопоставляем полученное значение λ^2 с критическими значениями: λ^2_{05}= 1,84 и λ^2_{01}=2,65. Если $\lambda^2 < \lambda^2_{05}$, то принимается нулевая гипотеза, если $\lambda^2 > \lambda^2_{05}$ или $\lambda^2 > \lambda^2_{01}$, то различия признаются существенными ($p < 0{,}05$ или $p < 0{,}01$). В нашем примере $\lambda^2 > \lambda^2_{01}$ Следовательно, различия в росте культур фибробластов существенны ($p< 0{,}01$).

Если число наблюдений в сравниваемых группах одинаково, то критерий Колмогорова — Смирнова можно рассчитать по формуле: $\lambda_2 = D^2 n/2$. Действительно в нашем случае; $\lambda^2 = 0{,}684 \cdot 11/2 = 3{,}76$.

1.10.3 Непараметрические методы изучения связи

Коэффициент корреляции рангов Спирмена. Рассмотрим методику вычисления этого коэффициента на примере определения связи между количеством эритроцитов и процентом гемоглобина в крови у человек (табл. 1.53).

Таблица 1.53 - Вычисление коэффициента корреляции рангов Спирмена между количеством эритроцитов и процентом гемоглобина в крови

Обследованные	Количество эритроцитов, x	Гемоглобин в %, y	Ранги вариант		Разности рангов, d	Квадраты разностей рангов, d^2
			r_x	r_y		
1	**2**	**3**	**4**	**5**	**6**	**7**
А.	1,98	40	1	1	0	0
К.	2,5	47	2	2	0	0
Б.	2,94	60	3	3	0	0
З.	3,25	62	4	4	0	0
С.	3,64	74	5	6,5	-1,5	2,25
И.	3,7	65	6	5	+1	1
Ж.	3,86	78	7	8	-1	1
И.	4,29	74	8	6,5	+1,5	2,25
	n=8					$\sum d^2$=6,5

Методика расчета критерия коэффициент корреляции рангов Спирмена будет следующая:

1. Располагаем данные обследованных в порядке возрастания вариант первого признака (в нашем примере - количества эритроцитов) (графы 1, 2,3);

2. Заменяем значения вариант в каждом ряду их рангами (графы 4 и-5). Если встречаются одинаковые варианты, то каждой из них присваивается средний ранг. В нашем примере варианта 74% гемоглобина встречается 2 раза и занимает порядковые места 6 и 7, следовательно, средний ранг равен

$$\frac{6+7}{2}=6{,}5;$$

3. Находим разности между смежными рангами сравниваемых рядов (графа 6) Сумма этих разностей должна равняться нулю;

4. Возводим п6лученные разности в квадрат и суммируем их(графа 7);

5. Вычисляем коэффициент корреляции рангов Спирмена по формуле:

$$\rho=1-\frac{6\sum d^2}{n(n^2-1)} \qquad (1.65)$$

где Σd^2 —сумма квадратов разностей рангов;

n—число сравниваемых пар.

В нашем примере:

$$\rho = 1-\frac{6*6{,}5}{8*(8^2-1)}=1-\frac{39{,}0}{504}=1-0{,}077=0{,}923$$

Таким образом между числом эритроцитов и содержанием гемоглобина в крови существует высокая прямая корреляционная связь. Величина коэффициента корреляции рангов Спирмена оценивается так же, как и величина параметрического коэффициента корреляции.

Оценка надежности коэффициента корреляции рангов Спирмена при числе наблюдений 10 и более производится с помощью критерйя t по формуле

$$t=\rho\sqrt{\frac{n-2}{1-\rho^2}} \qquad (1.66)$$

Вероятность, соответствующую полученному *t* определяем по таблице Стьюдента при числе степеней свободы *n*-2. В тех случаях, когда число наблюдений меньше 10, оценка значимости производится по табл. 1.54. Если вычисленный коэффициент корреляции ρ больше табличного $\rho_{05,}$ то он значим с вероятностью 95%. В нашем примере при n=8 $\rho>\rho_{01}$ (0,92>0,833) следовательно, полученный нами коэффициент корреляции значим с высокой вероятностью ($\rho<0{,}01$).

Таблица 1.54 - Критические значения коэффициента корреляции рангов Спирмена ρ

Число парных наблюдений,	Уровень значимости		Число парных наблюдений,	Уровень значимости	
n	0,05	0,01	*n*	0,05	0,01
4	1	—	16	0,425	0,601
5	0,9	1	18	0,399	0,564
6	0,829	0,943	20	0,377	0,534
7	0,714	0,893	22	0,359	0,508
8	0,643	0,833	24	0,343	0,485
9	0,6	0,783	26	0,329	0,465

10	0,564	0,746	28	0,317	0,448
12	0,506	0,712	30	0,306	0,432
14	0,456	0,645			

В ряде случаев коэффициент корреляции рангов, можно использовать для выяснения связи между качественными и количественными признаками (табл. 1.55). Коэффициент корреляции рангов в этом случае равен

$$\rho = 1 - \frac{6*44,5}{8*(64-1)} = 1 - \frac{268}{504} = 0,470$$

Критическое значение ρ_{05} при n=8 равняется 0,643. Таким образом $\rho < \rho_{05}$. Следовательно, результат оценки связи при данном числе наблюдений признать достоверным нельзя.

Таблица 1.55 - Вычисление коэффициента корреляции *p* между количеством креатинина в крови больных хронической уремией и выраженностью у них рвоты

Больные	Количество креатинина (мг %) *x*	Выраженность рвоты *y*	Ранги		Разность рангов *d*	Квадраты разности *d*
			r_x	r_y		
А.	2,1	+	1	3,5	-2,5	6,25
К.	2,3	+	2	3,5	-1,5	2,25
Б.	3,6	-	3	1	2	4
С.	3,7	++	4,5	7	-2,5	6,25
З.	3,7	++	4,5	7	-2,5	6,25
И.	4,4	+	6	3,5	2,5	6,25
Ж.	4,5	+	7	3,5	3,5	12,25
Н.	4,95	++	8	7	1	1
						∑=44,5

Известно, что на величину коэффициента ранговой корреляции серьезно влияет усреднение рангов. Поэтому при наличии значительного числа усредненных рангов для вычисления коэффициента ρ применяется следующая формула:

$$\rho = \frac{\frac{n^3 - n}{6} - (T_x + T_y) - \sum d^2}{\sqrt{\left(\frac{n^3 - n}{6} - 2T_x\right) \cdot \left(\frac{n^3 - n}{6} - 2T_y\right)}} \qquad (1.67)$$

где $T_x(T_y) = \sum \frac{t^3 - t}{12}$,

t – численность каждой группы усредненных рангов.

В нашем примере

$$T_x = \frac{(2^3 - 2)}{12} = \frac{6}{12} = 0{,}5$$

$$T_y = \frac{(4^3 - 4) + (3^3 - 3)}{12} = \frac{60 + 24}{12} = 7$$

$$\rho = \frac{\frac{8^3 - 8}{6} - (0{,}5 + 7) - 44{,}5}{\sqrt{(\frac{8^3 - 8}{6} - 2*0{,}5)(\frac{8^3 - 8}{6} - 2*7)}} = \frac{32}{\sqrt{83*70}} = 0{,}42$$

Полученный скорректированный коэффициент ρ равен 0,42 и также не дает возможности надежно судить о наличии исследуемой связи. Очевидно, требуется увеличить число наблюдений.

1.11 Современное программное обеспечение для статистической обработки биомедицинских исследований

Одним из обязательных этапов любого научного исследования является статистический анализ данных. Продолжительное время анализ медицинских данных был уделом специалистов, так как это требовало серьезной предварительной подготовки. С появлением и совершенствованием современных программ обработки данных статистическая обработка поднялась на новый уровень. Теперь исследователь-медик может и не иметь математической подготовки. Достаточно оперировать статистическими понятиями и, самое главное, правильно выбрать метод анализа. Все осуществимо благодаря компьютеру и новейшим программам.

Все программы статистической обработки данных можно разделить на *профессиональные*, *полупрофессиональные* (популярные) и *специализированные*. Статистические программы относятся к наукоемкому программному обеспечению. Профессиональные пакеты имеют большое количество методов анализа, популярные пакеты - количество функций, достаточное для универсального применения. Специализированные же пакеты ориентированы на какую-либо узкую область анализа данных. Создатели программных статистических пакетов заявляют, что их продукт превосходит аналоги. Отсутствие у большинства исследователей времени для освоения нескольких программ, делает непростым ее выбор. Далее приведена базовая информация о присутствующих на рынке основных полупрофессиональных программных пакетах, пригодных для статистической обработки биомедицинских данных.

MS Excel. Самой часто упоминаемой (и используемой) в отечественных статьях является приложение MS Excel из пакета офисных программ компании Microsoft – MS Office. Причины этого кроются в широком распространении

этого программного обеспечения, наличии русскоязычной версии, тесной интеграцией с MS Word и PowerPoint. Однако, MS Excel - это электронная таблица с достаточно мощными математическими возможностями, где некоторые статистические функции являются просто дополнительными встроенными формулами. Расчеты сделанные при ее помощи не признаются авторитетными биомедицинскими журналами. Также в MS Excel невозможно построить качественные научные графики. Безусловно, MS Excel хорошо подходит для накопления данных, промежуточного преобразования, предварительных статистических прикидок, для построения некоторых видов диаграмм. Однако окончательный статистический анализ необходимо делать в программах, которые специально созданы для этих целей. Существует макрос-дополнение XLSTAT-Pro для MS Excel который, включает в себя более 50 статистических функций, включая анализ выживаемости, которых в основных случаях достаточно для обычного применения. Пробную версию макроса можно взять на сайте производителя.

STADIA. Программа российской разработки с 16-и летней историей. Включает в себя все необходимые статистические функции. Она прекрасно справляется со своей задачей - статистическим анализом. Но. Программа внешне фактически не изменяется с 1996 года. Графики и диаграммы, построенные при помощи STADIA, выглядят в современных презентациях архаично. Цветовая гамма программы (красный шрифт на зеленом) очень утомляет в работе. К положительным качествам программы можно отнести русскоязычный интерфейс и наличие книг описывающих работу. Например: Кулаичев А.П. Методы и средства анализа данных в среде Windows. - М: ИнКо, 2002. - 341 с.

SPSS (Statistical Package for Social Science). Самый часто используемый пакет статистической обработки данных с более чем 30-и летней историей. Отличается гибкостью, мощностью, применим для всех видов статистических расчетов, применяемых в биомедицине. Существует русскоязычное представительство компании которое предлагает полностью русифицированную версию SPSS 12.0.2 для Windows. Появился учебник на русском языке, позволяющий шаг за шагом освоить возможности SPSS, репетитор по статистике на русском языке, помогающий в выборе нужной статистической или графической процедуры для конкретных данных и задач, а также справка по SPSS Base и SPSS Tables. Российский офис SPSS регулярно проводит учебные курсы по анализу данных при помощи программного обеспечения SPSS. На русский язык переведена книга по SPSS, которая вышла в свет в 2002 году в Киевском издательстве «Диасофт» под названием «SPSS 10: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей».

STATA. Профессиональный статистический программный пакет, который может применятся для биомедицинских целей. Один из самых популярных в образовательных и научных учреждениях США наряду с SPSS. Программа хорошо документирована, издается специальный журнал для пользователей системы. Однако возможности предварительного ознакомления с демо-версией нет.

STATISTICA. Производителем программы является фирма StatSoft Inc. (США), которая выпускает статистические приложения, начиная с 1985 года. STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций), объединенных следующими специализированными статистическими модулями: Основные статистики и таблицы, Непараметрическая статистика, Дисперсионный анализ, Множественная регрессия, Нелинейное оценивание, Анализ временных рядов и прогнозирование, Кластерный анализ, Факторный анализ, Дискриминантный функциональный анализ, Анализ длительностей жизни, Каноническая корреляция, Многомерное шкалирование, Моделирование структурными уравнениями и др. Несложный в освоении этот статистический пакет может быть рекомендован для биомедицинских исследований любой сложности. Российское представительство компании (http://www.statsoft.ru/) предлагает полностью русифицированную 6-ю версию программы. Сайт компании содержит много информации по статистической обработке медицинских данных, учебник по статистике на русском языке. Сам пакет STATISTICA описан в нескольких книгах, одна из которых, для медицинских работников: О.Ю. Реброва «Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA.» Москва, МедиаСфера, 2002. 312 с.

JMR. Один из мировых лидеров в анализе данных. Развивает этот статистический пакет SAS Institute, который выкупил в конце 2002 года известную статистическую программу StatView. Однако особых преимуществ для медико-биологической статистики этот программный продукт не имеет.

SYSTAT Статистическая система для персональных компьютеров Последняя 11 версия обладает неплохим интуитивно понятным интерфейсом. Компания Systat Software также разрабатывает популярные у отечественных исследователей SigmaStat и SigmaPlot, которые являются соответственно, программой статистической обработки и программой построения диаграмм. При совместной работе становятся единым пакетом для статистической обработки и визуализации данных.

NCSS. Программа развивается с 1981 года и рассчитана на непрофессионалов в области статистической обработки. Интерфейс системы многооконный и как следствие этого явления - немного непривычный в использовании. Все действия пользователя сопровождаются подсказками. Сейчас доступна версия 2004 г. С сайта можно переписать полнофункциональную пробную версию, работающую 30 дней.

MINITAB 14. Статистический пакет MINITAB в настоящее время выпускается в версии 14. С сайта производителя можно взять полнофункциональный пробный вариант программы, которая работает 30 дней. Это достаточно удобный в работе программный пакет, имеющий хороший интерфейс пользователя, хорошие возможности по визуализации результатов работы. Имеет подробную справку.

STATGRAPHICS PLUS. Довольно мощная статистическая программа. Содержит более 250 статистических функций, генерирует понятные, настраиваемые отчеты. Последняя доступная версия - 5.1. Ее можно получить на сайте Есть возможность скачать демо-версию. Следует отметить, что ранние версии этой программы были весьма популярны у отечественных исследователей.

PRISM. Эта программа создавалась специально для биомедицинских целей. Интуитивно понятный интерфейс позволяет в считанные минуты проанализировать данные и построить качественные графики. Программа содержит основные часто применяемые статистические функции, которых в большинстве исследований будет достаточно. Однако, как отмечают сами разработчики, программа не может полностью заменить серьезных статистических пакетов. На сайте помимо возможности ознакомления с демо-версией Prism можно получить справочник в формате PDF по биомедицинской статистике.

Дополнительная информация. В настоящее время в Интернете доступны многие ресурсы, посвященные статистической обработке данных. Один из них - это статистический портал, созданный при содействии В. П. Боровикова, автора книг по программному пакету STATISTICA Российское представительство StatSoft Inc. предлагает на своем сайте бесплатный электронный учебник по статистике, который призван помочь разобраться с основными понятиями статистики и более полно представить диапазон применения статистических методов На этом же сайте существует «Статистический медицинский советник», который поможет правильно выбрать нужный статистический метод

Из ресурсов Интернет заслуживает внимания сайт с пятилетней историей – Биометрика

На какой программе остановить свой выбор? Безусловно, дороговизна программ не позволяет их менять. Поэтому имеет смысл посмотреть демо-версии, разобраться с работой и потом делать окончательный вывод. Русскоязычные версии (с документацией) имеют только SPSS и STATISTICA.

Что касается возможных рекомендаций, то они следующие:

- Если нужен мощный, общепризнанный пакет с простым и понятным даже начинающим пользователям интерфейсом, то лучше воспользоваться SPSS.

- Для начинающих и профессионалов, которым нужна подсказка и развитая документация на русском языке, можно рекомендовать STATISTICA. Это мощное приложение с профессиональными возможностями.

- Для непритязательных пользователей, которые ограничиваются в своих исследованиях стандартными статистическими методами можно рекомендовать англоязычную программу Prism.

2 ПРИНЦИПЫ ПОСТРОЕНИЯ БАНКОВ ДАННЫХ

2.1 Общие сведения о банках данных

Банк данных (БНД) – совокупность базы данных и системы управления базами данных

База данных (БД) – это структурированная совокупность данных. Данные в БД хранятся в виде записей. Запись представляет совокупность элементов описания данных, объединенных отношением принадлежности к одному описываемому объекту. Элемент описания данных представляет собой наименьшую единицу описания данных. Элементы описания соответствуют отдельным свойствам объекта, а запись описывает объект в целом.

Основным назначением БД в первую очередь является быстрый поиск содержащейся в ней информации. При значительном размере БД ручной поиск, а также модификация содержащейся информации занимает значительное время. Использование компьютера для ведения БД устраняет перечисленные выше проблемы - поиск и выборка информации, ее модификация осуществляются достаточно быстро и эффективно, а сама БД, состоящая из тысяч записей, может легко уместиться на дискете.

Система управления базами данных (СУБД) – совокупность языковых и программных средств, предназначенных для создания и использования базы данных прикладными программами, а также непосредственно пользователями – непрограммистами.

Основная особенность СУБД - это наличие средств для ввода и хранения не только самих данных, но и описаний их структуры.

Если говорить более детально, то к функциям СУБД относят следующие:

управление данными непосредственно в БД - функция, обеспечивающая хранение данных, непосредственно входящих в БД, и служебной информации, обеспечивающей работу СУБД;

управление данными в памяти компьютера - функция, связанная в первую очередь с тем, что СУБД работают с БД большого размера. В целях ускорения работы СУБД используется буферизация данных в оперативной памяти компьютера. При этом пользователь СУБД использует только необходимую для его конкретной задачи часть БД, а при необходимости получает новую "порцию" данных;

управление транзакциями - функция СУБД, которая производит ряд операций над БД, как над единым целым. Как правило, такие операции производятся в памяти компьютера. В первую очередь транзакции необходимы для поддержания логической целостности БД в многопользовательских системах. Если транзакция (манипуляция над данными) успешно выполняется, то СУБД вносит

соответствующие изменения в БД. В обратном случае ни одно из сделанных изменений никак не влияет на состояние БД;

управление изменениями в БД и протоколирование - функция, связанная с надежностью хранения данных, то есть возможностью СУБД восстанавливать состояние БД в аварийных ситуациях, например, при случайном выключении питания или сбое носителя информации. Очевидно, что для восстановления БД нужно располагать дополнительной информацией, по которой и осуществляется восстановление. С этой целью ведется протокол изменений БД, в который перед манипуляциями с данными делается соответствующая запись. Для восстановления БД после сбоя СУБД используется протокол и *архивная копия БД* - полная копия БД к моменту начала заполнения протокола;

поддержка языков БД - для работы с БД используются специальные языки, в целом называемые *языками баз данных*. В СУБД обычно поддерживается единый язык, содержащий все необходимые средства - от создания БД до обеспечения пользовательского интерфейса при работе с данными. Наиболее распространенным в настоящее время языком СУБД является язык SQL (Structured Query Language).

Применение банков данных позволяет решить следующие проблемы организации и обработки больших массивов информации:

1) сокращение избыточности;
2) обеспечение целостности;
3) разграничение доступа;
4) обеспечение независимости представления данных.

Избыточность, как правило, вызывается наличием разных форм представления одних и тех же данных, размножением части данных для дальнейшего использования прикладными программами, повторными записями одинаковых данных на различных физических носителях информации. Для сокращения избыточности производится объединение одинаковых по смыслу, но имеющих различный тип данных в единую БД с приведением к общему, стандартному виду. Процесс объединения данных, используемых различными пользователями, в одну общую БД называется *интеграцией базы данных*.

Целостностью называется свойство БД в любой момент времени содержать лишь достоверные данные, то есть обеспечение целостности предполагает отсутствие избыточных, противоречивых и неверно составленных данных в БД.

Режим разграничения доступа предполагает, что каждый конкретный пользователь должен получить доступ лишь к некоторому подмножеству данных из БД, необходимых для выполнения своих прикладных программ. Одновременно с этим обеспечивается режим секретности и повышается степень защищённости данных от несанкционированного доступа.

Применение БД даёт возможность *обеспечения независимости представления данных* в прикладных программах от типов запоминающих устройств и способов их физической организации. В основном это достигается построением двух уровней представления данных: логического и физического.

Логический уровень представления данных отражает вид представления данных удобный для использования их в прикладных программах или непосредственно пользователями.

Физический уровень представления данных отражает способ хранения и структуру данных с учётом их расположения на носителях информации в запоминающих устройствах ЭВМ.

Важнейшим понятием в теории БД является *модель данных* – формализованное описание, отражающее состав и типы данных, а так же взаимосвязи между ними. Модели данных классифицируются по ряду признаков.

В зависимости от объекта описываемой информации на логическом уровне различают внешнюю и внутреннюю модели данных

Внешняя логическая модель данных описывает структуру информации, относящейся к некоторой конкретной процедуре или группе родственных процедур обработки данных.

Внутренняя логическая модель данных объединяет все внешние модели данных.

По способам отражения связей между данными на логическом уровне различают модели – иерархическую, сетевую и реляционную.

Сетевой называется модель, в которой данные и их связи имеют структуру графа.

Иерархичной называют модель, в которой структура отражаемых связей представляется в виде дерева.

Реляционной называют модель, в которой представление данных осуществляется в форме таблиц.

Задание модели данных в БД осуществляется на специальном *языке описания данных* (ЯОД). Язык данных описания представляет собой совокупность директив, построенных в соответствии с выбранной моделью данных. Несмотря на то что ЯОД ориентирован на логический уровень, в него, как правило, включаются директивы, позволяющие управлять расположением данных на внешних носителях.

Прикладные программы, использующие БД, записываются на некотором алгоритмическом языке (например, ФОРТРАН, ПАСКАЛЬ СИ и др.), называемом *включающим языком*. Для обеспечения взаимодействия с БД в эти программы должны быть введены операторы обращения к СУБД. Совокупность операторов обращения к СУБД из прикладной программы составляет *язык манипулирования данными* (ЯМД). Основные операции с данными, выполняемые средствами ЯМД, следующие:

1)поиск информации по заданным поисковым признакам в БД;

2) включение в БД новых записей;

3) удаление из БД лишних или ненужных в дальнейшем записей;

4) изменение значений элементов данных в записях.

Совокупность модели данных и операций, определённых над данными, называется подходом. В соответствии с моделями данных различают реляцион-

ный, сетевой и иерархический подходы. Так как подход лежит в основе построения СУБД, различают реляционные, сетевые и иерархические СУБД. Иерархические и сетевые СУБД обладают возможностью обеспечения быстрого доступа к данным. Реляционные СУБД, несмотря на трудность их программной реализации, позволяют более удобно для пользователя описать структуру данных и манипулировать ими.

Тип организации СУБД определяется так же степенью структурированности записей в составе БД. Различают *сильноструктурированные и слабоструктурированные* записи.

Сильноструктурированная запись – запись, построенная в соответствии с фиксированным, заранее определённым форматом всех элементов описания. К таким данным относятся, например, сведения о микросхемах и других элементах в БД САПР, сведения о сотрудниках в БД отдела кадров организации и др. СУБД, предназначенные для хранения сильноструктурированных записей получили название *фактографических* СУБД.

Слабоструктурированная запись – это запись, у которой лишь отдельные элементы описания имеют фиксированный, заранее определённый формат. Например, в отчете о НИР фиксированный формат могут иметь лишь элементы описания, соответствующие заголовку, году регистрации НИР, организации – исполнителя. Сам текст отчета содержит информацию символьного типа переменной длины. СУБД, предназначенные для хранения таких записей получили название *документальных* или *информационно – поисковых* (ИПС).

Организация технического обеспечения АСОИ оказывает влияние на структуру информационного обеспечения и в первую очередь баз данных. Если БД сконцентрирована в одном узле вычислительной сети, то она называется *сосредоточенной*, в противном случае – *распределённой*. Если информационное обслуживание с помощью БД относится ко всей АСОИ, то БД называют *общей* (*интегрированной* или *центральной*), а если к отдельной подсистеме или к отдельному пакету прикладных программ, то *локальной* БД.

2.2 Типы баз данных

Для разных задач целесообразно использовать различные типы баз данных, поскольку, конечно, базу данных сведений о сотрудниках какого-то небольшого коллектива и базу данных о каком-нибудь банке, имеющем филиалы во всех концах страны, надо строить по разному. /

В настоящее время различают четыре типа баз данных:

- Автономные БД
- Файл-серверные БД
- Многоярусные БД
- БД на платформе «Клиент/сервер».

2.2.1 Автономные базы данных

Автономные локальные базы данных являются наиболее простыми. Они хранят свои данные в локальной файловой системе на том компьютере, на котором установлены. Система управления базой данных, осуществляющая к ним доступ, находится на том же самом компьютере. Сеть не используется. Поэтому разработчику автономной базы данных не приходится иметь дело с проблемой параллельного доступа, когда два человека пытаются одновременно изменить одну и ту же запись, потому что такого никогда не может быть. Автономные базы данных не используются для приложений, требующих значительной вычислительной мощности, потому что процессорное время будет потрачено на выполнение манипуляций с данными и в целом будет потеряно для приложения.

Автономные базы данных полезны для развития тех приложений, которые распространены среди многих пользователей, каждый из которых поддерживает отдельную базу данных. Это, например, приложения, обрабатывающие документацию небольшого офиса, кадровый состав небольшого предприятия, бухгалтерские документы небольшой бухгалтерии. Каждый пользователь такого приложения манипулирует своими собственными данными на своем компьютере. Пользователю нет необходимости иметь доступ к данным любого другого пользователя, так что отдельная база данных здесь вполне приемлема.

2.2.2 Файл-серверные базы данных

Файл-серверные базы данных отличаются от автономных тем, что они могут быть доступны многим клиентам через сеть. Это очень удобно, так как изменения в таких базах данных видят все пользователи. Например, базу данных сотрудников крупного учреждения целесообразно делать именно такой, чтобы администраторы отдельных подразделений обращались к ней, а не заводили у себя локальные базы данных (при этом можно сделать так, чтобы каждый администратор видел только ту информацию, которая относится к его подразделению).

Сама база данных хранится на сетевом файл-сервере в единственном экземпляре. Для каждого клиента во время работы создается локальная копия данных, с которой он манипулирует. При этом возникают проблемы, связанные с возможным одновременным доступом нескольких пользователей к одной и той же информации. Например, при проектировании приложений, работающих с подобными базами данных, должны быть решены такие проблемы: что делать, если пользователь прочел некоторую запись и, пока он ее просматривает и собирается изменить, другой пользователь меняет или удаляет эту запись.

Одним из недостатков файл-серверных баз данных является непроизводительная загрузка сети. При каждом запросе клиента данные в его локальной копии полностью обновляются из базы данных на сервере. Даже если запрос относится всего к одной записи, обновляются все записи данных. Если записей в базе данных много, то даже при небольшом числе клиентов сеть будет загружена очень основательно, что серьезно скажется на скорости выполнения запросов.

Другой недостаток связан с тем, что забота о целостности данных при такой организации работы возлагается на программы клиентов. Если они недостаточно тщательно продуманы, в базу данных легко занести ошибки, которые могут отразиться на всех пользователях.

2.2.3 Многоярусные базы данных

Это новый и многообещающий путь обработки данных в сети. Иногда этот способ организации баз данных называется multi-tier — многонитевые. В этом термине под нитью понимается один из множества потоков данных, обменивающихся одновременно с базой данных.

Наиболее распространен трехярусный вариант:

- На нижнем уровне на компьютерах пользователя расположены приложения клиентов, обеспечивающие пользовательский интерфейс.
- На втором уровне расположен сервер приложений, обеспечивающий обмен данными между пользователями и распределенными базами данных. Сервер приложений размещается в узле сети, доступном всем клиентам
- На третьем уровне расположен удаленный сервер баз данных, принимающий информацию от серверов приложений и управляющий ими.

Подобную концепцию обработки данных пропагандируют, в частности, фирмы Oracle и Sun. Первый, элементарный уровень состоит из «тонких клиентов», то есть несложных терминалов, предназначенных, в основном, для ввода — вывода. Второй, средний (middleware) уровень — это рабочие станции и серверы приложений, то есть значительно более серьезные машины, на которых выполняются программы, критичные к загрузке процессора. Третий и последний уровень — мощные специализированные серверы баз данных.

Многоярусная организация — наиболее сложный, гибкий и эффективный способ работы с базами данных. При этом надо отметить, что на нижнем уровне — на компьютерах пользователя не требуется установки Borland Database Engine (BDE). В этом заключается одно из преимуществ многоярусных распределенных баз данных.

2.2.4 Базы данных клиент/сервер

Для больших баз данных с множеством пользователей часто используются

базы данных на платформе клиент/сервер.

Основой такой системы является *сервер БД*, представляющий собой приложение, осуществляющее комплекс действий по управлению данными - выполнение запросов, хранение и резервное копирование данных, отслеживание целостности, проверку прав пользователей, ведение журнала транзакций. В качестве *рабочего места (клиента)* при этом может быть использован обычный персональный компьютер.

Таким образом, информационная система построенная по принципу клиент/сервер, состоит обычно из трех основных компонентов:

- *сервер БД,* который и является собственно СУБД и управляет хранением данных, доступом, защитой, резервным копированием, отслеживает целостность данных и выполняет запросы клиента:
- *клиенты,* представляющие собой различные приложения пользователей и выполняющие запросы к серверу, проверяющие допустимость данных и получающие ответы от него;
- *сеть и коммуникационное программное обеспечение,* осуществляющее взаимодействие между клиентом и сервером с помощью сетевых протоколов.

В функции сервера БД входит не только непосредственное обслуживание данных. Обязательно предусматриваются системы блокировки и управления многопользовательским доступом, элементы ограждения данных от несанкционированного доступа, структуры оптимизации запросов к БД.

Кроме того, в задачи серверной части СУБД входит обеспечение *ссылочной целостности данных* и *контроль завершения транзакций*. *Ссылочная целостность данных* - это система специальных правил, обеспечивающих единство связанных данных в БД. *Контроль завершения транзакций* - задача СУБД по контролю и предупреждению повреждения данных в нештатных ситуациях, например, при аппаратном сбое.

Эти функции реализуются при помощи *хранимых процедур, триггеров и правил*. *Хранимые процедуры* - это набор особых действий и манипуляций с данными, который хранится на сервере, причем программы-клиенты способны их выполнять. *Триггеры* - это вид хранимых процедур. Они связаны с событиями, и запускаются автоматически, как только на сервере БД с данными происходит такое событие. *Правило* - это такой тип триггера, который проверяет данные до внесения их в БД.

Основные преимущества клиент/серверных по сравнению с аналогичными информационными системами заключаются в следующем.

Во-первых, это снижение количества передаваемой по компьютерной сети информации. Это происходит потому, что сервер обрабатывает запрос клиента и в качестве результата передает клиенту только интересующую информацию, а не всю БД.

Во-вторых, преимуществом архитектуры клиент/сервер является возможность хранения правил доступа и обработки на сервере, что позволяет избежать дублирования кода в различных приложениях, использующих общую БД. Кро-

ме того, любая манипуляция с данными может быть произведена только в рамках этих правил. Часть кода, связанного с обработкой данных, как правило, реализуется в виде *хранимых процедур* сервера, что позволяет еще более ускорить работу клиентского приложения за счет уменьшения его размеров, а это в свою очередь означает, что требования к рабочим станциям могут быть не такими высокими Это в конечном итоге снижает общую стоимость информационной системы даже при использовании дорогостоящей СУБД и мощного сервера БД.

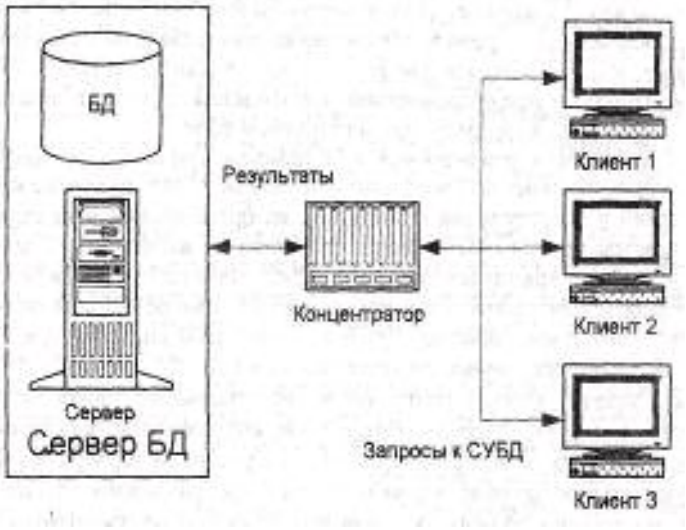

Рис.2.1. Упрощенная схема БД на платформе клиент/сервер

В третьих, современные СУБД, реализованные на платформе клиент/сервер, обладают мощными возможностями управления доступа к элементам БД, резервного копирования, архивации и параллельной обработки данных, что значительно улучшает работу.

Используя множество компьютеров, системы на платформе клиент/сервер распределяют прикладную задачу по различным рабочим станциям и серверам. Каждый элемент при этом берет на себя свою часть вычислительной нагрузки, используя информацию совместно с другими компьютерами сети, при этом мощность системы повышается без наращивания производительности одного отдельного компьютера, а получается как результат суммирования возможностей многих. Помимо всего, архитектура клиент/сервер является технологией, предоставляющей большую самостоятельность пользователям и возможность проявления творчества в создании клиентских приложений.

Подобная организация работы повышает эффективность выполнения приложений за счет использования мощности сервера, разгружает сеть, обеспечивает хороший контроль целостности данных.

В базах данных клиент/сервер возникает *дополнительная проблема — спроектировать приложение так, чтобы оно максимально использовало воз-*

можности сервера и минимально нагружало сеть, передавая через нее только минимум информации.

2.3 Реляционный подход к построению БД

2.3.1 Реляционная модель данных

Основу реляционной модели данных составляет совокупность данных, сформированных в виде таблицы. Такая форма представления данных привычна для специалиста, пользующегося различной справочной литературой.

Таблица рассматривается как непосредственное «хранилище» данных'. Традиционно в реляционных системах таблицу называют *отношением*. Строку таблицы называют *кортежем*, а столбец - *атрибутом*. При этом атрибуты имеют уникальные (в пределах отношения) имена. Количество кортежей в таблице называют *кардинальным числом*, а количество атрибутов - *степенью*. Для отношения предусматривают уникальный идентификатор, то есть один или несколько атрибутов, значения которых в одно и то же время не бывают одинаковыми. Множество допустимых однородных значений для того или иного атрибута представляет собой *домен*. Таким образом, домен можно рассмотреть как именованное множество данных, причем составные части этого множества являются логически неделимыми единицами (в качестве домена могут выступать, например, перечень фамилий сотрудников учреждения, однако не все фамилии могут присутствовать в таблице).

Отношение содержит две части - *заголовок* и *собственно содержательную часть*. Заголовок содержит конечное множество атрибутов, а содержательная часть (тело отношения) – множество пар имени атрибута и его значения. Например, на рис. 2.2 KOD, NAME и SUMM, содержащиеся в заголовке, являются атрибутами, а скажем, пары SUMM - 25.50 или KOD - 5216 являются элементами тела

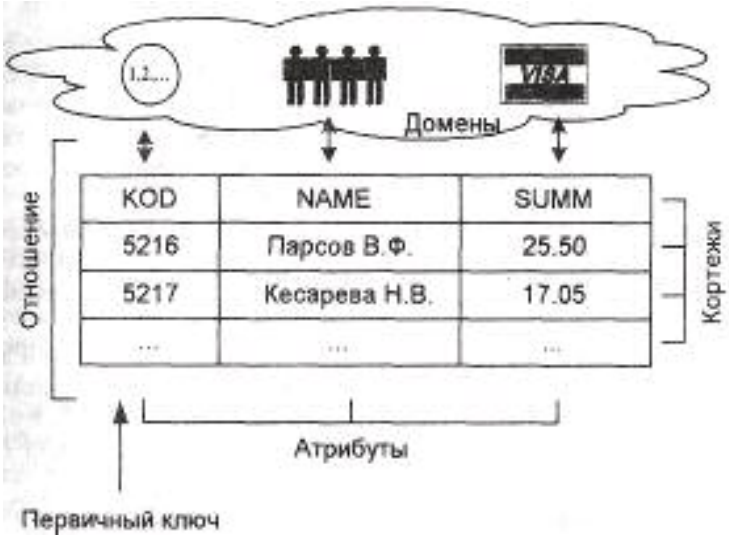

Рис.2.2. Элементы реляционной модели БД

Отношения имеют ряд основных свойств, а именно:

■ в самом общем случае в отношении не бывает двух одинаковых кортежей. Это следует из самого определения отношения, однако для некоторых СУБД в ряде случаев допускается отступление от этого свойства. Действительно поскольку в отношении имеет место первичный ключ, то одинаковые кортежи исключены;

■ кортежи не упорядочены сверху вниз - в отношении просто отсутствует понятие позиционного номера. В отношении без потери информации можно с успехом расположить кортежи в любом порядке;

■ атрибуты не упорядочены слева направо - атрибуты в заголовке отношения можно располагать в любом порядке. При этом целостность данных не нарушается. Поэтому понятия позиционного номера в отношении атрибута тоже не существует;

■ значения атрибутов состоят из логически неделимых единиц - это свойство есть следствие того, что значения берутся из доменов. Иначе, можно сказать, что отношения не содержат групп повторения, то есть являются нормализованными (о чем мы будем говорить ниже).

В реляционных системах поддерживаются несколько видов отношений.

Именованное - представляет собой переменное отношение, определяемое в СУБД путем использования операторов создания и необходимое для более удобного представления информации для пользователя.

Базовое отношение - являющееся непосредственной важной частью БД, поэтому при проектировании им дают собственное наименование.

Производное отношение - то которое было определено через другие (как правило, базовые) отношения путем использования средств СУБД.

Представление - это именованное производное отношение, которое выражается исключительно через операторы СУБД, примененные к именованным отношения. Представления физически в БД не существует.

Результат запроса - это неименованное производное отношение, содержащее данные - результат конкретного запроса. Результат запроса в БД не хранится, а существует только до тех пор, пока он необходим пользователю.

Хранимое отношение - то которое физически поддерживается в памяти компьютера. К хранимым, в большинстве случаев, относятся базовые отношения.

Исходя из вышесказанного, можно теперь определить реляционную БД как набор отношений, связанных между собой.

Связь в данном случае - это ассоциирование двух или более отношений. БД, не имеющая связей между отношениями, имеет очень простую структуру и в полной мере реляционной называться не может. Одно из основных требований к организации реляционной БД - это обеспечение возможности поиска одних кортежей по значениям других, для чего необходимо установить между ними связи. А так как в реальных информационных системах часто содержатся тыся-

чи кортежей, то теоретически между ними может быть установлено более миллиона связей. Наличие такого множества связей и определяет сложность реляционных моделей БД.

Существуют следующие основные виды связей:

- один к одному;
- один ко многим;
- многие к одному;
- многие ко многим

Связь "один к одному" предполагает, что в каждый момент времени каждому элементу (кортежу) A соответствует 0 или 1 элементов (кортежей) B (см. рис 2.3). Например, работник получает зарплату, и только одну.

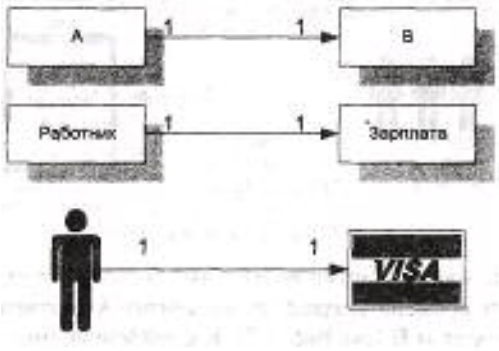

Рис 2.3 Связь один к одному

Связь "один ко многим" состоит в том, что в каждый момент времени каждому элементу (кортежу) A соответствует несколько элементов (кортежей) B (см. рис. 2.4). В качестве примера можно сказать, что в доме проживает много жильцов.

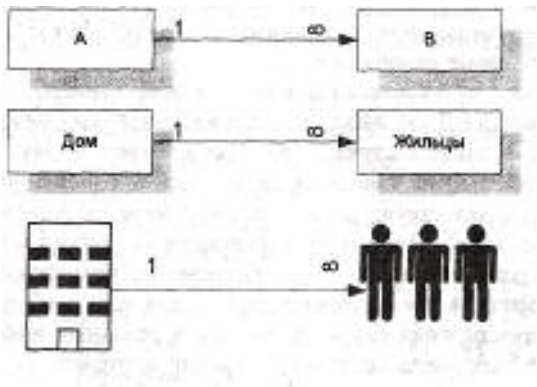

Рис 2.4 Связь один ко многим

Связь "многие к одному" предполагает, что в каждый момент времени множеству элементов A соответствует 1 элемент B. Например, несколько студентов представляют собой студенческую учебную группу (см. рис. 2.5).

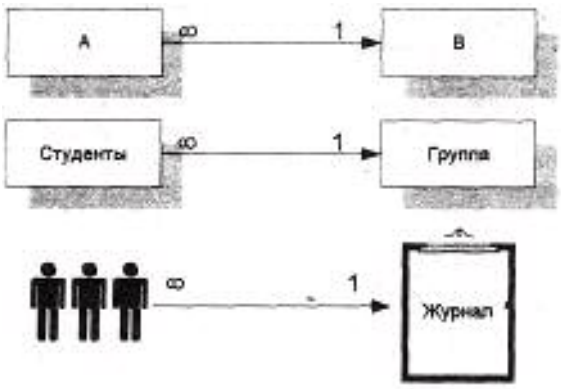

Рис 2.5 Связь многие к одному

Наконец, связь "многие ко многим" состоит в том, что в каждый момент времени множеству элементов A соответствует множество элементов B (см. рис. 2.6). К сожалению, этот тип связи в реляционных БД непосредственно не поддерживается. Примером такой связи может служить тот факт, что у студентов учебные занятия по дисциплинам ведут множество преподавателей.

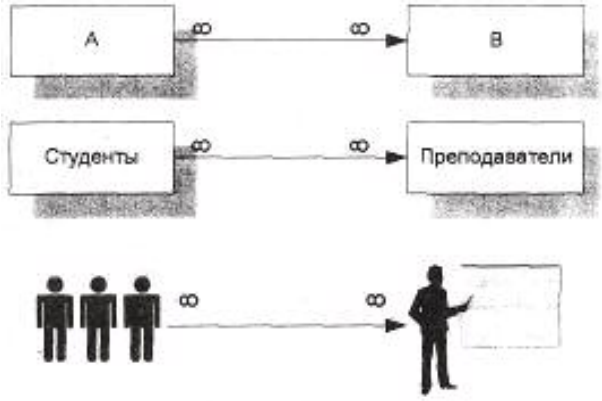

Рис 2.6 Связь многие ко многим

Помимо вышеперечисленных, еще могут существовать множественные связи между одними и теми же элементами и тренарные связи (см. рис. 2.7), которые, впрочем, могут быть выражены через уже рассмотренные.

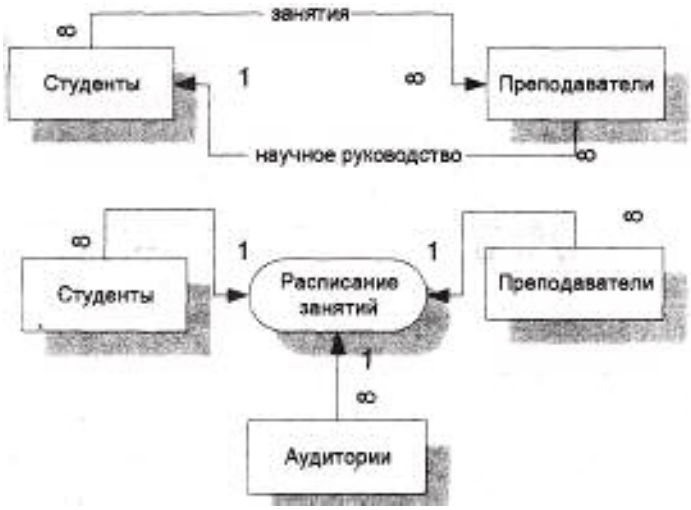

Рис 2.7 Множественные и тренарные связи

2.3.1.1 Целостность данных

В реляционных моделях вопросу целостности данных отводится особое место. Эта проблема решается за счет введения в отношение первичных и внешних ключей. Напомним, что *ключ* - это минимальный набор атрибутов, по значениям которых можно однозначно найти требуемый кортеж. Минимальность означает, что исключение из набора любого атрибута не позволяет идентифицировать кортеж по оставшимся атрибутам.

Каждое отношение обладает хотя бы одним возможным ключом. Один из них принимается за *первичный ключ*. Он служит для идентификации кортежей в самом отношении. При выборе первичного ключа следует отдавать предпочтение несоставным ключам или ключам, составленным из минимального числа атрибутов. Нежелательно также использовать ключи с длинными текстовыми значениями (предпочтительнее использовать целочисленные атрибуты). Так, для идентификации работника можно использовать либо уникальный табельный номер или номер паспорта, либо набор из фамилии, имени, отчества и номера отдела.

Не допускается, чтобы первичный ключ отношения, то есть любой атрибут, участвующий в первичном ключе, принимал неопределенное значение. В этом случае возникнет противоречивая ситуация: появится не обладающий уникальностью элемент первичного ключа.

Теперь поговорим о *внешних ключах*. Стоит иметь в виду, что если отношение С связывает отношения А и В, то оно должно включать внешние ключи, соответствующие первичным ключам отношений А и В, что представлено на рис. 2.8. Таким образом, при рассмотрении проблемы выбора способа связи отношений в БД возникает вопрос о том, каковы же должны быть внешние клю-

чи. При этом для каждого внешнего ключа необходимо решить проблему, связанную с возможностью (или невозможностью) появления во внешних ключах неопределенных значений *(NULL-значений* - значений атрибута для отсутствующей информации). Другими словами, может ли существовать некоторый кортеж в отношении, для которого неизвестен кортеж в связанном с ним отношении.

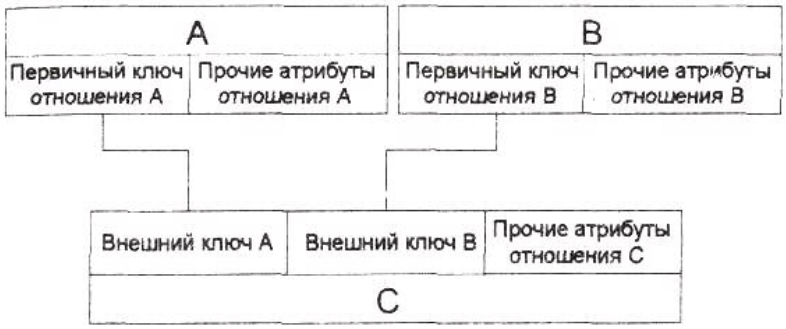

Рис. 2.8. Внешние ключи

С другой стороны, необходимо заранее обдумать вопрос о том. что произойдет при попытке удаления кортежей из отношения, на которое ссыпается внешний ключ. При этом существуют следующие вероятные возможности:

■ операция *каскадируется* - то есть удаление кортежей в отношении приводит к удалению соответствующих кортежей в связанном отношении. Например, удаление информации о фамилии, имени и т. п. сотрудника в одном отношении приводит к удалению информации о его заработной плате в другом;

■ операция *ограничивается* - то есть удаляются лишь те кортежи, для которых связанной информации в другом отношении нет. Если таковая информация имеется, то удаление осуществить нельзя. Например, удаление информации о фамилии, имени и т. п. сотрудника возможно лишь в том случае, если информация о его заработной плате в связанном отношении отсутствует.

Наконец, нужно предусмотреть технологию того, что будет происходить при попытке обновления первичного ключа отношения, на которое ссылается некоторый внешний ключ. Здесь имеются те же возможности, как и при удалении:

■ операция *каскадируется* - то есть при обновлении первичного ключа происходит обновление внешнего ключа в связанном отношении. Например, обновление первичного ключа в отношении, где хранится информация о сотруднике приводит к обновлении внешнего ключа в отношении с информацией о его заработной плате;

■ операция-*ограничивается* - то есть обновляются лишь те первичные ключи, для которых связанной информации в другом отношении нет. Если таковая информация имеется, то обновление сделать нельзя. Например, обновле-

ние первичного ключа в отношении, где хранится информация о сотруднике, возможно в том случае. если информация о его заработной плате в связанном отношении отсутствует.

Таким образом, для каждого внешнего ключа в БД должны предусматриваться не только атрибут или комбинация атрибутов, составляющих этот внешний ключ, и отношение, идентифицируемое этим ключом, но также и варианты "поведения" БД в рассмотренных выше случаях.

2.3.2 Реляционная алгебра

Формальной основой реляционной модели БД являются *реляционная алгебра*, основанная на теории множеств и рассматривающая специальные операторы над отношениями, и *реляционное исчисление*, базирующееся на математической логике.

Основных операторов в реляционной алгебре восемь, и схематически их можно представить так, как это показано на рис. 2.9

Надо отметить, что реляционная алгебра обладает большой мощностью - сложные запросы к БД могут быть выражены с помощью одного выражения. Именно по этой причине эти механизмы включены в реляционную модель данных. Конкретный язык манипулирования реляционными БД называется *реляционно-полным*, если любой запрос, выражаемый с помощью одного выражения реляционной алгебры или одной формулы реляционного исчисления, может быть выражен с помощью одного оператора этого языка.

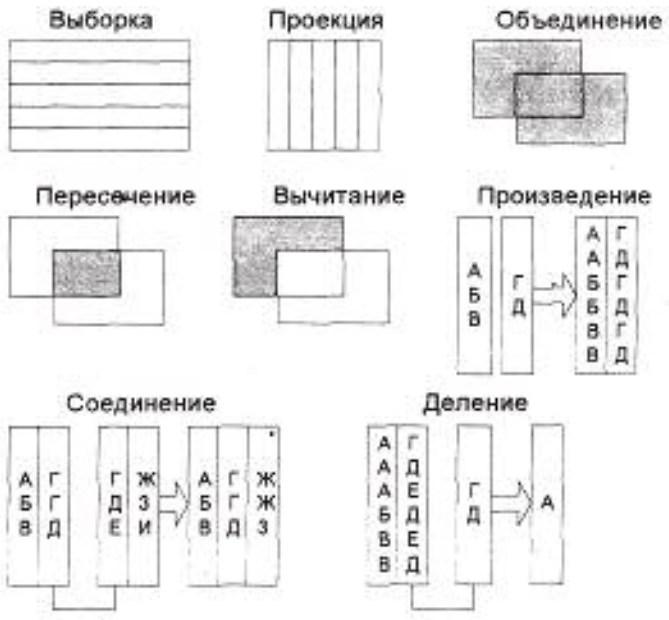

Рис.2.9. Основные операторы реляционной алгебры

Реляционная алгебра обладает важным свойством - она замкнута относительно понятия отношения. Это означает, что выражения реляционной алгебры выполняются над отношениями реляционных БД и результаты их вычисления также представляют собой отношения. Поэтому любое выражение может быть представлено как отношение, что позволяет использовать его в других выражениях реляционной алгебры.

Основная идея реляционной алгебры состоит в том, что средства манипулирования отношениями, рассматриваемыми как множества, основаны на традиционных множественных операциях, дополненных некоторыми специфичными операциями для БД.

Существует много подходов к определению реляционной алгебры, которые различаются набором операций и способами их интерпретации, но в принципе все они более или менее равнозначны.

Опишем вариант алгебры, который был предложен Коддом. В этом варианте, как уже было показано выше, набор алгебраических операций состоит из восьми основных:

- выборка отношения;
- проекция отношения;
- объединения отношений;
- пересечение отношений;
- вычитание отношений;
- произведение отношений.
- соединение отношений;
- деление отношений.

Эти операции можно объяснить следующим образом:

- результатом *выборки* отношения по некоторому условию является отношение, которое включает только те кортежи первоначального отношения, которые удовлетворяют этому условию;
- при осуществлении *проекции* отношения на заданный набор его атрибутов будет получено отношение, кортежи которого взяты из соответствующих кортежей первона-чального отношения;
- при выполнении операции *объединения* двух отношений будет получено отношение, включающее все кортежи, входящие хотя бы в одно из участвующих в операции отношений;
- при выполнении операции *пересечения* двух отношений получается отношение, включающее все кортежи, входящие в оба первоначальные отношения;
- при выполнении операции вычитания одного отношения из другого по лучается отношение, которое включает все кортежи, входящие во второе отношение но не входящие в отношение, являющееся первым;
- при выполнении прямого *произведения* двух отношений получается отношение, кортежи которого являются сочетанием кортежей первого и второго отношения;

■ при *соединении* двух отношений по некоторому условию образуется результирующее отношение, кортежи которого являются сочетанием кортежей первого и второго отношений, удовлетворяющим этому условию;

■ операция реляционного *деления* имеет два операнда - бинарное (т. е. состоящее из двух атрибутов) и унарное (содержит один атрибут) отношения. Результатом операции является отношение, состоящее из кортежей, включающих значения первого атрибута кортежей первого отношения, для которых множество значений второго атрибута совпадает со множеством значений второго отношения.

Помимо вышеперечисленных, есть ряд особых операций, характерных для работы с БД:

■ *переименование отношения;*

■ *присваивание*

В результате операции *переименования* получается отношение, набор кортежей которого совпадает с телом первоначального отношения, но имена атрибутов изменены;

Операция *присваивания* позволяет сохранить результат вычисления реляционного выражения в существующем отношении БД.

Таким образом, благодаря тому, что результатом реляционной операции является некоторое отношение, то имеется возможность образовывать реляционные выражения, в которых вместо первоначального отношения (отношения-операнда) будет использоваться вложенное реляционное выражение.

Приведенные выше объяснения 8 основных операций являются общими с точки зрения теории множеств. В реляционной алгебре эти операции имеют некоторые специфические ограничения. Кратко рассмотрим их

Операция выборки требует наличия двух отношений: первоначального отношения-операнда и простого условия ограничения. В результате выполнения операции выборки производится отношение, заголовок которого совпадает с заголовком отношения-операнда, а в тело входят те кортежи отношения-операнда, которые удовлетворяют значениям условия ограничения.

Операция взятия проекции также требует наличия двух операндов - проецируемого отношения А и списка имен атрибутов, входящих в заголовок отношения А. *Результатом проекции отношения А по списку атрибутов* a_1, a_2, ..., a_n будет отношение, заголовком которого является множество атрибутов a_1, a_2,...., a_n. Тело результата будет состоять из кортежей, для которых в отношении А имеется кортеж, атрибут a_i, которого имеет значение v_i, атрибут a_2 имеет значение v_2, ..., атрибут $a_п$ имеет значение v_n,По сути, при выполнении операции проекции определяется "вертикальная" вырезка отношения-операнда с удалением возникающих кортежей-дубликатов.

При выполнении операции объединения результатом должно являться отношение. Если допустить в реляционной алгебре возможность объединения произвольных двух отношений с разными наборами атрибутов, то результатом такой операции будет множество, однако множество разнотипных кортежей, то

есть, не отношение. Если исходить из требования замкнутости реляционной алгебры относительно понятия отношения, то такая операция объединения является бессмысленной. Это приводит к появлению понятия *совместимости отношений по объединению*: два отношения совместимы по объединению в том и только в том случае, когда обладают одинаковыми заголовками Более точно это означает, что в заголовках обоих отношений содержится один и тот же набор имен атрибутов, и одноименные атрибуты определены на одном и том же домене.

Приведенные рассуждения в равной мере относится к операциям пересечения и вычитания. При условии того, что два отношения совместимы по объединению, то при обычном выполнении над ними операций объединения, пересечения и вычитания результатом операции является отношение с корректно определенным заголовком, совпадающим с заголовком каждого из отношений-операндов. Если же два отношения не полностью совместимы по объединению, то есть совместимы во всем, кроме имен атрибутов, то до выполнения операции типа соединения эти отношения можно сделать полностью совместимыми по объединению путем применения операции переименования.

Операция прямого произведения двух отношений вызывает новые проблемы. В теории множеств прямое произведение может быть получено для любых двух множеств. Элементами результирующего множества будут являться пары, составленные из элементов первого и второго множеств. Поскольку отношения являются множествами, то и для любых двух отношений возможно получение прямого произведения, однако, результат не будет отношением. Элементами результата будут являться не кортежи, а *пары кортежей*. Поэтому в реляционной алгебре используется специальная форма операции взятия прямого произведения -*расширенное прямое произведение отношений* При взятии расширенного прямого произведения двух отношений элементом результирующего отношения является кортеж, формирующийся при слиянии одного кортежа первого отношения и одного кортежа второго отношения. Тут же возникает вторая проблема, связанная с получением корректно сформированного заголовка результирующего отношения. Это приводит к необходимости ввода понятия *совместимости отношений по взятию расширенного прямого произведения.* Два отношения *совместимы по взятию прямого произведения* в том и только в том случае, если множества имен атрибутов этих отношений не пересекаются. Любые два отношения могут быть преобразованы к совместимому виду по взятию прямого произведения путем применения операции переименования к одному из этих отношений.

Операция соединения, называемая иногда *соединением по условию*, требует наличия двух операндов - соединяемых отношений и третьего операнда - простого условия. Пусть соединяются отношения A и B. Как и в случае операции выборки, условие соединения C имеет вид либо (a comp-op b), либо (a comp -op const), где a и b - имена атрибутов отношений A и B, const - литерально заданная константа, a comp -op - допустимая в данном контексте операция сравне-

ния. Тогда по определению результатом операции соединения является отношение, получаемое путем выполнения операции ограничения по условию C прямого произведения отношений A и B.

Имеется важный частный случай соединения - *естественное соединение. Операция соединения* называется *операцией естественного соединения*, если условие соединения имеет вид (a = b), где a и b - атрибуты разных операндов соединения Этот случай важен потому, что он особо часто встречается на практике и для него существуют эффективные алгоритмы реализации в СУБД. Операция естественного соединения применяется к паре отношений A и B, обладающих общим атрибутом R, то есть атрибутом с одним и тем же именем и определенным на одном и том же домене. Пусть ab обозначает объединение заголовков отношений A и B. Тогда естественное соединение A и B - это спроектированный на ab результат соединения A и B. Операция естественного соединения не включается прямо в состав набора операций реляционной алгебры, но она имеет очень важное практическое значение.

Операция деления отношений нуждается в более подробном объяснении, поскольку наиболее трудная для понимания. Пусть заданы два отношения - A с заголовком {a,, a_2,..., a_n, $b_ь$ b_2,..., b_m} и B с заголовком {$b_ь$ b_2, ..., b_m}. Будем полагать, что атрибут b, отношения A и атрибут b, отношения B обладают одним и тем же именем и определены на одном и том же домене. Назовем множество атрибутов {a_i} составным атрибутом a, множество атрибутов {b_i,} - составным атрибутом b. После этого будем говорить о реляционном делении бинарного отношения A(a,b) на унарное отношение B(b).

Результатом деления A на B является унарное отношение C(a), состоящее из таких кортежей v, что в отношении A имеются кортежи <v, w> которые во множестве значений {w} включают множество значений атрибута b в отношении B.

Поскольку деление наиболее трудная операция, поясним ее примером. Пусть в БД студентов имеются два отношения: СТУДЕНТЫ (ФИО, НОМЕР) и ИМЕНА (ФИО), причем унарное отношение ИМЕНА содержит все фамилии, которыми обладают студенты университета. Тогда после выполнения операции реляционного деления отношения СТУДЕНТЫ на отношение ИМЕНА будет получено унарное отношение, содержащее номера студенческих билетов, принадлежащих студентам со всеми возможными в этом университете фамилиями.

2.3.3 Реляционное исчисление

Допустим, что имеется БД. обладающая следующей структурой: отношение СТУДЕНТЫ (СТУД_НОМ., СТУД_ЙМЯ, СТУД_СТИП, ГР_НОМ) и отношение ГРУППЫ (ГР_НОМ., ГР_КОЛ., ГР_СТАР). Предположим, что необходимо узнать имена и номера студенческих билетов у студентов, являющихся старостами групп с количеством студентов больше 25.

Если бы для формулировки такого запроса использовалась реляционная алгебра, то мы получили бы алгебраическое выражение, которое читалось бы, например, следующим образом:

- выполнить соединение отношений СТУДЕНТЫ и ГРУППЫ по условию СТУД_НОМ = ГР_СТАР:
- ограничить полученное отношение по условию ГР_КОЛ>25;
- спроецировать результат предыдущей операции на атрибут СТУД_ИМЯ. СТУД_НОМ.

Здесь пошагово сформулирована последовательность выполнения запроса к БД, каждый из которых соответствует одной реляционной операции. Если же сформулировать тот же запрос с использованием реляционного исчисления, то мы получили бы формулу', которую можно было бы прочитать, например, следующим образом: Выдать СТУД_ИМЯ и СТУД_НОМ для таких студентов, чтобы существовала группа с таким же значением ГР_СТАР и значением ГР_КОЛ, большим 25.

Во второй формулировке мы указали лишь характеристики результирующего отношения, но ничего не сказали о способе его формирования. В этом случае СУБД должна сама решить, что за операции и в каком порядке нужно выполнить над отношениями СТУДЕНТЫ и ГРУППЫ. Оба рассмотренных в примере способа на самом деле эквивалентны, и существуют не очень сложные правила преобразования одного в другой.

Базисными понятиями реляционного исчисления являются *понятие переменной с определенной для нее областью допустимых значений* и *понятие правильно построенной формулы*, *опирающейся на переменные и специальные функции*.

В зависимости от того, что является областью определения переменной, различаются *исчисление кортежей* и *исчисление доменов*. В *исчислении кортежей* областями определения переменных являются отношения БД, то есть допустимым значением каждой переменной является кортеж некоторого отношения. В *исчислении доменов* областями определения переменных являются домены, на которых определены атрибуты отношений БД, то есть допустимым значением каждой переменной является значение некоторого домена.

Для определения кортежной переменной используется оператор RANGE. Например, для того, чтобы определить переменную СТУДЕНТ, областью определения которой является отношение СТУДЕНТЫ, нужно употребить конструкцию

RANGE СТУДЕНТ IS СТУДЕНТЫ

Из этого определения следует, что в любой момент времени переменная СТУДЕНТ представляет некоторый кортеж отношения СТУДЕНТЫ. При использовании кортежных переменных в формулах можно ссылаться на значение атрибута переменной. Например, для того, чтобы сослаться на значение атрибута СТУД_ИМЯ переменной СТУДЕНТ, нужно употребить конструкцию СТУДЕНТ.СТУД. ИМЯ.

Правильно построенные формулы служат для выражения условий, накладываемых на кортежные переменные. В основе таких формул лежат простые сравнения, представляющие собой операции сравнения значений атрибутов переменных или литерально заданных констант. Например, конструкция "СТУДЕНТ СТУД_НОМ= 123456" является простым сравнением. Более сложные варианты правильно построенных формул реализуются с помощью логических связок NOT, AND. OR и IF ... THEN. Наконец, допускается построение правильно построенных формул с помощью *кванторов.* Если F - это правильно построенная формула, в которой участвует переменная **var**, то конструкции *EXISTS var (F)* и *FORALL var (F)* являются правильными. Здесь квантор *EXISTS* обозначает "существование", а *FORALL* - "для всех кортежей".

Переменные, входящие в правильно построенные формулы, могут быть свободными или связанными. Все переменные, входящие в состав формулы, при построении которой не использовались кванторы, являются *свободными.* Фактически, это означает, что если для какого-то набора значений свободных кортежных переменных при вычислении формул получено значение "истина", то эти значения кортежных переменных могут входить в результирующее отношение. Если же имя переменной использовано сразу после квантора при построении формул вида *EXISTS var (F)* или *FORALL var (F)* то здесь, и во всех формулах, где она использована, var - *связанная переменная*. При вычислении значения такой правильно построенной формулы используется не одно значение связанной переменной, а вся ее область определения.

Пусть СТУД1 и СТУД2 - две кортежные переменные, определенные на отношении СТУДЕНТЫ. Тогда, формула

EXISTS СТУД2(СТУД1. СТУД_СТИП>СТУД2.СТУД_СТИП)

для текущего кортежа переменной СТУД1 принимает значение "истина" только в том случае, если во всем отношении СТУДЕНТЫ найдется такой кортеж, связанный с переменной СТУД2, что значение его атрибута СТУД_СТИП удовлетворяет внутреннему условию сравнения.

Правильно построенная формула

FORALL СТУД2(СТУД1. СТУД_СТИП > СТУД2.СТУД_СТИП)

для текущего кортежа переменной СТУД1 принимает значение "истина" только в том случае, если для всех кортежей отношения СТУДЕНТЫ, связанных с переменной СТУД2. значения атрибута СТУД_СТИП удовлетворяют условию сравнения.

Таким образом, правильно построенные формулы обеспечивают средства выражения условия выборки из отношений БД. Чтобы можно было использовать реляционное исчисление для реальной работы с БД, требуется еще один компонент, который определяет набор и имена столбцов результирующего отношения. Этот компонент называется *целевым списком.*

Целевой список строится из целевых элементов, каждый из которых может иметь следующий вид:

■ var.attr. где var - имя свободной переменной соответствующей формуле, а attr - имя атрибута отношения, на котором определена переменная var;

■ var, что эквивалентно наличию подсписка $var.attr_1$, $var.attr_2$, ... $var.att_n$, где $attr_1$, $attr_2$. ... $attr_n$ включает имена всех атрибутов определяющего отношения;

■ new name = var.attr; new name - новое имя соответствующего атрибута результирующего отношения.

Последний вариант требуется в тех случаях, когда в формуле используются несколько свободных переменных с одинаковой областью определения.

В исчислении доменов областью определения переменных являются не отношения, а домены. Применительно к БД СТУДЕНТЫ-ГРУППЫ можно говорить, например, о доменных переменных ИМЯ (значения домена - допустимые имена) или НОМ_СТУД (значения домена - допустимые номера студентов).

Основным отличием исчисления доменов от исчисления кортежей является наличие дополнительного набора предикатов (см. ниже), позволяющих выражать так называемые условия членства. Если R - это n-арное отношение с атрибутами a_1, a_2, .., $a_п$, то условие членства имеет вид

$$R \quad (a_{i1}: vi_1, \quad a_{i2}:v_{i2}, \quad ..., \quad a_{im}:v_{in}) \quad (m <= n),$$

где v_{ij} - это либо литерально задаваемая константа, либо имя кортежной переменной. Условие членства принимает значение ИСТИНА только в том случае, если в отношении R существует кортеж, содержащий соответствующие значения указанных атрибутов. Если v_{ij} - константа, то на атрибут a_{ij} задается жесткое условие, не зависящее от текущих значений доменных переменных; если же v_{ij} - имя доменной переменной, то условие членства может принимать различные значения при разных значениях этой переменной.

. *Предикатом* принято называть некую логическую функцию, которая для некоторого аргумента возвращает значение ИСТИНА или ЛОЖЬ. Отношение может быть рассмотрено как *предикат с аргументами*, являющимися атрибутами рассматриваемого отношения. Если заданный конкретный набор кортежей присутствует в отношении, то предикат выдаст истинный результат, в противном случае - ложный.

Во всех остальных отношениях формулы и выражения исчисления доменов выглядят похожими на формулы и выражения исчисления кортежей. Реляционное исчисление доменов положено в основу большинства языков запросов, основанных на использовании форм.

2.4 Иерархический и сетевой подходы

В реляционном исчислении и алгебре полностью отсутствуют указания на то каким образом производить поиск необходимых данных. Оперирование отношениями (таблицами) предполагает просмотр всех записей. Когда БД велика,

то невозможно производить полный просмотр всех ее записей. Поэтому необходимо предварительное упорядочивание и объединение в группы записей по признакам поиска.

Для упорядочивания записей и организации поиска нужных записей используется *ключ и связи*.

Ключ – это уникальное имя записи, в качестве которого может выступать как один какой-либо атрибут записи (*простой ключ)*, так и совокупность нескольких атрибутов (*составной ключ)*. С помощью ключа производится идентификация каждой конкретной записи, а также упорядочение записей в файле. Упорядочение по ключу может быть либо прямым, либо выполнено с помощью хеш-функции.

Прямое упорядочение предполагает лексикографическое расположение записей: записи могут располагаться либо в порядке увеличения значения ключа, либо в алфавитном порядке.

Хеш-функция производит пересчет ключа в адрес записи в файле. Эта операция выполняется СУБД всякий раз при поиске нужной записи по ключу.

Связи позволяют осуществить группирование записей во множества (сегменты записей), а так же указывать взаимоотношения между этими множествами (сегментами). На практике связь реализуется, как правило, в виде указателя.

Так, если необходимо указать связь между записью а1 в отношении А с записью в1 отношения В, то в состав записи а1 необходимо включить указатель на запись в1. В случае присутствия и обратной связи запись в1должна так же содержать указатель на запись а1.

При построении иерархических и сетевых баз данных используются все перечисленные ранее типы связей. Графически разным типам связей соответствуют обозначения в виде различных стрелок: «→» - связь с одной записью; «→→» - с несколькими записями. Взаимные связи имеют следующие названия: «←→» - «один к одному», «←→→» - «один ко многим», «←←→→» - «многие ко многим». На рис.2.10 записи изображены в виде прямоугольников, а сегменты представлены кружками. Например, обозначение, приведенное на рис.2.10, в следует понимать так: каждая запись сегмента А связана с некоторой группой записей сегмента В, а каждая запись из сегмента В связана лишь с одной записью сегмента А.

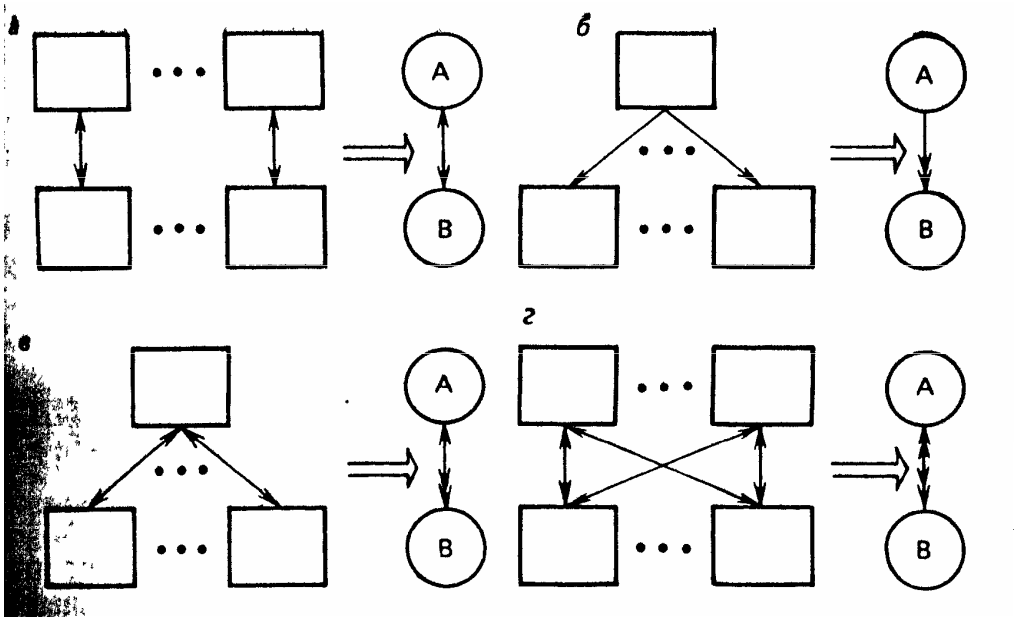

Рис2.10. Типы и обозначения связей
а-«один к одному»; *б, в*-«один ко многим»; *г*-«многие ко многим»

С помощью указанных обозначений строится граф логической схемы БД, вершины которого – сегменты, а дуги – обозначения типов связей между сегментами.

2.4.1 Иерархический подход.

Иерархическая БД имеет граф логической схемы в виде дерева, а тип связей соответствует «один ко многим» (рис.2.10,б). Пример логической схемы иерархической БД приведен на рис. 2.11. В иерархической БД связи направлены только от верхних сегментов к нижним, обратные указатели отсутствуют. Это объясняется принципиальным свойством иерархического представления данных: каждая запись приобретает смысл лишь тогда, когда она рассматривается в своем контексте, т. е. любая запись не может существовать без предшествующей ей записи по иерархии.

Таким образом, БД, основанная на иерархической модели, состоит из упорядоченного набора деревьев. Каждое дерево состоит из одного "корневого" (предок) и упорядоченного набора из нуля или более связанных с ним поддеревьев (потомки). Целостность связи между ними поддерживается автоматически.

При поиске в иерархической БД необходимо указывать значение ключа на каждом уровне иерархии.

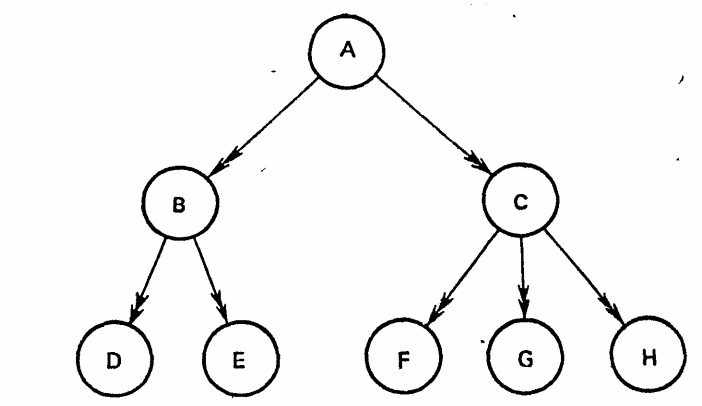

Рис.2.11. Пример логической схемы иерархической базы данных

Так, для доступа к записи из множества **G** (рис.2.11) должны быть последовательно указаны ключи записей из множеств **A**, **C** и **G**.

В таких БД поддерживаются следующие операторы манипулирования данными:

- найти дерево БД по заданному признаку;
- перейти от одного дерева к другому;
- перейти к записи внутри дерева или в порядке обхода иерархии (сверху вниз, слева направо):
- вставить новую запись в указанную позицию.
- удалить текущую запись.

Реляционная БД всегда может быть преобразована в иерархическую. Однако конкретный вид логической схемы будет зависеть от типа запросов.

Пусть имеется реляционная база представленная следующими отношениями.

Пациент:

ФИО	Пол	Возраст	Рост	Вес
1. 2. 3. 4.				

Участок:

№ участка	ФИО врача	Улица	Номера домов
1	Курильчик	Руссиянова	Нечетные
2	Кураева	Руссиянова	Четные
3	Коваленко	Руссиянова	Нечетные
4	…	Никифорова	Четные
5	…	Никифорова	Нечетные

6	…	Никифорова	Четные
7	…	Шагаева	Нечетные
8	…	Шугаева	Четные
9	…	Шугаева	Нечетные

Посещения:

№ участка	ФИО пациента	Дата посещения	Вид заболевания

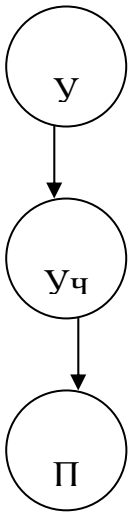

Рис. 2.12.Логическая схема для данного запроса:
У – улица; Уч – участок; П – посещения.

Тогда для запроса типа «Найти номера участков, расположенных на улице … », или «Перечислить ФИО пациентов, проживающих на данном участке и посещающих поликлинику» можно построить иерархическую БД (Рис.2.12). Для этого из записей, содержащие данные об участках, расположенных на определенной улице, образуем сегмент «улица». В этом случае реализация связей между двумя первыми сегментами будет следующая (Рис.2.13).

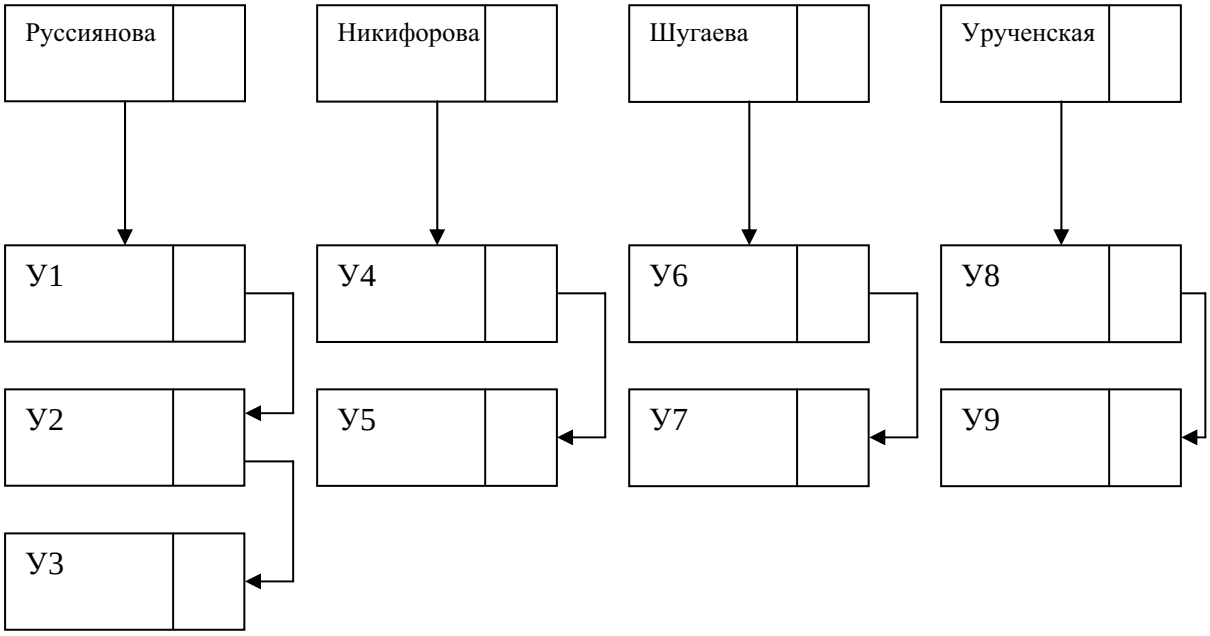

Рис.2.13. Схема реализации связей между сегментами «Улица» и «Участок»

Каждая запись сегмента «улица» связана с группой записей сегмента «участок». В данной конкретной реализации запись из сегмента «улица» ссылается на первую запись из сегмента «участок». Сами же записи сегмента «участок» объединены с помощью указателей в списки. Поиск с такой БД осуществляется по составленному ключу. Например, составляют ключ: «Руссиянова.У1» - позволяет найти сведения об участке номер 1, расположенный на улице Руссиянова. По организации связей между уровнями «участок – посещение» возникает проблема избыточности, которая заключается в необходимости повторения записей «ФИО пациента» и «вид заболевания». В конкретной реализации возможно избежать многократность повторения записи «ФИО пациента» и «вид заболевания» введением указателя на эти записи, хранящиеся отдельно. Однако значение указателя на запись «ФИО пациента» и «вид заболевания» должно быть разрознено, что усложняет решение проблемы целостности БД.

Трудности в иерархической БД возникают при изменении запроса. Так если появился запрос: «найти номера участков, на которых имело бы место то или иное заболевание», то пропадают преимущества предыдущего упорядочения. Для выполнения данного запроса необходимо просмотреть записи участков и связанные с ними записи вида заболевания. Для реализации такого запрөса было бы целесообразно переупорядочить БД и построить новую иерархию.

2.4.2 Сетевой подход.

Необходимость в организации различного упорядочения записей в БД с целью удовлетворения разных типов запросов привела к разработке *сетевых баз данных*. В сетевой модели данных в принципе разрешены любые группиро-

вания записей и организация произвольных связей между ними. Однако на практике целесообразно введение некоторых ограничений.

Набор – основная конструкция сетевых моделей, представляющая собой поименованное двухуровневое дерево. Каждое такое дерево состоит из одного "корневого" (предок) и упорядоченного набора из нуля или более связанных с ним поддеревьев (потомки). С помощью двухуровневых деревьев могут быть построены многоуровневые деревья и большинство сетевых структур. Если рассматривать логическую схему, то набор может интерпретироваться как имя связи между двумя сегментами записей типа «один ко многим». *Экземпляр набора* – конкретная реализация такой связи. На рис. 2.14 приведен пример набора и его двух экземпляров. Каждый экземпляр набора содержит одного *владельца* и нескольких *членов набора*. Например, на рис. 2.14 владельцем 1-го экземпляра набора является запись a_1, а членами набора – записи $b_1, \ldots, b_3$.

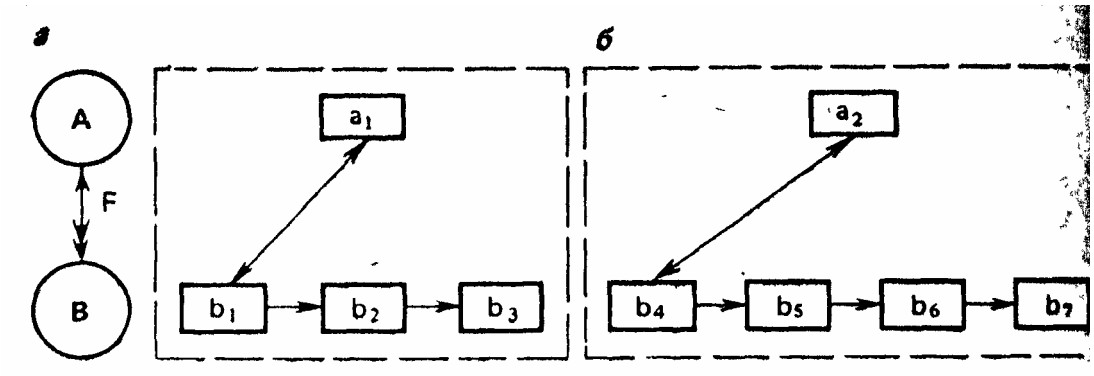

Рис.2.14. Пример набора (*а*) и экземпляра набора (*б*)
A, **B**-сегменты; a_1, a_2-записи сегмента **A**; b_1, b_2, …, b_7-записи сегмента **B**;
F-имя набора .

В сетевой модели данных допускается, чтобы записи одного сегмента были владельцами нескольких наборов, и, наоборот, записи некоторого набора могут быть членами различных наборов. На рис.2.15 показаны соответствующие конструкции логической схемы.

В логической схеме сетевых БД не разрешены связи типа «многие ко многим». Однако с помощью вышеуказанных свойств набора можно описать такую связь введением дополнительного сегмента записей, задающих взаимоотношение между исходными сегментам, и определить на них два набора так, как это показано на рис.2.16. Сегмент **C** содержит записи в форме адресных таблиц. Записи сегмента **C** входят в различные наборы для сегментов **A** и **B**. Если необходимо найти записи в **B**, связанные с данной записью из **A**, то с помощью набора **F**1 находят указатели на записи в сегменте **C**, а затем по набору **F2** нахо-

дят искомые записи в **B**. Если же требуется выполнить обратный поиск из **B** в **A**, то необходимо воспользоваться сначала набором **F**2, а затем набором **F**1.

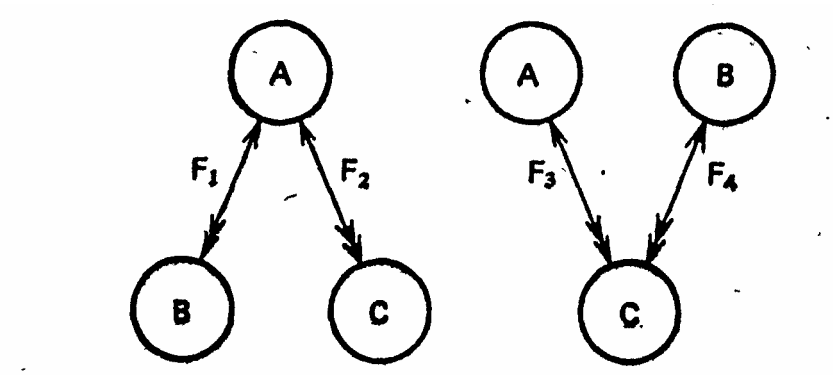

Рис.2.15 Допустимые варианты построения наборов.

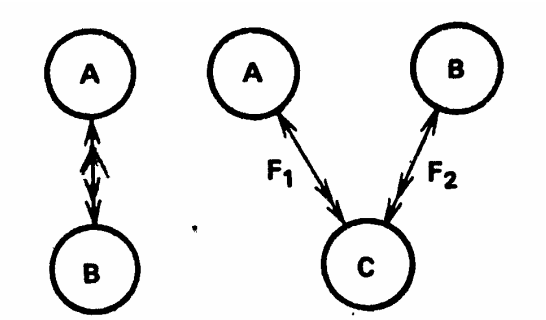

Рис.2.16. Представление отношения «многие ко многим».

Пример логической схемы сетевой БД приведен на рис. 2.16.

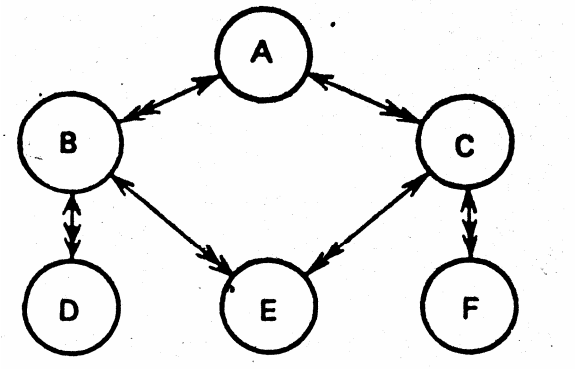

Рис. 2.16. Пример логической схемы сетевой БД

Таким образом, в БД с сетевой структурой данных, в отличие от иерархической, поддеревья (потомки) могут иметь любое число корневых деревьев (предков). Фактически сетевая БД состоит из набора записей и множества связей

между этими записями. Примерный перечень операций для сетевых БД может быть следующим:

- найти запись по заданному признаку;
- перейти от предка к потомку по указанной связи;
- перейти от потомка к предку по некоторой связи;
- создать новую запись или удалить существующую;
- модифицировать заданную запись;
- включить в связь или исключить из связи;
- переставить в другую связь.

2.5 Инвертированные базы данных

Ознакомление с различными моделями данных показало, что поиск необходимой информации требует значительных затрат времени даже для иерархических СУБД, особенно при больших объемах баз данных. Однако если удается выделить совокупность признаков, по которым формируется запрос, то можно предложить способ организации баз данных, значительно сокращающий время поиска затребованной информации. В основе такого способа лежит понятие *инвертированного списка*.

Инвертированный список представляет собой таблицу, в первом столбце которой помещены значения данного признака, а во втором – указатели на соответствующие записи в БД. Допустим, в примере базы данных, приведенной в таблице «Заболевания», в записях о видах заболеваний необходимо выявить заболевания, имеющие одинаковое количество обращений.

Таблица- Заболевания

№	№ участка	Вид заболевания	Количество обращений
1	1	ОРВИ	16
2	1	Ангина	4
3	2	ОРВИ	2
4	2	Грипп	3
5	3	Бронжит	16
6	3	Грипп	16
7	4	Ангина	11
8	4	ОРВИ	4

Тогда все записи о видах заболеваний могут быть скомпонованы плотно в памяти друг за другом и иметь порядковые номера, соответствующие их последовательности в таблице «Заболевания». Наряду с этим создается инвертированный список в таком виде:

Количество обращений	2	3	4	11	16
Список указателей	У3	У4	У2,У8	У7	У1,У5,У6

Обозначение УN представляет собой изображение указателя на N–ю запись. Таким образом, инвертированный список как бы заранее хранит ответ на вопрос «Назвать виды заболевания, имеющих определенное количество обращений пациентов». В примере количество обращений выступает в качестве признака в запросе. Конечно, можно выделить и другие признаки: например, на каком участке имеют место обращения по поводу того или иного заболевания. Для каждого признака поиска должен быть построен свой инвертированный список.

Наличие нескольких инвертированных списков позволяет строить запрос в виде некоторой логической функции от совокупности признаков (дизъюнкции, конъюнкции, отрицание).

Для поиска необходимой записи в этом случае должны быть выполнены следующие действия:

1) выделить в запросе признаки поиска;
2) определить вид логической функции между запрашиваемыми признаками;
3) для каждого признака в нужном инвертированном списке найти множество указателей на записи;
4) в соответствии с видом логической функции произвести операции над множествами указателей.

Соответствие операций над множествами указателей виду логической функции должно быть принято из разделов по реляционной алгебре. Так, логической функции дизъюнкции ставится в соответствие операция объединения множеств, конъюнкции – операция пересечения множеств, отрицанию – операция дополнения и т.д. Для запроса «Какие заболевания имеют количество посещений 16 и имели место на участке№1» в ходе поиска будут выполнены следующие действия:

1) в запросе два признака, количество посещений и № участка;
2) вид логической функции в запросе – конъюнкция;
3) для признака «количество посещений» в инвертированном списке значению 16 будет соответствовать множество указателей {У1, У5, У6}; признаку «№ участка» в соответствующем инвертированном списке для значения «№ участка 1» - множество указателей {У1, У2}.
4) над найденными двумя множествами указателей производится операция перечисления. В итоге получается искомый результат: {У1}. По этому номеру в файле записей будет найдена запись о заболевании ОРВИ.

Если для всех записей, хранящихся в базе данных, созданы инвертированные списки для возможных вариантов запросов, то такая база данных называется *инвертированной*. Инвертированные БД широко используются в информа-

ционно-поисковых системах (ИПС), предназначенных в основном для хранения текстовых документов. Признаки, по которым отыскивается необходимый документ в ИПС, называются *дескрипторами*. Для каждого дескриптора в ИПС строится инвертированный список, содержащий все возможные значения дескриптора и соответствующие им множества указателей на документы. Запрос в ИПС имеет вид логического высказывания относительно значений дескрипторов и их взаимосвязи. Так если в качестве документов выступают отчеты поликлиник, то дескрипторами могут быть такие понятия, как год подготовки отчета, объект исследования, объем отчета и др. Запрос мог бы выглядеть, например, так: «Найти отчеты подготовленные не позднее 2001 года., посвященные исследованию заболеванию ОРВИ».

Отчеты по НИР, как уже указывалось, слабо структурированы. Однако любой отчет может быть охарактеризован некоторой совокупностью признаков. Так, техническое задание на проведение исследований влияния условий проживания на результаты амбулаторного лечения описывается дескрипторами: техзадание, условия проживания, результаты амбулаторного лечения. Для хранения текучих отчетов исследований необходимо заранее заготовить совокупность дескрипторов, которые наряду с содержанием отражали бы и поэтапность составления соответствующих документов. Например, могут быть указаны этапы или стадии процесса исследования.

2.6 Принципы построения реляционных баз данных

Реляционная база данных представляет собой набор таблиц, хотя в базу данных могут входить также и ряд других объектов. *Таблицу* можно представлять себе как обычную двумерную таблицу с характеристиками (атрибутами) какого-то множества объектов. Таблица имеет *имя* — идентификатор, по которому на нее можно сослаться. В табл. 2.6.1 приведен пример фрагмента подобной таблицы с именем Pers, содержащей сведения о сотрудниках некоторой организации.

Таблица 2.6.1. Пример таблицы данных о сотрудниках Pers

Номер	Отдел	Фамилия	Имя	Отчество	Год рождения	Пол	Характеристика	Фотография
Num	Dep	Fam	Nam	Par	Year_b	Sex	Charact	Photo
1	Бухгалтерия	Иванов	Иван	Иванович	1950	м		

2	Цех 1	Петров	Петр	Петрович	1960	м	...	
3	Цех 2	Сидоров	Сидор	Сидорович	1955	м	...	...
4	Цех 1	Иванова	Ирина	Ивановна	1961	ж		...
...	...	...	...	...	...	...		...

Столбцы таблицы соответствуют тем или иным характеристикам объектов — *полям*. Каждое поле характеризуется именем и типом хранящихся данных.

Имя поля — это идентификатор, который используется в различных программах для манипуляции данными. Он строится по тем же правилам, как любой идентификатор, т.е. пишется латинскими буквами, состоит из одного слова и т.д. Таким образом имя — это не то, что отображается на экране или в отчете в заголовке столбца, а идентификатор, соответствующий этому заголовку. Например, в таблице 2.6.1 введены для последующих ссылок имена нолей Num, Dep, Fam, Nam, Par, Year_b, Sex, Charact, Photo, соответствующие указанным в ней заголовкам полей.

Тип поля характеризует тип хранящихся в поле данных. Это могут быть строки, числа, булевы значения, большие тексты (например, характеристики сотрудников), изображения (фотографии сотрудников) и т.п.

Каждая *строка таблицы* соответствует одному из объектов. Она называется *записью* и содержит значения всех полей, характеризующие данный объект.

При построении таблиц баз данных важно обеспечивать непротиворечивость информации. Обычно это делается введением *ключевых полей*, обеспечивающих уникальность каждой записи. Ключевым может быть одно или несколько полей. В приведенном выше примере можно было бы сделать ключевыми совокупность полей Fam, Nam и Par. Но в этом случае нельзя было бы заносить в таблицу сведения о полных однофамильцах, у которых совпадают фамилия, имя и отчество. Поэтому в таблицу введено первое поле Num — номер, которое можно сделать ключевым, обеспечивающим уникальность каждой записи.

База данных обычно содержит не одну, а множество таблиц. Например, база данных о некоторой организации помимо таблицы данных о сотрудниках Pers может содержать таблицу имеющихся в ней подразделений с характеристикой каждого из них Пример такой таблицы с именем Dep приведен в таблице 2.6.2. Имена полей этой таблицы: Dep и Proisv.

Отдельные таблицы, конечно, полезны, но гораздо больше информации можно извлечь именно из совокупности таблиц. Например, пользователю может требоваться узнать общее количество сотрудников, работающих в производственных цехах. Но ни одна из приведенных выше таблиц не поможет отве-

тить на этот вопрос, поскольку в таблице Pers отсутствуют сведения о типах отделов, а в таблице Dep — о сотрудниках. Для получения ответов на подобные запросы необходимо рассмотрение совокупности *связных таблиц*.

Таблица 2.6.2. Пример таблицы данных о подразделениях Dep

Отдел	Тип
Dep	Proisv
Бухгалтерия	управление
Цех 1	производство
Цех 2	производство

В *связных таблицах* обычно одна выступает как *главная*, а другая или несколько других — как *вспомогательные*, управляемые главной. Главная и вспомогательная таблицы связываются друг с другом *ключом*. В качестве ключа могут выступать какие-то поля, присутствующие в обеих таблицах. Например, в приведенных ранее таблицах головной может быть таблица Dep, вспомогательной Pers, а связываться они могут по полю Dep, присутствующему в обеих таблицах.. Каждой записи в главной таблице ключ ставит в соответствие в общем случае множество записей вспомогательной таблицы. Так в нашем примере каждой записи главной таблицы Dep соответствуют те записи вспомогательной таблицы Pers, в которых ключевое поле Dep с названием отдела совпадает с названием отдела в текущей записи главной таблицы. Иначе говоря, если в текущей записи главной таблицы в поле Dep написано «Бухгалтерия», то во вспомогательной таблице Pers выделяются все записи сотрудников бухгалтерии.

При работе с таблицей пользователь или программа как бы скользит курсором по записям. В каждый момент времени есть некоторая *текущая запись*, с которой и ведется работа. Записи в таблице базы данных физически могут располагаться без какого-либо порядка, просто в последовательности их ввода. Но когда данные таблицы предъявляются пользователю, они должны быть упорядочены. Например, пользователь может хотеть просматривать список сотрудников в алфавитном порядке, или рассортированными по отделам, или по мере нарастания года рождения и т.п. Для упорядочивания данных используется понятие *индекса*.

2.6.1 Процедура индексирования

Индекс показывает, в какой последовательности желательно просматривать таблицу. Он является как бы посредником между пользователем и таблицей Курсор скользит по индексу, а индекс указывает на ту или иную запись таблицы. Для пользователя таблица выглядит упорядоченной, причем он может сменить индекс и последовательность просматриваемых записей изменится. Но

в действительности это не связано с какой-то перестройкой самой таблицы и с физическим перемещением в ней записей. Меняется только индекс, т.е. последовательность ссылок на записи.

Для того чтобы детальнее разобраться с индексами, рассмотрим в качестве примера таблицу с данными об оценках и запрос на поиск студентов, сдававших тот или иной учебный предмет. При таких условиях в БД можно выбрать способ хранения данных, схематически показанный на рис. 2.17.

Он основан на двух хранимых файлах: файле с данными об успеваемости студентов USP и файле с данными об учебных предметах. При этом предполагается, что в файле предметов используется упорядочение по алфавитному перечню их названий, т.е. по ключевому полю PN с указателями на соответствующие записи в файле поставщиков.

Рис. 2.17. Индексирование таблицы оценок по предмету

Возможны следующие стратегии, которые можно применить для поиска всех студентов, сдававших физику:

- найти весь файл успеваемости, найти все записи, для которых названием дисциплины является строка Физика.
- найти файл предметов со строкой Физика, а затем согласно указателям извлечь все соответствующие записи из файла успеваемости.

Если доля всех студентов, сдавших физику, по отношению к общему количеству студентов невелика, то вторая стратегия будет гораздо эффективнее первой. Дело в том, что СУБД известна физическая последовательность записей в файле предметов, а поиск будет прекращен после извлечения следующего за физикой в алфавитном порядке названии предмета. Кроме того, даже если придется просмотреть файл предметов полностью, для такого поиска потребуется гораздо меньше операций ввода-вывода, поскольку физический размер файла предметов меньше, чем размер файла успеваемости из-за меньшего размера записей.

В рассмотренном примере файл предметов называют *индексным файлом* или *индексом* по отношению к файлу успеваемости, или наоборот - файл успеваемости *индексирован* по отношению к файлу предметов.

Индексный файл является хранимым файлом особого типа, в котором каждая запись состоит из двух значений: данных и указателя номера записи. При

этом данные необходимы для индексного поля из индексированного файла, а указатель - для связывания с соответствующей записью индексированного файла.

Если индексирование организовано на основе *ключевого поля*, например, на основе поля SN файла успеваемости, то индекс называется *первичным*. А если индекс организован на основе другого поля, например, поля PN, то он называется *вторичным*. Вторичные индексы могут создаваться как в процессе создания самой базы данных, так и позднее в процессе работы с ней Вторичным индексам присваиваются *имена — идентификаторы*, по которым их можно использовать.

Кроме того, индекс, организованный на основе ключевого поля или другого ключа, называется уникальным.

Основным преимуществом использования индексов является значительное ускорение процесса выборки или извлечения данных, а основным недостатком - замедление процесса обновления данных, т. к. при каждом добавлении новой записи в индексированный файл потребуется также добавить новый индекс в индексный файл.

Индексы можно использовать двумя разными способами:

- для последовательного доступа к индексированному файлу, т. е. в последовательности, заданной значениями индексного поля. Например индекс PN будет определять доступ к записям файла успеваемости согласно алфавитному перечню предметов;
- индексы могут использоваться и для прямого доступа к отдельным записям индексированного файла на основе заданного значения индексного поля, как это было сделано в приведенном примере.

Хранимый файл может иметь несколько индексов: например, хранимый файл успеваемости может иметь индекс PN и индекс OCENKA (см. рис. 2.18).

Рис. 2.18. Индексирование таблицы оценок по двум полям

Индексы могут использоваться как раздельно, так и совместно для более эффективного доступа к данным об успеваемости, например, при запросе на поиск студентов, сдавших физику на 5.

Тогда согласно индексу PN для студентов будут найдены записи с идентификационными указателями 3412 и 3414, а согласно индексу OCENKA - записи с указателями 3412 и 3415. Понятно, что на основе сравнения этих двух наборов записей условиям запроса удовлетворяет только запись с данными о студенте 3412 и только после этого в СУБД будет организован доступ к файлу успеваемости и будет извлечена данная запись.

Часто индекс создают на основе комбинации двух или более полей. Например, на рис. 2.19 показана схема индексирования файла успеваемости на основе комбинации полей PN и OCENKA. При такой организации индексов в СУБД можно выполнить запрос на поиск студентов, сдавших физику на 5 на основе однократного просмотра с помощью одного индекса, в то время как при использовании пары индексов требуется два отдельных просмотра, тем более, что скорость выполнения запроса может сильно зависеть от последовательности выполнения отдельных просмотров по индексам.

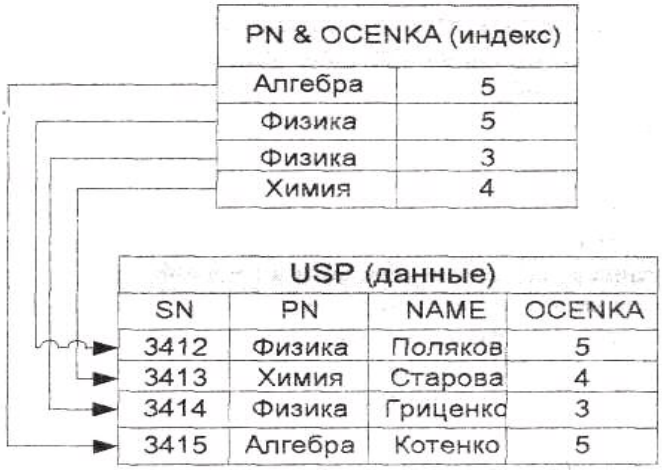

Рис. 2.19. Индексирование таблицы оценок по комбинации двух полей PN и OCENKA

Итак, основной целью использования индекса является ускорение процесса извлечения данных, за счет уменьшения числа дисковых операций ввода-вывода, для чего используются указатели.

2.6.2 Организация связи с базами данных прикладных программ

Создают базы данных и обрабатывают запросы к ним системы управления базами данных — СУБД. Известно множество СУБД, различающихся своими возможностями или обладающих примерно равными возможностями и конкурирующих друг с другом: Paradox, dBase, Microsoft Acess, FoxPro, Oracle, Inter-

Base, Sybase и много других.

Разные СУБД по разному организуют и хранят базы данных. Например, Paradox и dBase используют для каждой таблицы отдельный файл. В этом случае база данных — это каталог, в котором хранятся файлы таблиц. В Microsoft Access и в InterBase несколько таблиц хранится как один файл. В этом случае база данных— это имя файла с путем доступа к нему. Системы типа клиент/сервер, такие, как серверы Sybase или Microsoft SQL, хранят все данные на отдельном компьютере и общаются с клиентом посредством специального языка, называемого SQL.

Поскольку конкретные свойства баз данных очень разнообразны, пользователю было бы весьма затруднительно работать, если бы он должен был указывать в своем приложении все эти каталоги, файлы, серверы и т.п. Да и приложение часто пришлось бы переделывать при смене, например, структуры каталогов и при переходе с одного компьютера на другой, Чтобы решить эту проблему, используют *псевдонимы баз данных*.

Псевдоним (alias) содержит всю информацию, необходимую для обеспечения доступа к базе данных. Эта информация сообщается только один раз при создании псевдонима. А приложение для связи с базой данных использует псевдоним. В этом случае приложению безразлично, где физически расположена та или иная база данных, а часто безразлична и СУБД, создавшая и обслуживающая эту базу данных. При смене системы каталогов, сервера и т.п. ничего в приложении переделывать не надо. Достаточно, чтобы администратор базы данных ввел соответствующую информацию в псевдоним.

При работе с базами данных часто используется кэширование всех изменений. Это означает, что все изменения данных, вставка новых записей, удаление существующих записей, т.е. все манипуляции с данными, проводимые пользователем, сначала делаются не в самой базе данных, а запоминаются в памяти во временной, виртуальной таблице. И только по особой команде после всех проверок правильности вносимых в таблицу данных пользователю предоставляется возможность или зафиксировать все эти изменения в базе данных, или отказаться от этого и вернуться к тому состоянию, которое было до начала редактирования.

Фиксация изменений в базе данных осуществляется с помощью *транзакций*. Это совокупность команд, изменяющих базу данных. Пользователю предоставляется возможность завершить транзакцию или внесением всех изменения в реальную базу данных, или отказом от этого с возвратом к тому состоянию, которое было до начала транзакции.

Основой работы прикладных программ с базами данных является Borland Database Engine (BDE) — процессор баз данных фирмы Borland. BDE служит посредником между приложением и базами данных. Он предоставляет пользователю единый интерфейс для работы, развязывающий пользователя от конкретной реализации базы данных. Благодаря этому не надо менять приложение при смене реализации базы данных. Приложение никогда не обращается непо

средственно к базе данных, а только к BDE. Таким образом, общение с базами данных соответствует схеме, приведенной на рис. 2.20.

Приложение, когда ему нужно связаться с базой данных, обращается к BDE и сообщает обычно псевдоним базы данных и необходимую таблицу в ней. BDE реализован в виде динамически присоединяемых библиотек DLL. Они, как и любые библиотеки, снабжены API (Application Program Interface — интерфейсом прикладных программ), названным IDAPI (Integrated Database Application Program Interface). Это список процедур и функций для работы с базами данных, которым и пользуются приложения.

BDE по псевдониму находит подходящий для указанной базы данных драйвер. Драйвер — это вспомогательная программа, которая понимает, как общаться с базами данных определенного типа. Если в BDE имеется собственный драйвер соответствующей СУБД, то BDE связывается через него с базой данных и с нужной таблицей в ней, обрабатывает запрос пользователя и возвращает в приложение результаты обработки. BDE поддерживает естественный доступ к таким базам данных, как Microsoft Access, FoxPro, Paradox, dBase.

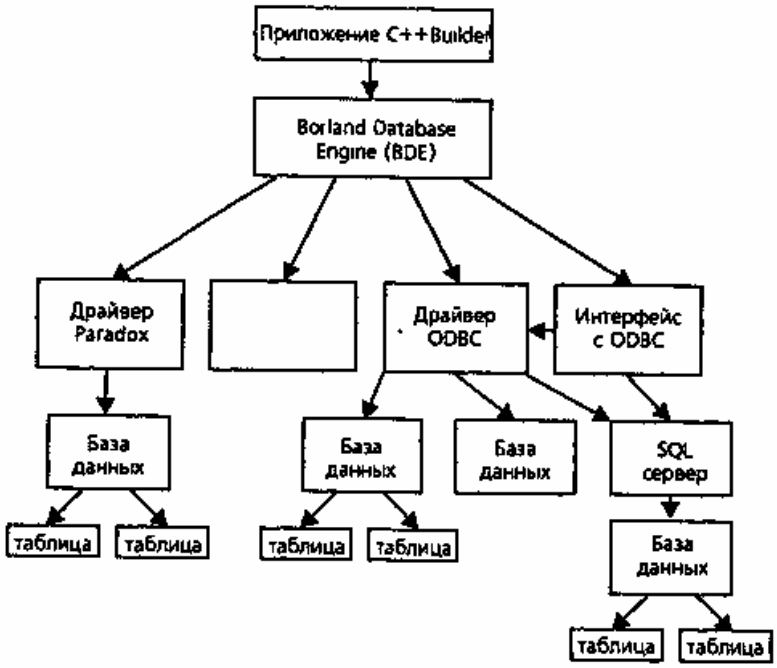

Рис. 2.20. Схема связи приложения с базами данных

Если собственного драйвера нужной СУБД в BDE нет, то используется драйвер ODBC. ODBC (Open Database Connectivity) — это DLL, аналогичная по функциям BDE, но разработанная фирмой Microsoft. Она хранится в файле ODBC.DLL. Поскольку Microsoft включила поддержку ODBC в свои офисные продукты и для ODBC созданы драйверы практически к любым СУБД, фирма Borland включила в BDE драйвер, позволяющий использовать ODBC. Правда, работа через ODBC осуществляется несколько медленнее, чем через собствен-

ные драйверы СУБД, включенные в BDE.

BDE поддерживает SQL — стандартизованный язык запросов, позволяющий обмениваться данными с SQL-серверами, такими, как Sybase, Microsoft SQL, Oracle, Interbase. Эта возможность используется особенно широко при работе на платформе клиент/сервер.

В C++ введена другая альтернативная возможность работы с базами данных, минуя BDE. Это разработанная в Microsoft технология ActiveX Data Objects (ADO). ADO — это пользовательский интерфейс к любым типам данных, включая реляционные и не реляционные базы данных, электронную почту, системные, текстовые и графические файлы. Связь с данными осуществляется посредством так называемой технологии OLE DB. Использование ADO обеспечивает более эффективную работу с данными. Для реализации этой возможности на компьютере пользователя должна быть установлена система ADO 2.1 или более старшая версия. Кроме того должна быть установлена клиентская система доступа к данным, например, Microsoft SQL Server, а в ODBC должен иметься драйвер OLE DB для того типа баз данных, с которым вы работаете.

Надо сказать, что возможности ADO в C++ пока в некоторых отношениях ниже, чем возможности BDE.