В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин

ПРИКЛАДНАЯ МЕДИЦИНСКАЯ СТАТИСТИКА

Учебное пособие

Рекомендовано Учебно-методическим объединением
по медицинскому и фармацевтическому образованию
Министерства образования и Министерства здравоохранения РФ
в качестве учебного пособия для студентов медицинских вузов

Санкт-Петербург
ФОЛИАНТ
2006

УДК 61 311-2
ББК 51 1 (2)

Рецензенты:
Заведующий кафедрой общественного здоровья
и здравоохранения СПбГМУ им И П Павлова
д м н , профессор Н И Вишняков
Заведующий кафедрой общественного здоровья
и здравоохранения СПбПМА, д м н , профессор В К Юрьев

Зайцев В. М., Лифляндский В. Г., Маринкин В. И.
Прикладная медицинская статистика —
СПб ООО «Издательство ФОЛИАНТ», 2003 — 432 с

ISBN 5-93929-056-6

Настоящее издание является пособием по применению статистичаского анализа в медицине В пособии рассматриваются источники получения информации о здоровье населения, некоторые понятия теории вероятности, необходимые для уяснения логики статистического анализа, а также конкретные методики математической статистики в приложении к практике медико-биологических, клинических, гигиенических и социально-гигиенических исследований На наглядных примерах рассмотрены приамы статистичаской обработки данных с помощью популярного программного паката *Microsoft Excel*

Издание предназначено для студентов аузов медико-биологического профиля, научных работников и ерачей всех специальностай При подготоаке издания использованы материалы А А Самуся

ISBN 5-93929-056-6

ОГЛАВЛЕНИЕ

1. Введение в медицинскую статистику

В здравоохранении, как в системе организации медицинской помощи населению, а также в профилактической и клинической медицине повсеместно используются различные численные методы Они применяются в клинической практике, когда врач имеет дело с отдельным больным, в организации медико-социальной помощи населению при прогнозировании, осуществлении и оценке результатов тех или иных медико-социальных программ Знание этих методов необходимо при планировании и проведении научных исследований, для правильного понимания их результатов, критической оценки публикуемых данных. Понимает врач это или нет, но в основе решения любого вопроса о применении способа, тактики лечения или профилактики патологии лежат численные методы Исторически сложилось так, что большой набор численных методов, применяемых в медицине, получил общее название — ***статистика***

Вообще, по своей природе термин ***статистика*** имеет несколько толкований Наиболее примитивное из них подразумевает под статистикой всякий упорядоченный набор числовых характеристик какого-либо явления Если рассматривать статистику как отрасль знаний или отрасль практической деятельности, то все выглядит значительно сложнее Считается, что корни термина ***статистика*** происходят от латинского слова «статус» *(status)* — состояние Несомненна связь и с итальянским *«stato»* — государство Собирание данных о материальном состоянии населения, случаях рождений и смерти, по свидетельству древнегреческого историка Геродота, существовало в Персии уже за 400 лет до Рождества Христова В Ветхом Завете Библии есть це-

лая глава (Книга чисел), посвященная таким статистическим выкладкам В эпоху Возрождения в Италии появились люди, которых называли *«Statisto»* — знаток государства Как синоним терминов *политическая арифметика и государствоведние* термин статистика стал впервые употребляться с середины XVII века

Государственная или политическая статистика остается и поныне большим самостоятельным разделом статистики В настоящее время эта статистика представляет собой сложную, разветвленную систему взаимосвязанных разделов и дисциплин Все они имеют в большей или меньшей степени самостоятельное научное и практическое значение в деле управления государством *Например:* статистика государственного бюджета, статистика сельского хозяйства, статистика промышленности, транспорта и т д В зависимости от отраслевой принадлежности выделяют статистику капитального строительства, материально-технического обеспечения, статистику материальной базы и т п Государственная медицинская статистика, тесно взаимоувязанная с этими разделами статистического учета, также является одной из многих отраслевых статистик, как раньше говорили, «статистикой отрасли народного хозяйства» Но отраслевая медицинская статистика, иногда называемая санитарной статистикой, имеет и свои специфические разделы В первую очередь это — ***статистика системы здравоохранения*** (обеспеченность населения врачами и другим медицинским персоналом, амбулаторно-поликлиническими и стационарными учреждениями, показатели эффективности деятельности этих учреждений и т п), а также ***статистика здоровья населения*** (медико-демографические характеристики населения, показатели его физического развития и т п)

В медицинской статистике, как отрасли знаний, нередко выделяют статистику клиническую, онкологическую, статистику инфекционной заболеваемости, заболеваемости особо опасными инфекциями и т. д Многообразие этих разделов медицинской статистики определяется многообразием разделов медицины как науки и разнообразием видов конкретной практической деятельности медиков Все разделы медицинской статистики тесно между собой взаимосвязаны, имеют единую методическую

основу, и их деление во многих случаях является весьма условным

Математическая статистика, как отрасль знаний, представляет собой специальную научную дисциплину и соответствующую ей учебную дисциплину Предмет этой дисциплины — явления, оценка которых может производиться только в массе наблюдений Эта ключевая особенность обусловлена тем, что изучаемые статистикой явления не имеют постоянных, всегда одних и тех же исходов. Например масса тела, даже у одного и того же человека постоянно меняется, состав клеточных элементов крови при каждом заборе анализа у одного и того же пациента будет несколько различаться, последствия применения одного и того же препарата у разных людей могут иметь свои индивидуальные особенности, и т п Однако многие хаотичные на первый взгляд явления имеют на самом деле вполне упорядоченную структуру и, соответственно, могут иметь вполне конкретные численные оценки Главное условие для этого — статистическая регулярность, статистическая устойчивость этих явлений, то есть существование строго определенных закономерностей, пусть даже скрытых на первый взгляд, которые можно описать математическими методами статистики

Френк Йейтс (1937) писал «Большинству биологических объектов свойственна вариабельность, и прелесть простоты и воспроизводимости физических или химических экспериментов утрачивается А значит, на передний план начинают выдвигаться статистические проблемы»

Фактором, оказавшим значительное влияние на развитие математических методов статистики, стало открытие закона больших чисел Яковом Бернулли (1654–1705) и появление теории вероятности, основы которой разработал французский математик и астроном Пьер Симон Лаплас (1749–1827) Заметным этапом в ряду этих событий для медицинской статистики стала публикация работ бельгийского ученого А Кетле (1796–1874), впервые применившего на практике математико-статистические методы исследования. В своей работе «О человеке и развитии его способностей» А. Кетле вывел тип среднего человека, наделенного, наряду со средними показателями физического развития (рост, вес),

средними умственными способностями и средними моральными качествами В этот же период времени в России выходит работа врача Д Бернулли «О прививках против оспы О смерти и теории вероятности»

На основе теории вероятности, которая позволяет выявлять определенные тенденции в кажущемся на первый взгляд хаосе случайных явлений, в последующие годы появилась *математическая статистика*. Предметом математической статистики стала формально-математическая сторона статистического анализа и количественная оценка на ее основе вероятностей различных явлений В зависимости от точки приложения существуют различные направления использования методов математической статистики.

Медицинская статистика как точка приложения методов математической статистики занимает особое место Это особое место обусловлено большой ролью медицины в возникновении статистики как самостоятельной науки и существенным влиянием научно-исследовательских разработок медико-биологических проблем на появление многих методов статистического анализа. В настоящее время, с целью подчеркнуть особый статус медико-биологической математической статистики, для ее обозначения все чаще используют термин — ***биометрия***

Необходимо помнить, что выбор тех или иных методов статистического исследования не может быть раз и навсегда очерчен рамками какого-либо раздела или отрасли медицины Окончательный выбор конкретных методик зависит от многих обстоятельств, не последним из которых является уровень подготовки специалиста-исследователя в области применения статистических методов

Большинство методов статистического анализа являются универсальными и могут применяться не только в разных отраслях медицинской статистики, но и в самых разнообразных отраслях человеческой деятельности Например, с точки зрения формальной логики статистический прогноз инфекционной заболеваемости и прогноз курса доллара — одна и та же задача По этой же причине большинство компьютерных статистических программ не являются чисто медицинскими прикладными про-

граммами Кроме того, нередко выполнение отдельных статистических функций предусмотрено программами, которые не являются по своей сути статистическими

Возможные области практического применения и доступность методов глубокого статистического анализа особенно резко возросли за последнее десятилетие в связи с появлением и бурным ростом парка персональных компьютеров Все программные средства, которые могут использоваться для статистической обработки данных на персональных компьютерах, можно разделить на

• специализированные пакеты (например, *Мезозавр* — программа анализа временных рядов) Применяются для решения узкого круга задач, с использованием специальных методов статистического анализа. Эксплуатация этих программ подразумевает высокий уровень подготовки пользователя в области определенных разделов статистики,

• статистические пакеты общего назначения (Диастат, STADIA, STATGRAPHICS) Они более доступны для практики и могут использоваться широким кругом специалистов различного профиля Такие пакеты получили достаточно широкое распространение в практике для анализа результатов медико-биологических исследований;

• табличные процессоры и электронные таблицы (QUATTRO PRO, *Excel*) Имеют возможности статистической обработки данных с помощью распространенных методов описательной статистики, сглаживания прогнозирования, регрессионного анализа и т п Имеют повсеместное распространение среди пользователей персональных компьютеров

Особенностью любого пакета статистических программ является выдача большого количества информации, которая описывает результат статанализа В этой ситуации, пользователь, не имеющий достаточной подготовки, зачастую оказывается неспособным правильно все воспринять и осмыслить Попытки употребить различные сопроводительные описания часто оказываются безуспешными, так как они подчас напоминают учебники по статистике, написанные для людей, имеющих специальную подготовку в области математики.

Настоящее пособие дает основы знаний, необходимых для практического использования элементарных методов статистического анализа При этом большое внимание уделяется разъяснению основных терминов и постулатов общей теории статистики Рассматриваются примеры использования для статистического анализа возможностей электронной таблицы *Microsoft Excel*, которая является составной частью одного из самых распространенных пользовательских пакетов *Microsoft Office*

В издании рассмотрены методики организации сбора, сводки и первичной статистической обработки, приемы группировки данных, способы повышения наглядности статистического материала, наиболее широко использующиеся в медицинской науке и практике Разъясняются методы математико-статистической обработки данных, способы практического использования и область их применения

2. Цели, задачи, план и программа статистического исследования

Традиционно считается, что классическое выполнение всех статистических исследований проходит строго по определенным этапам Собственно говоря, выполнение любого серьезного исследования имеет этапный характер Однако во всяком научном исследовании всегда имеется элемент неопределенности, который трудно предугадать и спланировать. Вместе с тем, статистический анализ (статистическое исследование) данных начинается, как правило, тогда, когда уже есть или наметились определенные результаты исследования Поэтому четкое плановое разбиение статистического исследования на определенные этапы вполне возможно и, самое главное, необходимо Необходимость такого разбиения объясняется тем, что стадии статисти-

ческого исследования тесно взаимоувязаны. Ошибки и просчеты на одном этапе неизбежно ведут к возникновению проблем на других этапах (если не провалу) статистического анализа Кроме того, существует много областей научных исследований, которые с самого начала по своей сути носят ярко выраженный статистический характер (медико-демографические исследования, изучение заболеваемости, оценка работы медицинских и социальных учреждений и т п) Начало таких исследований без этапного планирования невозможно уже в принципе

Обычно статистическое исследование включает следующие этапы:

1-й этап — определение целей и задач, составление плана и программы исследования;

2-й этап — наблюдение, сводка и группировка полученных статистических материалов Вычисление первичных итогов,

3-й этап — углубленная математико-статистическая обработка данных,

4-й этап — анализ полученных результатов, выводы

На 1-м этапе статистического исследования определяют

- что делать (цели и задачи исследования) Выбор конкретных методов статистического анализа из множества методов статистики и ход этого анализа напрямую определяются целью и задачами исследования;
- как делать (программа исследования);
- кто, когда и за чей счет будет делать (план исследования)

Цель исследования обусловливается конечным результатом, на достижение которого направлено это исследование, то ради чего оно проводится Цель большинства статистических исследований в медицине — раскрытие взаимосвязи и оценка влияния тех или иных *факторов* на здоровье человека Под *фактором* в этой ситуации могут пониматься фармакологические препараты, средства специфической и неспецифической профилактики, условия труда, быта, образ жизни, социально-демографические особенности и т п. Главная особенность таких статистических исследований — изучение тенденций и закономерностей, которые проявляются в массе наблюдений и не могут быть достоверно проанализированы в отдельно взятых случаях

Задачи исследования отражают частные вопросы, которые необходимо последовательно решить, чтобы достигнуть конечной цели исследования Например, при изучении влияния неблагоприятных производственных факторов на здоровье работников предприятия необходимо решить следующие задачи:

- получить санитарно-гигиенические оценки условий и характера труда на предприятии;
- изучить социально-демографический состав работников;
- изучить заболеваемость с временной утратой трудоспособности;
- по данным целевых медицинских осмотров выявить уровень патологической пораженности работников предприятия;
- выявить и оценить взаимосвязь неблагоприятных факторов производства и патологической пораженности работников основных профессий,
- разработать рекомендации по оздоровлению условий труда на предприятии.

При определении конкретных задач обязательно учитывается главный принцип любого статистического исследования — достоверность исходной информации и объективность отражения изучаемого явления.

План исследования систематизирует решение организационных вопросов В том числе выбор места и сроков наблюдения, источников финансирования, субъекта исследования (организации и лиц, осуществляющих основные работы), подбор и обучение кадров, подготовку необходимых аппаратных и программных средств, регистрационных бланков и т п. Для детальной проработки организационных и методических вопросов, планом предусматривается проведение пробных (пилотажных) исследований

Программа исследования включает программу сбора и программу разработки материалов исследования Обусловливает выбор объекта и единицы наблюдения, а также учетных признаков, подлежащих регистрации в ходе исследования

2.1. Статистическая совокупность, единица наблюдения, учетные признаки

Объект наблюдения — совокупность предметов и явлений, избранная для статистического наблюдения Эту совокупность в статистике принято называть **статистической совокупностью.**

Единица наблюдения — первичный элемент статистической совокупности Иногда единицу наблюдения называют единицей счета Суммарное число отдельных единиц характеризует *объем* статистического наблюдения *Например* при изучении влияния производственных факторов на работников промышленного предприятия объектом статистического исследования является совокупность (группа) работников данного предприятия. Каждый работник данного предприятия, включенный в исследование, является отдельной *единицей наблюдения.* Каждая из этих единиц характеризуется определенными признаками, называемыми *учетными признаками*

Учетные признаки — признаки, подлежащие регистрации в ходе статистического исследования (пол, возраст, профессия работника, стаж работы и т п) Особенностью большинства учетных признаков в медико-биологических исследованиях является их вероятностный, случайный характер Случайный характер учетных признаков объясняется индивидуальными особенностями анатомических, физиологических и других характеристик *Например,* уровень артериального давления даже у одного человека может колебаться в определенных пределах Величина роста человека изменяется в течение суток К вечеру она несколько меньше, чем утром Причем эти изменения у каждого индивидуальны Даже если тот или иной учетный признак является вполне определенным (пол, возраст и т п), эти характеристики в массе единиц наблюдений распределяются, как правило, случайно

По виду учетные признаки могут быть качественными или количесгвенными.

Качественные, описательные или *атрибутивные* — характеризуют качество отдельных единиц совокупности *Например* пол мужской или женский, образование начальное, среднее, высшее; диагноз заболевания и т п.

Количественные — признаки, значения которых имеют числовое выражение. *Например*: рост — см, масса тела — кг, частота пульса — уд/мин и т д.

Некоторые признаки можно рассматривать и как качественные, и как количественные *Например*, масса тела может выражаться количественно в килограммах, а может — качественно (есть избыток массы тела или нет) При выборе формы представления таких данных необходимо учитывать цели конкретного исследования, а также программу дальнейшей статистической обработки, поскольку для каждого вида признаков существуют свои правила статистической обработки. *Например:* даже простейшая арифметическая обработка качественных данных в ряде случаев не допускает расчета средних величин (средний пол) Здесь допустимы только арифметические операции сложения эквивалентных единиц (сложение всех случаев одинаковых заболеваний), либо объединение всех значений конкретного признака в суммарные итоги, либо расчет относительных величин (показатели структуры, частоты, распространенности, соотношения и т д) Аналитическая оценка взаимосвязи качественных и количественных признаков проводится только после разбиения количественных признаков на качественные группы.

По роли в статистической совокупности учетные признаки можно подразделить на факторные (факториальные) и результативные (результирующие) признаки *Результативный признак* — зависимый, изменяющий свое значение под влиянием другого, связанного с ним и действующего на него *факторного признака* Например концентрация сварочного аэрозоля в воздухе рабочей зоны — факторный признак, вероятность возникновения заболевания у сварщика — результативный признак Ролевая значимость этих признаков иногда может меняться *Например* концентрация инсулина в крови и концентрация сахара крови Высокий уровень сахара крови вызывает усиленный выброс ин-

сулина в кровь. В то же время, повышение концентрации инсулина ведет к снижению сахара крови

Все единицы наблюдения, относящиеся к одной статистической совокупности, имеют некоторое число общих учетных признаков, свидетельствующих о принадлежности конкретной единицы наблюдения к этой совокупности Такие признаки называются *признаками сходства* (место работы, время работы на предприятии, место жительства и т. п) Эти признаки описывают обязательное условие статистического наблюдения **единство места и времени исследования.**

Признаки различия представляют индивидуальные особенности (характеристики) каждой единицы наблюдения В медицинских исследованиях это могут быть пол, возраст, производственный или профессиональный стаж, заболеваемость и т п Строго говоря, признаки различия и являются конечным объектом статистического исследования.

Выбор единицы наблюдения и учетных признаков определяет весь ход и результаты статистического исследования *Например*, при изучении заболеваемости за единицу наблюдения может быть принят случай заболевания (случай обращения за медицинской помощью, случай смерти от заболевания, случай потери трудоспособности из-за заболевания и т п), а учетными признаками при этом обычно служат диагноз, пол, возраст заболевшего и т п Другой вариант изучения заболеваемости предусматривает так называемый полицевой учет. Единицей наблюдения в этом варианте является человек, а учетными признаками являются случаи его заболевания, диагноз этих заболеваний, случаи обращения за медицинской помощью, пол, возраст, и т. п. На первый взгляд существенной разницы при первом и втором подходе к сбору данных о заболеваемости нет, поскольку описывается одна и та же информация Однако получаемые в первом и во втором случае результаты, как правило, не имеют не только никакого численного сходства, но и их достоверность и объективность совершенно различны

2.2. Программа сбора. Генеральная и выборочная совокупности. Репрезентативность данных

В обобщенном виде программа сбора — это перечень учетных признаков наблюдения, которые позволяют достаточно полно характеризовать каждую единицу наблюдения и факторы изучаемых явлений. Конкретное воплощение этого перечня — набор вопросов, содержащихся в формуляре статистического наблюдения Таким формуляром может быть анкета опроса, регистрационная карта наблюдения и т п Правильно разработанный формуляр статистического наблюдения является ключом всего исследования. В правильно составленном учетном документе есть все признаки, существенные для данного исследования В нем не должно быть громоздких, малосодержательных, не информативных признаков, затрудняющих регистрацию и обработку материала и мало влияющих на конечные результаты исследования. Учетный документ должен предусматривать удобство работы регистратора, что уменьшает число ошибок регистрации Удобство обработки на ЭВМ гарантирует малую вероятность ошибок ввода данных в компьютер, возможность углубленной и быстрой обработки данных Для снижения числа ошибок и повышения скорости регистрации наблюдений используются специальные аппаратные средства автоматизации (автоматические регистраторы физиологических параметров, сканеры ввода анкетных данных и т. п.)

В целом, по способу наблюдения программой сбора могут предусматриваться следующие варианты получения исходных данных: непосредственное наблюдение (регистрация), выкопировка данных из отчетно-учетных документов, опрос

Непосредственное наблюдение предполагает непосредственную регистрацию единиц наблюдения и их характеристик в натуре, либо измерение параметров с помощью технических средств (измерение жизненной емкости легких, форсированного выдоха, параметров кардиограмм и т п)

Выкопировка данных из отчетно-учетной документации предполагает использование в виде источника информации различных документов (история болезни, история развития ребенка, больничный лист) Этот способ получения информации требует предварительной экспертной оценки наличия документации в полном объеме, правильности заполнения и полноты записей в документах. При этом имеется в виду, что, несмотря на декларируемую полную регистрацию данных, информация из официальных источников (заболеваемость, численность и состав населения) нередко имеет серьезные дефекты, появление которых связано с недоброкачественной работой регистраторов, местными условиями, традициями, экономической ситуацией и т п

Опрос обеспечивает получение информации со слов опрашиваемого (респондента) методом интервью или заочным путем (почтовые, телефонные, прессовые опросы). Регистрация такой информации производится на специальные опросные листы или анкеты. Для качественного проведения опроса рекомендуется привлекать специалистов по разработке опросных листов Выбор метода опроса диктуется организационными и финансовыми возможностями При этом любому из существующих методов присущи определенные недостатки *Например* существенным недостатком телефонного опроса является неразвитость в ряде регионов телефонных сетей, серьезно затрудняющая формирование представительной выборки Опросы первого встречного, интерактивные телевизионные опросы — методы, которые решают в основном журналистские задачи К методам получения достоверной информации такие опросы отнесены быть не могут

Наиболее точный, но весьма дорогостоящий вариант — опрос методом интервью Для такого опроса требуется иметь специально подготовленных людей (интервьюеров), способных устанавливать хороший непосредственный контакт с опрашиваемыми (респондентами) Производительность интервьюера не более 10—20 опросов за день по опросникам, содержащим не более 30—40 вопросов При заочном опросе информация собирается путем саморегистрации. Для этого требуются анкеты, вопросы в которых должны быть, безусловно, понятны всем (почти всем) респондентам

Заочный опрос, как правило, сопровождается большими потерями анкет, поскольку отвечают далеко не все получившие опросные анкеты. Кроме того, ответы, получаемые заочно, подчас отражают мнение не самого респондента, а ближайших членов его семьи (жены, мужа и т. п)

Обязательное условие успешного проведения опроса — организация пробного (пилотажного) опроса В ходе пробного опроса, включающего 10—15 единиц наблюдения, уточняются формулировки вопросов и ответов, проверяется обоснованность набора регистрируемых данных и т п Одним из самых распространенных промахов при составлении программы опроса является непродуманная подстраховка дополнительными вопросами Увеличение по этой причине объема регистрируемой информации, помимо удорожания и увеличения времени обработки, неизбежно ведет к росту числа погрешностей, резко снижающих ценность собранных данных.

Для получения оптимальных результатов посредством любой из существующих методик проведения опросов необходимо придерживаться следующих основных требований к содержанию опросного листа (анкеты)

• *простота и доступность изложения вопросов и вариантов ответов* Учет контингента, среди которого будет проводиться опрос *Например:* опросы сотрудников управленческого аппарата, врачей, инженерно-технических работников и опросы санитарок — не должны быть одинаковыми (Еще Иисус Христос, наставляя своих апостолов, обращал внимание на то, что они должны всегда помнить о том, кто их слушает);

• *удобство для последующего ввода и обработки на компьютере* Необходимо помнить, что формализованные, краткие ответы, используемые с этой целью, могут не всегда соответствовать представлениям респондентов Поэтому рекомендуется предусматривать варианты уклончивых или неопределенных ответов;

• *по возможности содержать контрольные вопросы*,

• *идентификационные данные* (возраст, место жительство, телефон и т п) желательно относить в конец опроса,

• *оптимальность объема опроса* С одной стороны, размер анкеты должен обеспечивать получение максимума информации

С другой — анкета не должна быть утомительной для заполнения (усталость респондента и опрашивающего резко повышает число неустранимых ошибок) и затруднять своей обширностью последующую обработку

Не в меру завышенный объем регистрируемых данных относится к числу серьезных ошибок, способных вызвать срыв всего исследования, несмотря на применение самых изощренных методов статистической обработки данных Авторы громоздких опросов обычно считают большое количество регистрируемой информации существенным достоинством исследования Но на практике все выглядит наоборот В частности, в большом опроснике все ответы, записанные в большом количестве пунктов, в действительности очень часто могут быть заменены крайне ограниченным набором «тривиальных» ответов

В то же время, как показывает накопленный нами опыт, получить неискаженную информацию по анкете, содержащей более 4—5 десятков вопросов, на практике невозможно. Респонденты и работники, проводящие длительный опрос, неизбежно утомляются Из-за этого лавинообразно нарастает число систематических ошибок в ответах респондентов и систематические ошибки регистраторов В итоге надежность такой программы сбора данных, как правило, оказывается за пределами допустимого уровня точности

Частично уменьшить число систематических ошибок, которые в отличие от случайных ошибок нельзя рассчитать и оценить, можно прибегнув к помощи специально подготовленных и поэтому дорогостоящих работников, проводящих опрос При этом большой объем наблюдений требует большого числа таких работников Даже несмотря на широкое использование высококвалифицированного персонала (что на практике выглядит весьма фантастично), при регистрации больших объемов информации все равно неизбежно накапливается масса систематических ошибок

Вообще чрезмерное увеличение объема любой исходной информации, в том числе получаемой и не путем опроса, неизбежно ведет к увеличению так называемого «информационного шума», т е к росту числа помех Этот «шум» складывается из

механических систематических ошибок регистрации и ошибок, связанных с вариабельностью исходной информации. Достигая известного предела, такой «шум» просто подавляет искомую исследователем информацию

Не является панацеей от упомянутого «шума» и использование самых точных и совершенных приборов и аппаратуры Это связано, прежде всего, с большой ***вариабельностью*** (изменчивостью, случайностью) медико-биологических процессов, которая представляет собой сложное явление, складывающееся из нескольких составляющих Одной из самых существенных среди них по негативным последствиям, с точки зрения результативности статистической обработки, является *аналитическая вариабельность* Она возникает из-за расхождения между результатами измерений в одной пробе Например, результаты подсчета клеточных элементов крови одного и того же зафиксированного мазка, проведенного несколько раз, даже одним и тем же специалистом, будут, хоть и не намного, отличаться

Внутрииндивидуальная вариабельность характеризует расхождение значений нескольких замеров какого-либо параметра у одного и того же человека Например масса тела у одного и того же здорового человека в течение суток изменяется

Межиндивидуальная, или внутригрупповая вариабельность характеризует расхождение значений замеров какого-либо параметра у разных людей, даже если они здоровы

К этим видам вариаций добавляется еще одна — *вариабельность межгрупповая*, изучение которой, собственно говоря, и является целью любого исследования

Высокую вариабельность медико-биологических данных можно снижать, используя специальные, порой весьма громоздкие, методики регистрации измеряемых параметров (стандартизация условий регистрации, многократная регистрация замеров и т д), а также употребляя специальные приемы статистической обработки Однако результативность таких операций остается во многих случаях невысокой. В лучшем случае удается уменьшить один из компонентов общей вариабельности Например, в эксперименте, в котором сравниваются два наблюдения одной и той же группы до и после воздействия изучаемого фактора, исключается

часть внутригрупповой вариации, связанной с индивидуальными различиями субъектов, поскольку рассматриваются просто разности между двумя измерениями («до» и «после») для каждой единицы наблюдения Именно так, в частности, и проводится вычисление *t*-критерия для зависимых выборок.

Считается, что аналитическая вариабельность не должна превышать 10% от среднего значения регистрируемой величины или 25% от ширины интервала колебаний нормы А величина индивидуальной вариабельности не должна составлять более 10% от межиндивидуальной [Власов В В , 1978] Однако значительное число клинико-физиологических параметров не укладываются в этот «коридор», несмотря на все методические ухищрения. Это относится, например, к данным о содержании натрия, калия, кальция, хлоридов, белка, альбумина, диоксида углерода, фосфора, магния, глюкозы, лактатгидрогеназы в плазме крови [Власов В. В , 1988; Ван дер Ваден Б Л , 1960] Другой пример вариации параметров форсированного выдоха в течение дня у здоровых лиц составляют около 25% от межиндивидуалыых У лиц с хроническими заболеваниями легких это соотношение достигает 50% и более [Власов В В , 1988 и др]

По времени наблюдение может быть текущим или единовременным

Текущее (непрерывное) наблюдение предусматривает регистрацию данных по мере их возникновения за какой-либо промежуток времени Данные при этом виде наблюдения накапливаются во времени *Например* данные о заболеваемости по обращаемости, инфекционной заболеваемости, смертности, регистрация актов гражданского состояния ЗАГСами за год и т д Отдельные результаты, полученные таким путем, можно суммировать (помесячную заболеваемость можно суммировать по кварталам, за год)

Единовременное (прерывное) наблюдение предусматривает регистрацию данных в один момент времени, или по состоянию на один момент времени, так называемый *критический момент наблюдения* Таким образом проводится сбор данных при переписях населения Данные Всесоюзной переписи населения 1989 года, *например*, регистрировались по состоянию на 12 ч ночи с 11 на

12 января Несмотря на то что переписчик, проводивший регистрацию, мог получить данные в любой из 8 последующих дней, данные регистрировались строго на критический момент времени и изменения, произошедшие после этого момента, во внимание не принимались Если за эти 8 дней кто-либо умирал, то он регистрировался, как живой И наоборот, если за это время рождался ребенок, то он не учитывался

На определенный момент времени (на конец года) учитывается численность населения, численность медицинского персонала, количество учреждений медицинской помощи Заболеваемость по данным профилактических осмотров (патологическая пораженность) так же регистрируется на определенный момент времени Особенность данных полученных таким путем — их нельзя просто суммировать *Например* численность населения района в 1998 году составила 2,5 млн человек, в 1997 году — 2,6 млн человек. За суммарный период 1997–1998 годов можно рассчитать только среднегодовую численность населения Если единовременное наблюдение регулярно повторяется (ежегодный учет численности населения, предприятий и т п), то такое наблюдение называется периодическим

По охвату статистической совокупности исследование может быть сплошное или не сплошное. Эта методическая особенность сбора данных определяет весь дальнейший ход и методику статистического анализа

При ***сплошном*** статистическом исследовании группа наблюдения формируется путем полного охвата всех единиц изучаемого явления Множество всех единиц наблюдения, охватываемых таким сплошным наблюлением, называется **генеральной совокупностью** На практике сплошное исследование проводится крайне редко, поскольку осуществить такое наблюдение организационно очень трудно или физически невозможно из-за больших размеров генеральной совокупности или из-за отсутствия определенных границ этой совокупности В ряде ситуаций, даже если генеральная совокупность ограничена в своих размерах, исследование объекта приводит к его уничтожению (анализ качества промышленных партий вакцин, сывороток, медикаментов) В этой ситуации проведение сплошных исследований также не-

возможно К тому же, сплошные исследования во много раз дороже несплошных Существенный недостаток сплошных исследований — большие затраты времени на сбор данных, что приводит к затягиванию исследований на продолжительное время Кроме того, формирование в ходе таких наблюдений больших массивов данных вызывает затруднения при обработке собранных материалов Такие наблюдения используют, как правило, только для решения общегосударственных задач переписи населения, сбора информации об инфекционной заболеваемости и некоторых других задач

К методам **не сплошного** наблюдения относятся монографический метод, метод основного массива и собственно выборочный метод

Монографический метод применяется для подробного описания объекта, имеющего какие-либо яркие особенности *Например.* медико-социальное обследование национальностей Крайнего Севера или социально-гигиеническое описание промышленного центра Выводы, которые получаются путем таких обследований, относятся либо только к конкретному объекту исследования, либо могут быть распространены на весьма ограниченную группу аналогичных объектов

Метод основного массива предусматривает обследование контингентов, которые могут быть сосредоточены на конкретном объекте *Например.* изучение госпитализированной заболеваемости в стационаре. Данные о структуре заболеваний, тяжести течения и их прогнозе, полученные в этом исследовании, могут иметь значение только для решения частных вопросов Судить о распространенности патологии за пределами такого стационара по этим данным нельзя.

Собственно **выборочное исследование** охватывает **выборочную совокупность** или просто **выборку** из генеральной совокупности. Такое исследование имеет ряд весьма существенных преимуществ перед сплошным наблюдением Во-первых, оно дает значительную экономию средств и требует существенно меньше времени, чем сплошное Во-вторых, при выборочном исследовании может быть достигнута большая глубина и детальность изучения вопроса В-третьих, при меньшем числе наблюдений

уменьшаются вероятности *систематических ошибок наблюдения*, возникающих, когда объектам приписывают искаженные данные

Конечной целью изучения выборочной совокупности всегда является получение информации о генеральной совокупности Для этого выборочное исследование должно удовлетворять определенным условиям Одно из главных условий — представительность выборки

Каждое выборочное исследование несет с собой некоторую погрешность, вытекающую из самого факта выборочности, когда результаты, полученные на выборке, переносят на всю генеральную совокупность, т е ***по части судят о целом***

В этой связи следует отметить, что в настоящее время действует Федеральный закон, касающийся в связи с проведением опросов общественного мнения референдумов и т. п ответственности за информацию, получаемую путем опросов Этот закон (Ст 37) гласит «При публикации результатов опросов средства массовой информации обязаны указывать организацию, проводившую опрос, время его проведения, число опрошенных (выборку), метод сбора информации, точную формулировку вопроса, статистическую оценку возможной погрешности» В статистике подобного рода погрешность называется **ошибкой репрезентативности** (представительности)

Выделяют репрезентативность количественную и качественную (структурную) *Количественная репрезентативность* определяется числом наблюдений, гарантирующим получение статистически достоверных данных В общем, здесь действует основной постулат **закона больших чисел** — *«чем больше наблюдений — тем результаты достоверней»* или *«чем больше число наблюдений, тем больше значения характеристик выборки приближаются к соответствующим характеристикам генеральной совокупности»*

Это обстоятельство существенно для *случайных ошибок*, которые имеют некоторые общие для данной совокупности свойства Число положительных случайных ошибок почти всегда равно числу отрицательных Чем больше проведено наблюдений, тем ближе к нулю разность между теми и другими случайными ошибками *Систематические ошибки* наблюдения, которые мо-

гут возникать как при сборе, так и при сводке информации, искажают результат наблюдения в одном направлении Они могут возникать, например, в результате желания опрашиваемых или исследователей представить все в лучшем свете, чем есть на самом деле Ошибки такого рода не могут быть устранены увеличением объема выборки

Величина ошибки репрезентативности зависит также и от изменчивости изучаемого признака Если бы все единицы совокупности были одинаковы, то результаты, полученные на одной единице наблюдения, можно было бы распространить на все остальные Однако реально всегда имеется какой-то разброс значений изучаемых признаков. И чем он больше, тем больше статистическая ошибка Именно поэтому при анализе статистических данных необходимы характеристики изменчивости (разброса) значений, составляющих ряды распределений Подробно методика вычисления ошибок репрезентативности рассматривается в специальном разделе («Оценка статистических параметров по выборочным данным»)

Качественная репрезентативность — обозначает структурное соответствие выборочной и генеральной совокупностей *Например*· если в составе генеральной совокупности 50% — лица мужского пола, то и в выборочной группе их должно быть 50%

В силу закона больших чисел выборка будет качественно репрезентативной только в том случае, если ее осуществить случайно Проводить отбор случайно, значит обеспечить выполнение условия, что каждый объект выборки отбирается случайно из генеральной совокупности. При соблюдении этого условия можно определенно утверждать, что объекты выборки правильно представляют генеральную совокупность

Случайность, гарантирующая качественную (структурную) репрезентативность статистических исследований, достигается выполнением ряда условий формирования выборочных групп (совокупностей):

1. *Каждый член генеральной совокупности должен иметь равную вероятность попасть в выборку Например* если отбор историй болезней проводить по заглавным буквам фамилий больных, то вероятность попасть в выборку для разных фамилий

будет разная, т к частота встречаемости различных букв алфавита в началах фамилий разная

2 *Отбор единиц наблюдения из генеральной совокупности необходимо проводить независимо от изучаемого признака* Если отбор проводится целенаправленно, то и при этом необходимо соблюдать условия независимости распределения изучаемого признака *Например*, при изучении взаимосвязи курения и здоровья человека можно поступить двояко

- целенаправленно сформировать группы обследуемых в зависимости от их отношения к курению (не курят, курят мало, курят много и т п) В этом случае независимо должны формироваться показатели здоровья в этих группах,
- целенаправленно сформировать группы здоровья *Например* здоровые, редко болеющие острыми заболеваниями, хронические больные и т п В этом случае независимо должны формироваться показатели отношения к курению в этих группах *Например* не курят, курят мало, много и т п

В первом случае доказательством связи курения и здоровья будут разные уровни показателей здоровья в различных по отношению к курению группах Во втором — различная распространенность курения в различных группах здоровья

3 *Отбор должен проводиться из однородных групп. Например* показатели физического развития мужчин и женщин существенно отличаются друг от друга, поэтому для оценки физического развития необходимо брать либо однополые группы, либо группы с одинаковым соотношением полов

Соблюдение условий, гарантирующих максимальную близость выборочной и генеральной совокупностей, обеспечивается специальными способами отбора В зависимости от способа формирования различают следующие выборки

1 Выборки, не требующие разделения генеральной совокупности на части (собственно случайная повторная или бесповторная выборка)

2 Выборки, требующие разбиения генеральной совокупности на части (механическая, типическая или типологическая выборки, когортная, парно-сопряженная выборки)

Собственно **случайная выборка** формируется случайным отбором — наудачу В основе случайного отбора — перемешивание. *Например.* выбор шара в спортлото после перемешивания всех шаров, выбор выигрышных номеров лотереи, случайный выбор карточек больных для исследования и т п Иногда используют случайные числа, получаемые из таблиц случайных чисел или с помощью генераторов случайных чисел. Согласно этим числам из заранее пронумерованного массива генеральной совокупности выбираются единицы наблюдения с номерами, соответствующими выпавшим случайным числам

При составлении случайной выборки после того, как объект отобран и все необходимые данные о нем зарегистрированы, можно поступать двояко объект можно вернуть или не вернуть в генеральную совокупность В соответствии с этим выборку называют **повторной** (объект возвращается в генеральную совокупность) или **бесповторной** (объект не возвращается в генеральную совокупность) Отбор с возвращением (повторный отбор) гарантирует большую независимость выборки, однако этот вид отбора труден в организационном плане Вместе с тем, разность погрешностей бесповторного и повторного отбора тем меньше, чем больше объем генеральной совокупности В практике медико-биологических и гигиенических исследований, как правило, объем генеральной совокупности неизвестен или гипотетически достаточно велик В предельном случае генеральная совокупность бесконечно велика. *Например*: при оценке профилактических свойств фармацевтического препарата исследователь, как правило, не знает точно числа людей, которые в объективной реальности могут его принимать Поэтому в большинстве статистических исследований разница между повторной и бесповторной выборками практически отсутствует и априорно принимается условие, что выборка повторная

При неизвестной величине генеральной совокупности величину повторной выборки, гарантирующую репрезентативные результаты, если результат отражается показателем в виде относительной величины, определяют по формуле:

$$n=\frac{t^2pq}{\Delta^2},$$

где p — величина показателя изучаемого признака; $q=(100-p)$; t — доверительный коэффициент, показывающий какова вероятность того, что размеры показателя не будут выходить за границы предельной ошибки (обычно берется $t=2$, что обеспечивает 95% вероятность безошибочного прогноза); Δ — предельная ошибка показателя

Например: одним из показателей, характеризующих здоровье рабочих промышленных предприятий, является процент не болевших в течение года работников Предположим, что для промышленной отрасли, к которой относится обследуемое предприятие, этот показатель равен 25% Предельная ошибка, которую можно допустить, чтобы разброс значений показателя не превышал разумные границы, 5% При этом показатель может принимать значения 25% ±5%, или от 20% до 30%. Допуская $t=2$, получаем

$$n=\frac{2^2\times25\times75}{5^2}=300 \text{ рабочих}$$

В том случае, если показатель — средняя величина, то число наблюдений можно установить по формуле.

$$n=\frac{t^2\sigma^2}{\Delta^2},$$

где σ — показатель вариабельности признака (среднеквадратическое отклонение), который можно получить из предыдущих исследований либо на основании пробных (пилотажных) исследований.

При бесповторном отборе и при условии известной генеральной совокупности для определения необходимого размера случайной выборки в случае использования относительных величин применяется формула

$$n=\frac{t^2pqN}{\Delta^2N+t^2pq}$$

Для средних величин используется формула

$$n=\frac{t^2\sigma^2 N}{\Delta^2 N+t^2\sigma^2},$$

где N — численность генеральной совокупности Исходя из условий приведенного выше примера и принимая численность генеральной совокупности N=500 рабочих, получаем

$$n=\frac{2^2\times 25\times 75\times 500}{5^2\times 500+2^2\times 25\times 75}=187{,}5\approx 188 \text{ рабочих}$$

Нетрудно заметить — необходимая численность выборки при бесповторном отборе меньше, чем при повторном (соответственно 188 и 300 рабочих)

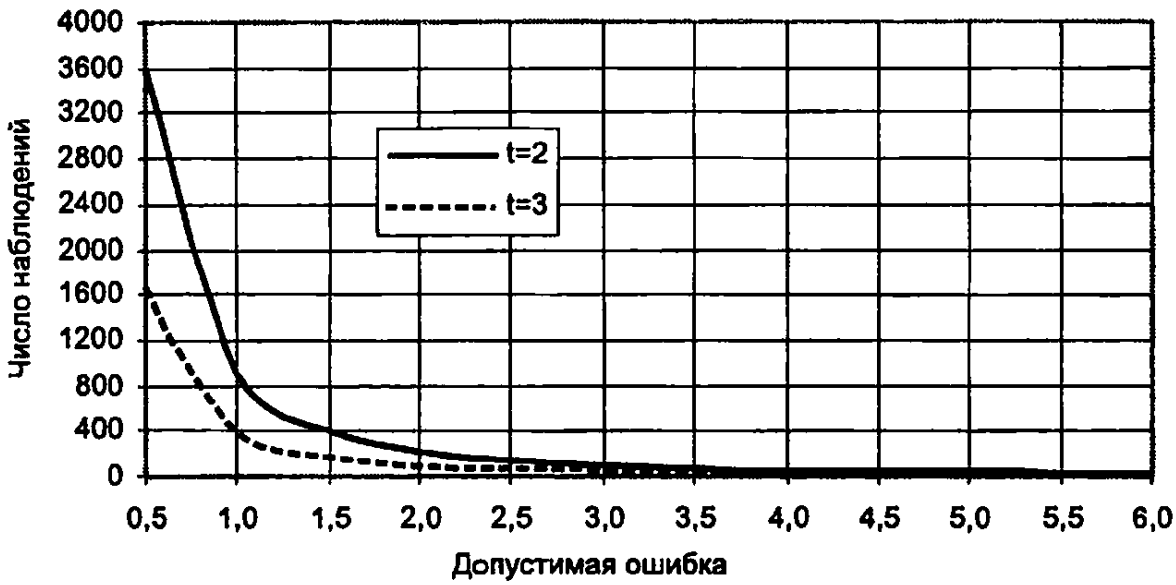


Рис. 1. **Зависимость числа наблюдений от величины допустимой ошибки выборочного исследования (при $\sigma=10$)**

В целом, число наблюдений, необходимое для получения репрезентативных данных, изменяется обратно пропорционально квадрату допустимой ошибки

Механическая выборка — выборка, когда из обследуемой совокупности единицы наблюдения отбираются механически *Например* отбор каждого пятого или каждого десятого рабочего по карточкам отдела кадров предприятия или по амбулаторным картам поликлиники МСЧ При этом надо помнить, что обра-

щаемость в поликлинику взрослого населения может зависеть от состояния здоровья (здоровые почти не обращаются) В результате чего, в поликлинике амбулаторные карты имеются, как правило, только на больных Кроме того, большая часть амбулаторных карт может находиться на руках у пациентов

Типическая, типологическая или ***районированная выборка*** предполагает разбивку генеральной совокупности на ряд качественно однородных групп *Например* при изучении заболеваемости студентов вуза для углубленного обследования на каждом курсе выбираются типичные по своему составу студенческие группы Часто этот способ отбора комбинируется с другими способами. *Например* территория города делится в зависимости от степени загрязнения на типичные районы, в этих районах путем случайного отбора формируются группы наблюдения

Когортный отбор относится к целенаправленным отборам, при этом способе из генеральной совокупности отбираются лица, объединенные моментом появления какого-либо признака, играющего существенную роль в исследовании (год рождения, начало болезни и т п)

Следует отдельно остановиться на использовании выборочного метода в санитарной статистике при изучении общей заболеваемости населения В отличие от других областей медицины выборочные исследования здесь не получили широкого распространения, хотя о необходимости таких исследований отечественные статистики говорили еще в 20-е годы прошлого XX столетия [Паевский В В , Сичинский М В , Смулевич Я М , Богословский С М , Мерков А М и др] Более того, теоретические предпосылки выборочного метода были проверены в ходе специальных исследований Так, В С Быховский и соавт в 1928 году сделали параллельную обработку 132,8 тыс карт с данными о заболеваниях сплошным методом и методом механического отбора каждой пятой карты Анализ результатов этой обработки показал высокую репрезентативность данных выборочного исследования заболеваемости Однако вплоть до сегодняшнего дня отсутствуют единые методические подходы проведения в широкой практике выборочных санитарно-статистических исследований В частности, нет четких критериев для

определения объема выборки Например, В В Паевский в опубликованной в 1928 году работе указывал, что при размере генеральной совокупности 50 тыс человек и выше для выборки необходимо брать не менее 25 тыс единиц наблюдения! А при больших объемах, по мнению автора, можно ограничиваться 10% выборкой В настоящее время при социологических опросах населения России численностью более 147 млн человек объем выборки обычно составляет около 2 гыс человек Если опираться на рекомендации В В Паевского, эта выборка должна бы составлять 14,7 млн человек Справедливости ради следует отметить, что в настоящее время рекомендуемые объемы выборки для изучения общей заболеваемости значительно ниже, однако они представляются во многом спорными и остаются значительно выше принятых при исследованиях в других областях знаний

2.3. Программа статистической разработки. Основы группировки данных

Программа разработки предусматривает реализуемые на втором этапе статистического исследования *сводку и группировку* статистических данных Эти операции, осуществляемые на основе статистических таблиц, позволяют систематизировать полученные в ходе наблюдения данные, провести обработку и подсчет групповых итогов, расчеты простейших производных величин (статистических коэффициентов, средних величин) На этом же этапе, для повышения наглядности данных, предусматривается использование графических изображений Иногда (в официальной статистике — почти всегда), на этом заканчивается весь процесс обработки собранных данных

Группировка — основа статистической разработки (систематизации) первичного материала Играет исключительную роль в статистике Группировка, правильно спланированная на этапе подготовки исследования, позволяет облегчить регистрацию или понизить точность измерений (*accuracy* — англ) на этапе сбора

исходных данных без снижения результативности исследования в целом Группировка собранной исходной информации определяет весь ход статистического анализа

В ходе статистической разработки исследователю приходится сталкиваться со следующими вариантами группировок, каждый из которых имеет свои методические особенности

• *разделение анализируемой статистической совокупности на группы по тем или иным признакам* С такого рода группировкой приходится сталкиваться уже при подготовке программы сбора и в ходе реализации программы разработки любого исследования;

• *объединение мелких однородных групп в более крупные* Этот вариант группировки применяется, как правило, уже в процессе статистической обработки данных, если выясняется несостоятельность мелких групп (малое число наблюдений, не четко выраженный характер распределений и т п.) Возможность такой группировки целесообразно предусмотреть уже на этапе подготовки программы сбора данных, т е. обеспечить возможность укрупнения групп в соответствии с общепринятыми границами групп,

• *комплексная группировка обеспечивает формирование комплексных оценок на основе многих учетных признаков, даже если они разнородны.* Такая группировка часто делается на основе специально разрабатываемых алгоритмов или экспертных оценок (по аналогии — постановка диагноза на основе многих симптомов и результатов лабораторных обследований) Такая группировка представляет весьма сложную задачу и нередко сама по себе является самостоятельной целью исследования

Выбор метода или способа группировки во многом определяется видом учетных признаков. Для группировки качественных признаков используются альтернативная шкала и шкала рангов.

Альтернативная шкала, шкала рангов, шкала номиналов, шкала категорий характеризуются тем, что для отличия одного «измерения» признака от другого «измерения» используются имена, метки, ярлыки (номер телефона, почтовый индекс и т п) Эти метки могут быть дихотомическими, т е допускать разбиение на

два, либо на несколько вариантов *Например* пол, диагноз заболевания, место рождения Измерение в такой шкале не содержит никаких указаний на величину признака или его качественную характеристику относительного другого признака (мужской пол лучше, женский — хуже). Вместе с тем, при статистической сводке для удобства работы все значения качественных учетных признаков часто кодируются (мужской пол — 1, женский — 2 и т п). Таким образом, эти характеристики приобретают некое количественное выражение, оставаясь, в принципе, качественными Это количественное выражение нередко побуждает исследователя механически включать такие данные в статистическую обработку вместе с другими количественными характеристиками изучаемого явления, что является грубой ошибкой

Шкала рангов (баллов) или порядковая шкала (шкала ординаров) представляет собой альтернативную шкалу, но дополнительно вводит ранговый порядок взаиморасположения Это упорядочение производится по определенному правилу *Например* от большего к меньшему или наоборот В этой шкале каждое конкретное значение признака может быть выше, ниже или равно другому значению *Например.* стадия онкологического заболевания, разделение территорий по уровню загрязненности (высокий уровень, средний, низкий). Значения рангов (баллы, классы) при этом остаются качественными. Разность между ними не соответствует действительному различию признаков Основой для выводов здесь является только соотношение «больше» — «меньше», т. е. анализируется только информация о взаимной упорядоченности признаков С этими числами, как и в случае шкалы номиналов, нельзя делать привычные арифметические операции *Например*, вряд ли справедливы утверждения, что знания отличника равны сумме знаний двоечника и троечника (хотя 5=2+3), или что сумма знаний двух двоечников равна знаниям одного хорошиста (4=2+2) Несмотря на это «средний балл» широко используется в отечественной педагогике

Достоинством балльных шкал является возможность получения интегрированных оценок (табл 1) *Например*

Таблица 1

Сравнительная оценка отношения к курению мужчин и женщин

Отношение к курению	Баллы	Мужчины		Женщины	
		%	баллов	%	баллов
Не курят	1	30	30×1=30	55	55×1=55
Выкуривают менее пачки сигарет в день	2	39	39×2=78	40	40×2=80
Выкуривают более пачки сигарет в день	3	31	31×3=93	5	5×3=15
Итого	—	100	201	100	150

Используются балльные шкалы и при получении комплексных оценок, когда составляющие этих оценок выражены разнородными величинами *Например* оценка физического развития может включать рост в сантиметрах, массу тела в килограммах, и т п Для получения комплексной оценки производится суммирование баллов Если рост ребенка оценен 2 баллами, вес 3 баллами, то его комплексная оценка по росту и весу будет равна (2+3) 5 баллам. (Более подробно см. в разделе «Квантили».)

Ключевым звеном успешного статистического анализа качественных ранжируемых признаков является выбор границ рангов Неправильно выбранные границы не позволят вскрыть истинную картину распределения изучаемых явлений (См также раздел «Ряды распределений»)

Оптимальное число рангов при распределении медико-биологической информации лежит в пределах 6—14 Чаще всего границы балльных оценок устанавливаются эмпирически, при этом желательно сохранить достаточную наглядность распределения изучаемой статистической совокупности Как показывает практика, формирование балльных оценок исследователями подсознательно происходит в соответствии с психофизическим законом Фехнера, согласно которому сила ощущения каких-либо раздражителей изменяется в арифметической прогрессии, в то время как сила явления, вызывающая раздражение, изменяется в геометрической прогрессии Учет этого обстоятельства помогает лучшему восприятию и осмыслению статистических данных

Интервальные шкалы позволяют получать количественные оценки объектов исследования (учетных признаков), упорядочивать объекты исследования, численно выражать их характеристики и проводить сравнение *Например*, увеличение при некоторой дозированной нагрузке диастолического давления в легочной артерии на 4,7 мм рт ст (с 10,5 до 15,2 мм рт ст)

Относительные шкалы очень похожи на интервальные шкалы. В дополнение ко всем свойствам переменных, измеренных в интервальной шкале, их характерной чертой является наличие нулевой точки Типичный пример шкалы отношений — температура по Кельвину Можно вполне определенно утверждать, что температура 200° вдвое выше, чем 100° Интервальные шкалы (шкала Цельсия) не обладают данным свойством шкалы отношения

Выделяют следующие основные виды группировки в зависимости от конкретной цели статистического исследования (табл 2).

Таблица 2

Классификация статистических группировок

Аналитические	Структурные	Типологические	Специальные (балансовые, матричные и т д)
Характеризуют взаимосвязи между признаками, проявляют основные тенденции	Выявляют состав, структуру обследованных групп	Характеризуют основные группы (типы групп признаков)	Используются при составлении балансов предприятий и учреждений, отраслей промышленности и т п

Аналитическая группировка выявляет взаимосвязи между явлениями (признаками их характеризующими) При этом они подразделяются на **факторные и результативные.** Взаимосвязь проявляется в систематическом изменении результативного признака в связи с изменением факторного *Например* температура тела влияет на частоту пульса, в зависимости от величины роста изменяется вес и т п

Структурная группировка выявляет состав, строение однородной в качественном отношении статистической совокупности *Например* состав больных по полу, возрасту, диагнозу и т п Сопоставление данных структурной группировки во времени дает представление о структурных сдвигах

Группировка типологическая — с ее помощью в статистической совокупности выделяются качественно однородные в существенном отношении группы *Например* группы больных с одинаковым диагнозом, с одинаковым исходом заболеваний и т п

При статистической разработке материала любого исследования необходимо учитывать существующие правила и стандарты определения группировочных признаков и границ групп (возрастно-половые группировки, группировки по категориям тяжести труда и т п) Не соблюдение правил формирования этих группировок ведет к потере ценности данных. Это обусловлено следующими причинами:

Во-первых, невозможностью сравнения полученных данных с данными других исследований *Например* для общей оценки возрастного состава применяется следующее укрупненное распределение на три группы 0—14 лет, 15—59 лет, 60 лет и старше Если использовать группы 0—16 лет или 15—50 лет, то полученные данные в этом случае будут просто несопоставимыми

Во-вторых, объединение в стандартные группы обусловлено определенными мотивами, игнорирование которых может лишить научной содержательности все исследование *Например*, группировка людей по возрастным группам производится с учетом физиологических особенностей развития организма человека, действующего законодательства (трудового, пенсионного и т п), практикой экономического анализа и демографических исследований В качестве примера рассмотрим границы и мотивы образования возрастно-половых групп, при формировании которых исследователями допускается наибольшее число ошибок В национальной статистике выделяют следующие возрастно-половые группы (в границах точного возраста):

1) дети до 3 лет Эта группа находится под наблюдением детских консультаций и обслуживается детскими яслями Из них часто выделяются дети в возрасте 0 лет и 1 год,

2) дошкольники — дети от 3 до 7 лет Обслуживаются детскими садами,

3) дети и подростки школьного возраста — 7—13 лет и 13—16 лет,

4) подростки — 16—18 лет,

5) трудоспособный контингент — мужчины 16—60 лет, женщины 16—55 лет;

6) лица пенсионного возраста — мужчины 60 лет и старше, женщины 55 лет и старше;

7) женщины репродуктивного возраста — обычно от 15 до 45 лет

Иногда для группировки по возрасту используют одногодичные или (для взрослых) пятилетние, реже десятилетние интервалы Группировка с пятилетним интервалом выглядит следующим образом до 20 лет, 20—24, 25—29, 30—35 и т д С десятилетним интервалом до 20 лет, 20—29, 30—39 и т д При изучении заболеваемости в связи с производственными факторами обычно используются аналогичные 5- и 10-летние интервалы группировок по стажу работы 1—4 года, 5—9 лет, 10—14 лет и т д При этом, в группу 1—4 года относятся лица со стажем от 1 года до 4 лет 11 мес 29 дней Аналогично в группу со стажем от 5 до 10 лет относятся лица, отработавшие от 5 лет до 9 лет 11 мес и 29 дней

В ряде случаев целесообразно расчленять отдельные крупные группы на более мелкие *Например* до 20 лет, 20—29, 30—39, 40—44, 45—49, 50—59, 60 и старше

Группировка данных по своей сути представляет собой процесс классификации, т е установление принадлежности явлений и объектов к определенным классам В государственной статистике для этого используются **классификаторы** — специальные справочники, инструкции и указатели в виде алфавитных и систематических словарей, дополняемых стандартным перечнем объектов и их групп Использование классификаторов в официальной статистике является обязательным и имеет силу государственного стандарта *Например* сводка данных о заболеваемости с временной утратой трудоспособности (форма 16-ВН) производится по специальным инструкциям, регламентирующим распределение нозологических форм заболеваний по определенным группам

диагнозов Аналогичным образом, по специальным инструкциям формируется статистика данных об общей, госпитализированной заболеваемости, инвалидности, смертности и т п.

Основные классификаторы рассчитаны на длительное использование Однако с течением времени они пересматриваются, дополняются, в них вносятся необходимые коррективы Так, с 1983 до 1999 года в СССР, а затем в России общепринятой считалась *Международная статистическая классификация болезней, травм и причин смерти* 9-го пересмотра С 1999 года вводится классификация 10-го пересмотра (с 01 01 99 г по Приказу Минздрава России № 3 от 12 01 98 г)

В медико-биологических научных исследованиях использование государственных и международных классификаторов болезней и причин смерти не является строго обязательным. Однако только их прямое применение или возможность путем вторичной группировки привести данные исследований в рамки, определяемые общепринятыми классификаторами, гарантирует сопоставимость статистических материалов о заболеваемости, причинах смерти разных категорий населения на различных территориях страны и за рубежом

2.4. Статистические таблицы. Правила оформления статистических таблиц

Возникновение статистики как науки тесно связано с появлением статистических таблиц Фактически вся программа разработки воплощается в статистических таблицах Более того, считается, что *в статистических таблицах — все красноречие статистики.* Статистические таблицы принципиально отличаются от перечневых таблиц **Перечневые** или описательно-информационные таблицы представляют собой простой перечень данных *Например* таблицы умножения, таблицы обратных величин, логарифмов, расписание железнодорожных перевозок и т п Перечневые таблицы могут использоваться на этапе предва-

рительной, первичной группировки исходных статистических данных

Приводимые в статистической таблице данные группируются особым способом, что дает возможность не просто их систематизировать, но и проводить анализ тенденций распределения, взаимосвязи различных явлений

Содержание и формы группировок соответствующих показателей в статистической таблице указываются в наименовании граф (столбцов) и строк, а величины показателей даются цифрами на пересечении соответствующих столбцов и строк (табл 3)

Таблица 3

Макет статистической таблицы

Название таблицы (общий заголовок)

Наименование подлежащего (верхний внутренний боковой заголовок)	Наименование сказуемого				
	Заголовки сказуемого (верхние внутренние заголовки)				
А	1	2	3	2+3	Итого
Подлежащее		Графо-клетки			
(боковые внутренние					
заголовки)					
Всего					

Статистическое подлежащее — это те группы или отдельные единицы статистической совокупности, которые характеризуются в таблице При статистическом анализе в подлежащем размещают показатели результативных признаков

Статистическое сказуемое — это признаки, которыми характеризуются группы или единицы совокупности В сказуемом, как правило, размещают числовые показатели факторных признаков Расположение и форма сочетания конструктивных элементов представляется в макете статистической таблицы

В целом, таблица должна быть по возможности небольшой по размеру, так как краткую таблицу легче анализировать Ино-

гда легче построить две-три небольшие таблицы, чем одну большую

Общий заголовок лаконично отражает содержание таблицы Если таблица приводится одна, вне текста или без пояснений в сопровождающем тексте, то обязательно указывается место и время отражаемого в ней явления Если в таблицу включаются данные из официального источника либо из опубликованной ранее научной работы, или данные других авторов, то обязательно указываются источники получения информации Для улучшения восприятия содержимого таблицы желательно указать в заголовке вывод, вытекающий из содержания таблицы *Например* заголовок «Заболеваемость населения обследованного района в 1992—1997 гг » целесообразно поменять на «Рост заболеваемости » или « Изменение структуры заболеваемости » В том случае, когда для всех клеток таблицы единицы измерения одинаковы, то в заголовке указывается размерность единиц («%», «на 1000 населения» и т п.)

Верхний внутренний боковой заголовок отражает содержание всех строк. По смыслу соответствует подлежащему статистической таблицы в целом Боковые внутренние подзаголовки, расположенные по строкам, раскрывают содержание отдельных строк или группы строк Если каждая строка имеет свою особую единицу измерения, то для их обозначения следует отводить особую графу

Верхние внутренние подзаголовки раскрывают содержание граф (столбцов) Если каждая графа имеет свою особую единицу измерения, то они приводятся в графе вместе с подзаголовком

Строки в подлежащем и графы в сказуемом часто нумеруются При этом в сказуемом нумеруются только графы В отдельную графу выносятся номера строк подлежащего В отдельную строку — номера граф сказуемого Графы подлежащего (они могут быть в случае сложного подлежащего) либо совсем не нумеруются, либо обозначаются буквами («а», «б», «в» и т д) Все это облегчает пользование таблицей Во-первых, при разрыве таблицы в случае ее размещения на двух или более страницах можно не повторять названия граф на последующих страницах, а указать только их номера Во-вторых, это дает возможность пока-

зать способ расчета производных показателей Порядок вычисления показателя можно указать тут же, в подзаголовке графы *Например.* для вычисления суммарных показателей данные в графах 2 и 3 суммируем и размещаем в графе, обозначенной «2+3» (см табл 3)

Не следует строить очень громоздких таблиц с большим числом граф Иногда целесообразно некоторые графы объединить под рубрикой «прочие» Однако не рекомендуется, чтобы эта группа охватывала более 10% итогов

Строки подлежащего и графы сказуемого обычно размещаются по принципу от частного к общему, т е сначала показывают слагаемые, а в конце подлежащего или сказуемого подводят итоги Если не приводятся все слагаемые, а выделяются наиболее важные из них, то сначала показывают общие итоги, а затем выделяют наиболее важные составляющие их части, для этого после итоговой строки дают «в том числе»

Итоги в статистической таблице имеют ключевое значение Без них статистическая таблица не считается законченной Анализ любой статистической таблицы следует начинать именно с итогов Ознакомление с итогами дает общее представление о данных таблицы. Этот порядок соответствует методологии анализа от *общего к частному.* Затем анализируют данные отдельных граф и строк, но их следует читать не подряд, а выбирать сначала наиболее характерные данные, а затем анализировать все остальные.

Вертикальные итоги получаются в результате суммирования чисел граф, ***горизонтальные*** — суммированием всех чисел строк Иногда вместо суммы чисел вычисляются средние показатели Если таблица групповая или комбинационная, то обязательно вычисляются промежуточные итоги в анализируемых подгруппах Особое значение это обстоятельство имеет для показателей структуры Следует обязательно указывать откуда ведется отсчет долей Если показатели доли выражаются в процентах, *например,* то в соответствующих строках или столбцах указывается 100%

Отсутствие итоговых показателей является методической ошибкой (см табл 8), поскольку не дает возможность ответить на вопрос о распределении показателей в целом и по отдельным

группам (возрастным и половым) В приведенном примере затруднен и сравнительный анализ различий показателей между мужчинами и женщинами в целом, вне зависимости от возраста Расчет промежуточных итогов по отдельным возрастным группам позволяет наглядно представить эти данные (см табл 9)

Следует различать «итого» и «всего». **«Итого»** является итогом для некоторой части совокупности **«Всего»** — для всей представленной в таблице совокупности

Цифровые данные целесообразно округлять При этом округление должно быть единообразным для всех данных в графе (столбце) Размерность приводимых показателей указывается либо в общем заглавии таблицы, либо, если они различны, в подзаголовках граф и строк При этом однотипные показатели приводятся с одинаковой точностью

Текстовые сокращения в графах и подзаголовках используются также единообразно, на основе общепринятых правил или стандартов Если возможно, расшифровка используемых сокращений приводится в общем заголовке таблицы. *Например* «Динамика показателей жизненной емкости легких (ЖЕЛ) »

Для выделения особо важных показателей или итогов по группам применяются различные технические приемы акцентирования (подчеркивание, изменение шрифта, выделение строк и граф двойной или жирной линией)

В каждой клетке таблицы должно стоять какое-либо число Иногда клетки могут оставаться пустыми Причины отсутствия чисел в таблице обязательно определенным образом указываются При *отсутствии сведений* о данном факте рекомендуется ставить три точки (.) или указывать «нет сведений» В случае сомнительных данных ставится знак вопроса (?) Если в клетке указан знак (*), то данные предварительные В подсчет итогов они, как правило, не включаются или их включение специально оговаривается При *отсутствии самого явления*, в связи с невозможностью его появления в принципе, ставится тире (—) Если изучаемое явление наблюдается в очень *малых размерах* и составляет менее половины последней значащей цифры в условиях принятой точности, в клетке выставляется ноль (0,0) Это говорит о том, что при меньшем округлении какая-либо значащая

цифра может появиться Если данная клетка *не подлежит заполнению*, то в ней ставится X

Примечания относятся обычно ко всей таблице и содержат, как правило, указания о методах получения исходной информации Сноски относятся либо к отдельным клеткам, либо к графам или строкам таблицы Отсутствие сноски или примечаний может приводить к серьезным погрешностям в трактовке табличных данных *Например* сноска может указывать на то, что продолжительность периода, за который оценивается заболеваемость детей в возрасте до 1 года, достаточно велика, чтобы сравнивать показатели накопленной заболеваемости с показателями других возрастных групп (больше период наблюдения — больше вероятность проявления заболевания)

В зависимости от насыщенности данными выделяют простые, групповые и комбинационные таблицы

В ***простой таблице*** дается распределсние, ряды распределения одного учетного признака (табл 4) Часто это распределение приводится вместе с одной или несколькими количественными характеристиками этого распределения Как правило, эти характеристики приводятся в виде относительных величин (интенсивных, экстенсивных, коэффициентов соотношения и т п)

Таблица 4

Распределение валового выброса антропогенных загрязнителей по обследованным районам города

Номер района	Валовой выброс (тонн в год)	По районам (в % к итогу)
1	113377	31,9
2	152962	43,0
3	10164	2,9
4	10444	2,9
5	69400	19,3
Итого	355347	100,0

В ***групповой таблице*** дается распределение единиц наблюдения по двум признакам (табл 5) Наряду с абсолютными показателями в таблице могут быть любые производные величины В том случае, когда в таблице приводятся показатели структуры, рекомендуется

указывать графу или строку, от которой идет отсчет частей (100%) При этом не имеет значения, как расположены показатели структуры — по строкам или столбцам Порядок расположения диктуется логикой проводимого анализа Например· на основании показателей структуры, вычисленных по строкам (см табл 5), можно говорить о структуре загрязнителей в том или ином обследованном районе промышленного города По показателям структуры, вычисленных по столбцам (табл 6), можно анализировать распределение того или иного загрязнителя по районам

Таблица 5

Структура антропогенных загрязнителей в обследованных районах по данным 1995 года в % к валовому выбросу в каждом районе (итоги по строкам)

Номер района	Загрязнители				Валовой выброс	
	Специфические		Не специфические			
	тонн/год	%	тонн/год	%	тонн/год	%
1	89568	79,0	23809	21,0	113377	100,0
2	136422	82,9	16520	17,1	152962	100,0
3	8812	86,7	1352	13,3	10164	100,0
4	10256	98,2	188	1,2	10444	100,0
5	60124	87,9	8276	12,1	68400	100,0
Итого	305202	85,9	50145	14,1	355347	100,0

Таблица 6

Структура антропогенных загрязнителей в обследованных районах по данным 1995 года в % к суммам различных видов загрязнителей (итоги по столбцам)

Номер района	Загрязнители				Валовой выброс	
	Специфические		Не специфические			
	тонн/ год	%	тонн/ год	%	тонн/ год	%
1	89568	29.3	23809	47,5	113377	31,9
2	136422	44,7	16520	32,9	152962	43,0
3	8812	2,9	1352	2,7	10164	2,9
4	10256	3,4	188	0,4	10444	2,9
5	60124	19,7	8276	16,5	68400	19,2
Итого	305202	100,0	50145	100,0	355347	100,0

Допускается вычисление показателей структуры и «уголком» (табл 7)

Таблица 7

Структура антропогенных загрязнителей в обследованных районах по данным за 1995 год в % к сумме всего валового выброса (итоги «уголком»)

Номер района	Загрязнители				Валовой выброс	
	Специфические		Не специфические			
	тонн/год	%	тонн/год	%	тонн/год	%
1	89568	25,2	23809	6,7	113377	31,9
2	136422	38,4	16520	4,6	152962	43,0
3	8812	2,5	1352	0,4	10164	2,9
4	10256	2,9	188	0,1	10444	2,9
5	60124	16,9	8276	2,3	68400	19,2
Итого	305202	85,9	50145	14,1	355347	100,0

Простые и групповые статистические таблицы дают возможность проводить, как правило, только ориентировочный, предварительный анализ В таких таблицах невозможно проследить взаимосвязи нескольких отдельных факторов. Более того, увиденные в этих таблицах тенденции могут быть ошибочно приписаны одному группировочному признаку Для изучения сравнительной роли различных факторов используется система комбинационных группировок, выполняемая с помощью ***комбинационных таблиц*** В таких таблицах оценивается распределение трех и более признаков (табл 8, табл 9)

Таблица 8

Неполная форма представления табличных данных о распределении рабочих по кратности заболеваний в течение года (в % к всего)

Группы по кратности заболеваний	Возраст и пол рабочих					
	До 35 лет		35 лет и старше		Итого	
	Муж	Жен	Муж	Жен	Муж	Жен
Не болели	31,0	12,0	18,2	25,7	22,4	22,1
Болевшие, в том числе	69,0	98,0	81,8	74,3	77,6	77,9
1 раз	27,7	23,0	28,6	13,5	28,3	16,3
2 раза	29,8	14,0	24,1	29,0	25,8	24,3
3 раза	17,0	23,0	17,9	23,1	17,6	23,0
4 и более	25,5	40,0	29,4	34,4	28,3	36,4
Всего	100,0	100,0	100,0	100,0	100,0	100,0

Таблица 9

Полная форма представления табличных данных о распределении рабочих по кратности заболеваний в течение года (в % к всего)

Группы кратности заболеваний	Возраст и пол рабочих								
	До 35 лет			35 лет и старше			Итого		
	Муж	Жен	Оба пола	Муж	Жен	Оба пола	Муж	Жен	Оба пола
Не болели	31,0	12,0	27,4	18,2	25,7	25,1	22,4	22,1	22,2
Болевшие, в том числе	69,0	98,0	72,6	81,8	74,3	74,9	77,6	77,9	77,8
1 раз	27,7	23,0	25,5	28,6	13,5	23,5	28,3	16,3	21,7
2 раза	29,8	14,0	21,0	24,1	29,0	25,6	25,8	24,3	24,9
3 раза	17,0	23,0	19,0	17,9	23,1	21,4	17,6	23,0	22,0
4 и более	25,5	40,0	34,5	29,4	34,4	29,5	28,3	36,4	31,4
Всего	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Существенным недостатком комбинационных таблиц является большое число наблюдений, которое требуется для получения с их помощью статистически достоверных данных Чтобы вычислить показатели в приведенном примере (табл 9), потребовалось более 1000 единиц наблюдения

2.5. Основы работы с электронной таблицей *MS Excel*

Microsoft Excel — это программа управления электронными таблицами, позволяющая обрабатывать числовые данные, создавать графики и анализировать информацию баз данных

Для запуска *Excel* нажмите кнопку **[Пуск]** на панели задач *Windows*, в Главном меню выберите строку ***Программы*** и затем щелкните по значку ***Microsoft Excel***. Иногда эта программа располагается в пакете *Microsoft Office* и для запуска *Excel* необходимо предварительно открыть эту строку После запуска на экране появится окно программы *Excel* (рис 2).

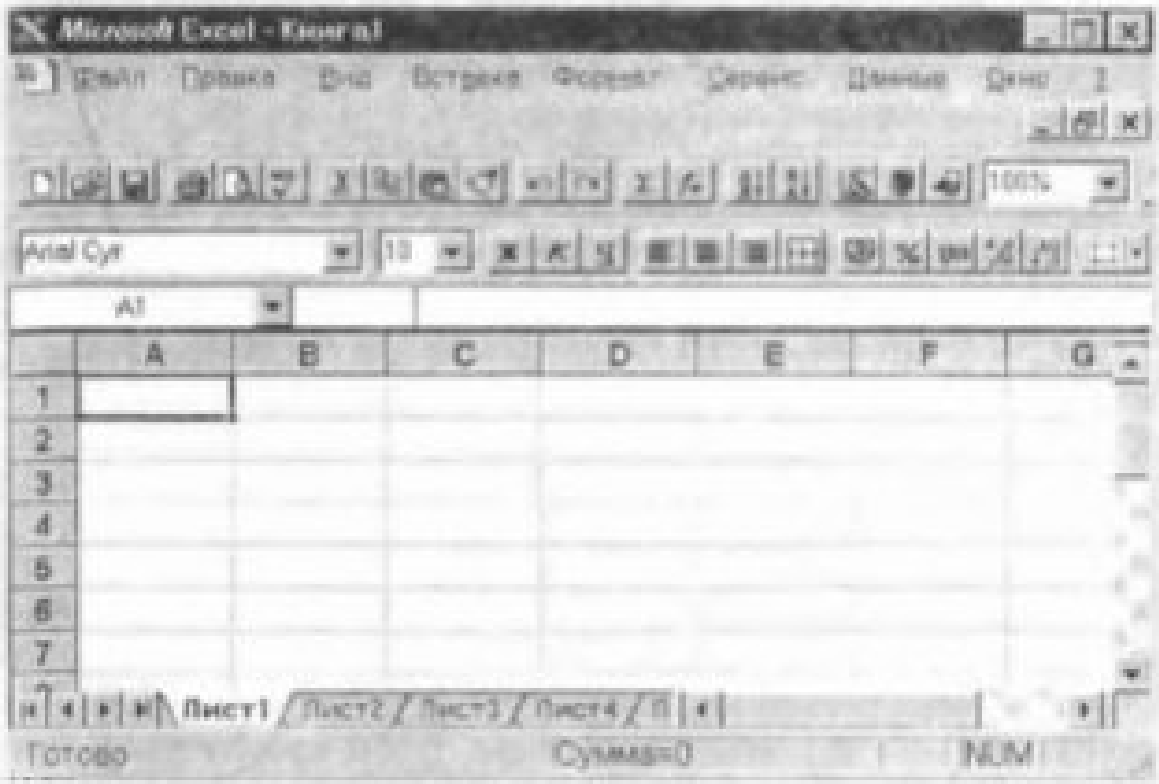


Рис 2 **Окно программы *Excel***

В заголовке окна *Excel* кроме названия программы отражается и имя редактируемого файла Вторая строка сверху в окне — это строка меню, обращение к которым открывает доступ ко всем командам и параметрам приложения

Под строкой меню находятся так называемые **панели инструментов** Кнопки на этих панелях называются инструментами Инструменты предназначены для выполнения самых распространенных команд и действий *Excel*

Чуть ниже панелей инструментов находится строка формул, предназначенная для ввода текста, чисел и формул в ячейки таблицы Если ввод данных осуществляется непосредственно в ячейки таблицы *Excel*, то в этой строке отображается вся вводимая информация

Основную часть экрана занимает окно рабочей книги, которое делится на строки и столбцы Столбцы обозначаются буквами латинского алфавита от A до ZZ, строки — цифрами Место пересечения строки и столбца называется ячейкой Адресом

ячейки служат буква и число, соответствующие пересекающимся в ячейке столбцу и строке Например, B5 обозначает ячейку на пересечении столбца B и строки 5

Один файл *Excel* может содержать несколько рабочих листов и называется рабочей книгой Для выбора нужного рабочего листа в нижней части окна находятся ярлычки рабочих листов На этих ярлычках написаны названия рабочих листов **Лист1, Лист2** и т д Щелкнув мышью на нужном ярлычке, вы вызовете на экран соответствующий рабочий лист

2.5.1. Ввод и редактирование данных

В электронной таблице одна из ячеек всегда является активной. ***Активная ячейка*** — это ячейка, выделенная указателем ячейки Смена активной ячейки производится с помощью клавиш управления курсором или мыши Чтобы сделать ячейку активной, достаточно выполнить щелчок мышью на этой ячейке

Ввод информации в активную ячейку выполняется в строке формул и заканчивается нажатием клавиши **[Enter]** или кнопки **[Ввод]**, которая находится слева от строки формул Для отказа от ввода в строке формул предназначена соседняя кнопка **[Отмена]**

Если длина введенного в ячейку текста превышает текущее значение ширины этой ячейки, то после завершения ввода текст либо будет полностью представлен в таблице, закрывая собой незаполненные ячейки, которые расположены справа, либо будет урезан по правому краю ячейки, если смежная с ней ячейка содержит какую-либо информацию Весь текст полностью можно увидеть в строке формул при помещении указателя ячейки на ячейку с данным текстом

Если же вследствие недостаточной ширины ячейки числовые значения в ней не могут быть представлены полностью, то на экране будет отображено соответствующее число символов диез (#), при этом содержимое ячейки полностью будет представлено в строке формул

Если в какую-либо ячейку введены неверные данные, то ошибка может быть устранена либо путем повторного ввода в ту же ячейку правильной информации, либо включением режима редактирования

Для редактирования содержимого ячейки необходимо

- установить указатель ячейки на данную ячейку;
- дважды щелкнуть мышью или нажать клавишу **[F2]**;
- изменить содержимое ячейки в строке ввода;
- для сохранения сделанных изменений нажать **[Enter]**

Для удаления содержимого ячейки установите указатель ячейки в эту ячейку и нажмите клавишу **[Delete]**.

2.5.2. Выделение блока ячеек

При работе с электронной таблицей одной из наиболее часто используемых операций является выделение блока ячеек Выделение блока ячеек служит для обозначения ячеек, к которым должна относиться следующая команда или функция Например, блок выделяется при копировании формул, форматировании таблицы, создании графиков и др

Для выделения (маркировки) блока ячеек с помощью клавиатуры необходимо разместить указатель ячейки на одной из угловых ячеек маркируемой области, нажать клавишу **[Shift]**, после чего передвинуть указатель ячейки с **помощью** клавиш управления курсором После того, как блок ячеек будет выделен, отпустить клавишу **[Shift]**

С помощью мыши выделение интервала выполняется путем нажатия левой кнопки на угловой ячейке и перетаскивания указателя мыши по остальным ячейкам интервала.

Для выделения несмежных диапазонов ячеек необходимо вначале мышью выделить первый блок ячеек, затем нажать клавишу **[Ctrl]** и, удерживая нажатой клавишу **[Ctrl]**, мышью выделить другие блоки ячеек

При выделении блока ячеек происходит их подсвечивание Чтобы отменить выделение блока ячеек, достаточно выполнить щелчок мышью вне выделенного фрагмента таблицы или нажать одну из клавиш управления курсором.

2.5.3. Ввод математических формул

Ввод формулы должен всегда начинаться со знака = (равно) или со знака + Формула может содержать обычные арифметические операторы, например, + (плюс), — (минус), * (умножить), / (разделить) Например, для получения в ячейке С1 суммы двух чисел, находящихся в ячейках А1 и В1, следует в ячейку С1 ввести формулу **=А1+В1.**

Для задания формулы можно использовать различные технические приемы Формула может быть задана путем ввода с клавиатуры Однако существует и другой способ задания формулы после ввода знака равенства следует выполнить щелчок мышью на ячейке, которая должна быть указана в формуле первой (А1) Адрес данной ячейки появится в итоговой ячейке (С1) Далее следует ввести оператор сложения, а затем выполнить щелчок на следующей ячейке (В1) Использование этого способа значительно упрощает ввод адресов ячеек

Кроме того, они могут использовать специально встроенные функции, которые облегчают процесс вычисления Например, статистические функции выполняют операции по вычислению различных величин или параметров их распределений стандартного отклонения, дисперсии, медианы и т. п

В формулах можно указывать не только отдельные ячейки, но и целые блоки ячеек Блоком называется прямоугольная группа ячеек Как ячейка определяется своим адресом, так и блок определяется своими координатами. В качестве координат блока указывается адрес левой верхней ячейки и через разделитель (точку или двоеточие) адрес правой нижней ячейки Например **=СУММ(А1:С5)** — сумма чисел, расположенных в 15 ячейках

Если в формуле указываются несмежные ячейки, то их адреса следует разделить точкой с запятой Например **=СРЗНАЧ (А1;В3;С5)** — среднее арифметическое чисел, расположенных в ячейках А1, В3 и С5

Если Вы знаете название функции, то можете ввести его в ячейку с клавиатуры. Аргументы функции должны быть указаны после ее названия в круглых скобках Поэтому после ввода от-

крывающейся круглой скобки следует выделить с помощью мыши ячейки, содержимое которых должно использоваться в качестве аргументов Адрес выделенного блока ячеек будет сразу же представлен в строке ввода Завершите задание функции вводом закрывающей скобки и нажатием клавиши **[Enter]**.

При вводе функции удобно работать с *Мастером функций* В этом случае для вставки функции в строку ввода необходимо

- выбрать команду <**Функции**> из меню <**Вставка**> или нажать кнопку **[Вставка функции]** на стандартной панели инструментов,
- в открывшемся окне "«Мастер функций» шаг 1 из 2" (рис 3) сначала выделить **категорию**, а затем выбрать требуемую **функцию**; выполнить щелчок на кнопке **[ОК]** для перехода в следующее диалоговое окно «Мастера функций»,

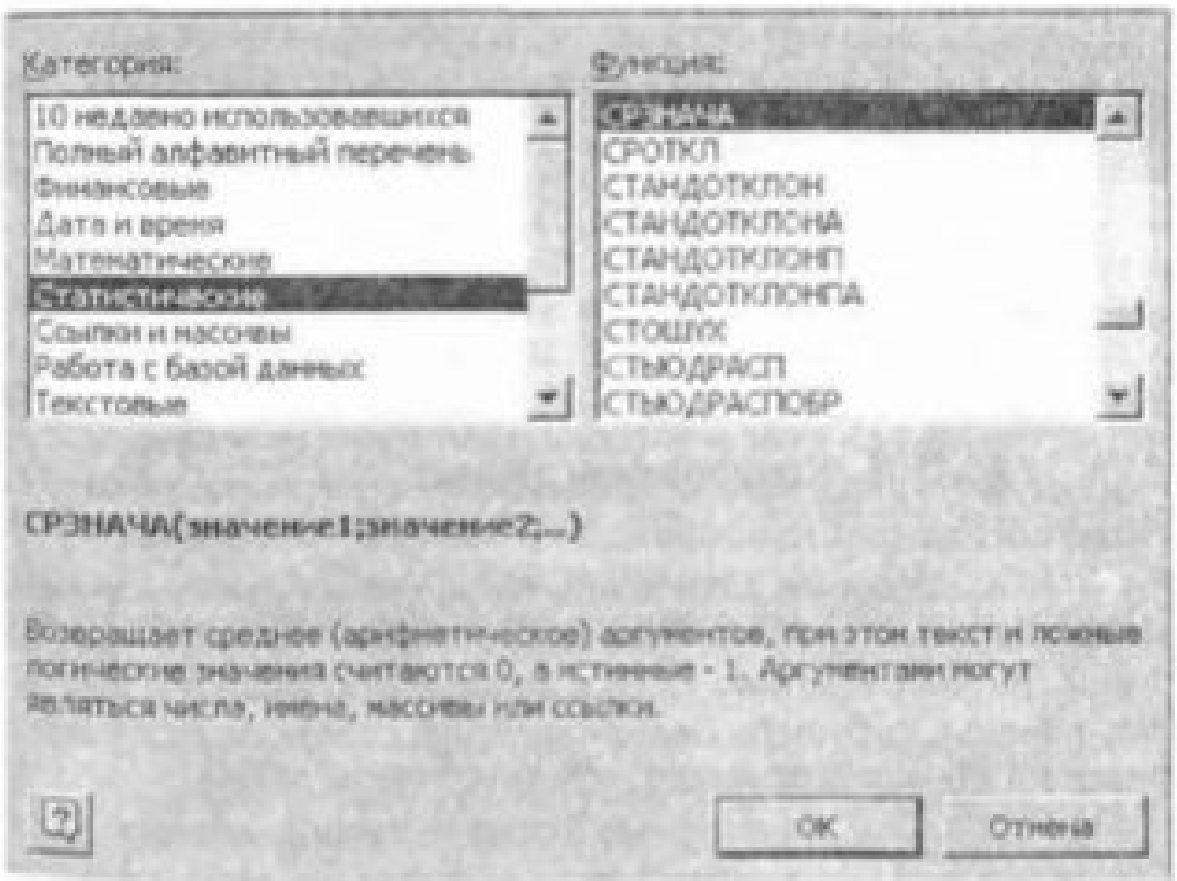


Рис 3. **Окно мастера функций. Шаг 1**

- во втором окне (рис. 4) указать аргументы функции — диапазон ячеек Для этого шелкнуть кнопку свертки — маленькую красную стрелку в правой части поля **Число1**. Окно диалога сворачивается, что позволяет выделить нужный диапазон ячеек в таблице (адреса первой и последней ячеек выделенного блока будут автоматически представлены в поле аргумента) По завершении ввода аргументов следует нажать клавишу **Enter**,

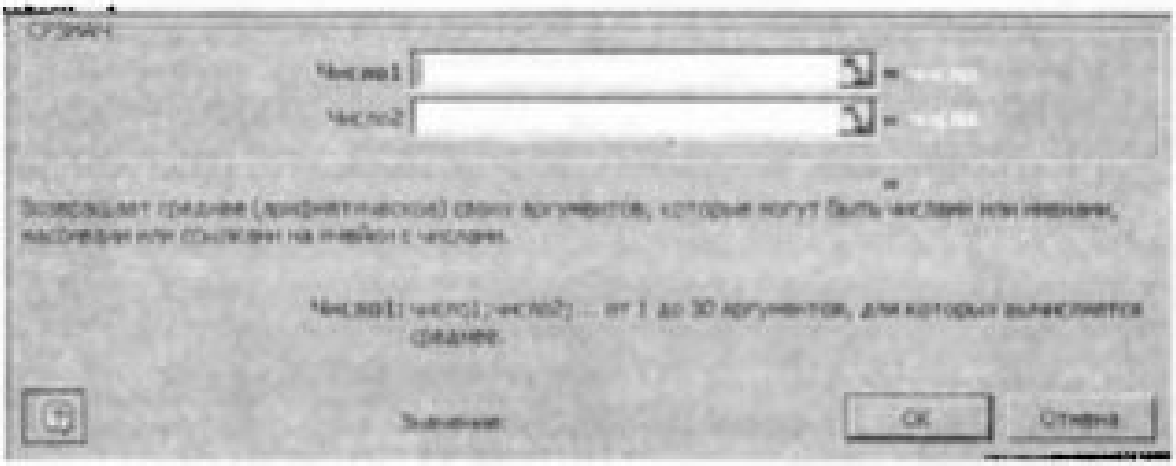


***Рис. 4.* Окно мастера функций. Шаг 2**

- нажать кнопку **[ОК]** для ввода функции и результат вычислений будет представлен в соответствующей ячейке

2.5.4. Копирование данных

Содержимое каждой отдельной ячейки или блока ячеек может быть скопировано. Операция копирования часто используется для многократного ввода в электронную таблицу одинаковых данных или формул Для копирования содержимого ячейки в интервал ячеек необходимо:

1 Установить указатель ячейки в ту ячейку, которую надо скопировать

2 Выбрать команду <**Копировать**> из меню <**Правка**> или нажать кнопку **[Копировать]** на панели инструментов

3 Выделить интервал ячеек, в который нужно скопировать данные, и нажать клавишу **[Enter]**.

При копировании формул в другое место таблицы необходимо управлять формированием адресов исходных данных Поэтому в электронной таблице при написании формул используются понятия относительной и абсолютной адресации

Абсолютный адрес — это не изменяющийся при копировании формулы адрес ячейки, содержащей исходные данные Для указания абсолютной адресации вводится символ $ Например, адрес ячейки B4 при копировании не будет изменяться, если в формулу записать ее адрес в виде $B$4 **Обратите внимание!** Если указать B$4, то столбцы при копировании могут меняться, а строки нет. И наоборот, если указать $B4, то столбцы меняться не будут, а строки — будут

Относительный адрес — это изменяющийся при копировании формулы адрес ячейки, содержащей исходные данные Такой адрес в своем имени не содержит символ $

2.5.5. Дублирование формул

Для дублирования (копирования) формулы в соседние ячейки можно использовать также команду **Заполнения** Для этого

1 Выделите ячейку с исходной формулой

2 Поместите курсор мыши на маркер заполнения (маленький квадратик) в правом нижнем углу копируемой ячейки При этом курсор должен приобрести вид маленького черного крестика

3 Перетащите мышью маркер заполнения по ячейкам, которые требуется заполнить

2.5.6. Формирование границ таблицы

Для выделения информации на листе полезно обводить ячейки рамками Для формирования рамки выполните следующие шаги

1 Выделите блок ячеек, вдоль границ которых должна быть проведена линия

2 Выполните команду <**Ячейка...**> из меню <**Формат**>

3 В диалоговом окне **Формат ячеек** перейдите на вкладку **Граница**

4 На вкладке **Граница** выберите местоположение, тип и цвет линии, а затем активизируйте кнопку **[OK]**

Если рамка сформирована не полностью, повторите описанные действия

Полезным заменителем вкладки Граница является кнопка **[Граница]** на панели инструментов форматирования, работа с которой происходит значительно быстрее

2.5.7. Формирование баз данных и сводных таблиц в *MS Excel*

В программе *MS Excel* имеется специальный пакет функций, которые позволяют создавать обширные базы данных, производить из них отбор необходимой информации, проводить ее сводку в статистические таблицы, вычислять производные величины и другие характеристики, необходимые для статистического анализа

В настоящем издании будут рассмотрены только некоторые приемы работы с указанным пакетом, позволяющие проводить минимально необходимый объем операций для решения задач формирования и обработки баз данных, сводки и статистического анализа табличных данных Знакомство с этими возможностями значительно облегчит более углубленное самостоятельное освоение указанного пакета программ

Информация в любой базе данных всегда располагается строго в определенном порядке В первой строке окна *MS Excel* обязательно указываются наименования столбцов, куда будут заноситься значения соответствующих учетных признаков Отдельная клетка этой строки, которая является заглавием конкретного столбца, называется **Поле** базы данных В данном примере (рис 5)

используются следующие поля: Возраст, Пол, Профессия, ЖЕЛ факт (Фактическая жизненная емкость легких), ЖЕЛ долж (Должная жизненная емкость) и Ф/Д (Отношение фактической к должной жизненной емкости легких)

	A	B	C	D	E	F
1	Возраст	Пол	Профессия	ЖЕЛ факт	ЖЕЛ долж	Ф/Д
2	49	Муж	Основная	4,02	3,33	120,72%
3	47	Жен	Вспомогательная	3,12	3,39	92,04%
4	51	Муж	Основная	3,23	3,32	97,29%
5	53	Муж	Вспомогательная	2,59	3,25	79,69%
6	52	Муж	Вспомогательная	3,13	3,53	88,67%
7	50	Жен	Вспомогательная	3,94	3,57	110,36%
8	42	Жен	Основная	3,91	3,4	115,00%
9	50	Жен	Основная	3,61	3,34	108,08%
10	43	Муж	Основная	5,56	4,78	116,32%
11	44	Жен	Основная	3,71	3,21	115,58%
12	50	Муж	Вспомогательная	3,76	3,19	117,87%
13	53	Муж	Вспомогательная	2,52	3,21	78,50%
14	47	Муж	Вспомогательная	3,49	3,43	101,75%
15	59	Жен	Основная	3,04	3,22	94,41%
16	40	Жен	Основная	3,28	3,43	95,63%
17	40	Жен	Вспомогательная	2,78	3,26	85,28%

Рис 5. **Пример базы данных**

Все учетные признаки по каждой отдельной единице наблюдения записываются в отдельную строку Содержимое этой строки называется **Записью** В данном примере представлено 16 записей, расположенных начиная со 2 по 17 строку, т е во второй строке записаны данные Иванова, в третьей — Петрова и т д

База данных может постоянно пополняться или модифицироваться *Например* в качестве исходной информации не обязательно вводить заранее вычисленное значение Ф/Д Можно ввести только исходные поля ЖЕЛ факт и ЖЕЛ долж , а потом вычислить их соотношение. Для этого необходимо

- в клетке F1 обозначить поле Ф/Д (если оно не было обозначено ранее);

- в клетку F2 ввести формулу =D2/E2*100,
- скопировать содержимое клетки F2 в блок ячеек F3:F17

После того, как исходная информация размещена на первом листе (в нижней части листа установлен переключатель **Лист1**) в соответствии с образцом (см рис 5), можно приступить к работе с базой данных. Значительно упростить эту работу можно пользуясь для просмотра, ввода и обработки данных окном формы, в котором база данных представлена как картотека (в каждый момент отображается карта только одной единицы наблюдения) В этом окне можно вести поиск записей по заданному критерию, дополнять список новыми записями, удалять ненужные, а также редактировать записи.

Окно формы данных можно открыть с помощью выбора команды **Форма** из меню <**Данные**>

В окне формы данных указывается общее количество записей в списке и позиция текущей записи С помощью полосы прокрутки можно перейти к любой записи списка Для перехода к предыдущей/следующей записи можно использовать также кнопки **[Назад]**/**[Далее]** или клавиши клавиатуры со стрелками, направленными вверх/вниз Кнопка **[Добавить]** позволяет ввести в базу данных новую запись Кнопка **[Удалить]** позволяет стереть из базы ненужные записи Если при редактировании данных были случайно стерты важные данные в каком-то из полей, их можно восстановить с помощью кнопки **[Вернуть]**

Для поиска определенных записей с помощью формы данных необходимо задать критерий поиска Перед тем как начать поиск целесообразно установить первую запись базы данных Допустим, что необходимо просмотреть данные о пациентах в возрасте 47 лет (см рис 5) Для решения этой задачи воспользуйтесь кнопкой **[Критерий]** В результате на экране будет открыта форма данных без записей Введите в поле Возраст в качестве критерия число 47 и нажмите кнопку **[Правка]** Начните поиск записи нажатием кнопки **[Далее]** После этого в окне будет представлена первая найденная запись, соответствующая заданному критерию Другие найденные в соответствии с критерием записи можно увидеть после повторного нажатия кнопки

[Далее] Кнопка **[Назад]** служит для просмотра найденных записей в обратном порядке

Могут задаваться и более сложные, составные критерии поиска Например, выполните поиск всех записей для женщин в возрасте меньше 40 лет. В этом случае необходимо при поиске по полю Пол использовать символ подстановки *, заменяющий любое количество любых букв, а при поиске по возрасту — оператор сравнения < Поэтому в качестве критерия поиска в поле Пол введите Ж*, а в поле Возраст <40

Сортировка баз данных. При сортировке базы все записи автоматически переставляются в соответствии с заданным порядком Порядок сортировки задается по содержимому любого из полей по возрастанию / убыванию (для числовых данных) или по алфавиту «от А к Я» / «от Я к А» (для текстовых данных)

Чтобы выполнить сортировку по выбранному полю базы данных

- выберите команду **Сортировка** из меню <**Данные**>, чтобы открыть диалоговое окно **Сортировка данных**,
- в окне откройте список **Сортировать по:** и выберите название нужного поля,
- установите переключатель **По возрастанию** или **По убыванию**;
- активизируйте кнопку **[ОК]**;
- убедитесь, что все записи расположились в требуемом порядке

Всего можно реализовать до трех уровней сортировки, например, выполнить сортировку сначала по полу, затем по профессии и затем по возрасту. Следует иметь в виду, что при сортировке по полям с повторяющимися значениями происходит фактически группировка записей (например, в верхней части таблицы будут показаны все записи для женщин и ниже все записи для мужчин и т. д)

Можно выполнять сортировку, используя кнопки панели управления Установите указатель ячейки в первой строке таблицы на заголовок нужного поля и, нажимая указанные кнопки, отсортируйте записи по возрастанию и убыванию

Применение функции автофильтрации. С помощью автофильтрации можно производить отбор записей, удовлетворяющих заданному критерию

Чтобы отфильтровать список, необходимо выполнить следующие действия

- выбрать команду **Фильтр → Автофильтр** из меню <**Данные**>, чтобы на именах полей в первой строке таблицы появились кнопки открытия списка столбца;
- щелкнуть мышью на кнопке списка столбца в требуемом поле, при этом должен появиться список значений, встречающихся в столбце;
- выбрать нужное значение,
- убедиться, что в таблице отображаются записи только с указанным значением

Обратите внимание, что после применения функции автофильтра в строке состояния появляется сообщение о количестве найденных записей

Можно выполнять отбор одновременно по нескольким полям Все кнопки списков столбца, на которых заданы условия отбора, окрашиваются в синий цвет

Примечание. Список столбца содержит значение **(Все)** Используйте эту позицию для отмены результатов фильтрации столбца

Можно выполнять отбор записей формируя и более сложные логические условия, для этого в списке столбца нужно выбрать позицию **(Условие...)** и в появившемся диалоговом окне **Пользовательский автофильтр** задать пару условий, объединяя их логической связкой «И» или «ИЛИ» Например, если нужно получить список пациентов в возрасте от 30 до 39 лет, нужно в поле Возраст вывести диалоговое окно **Пользовательский автофильтр**, в котором задать условие Возраст [больше] [29] [И] [меньше] [40], выбирая соответствующие значения из раскрывающихся списков или вводя с клавиатуры

Чтобы выйти из режима автофильтрации, выполните повторно команду **Фильтр → Автофильтр** из меню <**Данные**>

Вычисление промежуточных итогов. Для предварительно сгруппированных баз данных можно выполнять автоматическое вычисление некоторых итоговых результатов для каждой из групп записей Например, можно рассчитать средний возраст мужчин и женщин, среднеквадратическое отклонение от среднего ЖЕЛ факт в разных профессиональных группах и т д

Для подведения итогов по заданному столбцу необходимо выполнить следующие действия

- выполнить сортировку списка по заданному столбцу (с помощью команды **Сортировка** из меню <**Данные**>), образовав группы записей;
- выбрать команду **Итоги** из меню <**Данные**>;
- заполнить диалоговое окно **Промежуточные итоги,**
- активизировать кнопку **[ОК]**

Пример Необходимо получить данные о среднем возрасте пациентов разных профессиональных групп

Для решения этой задачи

- отсортируйте список по полю Профессия;
- выполните команду **Итоги** из меню <**Данные**>, чтобы открыть диалоговое окно **Промежуточные итоги,**
- выберите в поле **При каждом изменении в:** заголовок столбца, для групп которого необходимо вычислить итоги, то есть Профессия,
- для вычисления среднего арифметического значения при определении итогов в поле **Операция:** выберите функцию **Среднее,**
- в поле **Добавить итоги по:** включите индикатор (птичку) для столбца, ячейки которого будут использоваться для вычисления итогов, т е. для столбца **Возраст,**
- установите флажки в опциях **Заменить текущие значения** и **Итоги под данными;**
- активизируйте кнопку **[ОК].**

В результате таблица будет дополнена строками, в которых для каждой из профессий отображается итог' средний возраст В последней из вставленных в таблицу строк содержится общий итог для всех профессий вместе

Кроме того, при вычислении итогов таблица структурируется и в верхнем левом углу таблицы появляются кнопки уровней структуры Чтобы отобразить на экране только итоговые данные, следует выполнить щелчок на кнопке для второго уровня структуры Чтобы снова вывести на экран весь список, необходимо выполнить щелчок на кнопке для третьего уровня Кнопка первого уровня отображает на экране только общий итог по всей таблице

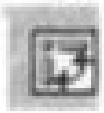
Для удаления строк с итогами, следует использовать кнопку **[Убрать все]** в диалоговом окне **Промежуточные итоги**

Для размещения **сводной таблицы** лучше всего использовать новый лист С этой целью переключите указатель в нижней части листа *MS Excel*, на котором Вы находитесь, в положение **Лист2** Затем, установите курсор в верхний левый угол листа (позиция A1). После этого через команду **Вид** и **Панели инструментов** нужно вызвать пакет **Сводные таблицы** На экране появится окно инструментов сводных таблиц (рис. 6).

***Рис. 6.* Окно инструментов сводных таблиц**

Окно мастера сводных таблиц может располагаться непосредственно на листе, с левого или правого края, снизу или вверху листа. Для того чтобы переместить его в наиболее удобное для работы место, нужно установить указатель мыши на поверхность окна мастера (см рис 6) и нажав левую клавишу мыши перетащить его в нужное место.

Затем последовательно выполните следующие действия:

1. Установите указатель мыши на значок мастера таблиц  и щелчком левой кнопки запустите мастер

2 В открывшемся окне мастера сводных таблиц (**Шаг 1 из 4**) укажите источник информации, откуда будут поступать исходные данные В данном примере необходимо указать **в списке или базе данных** *Microsoft Excel*, после чего нажмите кнопку **[Далее>]**

3 В открывшемся окне **шаг 2 из 4** укажите диапазон исходных данных В этот диапазон обязательно должна попасть первая строка с именами полей базы данных (рис 7) После этого нажмите кнопку **[Далее>]**.

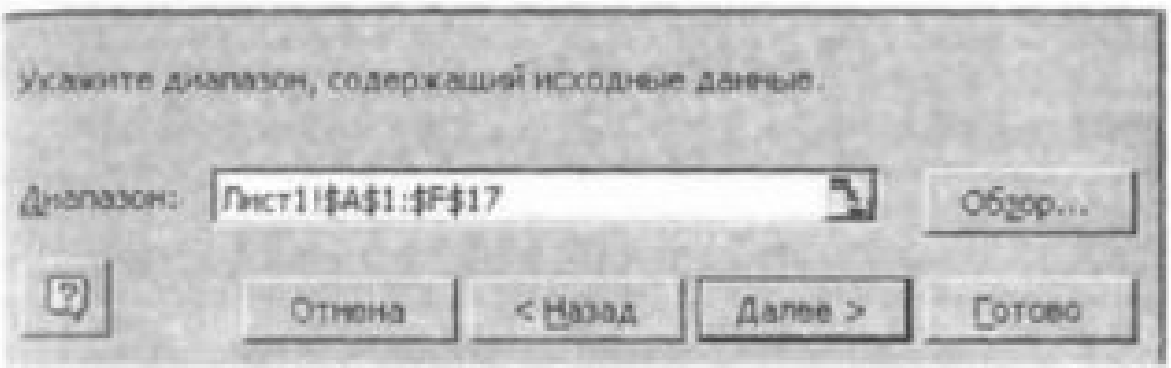


Рис. 7. **Окно указателя диапазона исходных данных мастера таблиц (шаг 1 из 4)**

4. На экране появится следующее окно мастера таблиц. В этом окне, справа, указан перечень учетных признаков, которые можно включать в таблицу Установите указатель мыши на нужном вам признаке, нажмите левую клавишу мыши и не отпуская ее перетащите признак в нужное место Расположите информацию так, как указано на рис 8. Признак *Пол* нужно расположить в области столбцов, признак *Возраст* — в области строк, признак *Профессия* — в области страниц, признак *Ф/Д* располагается в центральной области данных. Обратите внимание для признака *Ф/Д* предусматривается вывод информации в виде количества значений (**Количество значений признака**) Если у Вас установлен вывод другой информации (среднее, сумма и т д), тогда установите указатель мыши на сообщение *Ф/Д*, дважды быстро щелкните правой клавишей и в открывшемся окне **Вычисление поля сводной таблицы** выберите операцию **Количество значений** и нажмите клавишу **[ОК]**

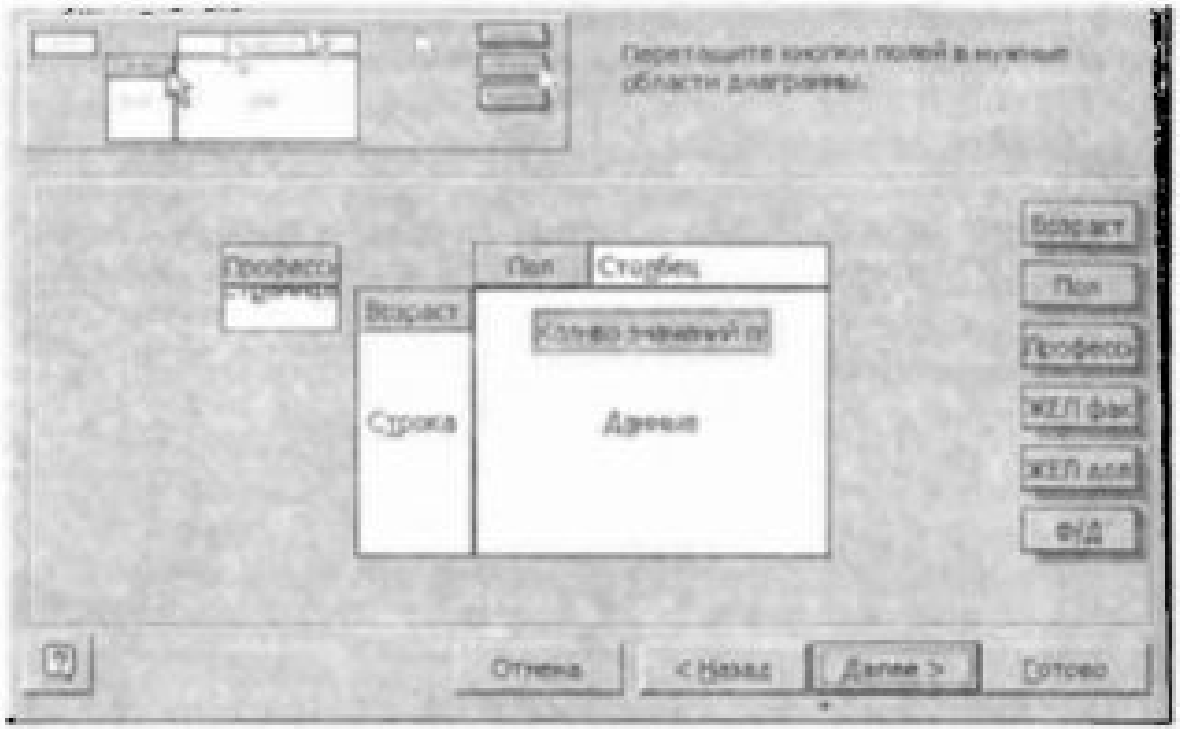


Рис 8 **«Мастер сводных таблиц» (шаг 2 из 4)**

После подготовки макета таблицы нажмите клавишу **Готово** На экране появится групповая таблица распределения в абсолютных числах (рис 9) Групповой эта таблица считается потому, что в ней приводится распределение по двум учетным признакам пол и возраст. Если использовать переключатель ***Профессия***, который установлен выше полученной таблицы, то можно рассматривать результаты распределения с учетом той или иной профессии Таким образом, распределение будет уже учитывать три группировочных признака Таблица в этом случае становится комбинационной

Для того чтобы выбрать с помощью переключателя третий группировочный признак, установите указатель мыши на значок ▼ и щелкните правой клавишей. Затем выберите из появившегося перечня нужный вам признак

Полученные результаты основаны на распределении показателей *Ф/Д* Поскольку эти данные изначально были у каждой единицы наблюдения, то по ним можно судить о распределении всей анализируемой совокупности

1	Профессия	(Все)	
2			
3	Кол-во значений по полю Ф/Д	Пол	
4	Возраст	Жен	Муж	Общий итог
5	40	2		2
6	42	1		1
7	43		1	1
8	44	1		1
9	47	1	1	2
10	49		1	1
11	50	2	1	3
12	51		1	1
13	52		1	1
14	53		2	2
15	59	1		1
16	Общий итог	8	8	16

Рис. 9. **Таблица распределения в абсолютных числах**

Для того чтобы получить распределение в любой другой форме производных величин, нужно воспользоваться функцией **Вычисление поля сводной таблицы**

Обратите внимание! Добиваться каких-либо изменений в таблице можно только тогда, когда Вами выделена хотя бы одна клетка этой таблицы.

В окне мастера таблицы функцию вычисления поля таблицы можно вызвать значком  .

В открывшемся вторичном окне **Вычисление поля сводной таблицы** выберите операцию **Среднее** и нажмите клавишу **[ОК]** В таблице после этой операции будет представлено распределение средних показателей *Ф/Д* (рис 10) Для того чтобы изменить числовой формат полученных результатов, нужно в окне **Вычисление поля сводной таблицы** выбрать операцию **Формат.....** и задать необходимую размерность числового формата

Профессия	(Все)		
Среднее по полю Ф/Д	Пол		
Возраст	Жен	Муж	Общий итог
40	0,904514479		0,904514479
42	1,15		1,15
43		1,163179916	1,163179916
44	1,15576324		1,15576324
47	0,920353982	1,017492711	0,968923347
49		1,207207207	1,207207207
50	1,09223989	1,178683386	1,121054389
51		0,972891566	0,972891566
52		0,886685552	0,886685552
53		0,790984903	0,790984903
59	0,944099379		0,944099379
Общий итог	1,020465667	1,001013768	1,010739718

Рис. 10 **Таблица распределений в средних величинах**

Аналогичным образом можно использовать для представления данных и другие производные величины (дисперсии, среднеквадратические отклонения и т. п)

Вычисление некоторых производных величины можно устанавливать через опцию **Дополнительно>>** в окне **Вычисление поля сводной таблицы** Для того чтобы получить, например, процентное распределение (распределение в экстенсивных показателях), необходимо предварительно указать поле, т е группировочный признак, относительно которого будет производиться расчет указанных величин.

Одной из самых употребительных функций мастера сводных таблиц является функция дополнительной группировки анализируемых признаков. Например, данные, представленные в полученных ранее таблицах (рис 9 и рис 10), показывают распределение по достаточно мелким возрастным интервалам, т е практически по каждому отдельно взятому возрасту Для объединения всех возрастных интервалов в 2 группы (от 40 до 49 лет и от 50 до 59 лет) необходимо выполнить следующие операции

1 Установить указатель мыши в ту клетку столбца **Возраст**, в которой указан возраст **40** Затем нажав правую клавишу мыши и не отпуская ее выделите диапазоны **40 42 43 44 47 49**

2. Вызовите из окна мастера сводных таблиц нажатием на клавишу → команду **Сгруппировать** В таблице появится еще один столбик **Возраст2** и в нем **Группа1** Аналогичным образом проведите группировку остальной части возрастных интервалов

3. После того как эта группировка выполнена, уберите из таблицы столбик **Возраст** Для этого установите указатель мыши на подпись столбца **Возраст** и не отпуская клавишу перетащите надпись на любое свободное место листа После этого в таблице будет оставлен только столбик сгруппированных данных Названия **Группа1** и **Группа2** можно изменить Для этого выделите клетку таблицы с названием, нажмите на клавишу (клавиатуры компьютера) F2 и введите нужное вам название, удалив все ненужное (рис 11)

1	Профессия	(Все)		
2				
3	Среднее по полю Ф/Д	Пол		
4	Возраст2	Жен	Муж	Общий итог
5	Группа1	1,007029236	1,129293278	1,052878252
6	Группа2	1,04285972	0,924046062	0,968601184
7	Общий итог	1,020465667	1,001013768	1,010739718

Рис. 11. **Сводная таблица сгруппированных по возрасту данных**

Изменения в группировках будут автоматически сохранены и ими можно пользоваться при дальнейшей работе, вызывая их в необходимых случаях в окне мастера сводных таблиц (шаг 3 из 4)

2.6. Графические изображения. Правила построения графических изображений

Графические изображения, использующиеся для более наглядного отображения статистических данных, называются **диаграммами** В некоторых случаях диаграммы позволяют прово-

дить более точный анализ, поскольку при их помощи легче уяснить закономерности развития, распределения и размещения явлений

Часто разного рода ошибки и неточности выявляются именно при применении диаграмм Весь вопрос в том, как найти правильное графическое решение для анализа данных.

При построении статистической диаграммы необходимо правильно выбрать графический образ диаграммы и ее экспликацию. **Экспликация** включает словесные пояснения к помещенным на графике геометрическим фигурам и вспомогательные изобразительные средства (системы координат, шкалы, масштабные сетки, наименование графика, единиц измерения, числовых данных и отдельных деталей) Целесообразно придерживаться следующих правил построения диаграмм

1 Общая структура диаграммы должна предполагать чтение слева направо

2. Следует избегать попыток изображения линейных величин с помощью площадей и объемов, как не соответствующих сути показателей Кроме того, следует помнить, что из-за обмана зрения могут возникать ошибки сравнительного восприятия отображаемых величин *Например* на приведенном рис 12 все три фигуры имеют одну площадь.

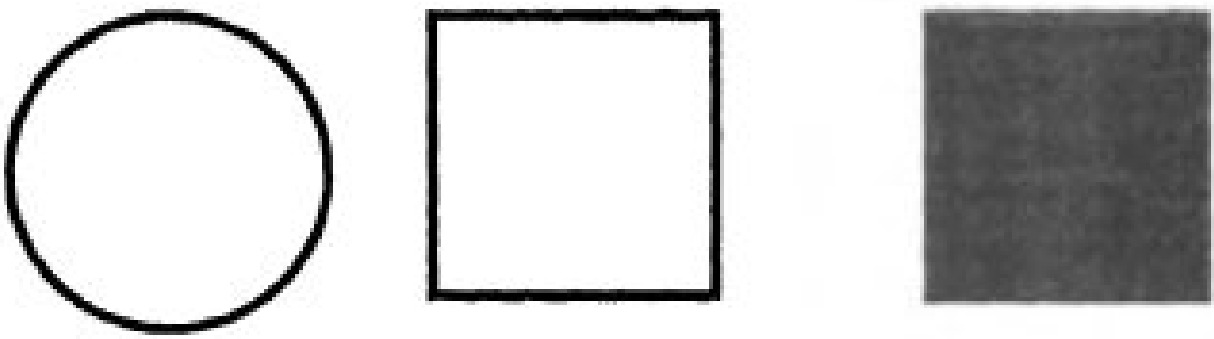


Рис. 12. **Фигуры одинаковых площадей**

3. Вертикальную шкалу для кривой независимо от ее назначения следует выбрать так, чтобы на диаграмме оказалась нулевая отметка. Иногда это невозможно, *например*, из-за больших значений показателей (рис. 13)

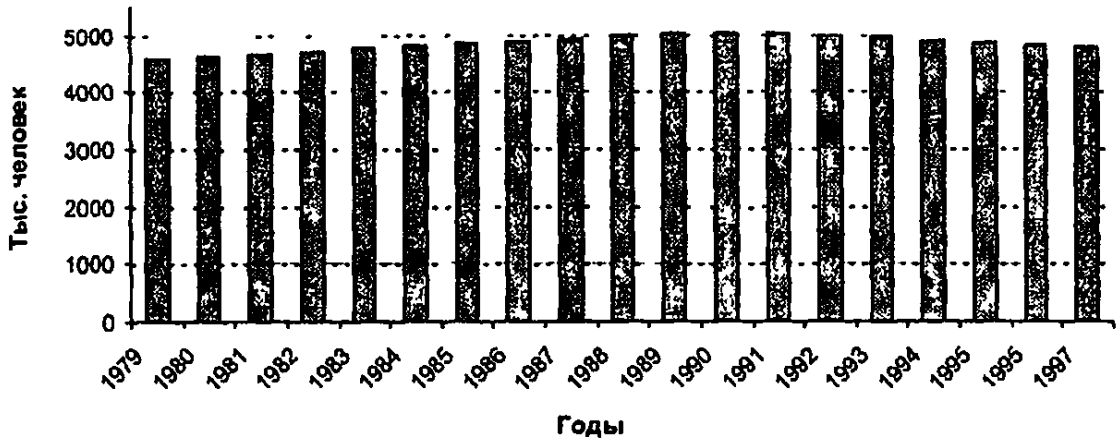


Рис. 13 **Динамика численности населения Санкт-Петербурга в 1979—1997 годах**

В этом случае отсчет шкалы целесообразно делать по возможности от круглого числа, либо от уровня имеющего какое-либо смысловое значение (стандарт, среднее и т п), рис 14

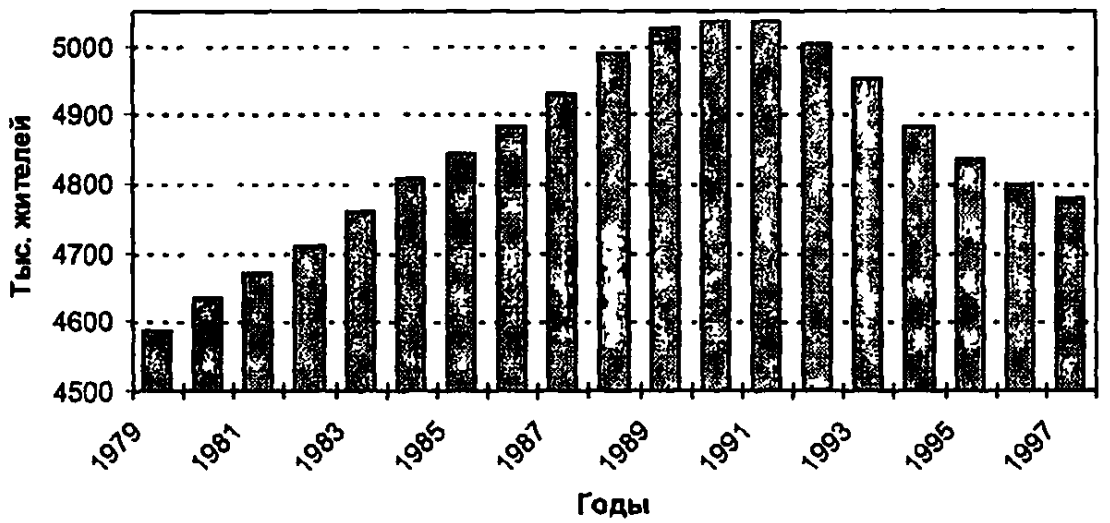


Рис 14 **Динамика численностн населения Санкт-Петербурга в 1979—1997 годах**

4 Для кривых, имеющих шкалу, изображающую проценты, промилле и т п , каким-то образом выделяются соответственно 100, 1000, 10000 и т д Целесообразно выделять величины, обозначающие норму, стандарт или средний уровень показателей

5 Когда шкалы относятся к датам, лучше не выделять первые и последние ординаты, т. к подобные диаграммы, как правило, не отражают начало и конец времени.

6 Для кривых, характеризующих группы наблюдений, рекомендуется по возможности ясно указывать на диаграмме все кривые, представляющие отдельные наблюдения

7 Горизонтальную шкалу для кривых следует читать, как правило, слева направо, а вертикальную — снизу вверх Если отображаемые данные резко отличаются друг от друга по своей величине, рекомендуется делать разрыв масштабной шкалы Этот же прием применяется если нет данных за какой-либо отрезок анализируемого периода При этом необходимо соблюдение двух условий Во-первых, данные должны быть однородны, во-вторых, разрыв должен быть обозначен и на построенной кривой В том случае, когда вырезки делать нецелесообразно (необходим анализ всего числового ряда без промежутков), рекомендуется использовать логарифмические шкалы *Например* (табл 10, рис 15)

Таблица 10

Рост первичной заболеваемости населения Санкт-Петербурга сифилисом

Год	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Случаев на 100 тыс чел	10,1	7,7	7,7	13,7	20,8	29,9	76,0	173,0	267,8	239,6

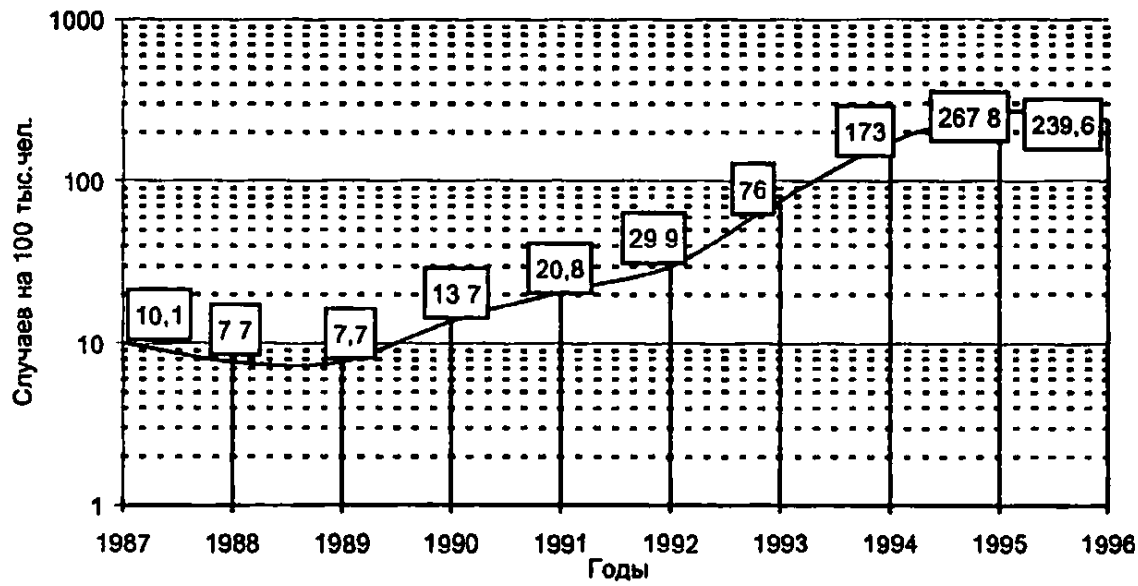


Рис. 15 **Рост первичной заболеваемости населения Санкт-Петербурга сифилисом**

8 Цифры на шкалах следует располагать слева и снизу вдоль соответствующих осей Если цифровые данные не попали на диаграмму, желательно привести данные в таблице, сопровождающей диаграмму

9 Желательно включать в диаграмму цифровые данные или используемые формулы

10 Наименования следует давать возможно яснее и полнее Если это требуется, необходимо вводить подзаголовки и пояснения

11. При использовании условных обозначений необходимо давать пояснения к ним.

12 Наименования графических изображений в книгах, журналах и т п обычно указывают снизу от рисунка Названия таблиц — вверху В диаграммах, не предусмотренных для печати, *например* настенных диаграммах, слайдах, целесообразно писать заголовки сверху

13 При построении линейной диаграммы в двухосной системе координат соотношение горизонтальной и вертикальной осей по длине целесообразно выбирать на основе принципа золотого сечения Это такое сечение, при котором отношение целого отрезка к большей его части равняется отношению большей части к меньшей В наиболее обобщенном виде это соотношение равно 3 к 2 (рис 16)

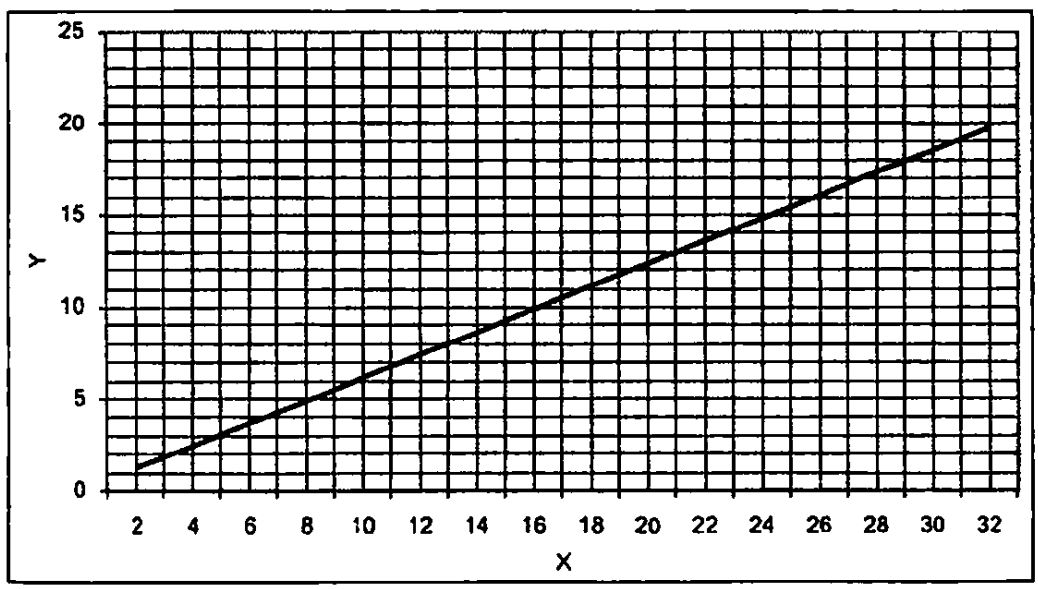


Рис 16 **Соотношение осей X и Y при соблюдении пропорций «золотого сечения»**

14 При использовании в нескольких последовательно расположенных диаграммах одних и тех же учетных признаков, применяются обязательно одни и те же условные обозначения для этих признаков

При построении секторной диаграммы начало отсчета производится от верхней точки («12 часов») и по ходу часовой стрелки. Следует помнить, что секторная диаграмма не допускает разбиения на большое число секторов (частей) Не рекомендуется использовать эту диаграмму для отображения более 5—7 показателей Если такая необходимость существует, то нужно использовать другой тип диаграмм Целесообразно откладывать числовые значения признака от большего к меньшему Если этот порядок противоречит логической последовательности данных, то он может быть нарушен *Пример* секторные диаграммы отражают распределение обследованных женщин по числу прерванных беременностей (рис 17)

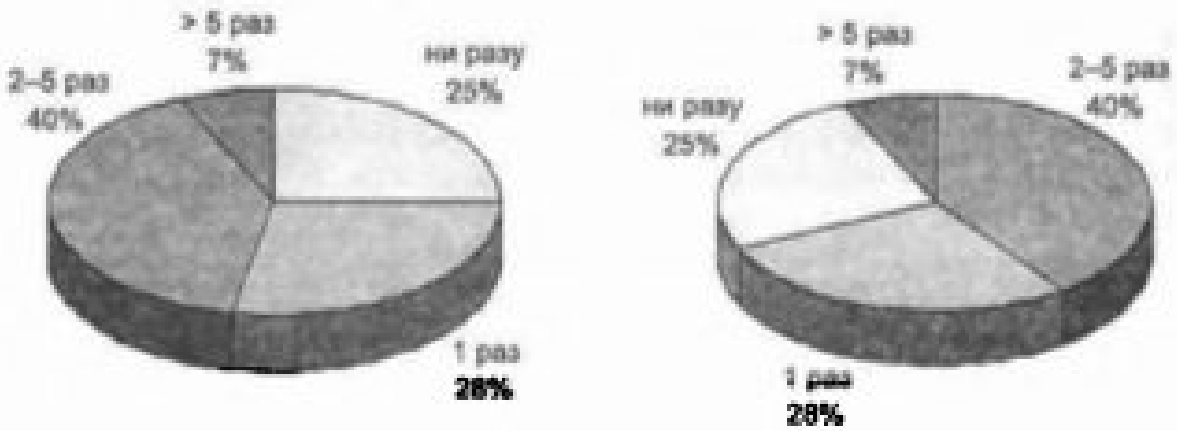


Рис 17. **Секторные диаграммы**

На диаграмме отображена ситуация, когда группы распределены по возрастанию их доли в общем числе обследованных В другом варианте — ранжирование по кратности прерванных беременностей Правомерность выбора того или иного приведенного изображения зависит только от конкретных задач, решаемых исследователем

2.6.1. Основные типы диаграмм

В настоящее время, благодаря широкому использованию персональных компьютеров и пакетов специализированных прикладных программ, фактически не существует никаких ограничений, которые ранее диктовались трудоемкостью создания тех или иных типов диаграмм Для создания графических изображений без помощи компьютеров, а также для первичного, ориентировочного анализа статистических данных, могут использоваться традиционные приемы с малыми подручными средствами Одним из простейших вариантов является построение диаграммы с помощью пишущей машинки (рис 18)

```
У !   xxxxxxxxxxxxx
  !  xxxxxxxxxxxxxxxxx
  !   xxxxxxxx
О !--------------------------X
```

Рис. 18. **Диаграмма, выполненная на пишущей машинке**

Нетрудно заметить, что приведенная диаграмма выполнена с помощью буквы «х» на пишущей машинке Естественно, условному значку «х» соответствует определенный масштаб отображаемых значений Можно пойти еще дальше, если абстрактные геометрические фигуры заменить рисунками, соответствующими определенному содержанию статистических данных Так, несложно выполнить диаграмму с помощью каких-либо трафаретных фигур *Например*: сравнительное потребление чая мужчинами и женщинами Каждая фигура — определенное количество чая (рис 19)

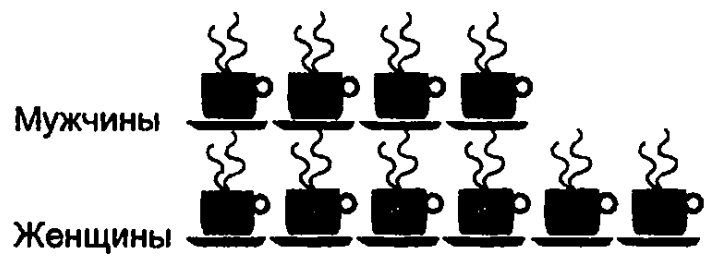


Рис 19. **Сравнительное потребление чая мужчинами и женщинами**

Фигурные диаграммы наиболее целесообразно применять при демонстрации каких-либо данных для широкой аудитории, не имеющей специальной подготовки (санитарно-просветительная работа, массовая агитация и т. п)

Линейные диаграммы — наиболее распространенный вид диаграмм. Применяется для отображения практически любых статистических величин. Этот вид графических изображений относится к координатным диаграммам, то есть диаграммам, использующим координатную систему Для более наглядного отображения различий кроме обычных координатных осей рекомендуется использовать координатную сетку (рис 20)

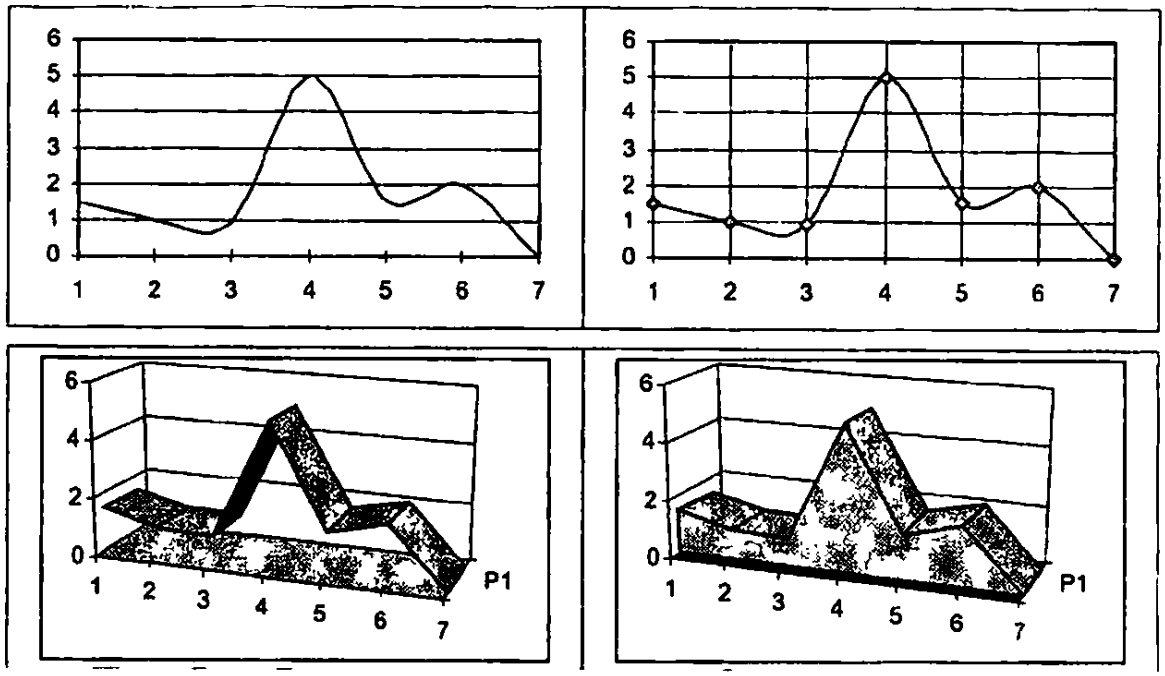


Рис 20. **Примеры компоновки (экспликации) линейных диаграмм**

Если графическое изображение плоскостное, то применяется двухмерная система координат. Если объемное — трехмерная Вариант трехмерной линейной диаграммы называется ленточной диаграммой.

Несмотря на большую изобразительную привлекательность, трехмерные диаграммы менее наглядны и не допускают включения более 2—3 аналитических рядов Достоинством плос-

костных линейных диаграмм является возможность наглядно отражать распределение большего количества рядов, в том числе рядов, состоящих из несопоставимых показателей Это могут быть разнородные показатели или однородные, но сильно отличающиеся по своим значениям показатели *Например*, если на диаграмме требуется представить динамику двух разнородных показателей, то используют не одну, а две масштабных шкалы. Одну из них размещают справа, а другую — слева (рис. 21).

Аналогичным образом поступают, если требуется изобразить очень растянутую диаграмму и для одной какой-либо числовой последовательности

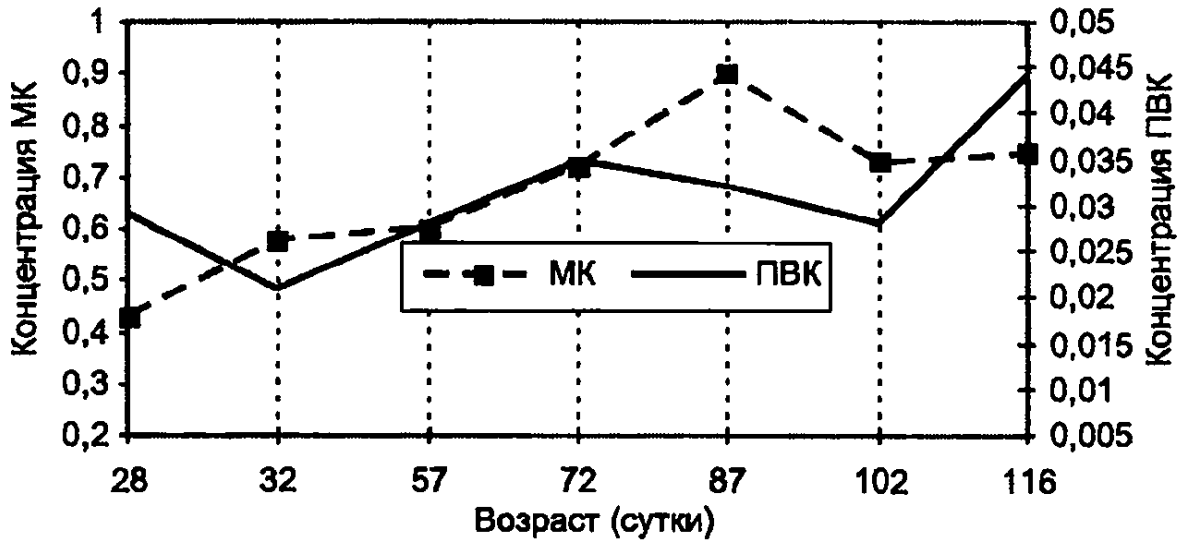


Рис. 21. **Динамика концентраций молочной (МК) и пировиноградной (ПВК) кислот у подопытных животных (в мг/л)**

Одним из способов повышения наглядности диаграмм с разнородными показателями является использование комбинированных графиков (рис 22)

Однако подобное сравнение кривых не всегда дает достаточно яркой картины динамики. Кроме того, применение этого метода не возможно в тех случаях, когда необходим одновременный сравнительный анализ многих разнородных показателей (табл. 11, рис 23).

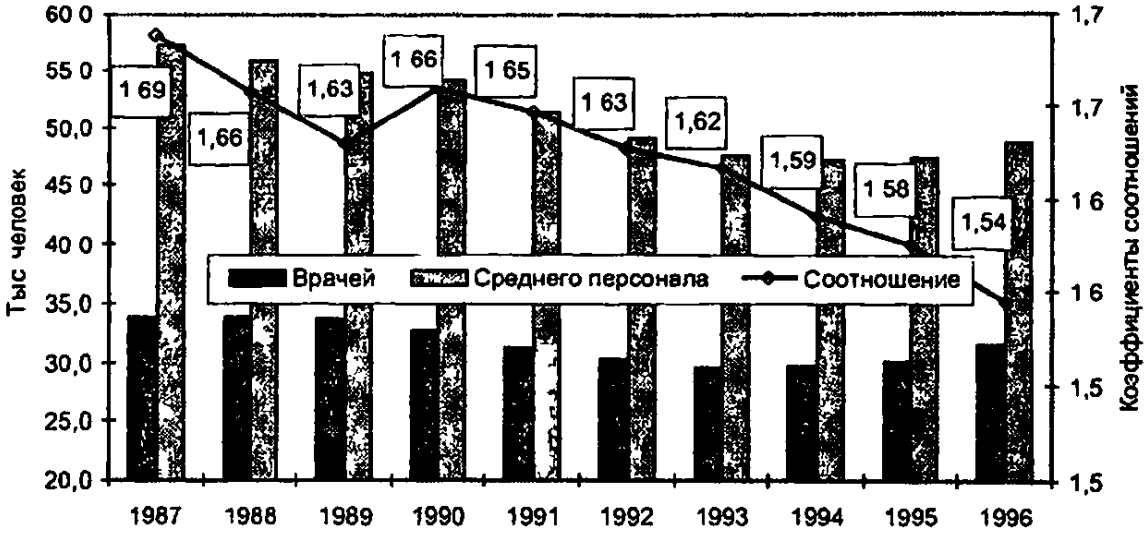


Рис. 22. **Динамика численности врачей и среднего медперсонала в Санкт-Петербурге**

Таблица 11

Динамика заболеваемости в показателях наглядности (в % к уровню 1987 года)

Заболевания	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Злокачественные новообразования	323,6	319,6	323,9	326,1	324,2	343,6	355,6	348,2	350,6	374,5
Острая гонорея	168,0	185,7	206,5	185,7	207,9	309,0	428,0	341,5	259,6	178,4
Активный туберкулез	25,6	24,7	22,8	20,5	25,7	26,7	34,5	41,9	40,3	43,0

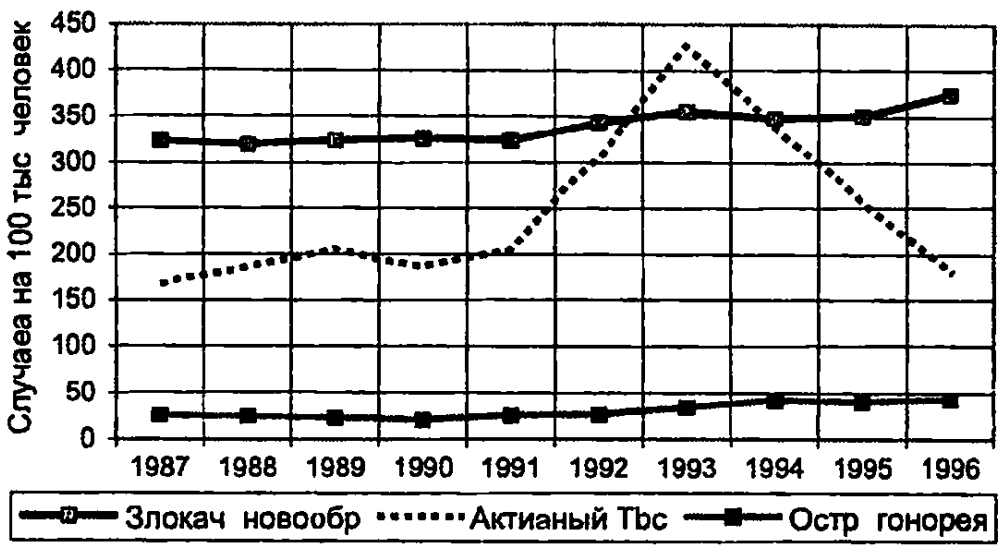


Рис 23. **Динамика первичной заболеваемости жителей Санкт-Петербурга (случаев на 100 тыс. населения)**

В этой ситуации диаграмму наиболее целесообразно строить на основе использования одного масштаба, после преобразования исходных данных в относительные величины В приведенном примере исходные данные о заболеваемости переведены в показатели наглядности За исходный, нулевой уровень принят 1987 год. Остальные по каждой нозологической форме рассчитываются в процентах относительно этого исходного уровня (табл 12, рис. 24)

Таблица 12

Динамика заболеваемости в показателях наглядности (в % к 1987 году)

Заболевания	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Злокачественные новообразования	0,0	−1,2	0,1	0,8	0,2	6,2	9,9	7,6	8,3	15,7
Острая гонорея	0,0	10,5	22,9	10,5	23,8	83,9	160,7	103,3	54,5	6,2
Активный туберкулез	0,0	−3,5	−10,9	−19,9	0,4	4,3	34,8	63,7	57,4	68,0

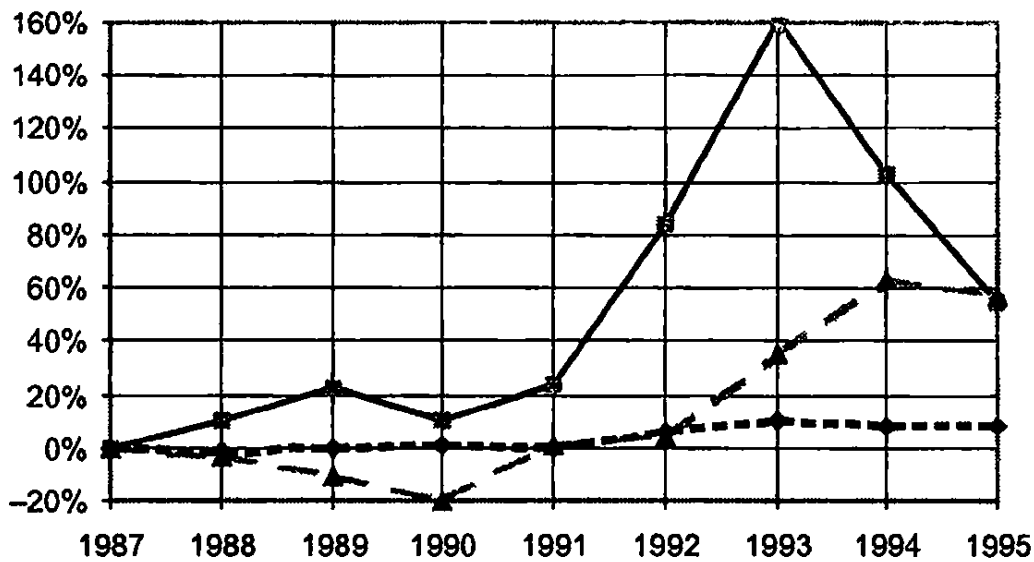


Рис. 24 **Динамика первичной заболеваемости жителей Санкт-Петербурга в показателях наглядности**

Столбиковые диаграммы представляют собой изображения различных величин в виде расположенных в высоту прямоугольников одинаковой толщины и разной высоты Построение столбиковой диаграммы требует только одной масштабнои шкалы,

которая задает высоту столбика Такие диаграммы применяются для отображения практически всех абсолютных и производных статистических показателей Исключения составляют экстенсивные показатели Для них более целесообразно использовать внутристолбиковые или секторные диаграммы Столбиковые диаграммы могут быть представлены в плоскостном варианте (рис 25, 26, 27) или объемном (рис 28)

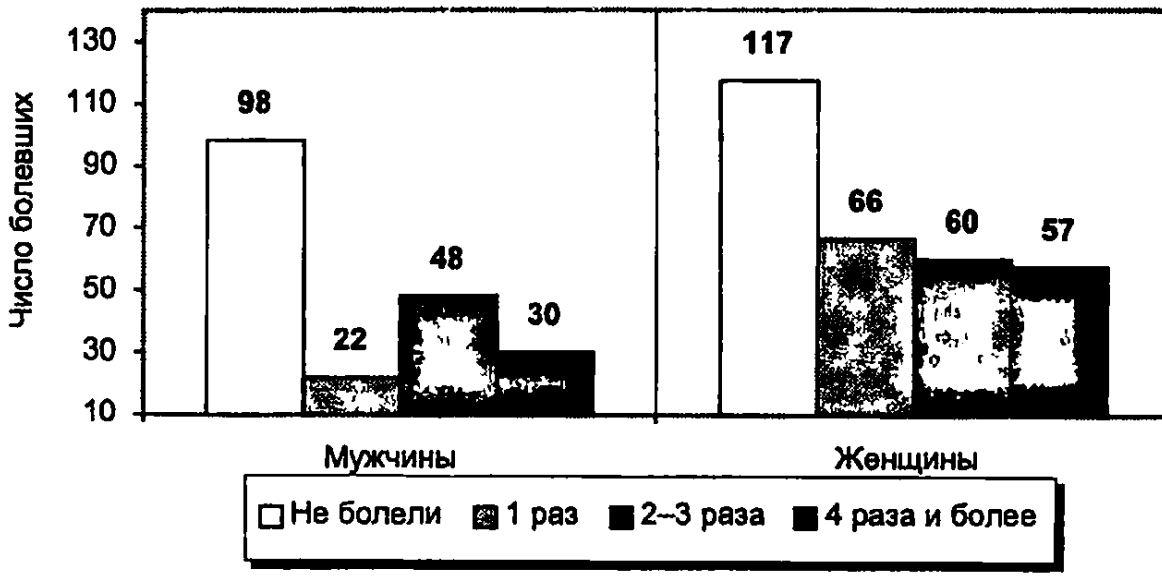


Рис 25 **Сравнительное распределение мужчин и женщин по кратности заболеваний в году**

Между столбиками диаграммы обязательно оставляются просветы Если столбики группируются по каким-либо признакам, то просвет целесообразно устанавливать между группами, а столбики в группах отображать слитно (см рис 25). Надписи в таких диаграммах располагают на столбиках Внутри столбиков надписи располагают только в случае приблизительного равенства высоты столбиков и отсутствия достаточного места над ними

Вместе с тем, главным критерием выбора той или иной диаграммы для отображения статистических показателей является наглядность и удобство анализа результатов *Например* если анализируется сравнительная заболеваемость мужчин и женщин, то более целесообразно представить попарно сгруппированные показатели мужчин и женщин (см рис 27).

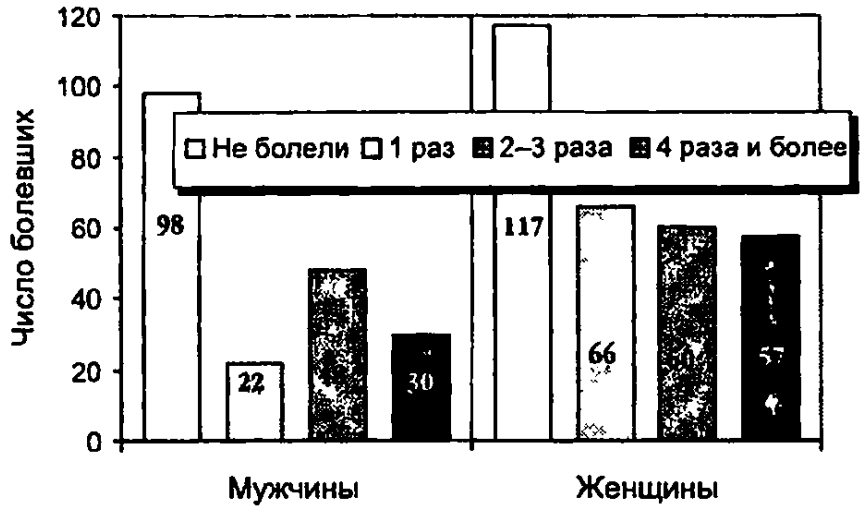


Рис. 26. **Сравнительное распределение мужчин и женщин по кратности заболеваний в году (менее удачный вариант экспликации)**

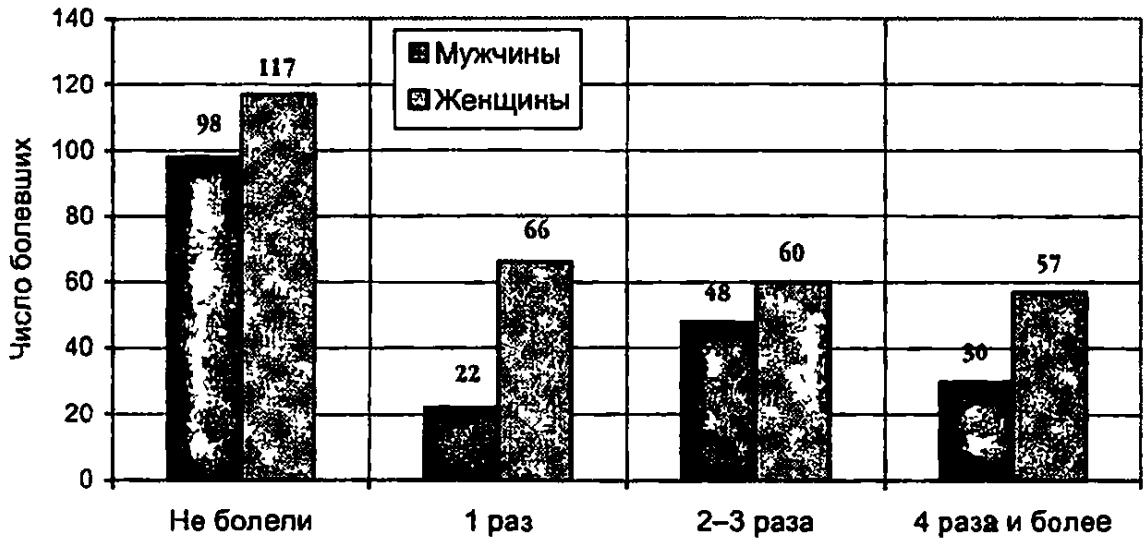


Рис 27. **Сравнительное распределение мужчин и женщин по кратности заболеваний в году**

Объемные диаграммы могут давать более насыщенное отображение показателей Однако их использование не гарантирует само по себе увеличения наглядности и ограничивает число данных, которые можно выводить на диаграмму. На приведенном примере (см рис 28) видно, что наглядность диаграммы не возросла Кроме того, на диаграмме явно необходимо поменять масштаб изображения

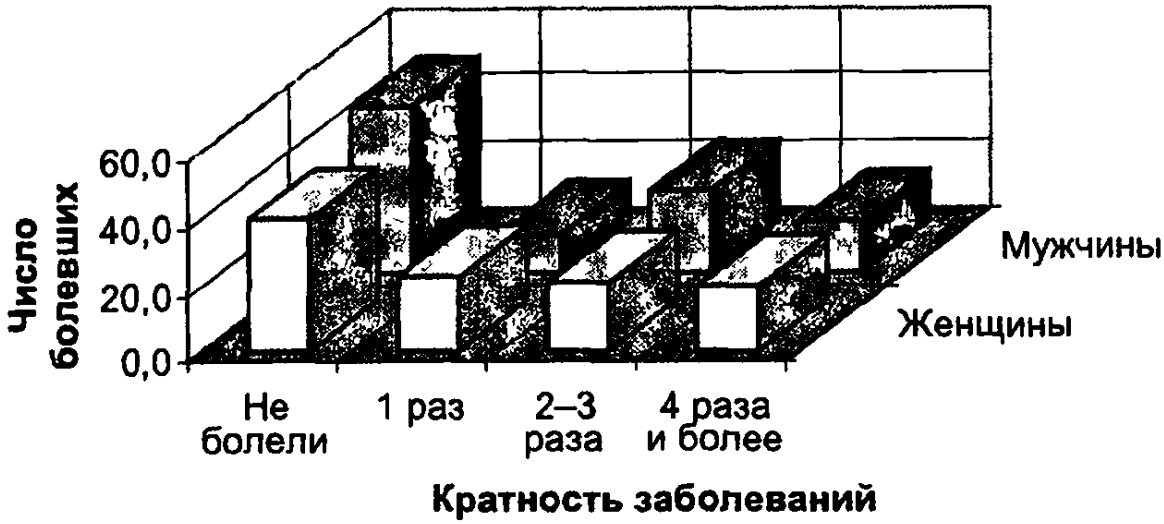


Рис 28. **Распределение мужчин и женщин по кратности заболеваний в году**

Для отображения экстенсивных показателей, характеризующих часть от целого (долю, удельный вес) в %, используются **внутристолбиковые** диаграммы При этом вся площадь столбика принимается за 100% Затем из этой площади выделяются части, размеры которых пропорциональны величинам показателей (рис 29)

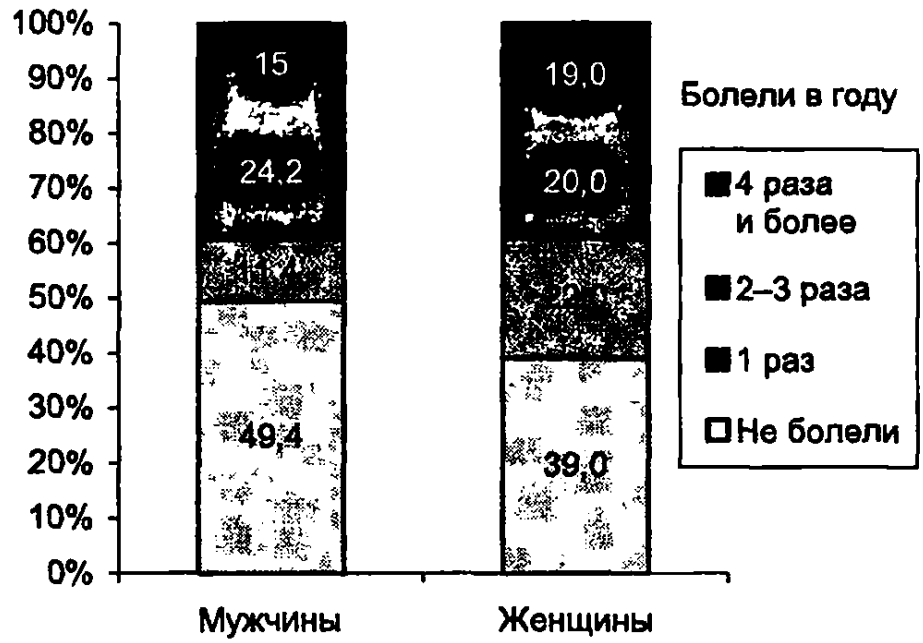


Рис 29 **Распределение мужчин и женщин по кратности заболеваний в году**

В ряде случаев для отображения экстенсивных показателей вместо внутристолбиковой может использоваться более простая столбиковая диаграмма Выбор конкретной формы представления данных определяется особенностями отображаемых числовых рядов и изобразительными возможностями графических редакторов (рис 30, 31)

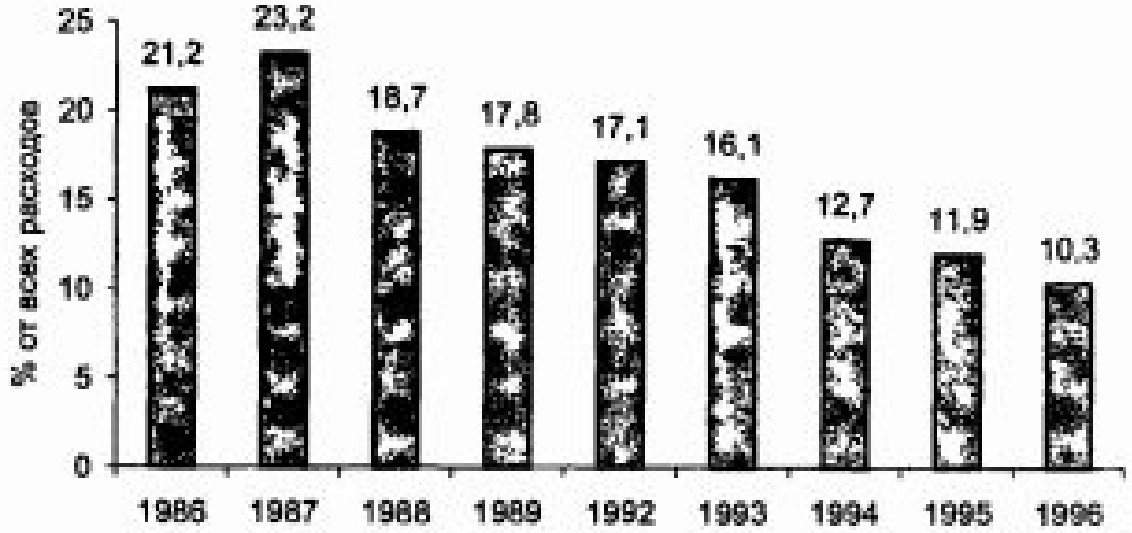


Рис. 30. **Доля затрат на здравоохранение в структуре всех затрат бюджета Санкт-Петербурга (плоскостная диаграмма)**

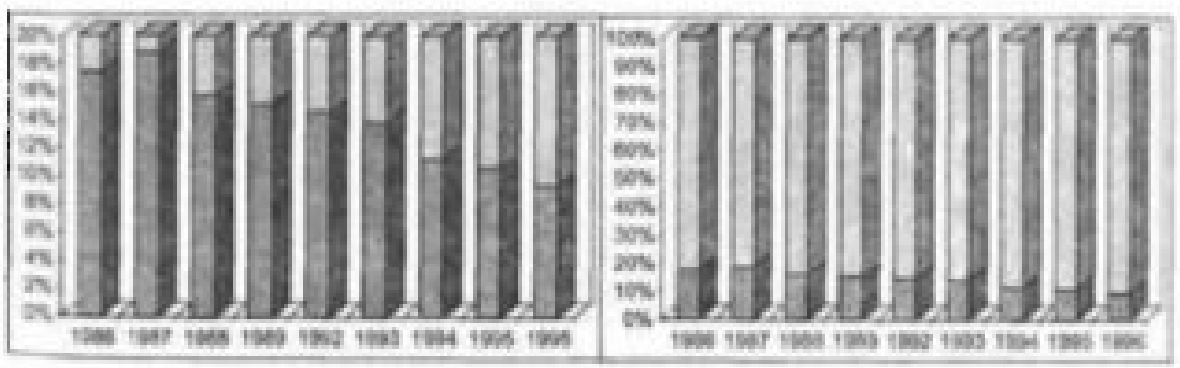


Рис 31. **Доля затрат на здравоохранение в структуре всех затрат бюджета Санкт-Петербурга (варианты объемных диаграмм)**

Одним из вариантов столбиковых диаграмм являются ленточные диаграммы Практически они представляют столбиковую диаграмму, «уложенную на бок» (рис 32)

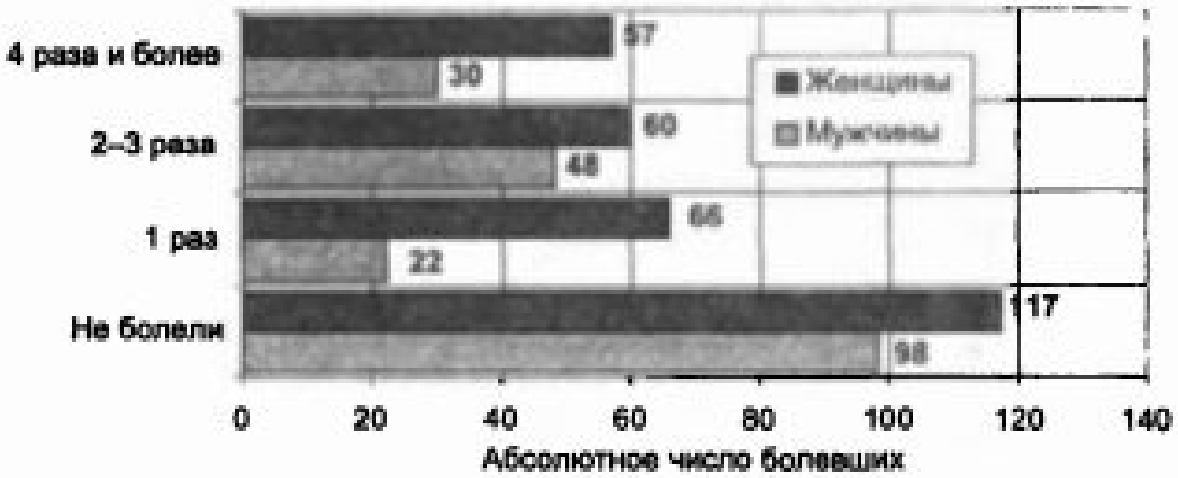


Рис 32 **Сравнительное распределение мужчин и женщин по кратности заболеваний в году**

Для отображения экстенсивных показателей по аналогии с внутристолбиковыми могут использоваться внутриленточные диаграммы (рис. 33)

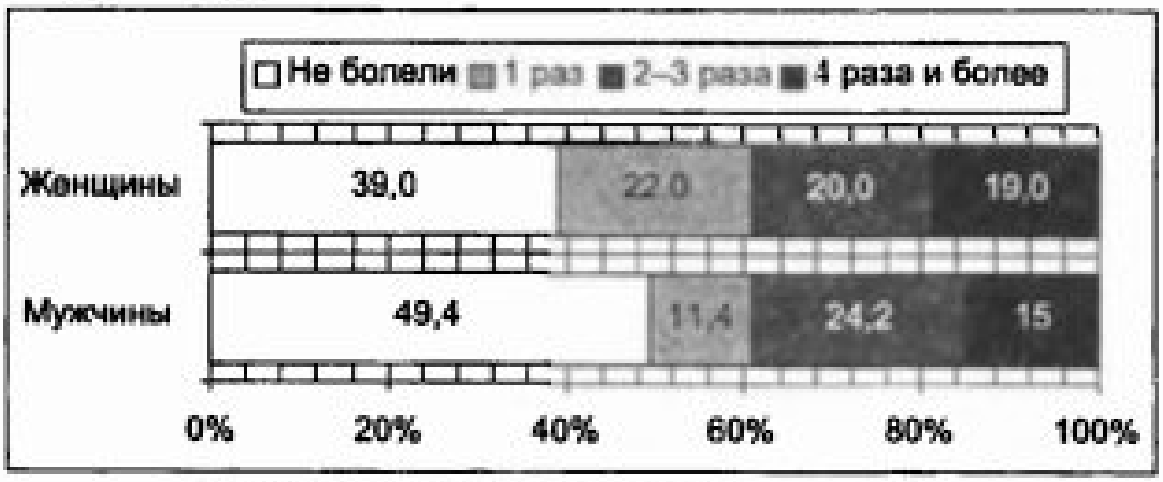


Рис. 33 **Сравнительное распределение мужчин и женщин по кратности заболеваний (Внутриленточная диаграмма)**

Наибольшая компактность обеспечивается одной из разновидностей ленточной — пирамидальной диаграммой Благодаря компактности изображения с ее помощью удается отобразить в диаграмме большое количество исходных данных Наиболее удобно применение пирамидальной диаграммы для отображения сложных динамических рядов и демографических данных (рис 34)

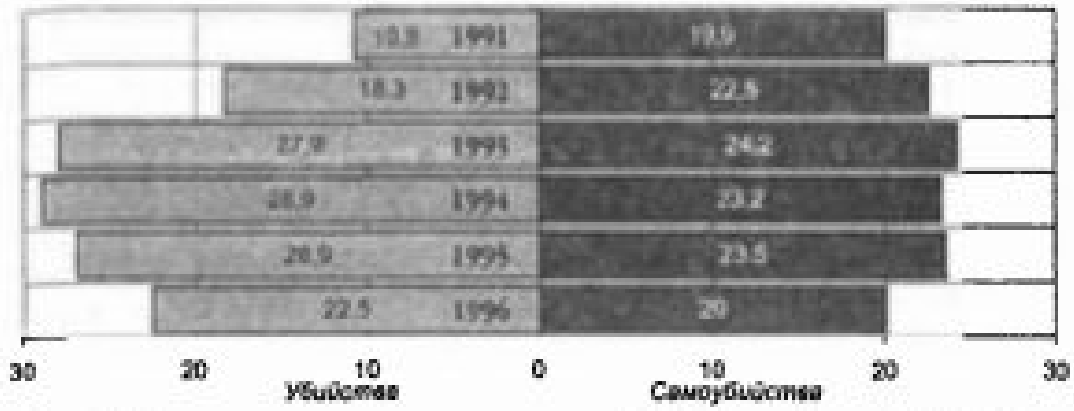


Рис 34. **Некоторые причины смерти жителей С -Петербурга в 1991—96 гг (случаев на 100 тыс населения)**

Более насыщенное трехмерное изображение менее компактно и занимает, соответственно, бóльшую площадь, поэтому требуется другой масштаб изображения (рис 35)

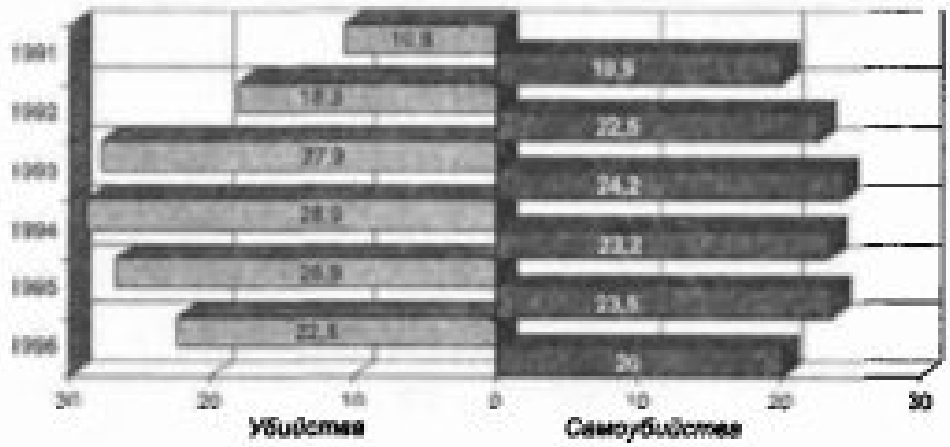


Рис. 35. **Некоторые причины смерти жителей С.-Петербурга в 1991—96 гг. (случаев на 100 тыс. населения)**

2.6.2. Специальные диаграммы

Приемы их построения и принципы анализа отображаемых результатов имеют свои жесткие правила Любое отступление от них может привести к серьезным методическим погрешностям К таким диаграммам относятся гистограмма, полигон распределения, огива, кумулята (кумулирующая кривая), радиальная диаграмма сезонности

При построении **радиальной диаграммы сезонности:**

1 Формируется окружность, радиус которой принимается за 100%.

2. Далее, по длине каждого из 12 радиусов (по числу месяцев в году) от центра к периферии на радиусах, последовательно по ходу часовой стрелки от «12 часов», откладываются значения сезонных показателей. (Подробнее расчет показателей сезонности смотри в разделе «Динамические ряды».) При превышении среднегодовых показателей кривая выходит за пределы окружности. В случае более низких уровней помесячных показателей — кривая оказывается внутри пределов окружности. При правильном построении диаграммы площадь поверхности, которая выделена кривой, равна площади круга (табл. 13, рис 36)

Таблица 13

Показатели частоты кишечных инфекций в году

Месяц	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
%	81,5	77,9	81,5	84,2	107,4	118,6	118,5	122,2	114,8	103,7	95,6	92,6

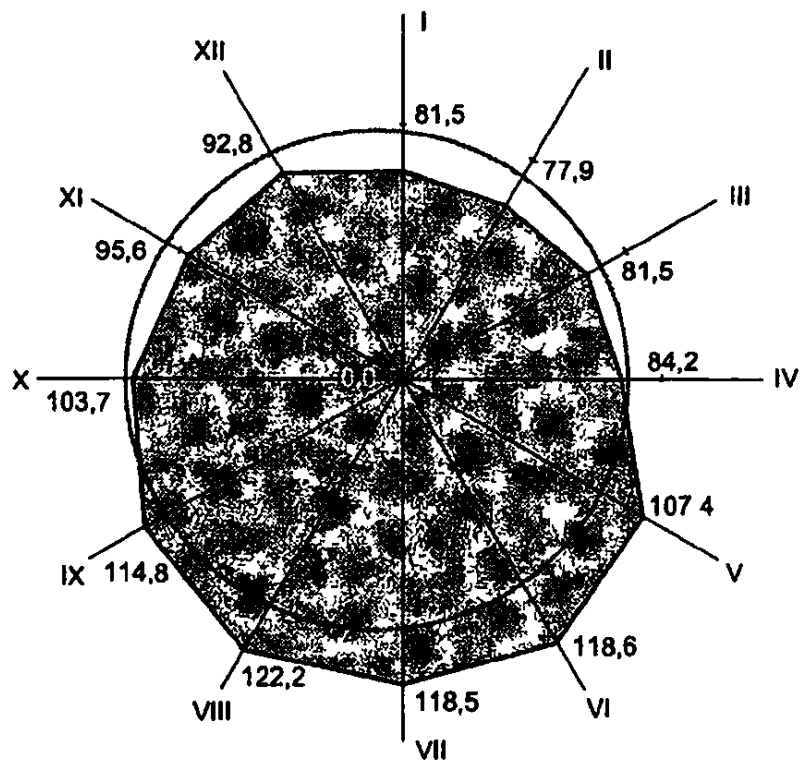


Рис. 36. **Диаграмма сезонности острых кишечных инфекций**

Гистограмма применяется для графического изображения интервальных рядов распределения Она позволяет качественно оценить различные распределения Например, на ней можно увидеть, что распределение бимодально (имеет 2 пика), что обычно является причиной неоднородностей выборки и г п

Внешне гистограмма представляет собой многоугольник, построенный с помощью смежных четырехугольников Ширина основания каждого четырехугольника соответствует границам группы вариант, определяемых интервалом Когда интервалы во всех группах равны, то ширина столбиков принимается одинаковой Если интервалы распределения не равны между собой, то ширина столбиков выбирается пропорционально границам конкретной группы вариант Высота столбика определяется частотой группы. *Например* в интервале от 165 до 170 см находится 8 вариант При построении графического изображения гистограмм между отдельными четырехугольниками, образующими гистограмму, не должно быть интервалов (рис 37)

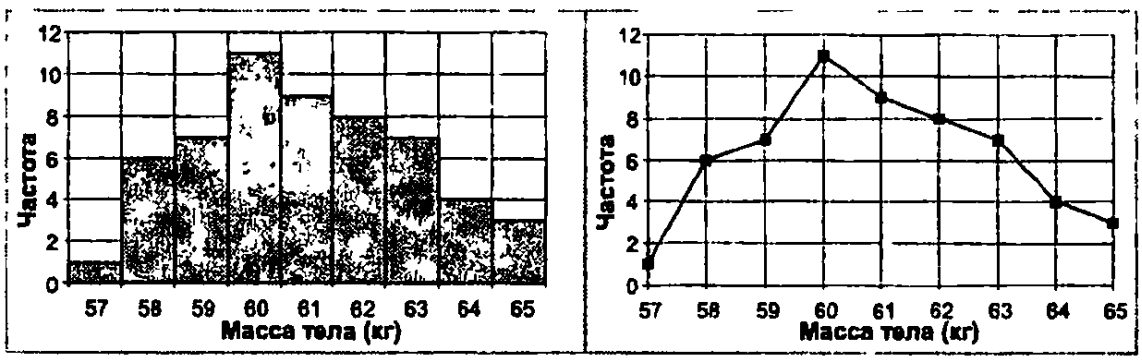


Рис 37. **Гистограмма и полигон распределения**

Для графического изображения дискретных рядов обычно применяются **полигоны распределения.**

В некоторых случаях при статистическом анализе распределений используется ряд накопленных частот — кумулятивный ряд, который отображается **кумулятой** Она позволяет наглядно показать число случаев ниже или выше определенного уровня Наглядность кумуляты повышается если представить накопленные частоты в виде *частостей*, т е долей от общего числа наблюдений *Например* лиц с массой тела 51 кг было 60% от числа обследованных (рис 38)

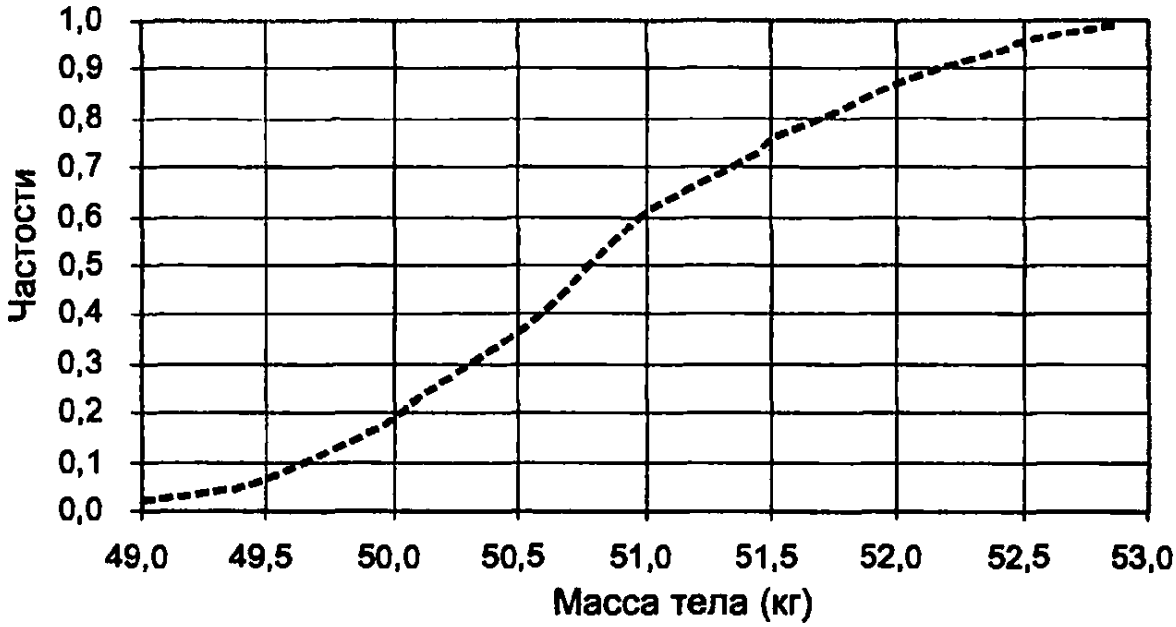


Рис. 38. **Вариант графического изображения кумуляты распределения**

Кумуляту целесообразно использовать для наглядного представления различий статистических распределений (рис. 39).

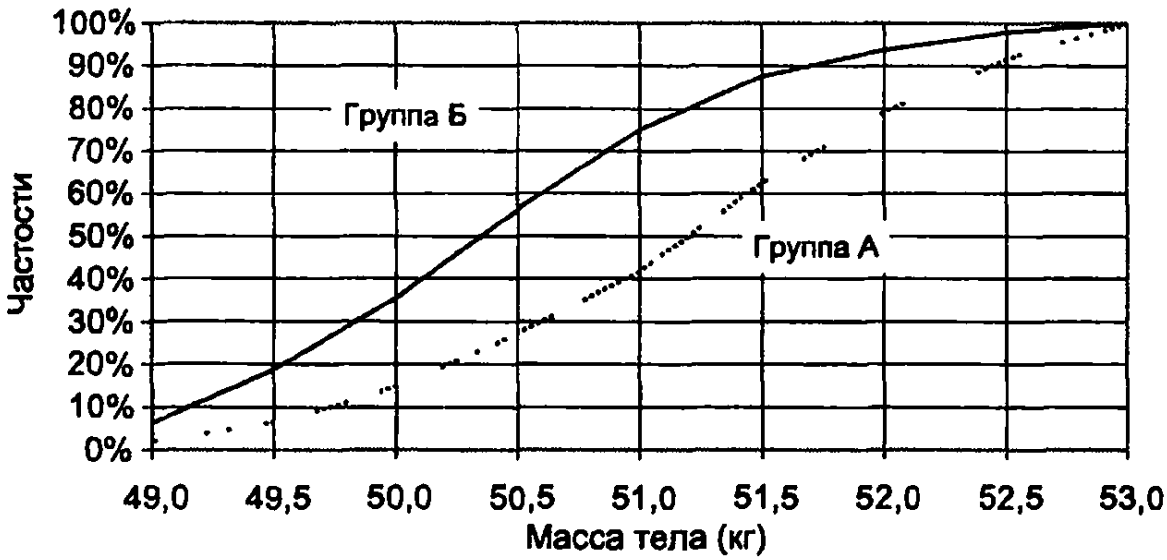


Рис. 39. **Сравнительное изображение кумулят распределения**

Приведенный пример кумуляты весьма близко подходит к **огиве** (рис 40) Различия состоят только в том, что огива дает значение варьирующего признака у каждой ранжированной единицы статистического наблюдения Кумулята дает по существу аналогичную диаграмму, но только в отношении групп наблюдения *Например* огива дает представление о массе тела каждого подростка из группы, ранжированной в порядке возрастания значений 48,7 49,3 49,5 49,5 49,5 49,5 49,5 49,7 49,7 50,0 50,0 50,0 50,0 50,1 50,2 50,2 50,4 50,5 50,5 50,5 51,0 51,0 51,0 51,0 51,0 51,0 51,0 51,1 51,2 51,2 51,5 51,6 51,9 52,2 52,3 52,5 52,8 53,1 53,7 54,5 кг

Рис 40. **Огива распределения**

Специальная плоскостная диаграмма, в виде прямоугольника, с помощью которой можно одновременно отображать три величины, называется **Знак Варзара** по фамилии русского статистика В Е. Варзара. Одна отображаемая величина — основание прямоугольника (a), другая — его высота (h), третья — площадь прямоугольника (S), произведение $S=ah$

2.6.3. Построение диаграммы в *MS Excel*

Диаграммы в *Excel* создаются на основе данных, уже введенных в таблицу. В качестве примера рассмотрим этапы создания линейной диаграммы, отражающей динамику заболеваемости населения города **Н** пневмонией в период с 1987 по 1992 гг. Для того чтобы приступить к созданию диаграммы, необходимо подготовить исходные данные на **Листе1** (рис. 41)

	A	B	C	D	E	F
1	1987	1988	1989	1990	1991	1992
2	6,1	5,7	4,7	4,2	5,3	6,8

Рис. 41 **Исходные данные для построения диаграммы**

Для создания диаграммы вызовите *Мастер диаграмм* С этой целью выполните команду <**Диаграмма**> из меню <**Вставка**> или нажмите кнопку [**«Мастер диаграмм»**], которая обычно располагается на стандартной панели инструментов 

После активизации *Мастера диаграмм* на экране появляется первое диалоговое окно, в котором необходимо выбрать тип диаграммы.

В этом окне (рис. 42) имеется два списка (закладки), в которых перечислены возможные варианты диаграмм **Стандартные** и **Нестандартные.** Из списка **Стандартные** с помощью указателя мыши выберите график, отображающий развитие процесса во времени или по категориям Затем нажмите расположенную внизу окна кнопку [**Далее**]

Когда откроется следующее окно, перейдите на вкладку ряд, где нажмите на кнопку [**Добавить**] под полем **Ряд** Затем в поле **Значения** введите диапазон ячеек с исходными данными: =Лист1!$A$2 $F$2, а в поле **Подписи оси X** — диапазон ячеек: =Лист1!$A$1:$F$1 (рис. 43)

Примечание: для ввода диапазона ячеек щелкните мышкой кнопку свертки (с маленькой красной стрелкой) в правой части поля, затем выделите нужные ячейки и нажмите клавишу [**Enter**]

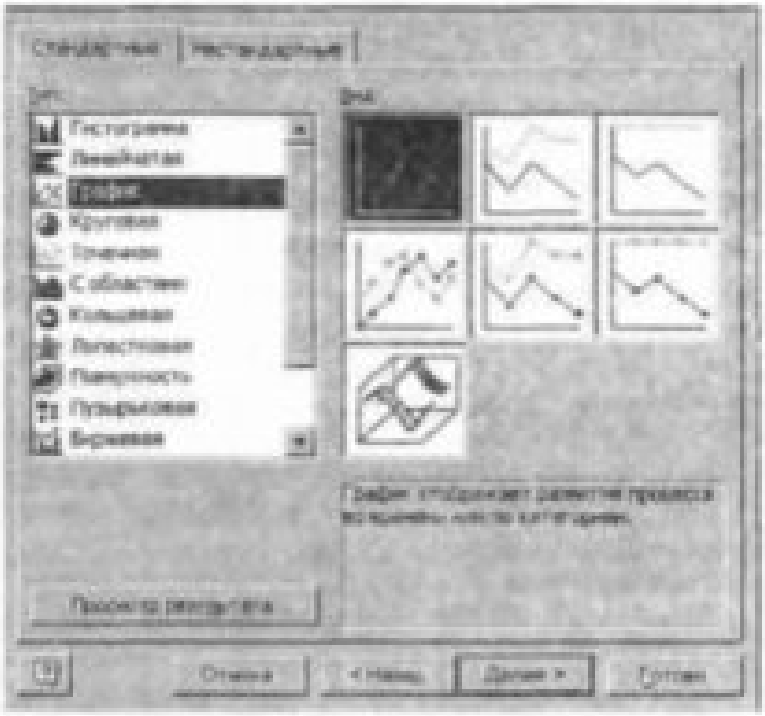


Рис. 42. Первое окно мастера диаграмм

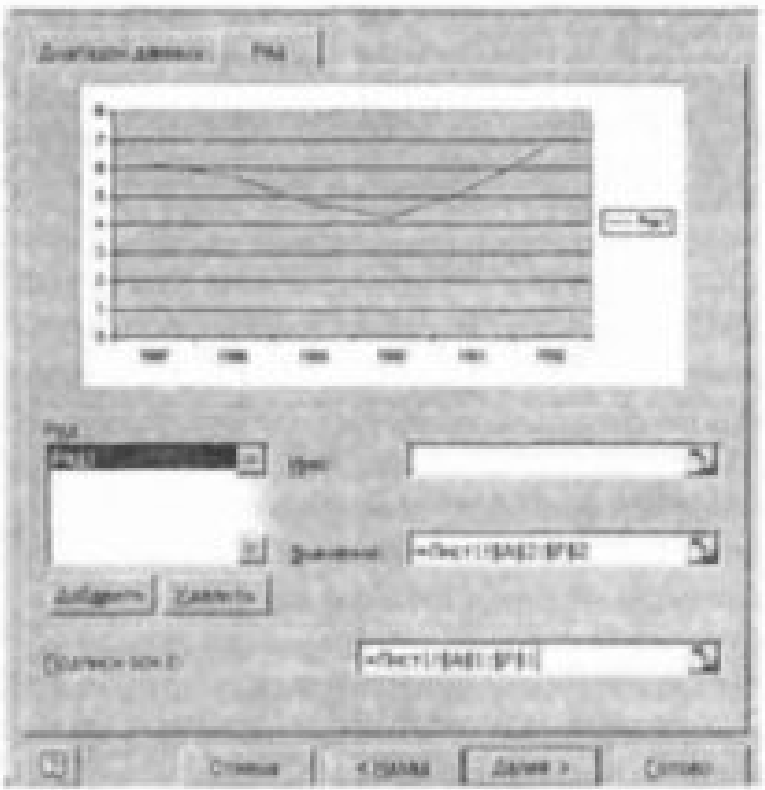


Рис. 43. Второе окно мастера диаграмм

Следующий шаг позволяет указать название диаграммы и подписей к осям, выбрать параметры осей и линий сетки, а также добавить легенду (подписи полученных линий или столбцов распределений) и включить в диаграмму в качестве составной части таблицу исходных данных (рис. 44) В случае необходимости с помощью кнопки **[Назад]** можно вернуться к предыдущим шагам мастера диаграмм

Рис. 44. **Третье окно мастера диаграмм**

В последнем диалоговом окне мастера диаграмм появляется запрос где должна располагаться созданная диаграмма? Ее можно разместить на отдельном листе, уже существующем (используемом для исходных данных) листе. Для завершения создания диаграммы необходимо нажать кнопку **[Готово]**

Следует отметить, что внешний вид полученной диаграммы может Вас не удовлетворять Чтобы получить диаграмму в виде, максимально приближенном к Вашим запросам, после создания диаграммы можно изменять ее экспликацию (гип, текст названий и подписей, их шрифт, числовой формат и т п) Для этого используйте следующие возможности редактирования диаграмм.

1 Чтобы пропорционально изменить размеры диаграммы, установите указатель мыши на углу рамки диаграммы (если рамки нет — вызовите ее щелчком левой кнопки мыши по краю диаграммы), нажмите кнопку мыши и не отпуская ее перетащите на необходимое расстояние Аналогично изменяются продольные и поперечные размеры диаграммы

2. Чтобы изменить внешний вид диаграммы, щелкните левой кнопкой мыши по центру диаграммы, а затем еще раз щелкните, но уже правой кнопкой мыши. На экране появится перечень команд для правки диаграммы

В качестве примера рассмотрим этапы построения диаграммы сезонности острых кишечных инфекций. Вначале введите исходные данные в электронную таблицу (рис. 45)

	A	B	C	D	E	F	G	H	I	J	K	L
1	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
2	81,5	77,9	81,5	64,2	107,4	118,6	118,5	122,2	114,8	103,7	92,6	96,6

***Рис 45.* Исходные данные для построения диаграммы сезонности**

Для создания диаграммы выполните команду <**Диаграмма**> из меню <**Вставка**>. После активизации «Мастера диаграмм» на экране появляется первое диалоговое окно (см рис 42), в котором необходимо выбрать тип диаграммы Выберите из списка **Тип Лепестковая**, затем нажмите на кнопку [**Далее>**]

Когда откроется окно Шаг 2 из 4, перейдем на вкладку **Ряд**, где нажмем на кнопку [**Добавить под полем Ряд**]. Затем в поле **Значения** введите диапазон ячеек $A$2:$L$2, а в поле **Подписи оси X** — диапазон ячеек **$A$1:$L$1**.

В следующем диалоговом окне «Мастера диаграмм» параметры диаграммы на вкладке **Заголовки** введите название диаграммы *Показатели сезонности острых кишечных инфекций*. Затем нужно сбросить флажок **Добавить легенду** на вкладке **Легенда** и все флажки на вкладке **Линии сетки** (рис. 46)

В последнем диалоговом окне «Мастера диаграмм» (шаг 4 из 4) выбираем переключатель **Отдельном** и для завершения создания диаграммы на отдельном листе нажимаем кнопку [**Готово**].

Рис. 46. «Мастер диаграмм»: параметры диаграммы

3. Источники статистической информации о здоровье

В медицинской статистике довольно часто встречаются статистические выкладки, получение и интерпретация которых имеют ряд особенностей. На эти особенности следует обратить внимание, поскольку они имеют большое значение для правильного понимания статистических данных Учитывая, что неумелое, а подчас просто безграмотное пользование такими данными получает определенное распространение в отечественных и переводных публикациях и, кроме того, нередко в официальных изданиях, следует более детально рассмотреть особенности получения статистической информации о здоровье населения и его современных тенденциях

В этой связи следует особо отметить, что никаких принципиальных отличий в методиках использования и интерпретации статистических коэффициентов (как и большинства других статистических методов и отдельных статистических характеристик) в отечественной и зарубежной мед. статистике нет Вместе с тем, отечественная статистика имеет свой богатый практический опыт проведения широких, комплексных социально-гигиенических, эпидемиологических и т п исследований здоровья больших групп населения.

Такие исследования за рубежом последние 50—40 лет вообще не проводились или проводились в весьма ограниченном объеме Поэтому мы настоятельно рекомендуем прежде, чем обращаться к зарубежным авторам, ознакомиться с имеющимся опытом статистических исследований в России или в прошлом СССР. Это избавит от повторения тех ошибок, которые давно пройдены отечественной статистикой, и облегчит исследователю понимание описаний статистических методик, предлагаемых зарубежными авторами Несмотря на запутанную терминологию, осложненную нередко качеством перевода, эти методики чаще всего не несут ничего принципиально нового или декларируют весьма спорные вещи *Например*: в современной зарубежной литературе почти повсеместно при описании распространенности тех или иных заболеваний или при описании с помощью относительных величин показателей здоровья используется термин **популяция** Однако изначально этот термин начал применяться в биологии, где популяцией, строго говоря, называют совокупность особей одного вида, занимающую одну территорию и воспроизводящую себя в нескольких поколениях. Таким образом, перенос этих дефиниций на совокупности людей явно не вполне корректен Поэтому следует избегать употребления этого термина в медицинской статистике

Среди статистических показателей здоровья населения особое место занимает информация о заболеваемости.

3.1. Заболеваемость

Статистические показатели заболеваемости (*health indicator* — англ) позволяют оценить частоту встречаемости и характер заболеваний, закономерности их распространения среди населения Это необходимо для выявления факторов, отрицательно влияющих на здоровье и, наоборот, способствующих укреплению здоровья населения На основе динамики показателей заболеваемости планируется и оценивается эффективность деятельности медико-социальных учреждений, результативность лечебно-профилактических мероприятий, экологические ситуации Характеристики заболеваемости косвенно отражают и уровень социально-экономического благополучия населения в целом.

В государственной статистике России для регистрации первичных данных о заболеваемости используют большое количество различных специальных учетных документов медицинские карты, экстренные извещения, листки нетрудоспособности, карты выбывших из стационара, врачебные свидетельства о смерти и другие учетные и отчетные статистические формы При проведении научно-исследовательских работ, как правило, используются особые, специально создаваемые для каждого исследования учетные формы

Важнейшим, ключевым звеном любого статистического анализа является сводка и группировка Для того чтобы данные о заболеваемости независимо от места получения (разные регионы России или другие государства), а также характера происхождения (официальная статистика или научные разработки) были сопоставимы, необходима единая методическая основа статистической сводки и группировки показателей Такой общепринятой основой является «Международная статистическая классификация болезней и причин смерти»

Международная статистическая классификация (номенклатура) болезней (МСКБ) — это система группировки болезней и патологических состояний, отражающая современный этап развития медицинской науки МСКБ является общепринятым нор-

мативным документом при изучении здоровья населения в странах — членах Всемирной Организации Здравоохранения (ВОЗ)

Структура МСКБ создана на основе принципов, предложенных У Фарром и М д'Эспином более 100 лет тому назад Эта структура, охватывая все известные болезни и патологические состояния, обеспечивает их четкую систематизацию в перечне рубрик Болезни, имеющие особую значимость из-за высокой распространенности, большой эпидемиологической опасности, высокой летальности и т п , представлены отдельными рубриками Первая классификация для практического использования была подготовлена Ж Бертильоном и утверждена на международной конференции в Париже в 1900 году

По мере развития медицинской науки, изменения теоретических воззрений на патологический процесс, а также с учетом приобретаемого практического опыта МСКБ периодически уточняется и дополняется примерно раз в 10 лет

В России до 1918 года существовали одновременно две собственные классификации: официальная номенклатура врачебного департамента Министерства внутренних дел, применявшаяся для официальной отчетности, и классификация, утвержденная в 1889 году Пироговским съездом врачей. С 1918 года в России Пироговская классификация была введена повсеместно В 1924 году утвердилась первая советская номенклатура болезней и причин смерти, близкая к действующей тогда международной классификации А с 1930-го в стране действовала собственная классификация, просуществовавшая с небольшими изменениями по 1970 год С 1970 года в СССР и, соответственно, в России начала использоваться международная классификация (сначала 8-го, затем 9-го пересмотра). С 1999 года в России поэтапно вводится классификация 10-го пересмотра

Статистический анализ любого вида заболеваемости обычно проводится с помощью общих и специальных показателей, изменение которых оценивается в динамике в сравнении с предыдущими годами, кварталами, месяцами (*person-time rate* — англ) Показатель многолетней, длительной динамики (тренда) изменения заболеваемости в зарубежной статистике называется *секулярный тренд* (*secular trend* — англ)

Общие показатели заболеваемости дают представление об уровне, распространенности, частоте всей совокупности или отдельных нозологических групп заболеваний, зарегистрированных за определенный период времени и рассчитанных на 100 тыс (‰) или 10 тыс (‰) населения Вычисляется также показатель структуры заболеваемости (доля в процентах того или иного заболевания среди всех заболеваний).

Специальные показатели характеризуют уровень и структуру заболеваемости по определенным нозологическим формам, а также различным возрастным, половым или социальным группам Такой подход позволяет установить причины повышенной заболеваемости

В зависимости от структуры общей заболеваемости населения региона или страны выделяют следующие типы заболеваемости:

- эпидемический тип,
- переходный тип;
- неэпидемический тип

Эпидемический тип заболеваемости характерен для слаборазвитых, экономически отсталых стран, имеющих слаборазвитую систему медико-социальной помощи и низкую санитарную культуру населения Этот тип заболеваемости характеризуется высоким уровнем инфекционных болезней, занимающих ведущее место среди всех остальных заболеваний и являющихся одной из главных причин высокой смертности населения

Неэпидемический тип заболеваемости характерен для экономически развитых стран, имеющих высокоразвитую систему оказания медико-социальной помощи Характеризуется относительно высокой долей среди всех заболеваний болезней сердечно-сосудистой системы, онкологической патологии, нервно-психических расстройств, травм, заболеваний органов дыхания

Переходный тип заболеваемости представляет промежуточный вариант между эпидемическим и неэпидемическим типом заболеваемости Характерен для развивающихся стран или стран переживающих затяжные социальные катаклизмы

В статистике заболеваемости применяются различные методики получения исходных данных, оказывающих большое влияние на содержательную сущность данных и полноту выявления

истинной распространенности болезней Большинство показателей заболеваемости, в зависимости от метода получения исходной информации, можно условно разбить на две группы показатели, полученные путем регистрации обращений за медицинской помощью, и показатели, полученные путем активного выявления патологии на медицинских осмотрах

Заболеваемость по обращаемости — основа официальной статистики здоровья населения и здравоохранения в России В зарубежной статистике эта группа показателей применяется редко Это связано с тем, что исходные данные о случаях заболеваний по обращаемости формируются при обращении за медицинской помощью, их полнота и достоверность зависят от доступности медицинской помощи, ее качества (радиус обслуживания, наличие врачей-специалистов и т п.) Отсюда вытекает важное обстоятельство проводить сравнение показателей заболеваемости по обращаемости, полученных из разных стран или разных регионов одной страны, имеющих существенные различия в организации здравоохранения и доступности медицинской помощи, следует весьма осторожно.

Затруднения, с которыми приходится сталкиваться при анализе заболеваемости по обращаемости в России, обычно связаны с неполной регистрацией случаев заболеваний по причине плохо налаженной работы соответствующих местных управленческих и организационных структур здравоохранения, из-за желания приукрасить истинное положение вещей и т д Большое значение имеют и другие местные особенности организации медицинской помощи Например, в сельской местности, где как правило нет возможности получить высококвалифицированную медицинскую помощь, определенную роль играет близость крупных мегаполисов с развитой сетью специализированных медучреждений Чем ближе крупный город, тем чаще сельские жители обращаются в городские медицинские центры и их обращения регистрируются не по месту жительства, а в городских учреждениях Это обстоятельство во многом объясняет тот факт, что в подавляющем большинстве сельских районов России статистические показатели заболеваемости по обращаемости ниже, чем в городах, в то время как истинная заболевае-

мость жителей села, устанавливаемая углубленными медицинскими осмотрами, выше, чем в городах

Значительно уменьшает достоверность официальной статистики по обращаемости прямое желание больных скрыть случаи обращений за медицинской помощью (венерические болезни, случаи производственного травматизма), желание частнопрактикующих врачей укрыть от налоговых органов объем оказываемых ими платных услуг и т п Именно поэтому в странах, где медицинская помощь (в первую очередь амбулаторно-поликлиническая помощь) предоставляется чаще всего в негосударственных учреждениях и частнопрактикующими врачами, статистика обращаемости либо отсутствует вовсе, либо имеет весьма сомнительную достоверность

Вместе с тем, данные, полученные при самом тщательном соблюдении всех правил регистрации случаев обращений за медицинской помощью, нельзя считать полным отражением истинного распространения болезней среди населения Например, по поводу хронического заболевания человек может не обращаться в течение нескольких лет Поэтому данные, полученные по обращаемости, могут характеризовать только общие закономерности распространения патологии Тем не менее, статистический анализ обращаемости сохраняет в условиях России свое значение для оценки здоровья населения, решения конкретных задач планирования и оценки результатов деятельности системы медико-социальной помощи

Группа статистических показателей по обращаемости представлена следующими вариантами

Собственно заболеваемость (первичная заболеваемость, инцидентность — incidence rate) — частота ни где ранее не зарегистрированных и впервые в данном году выявленных заболеваний среди населения обратившегося за медицинской помощью

Болезненность (накопленная заболеваемость, превалентность — prevalence англ) — это частота всех имеющихся среди населения заболеваний, как впервые в данном году выявленных, так и зарегистрированных в предыдущие годы, по поводу которых больной вновь обратился за помощью в данном году

В зарубежной статистике заболеваемости, особенно статистике инфекционных болезней, большое распространение получил специальный показатель называемый ***превалентность периода*** (*period prevalence* — англ.) — общее число случаев какого-либо (обычно длительно текущего хронического) заболевания, выявленного или имеющегося в группе людей в течение определенного периода времени Например, общее число всех случаев туберкулеза в период с 1 октября 1999 г по 31 сентября 2000 г Частным случаем этого показателя является *моментальная превалентность* (*point prevalence* — англ) — общее число случаев какого-либо (обычно длительно текущего хронического) заболевания, выявленного (имеющегося) на конкретный момент времени Например, общее число всех случаев туберкулеза по состоянию на 1 октября 1999 года

В зависимости от вида конкретного источника получения информации и характера регистрируемых заболеваний выделяют и другие группы показателей.

Обращаемость за медицинской помощью — число больных, впервые в году обратившихся за медицинской помощью по данному заболеванию При этом хронические заболевания могут быть зарегистрированы в году только один раз, а острые регистрируются при каждом их новом возникновении (регистрируется только первое, первичное обращение в случае заболевания) Все первичные и повторные посещения врача по поводу заболеваний в году характеризуют *посещаемость* На одно обращение в городе обычно приходится 3—4 посещения В сельской местности 1—2 посещения

Заболеваемость по данным обращаемости в поликлинику Основная составляющая показателя общей заболеваемости по обращаемости. Учитывая, что 70—80% больных начинают и заканчивают лечение в амбулаторно-поликлинических учреждениях*, при высокой доступности этого вида медицинской помо-

* Следует отметить, что с финансированием прямо противопопожная ситуация на амбулаторно-поликлиническую помощь в России в 1999 году тратилось 20–26% всех средств, выделяемых на здравоохранение

щи, показатель может давать достаточно объективную информацию о распространенности болезней

Источником информации служит медицинская карта (учетная форма № 25/у), статистический талон для регистрации уточненных, заключительных диагнозов

Госпитализированная заболеваемость. Источники информации — истории болезни или карты выбывших из стационара (учетная форма № 266/у), которые заполняются на каждый случай госпитализации Этот вид статистики заболеваемости характеризует совокупность болезней, послуживших причиной госпитализации в стационар Показатель используется в основном для анализа эффективности работы стационарных учреждений (качество и своевременность диагностики, длительность лечения, частота осложнений, больничная летальность и т. п) При оценке госпитализированной заболеваемости используются общие коэффициенты частота госпитализированной заболеваемости (число госпитализированных на 1000 населения), охват госпитализацией (число госпитализированных на 100 больных). В целом, по России из каждой тысячи человек в 1999 году 228 проходили стационарное лечение (т е больше, чем каждый четвертый) Средний срок госпитализации составил 18 дней

В качестве специальных рассчитываются показатели госпитализированной заболеваемости по отдельным возрастным и половым группам, по отдельным нозологическим формам и профилям медицинских учреждений, тяжести заболевания, длительности и исходов лечения

Среди всех показателей по обращаемости госпитализированная заболеваемость как критерий оценки здоровья населения представляет наименьший интерес по нескольким причинам. Одна из них — неполная регистрация случаев заболеваний, так как большинство больных начинают и заканчивают свое лечение в амбулаторно-поликлинических учреждениях

В отличие от обращаемости, *активное выявление заболеваний* дает наиболее объективную картину распространения болезней Показатели заболеваемости, полученные путем активного выявления, или ***заболеваемость по данным медицинских осмотров (патологическая пораженность)*** — совокупность болезней и патоло-

гических состояний, выявленных при профилактических осмотрах Существенным достоинством такого способа получения информации является возможность учета заболеваний в ранних стадиях или преморбидных состояний Однако получение таких данных связано с крупными материальными затратами из-за высокой стоимости труда врачей-специалистов, применяемого лабораторно-диагностического оборудования и т п. По этим причинам в широкой практике эти показатели используются крайне редко Кроме того, данные о заболеваемости, полученные путем активного выявления, сильно зависят от глубины и качества медицинского обследования (количества специалистов, их квалификации, набора инструментальных и лабораторных методик обследования и т п) К недостаткам показателей о заболеваемости, полученных путем активного выявления, следует отнести и недоучет острых заболеваний, поскольку во время медицинских осмотров регистрируются, в основном, заболевания, имеющиеся на момент освидетельствования

Регистрируются выявленные на медицинских осмотрах случаи заболеваний в «Карте подлежащего периодическому осмотру» (учетная форма № 046/у), «Медицинской карте амбулаторного больного» (учетная форма № 025/у), «Истории развития ребенка» (учетная форма № 112/у), «Медицинской карте ребенка» (учетная форма № 026/у), «Медицинской карте студента вуза» (учетная форма 025-3/у), контрольной карте диспансерного наблюдения (учетная форма № 30) В небольших лечебно-профилактических учреждениях при проведении целевых осмотров вместо специальных карточек могут составляться списки лиц, подлежащих целевому медицинскому осмотру.

Медицинские осмотры подразделяют на предварительные осмотры, периодические и целевые

Предварительные осмотры проводятся при поступлении на работу или учебу с целью определения соответствия состояния здоровья конкретного человека требованиям, предъявляемым условиями и характером труда, а также для выявления заболеваний, которые могут прогрессировать в процессе профессиональной деятельности или учебы

Основная задача ***периодических медицинских осмотров*** — выявление заболеваний или их ранних признаков, как этиологически связанных, так и не связанных с профессией, но при которых данная профессиональная деятельность представляет опасность для здоровья работника

Действующим в настоящее время законодательством выделяются несколько контингентов (так называемых «декретированных контингентов»), подлежащих обязательным предварительным и периодическим медицинским осмотрам.

- работники предприятий, учреждений и организаций, имеющие контакт с профессиональными вредностями;
- работники, чья профессиональная деятельность связана с постоянным риском для себя и других лиц, и поэтому к здоровью этих работников предъявляются повышенные требования (водители автотранспортных средств, работники железнодорожного, авиационного транспорта и т п);
- работники пищевых, детских и некоторых других коммунальных учреждений, а также торговли, которые проходят специальное обследование (на бактерионосительство, наличие кожных и венерических заболеваний и т п), поскольку они могут стать источником массового заражения;
- дети всех возрастных групп, рабочие-подростки, учащиеся средних специальных учебных заведений, студенты;
- контингенты, находящиеся под динамическим (диспансерным) наблюдением.

Целевые осмотры Это осмотры, проводимые с целью выявления конкретной патологии (гинекологической, онкологической и т п.) или с целью медицинского освидетельствования относительно небольших групп с высокой вероятностью возникновения одного или нескольких этиологических связанных заболеваний (групп повышенного риска).

Особый раздел статистики составляют показатели заболеваемости наиболее важными для общественного здоровья болезнями К ним относятся показатели заболеваемости по данным о причинах смерти, заболеваемости важнейшими неэпидемическими болезнями, инфекционной (эпидемической заболеваемости), заболеваемости с временной утратой трудоспособности,

заболеваемости со стойкой утратой трудоспособности (инвалидности) и др

Заболеваемость по данным о причинах смерти — совокупность заболеваний, зарегистрированных по данным о причинах смерти Источник информации в России врачебное свидетельство о смерти, учетная форма № 246 или фельдшерская справка о смерти, учетная форма № 246-ф В практике этот показатель чаще всего трактуется как показатель причин смертности населения (*proportionate mortality* — англ) и используется при анализе смертности населения Так как причиной смерти человека почти всегда являются заболевания (за исключением травм, отравлений, убийств и самоубийств), считается что показатель достаточно полно отражает состояние здоровья населения Этот показатель является одним из самых распространенных в государственной и международной статистике здоровья, поскольку на него в меньшей степени, чем на статистику заболеваемости по обращаемости, оказывает влияние уровень развития системы здравоохранения

Основным недостатком данных о причинах смерти является их ретроспективность По этой причине их трудно использовать для оперативного вмешательства с целью устранения или снижения влияния негативных факторов на здоровье человека Кроме того, статистике причин смерти свойственна неполная регистрация патологических состояний Это связано с тем, что если после смерти человека были обнаружены несколько патологических состояний, то регистрации подлежит только одно состояние или его осложнение, послужившее причиной смерти, а остальные, которые по мнению врача-патологоанатома или эксперта, не играли существенной роли в наступлении смерти, не регистрируются Указанное обстоятельство особенно сильно сказывается в старших возрастных группах, т к у человека в этот период жизни обычно имеется не одно, а несколько заболеваний

Следует помнить, что на показатели распространенности какого-либо заболевания, полученные на основе изучения причин смерти, значительно влияет летальность этого заболевания (летальность — частота смерти от данного заболевания) Например, летальность таких распространенных заболеваний, как

острые респираторные инфекции, значительно ниже, чем летальность онкологических заболеваний Таким образом, соотношение показателей смертности от респираторных инфекций и от онкологических заболеваний явно не будет соответствовать истинной распространенности этих болезней В санитарной статистике считается, что в полной мере распространенность болезней может быть оценена только при комплексной оценке показателей заболеваемости по обращаемости и по данным о причинах смерти

По данным Всемирной Организации Здравоохранения, среди причин смерти к 2000 году в экономически развитых странах мира первые десять мест занимали ишемическая болезнь сердца, депрессии, дорожный травматизм, инсульт, обструктивные заболевания легких, нагноительные заболевания легких, туберкулез, война и диарейные заболевания, ВИЧ и СПИД

Заболеваемость важнейшими неэпидемическими болезнями Сюда включаются туберкулез, сердечно-сосудистые заболевания и болезни системы кровообращения, злокачественные новообразования, психические расстройства, заболевания, передающиеся половым путем, и т п заболевания, имеющие большую медико-социальную значимость

Туберкулез В группу важнейших заболеваний туберкулез отнесен потому, что в этиологии и организации эффективной борьбы с ним ведущую роль играют социально-экономические факторы Социальная значимость этой патологии определяется и тем, что 75% всех случаев заболеваний туберкулезом происходит в наиболее трудоспособном возрасте от 20 до 49 лет По данным ВОЗ, к концу XX столетия около 1/3 населения планеты было инфицировано туберкулезом По прогнозу ВОЗ к 2004 году число инфицированных туберкулезом во всем мире составит 300 млн человек, разовьется заболевание у 90 млн человек, умрут от него 30 млн человек

Несмотря на успехи, достигнутые в России в борьбе с этим недугом к 1970-м годам, в последнее десятилетие XX века заболеваемость туберкулезом опять начала расти Если в 1991 году она составляла 34 случая на 100 тыс населения, то в 1999 году — 74,4 случая на 100 тыс населения Смертность по этой причине,

несмотря на имевшее место некоторое снижение в 1996–1998 годах, в 1999 возросла до 20,0 на 100 тыс жителей

Причинами роста заболеваемости туберкулезом в России в этот период являлись ухудшение социально-экономических условий, снижение жизненного уровня населения, рост числа лиц без определенного места жительства и занятий, активизация миграционных процессов Определенная ответственность за эти негативные тенденции лежала и на здравоохранении (В 1997 году в России 5,4 тыс. случаев туберкулеза были диагностированы после смерти больных)

В 1999 году в мире ежегодно погибало от *сердечно-сосудистых заболеваний и болезней системы кровообращения* 10 млн человек. Распространенность этой патологии в России в 1999 году составила 813 случаев на 100 тыс. населения. О высокой социальной значимости этой группы заболеваний говорит и тот факт, что на болезни системы кровообращения пришлось 14% общей заболеваемости в стране, 12% случаев временной утраты трудоспособности, 50% всех случаев инвалидности и 55% всех случаев смерти Всего в 1999 году в России умерли от этих болезней 1 862 600 человек, что в 2–3 раза превысило аналогичные показатели стран Европы и США

По оценкам Всемирной Организации Здравоохранения, *нервно-психические заболевания* к 2020 году займут второе место после сердечно-сосудистой патологии В 1999 году в мире нервно-психические расстройства имели 400 млн человек В России за 1990-е годы частота этих заболеваний увеличилась в 1,5 раза, а среди детей и подростков в 2,5 раза К концу XX столетия в России имели психические и нервные расстройства около 10 млн человек, то есть каждый 15-й россиянин имел нарушения психики За 1990-е годы более чем на 8% увеличилось число пациентов состоящих на учете у психиатров К 2000 году в Российской Федерации из каждых 100 тыс жителей свыше 745 человек страдали тяжелыми формами психозов и слабоумия, из них более половины шизофренией, 1257 человек — пограничными психическими расстройствами, 642 — умственной отсталостью

Другой, отнесенной в группу важнейших, является *онкологическая заболеваемость* В России ежегодно регистрируется более

400 тыс злокачественных новообразований О медико-социальной значимости этих болезней говорит тот факт, что по этой причине около 120 тыс человек ежегодно становятся инвалидами, около 300 тыс человек ежегодно погибает

Лица, имеющие важнейшие заболевания, берутся на учет в специализированных диспансерах либо ставятся под динамическое наблюдение в территориальном амбулаторно-поликлиническом учреждении по месту жительства Распространенность таких болезней изучают на основании «Извещения о больном с впервые в жизни установленным диагнозом активного туберкулеза, венерической болезни, трихофитии, микроспории, фавуса, трахомы, рака и других злокачественных новообразований» (учетная форма № 281) Извещение посылается в трехдневный срок в районный (городской) кожно-венерологический кабинет (диспансер) Копия извещения посылается в центр госсанэпиднадзора по месту жительства больного, а для лиц, не имеющих определенного места жительства, — по месту выявления заболевания На основании этих извещений и данных диспансерного наблюдения два раза в год составляется отчетная форма № 61 При выявлении больных с бациллярной формой туберкулеза составляют дополнительное извещение (учетная форма № 58), которое в течение 24 часов посылают в Центр госсанэпиднадзора по месту жительства больного

Инфекционная (эпидемическая) заболеваемость учитывает важнейшие эпидемические заболевания (или даже подозрения на эти заболевания), подлежащие обязательной регистрации и учету на территории страны не зависимо от места заражения и гражданства заболевшего

О социальной значимости этих болезней говорит тот факт, что к концу XX века ежегодно в мире от инфекционных и паразитарных заболеваний ежегодно погибало 16,6 млн человек В России в 1999 году было зарегистрировано 35,5 млн случаев инфекционных заболеваний (на 11 % больше, чем в 1988 году)

Основным учетным документом для регистрации инфекционных болезней в России является «Экстренное извещение об инфекционном заболевании, пищевом, остром профессиональном отравлении, необычной реакции на прививку» (учетная

форма № 058/у) Это извещение составляется каждым врачом или фельдшером, обнаружившим инфекционное заболевание или заподозрившим такое заболевание, и направляется в течение 12 часов в территориальный Центр госсанэпиднадзора по месту регистрации заболевания не зависимо от места жительства больного Медицинские работники фельдшерского звена составляют экстренное извещение в двух экземплярах первый экземпляр отсылают в территориальный Центр госсанэпиднадзора, второй — во врачебное учреждение, в ведение которого находится данный населенный пункт (сельскую, участковую, районную или городскую больницу, амбулаторию, территориальную поликлинику и т д) До отправления экстренное извещение регистрируется в журнале инфекционных заболеваний (учетная форма № 060/у) На основе записей в журнале составляется отчет о движении инфекционных заболеваний за каждый месяц, квартал, полугодие и год (учетная форма № 85 инф), который является основным источником информации об инфекционной заболеваемости для вышестоящих органов управления здравоохранением.

На основании экстренного извещения Центры санэпиднадзора проверяют полноту и своевременность проведенных мероприятий по каждому зарегистрированному случаю (госпитализация, дезинфекция, карантин и т. п), выполняют эпидобследование очага Результаты обследования заносятся в «Карту эпидемиологического обследования очага инфекционного заболевания» (учетная форма № 171).

Все подлежащие оповещению случаи инфекционных заболеваний делят на несколько групп:

• *особо опасные инфекции*, так называемые карантинные заболевания (чума, холера, натуральная оспа, желтая лихорадка и т. п) Извещения о них срочно направляются в вышестоящие органы в специально установленном порядке;

• *инфекции, передающиеся половым путем,* рассматриваются как одни из главных причин, негативно влияющих на репродуктивное здоровье населения.

В группу этих заболеваний, которые до настоящего времени часто называют венерическими, врачи-специалисты относят бо-

лее 20 нозологических форм Однако статистически учитываются в группе инфекций, передающихся половым путем, не все из этих заболеваний. С 1994 года в России подлежат обязательной регистрации в этой группе 9 инфекций До этого в этой группе болезней обязательно регистрировались только 3 заболевания — сифилис, гонорея и трихомониаз. (В США регистрируются только 4 заболевания)

В целом, инфекции, передающиеся половым путем, больше распространены среди женщин Девочки-подростки в 1990-е годы болели в три с лишним раза чаще, чем мальчики. У 18—19-летних это соотношение сохранялось, и только к 30—39 годам заболеваемость мужчин и женщин становилась примерно одинаковой

Следует отметить, что неблагоприятная эпидемиологическая ситуация с инфекциями, передающимися половым путем, была характерна не только для России, но и для всех стран бывшего СССР, включая республики Прибалтики

В 1990-е годы особо негативные тенденции в этих странах выявились с заболеваемостью сифилисом.

В СССР и России в 1972—1978 годах число болевших сифилисом было незначительным Некоторое повышение было отмечено в 1978 году. В 1988 показатели заболеваемости сифилисом были даже более низкими по сравнению с этими же показателями в экономически развитых странах Европы Однако в 1990-е годы положение в корне изменилось За период с 1989 по 1999 год заболеваемость сифилисом в России возросла в 44 раза с 5,3 до 234,8 случая на 100 тыс населения, в том числе у детей в 77 раз В значительной степени увеличилось число врожденных случаев — от единичных в начале 1990-х до 400 к концу 1990-х годов Некоторое снижение темпов роста венерических заболеваний (в том числе и сифилиса) с 1996 года во многом объясняется недоучетом данной категории больных Это связано с началом широкого обращения пациентов к частнопрактикующим специалистам

Динамика числа больных *гонореей* в 1990-е годы была более благоприятной Пик заболеваемости был отмечен в 1994 году, далее пошло ее резкое снижение По мнению специалистов, это связано с неполной регистрацией случаев заболеваний

ВИЧ-инфекции Количество зараженных вирусом СПИДа в мире в 2000 году составляло 36 млн человек В странах Восточной Европы и Центральной Азии количество вновь заболевших оценивалось в 250 тыс , а всего в регионе насчитывалось 700 тыс инфицированных Основными переносчиками болезни, по данным ВОЗ, являются наркоманы, использующие один и тот же шприц помногу раз

В 2000 году в мире зарегистрировано 5,3 млн новых случаев ВИЧ, включая 600 тыс детей в возрасте до 15 лет 3 млн человек умерли от СПИДа.

Наихудшая обстановка по-прежнему сохраняется на юге Африки В 2000 году на страны этого региона пришлось 72% всех новых заболеваний и 80% всех погибших от СПИДа Существует прогноз, что к концу года в регионе будет проживать 25,3 млн людей — носителей СПИДа и ВИЧ-инфекции, из которых 55% составят женщины Это означает, что 8,8% взрослого населения Африки будут носителями смертоносного вируса

В Азии же количество ВИЧ-позитивных случаев в 2000 году составило 6,4 млн человек

В промышленно развитых странах Северной Америки, Западной Европы и Тихого Океана к концу 2000 года проживало 1,5 млн ВИЧ-инфицированных. По прогнозам ВОЗ, к концу года 45 тыс новых случаев ожидается в Северной Америке и 30 тыс — в Западной Европе

За 1990-е годы в России значительно ухудшилась эпидемиологическая обстановка по ВИЧ-инфекции Только за 1999 год было зарегистрировано 15 674 новых случая, что в 3,9 раза больше чем в 1998 году Основные причины заражения ВИЧ — раскованное сексуальное поведение и наркомания, а также недостаточная эффективность гигиенического воспитания В России главной причиной распространения ВИЧ, в отличие от большинства развитых Европейских стран, было парентеральное введение наркотических средств. Более 90% из вновь выявленных больных ВИЧ к 2000 году в России составляли лица, употребляющие наркотики внутривенно

Всего на 1 января 2000 года в России было зарегистрировано свыше 26,6 тыс ВИЧ-инфицированных, из них заболело СПИДом 383 человека, умерло 267 человек

Заболевания, которые регистрируются как ***важнейшие неэпидемические болезни*** с одновременной информацией органов госсанэпиднадзора, подлежат индивидуальному учету в центрах Госсанэпиднадзора Основная информация об этих заболеваниях собирается специализированными лечебно-профилактическими учреждениями.

Заболевания, подлежащие индивидуальному учету в лечебно-профилактических учреждениях, которые представляют в органы санитарно-эпидемиологического надзора об этих заболеваниях только суммарную (цифровую) информацию, — грипп, ОРВИ и др

Заболевания, о каждом случае которых делаются сообщения в центры санэпиднадзора с приведением детальных сведений о заболевании, — брюшной тиф, паратифы, сальмонеллезы, бактериальная дизентерия, пищевые токсикоинфекции, колиты, энтериты, туляремия, менингиты, энцефалиты, инфекционный гепатит, столбняк, орнитоз и т п

Все медицинские данные о больном, включая и некоторые эпидемиологические характеристики, заносятся в зависимости от места регистрации в «Карту стационарного больного», «Индивидуальную карту амбулаторного больного», «Историю развития ребенка» и т д

При анализе инфекционной заболеваемости большое внимание уделяется бактерионосительству. На основании отчетных данных микробиологических, серологических, паразитологических и других лабораторий вычисляются показатели частоты выявленных бактерионосителей

При углубленном изучении инфекционной заболеваемости анализируется сезонность, источники заражения, эффективность профилактических прививок и т д Качественно проведенный анализ инфекционной заболеваемости позволяет правильно оценить медицинскую, социальную и экономическую эффективность не только учреждений здравоохранения, но и

дать оценку социально-экономического благополучия той или иной территории.

Заболеваемость с временной утратой трудоспособности Трудоспособность — совокупность духовных и физических возможностей человека, зависящих от состояния его здоровья, позволяющих заниматься ему трудовой деятельностью (Энциклопедический словарь медицинских терминов — М , 1984 Т 3 С 202) Если изменения в состоянии здоровья, повлекшие утрату трудоспособности, носят временный, обратимый характер и в ближайшее время ожидается выздоровление или значительное улучшение, а также восстановление трудоспособности, то такой вид утраты трудоспособности считается временной

Существенным недостатком статистики заболеваемости с временной утратой трудоспособности является то, что основанием для выдачи больничного листка является не заболевание, а случай трудопотери из-за заболевания То есть, если заболевание не повлекло утраты трудоспособности, то оно зарегистрировано не будет, несмотря на обращение за медицинской помощью

Заболеваемость с временной утратой трудоспособности занимала особое место в санитарной статистике СССР. Информацией о заболеваемости с временной утратой трудоспособности широко пользовались при планировании и оценке эффективности медико-социальных мероприятий на производствах и в учреждениях, при изучении влияния неблагоприятных производственных факторов и т п.

К началу 1990-х годов этот вид статистики заболеваемости потерял свое значение из-за крупных преобразований, произошедших в России Эти преобразования, сопровождавшиеся острыми социально-экономическими проблемами, привели к тому, что регистрация случаев временной нетрудоспособности перестала быть полной из-за массовых нарушений трудового законодательства, особенно в частном секторе экономики Кроме того, сумма выплачиваемых пособий в негосударственном секторе экономики в подавляющем большинстве случаев стала существенно меньше фактической, не декларируемой заработной платы работника, что сплошь и рядом стало приводить к отказу от «ухода на больничный лист»

По мере улучшения экономической ситуации и упорядочения взаимоотношений между работодателями и работниками на производстве, следует ожидать снижения влияния указанных факторов на полноту и достоверность регистрации случаев временной нетрудоспособности

Другим обстоятельством, сыгравшим заметную роль в снижении ценности заболеваемости с временной утратой трудоспособности как источника изучения распространенности той или иной патологии, стало то, что с конца 1980-х годов с целью соблюдения врачебной тайны в больничных листках диагноз заболевания указывается только с согласия пациента В случае его несогласия указывается только причина нетрудоспособности (заболевание, уход за ребенком и т п) В случае прерывания беременности по медицинским показаниям в графе «вид нетрудоспособности» делается запись «нетрудоспособна по медицинским показаниям» В ряде случаев в листках нетрудоспособности проставляется шифр (травмы и отравления, аборты, освобождение в связи с карантином и бактерионосительством)

Такой отказ от обязательной формулировки диагноза на практике полностью исключил возможность проводить анализ структуры заболеваемости и распространенности отдельных групп заболеваний с временной нетрудоспособностью.

Первичный учетный документ, фиксирующий факт заболевания повлекший утрату трудоспособности — листок временной утраты трудоспособности (в отдельных случаях справка) Листок нетрудоспособности выполняет функции не только первичного носителя статистической информации, но и функции юридического документа (подтверждает уважительную причину отсутствия на работе, учитывается при постановке диагноза профессионального заболевания и установления группы инвалидности, служит аргументом при судебных разбирательствах) Кроме того, он является финансовым документом (служит основанием для выплаты пособия по болезни)

На основании листков проводится статистическая разработка заболеваемости по форме № 16(ВН) С 1996 года Министерством здравоохранения РФ введена учетная форма № 025-9/у-96 «Талон на законченный случай временной нетрудоспособно-

сти» Эта форма заполняется врачом при завершении случая временной нетрудоспособности по «Медицинской карте амбулаторного больного» — (учетная форма 025/у), «Медицинской карте студента вуза, учащегося среднего специального заведения» — (учетная форма 025-3/у), «Истории развития ребенка» — (учетная форма 112/у) и прочей медицинской документации, в которой регистрируются случаи временной нетрудоспособности

При заболеваниях (травмах) листок нетрудоспособности выдается в день установления нетрудоспособности, включая праздничные и выходные дни Не допускается его выдача за прошедшие дни, когда больной не был освидетельствован врачом Гражданам, направленным здравпунктом в лечебно-профилактическое учреждение и признанным нетрудоспособными, листок нетрудоспособности выдается с момента обращения в здравпункт

Выдается листок временной нетрудоспособности в следующих случаях заболевание, травма, аборт, беременность и роды, усыновление из роддома, санаторно-курортное лечение, на период медицинской реабилитации, карантин, протезирование, уход за больным членом семьи, при переходе на легкий труд.

Листки нетрудоспособности не выдаются не работающим, при наступлении нетрудоспособности в период отпуска без сохранения содержания, по уходу за больным членом семьи во время очередного отпуска, при протезировании в амбулаторно-поликлинических условиях, на период проведения периодических медицинских осмотров

Право на выдачу листков нетрудоспособности имеют лечащие врачи государственной, муниципальной и частной системы здравоохранения на основании лицензии на проведение экспертизы временной нетрудоспособности В отдельных случаях (в местах, где нет лечащих врачей), по решению местных органов здравоохранения, выдача листков нетрудоспособности и справок может быть разрешена медицинскому работнику со средним медицинским образованием, в том случае, если он ведет самостоятельный лечебный прием

Основанием для выдачи больничного листка является экспертиза трудоспособности, которая устанавливает наличие и тяжесть заболевания, а также его прогноз Основная задача экспертизы — определение на основе медицинского, социального и трудового критериев возможности человека выполнять свои профессиональные обязанности.

Временная нетрудоспособность может быть полной или частичной

Полная нетрудоспособность — человек не может выполнять никакой работы и нуждается в специальном лечебном режиме

Частичная нетрудоспособность — нетрудоспособность в своей профессии, при сохранении способности выполнения другой работы в облегченных условиях или в меньшем объеме

Врач, проводящий экспертизу трудоспособности, при решении о выдаче больничного листка учитывает не только медицинский критерий (степень и выраженность функциональных нарушений, тяжесть патологического процесса, его прогноз), но также социальный и трудовой (условия и характер труда) критерии При этом в ряде случаев больной человек может быть не признан утратившим трудоспособность и, наоборот, физически здоровый человек может быть признан временно утратившим трудоспособность

Например: закрытый перелом одного из пальцев правой кисти у сторожа автобазы и водителя грузовика этой же автобазы Травма есть в обоих случаях Однако водитель не может выполнять свою работу, а сторож выполнять свои функции может Другой пример жена повара заболела гепатитом Сам повар здоров, однако он не может готовить пищу, так как у него был контакт по гепатиту Эти особенности выдачи больничных листков накладывают свой отпечаток на полноту данных о заболеваемости, полученных на основе учета листков нетрудопсобности

Основные показатели заболеваемости с временной утратой трудоспособности представляются в учетной форме №16 Представляются в абсолютных числах и в относительных показателях (на 100 работающих). Согласно действующим инструкциям эти показатели рассчитываются за квартал, полугодие и за год

Число случаев временной нетрудоспособности характеризует число трудопотерь, связанных с заболеваниями.

Число дней нетрудоспособности и средняя длительность нетрудоспособности характеризуют тяжесть заболевания и непосредственно определяют величину экономических потерь

Структура заболеваемости в случаях и днях временной утраты трудоспособности характеризует удельный вес, долю того или иного заболевания среди всех заболеваний

Следует помнить, что данные за квартал и за полугодие всегда являются неполными, так как часть больничных листов не успевает быть зарегистрированными в указанные периоды из-за того, что еще находятся на руках у болеющих В квартальный отчет может не попадать до 30 % больничных листов Особенно сильно этот фактор сказывается при малом числе работающих

Официальный отчет о заболеваемости с временной утратой трудоспособности не отражает особенностей заболеваемости по полу, возрасту, профессии, стажу работы, условиям и характеру труда, несмотря на то что эти факторы оказывают весьма существенное влияние на величину итоговых статистических данных. Именно поэтому сравнивать заболеваемость работников различных предприятий, даже одной отрасли, можно только убедившись, что состав сравниваемых контингентов по этим учетным признакам (пол, возраст, профессия и т п) одинаков или существенно не различается При существенных различиях состава работающих, необходимо проводить стандартизацию показателей

Методика углубленного изучения заболеваемости с выделением физических лиц позволяет проводить детальный статистический анализ с учетом основных социально-демографических, производственно-профессиональных и других факторов, способствующих возникновению заболеваний, анализировать распределение работников по кратности заболеваний в году, определять группу ни разу не болевших (индекс здоровья), группу часто и длительно болеющих (ЧДБ). Обычно в группу ЧДБ относят лиц, терявших трудоспособность в году 4 раза и более или болевших в году не менее 40 дней.

Подобное разделение работающих позволяет дифференцировано планировать и проводить необходимый комплекс лечебно-профилактических и оздоровительных мероприятий

Обычно для углубленного анализа отбираются лица, которые проработали на данном производстве не менее 1 года Так называемые «круглогодовые работники» Ранее такие исследования проводились на основе карт полицевого учета, которые заполнялись на каждого из работников предприятия Эти карты включали общие сведения (пол, возраст, профессия, стаж), данные о случаях заболеваний, результаты профилактических осмотров, диспансеризации и т. п Однако из-за большой трудоемкости учета и обработки таких данных в широкой практике полицевой учет распространения не получил. По мере создания автоматизированных информационных баз данных системы обязательного медицинского страхования возникает возможность возрождения статистики полицевого учета

Инвалидность (заболеваемость со стойкой утратой трудоспособности).

Инвалидность — это постоянная или длительная, полная или частичная потеря трудоспособности **Инвалид** (лат *Invalidus* — бессильный, слабый) — лицо, навсегда или на длительное время потерявшее трудоспособность, частично или полностью, в результате болезни или травмы Всестороннее изучение причин инвалидности, предупреждение ранней инвалидности, восстановление и использование остаточной трудоспособности населения важны для оценки здоровья населения и сохранения трудовых ресурсов.

По данным ВОЗ, в 1990-е годы каждый пятый человек в мире становился инвалидом из-за недостаточности питания, около 15% стали инвалидами вследствие вредных привычек (алкоголизм, наркомания, злоупотребление лекарственными средствами), а 15,1% — вследствие травм в быту, на производстве и в дороге В среднем инвалиды к концу XX века составляли около 10% от всего населения мира

В России численность инвалидов в последние десятилетия XX века имеет стойкую тенденцию к росту (с 3,9 млн человек в 1985 году до 10,1 млн человек в 1999 году). При сохранении

основных тенденций роста число инвалидов к 2005 году достигнет 12,4 млн человек

Большинство людей (80—90%) становились инвалидами в трудоспособном возрасте. При этом уровень реабилитации и восстановления трудоспособности был незначителен (10—12%)

Причинами инвалидности являются общее заболевание, трудовое увечье; профессиональное заболевание, инвалидность с детства; инвалидность с детства вследствие ранения (контузии, увечья), связанная с боевыми действиями в Великой Отечественной войне; военная травма или заболевание, полученное в период военной службы; при ликвидации аварии на Чернобыльской АЭС, последствиями радиационных воздействий и участием в деятельности подразделений особого риска, а также другие причины, установленные законодательством Российской Федерации.

Основаниями для признания гражданина России инвалидом являются.

- нарушение здоровья со стойким расстройством функций организма, обусловленное заболеваниями, последствиями травм или дефектами;
- ограничение жизнедеятельности (полная или частичная утрата лицом способности или возможности осуществлять самообслуживание, самостоятельно передвигаться, ориентироваться, общаться, контролировать свое поведение, обучаться или заниматься трудовой деятельностью);
- необходимость осуществления мер социальной защиты гражданина Наличие одного из указанных признаков не является условием, достаточным для признания лица инвалидом

За установлением инвалидности следует прекращение или изменение характера и условий труда, назначение различных видов государственного социального обеспечения (пенсия, профессиональное обучение, протезирование, трудоустройство и т д)

В зависимости от степени нарушений тяжесть инвалидности может быть различной — от ограничения трудоспособности в основной профессии до полной утраты трудоспособности в лю-

бом виде деятельности С учетом этого устанавливается I, II или III группа инвалидности, а лицу в возрасте до 16 лет — категория *«ребенок-инвалид»*

Категория *«ребенок-инвалид»* может устанавливаться сроком от 6 месяцев до 2 лет, от 2 до 5 лет и до достижения ребенком 16-летнего возраста

Первая группа инвалидности устанавливается больным с тяжелыми и стойкими нарушениями функций организма сроком на 2 года Эти лица не могут себя обслужить и нуждаются в постоянной помощи, уходе или надзоре К этой группе относятся также лица, которые могут быть приспособлены к отдельным видам трудовой деятельности в особо созданных условиях Например, слепые, слепо-глухие и т д

Вторая группа инвалидности устанавливается на год при выраженных постоянных и длительных нарушениях функций организма, не вызывающих, однако, полной беспомощности. К этой группе относятся лица, которые не нуждаются в постоянном уходе, а также лица, у которых в момент освидетельствования нарушения функций не столь тяжелы, но тем не менее им все виды труда на длительный период противопоказаны вследствие возможности ухудшения течения заболевания под влиянием трудовой деятельности В отдельных случаях больным разрешают труд в специально созданных условиях или на дому

Третья группа инвалидности устанавливается на год при значительном снижении трудоспособности вследствие хронических заболеваний или анатомических дефектов, когда

- по состоянию здоровья необходим перевод на работу по другой профессии более низкой квалификации;
- необходимы значительные изменения условий труда по своей профессии, приводящие к значительному сокращению объема производственной деятельности;
- когда не допускаются к работе по эпидемическим показаниям (например больные туберкулезом);
- значительно ограничены возможности трудоустройства вследствие выраженных функциональных нарушений у лиц в возрасте до 40 лет с низкой квалификацией или

ранее не работавших на время обучения или переобучения

Помимо перечисленных случаев, третья и вторая группы инвалидности устанавливаются независимо от выполняемой работы при наличии у больных дефектов и деформаций, влекущих за собой нарушение функций, которые приведены в специальном перечне Инструкции по определению групп инвалидности.

Признание лица инвалидом осуществляется при проведении *медико-социальной экспертизы* (МСЭ) исходя из комплексной оценки его здоровья и степени ограничения жизнедеятельности в соответствии с классификациями и критериями, утвержденными Министерством социальной защиты населения Российской Федерации (РФ) и Министерством здравоохранения РФ

Гражданин направляется на МСЭ клинико-экспертной комиссией ЛПУ или органом социальной защиты населения Лицо без определенного места жительства принимается на МСЭ по направлению органа социальной защиты населения

Специалисты, принимающие экспертное решение, дают по нему разъяснения гражданину или его законному представителю

В случаях, требующих применения сложных видов экспертно-реабилитационной диагностики, специальных методов обследования, тестирования, а также получения дополнительных сведений, составляется программа дополнительного освидетельствования, которая утверждается руководителем БМСЭ и доводится до сведения заявителя в доступной для него форме

Переосвидетельствование инвалидов I группы проводится один раз в 2 года, инвалидов II и III групп — один раз в год, а детей-инвалидов — в сроки, установленные в соответствии с медицинскими показаниями

Без указания срока переосвидетельствования инвалидность устанавливается мужчинам старше 60 лет и женщинам старше 55 лет, инвалидам с необратимыми анатомическими дефектами, другим инвалидам в соответствии с критериями утверждаемыми Министерством социальной защиты населения Российской

Федерации и Министерством здравоохранения и медицинской промышленности Российской Федерации

Переосвидетельствование инвалида ранее установленных сроков проводится в связи с выявлением подложных документов, на основании которых принималось решение о назначении группы инвалидности, либо по направлению учреждения здравоохранения в связи с изменением состояния его здоровья

При статистическом анализе инвалидности рассчитываются показатели во многом аналогичные показателям заболеваемости по обращаемости

3.2. Физическое развитие

Физическое развитие — комплекс морфологических и функциональных показателей, которые определяют физическую работоспособность и уровень биологического состояния индивидуума в момент обследования.

Физическое развитие, отражая процессы роста и формирование организма, непосредственно зависит от состояния здоровья, так как серьезное заболевание задерживает процесс физического развития, особенно у детей и подростков С другой стороны, течение и исход болезни во многом зависят от физического развития больного

Физическое развитие является одним из самых информативных критериев здоровья человека В статистике показатели физического развития употребляются для обобщенных оценок здоровья различных групп населения и в клинической практике при индивидуальной оценке здоровья отдельно взятого пациента

Физическое развитие человека динамически изменяется На каждом возрастном этапе оно характеризуется определенным комплексом свойств организма и обусловленных этим комплексом запасом физических сил Хороший уровень физического развития, как правило, сочетается с высокими показателями

мышечной и умственной работоспособности Процессы физического развития существенно зависят от социальных, экономических, санитарно-гигиенических и других условий Влияние факторов внешней среды наиболее выражено в так называемые сенситивные возрастные периоды:

- грудном и подростковом возрасте, когда интенсивно протекают процессы роста и развития;
- в пожилом и старческом возрасте, для которого характерны процессы инволюции

Исторически сложилось так, что о физическом развитии судят главным образом по внешним морфологическим характеристикам Ценность таких данных неизмеримо возрастает в сочетании с данными о функциональных параметрах организма Поэтому в клинической практике их рассматривают совместно в виде комплекса морфофункциональных показателей Морфофункциональные показатели характеризуют: антропометрические параметры (массу тела, рост, окружности и т д), параметры соматоскопии (телосложение, тургор, форму ног, грудной клетки и т п), физиометрические характеристики (частоту пульса, жизненную емкость легких, артериальное давление и т п)

Одной из самых распространенных комплексных характеристик физического развития является такое понятие, как *конституция* Эта обобщенная характеристика отражает не только особенности телосложения, но также и психическую деятельность, метаболизм, вегетативные реакции, адаптационные и патологические реакции индивидуума.

В настоящее время не существует единой, общепринятой схемы конституциональной типологии В повседневной медицинской практике России наиболее широко применяется схема конституциональных типов по М. В Черноруцкому, выделяющая три основных типа:

- *нормостенический тип* — характеризуется пропорциональными размерами тела и гармоничным развитием костно-мышечной системы,
- *астенический тип* — отличается стройным телом, слабым развитием мышечной системы, преобладанием продоль-

ных размеров тела и размеров грудной клетки над размерами живота, а длины конечностей над длиной туловища,

- *гиперстенический тип* — отличается хорошей упитанностью, длинным туловищем и короткими конечностями, относительным преобладанием поперечных размеров тела и размеров живота над размерами грудной клетки

Одним из неоспоримых доказательств влияния внешних факторов на физическое развитие человека является наблюдаемая последние 100 лет в развитых странах акселерация

Акселерация — ускорение по сравнению с предыдущими поколениями темпов роста и развития, увеличение размеров тела человека, наступление полового созревания в более ранние сроки По поводу причин акселерации большинство ученых считает, что это явление обусловлено комплексом генетических и внешних факторов, среди которых ведущее значение имеют социально-экономические условия

Ретардация — замедление биологического развития организма, явление противоположное акселерации, обусловлено изменчивостью индивидуальных темпов роста К концу пубертатного возраста ретарданты, как правило, догоняют своих сверстников по календарному возрасту

В последнее десятилетие XX века в России отмечается ухудшение физического развития подрастающего поколения В 1,5 раза увеличилось число школьников с дефицитом массы тела, более чем у 25% школьников-подростков отмечалась задержка полового созревания. До 40% пополнения призывов в армию не могли выполнить нормативы по физической подготовке, 11,5% имели дефицит веса, а у 28% обнаруживалось отставание умственного развития Одной из актуальнейших проблем при этом являлось нарастание разрыва между так называемым паспортным и биологическим возрастом у детей и подростков страны

Основанием для введения такого понятия, как биологический возраст послужили индивидуальные особенности роста и развития человека Главными критериями биологического возраста считаются

- зрелость, оцениваемая по степени развития вторичных половых признаков,

- скелетная зрелость (порядок и сроки окостенения скелета);
- зубная зрелость (сроки прорезывания молочных и постоянных зубов)

Чаще всего биологический возраст определяют по степени развития вторичных половых признаков При этом учитываются следующие признаки развитие волос на лобке (*P*) и в подмышечных впадинах (*Ax*); развитие молочных желез (*Ma*) и наступление менархе (*Me*) у девочек, пубертатное набухание сосков (*C*) и перелом голоса у мальчиков

Костный возраст определяется по стадиям оссификации скелета учитываются число точек окостенения, время и последовательность их появления, а также сроки наступления синостозов Для определения костного возраста на практике в большинстве случаев используют стадии оссификации костей кисти и запястья

Зубная зрелость определяется путем подсчета числа прорезавшихся зубов и сопоставления его с существующими нормативами Молочные зубы появляются у детей с 6 месяцев до 2 лет, а постоянные зубы — в возрасте от 6 до 13 лет, за исключением третьих моляров Поэтому зубная зрелость может быть показателем биологического возраста только до 13—14 лет Принято считать, что сроки прорезывания зубов более постоянны, чем сроки оссификации скелета и появления вторичных половых признаков

Процессы роста проявляются увеличением линейных размеров и веса тела Прекращение роста и накопление массы тела говорит о наступлении зрелости, при этом вес тела продолжает увеличиваться только за счет отложения жира, что нельзя рассматривать как проявление роста

Для индивидуальной оценки физического развития многие годы использовался метод сигмальных отклонений В основу этого метода положены таблицы стандартных антропометрических показателей для различных возрастно-половых и этнических групп Стандарты в этих таблицах представлены в виде интервалов $M \pm \sigma$ Где M — стандартная величина показателя (вес, рост и т п), σ — среднеквадратическое (стандартное) отклонение от этой величины Физическое развитие конкретного инди-

видуума оценивалось на основании того, в каком интервале находилось числовое значение его антропометрических данных За норму принимался интервал в пределах $M \pm 1\sigma$ Если оцениваемый антропометрический параметр ребенка оказывался в пределах от $\pm 1\sigma$ до $\pm 2\sigma$, то такого ребенка считали практически здоровым При попадании оцениваемого параметра в интервал от $\pm 2\sigma$ до $\pm 3\sigma$ и более, отклонение параметра считалось высоким Ребенок, получивший такую оценку, считался нуждающимся в специальном углубленном обследовании из-за высокой вероятности наличия патологии

Практика использования стандартов физического развития показала, что эти среднестатистические нормативы не дают всей информации для всесторонней, исчерпывающей оценки физического развития

Одним из комплексных методов оценки физического развития считается метод регрессионного анализа (по шкалам регрессии) Данный метод с помощью простейших математических выражений позволяет выявлять соотношение соразмерных антропометрических признаков, где отдельные признаки физического развития даются в соответствующей зависимости — длина тела и масса, длина тела и окружность грудной клетки (табл. 14)

Благодаря несложности эти методы до недавнего времени пользовались большой популярностью Несмотря на ряд недостатков, некоторыми индексами пользуются и сейчас для ориентировочной оценки отдельных показателей физического развития. Так, для определения должного веса (M) с учетом роста (L) и возраста человека используются следующие выражения, предложенные Броком (индекс Брока)

$M=L-100$ (кг) при росте 155–165 см,

$M=L-105$ (кг) при росте 166–175 см,

$M=L-110$ (кг) при росте более 175 см

Индекс Кетле, или весо-ростовой индекс, получается при делении веса в (г) на рост (см) и равен в среднем для мужчин 350–400 г/см, для женщин — 325–375 г/см

Таблица 14

Наиболее известные антропометрические индексы для детей
(P — вес тела; L — длина тела; Si — рост сидя; C — окружность груди)

Авторы	Индекс	До 1 года	2–3 года	6–7 лет	8–15 лет
Ливи	$\sqrt[3]{\frac{P}{L}}$	2,9	2,7–2,8	2,3–2,5	2,2–2,3
Ропер	$\frac{P}{L}100$	2,5	2,0–2,2	1,2–1,3	1,2–1,5
Пирке (Pelidisi)	$\sqrt[3]{\frac{10P}{Si}}100$	98–100	97	95–98	92–96
Пирке (Bedusi)	$\frac{L-Si}{Si}100$	54–58	68–70	78–80	80–95
Пинье	$L-(P+C)$	15–16	23	30–35	26–35
Бругш	$\frac{C}{L}100$	65–68	–	63–51	49–53
Эрисман	$C-L/2$	От +10 см до +13,5 см	От +6 см до +9 см	0	От −1 см до −3 см
Пейзар	$\frac{Si\ 100}{L}$	У новорожденного около 70, постепенно с возрастом педеет, у взрослого около 50			

В последнее десятилетие в практике широко используется более простой центильный метод оценки индивидуального физического развития, который широко применяется и за рубежом

В основу метода положено процентное (центильное) распределение величины данного параметра физического развития При этом, для каждого исследуемого возраста выделяются неодинаковые по величине центильные интервалы («коридоры», «зоны»). За норматив принимается «коридор» в пределах с 25-го по 75-й центиль. Величины ниже этого коридора распределяются по центильным интервалам следующим образом 1-й — очень низкие, во 2-й входят низкие величины, 3-й интервал включает сниженные показатели. Соответственно распределяются величины, превышающие средние значения: 6-й интервал включает повышенные показатели, 7-й — высокие показатели, а 8-й интервал включает очень высокие величины. Таким образом, зоны ниже 10-го и выше 90-го центиля свидетельствуют о выраженном снижении или, соответственно, повышении измеряемого показателя.

По центильным таблицам длина тела, в зависимости от попадания величины показателя в тот или иной интервал, характеризуется как средняя или нормальная, сниженная, повышенная, низкая, высокая. Аналогично оценивается масса тела и другие антропометрические величины

Центильные закономерности соотношений между массой тела и длиной применяются в качестве оценки гармоничности физического развития детей и подростков (гармоничное, дисгармоничное и резко дисгармоничное).

В настоящее время физическое развитие ребенка педиатр начинает прослеживать с детской поликлиники, определяя комплексную оценку состояния здоровья Оценка состояния здоровья проводится всем детям в определенные эпикризные сроки жизни

Эпикризные сроки — это промежутки времени, через которые проводится обязательная комплексная оценка состояния здоровья:

- на 1-м году жизни — 1 месяц (1 раз в месяц);
- на 2-м году — 3 месяца (1 раз в 3 месяца);
- на 3-м году — 6 месяцев (1 раз в 6 месяцев);
- с 4-го по 7-й год и старше — 1 год (1 раз в год).

3.3. Медицинская демография

Демография — наука о народонаселении Ее основателем считается Дж Граунт (1620—1674). Однако становление демографии как полноправной науки, имеющей свою специфическую терминологию, свои специфические методы, произошло в конце XIX — начале XX веков.

Слово «*Демография*» произошло от греческих слов «*demos*» — народ и «*grapho*» — писать, изображать. Демография исследует закономерности явлений и процессов, происходящих в структуре, размещении и динамике народонаселения Первое появление термина «*Демография*» в 1855 году связано с именем французского ученого А Гийяра, который первым дал определение

этой науки, как совокупности математических знаний о населении, его движении, физическом, умственном и духовном состояниях

В настоящее время в англоязычной литературе нередко используются синонимы «формальная демография» или «социальная демография». Во франкоязычной специальной литературе — «демографический анализ» Учеными США часто употребляется термин «система знаний о населении» (*population studies*).

Медицинская демография как прикладная дисциплина статистики здоровья в качестве критериев здоровья оперирует специальными демографическими показателями Показатели медицинской демографии неразрывно связаны с показателями общей демографической ситуации, и выделение медицинской демографии в отдельную отрасль демографии во многом носит условный характер Например, уровень смертности населения (медико-демографический показатель) обусловлен возрастно-половым составом, процессами естественного старения населения или отдаленными демографическими последствиями тяжелых войн.

Многие общие демографические показатели (*demographic information* — англ), не являясь напрямую показателями здоровья, являются базой для статистики здоровья населения и широко используются в организации здравоохранения Например, без знания территориального размещения, возрастно-полового и социального состава населения невозможно получение многих статистических коэффициентов заболеваемости и в первую очередь интенсивных показателей, которые рассчитываются относительно количества населения, проживающего на определенной территории (на 1000, 10 000 человек различного возраста, пола, социального положения и т п) Отсутствие исходных демографических данных послужило причиной того, что в трудах статистиков дореволюционной России, посвященных заболеваемости, почти все данные приводились либо в абсолютных числах, либо в показателях структуры (экстенсивных показателях).

Знание демографического состава населения необходимо и при анализе причин повышенного или пониженного уровня заболеваемости, причин преобладания той или иной патологии Это обусловлено тем, что рост частоты регистрируемых заболе-

ваний может быть следствием не ухудшившейся экологической ситуации или падения уровня жизни, а последствием постарения населения и, наоборот, низкий уровень заболеваемости может быть объяснен молодым составом населения Состав населения, при прочих равных, во многом определяет структуру заболеваемости, преобладание тех или иных причин смерти Например, повышение в составе населения доли лиц пожилого и старческого возраста увеличивает число болеющих и умирающих от болезней, свойственных этим возрастным группам (сердечно-сосудистые заболевания, сосудистые поражения центральной нервной системы, злокачественные заболевания).

Демографические характеристики являются основой планирования медицинской помощи населению Нормативы по обеспечению населения всеми видами амбулаторно-поликлинической и стационарной помощи базируются на знании численности и возрастно-полового состава населения (численность взрослого населения на один территориальный терапевтический участок, число больничных коек на 1000 населения, число врачей-педиатров на 100 детей, число акушер-гинекологов на 1000 женщин и т п.)

Важнейшим источником получения объективной демографической информации являются переписи населения

Перепись населения — это специальная научно организованная государственная статистическая операция по учету и анализу данных о численности населения, его составе и распределении по территории Перепись населения, при ее правильной организации и проведении, является наиболее достоверным источником сведений о демографической ситуации

Различные формы численного учета населения существовали с древних времен. Достоверно известно, что учеты населения проводились в Древней Греции, Египте, Вавилоне, Мессопотамии, Китае и Японии Они проводились, как правило, с фискальными (налоговыми) целями и в целях пополнения армии Поэтому этот учет охватывал только мужское население

Классическим примером такого учета служит Книга Чисел Ветхого Завета Библии «— *И сказал Господь Моисею в пустыне Синайской . исчислите все общество сынов Израилевых по родам*

их, по семействам их, по числу имен, всех мужского пола поголовно от двадцати лет и выше, всех годных для войны у Израиля »

Относительно систематизированный учет народонаселения в России начали вести со времен Петра I, когда были введены подушные переписи, называемые ревизиями На основе этих ревизий определялась численность мужского населения, облагаемого податями. Единицей учета являлась «ревизская душа», которая служила основой списков («ревизских сказок») Ревизские души числились в этих списках до следующей ревизии, не зависимо от того, живы они или уже несколько лет как умерли Поскольку проведение ревизий было связано с обложениями «тяготами» (налоги, рекрутские наборы и т п), возвращением «беглых», оказавшихся на мануфактурах, к своим прежним владельцам и т п., заинтересованные стороны всячески старались исказить «ревизские сказки» в выгодном для себя направлении Несмотря на то что указ Петра I от 1698 года, положивший начало проведению ревизий, и последующие указы предусматривали суровые наказания всех допускавших преднамеренный недоучет, вплоть до смертной казни, размах искажений был весьма широк Механизм такого рода деятельности и достоверность «ревизских сказок» достаточно наглядно описаны Н. В. Гоголем в знаменитом произведении «Мертвые души»

Справедливости ради следует отметить, что случаи массового преднамеренного искажения информации при переписях отмечались и в зарубежных странах

И в наше время продолжают встречаться аналогичные ситуации. Так, в Турции при проведении переписи 2000 года в некоторых городах был замечен значительный прирост населения по сравнению с данными переписи 1997 года Выяснилось, что сборщики информации приписали к живым и обитателей кладбищ К примеру, в городе Битлис в 1997 году насчитали 330 тыс жителей В 2000 году жителей города стало уже 560 тыс Похожая ситуация была отмечена в еще нескольких городах Мотив этой «операции» был довольно прост согласно турецким законам государственные дотации, выдаваемые муниципальным округам, прямо пропорциональны населению округов. За каждого вос-

крещенного мертвеца город получал около 36 долларов США (по данным Lenta.ru от 22.06 2001)

Однако при всех недостатках и искажениях переписи населения играли и играют существенную роль в оценке численности населения стран, социальном и национальном составе, оценке численности городского и сельского населения и т п

Положение о проведении первой всеобщей переписи населения Российской Империи было утверждено в 1895 году Проведена она была по состоянию на 9 февраля (28 января) 1897 года В программах сбора и разработки ее материалов было много дефектов, что привело к многочисленным неточностям Несмотря на это, она является единственным более или менее достоверным источником о численности и составе населения России в конце XIX века Помимо всеобщей переписи в некоторых регионах Российской Империи проводились свои, местные, переписи. В Москве (1871, 1882, 1902, 1912), Петербурге (1862, 1863, 1864, 1869, 1881, 1890, 1900, 1910, 1915), в Астраханской (1873), Акмолинской (1877), Псковской (1870) и некоторых других губерниях В 1863 и 1881 годах было переписано население всей Курляндской, а в 1881 году — Лифляндской и Эстляндской губерний Таких местных переписей было проведено около 200

На протяжении последующего XX столетия в России произошло 8 всеобщих переписей· в 1920, 1926, 1937, 1939, 1959, 1970, 1979 и 1989 годах

Первые переписи РСФСР (1920 и 1926) организовывались для решения неотложных задач преодоления последствий гражданской войны и разрухи. Перепись 1937 года проводилась в период развернувшихся в стране социальных преобразований Ее результаты оказались ошеломляющими Выявились многомиллионные потери населения в результате гражданской войны, голода, коллективизации и репрессий Вследствие этого все руководство переписи было репрессировано, а результаты переписи уничтожены Однако насущные интересы государства требовали точных данных о демографическом составе страны, поэтому в 1939 году была проведена еще одна перепись С тех пор переписи в России проводились регулярно раз в 10 лет до 1989 года Последняя перепись населения РФ проводилась в 2002 году

Перепись населения в современных условиях представляет собой важную статистическую операцию Среди научно-организационных основ переписи выделяется несколько основных и обязательных

Всеобщность — охват без исключения всех лиц, проживающих на территории, где проводится перепись, независимо от пола, возраста, социального положения, вероисповедания

Периодичность — в большинстве стран перепись проводится один раз в 10 лет (в экономически развитых странах — через 5 лет).

Единство методики сбора и обработки данных — всем обследуемым лицам независимо от их пола, возраста, места жительства, семейного положения и т п задаются одни и те же вопросы В некоторых случаях часть населения обследуется более подробно, с целью получения дополнительной, детальной информации. Например, в переписи населения СССР 1979 года 11 вопросов из 16 задавались всему населению, а остальные 5, касающиеся занятости населения, общественного положения, миграции и т п , задавались только 25% населения Аналогичные выборочные опросы проводились и в последующих переписях

Единовременность — данные переписи регистрируются на определенный момент времени (критический момент переписи) несмотря на то, что сбор данных продолжается несколько дней, поскольку перепись провести одномоментно невозможно Перепись 1989 года, например, проводилась 8 дней Единовременность соблюдается для того, чтобы повысить точность данных, поскольку состав и численность населения непрерывно меняется Так, к моменту всеобщей переписи в 1998 году в России ежедневно рождалось за один час в среднем 147 человек, а умирало — 227 Критический момент переписи, как правило, назначается на зимнее время, в середине недели, когда миграция населения наименьшая, число отпусков незначительное, а школьные каникулы уже кончились Например: критическии момент переписи 1970 года был 14 января (точнее в ночь с 14 на 15 января, в 0 часов), переписи 1989 — в ночь на 17 января Если в критический момент переписи человек присутствовал, а затем к моменту заполнения переписного листа по какой-либо причине выбыл

(уехал или умер), то он регистрировался как благополучно проживающий в данном месте И наоборот, ребенок, родившийся в 10 часов утра 17 января 1989 года, то есть после критического времени, не включался в число жителей страны

Сбор сведений методом опроса, без обязательного документального подтверждения. В случае расхождения этих данных с юридическим статусом опрашиваемого в переписной лист заносятся данные, полученные путем опроса Например· лица указавшие, что они состоят в браке, если даже он юридически не оформлен, будут зарегистрированы как супруги

Строгое соблюдение тайны переписи. Все собранные при переписи сведения используются только статистикой для получения итоговых данных Это требование, называемое «статистической тайной», является одним из главных условий получения достоверной информации До конца 80-х годов в СССР и, соответственно, в России доступ к детальным статистическим данным переписи был весьма ограничен В открытой печати публиковались или неполные, или уже устаревшие данные о составе и территориальном размещении населения страны. В 90-е годы эти ограничения были в значительной степени отменены.

В годы между переписями, в так называемый межпереписной период, оценка численности населения проводится двумя способами Во-первых, к итогам последней переписи населения добавляются лица, родившиеся и прибывшие в данный регион, и вычитаются лица, умершие и выбывшие из данного региона. В России для этого используют данные актов регистрации гражданского состояния (случаи рождения, смерти, браков, разводов) и статистических талонов прибытия или убытия, которые заполняются в отделах милиции, занимающихся регистрацией приезжающих или убывающих

Другой способ получения данных о составе населения в межпереписной период — проведение специальных вычислений ***Интерполяция*** — используется для получения данных в промежутке между двумя уже прошедшими переписями ***Экстраполяция*** — получение на основании результатов последней переписи данных на последующие годы.

Во время Великой Отечественной войны в СССР для получения оперативной информации о массовом перемещении населения из-за эвакуации практиковались *моментные переписи* С их помощью по очень короткой программе (4–5 вопросов) за 2–3 дня в отдельных регионах проводилась сплошная регистрация местного населения. Всего за годы войны было проведено 100 таких переписей

Статистический анализ народонаселения ведут в двух основных направлениях. статика населения и динамика населения

3.3.1. Статика населения

Статика — численность и состав населения на определенный момент времени по таким основным учетным признакам, как пол, возраст, социальная группа, профессия и занятие, семейное положение, национальность, язык, культурный уровень, грамотность, образование, место жительства (город или село), географическое размещение и т д

По данным ООН, численность населения Земли в 1999 году превысила 6 млрд человек и продолжает стремительно расти

Считается, что численность жителей Земли достигла первого миллиарда в 1804 году, второго миллиарда почти в пятнадцать раз быстрее — 123 года спустя. Чтобы численность населения увеличилась с 5 до 6 млрд, потребовалось еще меньше времени — всего 12 лет. По прогнозу ООН, дальнейший рост населения Земли будет немного замедляться 7 млрд — 2013 год, 8 млрд — 2028 год, 9 млрд — 2054 год

В процентном отношении население Земли к концу XX века распределялось следующим образом: Азия — 61%, Европа и Северная Америка — 16%, Африка — 13%, Латинская Америка — 8%, Океания — 0,5%

Всего в мире к концу XX века насчитывается 10 государств с населением более 100 млн человек, из них самые населенные — Китай (1200 млн чел) и Индия (900 млн человек) Если в 1950 году государств с населением более 100 млн человек было 4

(Китай, Индия, США, СССР), то, по прогнозу ООН, в 2025 году количество таких стран может достигнуть 18

Численность населения в России составляла на конец 2000 года 145,7 млн человек За последние 10 лет число россиян сократилось на 2,3 млн человек (рис 47, табл 15)

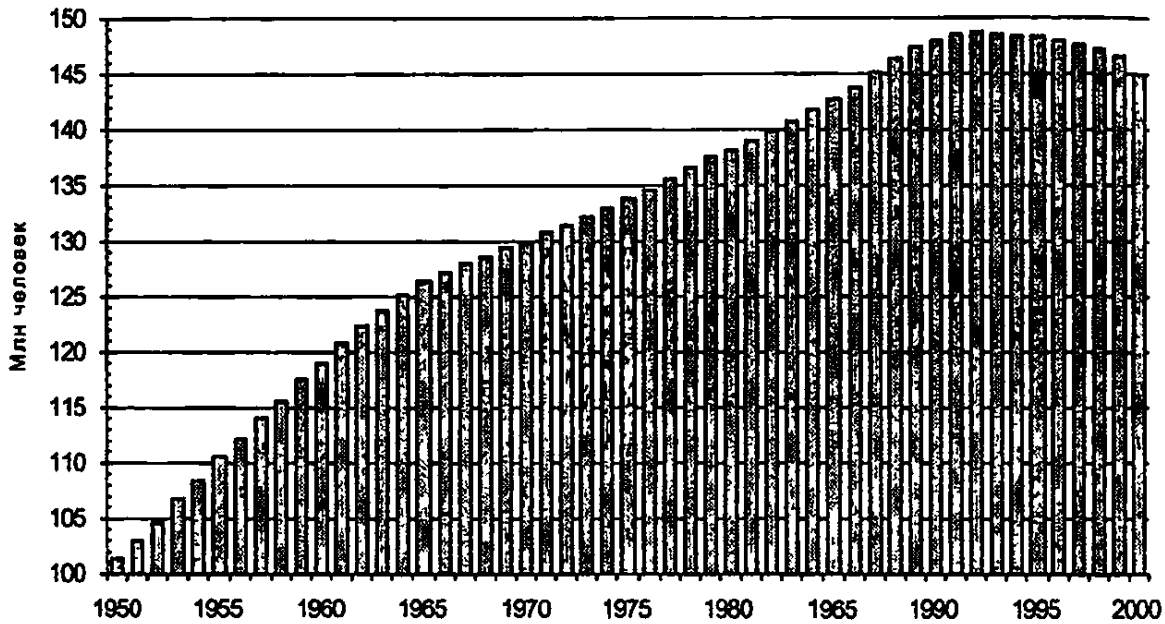


Рис. 47. **Динамика численности населения России во второй половине XX века**

Таблица 15

Динамика общей численности населения Российской Федерации за 1990–2000 годы

Годы	Число жителей (млн чел)	Годы	Число жителей (млн чел)
1990	147,1	1996	147,6
1991	148,0	1997	146,1
1992	148,5	1998	146,7
1993	148,3	1999	146,3
1994	148,0	2000	144,9
1995	147,9		

Снижение численности населения — тенденция характерная для всех экономически развитых стран Согласно прогнозам ООН, например, в таких странах, как Болгария, Италия, Эстония, к 2050 г численность населения сократится на 25—30%. Только население США в ближайшие 50 лет будет расти за счет миграции и, согласно прогнозным оценкам, увеличится почти на четверть

Весьма существенной статистической медико-демографической характеристикой является тип возрастной структуры населения Он определяется на основании соотношения среди всего населения страны долей лиц в возрасте 0—14 лет и 50 лет и старше. В зависимости от этого соотношения различают прогрессивный, регрессивный и стационарный типы населения

Прогрессивным считается тип населения, в котором доля лиц в возрасте 0—14 лет превышает долю населения в возрасте 50 лет и старше

Регрессивным считается тип населения, в котором доля лиц в возрасте 50 лет и старше превышает долю лиц в возрасте от 0—14 лет.

Стационарным типом принято считать население, в котором доля лиц в возрасте 0—14 лет равна доле лиц в возрасте 50 лет и старше.

В XIX веке число людей в возрасте старше 65 лет на Земле не превышало 2—3% К 2000 году их доля составляла уже 14% от общей численности населения земного шара. К 2030 году эта доля, по прогнозу ООН, возрастет до 25—30% В том числе в Австрии и Швейцарии — до 30 %, Бельгии — до 28%, Китае — до 33%, Дании — до 24%, Украине — до 27%, Беларуси — до 25%, Эстонии — до 29%, Латвии — до 27%.

Если в настоящее время на каждого пожилого человека в мире приходится 9 трудоспособных, то уже в 2030 году их будет всего 2. В России за последние годы также наблюдается увеличение удельного веса старших возрастных групп (табл 16)

Таблица 16

Возрастная структура населения России в 1897–1999 годах (в %)

Годы переписей	Возрастные группы			Итого
	До 14 лет	15–59 лет	60 лет и старше	
1897	48,7	44,2	7,1	100,0
1927	49,0	44,2	6,8	100,0
1939	45,4	47,9	6,7	100,0
1959	36,8	54,2	9,0	100,0
1989	29,9	54,8	15,3	100,0
1999	19,0	62,9	18,1	100,0

Сокращение численности и старение населения является одним из главных побуждающих мотивов многих социально-экономических реформ, в том числе повышение нижней границы пенсионного возраста; изменение причин преждевременного выхода на пенсию, создание системы медицинского обеспечения пожилого населения, изменение доли отчислений из заработков в пользу пенсионного и медицинского обеспечения пожилых и т п

Возрастно-половой состав населения определяет распространенность и структуру патологии, уровень рождаемости, смертности населения Уровень обращаемости за медицинской помощью различных социально-демографических групп также имеет существенные отличия Например, неработающее население (в первую очередь пенсионеры и дети) пользуется медицинской помощью в 2–2,5 раза чаще, чем работающее население

Удельный вес женского населения в России составляет 53,1%, доля мужчин 46,9%. Однако в разные периоды жизни людей соотношение полов различно. Если в возрасте 0–5 лет доля мальчиков составляет 52,2%, то к 30–34 годам соотношение мужчин и женщин практически становится одинаковым В возрасте 70 лет и старше удельный вес мужчин в структуре населения снижается и составляет всего 26,8%

3.3.2. Динамика населения. Естественное и механическое движение

Динамика населения — это изменение (движение) количества населения вследствие естественных биологических и социально-экономических процессов. Динамика населения подразделяется на механическое и естественное движение.

Механическое движение населения происходит в результате переселения или миграции, связанной чаще всего со сменой жительства Слово миграция происходит от латинского слова «*migratio*» (*migro* — перехожу, переселяюсь) Миграция подразделяется на

- безвозвратную (постоянную);
- временную (переселение на длительный, но ограниченный срок);
- сезонную (перемещение в определенные периоды года);
- маятниковую (регулярное перемещение к месту работы или учебы за пределы своего населенного пункта)

Кроме того, различают *внешнюю миграцию* (переселение за пределы своей страны) и *внутреннюю* (перемещение внутри страны) Примером внешней миграции являются *эмиграция*, т е выезд граждан из своей страны на постоянное место жительства в другую страну и *иммиграция* — въезд граждан другой страны в данную страну Внутренняя миграция является частью процесса урбанизации

Эмиграция населения может быть вызвана многими причинами экономическими, социальными, политическими (преследования и др), в том числе и демографическими (перенаселенность) На территории России к концу 90-х годов XX века находилось большое число выходцев из стран дальнего зарубежья (Афганистан, Шри-Ланка, Бангладеш, Эфиопия, Ангола, Вьетнам и др) и ближнего зарубежья (главным образом, таджики, украинцы, молдаване, грузины, армяне, азербайджанцы) Причина практически одна и та же — дисбаланс между численностью населения в странах исхода и потребностями национальных рынков труда, что вынуждает «лишних» людей искать источники существования за рубежом

Иммиграционный приток приходится в основном на страны, испытывающие дефицит рабочей силы, либо в те государства, где ситуация позволяет иностранцам добывать средства к существованию нелегально

США, Израиль, Германия, Италия, Греция, Португалия, Испания, Австралия и некоторые другие повышают квоты на прием иммигрантов, главным образом трудоспособного возраста Прием таких лиц позволяет не только заполнять образовавшийся вакуум на рынках труда, но и омолаживать нацию, поддерживать на должном уровне соотношение трудоспособных и нетрудоспособных лиц.

В Кувейте, например, численность населения на июль 1998 года составляла 2 млн 238 тыс человек, из которых коренных жителей насчитывалось только 772 тыс человек, остальные — иммигранты При этом в государственном секторе задействованы, главным образом, кувейтские граждане, в частном — иностранцы Причина — более высокие социальные гарантии государственных служащих

Многие западноевропейские государства широко используют граждан других стран на сезонных работах, на строительстве, в торговле, в качестве прислуги в частных домах

Большой размах приобрела иммиграция в Россию Миграционный прирост в России (суммарный итог миграции и эмиграции), начиная с 1960-х годов, имел четкую тенденцию к росту, в отличие от прямо противоположной тенденции естественного прироста, достигнув к 2000 году почти 202,1 тыс человек в год При этом эмигрировало из страны 136 тыс человек, иммигрировало в страну — 338,1 тыс человек. Положительное сальдо миграции образовалось за счет приезда из стран СНГ и Балтии Число эмигрантов в страны «дальнего зарубежья» превышало число иммигрантов на 50 тыс человек (въехало в Россию 8,3 тыс , выехало 58,1 тыс человек) Среди стран «дальнего зарубежья» наибольшие иммиграционные потоки из России пришлись на Германию, куда выехало в 2000 году на постоянное место жительства 37,5 тыс человек (для сравнения, в Израиль в 2000 году выехало из России 9 тыс человек, в США — 4,4 тыс человек).

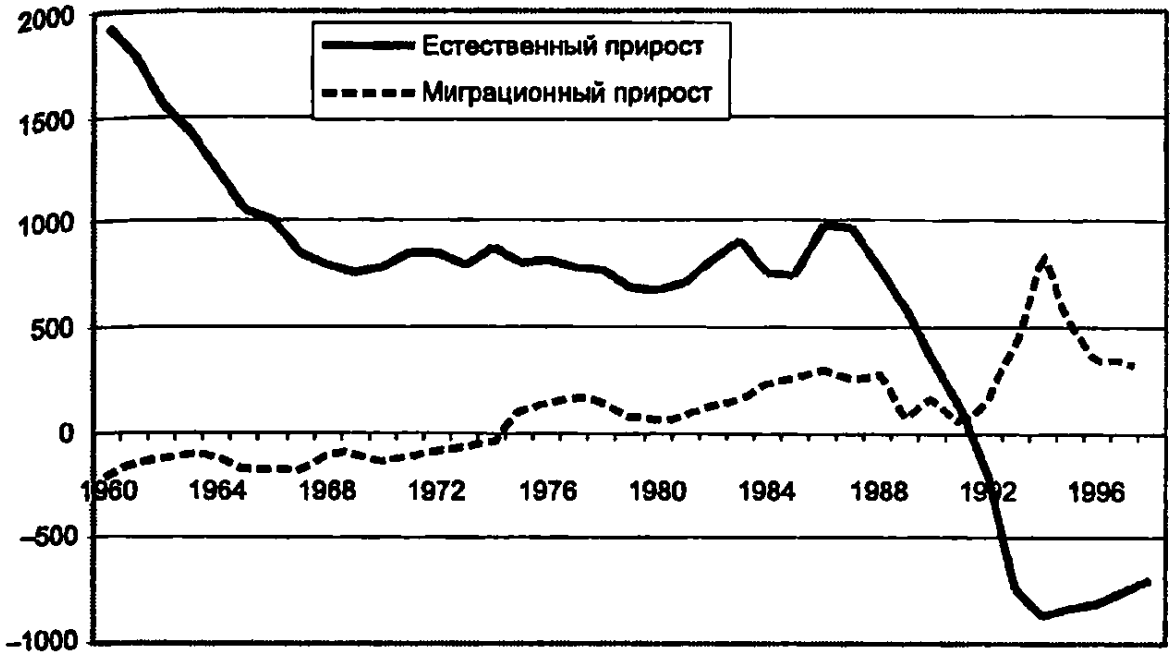


Рис. 48. **Динамика естественного и миграционного прироста в России в 1960–1990 годах**

Урбанизация (от латинского «*urbs*» — город) — это процесс миграции сельских жителей в развитые промышленные центры Имеет место в большинстве государств мира. В результате число жителей села сокращается, а города — стремительно растет Население Токио, например, к концу 2000 года подошло к отметке в 30 млн чел , Бомбея, Шанхая, Сан-Пауло — к 20 млн чел , Нью-Йорка, Мехико — к 16 млн чел , Пекина и Джакарты — к 14 млн человек. Эти города уверенно догоняет Москва

В целом, урбанизация представляет сложное комплексное явление, которое рассматривается как некий объективный процесс, связанный с обеспечением все большего числа жителей планеты условиями более удобной и комфортабельной жизни, обстановки для более полного развития способностей, для более защищенной и здоровой жизни

В России доля городского населения составляла в 1897 году 15,0%, увеличившись к 1913 году до 18,0% С 1917 по 1989 год она выросла до 73,5% С 1993 года численность городского насе-

ления стала несколько снижаться (73,0%), а с 1994 года удельный вес городских жителей фактически остается стабильным

Влияние урбанизации на здоровье населения неоднозначно и сложно С одной стороны, для большей части жителей городов присущ более высокий уровень социально-бытовых условий, чем на селе (рациональное водоснабжение, централизованное отопление, комфортабельное жилище, высокая доступность квалифицированной и специализированной медицинской помощи и т п) С другой стороны, большая плотность населения, загрязнение атмосферного воздуха, повышенный шумовой фон, усиление нервно-психических нагрузок с одновременным снижением мышечной активности и многое другое могут оказывать неблагоприятное воздействие на здоровье людей

Практически все виды миграции меняют уровень и структуру заболеваемости, смертности населения, приводят к изменению нагрузки на учреждения здравоохранения, изменяют экологическую и эпидемическую ситуации в регионах Все это обусловливает необходимость пересмотра организации медицинской помощи

Например, маятниковая и сезонная миграция способствуют распространению инфекционных заболеваний и ухудшению эпидемиологической ситуации, травматизму

Нарастание миграционных процессов в России с начала 1990-х годов оказывает негативное влияние на здоровье населения, сменившего место жительства из-за частой необустроенности нового жилья, длительной адаптации к новым социальным, экономическим, природно-климатическим и другим условиям

Естественное движение — это совокупность таких демографических явлений, как рождаемость, смертность, в том числе младенческая и материнская смертность, естественный прирост населения, брачность, разводимость, средняя продолжительность предстоящей жизни и т п

Рождаемость населения является важнейшим критерием жизнеспособности и воспроизводства населения Именно от рождаемости зависит замещение выбывающих поколений

В демографической статистике следует различать понятия рождаемости и плодовитости. Плодовитость — это биологиче-

ская способность к деторождению, а рождаемость относится к фактическому деторождению, т е реализации этой способности

Как статистический показатель рождаемость определяется либо абсолютным числом рождений, либо относительным коэффициентом, который называется — ***общий показатель рождаемости*** Он определяется отношением числа родившихся живыми в данном году на 1000 человек населения (‰).

$$\frac{\text{Общее число родившихся за год}}{\text{Среднегодовая численность населения}} \times 1000$$

Среднегодовая численность населения рассчитывается как полусумма численности населения на начало и конец календарного года.

Оценивается уровень общей рождаемости в следующих границах показателей: низкий уровень — до 19,9 ‰, средний — 20,0–29,9 ‰, высокий — 30,0 ‰ и выше Более детальные оценки приведены в табл. 17

Таблица 17

Критерии оценки уровней рождаемости (Минздрав РФ 1998 г)

Оценка	Общий коэффициент (‰)
Очень низкая	До 10,0
Низкая	10,0–14,9
Ниже среднего	15,0–19,9
Средняя	20,0–24,9
Выше среднего	25,0–29,9
Высокая	30,0–39,9
Очень высокая	40,0 и более

Полный показатель рождаемости вычисляется с учетом детей родившихся живыми и мертвыми.

Источником получения информации для вычисления этих показателей в России служит «Медицинское свидетельство о рождении» (учетная форма № 103/у-98), которое выдается во всех случаях живорождения при выписке матери из лечебно-профи-

лактического учреждения стационарного типа, в котором произошли роды Если роды произошли на дому, то «Медицинское свидетельство о рождении» выдает учреждение, медицинский работник которого принимал данные роды В сельской местности оно может быть выдано акушеркой или фельдшером, принимавшим роды, если в учреждениях здравоохранения этой местности нет врачей. При многоплодных родах «Медицинское свидетельство о рождении» выдается на каждого родившегося ребенка в отдельности В «Истории развития новорожденного» (учетная форма № 097/у) должна быть сделана запись о выдаче «Медицинского свидетельства о рождении» с указанием его номера и даты выдачи, а в случае мертворождения такая запись производится в «Истории родов» (учетная форма № 096/у). Затем, в течение месяца со дня рождения ребенка, на основании «Медицинского свидетельства о рождении» в государственных органах записи актов гражданского состояния по месту рождения ребенка или по месту жительства родителей производится регистрация новорожденного

Для более точного определения интенсивности процесса воспроизводства в основу расчета рождаемости принимаются случаи рождений живых детей у женщин детородного возраста Отношение числа родившихся живыми на 1000 женщин детородного возраста (15–49 лет), называется показателем плодовитости или фертильности Возрастной интервал 15–49 лет называется плодовитым (генеративным) периодом женщины Между общим коэффициентом рождаемости и коэффициентом плодовитости существует взаимосвязь

$$n = Fds,$$

где n — общий коэффициент рождаемости, F — коэффициент плодовитости; ds — доля женщин возраста 15–49 лет в общей численности населения

Кроме того, учитывая, что в различные возрастные периоды число рождений у женщин различно, вычисляют повозрастные показатели плодовитости Для этого весь генеративный период женщины условно подразделяют на отдельные возрастные интервалы (15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49 лет)

и, соответственно, в расчет повозрастных показателей фертильности принимают только число рождений у женщин соответствующего возраста

Более детальную характеристику рождаемости дают специальные показатели воспроизводства брутто- и нетто-коэффициенты воспроизводства Брутто-коэффициент воспроизводства учитывает только родившихся у женщин детородного возраста девочек, для этого число родившихся детей умножают на долю девочек среди новорожденных Нетто-коэффициент фертильности учитывает только девочек, доживающих до генеративного возраста.

Рождаемость в Российской Федерации за последние десятилетия XX века снижалась. В 1999 году ее уровень составлял около 8,4 на 1000 населения (8,4 ‰) (табл. 18). Этот уровень рождаемости в 2 раза ниже необходимого для замещения поколений родителей их детьми и составляет около 1,3 рождений на одну женщину в течение ее жизни Для простого воспроизводства населения, каким бы низким ни был уровень смертности, необходимо 2,15 рождений на одну женщину.

Таблица 18

Снижение рождаемости в России за 1988–1999 годы

Годы	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Рождаемость в ‰	16,0	14,6	13,1	12,1	10,7	9,4	9,6	9,3	8,9	8,6	8,8	8,4

Наиболее низкие показатели рождаемости, не превышающие 7,0 ‰ ,к концу 2000 года были отмечены в Санкт-Петербурге, Ленинградской, Московской, Смоленской и Тульской областях. В конце 1990-х годов только в восьми регионах России — в республиках Дагестан, Тува, Ингушетия, Саха (Якутия), Калмыкия, Алтай, Кабардино-Балкария, Карачаево-Черкессия — еще обеспечивалось замещение поколений родителей их детьми

Россия не единственная страна, переживающая подобные тенденции К концу XX века почти 45% всего населения земного шара живет в странах, где уровень рождаемости не обеспечивает режим простого воспроизводства населения (почти все страны Европы, Восточной Азии, Таиланд и Куба)

Обеспокоенные постоянным снижением уровня рождаемости многие экономически развитые страны, в первую очередь европейские, принимали меры по материальной поддержке семей, стимулируя тем самым и увеличение количества детей Швеция, Дания, Финляндия расходовали в 90-х годах на семейную политику более 4% валового внутреннего продукта (ВВП). Страны Южной Европы — от 0,4% до 1,1% ВВП, Германия — 2,1%, Франция — 2,6% Пакет детской помощи, состоявший из денежных выплат, налоговых льгот, различных услуг, в той или иной форме существовал во всех странах Европейского Союза, а также в России. Размеры помощи различались как в зависимости от экономических возможностей стран, так и внутри одной страны в зависимости от типа семьи, числа и возраста детей.

В XX столетии отмечалось снижение уровня рождаемости во многих странах и в том числе в России, за исключением некоторого общего подъема рождаемости после второй мировой войны Данный факт влечет за собой целый ряд негативных последствий. нарастание дефицита трудовых ресурсов, постарение населения, уменьшение числа женщин фертильного (детородного) возраста

Например, для полноценного воспроизводства населения на одну женщину фертильного возраста должно приходиться 2,5—3 рожденных детей. В 1999 году в России на одну женщину приходилось 1,2 рождений

Среди социальных факторов, объясняющих низкую рождаемость в России, можно выделить следующие: социально-экономический кризис страны, неудовлетворительные жилищные условия Вместе с тем, низкая рождаемость в экономически развитых странах свидетельствует о том, что преодоление экономического кризиса и рост благосостояния населения сами по себе не гарантируют рост рождаемости. На первое место в этих условиях будут выдвигаться такие факторы, как общественное положение женщин, их занятость в производстве, ориентация молодых семей на мало- или многодетность, национальные особенности, психологические и религиозные факторы

Снижение рождаемости в России имело и объективные предпосылки, порожденные второй мировой войной Уже во второй

половине 1960-х годов отмечалось значительное снижение рождаемости, обусловленное отдаленным последствием войны От малочисленного поколения, родившегося в военные годы, Россия имела малочисленное поколение их детей, появившихся в 60-е годы Именно это поколение (внуки тех, кто родился в годы войны) в начале 90-х годов и вошло в наиболее активный репродуктивный возраст

Таким образом, резкое снижение рождаемости в 1990-е годы является следствием сочетания двух негативных факторов негативные последствия коренных преобразований и отдаленные демографические последствия Великой Отечественной войны

Важнейшим демографическим фактором, во многом определяющим интенсивность воспроизводства населения, является брачность и его антагонизм — разводимость В России, которая традиционно была страной с более высоким, чем в странах Запада, уровнем брачности, с начала 1960-х годов наблюдается резкое снижение брачности

В 1999 году в России на 1000 населения ежегодно регистрировалось 6,2 браков и 4,3 разводов, то есть значительное число браков распадались. Следует отметить, что в общую статистику разводимости за каждый год попадают случаи расторжения браков, заключенных и в предыдущие годы. Однако динамика показателей брачности и разводимости за последние 40 лет в России позволяет сделать вывод о стабильном снижении устойчивости семьи Так, если в 1960 году на 1000 населения было зарегистрировано 12,5 браков и 1,5 разводов, то за 40 лет частота браков снизилась в 2 раза, а разводов, наоборот, увеличилась более чем в 2 раза.

Снижение интереса к браку выражается также в динамике и других показателей, например, уровня окончательного безбрачия В конце 80-х годов доля женщин, так и не вышедших замуж к возрасту 45—49 лет, в России была в 1,5—3 раза ниже, чем в таких странах, как Франция, Германия, Швеция В начале 90-х годов уровень окончательного безбрачия в России стал стремительно приближаться к показателям стран Запада При этом в России не происходит «компенсации» официальной, зарегистрированной брачности, брачностью фактической

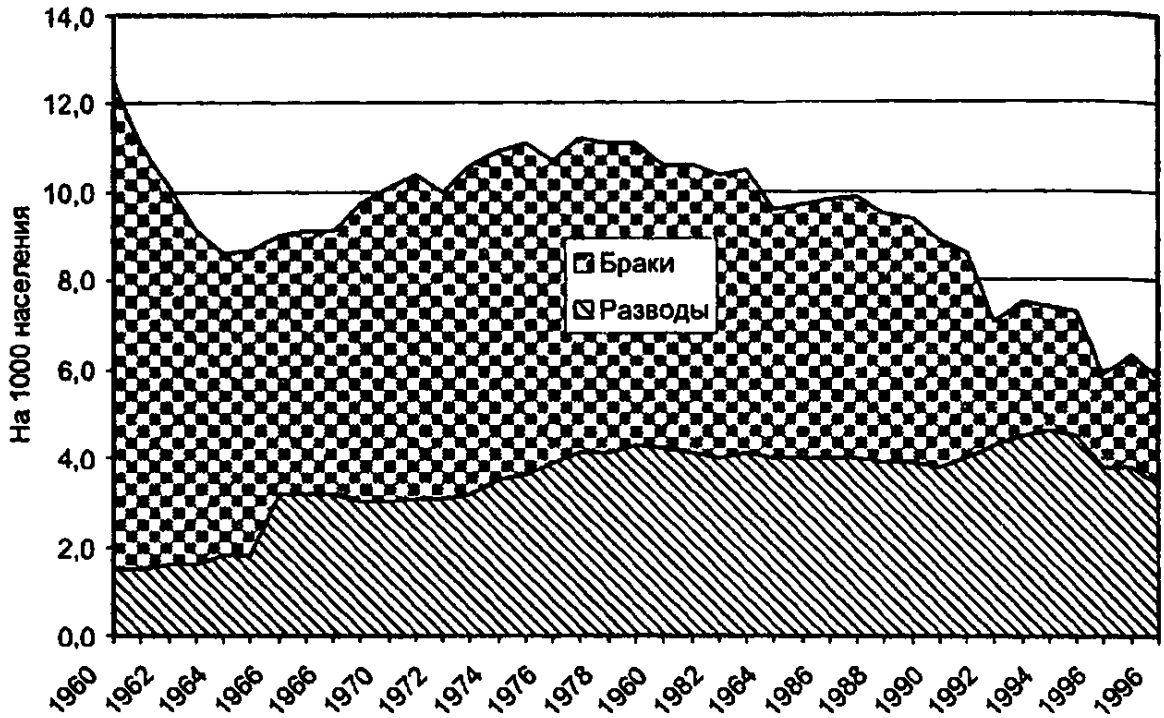


Рис. 49. **Распределение общих коэффициентов брачности и разводимости в России в 1960–1998 годах**

Как видно из графика (рис. 49), в динамике разводимости четко выделяются два периода ее роста середина 60-х годов и начало 90-х. Рост разводимости в первый из этих двух периодов можно объяснить реализацией «отложенного спроса» на разводы, связанной с изменением законодательства о разводах в 1965 году в сторону их «либерализации»

Динамика разводимости в 90-е годы определяется ослаблением интереса к браку, «бегством от брака», которое является одним из выражений кризиса семьи Эта тенденция особенно усилилась, когда ухудшились условия жизни большей части населения России

Нельзя забывать и об одном из главных последствий разводов только в 1997 году в семьях, распавшихся в результате развода, было более 450 тыс детей А всего детей, оставшихся сиротами и без попечения родителей, к концу 90-х годов XX столетия в России было более 620 тыс , из них около 90% составляли дети при живых родителях Для сравнения в труд-

ном для страны 1945 году, году тяжелейшей послевоенной разрухи, было 200 тыс детей-сирот (в пересчете на 1000 населения в полтора раза меньше, чем в 1999 году), и тогда это были истинные сироты, оставшиеся без родителей в результате войны

Смертность населения относится к важнейшим статистическим показателям, характеризующим санитарное состояние населения. Этот демографический показатель — один из самых распространенных в международной статистике здоровья Общий показатель смертности (*crude mortality rate* — англ.), для краткости часто называемый «показатель смертности» (*mortality rate* — англ.), показывает частоту смертей среди всего населения и рассчитывается следующим образом:

$$\frac{\text{Общее число умерших за год}}{\text{Среднегодовая численность населения}} \times 1000$$

Среднегодовая численность населения считается как полусумма численности населения на начало и конец года

Так как общий коэффициент смертности в значительной степени зависит от особенностей возрастного состава населения, он мало пригоден для каких-либо сравнений Например, в регионе, среди жителей которого преобладают лица пожилого возраста, общий коэффициент смертности будет более значительным, чем в регионе, где доля лиц пожилого возраста меньше. Более точными являются показатели *повозрастной смертности* (*age specific mortality rate* — англ.), которые рассчитываются по отдельным возрастным группам Иногда для устранения влияния на показатели неоднородного возрастного состава применяется *стандартизованный показатель смертности* (*age adjusted mortality* — англ.).

При углубленном статистическом анализе смертности рассчитывают специальные (частные *cause-specific mortality rate* — англ.) коэффициенты с учетом пола (*sex-specific mortality rate* — англ), профессии, причин смерти (*proportionate mortality* — англ) и т д.

Таблица 19

Общая смертность в России за 1988–1999 годы

Год	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Смертность в ‰	10,7	10,7	11,2	11,4	12,2	14,5	15,7	15,0	14,2	13,8	13,6	14,7

За последнее десятилетие XX столетия динамика показателей смертности в России была столь отрицательной, что некоторые ученые характеризовали ситуацию как кризисную (табл 19, 20) Возник даже специальный термин для ее обозначения: «депопуляция»

Таблица 20

Оценки уровня общей смертности

Общий коэффициент смертности, в ‰	Оценка уровня смертности
До 10	Низкий
10–14,9	Средний
15–24,5	Высокий
25–34,9	Очень высокий
35 и более	Чрезвычайно высокий

К концу XX столетия в России, как и в других развитых странах, первое место в структуре причин смертности занимали болезни органов кровообращения, на долю которых приходилось 55,0% всех случаев смертей, на втором — новообразования (14,7%). Травмы и отравления в структуре причин смерти стояли на третьем месте (13,8%) (табл 21)

Источником информации о причинах смерти в России являются записи во «Врачебных свидетельствах» (учетная форма № 106/у-84), заполняемых врачом. В сельской местности, имеющей в учреждении здравоохранения только одного врача, в случае его отсутствия (болезнь, отпуск и т. п), а также в учреждениях, где нет врача (фельдшерско-акушерский пункт), фельдшером выдается «Фельдшерская справка о смерти» (учетная форма № 106-1/у-84) При наличии подозрения на насильственную смерть, после искусственного аборта, произведенного вне лечебного учреждения, при внезапной смерти детей и взрослых, не

находившихся под медицинским наблюдением, а также невозможности установить причину смерти или личность умершего врачебное свидетельство или фельдшерская справка о смерти не выдается А все соответствующие документы выдаются только после вскрытия судебно-медицинским экспертом

Таблица 21

Распределение умерших по причинам смерти в России в 2000 году в % к 1999 году

Показатели	Тыс человек			На 100 тыс населения		
	2000 г	1999 г	прирост (+), снижение (−)	2000 г	1999 г	2000 г в % к 1999 г
Всего умерших	**2032,4**	**1953,0**	**+79,4**	**1532,7**	**1467,0**	**104,5**
В том числе от						
болезней системы кровообращения	1118,9	1078 6	+40,3	843,8	810,2	104,1
новообразований	273,1	273,0	+0,1	206,0	205,0	100,5
несчастных случаев. отравлений и травм	284,9	267,3	+17,6	214,9	200,8	107,0
из них от						
транспортных (всех видов) травм	36,2	34,6	+1,6	27 3	26,0	105,0
случайных отравлений алкоголем	31,0	24,1	+6.9	23,4	18 1	129,3
самоубийств	52,9	52,8	+0,1	39,9	39,7	100,5
убийств	36,9	34,1	+2,8	27,9	25,6	109,0
болезней органов дыхания	92,7	85,3	+7,4	69,9	64,1	109,0
болезней органов пищеварения	58,9	55,3	+3,6	44,5	41,6	107,0
инфекционных и паразитарных болезней	33,1	32,3	+0,8	25,0	24,3	102,9

В соответствии с законодательством случай смерти подлежит регистрации в органах ЗАГСа по месту жительства умершего или по месту наступления смерти на основании заключения медицинского учреждения не позднее 3 суток с момента наступления смерти или обнаружении трупа

В практическом здравоохранении для характеристики качества медицинской помощи широко используется статистический показатель летальности, который необходимо отличать от показателя смертности.

Летальность (*case-fatality rate* — англ) — это частота смерти от какого-либо заболевания Анализ летальности проводят по отдельным нозологическим формам по данным отчетов лечебно-профилактических учреждений

Материнская смертность — один из основных интегрирующих показателей здоровья женщин репродуктивного возраста и качества работы родовспомогательных учреждений В Международной классификации болезней X пересмотра материнская смертность определяется как «обусловленная беременностью смерть женщины, наступившая в период беременности или в течение 42 дней после ее окончания от какой-либо причины, связанной с беременностью, отягощенной ею или ее ведением, но не от несчастного случая или случайно возникшей причины»

Данный показатель позволяет оценить все потери беременных от абортов, внематочной беременности, от акушерской и экстрагенитальной патологии в течение всего периода гестации рожениц и родильниц в течение 42 дней после прекращения беременности

$$\frac{\text{Число умерших беременных (с начала беременности), рожениц, родильниц в течение 42 дней после прекращения беременности} \times 1000\ (100\ 000)}{\text{Число живорожденных}}$$

В соответствии с Международной классификацией болезней показатель материнской смертности должен рассчитываться на 1000 живорожденных Однако ВОЗ, учитывая небольшое число случаев материнской смертности в развитых странах и соответственно незначительную величину показателя при расчете на 1000 новорожденных, в статистических показателях приводит расчеты на 100 000 новорожденных.

По данным ВОЗ за 1996 год, мировая статистика материнской смертности свидетельствует, что в мире ежегодно умирали

585 000 женщин в связи с беременностью и родами. В странах с развитой экономикой материнская смертность составляла 27 на 100 000 живорожденных, в низкоразвитых странах этот показатель был значительно больше — 480 на 100 000 живорожденных (табл. 22)

Таблица 22

Материнская смертность по регионам мира по данным ВОЗ

Регионы	Число смертей на 100 000 живорожденных
Мир в целом	430
Высокоразвитые страны	27
Низкоразвитые страны	480
Африка	870
Азия	390
Европа	36
Восточная	62
Северная	11
Южная	14
Западная	17
Латинская Америка и страны Карибского бассейна	190
Америка Центральная	140
Америка Южная	200
Америка Северная	11
Австралия — Новая Зеландия	10
Оксания	680

Показатель материнской смертности в России (табл 23) превышает аналогичный показатель в среднем по развитым странам более чем в 2 раза, а по ряду стран Европы и США — в 4 раза

Таблица 23

Материнская смертность в России за 1988–1998 годы

Год	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Материнская смертность	50,0	49,0	47,4	52,4	50,8	51,6	52,3	53,2	48,9	50,2	44,0

В РФ уровень материнской смертности на протяжении последнего десятилетия остается высоким (см табл 23).

Регистрация и учет материнской смертности ведется в соответствии с теми же правилами, что и общей смертности Важное значение для предотвращения случаев материнской смертности имеет детальный анализ каждого случая учреждением, где произошла смерть матери. Кроме того, необходимо анализировать структуру причин материнской смертности В настоящее время неблагоприятным фактором следует считать то, что в большинстве случаев причины материнской смертности являются управляемыми. В современных условиях в структуре материнской смертности ведущее место занимают аборты вне лечебного учреждения, на втором месте среди причин смерти матери кровотечения при беременности и родах, токсикоз беременности занимает третье место, на четвертом месте — внематочная беременность

В статистике смертности детей принято выделять следующие показатели детскую или младенческую смертность — смертность детей на первом году жизни; смертность детей в возрасте до 5 лет; смертность детей в возрасте от 1 года до 15 лет

Детская смертность (*mortality rate, infant* — англ) рассматривается как оперативный и весьма информативный статистический критерий оценки санитарного благополучия населения, уровня и качества медицинской помощи, в том числе акушерской и педиатрической службы

Существуют несколько различных способов расчета коэффициента младенческой смертности. Самым простым из них считается способ расчета по следующей формуле

$$\frac{\text{Число детей, умерших на 1-м году жизни в течение года} \times 1000}{\text{Число родившихся живыми в данном календарном году}}$$

Однако среди детей, которые умерли в возрасте до 1 года в течение календарного года, могут быть родившиеся как в данном календарном году, так и в прошлом году Поэтому в практи-

ке для более точного расчета показателя младенческой смертности используют рекомендованную ВОЗ формулу Ратса

$$\frac{\text{Число детей, умерших на 1-м году жизни в течение года} \times 1000}{\text{2/3 родившихся живыми в данном календарном году + 1/3 родившихся живыми в предыдущем году}}$$

В наиболее экономически развитых странах (Япония, Германия, Канада, США) показатель детской смертности составляет от 4 до 7‰

В годы после второй мировой войны Россия достигла значительных успехов в снижении детской смертности Даже после ее кратковременного повышения в 1992–93 годах (которое могло быть связано с изменением практики регистрации и с изменением социально-демографического состава контингента рожающих женщин в момент резкого сокращения рождаемости), детская смертность не превысила уровня 1984 года А после 1993 года стало наблюдаться снижение этого показателя (табл 24).

Таблица 24

Детская смертность в России за 1990–2000 годы (‰)

Год	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Показатель	17,4	17,8	18,0	19,9	18,6	18,1	17,4	17,2	16,5	16,4	15,6

Уровень младенческой смертности оценивается в следующих границах. низкий уровень — до 10,0‰, средний — 10,1–19,9‰, высокий — 25,0‰ и более.

Вероятность смерти ребенка в течение первого года жизни распределяется весьма неравномерно Максимальная смертность наблюдается в первые сутки после рождения ребенка, а в последующие дни, недели, месяцы его жизни вероятность смерти снижается Причины, влияющие на смертность детей в эти периоды жизни, также существенно различаются Кроме того, подготовкой матери к родам, охраной здоровья ребенка занимаются медики различных специальностей. Поэтому, чтобы наиболее пол-

но и объективно оценить причины младенческой смертности и роль работников родовспоможения и педиатрической службы, рассчитывают специальные показатели детской смертности, подразделяя их с учетом момента наступления смерти В том числе для детей родившихся живыми

- *ранняя неонатальная* смертность — наступление смерти в первые 168 часов жизни ребенка;
- *поздняя неонатальная* смертность — наступление смерти на 2, 3, 4-й неделях жизни;
- *неонатальная* смертность (*mortality rate, neonatal* — англ.) — наступление смерти в первые 4 недели, или в первые 28 дней жизни ребенка;
- *постнеонатальная смертность* (*mortality rate, postneonatal* — англ) — наступление смерти с 29-го дня жизни до 1 года

По уровню смертности детей в возрасте до 5 лет Россия занимает одно из последних мест среди экономически развитых стран В 1995 году этот показатель составил 22,70, в 1998 году — 20,30 В структуре смертности детей старше года ведущее место занимают несчастные случаи, травмы и отравления, на втором месте — заболевания центральной нервной системы, на третьем месте — злокачественные новообразования и лейкозы

Для более полной оценки здоровья беременных, рожениц и новорожденных, а также качества медицинской помощи на этапе подготовки, в родах и в ранний неонатальный период в статистику смертности включают показатель перинатальной смертности

Перинатальная смертность — наступление смерти плода до родов с 22 недель беременности, в родах и в течение 168 часов после рождения ребенка Перинатальный период подразделяется на три периода антенатальный (внутриутробный) — с 22-й недели беременности до родов, интранатальный — (период родов) и постнатальный период — (первые 7 дней или первые 168 ч после рождения)

Смерть в антенатальный период (до начала родовой деятельности) и в интранатальный (смерть в родах) составляют мертворожденность, которую рассчитывают следующим образом

$$\frac{\text{Число мертворожденных} \times 1000}{\text{Число детей, родившихся живыми и мертвыми}}$$

В расчет перинатальной смертности дополнительно к мертворожденным включают умерших в первые 168 часов (в постнатальный период):

$$\frac{(\text{Число мертворожденных + число детей, умерших в первые 168 часов жизни}) \times 1000}{\text{Число детей, родившихся живыми и мертвыми}}$$

По данным ВОЗ, в развитых странах доля мертворождений среди перинатальных потерь составляла 45,8% В развивающихся странах — 56,1%. В последние 5 лет XX века в Российской Федерации эта доля колебалась в пределах 50,0–54,5%

Таблица 25

Перинатальная смертность, мертворожденность, ранняя неонатальная смертность в Росснйской Федерации за 1987–1999 годы (‰)

Показатели	Годы											
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Перинатальная смертность	18,6	17,9	17,6	18 4	17,9	17,6	17,9	17 4	16,1	16,0	15,8	15,0
В том числе мертворождаемость	9,6	9,3	9,0	9,1	8,8	8,4	8,0	7,9	7,6	7,9	8,0	7 5
ранняя неонатальная смертность	9,0	8,7	8,6	9,3	9,1	9,2	9,9	9,5	8,5	8,1	7,8	7,5

Приведенные данные (табл 25) свидетельствуют, что в последнее десятилетие XX века, несмотря на снижение рождаемости, показатель перинатальной смертности практически не изменялся В целях международной сопоставимости отечественной статистики с зарубежной, начиная с 1993 года, в соответствии с рекомендациями ВОЗ в мертворождаемость включают все случаи смерти плода и новорожденного с массой тела 500 г и более или

длиной тела 25 см и более, или сроком беременности 22 нед и более Ранее отечественная статистика учитывала все случаи смерти плода с массой тела 1000 г и более, длиной тела 35 см и более, сроком беременности 28 нед и более Изменились также и критерии живорождения и мертворождения

Живорождением является полное изгнание или извлечение продукта зачатия из организма матери вне зависимости от продолжительности беременности, причем плод после такого отделения дышит или проявляет другие признаки жизни, такие как сердцебиение, пульсация пуповины или произвольные движения мускулатуры, независимо от того, перерезана пуповина и отделилась ли плацента Каждый продукт такого рождения рассматривается как живорожденный

Мертворождением является смерть продукта зачатия до его полного изгнания или извлечения из организма матери вне зависимости от продолжительности беременности На смерть указывает отсутствие у плода после такого отделения дыхания или любых других признаков жизни, таких как сердцебиение, пульсация пуповины или произвольные движения мускулатуры

Массой при рождении считается результат первого взвешивания плода или новорожденного, зарегистрированный после рождения. Эта масса должна быть установлена предпочтительно в течение первого часа жизни, до того как в постнатальном периоде произойдет значительная потеря массы Измерение длины новорожденного (плода) должно обязательно производиться при вытянутом его положении на горизонтальном ростомере

Новорожденные (плоды), родившиеся с массой тела до 2500 г, считаются плодами с низкой массой при рождении; до 1500 г — с очень низкой, до 1000 г — с экстремально низкой

Учет перинатальной смертности ведется на основании анализа «Медицинских свидетельств о перинатальной смертности» (ф. № 106-2/у-98) В сельской местности, где в штатах учреждений здравоохранения нет врача, на случаи смерти в перинатальном периоде фельдшером заполняется «Фельдшерская справка о смерти» Порядок оформления и хранения этих справок имеет много общего с оформлением документации о смерти Однако

есть и некоторые особенности в случае смерти детей (плодов) при многоплодных родах свидетельство заполняется на каждого ребенка (плод) отдельно. Во всех случаях перинатальной смерти ребенка (плода) на дому или в учреждениях здравоохранения для установления причины гибели ребенка (плода) производится вскрытие. «Медицинское свидетельство о перинатальной смерти» («Фельдшерская справка о смерти») и корешок к нему оформляются в день вскрытия· клинические данные о патологии матери, ребенка (плода) во время беременности и родов берутся из медицинской документации («Истории родов» — ф № 096/у, «Истории развития новорожденного» — ф № 097/у). В случае мертворождения при родах, проведенных без помощи медицинского персонала, или в случае смерти ребенка на 10-й неделе жизни, не наблюдавшегося медицинским работником, вскрытие производится только судебно-медицинским экспертом, которым заполняются все соответствующие документы.

Естественный прирост населения является обобщающей характеристикой роста или убыли населения. Только за последние 40 лет XX столетия количество землян удвоилось При этом 96% прироста приходилось на развивающиеся страны Каждую минуту к концу XX века на планете рождалось примерно 250 и умирало 103 человека То есть число жителей увеличивалось каждые 60 секунд на 147 человек

Естественный прирост не всегда отражает демографическую обстановку в обществе и поэтому его необходимо оценивать только в соотношении с показателями рождаемости и смертности. Показатель естественного прироста населения может вычисляться следующим образом

1-й способ показатель рождаемости – показатель смертности;

$$\text{2-й способ: } \frac{\text{Абс число родившихся за год} - \text{Абс. число умерших за год}}{\text{Среднегодовая численность}} \times 1000$$

Отрицательный естественный прирост (убыль) населения свидетельствует о неблагополучной обстановке в обществе (наличие

социально-экономических кризисов, войны, землетрясений и т д.). Высокий естественный (положительный) прирост населения свидетельствует о благоприятной демографической обстановке только в случае низкого уровня смертности Если же при высоком уровне смертности наблюдается высокий естественный прирост населения (высокий показатель рождаемости), говорить о благоприятной демографической ситуации нельзя, так как данный факт характеризует неблагоприятную обстановку с воспроизводством населения.

Впервые в 1992 году в России отмечен отрицательный естественный прирост (противоестественная убыль) населения, что ведет к сокращению численности населения нашей страны (табл 26).

Таблица 26

Естественный прирост населения России за 1989–1999 годы (‰)

Годы	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Показатель	+3,9	+2,2	+0,7	−1,5	−5,1	−6,1	−5,7	−5,3	−5,2	−4,8	−6,3

По прогнозу ООН, неблагоприятные тенденции в рождаемости и смертности приведут к тому, что и во многих развитых странах к 2050 году численность населения сократится В Италии, например, ожидается уменьшение численности населения на 28%, в Румынии — на 26%, в Словении и Венгрии — на 25%, Испании и Чехии — на 24%, Греции — на 23%, Украине — на 22%, Японии и России — на 17%, Австрии — на 14%, Германии — на 11%, Нидерландах, Словакии — на 10%, Дании, Швейцарии — на 9%, Финляндии — на 5%, Китае — на 4%, Швеции — на 3% Исключение составят лишь США, где численность населения за счет значительного приема иммигрантов постоянно растет и к 2050 году должна увеличиться примерно на 80 млн человек.

Для интегральной медико-демографической оценки широко используют показатели продолжительности жизни населения В международной демографической статистике для этого применяется показатель **средней продолжительности предстоя-**

щей (ожидаемой) жизни Под этим показателем понимают число лет, которое в среднем предстоит прожить данному поколению родившихся при условии, что на всем протяжении жизни смертность в каждой возрастной группе будет такой же, какой она была в том году, для которого производилось исчисление показателя Аналогичные расчеты могут производиться и при определении средней продолжительности предстоящей жизни для любого поколения сверстников (лиц одного года рождения)

Статистический показатель — средняя продолжительность предстоящей жизни, для краткости часто называемый средней продолжительностью жизни, нельзя путать со средним возрастом умерших или средним возрастом живущих Оба этих показателя непригодны для характеристики здоровья населения и поэтому не используются в практике медико-демографических исследований Это связано с тем, что на их величину существенное влияние оказывает возрастной состав населения. Средний возраст умерших и живущих понижается из-за увеличения среди населения доли лиц молодого возраста

Для получения показателей средней продолжительности жизни используется исчисление специальных таблиц — таблиц смертности (таблиц доживаемости) Таблицы смертности, или доживаемости — это система взаимосвязанных показателей, характеризующих порядок вымирания населения при данном уровне смертности в отдельных возрастных группах Они показывают, как число одновременно родившихся лиц, условно принятое за 10 000 или 100 000, постепенно уменьшается с увеличением возраста из-за смертности

В подавляющем большинстве стран Европы средняя продолжительность предстоящей жизни, начиная с 1900 года, неуклонно росла. Исключение составляли Румыния и Россия.

В 1979 году средняя продолжительность жизни на нашей планете приближалась к 59 годам, в основном благодаря Европе и Америке, где она перевалила за семь десятков В Азии она достигла только 56 лет, а в Африке — 45 лет.

В России пик относительного благополучия приходился на 1979 год — 68 лет, к 1992 году показатель снизился до 65 лет у женщин и 56 лет у мужчин

В последние годы средняя продолжительность предстоящей жизни в России уменьшилась и в 1999 году была значительно ниже, чем в большинстве экономически развитых стран В Японии в 1999 году этот показатель составлял 80, Канаде — 79, США — 77, Афганистане — 45 лет Среди стран Европы наиболее благополучное положение в Швеции Там средняя продолжительность предстоящей жизни составляла для женщин 82 года, а мужчин — 77,5 В России для мужчин — 61 год, и для женщин — около 73 лет

Большая продолжительность жизни женщин, чем мужчин, отмечается практически везде, но в России этот разрыв значительно превышает различия имеющиеся в других странах Если в период с 1960-х до 1980-х годов во всем мире ожидаемая продолжительность жизни мужчин выросла на 13 лет, а женщин — на 14,1, то в России за этот же период она возросла всего на 1,8 года среди мужчин и на 2,9 года среди женщин (табл. 27)

Таблица 27

Средняя продолжительность предстоящей жизни (лет) мужчин и женщин в России в 1897–1999 годах

Годы	Мужчины	Женщины
1897	29,4	31.7
1927	33,7	37,9
1939	34,9	42,6
1959	63,0	71,5
1989	64,2	74,5
1996	61,0	73,1
1999	61,3	72,9

В 1985–1987 годах в России было отмечено неожиданное и значительное повышение ожидаемой продолжительности жизни на 3,2 года у мужчин и на 1,3 года у женщин. Большинство специалистов считают, что это кратковременное улучшение связано, главным образом, с широкомасштабной антиалкогольной

кампанией (1985 г) В 1988–1991 годах позитивный эффект кампании был исчерпан и показатели ожидаемой продолжительности жизни опять стали ухудшаться В 1994 году ожидаемая продолжительность жизни мужчин упала до самой низкой отметки в условиях отсутствия военных действий и голода — 58 лет В 1998–1999 годах наметилось некоторое улучшение показателей ожидаемой продолжительности жизни при рождении для мужчин (61,3) и женщин (72,9).

Самым старым мужчиной на земле был японец Сигечио Идзуми, родившийся на небольшом островке возле Окинавы 29 июня 1865 года Когда в Японии проводилась первая перепись населения (1871 г), ему было шесть лет, а умер он от воспаления легких 21 февраля 1986 года Таким образом, Идзуми прожил 120 лет и восемь месяцев, достигнув предела долголетия, установленного наукой для человеческого рода Родители Идзуми умерли, когда он был совсем молодым Имея небольшой рост (1,51 м) и малый вес (46,5 кг), он отличался хорошо развитой мускулатурой, так как работал в порту грузчиком Этот трудолюбивый бедняк ел мало, довольствуясь овощами и сырой рыбой Он был дважды женат и два раза становился вдовцом (последний раз в возрасте девяноста одного года)

В мире на одного столетнего мужчину приходится в среднем шесть столетних женщин Старейшей женщиной на земле была француженка Жанна Кальман. Она родилась в 1875 году в Арле и скончалась в 1997 году в возрасте 122 лет и почти восьми месяцев Французский врач Жорж Гаруайан посвятил Жанне Кальман свою докторскую диссертацию Он допускал, что немалую роль в долголетии Жанны Кальман могла сыграть наследственность ее мать скончалась в возрасте 90 лет, а отец прожил более 96

Наследственность, разумеется, всего лишь один из факторов Подобно японцу Идзуми, Жанна Кальман была мала ростом Она никогда не грешила излишествами и не перенесла ни одной серьезной болезни Лечащий врач Знаменитой пациентки видел главную причину ее долголетия в «счастливом» характере

Таблица 28

Документально подтвержденные рекорды по продолжительности жизни

Страна	Возраст	Имя	Родился	Умер
Франция	122	Жанна Кальман	23 02 1875	4 08 1997
Япония	120	Сигечио Идзуми	29 06 1865	21 02 1986
США	114	Марта Грэм	12 1844	25 06 1959
Великобритания	114	Марта Элиза Уильямс	2 06 1873	2 06 1987
Канада	113	Пьер Жубер	15 07 1701	16 11 1814
Испания	112	Жозефа Салас Матео	14 06 1860	27 02 1973
Франция	112	Августина Тессье	2 10 1869	9 03 1981
Марокко	112	Эль Хадж Мохаммед Эль Мокри (Великий Визирь)	1844	16 09 1957
Польша	112	Росалия Милщарак	1868	7 01 1981
Ирландия	111	Кэтрин Планкет	22 11 1820	14 10 1932
СССР	110	Хасако Дзугаев	7 08 1860	авг 1970

Интенсивный рост численности населения на Земле привел к возникновению различных теорий об угрозе перенаселения «Крестным отцом» многих теорий, посвященных решению демографических проблем в XX веке, является английский священник и экономист Томас Мальтус, опубликовавший в 1798 году свой памфлет «*Essay on the Principles of Population*». Согласно Мальтусу, человечеству грозит катастрофа, поскольку народонаселение растет в геометрической прогрессии («плодится с безответственностью трески»), а средства существования — лишь в арифметической прогрессии

Эти идеи нашли неожиданное развитие с появлением трудов Ч Дарвина, когда разработанные им принципы «естественного отбора» и «борьбы за существование» в животном мире перенесли на человечество. Любой слабый, нездоровый человек воспринимался как «элемент», подлежащий «естественному отбору» Войны, голод, эпидемии расценивались как дар природы, как естественный демографический регулятор

Евгеника, во многом базирующаяся на мальтузианстве и дарвинизме, предлагала активное решение демографических проблем путем ограничения рождаемости и улучшения рас селекцией человека. Френсис Гальтон (двоюродный брат Ч Дарвина), ав-

тор этой науки, определял евгенику как науку об «улучшении рода». «Слабые нации мира неизбежно должны уступить дорогу более благородным вариететам человечества .» Тем не менее евгеника по Ф Гальтону носила чисто академический характер Сам Ф Гальтон никогда не был сторонником рабства или уничтожения «низших рас» Однако очень быстро евгеника начала терять свое первичное академическое и гуманистическое начало

У истоков идей «активного регулирования и улучшения рас» стояли две леди. англичанка Мэри Степс и американка Маргарет Сангер Обе требовали ввести обязательную стерилизацию женщин, «неполноценных» для воспроизведения потомства, и были ярыми сторонницами политики сегрегации Они предлагали организовать специальные закрытые фермы, где изолированные от общества второсортные люди должны работать под наблюдением компетентных специалистов (Сам Мальтус считал единственно возможным, «не греховным» способом ограничения прироста населения — откладывание вступления в брак)

Наиболее чудовищное воплощение этих идей на практике произошло в фашистской Германии и полпотовской Кампучии Там миллионами истребляли «неполноценных» в концентрационных лагерях, умерщвляли умственных и физических инвалидов, стерилизовали женщин и мужчин Справедливости ради надо сказать, что первый закон о принудительной стерилизации был принят в США в 1907 году в штате Индиана, так называемый «индианский» закон. Такими же законами до 1937 года «обзавелись» еще 26 штатов В соответствии с этими законами в США было насильно стерилизовано свыше 100 тыс человек, в том числе много негров. Как правило, стерилизации подлежали психически ненормальные, умственно отсталые, осужденные за половые преступления В некоторых штатах, кроме того, — хронические алкоголики, эпилептики, проститутки, круглые сироты, бродяги, умственно отсталые Только в 1935 году в США было проведено 21 539 операций по стерилизации умственно отсталых Аналогичные законы в 30-х годах XX века принимались и в других государствах (Норвегия, Швеция, Дания, Финляндия, Эстония, Швейцария, Англия, Бермуды, Канада, Мексика, Япония)

После второй мировой войны поборники идеи активной борьбы с перенаселением планеты стали разрабатывать новые программы активного воздействия на воспроизводство населения В том числе программы, целью которых являлось разрешение абортов, разработка фармацевтических средств предотвращения беременности, воспитание нового отношения к сексу, семье, рождению детей

Согласно одной из концепций современных неомальтузианцев, в сексуальной активности человека нет нравственной составляющей, нет ни добра, ни человеческой привязанности, лишь животный оргазм Такой подход снимает все внутренние ограничения, легализуя и фактически поощряя те варианты сексуальных контактов, которые не приводят к беременности: оральный, анальный, мастурбация и т п Этому подходу соответствует легализация гомосексуализма и порнографии Представления о сексе как составной части неделимой семейной жизни расшатывались и в ходе так называемой «сексуальной революции» 70—80 годов XX века, которая во многом была спровоцирована мальтузианством.

Вместе с тем, угрозу перенаселения опровергли многие ученые Опровергает эту угрозу и фактически складывающаяся ситуация с воспроизводством народонаселения планеты к концу XX века Например, в Китае и Индии, считающихся самыми «проблемными» из-за большой численности населения, плотность населения к концу XX века была такая же, как в Англии, и в 20 (!) раз ниже, чем в Гонконге В Южной Корее плотность населения была в 4 раза больше, нежели в Китае, а на Тайване — в 5 раз

По данным Колина Кларка, высокотехнологичные методы ведения сельского хозяйства еще 10 лет назад позволяли прокормить 35,1 млрд человек. Это если придерживаться американского типа питания Если же взять за основу менее дорогостоящую диету (например, японскую), то хлеб насущный могли бы обрести 105 млрд!

За 150 лет (с 1750 г по 1900 г) численность населения Земли выросла в 2 раза, а добыча энергетических ресурсов в мире за это время возросла в 10 раз За следующие 70 лет (с 1900 г. по 1970 г.)

численность населения Земли выросла еще в 2 раза, а добыча энергетических ресурсов — в 13 раз, т е темпы производства энергии намного превышали темпы роста населения

В целом, уже с начала XX века на планете стал проявляться закон саморегуляции больших человеческих общин В 2/3 стран к концу столетия он уже действовал на полную мощность. Суть саморегуляции заключается в том, что при определенном уровне благосостояния в обществе наблюдается так называемый демографический переход, когда за резким снижением смертности через некоторое время, по мере экономического развития страны, следует снижение рождаемости

Великобритания, например, прошла через этот рубеж еще в прошлом веке Такие страны, как Таиланд, Индонезия и Бангладеш приближаются к этой черте Причем происходит это автоматически, даже без всякого вмешательства правительства, как в Китае По прогнозам демографов, прирост населения там прекратится к 2015 году Примерно в это же время перестанет увеличиваться население и в Индии, где правительство фактически не занимается вопросами контроля над рождаемостью

По прогнозам к середине XXI века человечество достигнет своей максимальной численности — 9,5 млрд человек После этого прирост населения на планете навсегда прекратится

3.4. Планирование медико-биологического эксперимента с малым числом наблюдений

Использование классических приемов сплошного исследования или выборки из генеральной совокупности, гарантирующих репрезентативность результатов, часто не возможно по различным причинам Среди этих причин одна из главных — незнание размеров и структуры генеральной совокупности (всех больных данным заболеванием, всех, у кого есть тот или иной симптом и т п) Кроме того, из-за финансовых, организацион-

ных или других причин клинические и лабораторные исследования обычно отличаются малым числом наблюдений Правильная организация, которая способна обеспечить получение статистически достоверных, репрезентативных данных, является ключевой проблемой таких исследований.

Необходимость особенно четкой организации малых по численности наблюдений объясняется и тем, что возможность «потом поправить» (отбросить сомнительные, «выскакивающие» варианты, перегруппировать их и т п) в таких исследованиях весьма ограничена Ограничено здесь и обычно проявляющееся при достаточно большом числе наблюдений взаимное погашение случайных, непредвиденных факторов

Решить проблему малых групп позволяет соблюдение ряда ключевых правил, заведомо обеспечивающих минимальную вероятность получения нерепрезентативных данных. Один из методов наблюдения, опирающийся на такие правила, называется *методом копий-пары* А выборка, полученная таким путем, называется *парно-сопряженная выборка.*

Метод копий-пары предусматривает формирование собственно двух групп группы наблюдения и контрольной группы Эти группы уравновешиваются структурно по основным признакам, способным оказывать существенное влияние на результат Не уравновешенным остается главный фактор, действие которого изучается *Например*: при изучении токсичности какого-либо вещества, берутся две группы подопытных животных Эти группы генетически однородны (по возможности), животные в них одного возраста, пола, они одинаково содержатся и т. д Отличаются только тем, что опытная группа подвергается воздействию изучаемого фактора (токсического вещества), а контрольная группа — нет Этот способ формирования выборки позволяет обходиться малым числом наблюдений. Однако в практике тех исследований, которые проводятся среди людей, простой подбор копий-пар трудно осуществим

Р А. Фишер (1950) дал знаменитое толкование правил проведения экспериментов с малым числом наблюдений. В качестве примера он рассмотрел гипотетический случай, когда некой английский леди было предложено провести оценки, что было ра-

ньше налито в чашку, — чай или молоко? По мнению Фишера, такое исследование должно строиться с соблюдением следующих правил (цит по Э Ллойду и У Ледерману, 1989):

- *повторяемость* (дублируемость) Нельзя делать каких-либо выводов о верной или ошибочной идентификации порядка смешивания молока и чая по одной единственной чашке;
- *чувствительность* Р Фишер отмечал, что пока число чашек не превысит некий минимум, никаких разумных выводов делать нельзя, поскольку выборка слишком мала;
- *сбалансированность* Леди должна была попробовать равное число чашек с молоком, добавленным в чай, и с чаем, добавленным в молоко, чтобы в ее суждениях не возникло смещения;
- *рандомизация* (случайность). Относится к тому, в каком порядке следует поставлять чашки на дегустацию Рандомизация есть на самом деле необходимое условие для того, чтобы стало возможным использование статистического анализа,
- *однородность* Изложенные выше соображения нельзя распространять слишком далеко Утверждаемое различие может вызываться разностью температур (чай может остыть), эффектом настаивания чая, усталостью леди, ее насыщение чаем и т п

Следование указанным правилам в определенной степени позволяет решить проблемы формирования малых выборочных групп при неизвестных генеральных совокупностях

При известной генеральной совокупности можно попытаться целенаправленно сформировать выборочную группу путем уравновешивания факторов, которые явно исказят результат статистического анализа. *Например* при углубленном обследовании работников предприятия удалось провести полное обследование только 187 рабочих Одним из самых существенных факторов, оказывающих влияние на результат этого обследования, могло быть не соответствие возрастно-половой структуры обследованной группы (выборочная совокупность) возрастно-половой структуре всего состава работников предприятия (генеральная совокупность)

Для того чтобы поверить указанное соответствие следует выполнить следующие действия

1 Получить таблицу фактического распределения работников предприятия по возрасту и полу (генеральная совокупность), табл 29

Таблица 29

Распределение работников обследованного предприятия по возрасту и полу (в % от числа всех работающих)

Пол	Возраст (лет)			
	до 20	20–29	40 и старше	Итого
Мужской	5,9	9,9	9,9	25,7
Женский	10,9	11,9	51,5	74,3
Оба пола	16,8	21,8	61,4	100,0

2 Составить таблицу распределения обследованных работников по возрасту и полу (выборочная совокупность), табл 30

Таблица 30

Распределение выборочной группы по возрасту и полу (в % от численности выборки)

Пол	Возраст (лет)			
	до 20	20–29	40 и старше	Итого
Мужской	4,4	11,0	6,6	22,0
Женский	13,2	16,5	48,4	78,0
Оба пола	17,6	27,5	54,9	100,0

3 Проверить с помощью критерия Пирсона χ^2 распределение показателей структуры выборочной группы и показателей структуры генеральной совокупности (Подробно методику расчетов см в разделах «Критерии различий эмпирических распределений» и «Оценка различий эмпирических распределений с помощью *Excel*»)

4 В данном случае различия возрастно-половой структуры генеральной и выборочной совокупностей можно признать не существенными (P =0,123)

Обратить внимание: в данном случае речь идет только о соответствии (или не соответствии) показателей **структуры распределения** Не путать с оценкой соответствия распределения!

Аналогичным образом можно проверить соответствие структуры выборочной и генеральной совокупностей и по другим параметрам (стаж работы, производственные вредности и т д) В случае необходимости, можно скорректировать состав выборочной группы, т е добавить несколько единиц наблюдения с заведомо заданными, нужными характеристиками или, наоборот, отбросить.

Следует отметить, что требования к чистоте эксперимента, с точки зрения его статистической репрезентативности, и, соответственно, правомерность тех или иных выводов, могут иметь в каждом конкретном случае свои особенности

4. Основы математико-статистической обработки данных

Математическая статистика — раздел статистики, посвященный математическим методам систематизации, обработки и анализа статистических данных Методы математической статистики основаны на вероятностной природе этих данных Нередко термин «статистический характер данных» используется для того, чтобы подчеркнуть вероятностный, случайный характер тех величин, с которыми оперирует математическая статистика

Вероятностный характер медико-биологических данных вызывает и вероятностный, т е сопряженный с той или иной неопределенностью, характер заключений, будь это выводы исследований, охватывающие большие по численности группы, или суждение о здоровье отдельного человека

4.1. Относительные величины. Статистические коэффициенты

Уже на этапе сводки и группировки статистических данных вычисляются промежуточные итоги в виде абсолютных величин Среди них суммы и простейшие средние, а также относительные величины, называемые иначе статистическими коэффициентами.

Абсолютные величины могут быть простыми, которые всегда представляются в именованных единицах измерения (сантиметры, килограммы, дни и т п), или сложными, которые выражаются произведениями единиц различной размерности (тонно-километры, человеко-часы и т п) Кроме того, в ряде случаев используются *условные величины Например*, энергозатраты на одну единицу выполненной работы Использование таких величин позволяет преодолевать несравнимость простых абсолютных показателей из-за их разнородности

Относительные величины (статистические коэффициенты) широко используются в официальной статистике для оценки медико-демографической и санитарно-эпидемиологической ситуации, оценки деятельности медицинских учреждений и т. п. Вычисление и анализ этих коэффициентов является основой медицинских исследований, проводимых на уровне больших групп населения, населенных пунктов, городских и сельских районов, областей и регионов

Относительной статистической величиной в наиболее общем виде называется отношение двух чисел, выражающих меру каких-либо явлений Смысл получения относительных величин — нахождение общей меры, приведение к общему знаменателю *Например* численность врачей в Санкт-Петербурге в 1990 году составляла 32 671 чел , а в 1996 году — 31 549 На основании этих данных можно сделать вывод о снижении обеспеченности врачами населения города на (32 671—31 549) 1122 врача. Однако за указанный период времени, с 1990 по 1996 год, численность жителей города уменьшилась, соответственно, с 5 035 200 до

4 801 500 человек. Если теперь пересчитать обеспеченность врачами в относительных показателях, то окажется, что обеспеченность врачами населения не уменьшилась, а возросла с 65,3 до 66,5 врачей на 10 000 человек населения города

Среди относительных величин наибольшее практическое значение имеют: интенсивные коэффициенты, экстенсивные коэффициенты, показатели соотношения, показатели наглядности, показатели относительной интенсивности

Интенсивные коэффициенты показывают интенсивность развития (частоту, уровень, распространенность) явления в своей среде В среде, которая продуцирует это явление. Применяются интенсивные коэффициенты, за редким исключением, только в медицине и демографии (или на стыке наук — медицинской демографии). Эти коэффициенты отвечают на вопрос, как часто явление встречается в известной среде Различают *общие и специальные интенсивные* коэффициенты

Общие — характеризуют общую (усредненную) интенсивность явления, ***специальные*** — дают более детальную характеристику явлению. Вычисление всегда начинают с обших показателей Расчет производится через пропорцию

Например: Во Владимирском районе в 1996 году зарегистрировано 1800 случаев инфекционных заболеваний, в том числе 60 случаев инфекционного гепатита Всего в районе в 1996 году проживало 60 тыс. человек

1800 случаев заболеваний пришлось на 60 000 человек

X случаев заболеваний — на 1000 человек

Уровень инфекционной заболеваемости во Владимирском районе в 1996 году (общий показатель)=(1 800/60 000)×1000=30,0 случаев на 1000 жителей за 1996 год.

В данном примере явление, частоту которого определяем через интенсивные показатели, это случаи инфекционных заболеваний. Среда — жители района Следует обратить внимание на то, что случаи заболевания и численность жителей района берутся за один и тот же год Одной из самых грубых ошибок, допускаемых при вычислении статистических коэффициентов, является несоблюдение единства времени для исходных данных В

данном примере случаи заболеваний и численность населения могут быть взяты только за один и тот же 1996 год

Аналогичным образом вычисляются специальные показатели *Например* уровень заболеваемости инфекционным гепатитом во Владимирском районе (специальный показатель) = (60/60 000)×1000=1,0 случай инфекционного гепатита на 1000 жителей района в 1996 году

Выбор множителя для показателя обусловлен удобством пользования результатами вычислений Если в этом примере считать на 100 000 населения, то получим 100 случаев гепатита на 100 000 населения, если на 100 — 0,1 случая на 100 человек В данном примере разумнее всего в качестве множителя использовать 1000, т к при этом получается не громоздкое, но целое число Иногда для обозначения множителя используют сокращенные обозначения Если показатель вычислялся на 100 — проценты (%), если — на 1000 — промилле (‰), на 10 000 — продецимилле (‱) и т д При этом интенсивный показатель всегда остается величиной именованной (случаи заболеваний, рождений, смертей и т п)

Одной из особенностей интенсивных коэффициентов является невозможность их прямого сложения Простое сложение интенсивных коэффициентов допустимо только в особых случаях, о которых здесь не будет упомянуто

Например: среди жителей одного из микрорайонов была проведена профилактическая вакцинация Прошли вакцинацию 1420 человек, не прошли — 850 Заболеваемость среди жителей, прошедших вакцинацию, составила 12,7‰ Среди жителей, не прошедших вакцинацию, — 87,1‰. Требуется определить суммарный уровень заболеваемости среди жителей микрорайона

Таблица 31

Суммирование интенсивных коэффициентов

Отношение к вакцинации	Число жителей	Заболеваемость	
		абс	на 1000 чел
Привитые	1420	18	12,7
Не привитые	850	74	87,1
Итого	2270	92	40,5

Если определять суммарный уровень заболеваемости путем простого сложения интенсивных показателей (12,7‰+87,1‰= 99,8‰), то итоговый показатель будет завышен более чем в два раза Правильное вычисление сумм интенсивных показателей производится следующим образом: сначала суммируются исходные абсолютные числа (1420+850=2270 и 18+74=92), затем на основании этих сумм вычисляется итоговый показатель (92/2270×1000=40,5‰).

Экстенсивные коэффициенты отражают структуру, распределение Они характеризуют отношение части статистической совокупности к целой совокупности (долю, удельный вес, часть от целого), то есть отношение отдельного элемента к итогу Выражаются только в процентах к итогу

Например, в структуре инфекционной заболеваемости жителей Владимирского района в 1996 году доля инфекционного гепатита среди всех инфекционных заболеваний составила (число случаев инфекционного гепатита/число всех случаев инфекционных заболеваний)×100=3,3%

На графике представлена динамика показателей структуры (распределения) больных хроническими неспецифическими заболеваниями легких — ХНЗЛ (рис 50). Группа А — здоровые,

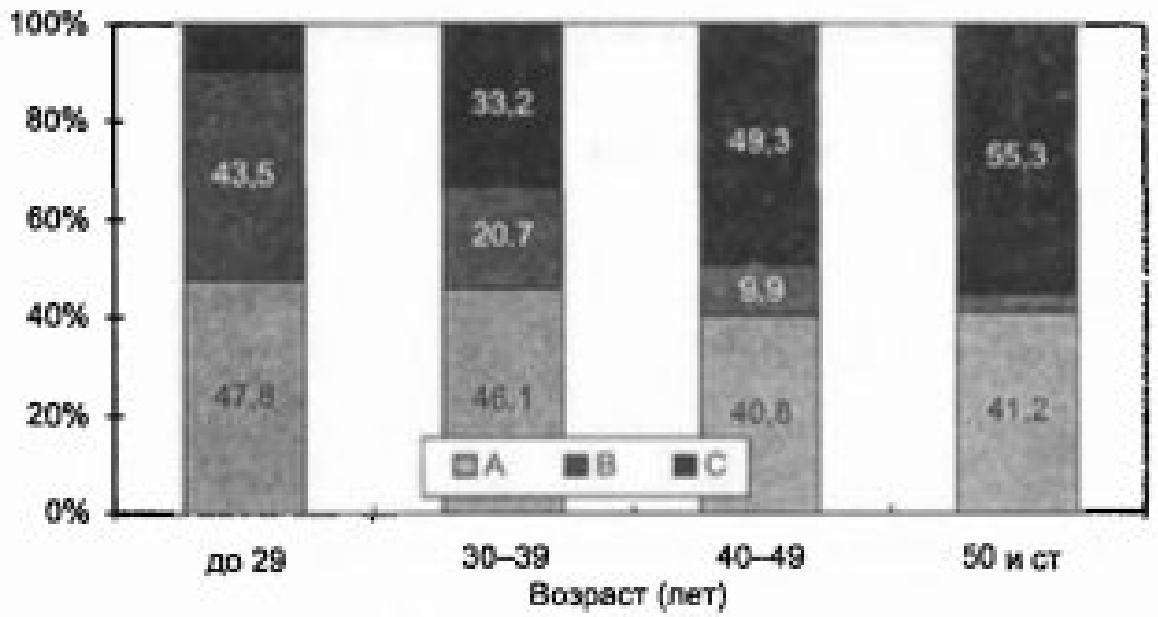


Рис 50 Структура пульмонологических групп

группа В — лица с начальными, обратимыми признаками заболевания, С — группа со сформировавшейся хронической патологией Из представленных данных наглядно видно, что доля (удельный вес) здоровых мало менялась на протяжении времени

Одной из самых распространенных ошибок, встречающихся в практике статистического анализа, является ошибочное использование интенсивных и экстенсивных коэффициентов В частности, по экстенсивным коэффициентам пытаются судить о величине или частоте явления *Например*:

Таблица 32

Доля затрат на здравоохранение в структуре расходов городского бюджета в С.-Петербурге

Год	Бюджет (расходы млн руб)		Доля на здравоохранение (%)
	На здравоохранение	Все бюджетные расходы города	
1994	488 100	3 839 064	12,7
1995	1 032 904	8 656 070	11,9
1996	1 224 200	11 923 689	10,3

По приведенным данным (табл 32) нельзя утверждать, что расходы на здравоохранение снизились в 1996 году по сравнению с 1994 годом (12,7% и 10,3%, соответственно) На самом деле, эти расходы увеличились почти в три раза с 488 100 млн руб. до 1 224 200 млн руб (инфляция в данном случае не покрывает всего прироста) Снизилась только доля расходов на здравоохранение в городском бюджете, по-видимому, из-за более интенсивного роста других расходов.

Методика вычисления коэффициента летальности (интенсивный коэффициент): среда — раненые, явление — гибель раненых Из 500 раненых, у которых ранение осложнилось газовой инфекцией, умерло 70 человек Или (70/500)×100=14%. Из всех 400 умерших из-за осложнений газовой инфекции погибло 70 человек, или (70/400)×100=17,5% (экстенсивный коэффициент — все умершие — 100%, часть из них погибла от газовой инфекции) и т. д (табл 33).

Таблица 33

Летальность при раневых осложнениях (числа условные)

Название осложнений	Число лечившихся (абс)	Число умерших (абс.)	На 100 раненых умерло (летальность в %)	Умерло раненых (в % к итогу)
Газовая инфекция	500	70	14,0	17,5
Столбняк	35	25	71,4	6,5
Шок	250	75	30,3	18,7
Кровопотеря	150	42	28,0	10,5
Сепсис	300	144	48,0	36,0
Остеомиелит	4000	20	0,5	5,0
Пневмония	200	24	12,0	6,0
Итого	5435	400	7,4	100,0

Если расположить полученные данные по убыванию интенсивных коэффициентов летальности, то получим следующее распределение, табл 34.

Таблица 34

Порядковое распределение величин коэффициентов

Название осложнений	Интенсивные (летальность)	Экстенсивные коэффициенты
Столбняк	1	5
Сепсис	2	1
Шок	3	2
Кровопотеря	4	4
Газовая инфекция	5	3
Пневмония	6	7
Остеомислит	7	7

Из представленных данных видно, что чаще всего (интенсивность явления) раненые погибали от столбняка Но среди всех умерших (часть от целого, доля, удельный вес) раненые с этим осложнением были на пятом месте

Коэффициенты наглядности — используются для облегчения сравнения и повышения наглядности Не изменяя по существу отношений между числами, они дают более отчетливое представление о характере изменения явления во времени Выражаются коэффициенты наглядности в процентах, которые вычисляют от исходного уровня, принимаемого за 100%

Поскольку эти коэффициенты являются неименованными величинами, их можно использовать для сравнения числовых рядов, которые состоят из разнородных величин, а также рядов из абсолютных и относительных чисел, табл 35

Таблица 35

Некоторые характеристики деятельности учреждений здравоохранения Санкт-Петербурга

Год	Число лиц, которым оказана внебольничная помощь		Всего операций в стационарах		Уровень госпитализации	
	Абс	Коэффициент наглядности	Абс	Коэффициент наглядности	На 1000 населения	Коэффициент наглядности
1992	675259	100,0	283100	100,0	170,9	100,0
1993	680057	100,7	274668	97,0	176,0	103,0
1994	589560	87,3	261651	92,4	170,5	99,8
1995	586180	86,8	252523	89,2	166,7	97,5
1996	561692	83,2	246340	87,0	161 8	94,7

За исходный уровень можно выбрать любой уровень, который имеет большее смысловое значение

Коэффициенты относительной интенсивности применяются, когда невозможно получить прямые интенсивные коэффициенты или когда необходимо измерить степень диспропорции в структуре двух или нескольких близких процессов В частности, эти коэффициенты используются в статистике здравоохранения, когда нет точных исходных данных о составе населения *Например* известно, что среди всех жителей района, которые обратились за медицинской помощью в связи с полученными травмами, мужчины составили 49%, а женщины — 51% Для того чтобы сделать заключение о более частом травматизме мужчин, необходимо рассчитать число случаев травм на 1000 мужчин и, соответственно, 1000 женщин (интенсивные коэффициенты) Однако для этого необходимо точно знать число мужчин и женщин, проживающих в районе В случае применения коэффициентов относительной интенсивности можно ограничиться только учетом структуры населения по полу

Таблица 36

Относительная частота обращений по поводу травм

Пол	Обратились по поводу травм (%)	Состав населения (%)	Коэффициенты относительной интенсивности
	A	B	A/B
Женщины	49	40	1,23
Мужчины	51	60	0,85
Оба пола	100	100	–

Из приведенных данных (табл 36) можно косвенным путем сделать заключение о более высоком травматизме женщин, хотя их доля среди обратившихся за медицинской помощью по этому поводу была несколько меньше

Таблица 37

Продажа алкогольных изделий в СССР (в % к 1940 году)

Годы	1958	1964	1965	1966	1967
Продажа %	166,0	261,0	279,0	306,0	351 0
Насселение %	105,5	116,5	118,1	119,4	120,7
Коэффициент относительной интенсивности	1,5	2,2	2 3	2,5	2,9

Коэффициент относительной интенсивности, представленный в табл 37, характеризует опережающий рост потребления алкоголя над ростом численности населения

Коэффициенты соотношения — применяются, когда приходится оценивать взаимосвязь разнородных величин *Например* обеспеченность населения больничными койками, соотношение средних медицинских работников и врачей, обеспеченность населения врачами и т д Коэффициенты соотношения, как и интенсивные коэффициенты, вычисляются через пропорцию Могут вычисляться на 100, на 1000, на 10 000 В отличие от интенсивных коэффициентов могут выражаться дробными числами, в которых дробная часть содержит одинаковое или большее количество значащих цифр, чем целая 1,53 медсестры на 1 врача (табл 38)

Таблица 38

Динамика показателей системы здравоохранения С -Петербурга

Показатели соотношения	1994	1995	1996
Обеспеченность больничными койками (на 10 тыс населения)	99,5	98,5	97,0
Обеспеченность средним персоналом (на 10 тыс населения)	98,4	99,4	102,7
Обеспеченность врачами (на 10 тыс населения)	61,9	63,1	66,5
Число средних медработников, приходящихся на 1 врача	1,53	1,62	1,54

4.2. Показатели описательной статистики

Решение научных и практических задач с помощью методов математической статистики связано с обязательным последовательным решением следующих вопросов

- установление закона распределения эмпирических (полученных опытным путем) статистических совокупностей и параметров этого распределения (свойств эмпирических совокупностей);
- числовая оценка причинно-следственных отношений и взаимосвязи между явлениями;
- решение проблем, связанных с репрезентативностью (представительностью) выборочных исследований и точностью статистического прогноза.

Одной из основных задач математико-статистической обработки является нахождение параметров, представляющих в обобщенном виде распределение данной статистической совокупности. Для решения этих задач используются методы описательной статистики (табл 39).

Примечание. Приведенные в таблице условные обозначения статистических показателей и статистических критериев не являются общепринятыми или стандартными *Например,* среднее арифметическое может обозначаться символами M или $\bar{X}$, коэффициент вариации C_V или V и т д Таким образом, использование условных обозначений в статистических работах обязательно должно сочетаться с пояснениями их смыслового значения.

Таблица 39

Статистические показатели распределения

Показатели	Назначение показателя	Примеры показателей
Средние величины	Описывают положение середины распределения	*Степенные средние* ▪ среднее арифметическое ▪ среднее гармоническое ▪ среднее квадратическое ▪ среднее геометрическое *Структурные средние* ▪ мода *Mo* ▪ медиана *Me*
Показатели разброса	Описывают степень разброса (вариабельности, изменчивости) данных	Лимит — *Lim* Амплитуда — *Ampl* Дисперсия — *D* Среднеквадратическое отклонение — σ Коэффициент вариации — *V* Квантили
Показатели формы распределения	Отвечают на вопрос о симметрии и островершинности распределения данных около центра	Коэффициент асимметрии — *As* Эксцесс — *E* Гистограмма Полигон распределения

4.3. Ряды распределений. Вариационные ряды

Значительную долю статистических данных составляют количественные признаки, принимающие некоторое числовое значение у каждой единицы статистического наблюдения Эти числовые значения выражаются в виде различных вариант Например: в качестве статистической совокупности рассматривается группа студентов вуза Каждый студент — отдельная единица наблюдения Если нас интересует вопрос о каком-либо показателе физического развития (учетном признаке единицы наблюдения), скажем, массе тела, то масса каждого

студента является в данном случае вариантой. Масса тела колеблется, варьирует от одного студента к другому. У первого студента масса тела 51 кг, у второго — 67 кг и т д Таким образом, варьирующий признак встречается в различных вариантах

Варьирующие признаки (варианты) могут быть двух видов — ***прерывные и непрерывные.***

Прерывный, или ***дискретный, признак*** — признак, принимающий конкретные значения в виде целых конечных чисел, между которыми нет промежутков *Например* число ударов пульса, число дней госпитализации и т п Прерывный признак всегда является результатом счета

Непрерывный — это признак, варианты которого могут принимать любые значения в некоторых пределах и выражаются лишь приближенно, с определенным приближением (точностью). Получаются эти признаки в результате измерения и могут выражаться дробно: вес, рост, длина и т д.

Первым шагом статистического анализа является построение *ряда распределения* Строго говоря, при использовании современной вычислительной техники и специальных программ статистической обработки данных, построение рядов распределения, а также проведение других вспомогательных операций почти полностью исключено Однако ключевые понятия и термины, используемые в процессе статистического анализа и при толковании отдельных результатов, могут быть поняты только на основе усвоения логической последовательности основных «ручных» операций статистической обработки данных

При наблюдении какого-либо варьирующего признака, так или иначе, ведется регистрация полученных значений *Например,* масса тела у обследованных студентов составила: 64, 57, 63, 62, 57, 61, 61, 59, 60, 60, 63, 59, 62, 59, 64, 60, 59, 60, 60, 60, 60, 63, 60, 59, 59, 61, 61, 58, 61, 61, 65, 61, 61, 58, 64, 62, 62, 60, 62, 62, 62, 58, 63, 63, 59, 60, 58, 63, 58, 60, 64, 63, 58, 61, 57 кг.

Здесь числа расположены в порядке регистрации данных Такой ряд называется *неупорядоченным рядом* отдельных наблюдений (рис. 51)

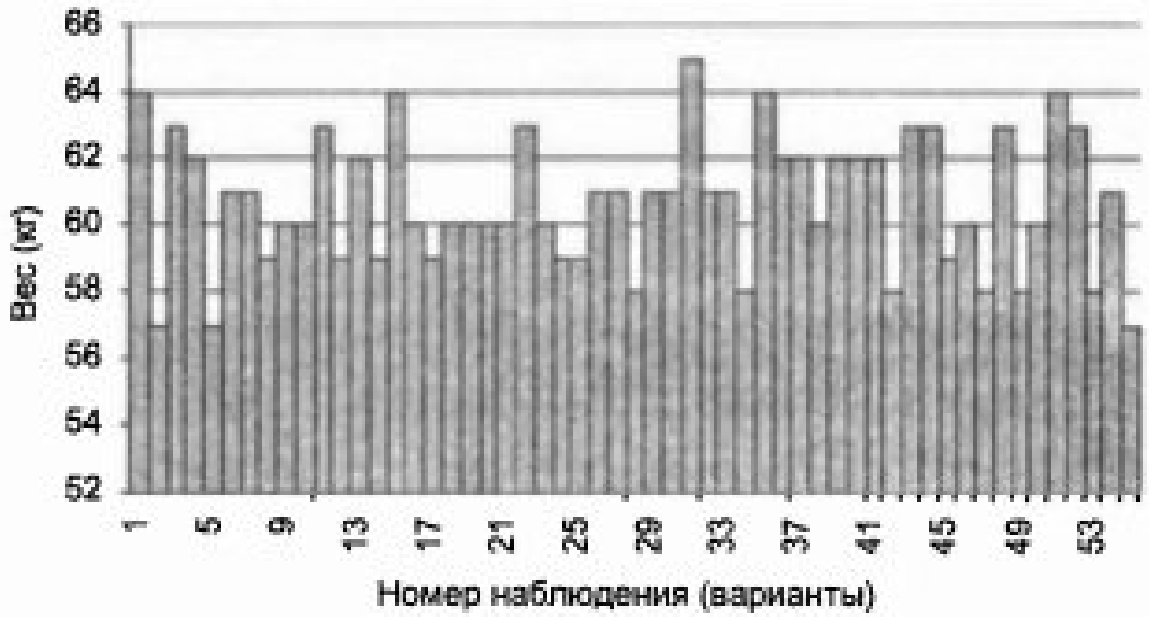


Рис. 51. **Неупорядоченный ряд исходных данных**

Началом статистического анализа числовых рядов является их упорядочение, *ранжирование*, в возрастающем или убывающем порядке 57, 57, 57, 58, 58, 58, 58, 58, 58, 59, 59, 59, 59, 59, 59, 59, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 61, 61, 61, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 65

В *ранжированном* ряду каждый отдельный случай еще сохраняет свою индивидуальность Более компактной формой описания вариации является образование рядов распределений, которые состоят из групп с одинаковыми или близкими значениями варьирующего признака По своей конструкции ряд распределения состоит из двух столбцов (*граф*). В одном столбце располагаются *варианты* (*V*), в другом — *частоты* (*P*) Частоты указывают, сколько раз встречаются одинаковые значения признака в этом ряду, т. е сколько студентов имели одинаковый вес (табл. 40)

Показатели частостей выражаются в относительных единицах — процентах от общего числа наблюдений или долях от единицы *Частости* показывают долю частот отдельных вариант от общего числа наблюдений.

Таблица 40

Пример вычисления накопленных частот и частостей

Вес V (кг)	Частоты (P)		Частости		
	Число студентов (частоты)	Накопленные частоты	%	p	Накопленные частости
57	3	3	5,5	0,05	0,05
58	6	9	10,9	0,11	0,16
59	7	16	12,7	0,13	0,29
60	11	27	20,0	0,20	0,49
61	9	36	16,4	0,16	0,65
62	7	43	12,7	0 13	0,78
63	7	50	12,7	0,13	0,91
64	4	54	7,3	0,07	0,98
65	1	55	1,8	0,02	1,00
	$n=\sum P=55$	–	100,0	1,00	–

Иногда бывает уместным осуществить еще одно преобразование вариационного ряда — построение ряда накопленных частот или частостей *Накопленные частоты и частости* позволяют при оценке распределений игнорировать неравную величину интервала в отдельных группах.

Если ряд распределения состоит из *дискретных величин*, то он называется *дискретным вариационным рядом* Наглядность тенденций распределений повышается при графическом представлении вариационных рядов. Графически дискретный вариационный ряд изображается в системе прямоугольных координат как многоугольник, так называемый ***полигон распределения*** (рис 52) По оси абсцисс откладываются различные возможные значения варьирующего признака (V), по оси ординат — частоты (P), число случаев Иногда в диаграмму включаются накопленные частоты или частости (см рис 53)

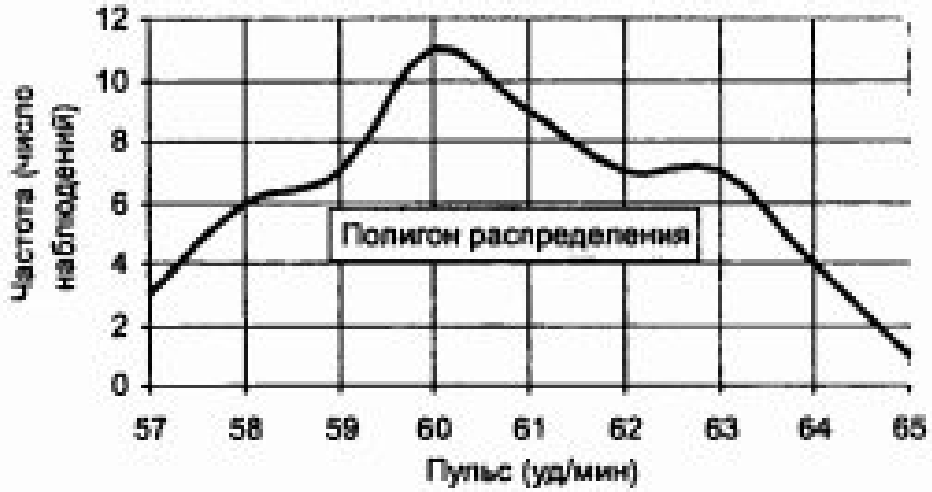


Рис. 52. **Полигон распределения**

Приведенная форма ряда распределения применима лишь для тех случаев, когда дискретный варьирующий признак принимает небольшое количество значений. Если таких вариант большое количество или бесконечно большое (в случае непрерывного ряда), то для каждой варианты образовать свою группу невозможно Объединение отдельных наблюдений в группы возможно лишь на базе *интервала (класса, разряда)*, т е в группы, имеющие определенные пределы значений. Эти группы образуют *интервальный сгруппированный вариационный ряд.* Графически такой ряд изображается **гистограммой** распределения (рис 53)

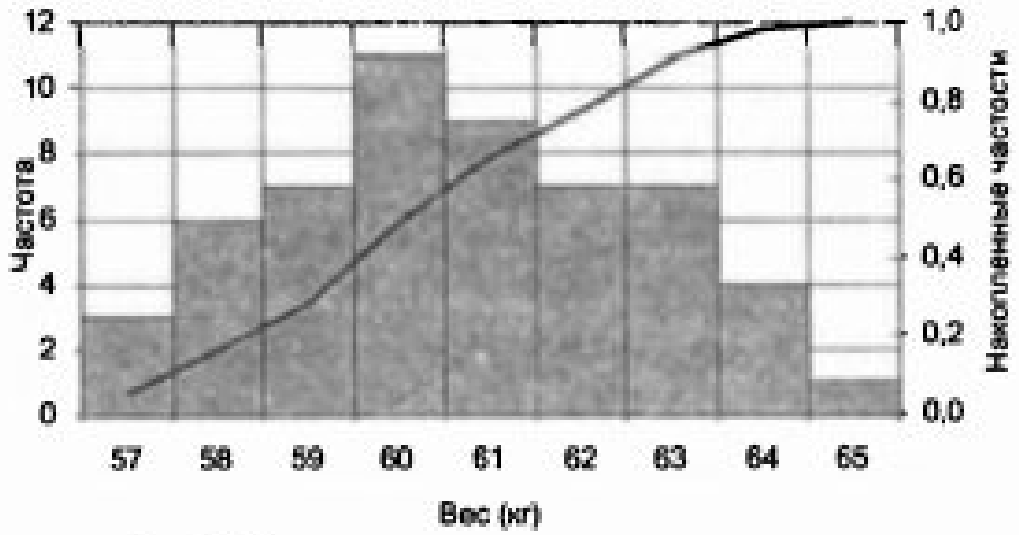


Рис 53 **Гистограмма распределения студентов по весу**

Всякая сводка или группировка уничтожает очертания отдельных единиц, растворяет их в группе, поэтому соблюдение правил формирования групп является необходимым условием сохранения основных тенденций распределения признака в сгруппированном ряду

В группах пределы обозначаются или подразумеваются «от» (верхняя граница) и «до» (нижняя граница) Желательно, чтобы интервалы во всех группах конкретного ряда были одинаковы Ряды распределений, где группировка данных проведена в неодинаковых интервалах, требуют применения специальных методик дальнейшей статистической обработки

Универсального ответа о величине интервала и, соответственно, о числе групп не существует. Этот вопрос решается отдельно в каждом конкретном случае Главное, чтобы характерные особенности распределения не были завуалированы, а не характерные, случайные колебания были бы сглажены Лучше допустить некоторую потерю в точности, но зато выиграть в наглядности, в аналитических возможностях

В принципе, *рекомендуется* руководствоваться следующими соображениями во-первых, число групп должно быть нечетным; во-вторых, желательно, чтобы при большом объеме наблюдений (более 100) число групп было больше (9—11—13), а при малом объеме — меньше (5—7—9). Если величина интервала берется равной для всех групп ряда, то размер интервала обычно устанавливается на основе крайних значений ранжированного ряда В этом случае, чтобы определить наиболее оптимальный интервал группировки, необходимо.

1. Найти разность между максимальным и минимальным значением вариант в ряду и разделить на число групп, которое хотят получить

2 Полученную в результате деления величину округлить и таким образом получить интервал

При незначительном разбросе вариант для определения интервала группировки можно воспользоваться формулой Стерджесса

$$i=\frac{V_{max}-V_{min}}{1+3{,}322\,Lg(n)},$$

где n — число наблюдений, V_{max} и V_{min} — соответственно, максимальное и минимальное значения вариант

Для подобных целей можно использовать и формулу, основанную на рекомендациях К Брукса и Н Краузерса

$$i=\frac{V_{max}-V_{min}}{5\times Lg(n)}.$$

Интервалы могут быть *открытыми* Такие интервалы имеют одну границу, либо верхнюю, либо — нижнюю *Например* «От 100 лет и более» Или «До 10 лет»

Закрытые интервалы имеют обе границы, нижнюю и верхнюю Соответственно формируются *открытые* или *закрытые вариационные ряды* (табл 41) Чтобы не возникало сомнений, в какую группу относится та или иная варианта, границы интервала (границы групп) не должны пересекаться, т е границы каждой группы должны отличаться от границ соседних групп *Например* если имеются группы 180—185, 185—189 см, то непонятно, в какую группу следует отнести варианту 185 см? Для исключения неопределенности, когда границы групп по каким-то причинам все же совпадают, делаются специальные оговорки («от» и «до») Оговорки однозначно указывают, в какую группу попадают пограничные значения вариант

Таблица 41

Примеры группировок вариант в вариационных рядах

Открытые ряды	Закрытые ряды*	Обозначения границ совпадают	Границы групп не совпадают
до 80	75—80	от 75 до 80	75—79
80—90	81—89	от 80 до 85	80—84
более 90	90—95	от 85 до 90	85—89

** В столбце приведен пример ряда с неравными интервалами в группах*

Иногда, в случаях неопределенности границ групп или неопределенности отдельной варианты, вызванной сомнениями в точности измерений, допускается использование половинных

частот *Например* если не понятно, куда отнести варианту 80 при групповых границах 75–80 и 80–85, то в обе группы добавляют по 0,5 частоты Варианта как бы делится между двумя группами (табл 42, табл 43)

Таблица 42

Исходный ряд

V	*P*
75–80	5
80–84	6
Итого	11

Таблица 43

После добавления одной варианты

V	*P*
75–80	5,5
80–84	6,5
Итого	12

В тех случаях, когда вариационный ряд представлен сгруппированным рядом распределения, для проведения дальнейшей статистической обработки (определение среднего взвешенного и т п) необходимо *определить середину* в каждой группе ряда. Если группа состоит из дискретных величин, то середина группы определяется как полусумма крайних значений группы Если ряд распределения непрерывный, то середина группы определяется как полусумма начальных вариант данной и последующей групп (табл 44)

Таблица 44

Примеры определения середины групп

Дискретный ряд		Непрерывный ряд	
Исходные группы *V*	Середина группы $V_{ср}$	Исходные группы *V*	Середина группы $V_{ср}$
150–151	(150+151)/2=150,5	150–151	(150+152)/2=151
152–153	(152+153)/2=152,5	152–153	(152+154)/2=153
154–155	(154+155)/2=154,5	154–155	(154+156)/2=155

Правильное применение интервалов позволяет построить компактный и наглядный *сгруппированный вариационный ряд*

Наиболее оптимальной из приведенных (табл 45–47) группировок является средняя группировка (табл 46)

Различные варианты группировок интервального ряда распределения женщин по росту

Таблица 45

Рост (см) *V*	Частота *P*
151,0–151,4	1
151,5–151,9	0
152,0–152,4	2
152,5–152,9	0
153,0–153,4	2
153,5–153,9	1
154,0–154,4	0
154,5–154,9	4
155,0–155,4	1
155,5–155,9	4
156,0–156,4	5
156,5–156,9	2
157,0–157,4	5
157,5–157,9	4
158,0–158,4	3
158,5–158,9	1
159,0–159,4	2
159,5–159,9	1

Таблица 46

Рост (см) *V*	Частота *P*
151	1
152	2
153	3
154	4
155	5
156	7
157	9
158	4
159	3

Таблица 47

Рост (см) *V*	Частота *P*
151–153	6
154–156	16
157–159	16

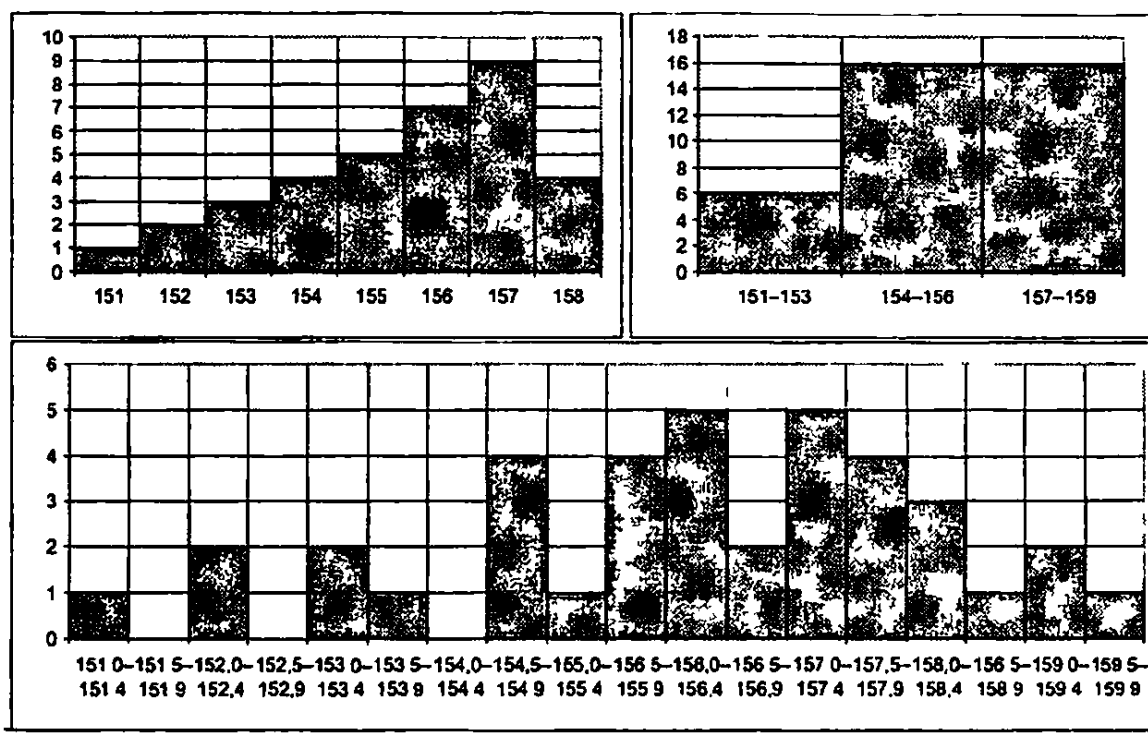


Рис. 54. **Графическое распределение группироаок из табл. 45–47**

Группировка, в которой величины интервалов завышены (см. табл 47), приводит к образованию крупных групп, в которых основные тенденции распределения теряются. Мелкая группировка (см. табл. 45) перегружает распределение частными деталями, не отражающими основных тенденций, что неизбежно затрудняет понимание характера вариации Кроме того, измерение роста с такой точностью, как правило, не производится, поскольку не имеет смысла

4.3.1. Построение вариационных рядов в *MS Excel*

В качестве примера создадим в *MS Excel* таблицу с результатами обработки измерений массы тела студентов. Для этого после запуска введите исходные данные (рис. 55)

	A	B	C	D	E	F
1	64	63	60	65	62	60
2	57	59	63	61	58	64
3	63	62	60	61	63	63
4	62	59	59	58	63	58
5	57	64	59	64	59	61
6	61	60	61	62	60	
7	61	59	61	62	58	
8	59	60	58	60	63	
9	60	60	61	62	58	
10	60	60	61	62	57	

Рис. 55. **Расположение исходных данных для построения вариационного ряда**

Чтобы выполнить поставленную задачу, перейдите на другой лист *MS Excel* . Для чего установите указатель мыши на надпись внизу окна таблицы **Лист2** и щелкните левой клавишей мыши Затем вызовите из меню <**Сервис**> прикладной пакет <**Анализ данных**>. В открывшемся окне <**Инструменты анализа**> выберите <**Гистограмма**>. Заполните поля в открывшемся окне как показано на рис 56

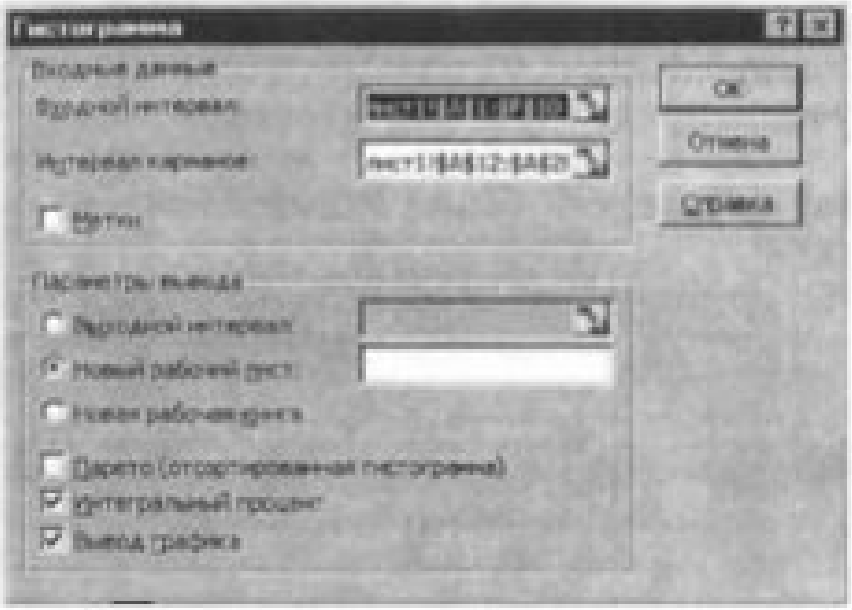


Рис. 56. **Вариант заполнения окна «Гистограмма»**

В окне (см. рис. 56) имеются следующие поля.

1. **Входной интервал**: содержит координаты ячеек электронной таблицы с исходными данными. Допускается и прямой ввод в это поле числовых значений анализируемого ряда.

2. **Интервал карманов**: содержит границы группировок (карманов) В данном случае числа 57, 58, 59, 60, 61, 62, 63, 64, 65 были введены заранее, последовательно в клетки первого листа (от A12 до A20). В процессе группировки все варианты наблюдения с весом 57 кг будут отнесены в группу 57 Все варианты с весом 58 кг будут отнесены в группу 58. Если варианты были бы представлены дробными числами, то варианта 58,8 кг, например, была бы отнесена в группу 58. Можно и не указывать границы группировок. В этом случае их подбор будет осуществлен автоматически.

3. **Метки**. отмечаются, когда в первой строке исходных данных имеются заголовки, это делается для того, чтобы они не включались в числовую обработку

4. В разделе **Параметры вывода** указывается место, куда будет выводиться результат В данном случае — на новый лист В этом же разделе указываются интервалы ячеек, в которых будут располагаться результаты.

5 **Парето**: группы, сортируются в порядке возрастания частоты.

6 **Интегральный процент** накопленные частости, выраженные в %.

7. **Вывод графика**: диаграмма распределения анализируемого ряда значений.

	A	B	C
1	*Карман*	*Частота*	*Интегральный %*
2	57	3	5,45%
3	58	6	16,36%
4	59	7	29,09%
5	60	11	49,09%
6	61	9	65,45%
7	62	7	78,18%
8	63	7	90,91%
9	64	4	98,18%
10	65	1	100,00%
11	Еще	0	100,00%
12			

Рис. 57. **Результат построения вариационного ряда**

Рядом выводится диаграмма полученного распределения (рис 58)

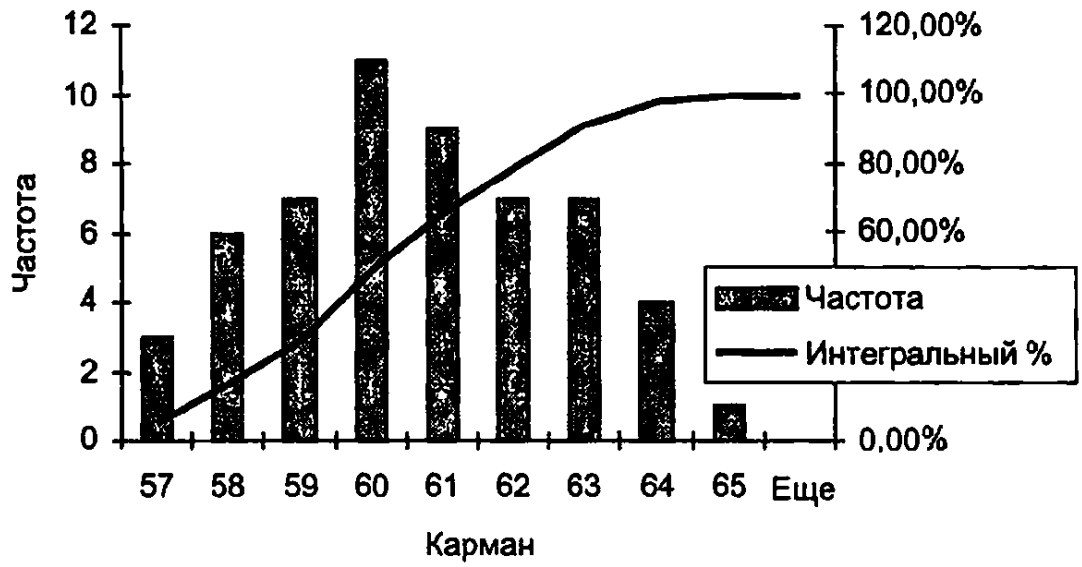


Рис 58. **Диаграмма распределения параметров сформированного вариационного ряда**

Внешний вид гистограммы и отдельные элементы экспликации дорабатываются пользователем в соответствии с его предпочтениями с помощью обычных приемов обработки графических изображений в *Excel*

4.4. Показатели центра распределения. Средние величины

Важным свойством статистической совокупности является положение центра ряда распределения. В примере (табл 48) представлены данные о распределении 80 мужчин и 33 женщин по росту и соответствующие полигоны частот (рис. 59).

Таблица 48

Распределение мужчин и женщин по росту

Рост (см) V	Рост (см) $V_{ср}$	Число		VP_1	VP_2
		женщин P_1	мужчин P_2		
150–151	151	1	1	151	151
152–153	153	3	4	459	612
154–155	155	5	8	775	1240
156–157	157	7	11	1099	1727
158–159	159	6	14	954	2226
160–161	161	5	17	805	2737
162–163	163	3	15	489	2445
164–165	165	2	9	330	1485
166–167	167	1	1	167	167
	Итого	$\sum P_1 = 33$	$\sum P_2 = 80$	$\sum VP_1 = 5229$	$\sum VP_2 = 12790$

Из этих данных видно, что кривые распределения показателей у мужчин и женщин имеют отличия Они связаны с различным числом наблюдений и поэтому имеют разные высоты кривых, а также сдвинуты одна относительно другой Параметры распределения, с помощью которых можно оценить величину

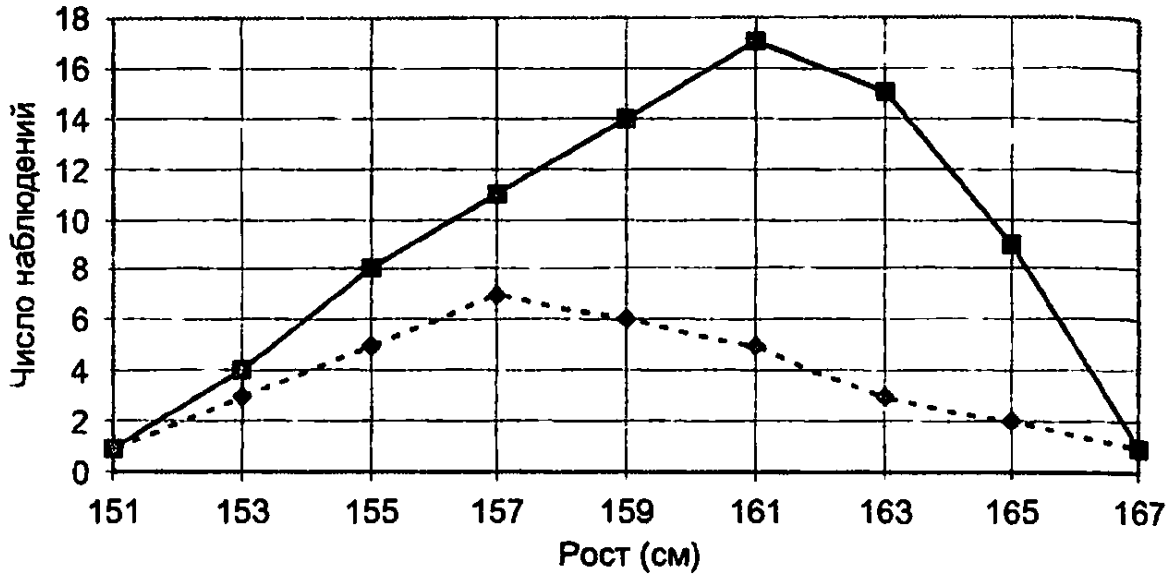


Рис. 59. **Распределение мужчин и женщин по росту**

этого сдвига, характеризуют распределение ряда положением его середины и называются **средними величинами.** *Средняя величина выражает характерную, типичную для данного ряда величину признака* Эта величина образуется в данных условиях места и времени под воздействием всей совокупности действующих факторов Средняя величина является равнодействующей всех этих факторов В средней величине погашаются индивидуальные различия отдельных единиц наблюдения, обусловленные случайными, привходящими обстоятельствами

4.4.1. Среднее арифметическое. Статистическое взвешивание

Наиболее употребительной из средних величин является среднее арифметическое Среднее арифметическое может обозначаться различным символом (*M*, *A* и др). *В медицинской статистике чаще всего для его обозначения применяется символ M (от латинского Медиа — середина).* Для простого среднего арифметического, которое вычисляется в простом, не сгруппирован-

ном вариационном ряду, используется формула: $M=\frac{1}{n}\sum_{i=1}^{i=n}V_i$, или в более упрощенном виде: $M=\frac{\sum V_i}{n}$, где n — число наблюдений, V_i — варианты (V_1, V_2, V_3, V_4 ... V_n) С арифметической точки зрения в основе вычислений лежат две простые операции сложение всех вариант и деление полученной суммы на число наблюдений (табл 49)

Сгруппированный вариационный ряд иногда называют *взвешенным рядом* Такое название определено ролью, которую играют частоты Понятно, чем больше частота той или иной варианты, тем большую роль, большую значимость, *больший вес*, она имеет в характере распределения числового ряда. Среднее арифметическое, рассчитанное в этом ряду, называют *взвешенным средним:* $M=\frac{1}{n}\sum V_i P_i$, где n — число наблюдений, V_i — варианты, P_i — их частоты Число наблюдений во взвешенном (дискретном) ряду определяется как сумма частот $n=\sum P_i$ Соответственно, формулу для вычисления среднего можно представить в виде $M=\frac{\sum V_i P_i}{\sum P_i}$. При вычислении среднего взвешенного последовательно выполняются следующие операции (табл 49, второй раздел):

1 Каждая варианта в таком ряду умножается на частоту ее встречаемости, как бы «взвешивается» ($V_1\times P_1$, $V_2\times P_2$, $V_3\times P_3$ $V_n\times P_n$) Чем больше частота варианты, тем больший «вес» она имеет при вычислении среднего. В том случае, когда среднее арифметическое определяется в интервальном ряду, т е. варианты разбиты на группы, частоты перемножаются на серединные значения этих групп

2 Полученные произведения суммируются $\sum V_i P_i$

3 Сумма произведений делится на число наблюдений, в результате чего получается среднее арифметическое

Таблица 49

Способы вычисления среднего арифметического

Простое среднее $M=\frac{\sum V_i}{n}$		Взвешенное среднее $M=\frac{\sum V_i P_i}{\sum P_i}$			Способ моментов $M=A+\frac{\sum d_i P_i}{\sum P_i}h$			
V	*P*	*V*	*P*	*VP*	*V*	*P*	*d*	*Pd*
15	1	15	1	15	15	1	−2	−2
16	1	16	3	48	16	3	−1	−3
17	1	17	5	85	A=17	5	0	0
18	1	18	4	72	18	4	1	4
19	1	19	2	38	19	2	2	4
$\sum V=85$	n=5		$\sum P=15$	$\sum VP=258$		$\sum P=15$	$\sum Pd=3$	
M=85/5=17		M=258/15=17,2			M=17+(3/15)×1=17,2			

Упрощенным вариантом вычисления среднего арифметического является вычисление по способу моментов. Не вдаваясь в математическое обоснование способа моментов, можно выделить следующие этапы вычисления среднего этим способом (табл 49, третий раздел):

1. В ранжированном ряду распределения выбирается *условное среднее A* За условное среднее можно принять любую варианту данного ряда. Для удобства вычисления лучше брать варианту ближе всего лежащую к центру ряда распределения и чаще всего встречающуюся (с наибольшей частотой *P*)

2 Выставляются *условные отклонения d* Их абсолютные значения последовательно увеличивают на единицу, начиная от 0, который соответствует варианте, принятой за условное среднее Знак минус обозначает уменьшение вариант от условного среднего Плюс — соответственное увеличение вариант

3. Произведения условных отклонений на соответствующие им частоты (*Pd*) суммируются с учетом отрицательных знаков $\left(\sum Pd\right)$.

4. Для того чтобы определить среднее арифметическое, полученная сумма делится на число наблюдений (*n*) $\frac{\sum Pd}{n}$ Частное

от этого деления умножается на величину интервала вариационного ряда (h): $\frac{\sum Pd}{n}h$, и к результату перемножения прибавляется условное среднее (A): $A+\frac{\sum Pd}{n}h$.

Нетрудно заметить, что вариационные ряды представляют собой арифметические прогрессии. В этих прогрессиях отдельные числовые значения или группы числовых значений признака располагаются строго упорядочено и с определенным интервалом. Вместе с тем, иногда встречается ситуация, когда вся совокупность разбита на несколько неравных по численности групп В этом случае среднее арифметическое вычисляют рассматривая каждую группу как самостоятельную совокупность В каждой из этих групп сначала вычисляется свое среднее Затем на основе этих данных определяют общее среднее, учитывая число наблюдений в каждой группе $M_{общ}=\frac{\sum M_j n_j}{\sum n_j}$

Этим правилом руководствуются и в случаях, когда надо найти общую среднюю из нескольких средних Для наглядности рассмотрим пример (табл 50).

Таблицы 50

Способ вычисления общего среднего арифметического

Варианты V_1	Частоты P_1	V_1P_1
2	6	12
3	7	21
5	5	25
$n_1=\sum P_1=18$		58

M_1=58/18=3,2

Варианты V_2	Частоты P_2	V_2P_2
3	26	78
6	30	180
9	29	261
$n_2=\sum P_2=85$		519

M_2=6,1

Порядок вычисления общего среднего арифметического следующий·

1. Вычисляем сумму произведений исходных средних на количество наблюдений в совокупностях, из которых они подсчитывались $n_1P_1+n_2P_2=3,2\times18+6,1\times85=577$

2 Находим сумму наблюдений в обеих совокупностях $n_1+n_2=18+85=103$

3 Затем находим общее среднее арифметическое: 577/103=5,6

Нетрудно заметить, что попытка вычислить общее среднее, как полусумму исходных средних, приводит к ошибочному результату $\left(\frac{3,2+6,1}{2}=4,7\right)$.

Таким образом, *общее среднее равно среднему арифметическому исходных средних, взвешенных по объемам наблюдений.*

Для наглядности рассмотрим еще один пример В клинической и лабораторной практике нередко возникает задача — получить смесь каких-либо компонентов с определенными свойствами. *Например:* Какая крепость спирта будет у смеси, состоящей из 10 литров 40%, 30 литров 70% и 50 литров 96% спирта?

Принцип решения этой и ей подобных задач состоит в применении суммирования взвешенных величин. Исходные данные в виде вариационного ряда будут выглядеть так (табл 51)

Таблица 51

Вычисленне средней концентрации спирта

Варианта (крепость спирта)	Вес (кг)
40	10
70	30
96	50

Вычисление среднего арифметического ряда и будет ответом на поставленную задачу. M=(40×10+70×30+96×50)/(10+30+50)=81,1. Путем несложного алгебраического преобразования можно решать и обратные задачи Например, сколько и какой крепости спирта надо добавить, чтобы получить необходимую концентрацию в заданном количестве.

Таким образом, при вычислении любого среднего арифметического, будь то среднее в отдельном вариационном ряду или групповое среднее, должны обязательно учитываться весовые значения отдельных вариант Это и происходит путем перемножения вариант на частоту их встречаемости Собственно говоря, именно поэтому среднее и называется взвешенным При таких вычислениях абсолютные значения частот могут заменяться их процентным выражением (частостями), т е. их удельными весами (см раздел «Ряды распределений Вариационные ряды»)

Иногда удельные веса сами играют роль величин, для которых нужно найти среднее. В таких случаях тоже прибегают к статистическому взвешиванию Рассмотрим условный пример.

Среди больных бронхиальной астмой города NN, состоявших под диспансерным наблюдением, удельный вес лиц, нуждавшихся в госпитализации, в течение года в одном районе составил 24%, в другом — 16%, а в третьем — 11% Нужно найти средний процент больных астмой, нуждавшихся в госпитализации. Самый простой вариант расчета — найти полусумму $\frac{24\%+16\%+11\%}{3}=17\%$

Однако при этом будет допущена ошибка, поскольку при таком расчете не учитывается число больных астмой, фактически состоящих на диспансерном учете в каждом из обоих районов города Для того чтобы избежать неточности, нужно взвесить процент нуждающихся в госпитализации по численности состоящих на учете (табл 52).

Таблица 52

Вычисление средневзвешенного процента нуждающихся в госпитализации

Район	%	Число больных	Произведение
I	24	77	24×77=1848
II	16	166	16×166=2656
III	11	296	11×296=3256
		539	7760

Средний взвешенный процент госпитализации
7760/539=14 4%

Средние величины и статистические коэффициенты (относительные величины) имеют общее родство Большинство относительных величин, по своей сути, являются средними. Например, интенсивный показатель число обращений за медицинской помощью на 1000 населения. То есть, число обращений, в среднем, на 1000 населения. Или показатель соотношения 2,9 штатных единиц среднего медперсонала на 1 врача Иначе, в среднем на 1 врача приходится 2,9 медсестры Такого рода относительные величины иногда называют изолированными средними Изолированными они называются, поскольку вычисляются без вариационного ряда и трактуются без его параметров (частота и т п). Аналогия статистических коэффициентов и средних предъявляет к коэффициентам при суммировании аналогичное требование: необходимость учета весовых значений суммируемых коэффициентов Это обеспечивается суммированием относительных величин путем пересчета исходных абсолютных чисел (см раздел «Относительные величины»)

Одним из способов сравнения динамики социально-экономических явлений является использование индексов, которые представляют собой синтез средних и относительных величин Несмотря на то что в данном издании использование индексов как метода статистического анализа не рассматривается, представляется целесообразным разобрать два примера, демонстрирующих роль весового оценивания и в этих статистических выкладках

Для примера рассмотрим условную индексную оценку изменения цен за 1998—2000 годы Для проведения сравнения будем использовать так называемый простой агрегатный индекс, который представляет собой отношение суммы цен за отчетный период к сумме цен за те же товары (услуги) в базисный, т е. исходный для расчетов, период (табл. 53) Для расчета обычно берется стандартный набор самых необходимых товаров В случае продовольственных продуктов, такой набор необходимых продуктов обычно называется потребительской корзиной продуктов В 2000 году в России потребительская корзина включала 33 наименования продуктов Для разных социальных, демографических категорий населения и места его проживания состав потребительской корзины может существенно различаться

Таблица 53

Расчет агрегатного индекса цен (цены условные)

Продукты	Цены (руб)	
	1998 год	2000 год
Мука пшеничная	4,5	8,55
Рис	7,2	20,54
Хлеб пшеничный	6,3	10,04
Картофель	2,8	7,41
Итого	20,8	46,54

Степень роста цен можно представить в виде отношения $\frac{46,5 \times 100}{20,8} = 224\%$ Это соотношение можно пояснить так стоимость указанного набора продуктов возросла на 124%. Как видно из представленных расчетов, простой агрегатный индекс представляет собой соотношение двух сумм. Однако этот показатель является весьма грубым и приблизительным, лишенным практического смысла Это связано с тем, что реальное потребление различных товаров и услуг различно Например: по данным Госкомстата России, за 2000 год один взрослый мужчина в стране потреблял риса в 1999 году 5 кг, а картофеля — 150 кг Таким образом, для получения более объективной картины динамики цен необходимо произвести расчет взвешенного агрегатного индекса, который бы учитывал весовое значение каждого потребляемого продукта (табл 54)

Таблица 54

Расчет взвешенного агрегатного индекса цен (цены условные)

Продукты	Потребительские цены		Объем потребления в год на 1 чел (мужчину)		
	1998	2000	Нормы (кг в год на 1 чел)	Стоимость (руб/год)	
Мука пшеничная	4,5	8,55	20	90,0	171,0
Рис	7,2	20,54	5	36,0	102,7
Хлеб пшеничный	6,3	10,04	75	472,5	753,0
Картофель	2,8	7,41	150	420,0	1111,5
Сумма	20,8	46,54		1018,5	2138,2

Рост взвешенного индекса цен составит. $\frac{2138{,}2\times 100}{1018{,}5}=209{,}9\%$.

Следовательно, реальный рост цен на указанный набор продуктов будет равен 109,9% Аналогичным образом можно просчитать рост реального потребления отдельных видов продуктов за сравниваемые годы в сопоставимых ценах Так получают взвешенный агрегатный индекс количеств (табл 55)

Таблица 55

Расчет взвешенного агрегатного индекса количеств (объем потребления и цены условные)

Продукты	Количество потребляемых продуктов (кг в год на 1 мужчину)		Стоимость за кг	Стоимость по ценам 1998 года	
	1998	2000	1998	1998	2000
Мука пшеничная	18	21	4,5	81,0	94,5
Рис	4,8	5,5	7,2	34,6	39,6
Хлеб пшеничный	75	81	6,3	427,5	510,3
Картофель	140	160	2,8	392,0	448,0
Сумма	237,8	267,5		980,1	1092,4

Рост взвешенного индекса количеств равен $\frac{1092{,}4\times 100}{980{,}1}=111{,}5\%$

Существует большое число и других методик расчетов индексов В этом издании они не рассматриваются.

4.4.2. Упрощенный способ «ручного» вычисления среднего арифметического

Расчет среднего арифметического с помощью современных статистических программ, установленных на компьютерах, в принципе, не требует построения рядов распределения. Однако когда вычисления производятся на основе рядов, варианты в которых — крупные числа, а также при большом объеме наблюдений, приемы упрощенного вычисления средних могут оказаться более быстрыми и более точными за счет сокращения

ошибок, неизбежно возникающих при вводе в компьютер больших числовых массивов

Например: требуется определить средний вес новорожденных детей. Варианты наблюдений — вес детей — представлены четырехзначными числами Объем наблюдений также достаточно велик — 2500 детей (табл 56) Если просто вводить в компьютер весь этот числовой массив или следовать обычному порядку вычислений среднего взвешенного, то придется оперировать большим количеством громоздких величин Так, для получения среднего взвешенного, частоты *P* необходимо перемножить на соответствующие им варианты *V* (вес в граммах) 100×3350, 150×3400, 175×3450 и т. д Затем суммировать эти произведения и разделить на сумму частот (число наблюдений) 9 046 250/2500=3618,5

При упрощенном вычислении проводить громоздких операций не требуется, поскольку вместо частот используются частости вариант Т. е промежуточные данные из столбика III (табл 56), заменяются другими, более простыми числами. Последовательность операций при вычислении среднего арифметического упрощенным способом выглядит следующим образом.

1 Определяем частости вариант в ряду распределения Вычислить эти частости достаточно просто Частоту конкретной варианты делим на общее число наблюдений. 100/2500=0,04, 150/2500=0,06 и т д.

2 Затем находим условные отклонения от условного среднего (*A*) За условное среднее можно принять любую варианту Лучше брать ту, которая находится ближе к середине ряда и чаще всего встречается (с наибольшей частотой) В нашем примере это варианта 3600 г Выставляем условные отклонения *(D)*, последовательно увеличивая их значения на единицу, начиная от 0, который соответствует варианте принятой за условное среднее, до самой большой (со знаком плюс) и самой малой (со знаком минус).

3. После этого находим произведения частостей на условные отклонения (*D*): 0,04×(−5)=−0,20, 0,06×(−4)=−0,24 и т д

4 Для получения искомого среднего арифметического эти произведения суммируются 0,37 После чего умножаются на ве-

личину интервала: 50×0,37 (в нашем примере интервал h=50 г), и к этой сумме прибавляется условное среднее (в данном примере 3600 г); M=50×0,37+3600=3618,5.

Таблица 56

Вычисление среднего арифметического упрощенным способом

Вес в граммах V	Число детей P	VP	Частости ω	Условные отклонения D	ωD
I	II	III	IV	V	IV×V
3350	100	335000	0,04	−5	−0,20
3400	150	510000	0,06	−4	−0,24
3450	175	603750	0,07	−3	−0,21
3500	250	875000	0,10	−2	−0,20
3550	275	976250	0,11	−1	−0,11
A=3600	300	1080000	0,12	0	0
3650	375	1368750	0,15	1	0,15
3700	275	1017500	0,11	2	0,22
3750	225	843750	0,09	3	0,27
3800	200	760000	0,08	4	0,32
3850	125	481250	0,05	5	0,25
3900	50	195000	0,02	6	0,12
Всего	2500	9046250	1,00	–	0,37

В итоге, несмотря на кажущуюся сложность расчетов, для нахождения среднего арифметического громоздких вычислений не понадобилось

4.4.3. Другие степенные средние

Помимо среднего арифметического для характеристики центра (середины) распределения используются и другие параметры (средние величины) К ним, в частности, относятся **среднее геометрическое, среднее гармоническое и среднее квадратическое** (табл 57) В математической статистике эти средние, как и среднее арифметическое, относят в группу так называемых **степенных средних** Они имеют единое математическое

выражение, отличающееся только коэффициентом k, который является *коэффициентом статистической размерности* признака Разница между этими средними тем больше, чем больше вариабельность признака в статистическом ряду. При небольшой вариабельности разница между этими средними практически незаметна.

Таблица 57

Виды степенных средних величин

Алгебраическое выражение	Степенные средние
$M=\left(\frac{\sum V_i^k}{n}\right)^{\frac{1}{k}}$	Общая формула степенных
$M=\frac{\sum V_i}{n}$	Арифметическое $K=1$
$M=\frac{n}{\sum \frac{1}{V_i}}$	Гармоническое $K=-1$
$M=\sqrt{\frac{V_i^2}{n}}$	Квадратическое $K=2$
$M=\sqrt[n]{V_1^{P_1} V_2^{P_2} \quad V_n^{P_n}}$	Геометрическое $K=0$

В приведенных формулах k — показатель степени (коэффициент статистической размерности); n — число наблюдений; V — варианты (если варианты представлены с частотами, то в формулы вводится P).

Среднее гармоническое — применяется, когда дело имеют с *обратными величинами* (коли-индексы), сложными абсолютными величинами (тонна-километр, килограммометр и т п.) Использование в этих случаях среднего арифметического приводит к ошибочным результатам *Например* В одном из районов к врачу психиатру в течение года из каждых 100 мужчин обратился 1 человек. Среди женщин 1 обратившаяся приходилась на 25 человек Необходимо определить, на сколько жителей, в среднем, приходился один обратившийся. Для простоты вычислений бу-

дем считать, что общее число мужчин и женщин одинаково Среднее арифметическое двух показателей (для мужчин и женщин) (25+100)/2=62,5 будет неверным Правильно в этом случае определять среднее из обратных величин, через среднее гармоническое $M = \frac{2}{\frac{1}{100} + \frac{1}{25}} = 40$

Таким образом, из каждых 40 человек 1 был посетителем психиатра Этот результат мог быть получен и через статистические коэффициенты (интенсивные показатели) Но в данной ситуации вычисление среднего гармонического значительно проще

Среднее квадратическое (не путать со среднеквадратическим отклонением!) — вычисляется, когда исходный ряд чисел представлен вариантами, отражающими значения площадей (площади ожогов, площади земельных участков и т п.)

Среднее геометрическое — вычисляется в тех случаях, когда дело имеют с числовым рядом, отдельные значения в котором распределяются в геометрической прогрессии (резко отличаются друг от друга). Наиболее целесообразно вычисление этого показателя при определении среднего во временных рядах распределения В целом, если при вычислении среднего арифметического подходят к рядам распределения с точки зрения разности между величинами, то при вычислении среднего геометрического подходят с точки зрения соотношения величин *Например* имеется два числа 4 и 16. Среднее арифметическое из них равняется 10, то есть 10 больше 4 на столько же, на сколько 10 меньше 16 Среднее геометрическое из этих чисел равно 8. Число 8 в два раза меньше 16 и в два раза больше 4

Чтобы лучше понять сущность среднего геометрического, рассмотрим *пример спора некого Ноццолини и Галилея (XVII век) Лошадь, стоящая 100 крон, одним лицом оценивается в 10 крон, другим в 1000 крон Какая из двух оценок менее ошибочна? Если рассматривать вопрос с арифметической точки зрения, на сколько ошибка велика, то получим в одном случае ошибку в 90 крон, а во втором — в 900 крон* Если оценивать, во сколько раз ошиб-

лись покупатели, то получим одинаковый ответ для обоих — в 10 раз*

Кроме упомянутых степенных средних величин, в практике медико-биологических исследований используются **среднее логарифмическое**, если ряды распределения представлены логарифмами чисел (децибелы, pH и т. п.), **среднее кубическое**, если ряды распределения — объемы (объемы плазмы, крови, объемы эритроцитарной массы и т п.) Таким образом, при вычислении среднего необходимо принимать во внимание фактический состав исходных данных

4.4.4. Мода и медиана

Большое значение при выборе характеристики среднего уровня имеет и распределение вариант в вариационных рядах. В ряде ситуаций вместо степенных средних более целесообразно использовать так называемые структурные средние К ним относятся мода и медиана

Мода (*Mo*) — наиболее часто встречающаяся в ряду распределения варианта Она дает представление о центре распределения вариационного ряда. Используется.

- для определения центра распределения в открытых вариационных рядах,
- для определения среднего уровня в рядах с резко асимметричным распределением

Например требуется определить среднюю длительность госпитализации рабочих промышленных предприятий в связи с производственным травматизмом (табл. 58). При визуальном анализе графического изображения распределения (рис 60) видно, что ряд распределения не симметричен вершина распределения сдвинута в начало ряда Если определять среднюю величину на основе среднего арифметического (*M*), то средняя длительность одной госпитализации составит 4,2 дня Однако чаще всего (*Mo*) длительность госпитализации составляла 3 дня

* Пример взят из «Общей теории стетистики» Ц Б Урланиса (1962 г)

Таблица 58

Распределение обследованных ио длительиости госпитализации

Число дней госпитализации v	Число рабочих *p*	Частости	Накопленные частости
2	6	0,10	0,10
3	18	0,30	0,40
4	14	0,23	0,63
5	10	0,17	0,80
6	6	0 10	0,90
7	3	0,05	0,95
8	2	0,03	0,98
9	1	0,02	1,00
Итого	60	1,00	–

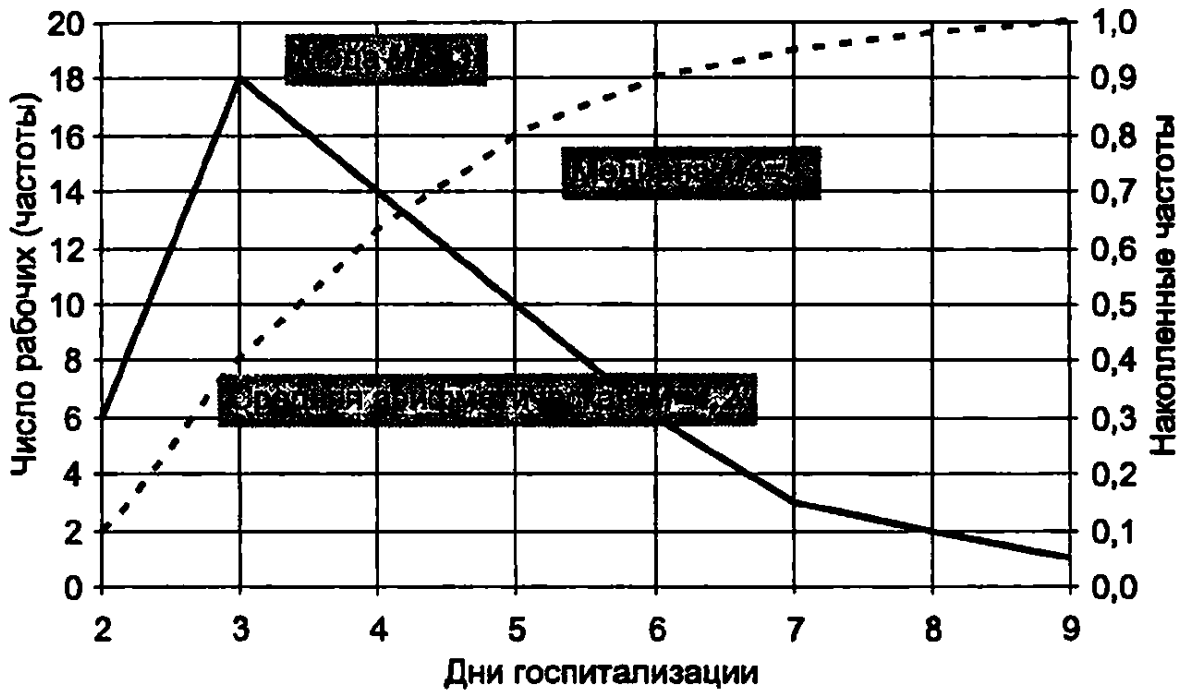


Рис. 60. **Распределение обследованных рабочих по длительности госпитализации**

Установить моду в дискретном вариационном ряду не представляется сложным — варианта, встречающаяся с наибольшей частотой, и есть мода В интервальном ряду нахождение моды сложнее

В грубом приближении в качестве моды можно принять середину группы, на которую приходится наибольшая частота *Например.* в вариационном ряду (см табл 48) наибольшая частота соответствует группе 156–157 см. Середина группы $\frac{156+158}{2}=157$ см

Более точный результат можно получить путем вычисления по формуле $Mo=V_0+h\frac{P_{Mo}-P_{Mo-1}}{2\times P_{Mo-1}-P_{Mo+1}}$, где V_0 — нижняя граница модального интервала; h — величина интервала, P_{Mo} — частоты модального интервала; P_{Mo-1} — частоты предмодального интервала; P_{Mo+1} — частоты послемодального интервала Итак, точное значение моды $Mo=156+2\frac{7-5}{2\times7-5-6}=157{,}3$ см. Практически, полученное значение равно приближенной оценке (157 см), полученной ранее

Медиана — это серединная варианта, центральный член ранжированного ряда Название медиана взято из геометрии, где так именуется линия, делящая сторону треугольника на две равные части В статистике медиана приходится на тот член ранжированного ряда, который «рассекает» совокупность на равные части *Например*, в совокупности (табл 59) медианой будет пятая по счету (ранг=5) варианта 21, ибо четыре значения (17, 18, 19, 20) лежат с одной стороны медианы, и столько же с другой (22, 23, 24, 25) Если вариант в ряду четное количество (табл 60), то медиана равна полусумме двух средних вариант (21+22)/2=21,5

Таблица 59

Нечетное число (9) ранжированных вариант

Ранг	1	2	3	4	5	6	7	8	9
Варианты	17	18	19	20	21	22	23	24	25

Таблица 60

Четное число (8) ранжированных вариант

Ранг	1	2	3	4	5	6	7	8
Варианты	18	19	20	21	22	23	24	25

Медиана в несгруппированном ряду для нечетного ряда — это варианта, имеющая ранг $\frac{N+1}{2}$, или $\frac{9+1}{2}=5$ (N — число вариант в ряду) Для четного ряда медианой является полусумма двух вариант с рангами $\frac{N}{2}+1$ и $\frac{N}{2}$, или в примере полусумма вариант с рангами $\frac{8}{2}+1=5$ и $\frac{8}{2}=4$, т. е. полусумма 5-й и 4-й по счету вариант

В сгруппированном интервальном вариационном ряду положение медианы устанавливается по *накопленным частостям* или *частотам* Варианта, соответствующая сумме частостей 0,5 (или 50% суммы частот), является медианой ряда. Наиболее наглядно положение медианы видно на диаграмме распределения с накопленными частотами (см рис 60)

Для точного определения медианы в интервальном ряду используется формула. $Me=V_0+h\left(\frac{\frac{N}{2}-S_{Me-1}}{P_{Me}}\right)$, где V_0 — нижняя граница медианного интервала, h — величина интервала, S_{Me-1} — накопленные частоты предмедианного интервала, S_{Me} — накопленные частоты медианного интервала, P_{Me} — частота медианного интервала, N — число наблюдений.

Медиана применяется·

- для определения среднего уровня признака в числовых рядах с неравными интервалами в группах,
- для определения среднего уровня признака, когда исходные данные представлены в виде качественных признаков и когда единственным способом указать некий центр тяжести совокупности является указание варианты (группы вариант), которая занимает центральное положение,
- при вычислении некоторых демографических показателей (средней продолжительности предстоящей жизни);
- при определении наиболее рационального места расположения учреждений здравоохранения, коммунальных учреж-

дений и т п. Имеется в виду учет оптимальной удаленности учреждений от всех объектов обслуживания

В настоящее время очень распространены различные опросы (маркетинговые, социологические и др), в которых опрашиваемых просят выставить баллы изделиям, политикам и т. п. Затем из полученных оценок рассчитывают средние баллы и рассматривают их как интегральные оценки, выставленные коллективом опрошенных. При этом обычно для определения средних показателей применяют среднее арифметическое Однако такой способ на самом деле применять нельзя, поскольку баллы — характеристики измеренные в порядковой шкале (см выше), а вычислять среднее арифметическое характеристик, измеренных в порядковых шкалах, некорректно Обоснованным в этом случае является использование в качестве средних баллов медианы или моды.

4.4.5. Вычисление средних в *MS Excel*

Для удобства работы возьмем уже использованные в предыдущем примере данные и расположим их в таблице следующим образом (рис. 61):

	A	B
1	Спирт (%)	Количество
2	40	10
3	70	30
4	96	50

Рис. 61

Последовательность решения задачи в *MS Excel*.

1 В таблице с исходными данными установите курсор в ячейку С2 и занесите в нее выражение =A2*B2 Напоминаем, что все адреса (A1, B1 и т. п) указываются латинским шрифтом! В ячейке С2 появится результат вычисления 400

2 Установите указатель мыши на нижний угол клетки С2 Она должна быть в это время активной, т е выделенной курсором (рис 62)

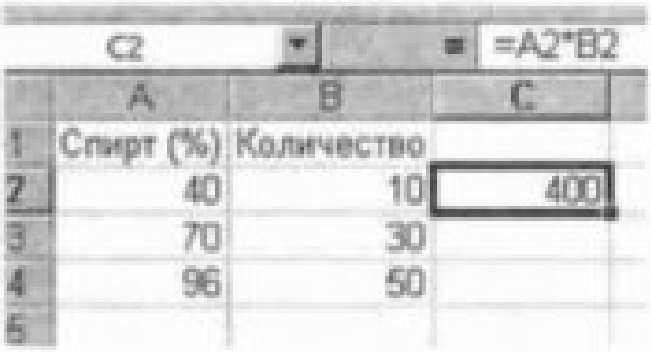


Рис 62

Затем, нажав правую кнопку мыши, перетащите выделенный фрагмент в клетки С3 и С4 В этих клетках появятся результаты вычислений, аналогично результатам клетки С2

3 Установите курсор в клетку С5. Выполните команду <**Функция**> из меню <**Вставка**> В открывшемся окне мастера функций выберите последовательно категорию «**Математические**», затем «**СУММ**» (рис 63)

Укажите координаты клеток, значения которых необходимо суммировать (С2 С4) и нажмите кнопку **[ОК]**. В клетке С5 появится сумма 7300.

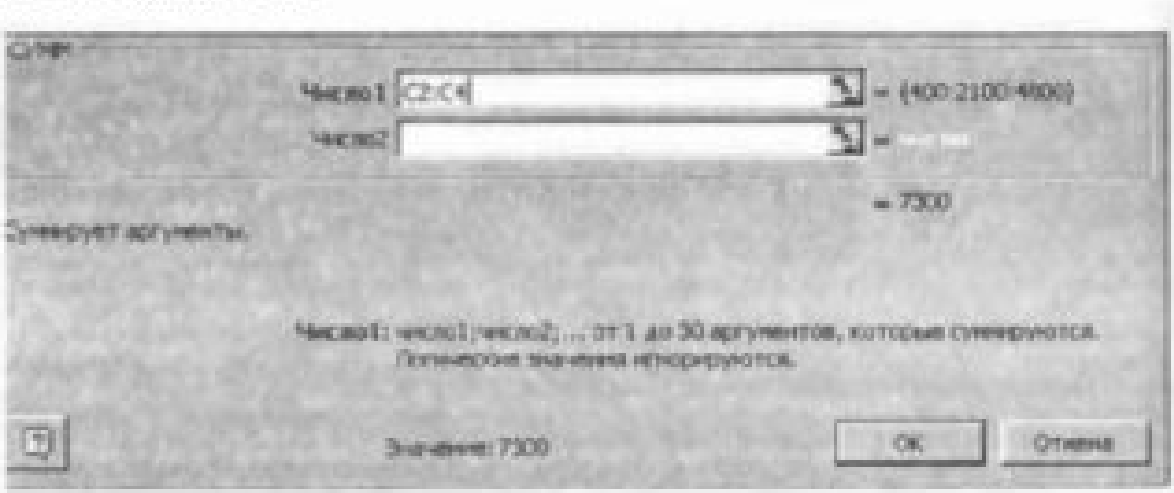


Рис. 63

4. Затем скопируйте (как в пп. 2) содержимое клетки С5 в клетку В5. После этого в клетке С5 будет виден результат суммирования всех произведений, а в клетке В5 — число наблюдений (сумма частот, «количеств»)

5. Затем в клетке С6 наберите выражение =С5/В5. Таким образом получите ответ: значение среднего арифметического взвешенного 81,1 (рис 64).

	A	B	C
1	Спирт (%)	Количество	
2	40	10	400
3	70	30	2100
4	96	50	4800
5		90	7300
6			81,11111

Рис 64

После этого, меняя числа в клетках А2, А3, А4 (крепость спирта) и В2, В3, В4 (количество спирта), можно добиваться любого нужного результата в итоговых клетках С6 (конечная концентрация раствора) и В5 (объем раствора).

В Excel имеется ряд статистических функций, предназначенных для вычисления некоторых степенных средних величин

СРЗНАЧ возвращает (синоним «позволяет получить») среднее арифметическое из нескольких массивов (аргументов) чисел. Число 1, число 2, ... — это от 1 до 30 массивов, для которых вычисляется среднее.

Пример: Если ячейки А1:А5 содержат числа 10, 7, 9, 27 и 2, то среднее арифметическое равняется 11 (рис. 65).

СРГАРМ возвращает среднее гармоническое множества данных. Среднее гармоническое — это величина, обратная к среднему арифметическому обратных величин

Пример: СРГАРМ(10; 7; 9; 27; 2) равняется 5,611

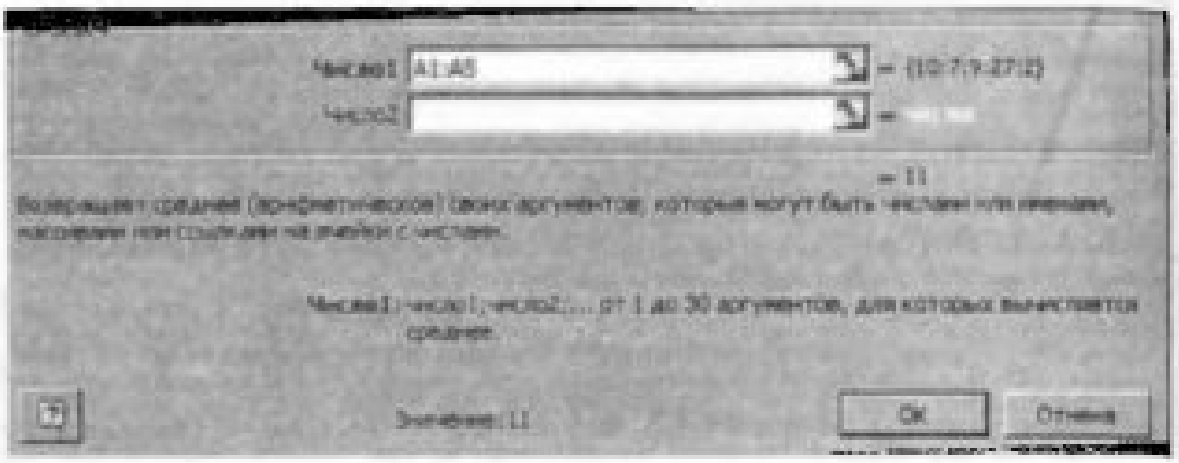


Рис. 65

СРГЕОМ возвращает среднее геометрическое значений массива или интервала положительных чисел. Например, функцию СРГЕОМ можно использовать для вычисления средних показателей динамического ряда

МЕДИАНА возвращает медиану заданных чисел Медиана — это число, которое является серединой множества чисел, т. е. половина чисел имеют значения бóльшие, чем медиана, а половина чисел имеют значения меньшие, чем медиана

Пример: МЕДИАНА 10; 7; 9; 27; 2 равняется 9

МОДА возвращает наиболее часто встречающееся значение в массиве или интервале данных.

Пример: МОДА 5,6; 4; 4; 3; 2; 4 равняется 4.

4.5. Показатели рассеяния вариант

В предыдущих разделах рассматривались методы построения рядов распределения и вычисления средних величин. Наряду с ними, большое значение имеют и характеристики распределения вариант в ряду распределения *Например*:

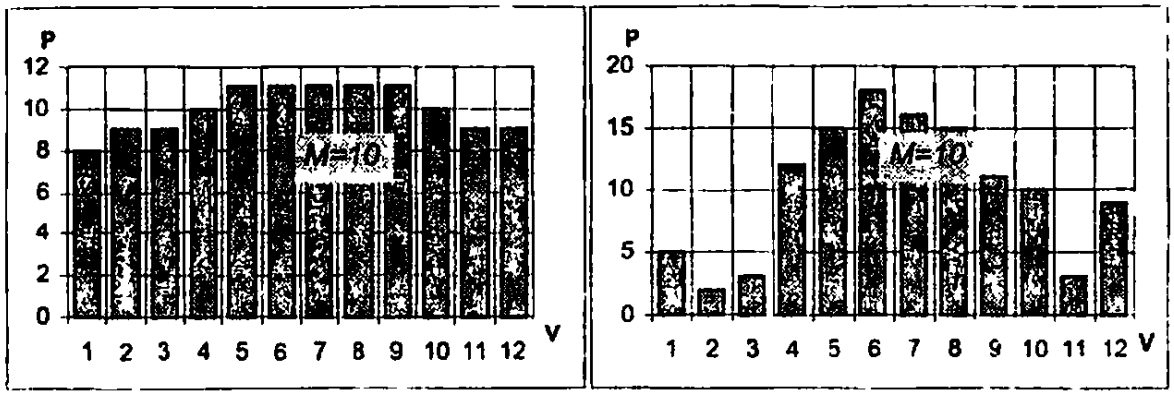


Рис 66 **Распределение числовых рядов при одинаковых средних арифметических**

Среднее арифметическое в двух представленных статистических совокупностях (рис 66) одинаково (M=10) Однако распределение отдельных числовых значений в этих совокупностях совершенно различно Более того, при решении практических задач возможна ситуация, когда в сравниваемых группах нет ни одного схожего значения, а средние величины этих групп одинаковы

Например: средние величины двух разных числовых последовательностей. −100; −20; 100; 20 и 0,1, −0,2; 0,1 абсолютно одинаковы и равны 0

Таблица 61

Основные статистические критерии разнообразия признака (разброса вариант) в рядах распределений

	Критерии	Формулы для расчетов
1	Лимит (*Lim*)	$Lim = V_{max} - V_{min}$
2	Амплитуда (*Ampl*)	$Ampl = V_{max} - V_{min}$
3	Дисперсия (D)	$D = \frac{\sum d_i^2}{n}$ для простого ряда $D = \frac{\sum d_i^2 p_i}{\sum p_i}$ для взвешенного ряда
4	Среднеквадратическое (стандартнос) отклонение (σ)	$\sigma = \sqrt{D}$
5	Коэффициент вариации (Cv)	$Cv = \frac{\sigma}{M} \cdot 100\%$

Таблица 62

Практическое применение параметрических критериев разнообразия признака

Дисперсия *D*	1 Для оценки вариабельности рядов распределения 2 Для факторного анализа (дисперсионный анализ) 3 Для статистической оценки двух совокупностей с одинаковыми или близкими значениями средних (критерий ***Фишера***)
Среднеквадратическое отклонение σ	1 Для оценки данных одноименных (однородных) числовых рядов при близких средних чем больше σ, тем больше разброс значений, соответственно среднее арифметическое менее типично для данного ряда 2 Для оценки типичности среднего (*правило трех сигм*) в изолированном ряду 3 Для определения доверительных интервалов статистических коэффициентов и репрезентативности выборочных исследований 4 Для диагностической оценки показателей физического развития
Коэффициент вариации *Cv*	1 Используется для сравнения вариабельности значений разнонменных признаков 2 Для получения нормированных значений вариабельности (малая, средняя, большая)

Простейшими количественными характеристиками рассеивания вариант являются лимит и амплитуда

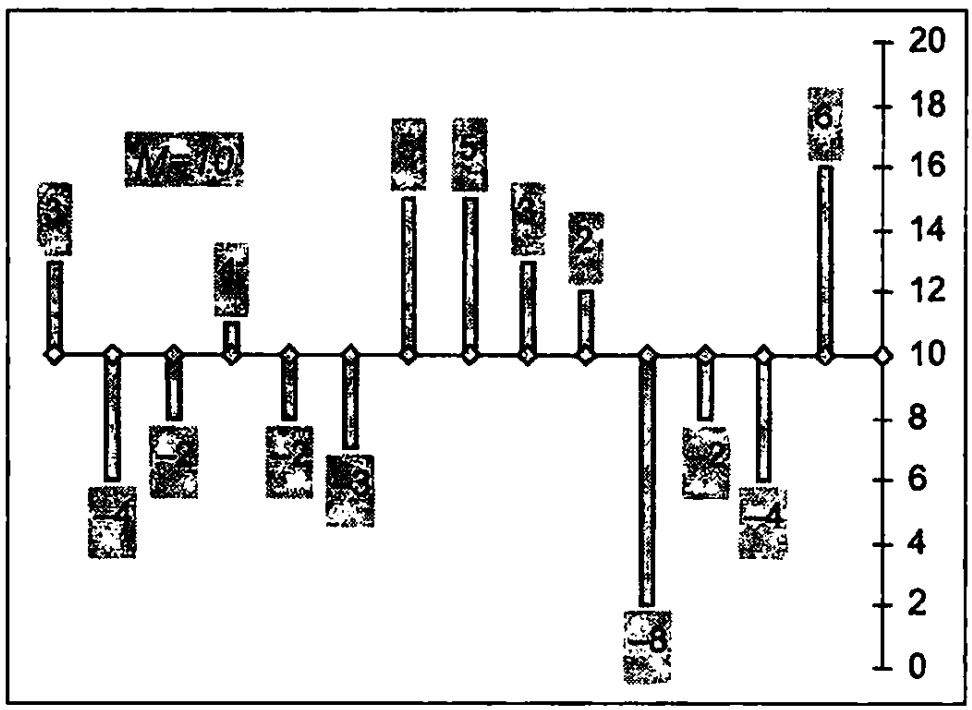
Лимит *(Lim)* указывает границы вариационного ряда *Например*, самый большой вес 95 кг, самый маленький 48 кг $Lim = (48 \div 95)$ кг

Амплитуда, или как еще говорят вариационный размах *(Ampl)*, исчисляется как разность между максимальным и минимальным значениями признака $Ampl = 95 - 48 = 47$ кг

4.5.1. Дисперсия

Существенным недостатком лимита и амплитуды как критериев вариабельности является то, что они полностью зависят от крайних значений признака в вариационном ряду При этом не учитываются колебания значений признака внутри ряда.

Наиболее просто определить однородность числового ряда с учетом всех значений, составляющих этот ряд, через отклонения всех вариант от центра ряда (среднего арифметического), поскольку каждое отдельное наблюдение на какую-то величину не совпадает со средним арифметическим. Разность между конкретной вариантой и средним арифметическим из этого ряда называется *отклонением от среднего* $d_i=(V_i-M)$ Такие отклонения от среднего *(M=10)* можно представить в графической форме



Для получения обобщающей характеристики числового ряда использовать сумму отклонений от среднего нельзя Это связано с тем, что сумма всех отрицательных и положительных отклонений от среднего всегда равна нулю. Можно избежать взаимной компенсации отклонений, беря квадраты отклонений, т. к при возведении в квадрат отрицательные и положительные числа дают только положительные значения При усреднении всех отклонений числового ряда получается средний квадрат отклонений, который называется **дисперсией** — *D* Алгебраическое выражение дисперсии $D=\sigma^2=\frac{\sum d_i^2}{n}$, где *n* — число наблюдений, *d* —

отклонения вариант от среднего $d_i=(V_i-M)$ Во взвешенном ряду дисперсия вычисляется по формуле $D=\frac{\sum d_i^2 P_i}{\sum P_i}$

Таблица 63

Способы вычисления дисперсии

Простой ряд $D=\frac{\sum d_i^2}{n}$			Простой ряд $D=\frac{\sum V_i^2}{n}-M^2$		Взвешенный ряд $D=\frac{\sum d_i^2 P_i}{\sum P_i}$ или $\frac{\sum V_j^2 P_j}{\sum P_j}-M^2$			
V	d	d^2	V	V^2	V	P	VP	V^2P
15	−2	4	15	225	15	1	15	225
16	−1	1	16	256	16	3	48	768
17	0	0	17	289	17	5	85	1445
18	1	1	18	324	18	4	72	1296
19	2	4	19	361	19	2	38	722
$M=17$ $\sum d=0$		$\sum d^2=10$	$\sum V^2=1455$		$\sum P=15$	$\sum VP=258$	$\sum V^2P=4456$	
$D=10/5=2$			$D=1455/5-17^2=2$		$M=258/15=17,2$ $D=4456/15-17,2^2=1,2$			

Упрощенные способы расчета дисперсии позволяют избежать вычислений отклонений d В этом случае, для не сгруппированного ряда $D=\frac{\sum V_j^2}{n}-M^2$, где $\sum V_j^2$ — сумма квадратов вариант ряда, M^2 — квадрат среднего арифметического, n — число наблюдений Для сгруппированного ряда формула вычисления дисперсии упрощенным способом выглядит следующим образом $D=\frac{\sum V_j^2 P_j}{\sum P_j}-M^2$, где $\sum V_j^2 P$ — сумма произведений квадратов вариант ряда на их частоту, M^2 — квадрат среднего арифметического, $\sum P_j$ — число наблюдений, определяемое как сумма частот

Если в результате статистического наблюдения получены несколько групп значений признака, то для вычисления общей дисперсии можно группы в единую совокупность не объединять Более того, если совокупность имеет большое число наблюдений (большой объем), то в случае «ручного» проведения вычислений

целесообразно ее разбить на несколько групп В том и другом случаях вычислением дисперсий отдельных групп можно заменить непосредственное вычисление общей дисперсии Поскольку *общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий* Это свойство дисперсий имеет большое теоретическое и практическое значение, являясь основой широко применяющегося в научных исследованиях **дисперсионного анализа**

Формула для расчета общей дисперсии представлена выражением $D_{общ}=D_{внгр}+D_{межгр}$, где:

$D_{общ}$ — общая дисперсия, дисперсия значений признака всей совокупности относительно общего среднего;

$D_{внгр}$ — внутригрупповая дисперсия, среднее арифметическое групповых дисперсий, взвешенных по объемам групп $D_{внгр}=\frac{\sum N_j D_j}{n}$, где n — объем всей совокупности, N_j — обьем группы j; D_j — дисперсия группы j,

$D_{межгр}$ — межгрупповая дисперсия. $D_{межгр}=\frac{\sum N_j\left(M_j-M\right)^2}{n}$,

где M_j — групповое среднее группы j , M — общее среднее, n — объем всей совокупности, N_j — объем группы j

Практически расчет общей дисперсии не представляет труда *Например* требуется найти общую дисперсию совокупности состоящей из двух групп Вычисления проходят по следующим этапам

1-й этап:

Таблица 64

Вычисление средних в первой и второй группе

Первая группа			Вторая группа		
V_1	P_1	V_1P_1	V_2	P_2	V_2P_2
2	1	2	3	2	6
3	4	12	5	4	20
4	5	20	7	6	42
6	3	18	8	2	16
$n_1=\sum P_1=13$		$\sum V_1P_1=52$	$n_2=\sum P_2=14$		$\sum V_2P_2=84$
$M_1=52/13=4$			$M_2=84/14=6$		

2-й этап Вычисление общего среднего всей совокупности (обеих групп) $M_{общ}=\frac{M_1 n_1 + M_2 n_2}{n_1 + n_2}=\frac{4\times13+6\times14}{13+14}=5{,}04\approx5$

3-й этап:

Таблица 65

Вычисление групповых дисперсий

Первая группа			Вторая группа		
V_1	P_1	$V_1^2P_1$	V_2	P_2	$V_2^2P_2$
2	1	4	3	2	6
3	4	36	5	4	20
4	5	80	7	6	42
6	3	108	8	2	16
$n_1=\sum P_1=13$		$\sum V_1^2P_1=228$	$n_2=\sum P_2=14$		$\sum V_2^2P_2=540$
$D_1=\sum V_1^2P_1/n_1-M_1=1{,}54$			$D_2=\sum V_2^2P_2/n_2-M_2=2{,}57$		

4-й этап Рассчитываем внутригрупповую дисперсию, как среднюю групповых дисперсий $D_{внгр}=\frac{1{,}54\times13+2{,}57\times14}{13+14}=2{,}07$

5-й этап. Определяем межгрупповую дисперсию, как дисперсию групповых средних относительно общего среднего: $D_{межгр}=\frac{13\times(4-5{,}04)^2+14\times(6-5{,}04)^2}{13+14}=1$

6-й этап Общая дисперсия $D_{общ}=D_{внгр}+D_{межгр}=1+2{,}07=3{,}07$

4.5.2. Среднеквадратическое отклонение

Существенный недостаток дисперсии, которая является именованной величиной, — несоответствие ее размерности и размерности отдельных единиц числового ряда. Так, если варианты выражены в метрах, то дисперсия дает квадратные метры, если варианты в килограммах, то дисперсия дает квадрат этой меры, и т д Указанного недостатка лишено **среднеквадратическое отклонение** σ (стандартное отклонение, стандарт распреде-

ления). Алгебраически среднеквадратическое отклонение представляет собой корень квадратный из дисперсии $\sigma=\sqrt{D}$ или $\sigma=\sqrt{\frac{\sum d^2}{n}}$.

Для взвешенного вариационного ряда формула, по которой вычисляется среднеквадратическое отклонение, выглядит так $\sigma=\sqrt{\frac{\sum d_1^2 P_1}{n}}$

При малом числе наблюдений можно нспользовать упрощенный способ приближенного определения величины среднеквадратического отклонения $\sigma=(V_{max}-V_{min})/6$, где V_{max} и V_{min} — максимальная и минимальная варианты ряда.

В педиатрии среднеквадратическое отклонение используется для оценки физического развития детей путем сравнения данных конкретного ребенка с соответствующими стандартными показателями За стандарт принимаются средние арифметические показатели физического развития здоровых детей Сравнение показателей со стандартами проводят по специальным таблицам, в которых стандарты приводятся вместе с соответствующими им **сигмальными шкалами**. Считается, что если показатель физического развития ребенка находится в пределах стандарт $\pm\sigma$, то физическое развитие ребенка (по этому показателю) соответствует норме Если показатель находится в пределах стандарт $\pm2\sigma$, то имеется незначительное отклонение от нормы Если показатель выходит за эти границы, то физическое развитие ребенка резко отличается от нормы (возможна патология).

Среднеквадратическое отклонение и дисперсия широко используются как *составляющие параметры нормального распределения* при вычислении различных сложных *параметрических статистических критериев* и проведения *параметрического статистического анализа*

В то же время, дисперсия и среднеквадратическое отклонение как статистические критерии рассеивания имеют следующие недостатки:

- эти критерии — абсолютные именованные величины, поэтому использовать их при сравнении разнородных рядов нельзя (сантиметры не сравнить с килограммами и т п);
- их размерность зависит, среди прочего, и от абсолютного значения среднего арифметического вариационного ряда *Например*, при одинаковой вариабельности признаков, D и σ будут больше в том ряду, где средний вес равен 120 кг, чем в ряду со средним весом 60 кг

Несмотря на то что дисперсия и среднеквадратическое отклонение зависят прямо пропорционально от разброса вариант (чем больше разброс, тем больше D и σ), эти критерии не дают оснований говорить, значителен или не значителен разброс в изолированном ряду Они могут использоваться только для сравнения двух или нескольких рядов распределения

4.5.3. Коэффициент вариации

Недостатков, свойственных дисперсии и среднеквадратическому отклонению, лишен **коэффициент вариации** (*Cv*) Этот коэффициент представляет процентное отношение среднеквадратического отклонения к среднему арифметическому $Cv = \frac{\sigma}{M} 100\%$. Арифметически отношение σ и M нивелирует влияние абсолютной величины этих характеристик, а процентное соотношение делает коэффициент вариации величиной не именованной Кроме того, этот коэффициент позволяет оценивать вариабельность (разброс) признака в нормированных границах Если его значение не превышает 10% , то можно говорить о слабом разбросе Если коэффициент вариации находится в пределах 10—20%, разброс средний, если превышает 20%, то разброс вариант считают большим Отличие коэффициента вариации от других критериев разброса наглядно демонстрирует *пример* (табл. 66).

Таблица 66

Состав работников промышленного предприятия *N*

Учетный признак	Среднее арифметическое *M*	Среднеквадратическое отклонение σ	Коэффициент вариации *Cv*
Стаж работы (лет)	8,7	2,8	32,1
Возраст (лет)	37,2	4,1	11,0
Образование (классов)	9,2	1,1	11,9

На основании приведенных в примере статистических характеристик можно сделать вывод об относительной однородности возрастного состава и образовательного уровня работников предприятия, при низкой профессиональной устойчивости обследованного контингента Нетрудно заметить, что попытка судить об этих социальных тенденциях по среднеквадратическому отклонению привела бы к ошибочному заключению, а попытка сравнения учетных признаков «стаж работы» и «возраст» с учетным признаком «образование» вообще была бы некорректной из-за разнородности этих признаков

4.5.4. Квантили

Для порядковых (ранговых) распределений, где критерием середины ряда является медиана, среднеквадратическое отклонение и дисперсия *D* не могут служить характеристиками рассеяния вариант. То же свойственно и для открытых вариационных рядов Указанное обстоятельство связано с тем, что отклонения, по которым вычисляются *D* и σ, отсчитываются от среднего арифметического, которое не вычисляется в открытых вариационных рядах и в рядах распределений качественных признаков. Поэтому для сжатого описания распределений используется другой параметр разброса — **квантиль** (синонимы «перцентиль», «персентиль»), пригодный для описания качественных и количественных признаков при любой форме их распределения Этот параметр может использоваться и для перевода

количественных признаков в качественные В этом случае такие оценки присваиваются в зависимости от того, какому по порядку квантилю соответствует та или иная конкретная варианта.

Например: В медико-биологических исследованиях нередко используют центили следующих размерностей 3-й ($V_{0,03}$), 10-й ($V_{0,10}$), 25-й ($V_{0,25}$), 50-й ($V_{0,50}$), 75-й ($V_{0,75}$), 90-й ($V_{0,90}$) и 97-й ($V_{0,97}$) В этом случае — 3-й ($V_{0,03}$) центиль равен варианте, чье значение не превышает значения у 3% всех вариант данного ряда. Варианты меньше 10-го ($V_{0,10}$) центиля встречаются у 10% всех вариант ряда и т д. При оценке показателей физического развития показатели, которые у конкретного пациента меньше 3-го центиля, оцениваются как резко пониженные, находящиеся по своему значению между 3-м и 10-м центилем, — как пониженные, между 10-м и 25-м центилем, — ниже среднего, между 25-м и 75-м — средние, между 75-м и 90-м — выше среднего, между 90-м и 97-м — повышенные, выше 97-го — резко повышенные.

В практике статистического анализа наиболее часто используются следующие квантили.

$V_{0,5}$ — медиана;

$V_{0,25}$, $V_{0,5}$, $V_{0,75}$ — квартили (четверти), где $V_{0,25}$ — нижняя квартиль, $V_{0,75}$ — верхняя квартиль;

$V_{0,1}$, $V_{0,2}$, $V_{0,3}$. $V_{0,9}$ — децили (десятые);

$V_{0,01}$, $V_{0,02}$, $V_{0,03}$ $V_{0,99}$ — процентили, или центили (сотые)

Квантили делят область возможных изменений вариант в вариационном ряду на определенные интервалы Медиана (квантиль $V_{0,5}$) — это варианта, которая находится в середине вариационного ряда и делит этот ряд пополам, на две равные части (0,5 и 0,5). Квартиль делит ряд на четыре части первая часть (нижняя квартиль $V_{0,25}$) — это варианта, отделяющая варианты, числовые значения которых не превышают 25% максимально возможного в данном ряду, квартиль $V_{0,5}$ отделяет варианты с числовым значением до 50% от максимально возможного. Верхняя квартиль ($V_{0,75}$) отделяет варианты величиной до 75% от максимально возможных значений

Одним из методов расчета, который позволяет понять статистическую суть квантилей, является графический метод В осно-

ве этого метода — построение диаграммы распределения числовых значений признака в квантилях, которые могут выражаться в процентах или долях единицы Для примера (рис 67) представлена кривая распределения подростков по весу в квартилях

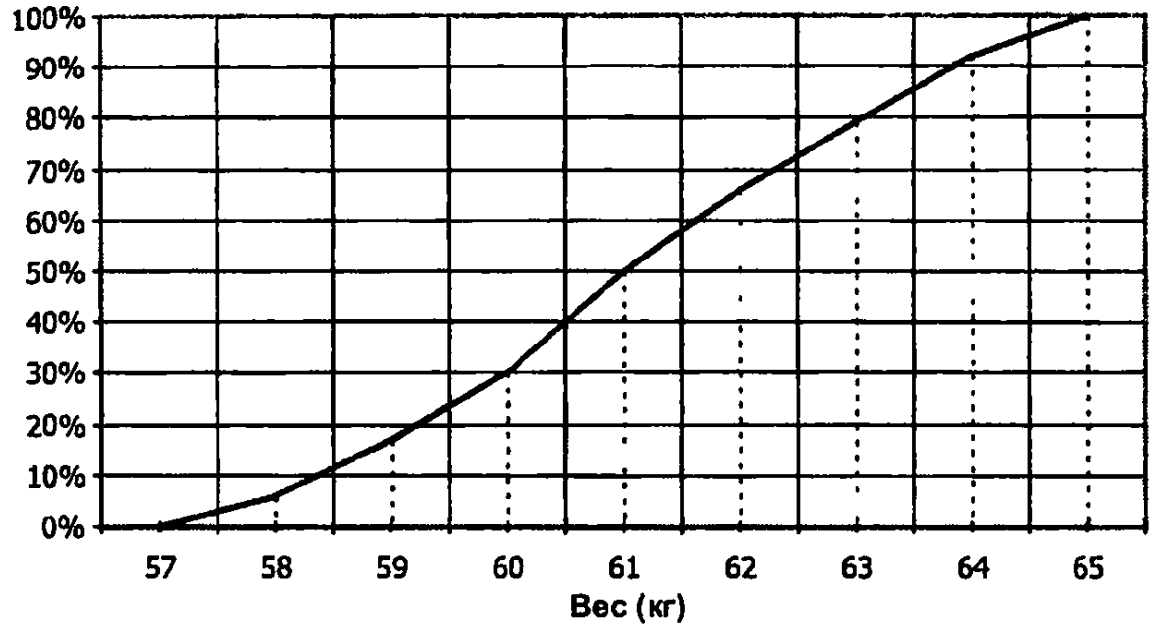


Рис. 67. **Распределение обследованных подростков по росту**

Для определения значения *медианы*, или *центрального квартиля*, проводим прямую от отметки накопленных частостей, равной 50%, до кривой распределения Затем от кривой распределения опускаем перпендикуляр на горизонтальную ось диаграммы, где отмечены числовые значения вариант В итоге получаем значение центрального квартиля $V_{0,5}$=61 кг Аналогичным образом определяются нижняя квартиль $V_{0,25}$=59 кг и верхняя квартиль $V_{0,75}$=62,5 кг

Более точным методом расчета квартилей и других видов квантилей являются вычислительные методы, основанные на использовании различных математических приемов, в том числе и с помощью пакета анализа *MS Excel.*

4.5.5. Использование *MS Excel* для нахождения квантилей

В *MS Excel* имеются статистические функции, предназначенные для нахождения любых квантилей.

Пример Необходимо рассчитать центили следующих значений 0,03, 0,1, 0,25, 0,50; 0,75, 0,90, 0,97 Введите в столбцы А и В *MS Excel* массив данных измерения роста группы рабочих В столбец С — значения квантилей, которые необходимо вычислить (рис. 68).

	A	B	C	D
1	164	184	0,03	157,0
2	157	160	0,10	157,9
3	163	157	0,25	160,0
4	162	166	0,50	161,0
5	166	182	0,75	163,3
6	161	160	0,90	167,6
7	161	161	0,97	182,9
8	158	159		
9	162	160		
10	160	163		

Рис. 68. **Исходные данные для вычисления квантилей**

Функция **ПЕРСЕНТИЛЬ** представляет квантили (синоним персентили) заданного числового массива Окно функции имеет поля ввода исходной информации **Массив** и **К**. В окно **Массив** могут вводиться числовые последовательности или ссылки на ячейки, содержащие числа В окно **К** вводятся искомые значения квантилей. Эти значения указываются в долях единицы 0,1; 0,25 и т. д. Для того чтобы вычислить ряд центилей

1 Введите в клетки таблицы *MS Excel* (от С1 и до С7) значения центилей 0,03; 0,10; 0,25; 0,50, 0,75; 0,90, 0,97

2. Установите курсор в клетку D1, куда после ввода исходных значений функции **ПЕРСЕНТИЛЬ** будет выведен результат вычислений, и вызовите эту функцию.

3 В окно **Массив** введите координаты исходного массива данных и в окно **К** — координаты клетки C1. В этой клетке расположено значение первого центиля (рис. 69)

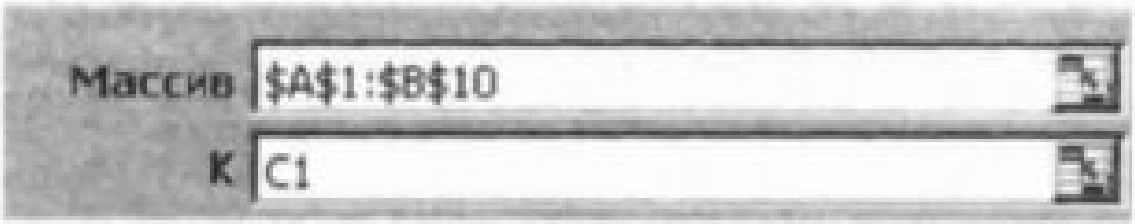


Рис. 69. **Пример заполнения рабочих окон функции ПЕРСЕНТИЛЬ**

После нажатия клавиши **[ОК]** в клетке D1 появится результат вычисления. Затем, установив указатель мыши на угол выделенной ячейки D1, нажмите левую клавишу мыши и не отпуская ее перетащите указатель до клетки D7. Тем самым будет скопировано содержимое первой клетки в клетки D2:D7 Из полученных данных в столбце D видно, что 3% обследованных рабочих имели рост не выше 157 см, 10% — не более 157,9 см и т. д

Для получения более детальной характеристики распределения центилей можно вызвать из меню <**Сервис**> <**Пакет анализа**>, в котором находится функция **Ранг и персентиль** (рис 70) Заполните поля окна функции в соответствии с расположением исходных данных:

Входной интервал — числовой массив или адреса ячеек, в которых расположен исходный массив,

Группировка по столбцам или строкам — расположение исходных данных по строкам или столбцам (в данном примере — по столбцам);

Метки в первой строке — указываются, если в первой строке имеются заголовки (метки);

Параметры вывода — указывают место вывода результатов

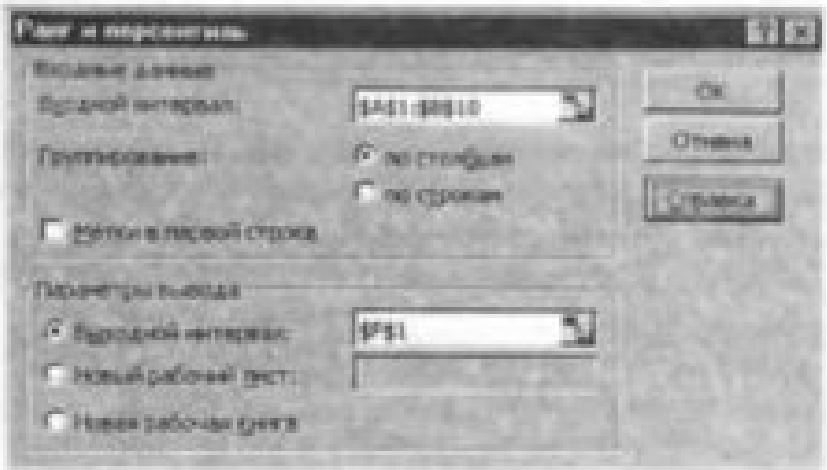


Рис. 70. **Заполнение окна исходных данных функции «Ранг и перцентиль»**

Результат представляется в виде таблицы (рис 71) **Обратите внимание!** Несмотря на то что исходные данные вводились одним массивом, результат представлен в виде двух блоков (см. рис. 71), каждый из которых относится к одному из столбцов исходных данных В данном примере один из них относится к первому столбцу (A1:A10), второй — к второму столбцу (B1 B10).

В ряду *Точка* указывается номер места, на котором располагается та или иная варианта в исходном массиве В ряду *Столбец1* указываются варианты, расположенные в ранговом порядке (в порядке убывания); в следующем ряду, *Ранг*, указывается ранговой номер варианты В ряду *Процент* указывается процент вариант, которые имеют такое же и меньше значение, чем варианта данного ранга.

Точка	*Столбец1*	*Ранг*	*Процент*	*Точка*	*Столбец2*	*Ранг*	*Процент*
5	166	1	100,00%	1	184	1	100,00%
1	164	2	88,80%	5	182	2	88,80%
3	163	3	77,70%	4	168	3	77,70%
4	162	4	55,50%	10	163	4	66,60%
9	162	4	55,50%	7	161	5	55,50%
6	161	6	33,30%	2	160	6	22,20%
7	161	6	33,30%	6	160	6	22,20%
10	160	8	22,20%	9	160	6	22,20%
8	158	9	11,10%	8	159	9	11,10%
2	157	10	,00%	3	157	10	,00%

Рис. 71. **Вывод результатов функции «Ранг и перцентиль»**

4.5.6. Статистические моменты. Асимметрия и эксцесс

Для вычисления многих статистических характеристик удобно пользоваться **эмпирическими** (т е вычисляемыми на основе результатов конкретных наблюдений) **момеитами.** Формула для вычислений обычных эмпирических моментов — $M_k^* = \frac{\sum P_j (V_j - C)^k}{n}$, где V_j — наблюдаемые варианты, P_j — их частоты, n — число наблюдений, k — степень (порядок) момента, C — произвольное постоянное число.

Если $C=0$, то момент считается начальным При $k=1$ (момент первого порядка) формула приобретает вид $M_1^* = \frac{\sum P_j (V_j - 0)^1}{n} = \frac{\sum P_j V_j}{n}$.

Нетрудно заметить, что эта формула алгебраически является полным аналогом формулы для вычисления среднего взвешенного Таким образом, *начальный момент первого порядка равен среднему арифметическому*

Если при вычислении моментов коэффициент C принимается равным центру распределения — среднему арифметическому ($C=M$), то момент называется центральным моментом При $k=2$ *(момент второго порядка)* алгебраически формула приобретает вид, который соответствует формуле, используемой для вычисления дисперсии (табл 67) Отсюда *центральный момент второго порядка равен дисперсии*

С помощью центральных моментов третьего и четвертого порядка удобно вычислять такие характеристики распределения, как **асимметрия и эксцесс.**

На представленном рис 72 полигон частот имеет явно скошенное в правую сторону асимметричное распределение Количественное описание такой скошенности дается с помощью *коэффициента асимметрии* (***As***)

Таблица 67

Начальные и центральные эмпирические моменты

Эмпирические моменты	Центр распределения C	k	Формула расчета	Область применения (статистический критерий)
Начальный момент (M_1^*)	0	1	$\frac{\sum P_j V_j}{n}$	Среднее арифметическое $M=M_1^*$
Центральный момент второго порядка (m_2)	Среднее арифметическое	2	$\frac{\sum P_j (V_j - M)^2}{n}$	Дисперсия $D=m_2$
Центральный момент третьего порядка (m_3)	Среднее арифметическое	3	$\frac{\sum P_j (V_j - M)^3}{n}$	Асимметрия $As=\frac{m_3}{\sigma^3}$
Центральный момент четвертого порядка (m_4)	Среднее арифметическое	4	$\frac{\sum P_j (V_j - M)^4}{n}$	Эксцесс $E=\frac{m_4}{\sigma^4}-3$

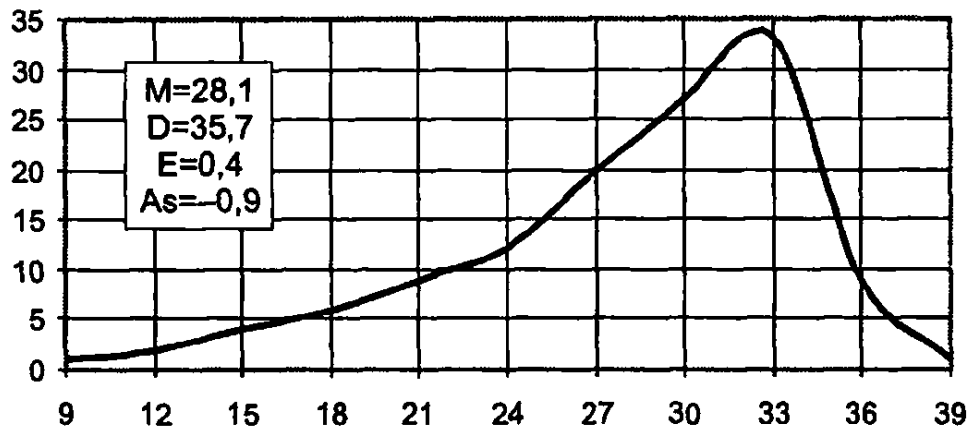


Рис 72. **Пример скошенного распределения**

Величина *As* тем больше, чем больше выражена асимметрия Знак величины этого коэффициента однозначно связан с направлением асимметрии Если распределение вытянуто в сторону отрицательных значений (центр распределения принимается за нуль), то коэффициент *As* — *положителен* (*As*>0), в противоположном случае — отрицателен (*As*<0) (рис 73) Количественное описание эксцесса дается с помощью коэффициента *эксцесса (E)*. При $E>0$ распределение принято считать островершинным, при $E<0$ — туповершинным (см рис 73)

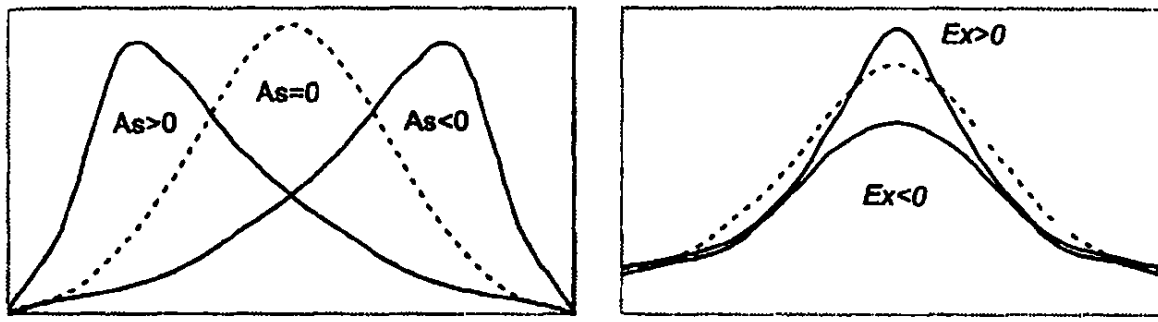


Рис 73

4.6. Статистическая проверка статистических гипотез

Методологической основой любого исследования является формулировка рабочей гипотезы При этом цель исследования — получение данных, на основании которых выдвинутую еще до начала исследования, как говорят априори, гипотезу можно было бы принять, т е. признать истинной, либо отвергнуть — признать ложной.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах распределений *Например*

- генеральная совокупность распределена по закону Пуассона;
- средние арифметические двух совокупностей не равны между собой,
- дисперсии (разброс значений) двух совокупностей равны между собой.

В первом случае выдвинута гипотеза о виде неизвестного распределения Во втором и третьем случаях — о параметрах двух известных распределений. Гипотеза «в четверг будет дождик» не является статистической, поскольку в ней речь не идет ни о виде, ни о параметрах статистического распределения

Выдвинутую гипотезу называют *основной или нулевой* (H_0) Гипотезу, которая противоречит нулевой и является ее логическим отрицанием, называют *конкурирующей или альтернативной*

(H_1) Гипотезы H_0 и H_1 предоставляют выбор только одного из двух вариантов *Например: если нулевая гипотеза предполагает, что среднее арифметическое* $M=15$, *то логическим отрицанием будет* $M\neq15$ Коротко это записывается так H_0 $M=15$, H_1 $M\neq15$ В медико-биологических исследованиях при оценках различий каких-либо параметров в качестве нулевой гипотезы обычно принимают гипотезу об отсутствии различий.

Простой называют гипотезу, содержащую только одно предположение Простая гипотеза прямо указывает на некий определенный закон распределения, точнее на его параметры (среднее арифметическое, дисперсию и т. п.).

Сложной называют гипотезу, которая состоит из конечного или бесконечного множества простых гипотез. *Например· сложная гипотеза H: D*>15 может состоять из бесчисленного множества простых *H. D*>16, *H. D*>17, *H: D*>18 и т д Таким образом, сложная гипотеза указывает не единственное распределение (параметры распределения), а какое-то множество, семейство распределений

Выдвинутая гипотеза может оказаться правильной или неправильной Поэтому она проверяется. Поскольку эту проверку делают статистическими методами, то ее называют *статистической проверкой.* При статистической проверке могут быть допущены ошибки двух родов

Ошибка первого рода — отвергается правильная гипотеза Вероятность совершить ошибку первого рода называют *уровнем значимости.* Этот параметр принято обозначать через α В биологии и медицине уровень значимости принимают не выше 0,05 Это означает, что в 5 случаях из 100 (в 5%) мы рискуем допустить ошибку первого рода.

Ошибка второго рода — принимается неправильная гипотеза. Значимость ошибки второго рода обозначают символом β

Последствия этих ошибок могут быть различны Если, например, отвергнут какой-либо весьма эффективный метод лечения (ошибка первого рода), то скорее всего будут применяться другие, может быть менее эффективные методы. Последствия неправильного лечения (ошибка второго рода) могут быть более

тяжкими. В другом варианте человек не уехал на нужном ему поезде (ошибка первого рода) или сел на поезд, следующий в другом направлении (ошибка второго рода) Последствия ошибок в этой ситуации так же явно не однозначны. Анализ потерь и выигрышей при принятии правильных и неправильных решений является задачей самостоятельной дисциплины — теории принятия решений.

Для проверки нулевых гипотез используют специально подобранные величины — *статистические критерии (К)* Например, если проверяют гипотезу о равенстве дисперсий в двух совокупностях H_0. $D_1=D_2$, то в качестве критерия (K) истинности нулевой гипотезы (H_0) принимают отношение этих дисперсий: $F=\frac{D_1}{D_2}$

Величина F, называемая *критерием Фишера*, подчиняется своему закону распределения — закону *Фишера—Снедекора* Зная этот закон, можно установить фактическое значение критерия при тех или иных параметрах этого распределения (число степеней свободы, объемы наблюдений и др.).

Для проверки гипотез величины критериев, полученные в результате наблюдений (опытов) ($K_{набл}$), сравнивают с уже известными (фактическими) значениями таких критериев ($K_{факт}$)

Главной проблемой здесь является правильный выбор статистического критерия с учетом допустимой области его применения Область же применения того или иного критерия задается законом его распределения Весьма существенное значение в этом аспекте имеет факт, относится ли избранный критерий к семейству параметрических или к семейству непараметрических критериев

Например В большинстве медико-биологических исследований для статистической проверки гипотез существенности различий средних используется параметрический критерий Стьюдента (t) Методология его применения основана на предположении о принадлежности сравниваемых выборочных совокупностей к нормальному распределению и равенстве дисперсий. На практике, зачастую, ни какой проверки соответствия исход-

ных данных этим условиям не проводят, что может вести к существенным просчетам

Для статистической проверки соответствия теоретического и эмпирического (полученного в результате опыта) распределений необходимо располагать теоретическими распределениями Получение последнего представляется некоторой проблемой Для ее решения могут быть использованы следующие способы

1 Наиболее оптимальный вариант — использование компьютерных статистических программ С их помощью все необходимые вспомогательные данные (теоретические распределения, критические значения статистических критериев и т п) получаются автоматически Результаты их сравнений с данными эмпирических наблюдений выводятся также автоматически по запросу пользователя

2 Вычисление теоретических значений критериев и отдельных статистических характеристик можно осуществлять используя встроенные функции математической статистики компьютерных программ (в том числе и *Microsoft Excel*) С помощью этих функций можно генерировать и числовые ряды различных теоретических распределений.

3 В случае «ручной» обработки, для нахождения теоретических значений можно использовать специальные таблицы (см Приложения) Иногда можно отказаться от таблиц, применяя приближенные оценки граничных значений критериев

Все возможные значения выбранного критерия (в нашем примере критерия Фишера F) разбиваются на две части (два подмножества) Одна часть, содержащая значения критерия $K_{кр}$, при которых нулевая гипотеза отвергается, называется *критической областью.* Другая часть возможных значений критерия, при которых нулевая гипотеза принимается, называется *областью принятия гипотезы.* Если полученные значения критерия $K_{набл}$ принадлежат критической области, то нулевую гипотезу отвергают Если полученные значения $K_{набл}$ принадлежат области принятия нулевой гипотезы, то ее принимают Критическая область и область принятия гипотезы являются интервалами, разделяемыми *критическими точками* ($k_{кр}$)

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области

Правосторонняя — критическая область определяется неравенством $K_{кр}>k_{кр}$ (рис 74*A*), где $k_{кр}$ — положительное число

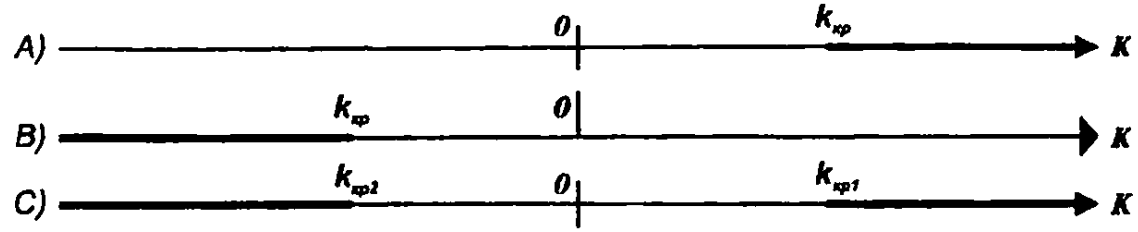


Рис. 74 **Схема распределения критических областей принятия гипотез**

Левосторонней называют критическую область, определяемую неравенством $K_{кр}< k_{кр}$ (рис 74*B*). *Двусторонней* называют область, определяемую неравенством $k_{кр1}< K_{кр}$ и $K_{кр}< k_{кр2}$ (рис 74*C*). Если вычисленное значение критерия $K_{набл}$ окажется в критической области, то есть $K_{набл} \geq K_{кр}$, то нулевую гипотезу отвергают Если же $K_{набл} \neq K_{кр}$, то оснований отвергнуть нулевую гипотезу — нет

В медико-биологических исследованиях используют как односторонние, так и двусторонние критерии На практике *односторонние критерии* применяются тогда, когда оценивается один из вариантов: либо одна выборка больше другой, либо одна выборка меньше другой *Двусторонние* когда оценивается соответствие одной выборки другой (отличается одна от другой?). На первый взгляд, логических различий здесь нет Однако они весьма существенны

Например при исследовании изменений веса печени подопытных животных, по двум группам животных (контрольной и опытной) найдены дисперсии $D_1=25$ и $D_2=7$. Требуется определить статистическую достоверность их различий Поскольку изменения могут касаться и увеличения и уменьшения веса, применение одностороннего критерия исключается Порядок рассуждений следующий Число наблюдений в первой выборке составило $n_1=13$, во второй $n_2=11$. Наблюдаемое значение критерия Фишера $F=\frac{25}{7}=3{,}6$ Поскольку распределение Фишера–Снедекора зависит от числа степеней свободы, для определения $Fk_{кр}$

вычисляются степени свободы: $k_1=n_1-1=11-1=10$ и $k_2=n_2-1=13-1=12$, где n_1 — объем выборки, по которой была вычислена бóльшая дисперсия, и n_2 — меньшая

При заданном уровне значимости $P=0,01$ и степенях свободы $k_1=10$ и $k_2=12$ находим правостороннюю критическую точку распределения $k_{кр}=4,3$ При $P=0,05$ критическая точка распределения $k_{кр}=2,7$ Для нахождения критических точек $k_{кр}$ с помощью *Microsoft Excel* можно использовать встроенную функцию **FРАСПОБР** из меню <**Вставка**>/<**Функция**> Эта функция воспроизводит $k_{кр}$ с заданной вероятностью и степенями свободы

Поскольку вычисленный критерий при $P=0,05$ $F>Fk_{кр}$, то нулевая гипотеза, говорящая о равенстве дисперсий (H_0 $D_1=D_2$), отвергается с вероятностью 0,05 При $P=0,01$ $F<Fk_{кр}$ В этом случае нулевая гипотеза не может быть отвергнута. Таким образом, с вероятностью ошибки α не более 0,05 можно признать различие этих дисперсий существенным или статистически достоверным Но при вероятности α не более 0,01, различие дисперсий существенным или статистически достоверным признать нельзя Окончательное заключение в данном случае зависит от исследователя, проводившего опытное наблюдение

Применение одностороннего критерия может быть следующим *Например:* под воздействием одного из неблагоприятных факторов городской среды отмечено снижение показателей здоровья в обследованной группе жителей города, по сравнению с контрольной группой жителей села. Значение критерия Стьюдента t, с помощью которого проверялась статистическая значимость выявленных различий, составило 1,8 Число степеней свободы в этом наблюдении было равно 25 Тогда критическая точка распределения t критерия при $P=0,05$ составляет 2,06 ($K_{набл}<K_{кр}$, 1,8<2,06) Таким образом, вероятность нулевой гипотезы слишком высока, чтобы ее можно было отвергнуть, т е *нулевая гипотеза не может быть отвергнута.* Следовательно, выявленные различия между жителями города и села статистически не значимы В случае использования одностороннего критерия критическая точка того же t критерия (1,8) будет равна 1,71 ($K_{набл}>K_{кр}$, 1,8>1,71) При таком уровне значимости *нулевая гипо-*

теза H_0 может быть отвергнута и, соответственно, различия можно признать статистически достоверными Поскольку в данной ситуации речь шла о воздействии неблагоприятного фактора внешней среды, ожидать улучшения здоровья в опытной группе под воздействием неблагоприятного фактора не представляется реальным. Поэтому можно ориентироваться на результаты анализа с помощью одностороннего критерия (в опытной группе может быть только хуже, чем в контрольной). В данном примере это существенно, так как при использовании двустороннего критерия признать результаты статистически достоверными нельзя.

Следует помнить о том, что если нулевая гипотеза все же принимается, то это не значит, что тем самым она доказана Один факт, подтверждающий какое-либо положение, еще не доказывает его. Т е в этой ситуации можно только утверждать, что полученные результаты не противоречат предположению об отсутствии различий Для подтверждения гипотезы, как правило, необходимы дополнительные исследования или подтверждающие гипотезу дополнительные данные, которые получены каким-либо другим путем. Отвергается гипотеза, как правило, более категорично, поскольку в математической статистике достаточно одного факта, чтобы отвергнуть любое сомнительное предположение

В отношении статистических гипотез следует придерживаться общего правила: **доказать** на основании однократной или косвенной проверки гипотезу **нельзя**, а **отвергнуть — можно**

Помимо вероятности (α) попадания критерия в критическую область при условии, что нулевая гипотеза справедлива, целесообразно учитывать и вероятность попадания критерия в критическую область, когда нулевая гипотеза неверна и, следовательно, справедлива конкурирующая Такая вероятность называется *мощностью критерия* Т. е *мощность критерия — есть вероятность попадания критерия в критическую область, при условии, что справедлива конкурирующая гипотеза.* Следовательно, если вероятность ошибки принять за неправильную гипотезу (ошибка второго рода равна β), то мощность равна $1-\beta$ Таким обра-

зом, чем больше мощность критерия, тем меньше вероятность ошибки второго рода.

На первый взгляд кажется, что для повышения точности статистических данных нужно просто уменьшить вероятности ошибок первого и второго рода Однако при неизменном объеме выборки, *одновременно увеличить* β и α невозможно При попытке уменьшить вероятность одной ошибки, неизбежно возрастает вероятность другой. Поэтому, вероятности ошибок выбирают с учетом возможных последствий этих ошибок Если к более тяжким последствиям ведут ошибки первого рода, то необходимо брать как можно меньшее значение α Если более тяжелые последствия вызывают ошибки второго рода, то стараются снизить β Единственный способ снизить одновременно вероятности статистических ошибок первого и второго рода — увеличение объема выборок

4.7. Оценка статистических параметров по выборочным данным

В практике медико-биологических исследований изучаются обычно выборочные, а не генеральные совокупности Естественно, что замена исследования генеральной совокупности исследованием выборки порождает ряд вопросов:

1 В какой степени выборка отражает свойства генеральной совокупности, т. е. в какой степени выборка репрезентативна по отношению к генеральной совокупности?

2 Какую информацию о значениях параметров генеральной совокупности могут дать параметры выборки?

3 Можно ли утверждать, что полученные выборочным путем статистические характеристики (средние величины, дисперсия или любые другие производные величины) равны тем характеристикам, которые могут быть получены из генеральной совокупности

Проверка показывает, что значения параметров, полученных для разных выборок из одной генеральной совокупности, обычно не совпадают Рассчитанные выборочным путем числовые значения параметров выборок являются лишь результатом приближенного **статистического оценивания** значений этих параметров в генеральной совокупности Статистическое оценивание, в силу изменчивости наблюдаемых явлений, позволяет получать только их приближенные значения

Примечание. *Строго говоря, в статистике* ***оценка*** *— это правило вычисления оцениваемого параметра, а термин* ***оценить****, т е. провести оценивание, означает указать приближенное значение.*

Различают оценки **точечные** и оценки **интервальные**. Проиллюстрируем точечные оценки простым условным примером Пусть мы имеем генеральную совокупность N, состоящую всего из 10 вариант Среднее значение генеральной совокупности составляет. M=(16+18+20+22+24+26+28+30+32+34)/10=25,0. Затем получим среднее арифметическое выборочным путем Для этого сформируем случайным способом три выборки с числом наблюдений равным 3, 4 и 5 (табл 68)

Таблица 68

Пример выборочных совокупностей

Генеральная совокупность N		1-я выборка	2-я выборка	3-я выборка
1	16			16
2	18		18	
3	20	20	20	
4	22	22		22
5	24			24
6	26	26	26	
7	28			
8	30			30
9	32		32	
10	34			34
Число наблюдений	10	3	4	5
Среднее арифметическое	25,0	22,7	24,0	25,2
Отклонения выборочных средних от генерального среднего		2,3	1,0	−0,2

Полученные выборочные средние (22,7; 24,0; 25,2) являются *точечными оценками* генерального среднего (25,0) Любая выборочная характеристика, используемая в качестве приближенного значения генеральной характеристики и получаемая вычислением одного числа (точки), называется **точечной статистической оценкой** При избрании способа получения точечных оценок учитывается, что они должны обладать свойствами *состоятельности, несмещенности* и *эффективности.*

Состоятельная оценка — точечная оценка, которая при неограниченном увеличении объема выборки приближается (*сходится*) к истинному значению оцениваемой генеральной характеристики *Например* по данным примера (см табл 68) среднее арифметическое в первой выборке составило 22,7, во второй — 24,0, в третьей — 25,2. Нетрудно заметить, что по мере увеличения числа наблюдений выборочные средние все больше приближаются к генеральному среднему (25,0) Соответственно абсолютные значения отклонений выборочных средних от генерального среднего уменьшаются (2,3; 1,0; 0,2). Таким образом, эти выборочные средние можно считать состоятельными точечными оценками генерального среднего. Следует отметить, что приведенный пример является условным На практике обнаружить схождение выборочных характеристик удается при значительно большем росте числа наблюдений

Несмещенная оценка — точечная оценка, лишенная *систематической ошибки Например* выборочное среднее арифметическое является несмещенной оценкой генерального среднего То есть выборочные средние могут иметь случайные отличия от генеральных Если рассматривать несколько выборок из одной генеральной совокупности, то отклонения точечных оценок из этих выборок будут взаимно погашаться, а их суммарная точность будет возрастать по мере увеличения числа этих оценок Выборочная *оценка дисперсии — смещенная оценка* Не вдаваясь в описание причин, вызывающих систематические ошибки при вычислении выборочной дисперсии, следует отметить, что она дает всегда несколько заниженные оценки генеральной дисперсии Поэтому, если для определения генеральной дисперсии по

выборочным данным используют формулу $D=\frac{\sum d_i^2}{n}$, то получают *смещенную точечную оценку генеральной дисперсии* Для получения несмещенной точечной оценки генеральной дисперсии из выборочных данных используют формулу расчета исправленной дисперсии $D=\frac{\sum d_i^2}{n-1}$ При сравнении формул видно, что они отличаются лишь знаменателями. Очевидно, что при больших объемах выборки смещенная и несмещенная (исправленная) дисперсия отличаются мало На практике пользуются исправленной дисперсией если число наблюдений в выборке не превышает 30 вариант ($n<30$), поскольку при большем числе наблюдений влияние -1 становится не существенным

Эффективная оценка — такая точечная оценка, которая гарантирует наименьшее отклонение выборочной оценки от такой же оценки генеральной совокупности.

4.7.1. Доверительная значимость, доверительная вероятность, доверительный интервал, доверительный предел

Оценки, рассмотренные выше, являются точечными. В связи с этим возникает вопрос можно ли по результатам точечной оценки одной лишь выборки судить о свойствах всей генеральной совокупности На первый взгляд кажется, что нельзя. На приведенном примере (см табл 68) видно, что выборочные средние не совпадают с генеральным средним. Однако каждый результат, полученный в отдельной выборке, можно рассматривать как случайную величину Соответственно, при увеличении числа выборок, распределение точечных оценок будет принимать характер нормального распределения Это значит, что в случае средних арифметических относительные отклонения выборочных средних от генерального среднего (характеристик не-

посредственно генеральной совокупности) распределяются так же, как относительные отклонения нормально распределенных вариант от среднего арифметического вариационного ряда

Отсюда в частности следует, что 68,3% всех выборочных средних находятся в пределах $\Delta = M \pm m$ Иными словами имеется вероятность 0,683, что выборочное среднее отличается от генерального не более чем на $\pm m$ В этой формуле Δ — предельная ошибка выборки, M — среднее выборочное, m — стандартное отклонение среднего значения (по аналогии со стандартным отклонением вариант от среднего вариационного ряда). В медико-биологической литературе параметр m принято называть **«стандартная ошибка среднего»** или **«ошибка среднего»** Вычисляется этот параметр в случае повторного отбора по формуле $m = \frac{\sigma}{\sqrt{n}}$, где σ — среднеквадратическое отклонение выборки, n — число наблюдений в выборке (объем выборки), или $m = \sqrt{\frac{\sigma^2}{n}}$, где $\sigma^2 = D$ — дисперсия

Если выборка, объем которой известен (n), сформирована из генеральной совокупности бесповторным отбором, то в формулу вводится поправочный множитель, и она приобретает вид $m = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$ Очевидно, что при большой генеральной совокупности, когда $N \to \infty$, этот множитель стремится к единице

При определении ошибки выборочной доли (*например:* 0,25 или 0,47) используют формулу $m = \pm\sqrt{\frac{P(1-P)}{n-1}}$ В случаях, когда доля выражена в % (*например:* 25% или 47%), $m = \pm\sqrt{\frac{P(100-P)}{n-1}}$. Указанным способом ошибки доли определяются, если число наблюдений достаточно велико. Необходимую величину выборки в этом случае можно найти из неравенства $Pn > 500$, т е произведение доли (в %) на число наблюдений не должно быть меньше 500 Кроме того, чтобы использовать указанную формулу, сами выборочные доли не должны намного отличаться от 0,5

(50%). В случае, когда доля меньше 0,2 (20%) или больше 0,8 (80%), следует использовать другую методику

Поскольку параметр *m* характеризует ошибку утверждения (ошибку прогноза) о том, что выборочное среднее равно генеральному среднему, то чем выше требование к вероятности этого вывода, тем шире должен быть обеспечивающий точность такого прогноза интервал, называемый **«доверительный интервал»**.

Статистическая оценка, которая определяется двумя числами — концами интервала, называется **интервальной оценкой.**

Величина доверительного интервала задается вероятностью безошибочного прогноза, эту вероятность принято называть **«доверительная вероятность»** или вероятностью безошибочного прогноза, а иногда надежностью. Величина доверительной вероятности может задаваться доверительным параметрическим коэффициентом *t* — коэффициентом Стьюдента (псевдоним английского химика У Госсета, 1908)

При достаточно большом числе наблюдений *(n>30),* значения доверительного коэффициента *t* и доверительной вероятности соотносятся следующим образом (табл. 69)

Таблица 69

Соотношение статистических критериев достоверности выборочных характеристик

Доверительный критерий *t*	Доверительиая вероятность (%)	Уровень значимости (*P*)
1	68,3	0,32
2	95,5	0,05
3	99,7	0,01

При малых числах наблюдений значения козффициента Стьюдента с учетом уровня доверительной вероятности можно установить по специальным таблицам.

Выбор того или иного уровня значимости или, соответственно, доверительной вероятности в общем является произвольным. В медико-биологических исследованиях допускается доверительная вероятность не менее 95,5% В этом случае доверительный интервал для средних при достаточно большом числе

наблюдений ($n>30$), равен $\pm 2m$. Предельная ошибка выборки $\Delta = M \pm 2m$ При доверительной вероятности 99,7% доверительный интервал составит $\pm 3m$, $\Delta = M \pm 3m$. В целом, чем больше доверительная вероятность, тем больше доверительный интервал и предельная ошибка

Граничные точки доверительного интервала называются **доверительными пределами.**

Каждому значению доверительной вероятности соответствует свой **уровень значимости (*P*).** Уровень значимости выражает вероятность нулевой гипотезы, т е вероятность того, что выборочная и генеральные средние не отличаются друг от друга Иначе говоря, чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют Для доверительной вероятности 0,95 (95%), *например*, уровень значимости $P=1-0,95=0,05$.

Таблица 70

Интервальная оценка среднего арифметического

$M=25,2$ $m=3,1$ $n=50$			
Критерий Стьюдента t	1	2	3
Доверительная вероятность	68 3%	95,5%	99,7%
Уровень значимости P	0,32	0,05	0,01
Доверительный интервал $\pm tm$	±3,1	±6,2	±9,3
Предельная ошибка выборки Δ	25,2±3,1	25,2±6,2	25,2±9,3
Доверительные пределы $M+tm \div M-tm$	28,3÷22,1	31,4÷19,0	34,5–15,9

Если выборки небольшие по объему, то распределение вероятностей не следует точно нормальному закону распределения В этом случае для определения величины доверительного коэффициента, соответствующей определенному значению доверительной вероятности или уровню значимости, пользуются специальными таблицами Очевидно, что в реальных исследованиях желательно иметь как можно меньший доверительный интервал при достаточно высокой доверительной вероятности

Таким образом, статистическая значимость выборочных характеристик представляет собой меру уверенности в их «истинности» Уровень значимости находится в убывающей зависимо-

сти от надежности результата Более высокая статистическая значимость соответствует более низкому уровню доверия к найденной в выборке характеристике Именно уровень значимости представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность.

Выбор порога уровня значимости, выше которого результаты отвергаются как статистически не подтвержденные, во многом произвольный Как правило, окончательное решение обычно зависит от традиций и накопленного практического опыта в данной области исследований Верхняя граница $P<0,05$ статистической значимости содержит довольно большую вероятность ошибки (5%) Поэтому в тех случаях, когда требуется особая уверенность в достоверности полученных результатов, принимается значимость $P<0,01$ или даже $P<0,001$

В практике медико-биологических исследований наиболее часто используются следующие значения показателей значимости: 0,1; 0,05, 0,01, 0,001. Традиционная интерпретация уровней значимости, принятая в этих исследованиях, представлена в табл. 71

Таблица 71

Интерпретация уровней значимости (*P*)

Показатели значимости (*P*)	Интерпретация
≥ 0,1	Данные согласуются с нулевой гипотезой (H_0)
≥ 0,05	Есть сомнения в истинности как нулевой (H_0), так и альтернативной гипотез (H_1)
<0,05	Нулевая гипотеза (H_0) может быть отвергнута
≤ 0,01	Нулевая гипотеза (H_0) может быть отвергнута Сильный довод
≤ 0,001	Нулевая гипотеза (H_0) почти наверняка не подтверждается Очень сильный довод

Из формулы стандартной ошибки среднего (m), от которой во многом зависит величина интервала, следует, что эта ошибка зависит как от числа наблюдений в выборке (n), так и от однородности выборки.

Интервальные оценки коэффициентов асимметрии, эксцесса и коэффициента вариации проводятся на основе стандартных ошибок этих коэффициентов.

Ошибка показателя асимметрии при очень малых объемах выборки может производиться по формуле: $m_{As} = \sqrt{\frac{6}{n}}$

Ошибка показателя эксцесса: $m_E = \sqrt{\frac{24}{n}} \approx 2m_{As}$

Ошибка коэффициента вариации: $m_{Cv} = \frac{Cv}{\sqrt{n}} \sqrt{\frac{1}{2} + \left(\frac{Cv}{100}\right)^2}$

В медико-биологических исследованиях объекты наблюдения, как правило, весьма вариабельны и не поддаются регулированию. Поэтому там, где это возможно, желательно брать выборки большего объема, что, к сожалению, в большинстве исследований весьма трудно или вообще невозможно. Поэтому значение выборочных статистических оценок в медико-биологических исследованиях не может быть определено с высокой точностью. Это в свою очередь делает бессмысленным точное измерение исходных данных. Поэтому при планировании исследований указанное обстоятельство необходимо обязательно учитывать, что позволит избежать ненужных материальных затрат (чем точнее измерения, тем они дороже) и неоправданных потерь времени при сборе данных

Иногда при статистических расчетах необходимо вычислять сумму, разность, произведение и частное от деления средних величин. В этих ситуациях необходимо соответствующим образом оперировать и с ошибками средних.

Для определения суммарной ошибки суммы средних $M_1 + M_2 + \ldots + M_n$ можно использовать формулу: $m = \pm\sqrt{m_1^2 + m_2^2 + \ldots + m_n^2}$. Следует иметь в виду, что более точный результат суммарной ошибки получается при пересчете объединенного массива исходных данных (всех вариант), а указанной формулой следует пользоваться в случае невозможности такого перерасчета

Формула ошибки разности средних арифметических $M_1 - M_2 - \ldots - M_n$ вычисляется аналогично: $m = \pm\sqrt{m_1^2 + m_2^2 + \ldots + m_n^2}$

Ошибка деления средних арифметических:

$$m = \pm \frac{M_1}{M_2} \sqrt{\left(\frac{m_1}{M_2}\right)^2 + \left(\frac{m_2^2}{M_1}\right)^2}.$$

Ошибка произведения средних арифметических

$$m = \pm M_1 M_2 \sqrt{\left(\frac{m_1}{M_2}\right)^2 + \left(\frac{m_2^2}{M_1}\right)^2}$$

4.8. Вычисление показателей описательной статистики в *MS Excel*

В *MS Excel* имеется надстройка (опция) «*Пакет анализа*», которая позволяет оперативно получить значения показателей описательной статистики При этом вычисляются следующие статистические показатели· среднее арифметическое, ошибка среднего, медиана, мода, стандартное отклонение (среднеквадратическое отклонение), дисперсия, эксцесс, асимметрия, интервал, минимум, максимум, сумма, счет (размер выборки) и уровень надежности.

Рассмотрим пример. В цехе синтеза аммиака производились замеры концентрации аммиака в воздухе рабочей зоны. Получен ряд значений (в мг/м^3) 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 13,17 15 и 14. Необходимо определить параметры описательной статистики для этого ряда

Порядок решения задачи в *MS Excel*:

1 Введите в ячейки A1:A15 исходные данные.

2. Выполните команду <Пакет анализа> из меню <Сервис>

Примечание: Если данная команда отсутствует, то необходимо с помощью команды <**Надстройки**> из меню <**Сервис**> открыть окно диалога «**Надстройки**» и в нем установить флажок для компоненты «**Пакет анализа**». После нажатия кнопки **[ОК]** меню <Сервис> будет дополнено командой <**Пакет анализа**>

Если появится сообщение, что выбранная надстройка не может быть найдена, то необходимо выполнить более полную инсталляцию пакета программ *MS Excel* на вашем компьютере.

3 Выберите в появившемся диалоговом окне метод «Описательная статистика» и нажмите кнопку [ОК]

4. В окне «Описательная статистика» установите следующие параметры

- Входной диапазон ($A$1:$A$15)
- Группирование (по столбцам)
- Метки (входной диапазон не содержит метки, т. е названий строк и столбцов)
- Альфа (уровень значимости=0,05).
- Выходной диапазон. Для выходного диапазона выберите ссылку на ячейку С1, расположенную в левом верхнем углу выходного диапазона. Полные размеры выходной области будут определены автоматически
- Итоговая статистика Установите флажок, чтобы в выходном диапазоне получить следующие характеристики: среднее (арифметическое), стандартную ошибку среднего, медиану, моду, стандартное отклонение, дисперсию выборки, эксцесс, асимметричность, амплитуду (интервал), минимум, максимум, сумму вариант и размер выборки (счет)
- Уровень надежности Установите флажок, чтобы в выходную таблицу включить строку для уровня надежности Уровень надежности — это половина доверительного интервала для генерального среднего арифметического

5. После завершения настройки параметров нажмите кнопку [ОК]

Результаты анализа принимают вид (рис. 75).

Из полученных данных следует, что с вероятностью 0,95 среднее арифметическое для генеральной совокупности находится в интервале 14,867±1,338 Амплитуда (интервал) разброса значений вариант равна 10.

	A	B	C	D
1	12		*Столбец1*	
2	16			
3	15		Среднее	14,867
4	14		Стандартная ошибка	0,624
5	10		Медиана	15
6	20		Мода	14
7	16		Стандартное отклонение	2,416
8	14		Дисперсия выборки	5,838
9	18		Эксцесс	0,931
10	14		Асимметричность	0,180
11	15		Интервал	10
12	13		Минимум	10
13	17		Максимум	20
14	15		Сумма	223
15	14		Счет	15
16			Уровень надежности(95,0%)	1,338

Рис. 75. **Результаты вычисления показателей описательной статистики**

5. Теоретические распределения

Как уже говорилось ранее, различные методы описательной статистики позволяют представить в обобщенном виде свойства статистической совокупности Важнейшее звено статистического анализа — аналитическое описание кривой конкретного распределения в виде закона распределения **Теоретическим называют распределение, которое выбирается как образец (стандарт) для описания закона фактического распределения.**

К задачам, решаемым на основе описания этой кривой, следует в первую очередь отнести

- обоснование выбора конкретных методов статистического анализа генеральной и выборочной совокупности,

- изучение причин, обусловливающих тот или иной закон распределения эмпирических данных (данных полученных опытным путем) Вскрытие на этой основе естественных причинно-следственных отношений и взаимосвязей различных явлений;
- построение статистических моделей, научное прогнозирование событий и процессов

В практике медико-биологических исследований встречаются как дискретные (прерывные), так и непрерывные распределения

Наиболее распространенные среди дискретных распределений — биноминальное и пуассоновское распределения Среди непрерывных — нормальное и связанные с ним распределения Стьюдента, *хи*-квадрат и *F* — распределение Фишера, которые особенно широко используются при построении доверительных интервалов и статистической проверке гипотез

Биноминальное распределение (*распределение Бернулли*) — возникает, когда оценивается, сколько раз происходит некоторое событие в серии определенного числа независимых, выполняемых в одинаковых условиях наблюдений При этом распределении разброс вариант (в простейшем случае есть событие или нет события) является следствием влияния ряда независимых и случайно сочетающихся факторов *Например* для контроля качества партии фармакологического препарата требуется подсчитать число изделий (упаковок), не соответствующих требованиям Все причины, влияющие на качество препарата, принимаются одинаково вероятными и не зависящими друг от друга Сплошная проверка качества в этой ситуации не возможна, поскольку изделие, прошедшее испытание, не подлежит дальнейшему использованию. Поэтому для контроля из партии наудачу выбирают определенное количество образцов изделий (*n*) Эти образцы всестороннее проверяют и регистрируют число бракованных изделий (X) Теоретически число бракованных изделий может быть любым, от 0 до *n*, но вероятности этих чисел различны В основе принятия решения лежит сравнение распределения результатов контроля, т е данных, полученных опытным путем, и теоретического распределения, при котором гарантиру-

ется необходимое качество всей партии — достаточно низкая вероятность брака $P(X=k)$, где k — число возможных событий (в примере одно событие — изделие с браком, второе — изделие без брака).

Распределение Пуассона — проявляется в ситуациях, когда в течение определенного отрезка времени или на определенном пространстве происходит случайное число каких-либо событий (число радиоактивных распадов, выпадение частиц аэрозоля, случаи заболеваний и т п) Основное отличие этого закона распределения — резко выраженная асимметрия Если обозначить вероятность одного события через p, то вероятность другого события q будет равна (1—p) Распределение этих событий будет тем асимметричней, тем больше различаются p и q В крайнем случае, когда p очень мало, а q, соответственно, близко к 1, получается крайне асимметричное распределение — распределение Пуассона Особенностью этого распределения является то, что дисперсия и среднее арифметическое при таком распределении равны между собой ($D=M$) Это избавляет от необходимости вычислять дисперсию, что очень удобно при определении ошибки репрезентативности (статистической достоверности) выборочных исследований, когда вычисление дисперсии крайне затруднено или вообще невозможно Необходимо только соблюдать главное условие распределения Пуассона — независимость событий При оценке заболеваемости, *например*, такая методика используется только в случае, если заболевания не связаны (или весьма слабо связаны) между собой причинно-следственными отношениями. Кроме того, равенство дисперсий и среднего в распределении Пуассона позволяет проводить ориентировочную оценку эмпирических (опытных) распределений *Например* если в ходе статистической обработки дискретных рядов получены среднее и дисперсия, которые равны между собой, то распределение можно считать подчиняющимся закону Пуассона

Распределение Пуассона и биноминальное распределение связаны между собой и с некоторыми другими типами распределений Среди них наиболее существенным представляется схожесть с нормальным распределением

5.1. Нормальное распределение

Нормальное распределение (распределение Гаусса) относится к числу законов распределения, используемых для приближенного описания явлений, которые носят вероятностный, случайный характер В связи с тем что предмет подавляющего большинства медико-биологических исследований — явления вероятностного характера, нормальное распределение в таких исследованиях встречается весьма часто. Особенностью этого распределения является и то, что многие другие формы распределения в предельных случаях (при предельно большом числе наблюдений) подчиняются гауссову закону Исторически приоритет в открытии этого одного из основополагающих законов математической статистики принадлежит Де Муавру (1733), но обычно его связывают с именем Гаусса, исследовавшего его в начале XIX века

Например возьмем некоторое биноминальное распределение с параметром K=11 Затем удвоим этот параметр (K=22) При этом уменьшим вдвое ширину каждого интервала (разряда), чтобы второе распределение имело бы ту же ширину, что и первое. Кроме того, если уменьшить и каждую частоту интервала вдвое, то оба распределения будут иметь одинаковый объем После этих операций над вторым распределением (сужение каждого разряда и уменьшение вдвое каждой частоты), оба распределения будут лишь несколько различаться своей формой (рис 76)

Если продолжить эту процедуру, беря все большие значения K (K=33), то в пределе $K = \infty$ верхние стороны столбиков сольются в гладкую кривую

Эту кривую называют **«кривая нормальиого распределения непрерывной величины»**. Как и всякую кривую, кривую распределения можно описать некоторой функцией $y = f(x)$. Эта функция указывает, чему равна ордината y, соответствующая заданному значению абсциссы x Выражение $f(x)$ называется **плотностью функции распределения** (плотностью вероятностей, плотностью распределения вероятностей) Плотность распределения характеризует *площадь*, которая ограничивается отрезком кривой распределения в интервале разряда Полная площадь, ограниченная кривой распределения, принимается равной 1

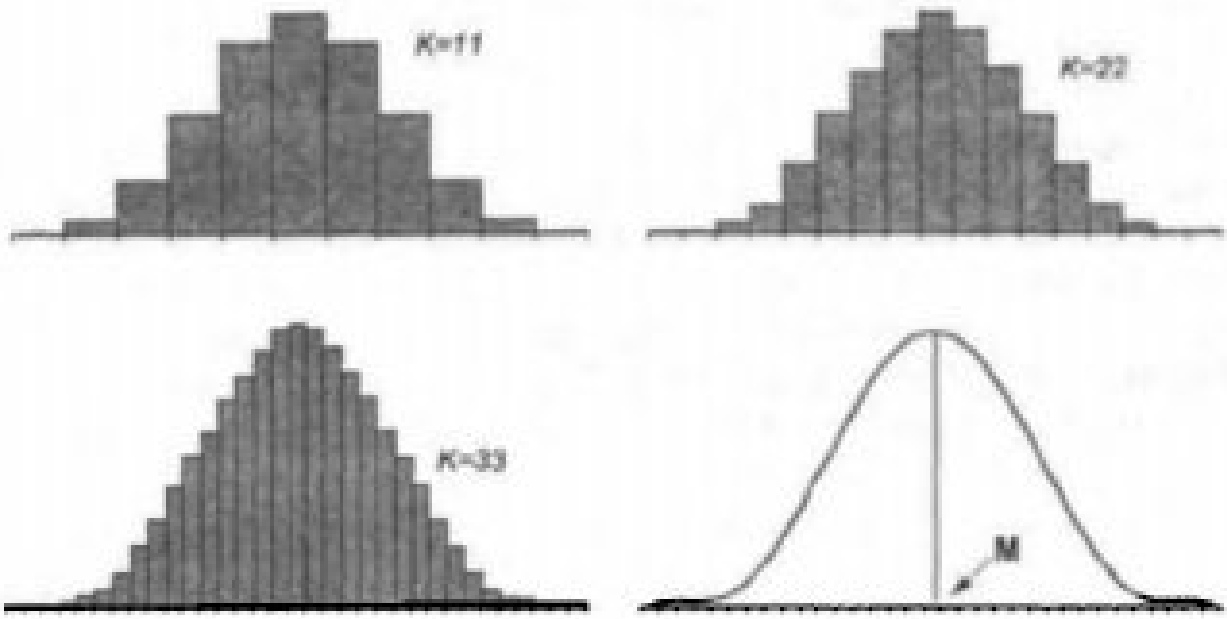


Рис. 76. **Трансформация различных вариантов распределений**

Такое распределение называется нормальным или гауссовым (распределением Гаусса) Следовательно, оно покоится на тех же предпосылках, что и биноминальное, т. е на гипотезе о малых независимых, случайно сочетающихся факторах Условие $K = \infty$ означает, что число этих факторов неограниченно велико Однако если какой-либо фактор играет преобладающую роль, то распределение не будет подчиняться гауссову закону Таким образом, нормальное распределение не следует считать универсальным.

Уравнение гауссовой кривой имеет вид

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(M-a)^2}{2\sigma^2}},$$

где M — среднее теоретическое значение, e — основание натуральных логарифмов (e =2,718..), σ — среднеквадратическое отклонение.

Диаграмма нормального распределения симметрична относительно точки M, т е положительные и отрицательные отклонения равной величины от центра распределения (среднего ариф-

метического) встречаются одинаково часто При этом в точке M функция достигает своего максимума Совокупность нормальных распределений представляет собой двухпараметрическое семейство Параметр M характеризует положение диаграммы функции на числовой оси (параметр положения) Параметр σ характеризует степень сжатия или растяжения (плотности) диаграммы Чем больше σ, тем «шире» кривая, а ее максимальная высота ниже Кривая как бы «расплывается» в стороны При малых σ кривая, наоборот, «стягивается» к середине При этом подразумевается, что совокупности имеют одинаковый объем (рис. 77).

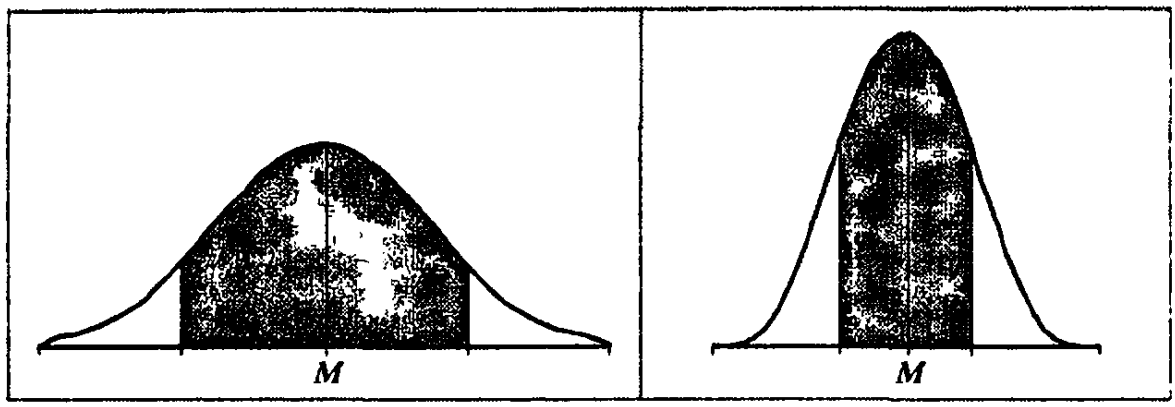


Рис. 77. **Различные варианты нормального распределения**

На представленных рисунках заштрихованная площадь отражает часть вариант, отклоняющихся от среднего значения не более чем на σ, т е от $-\sigma$ до $+\sigma$. В этой области при нормальном распределении всегда оказывается 68,3% всех вариант Таким образом, 68,3% всех вариант отклоняются от среднего значения не более чем на величину σ В пределах от -2σ до $+2\sigma$ лежат 95,5% всех вариант В пределах от -3σ до $+3\sigma$ — 99,7% и т д Эта закономерность трактуется как **правило трех сигм.**

Параметры M и σ являются параметрами, описывающими различные модификации нормального распределения. Те статистические критерии, которые используют эти параметры, принято называть **параметрическими критериями**

5.2. Критерии совпадения эмпирических и теоретических распределений. Статистические оценки нормальности распределения

Оценка меры совпадения теоретических и эмпирических распределений, полученных опытным путем, одна из ключевых задач любого статистического исследования Вместе с тем, самый распространенный дефект статистического анализа в научных работах медико-биологического профиля — оценка только средних величин Из-за этого под сомнение могут ставиться все статистические выкладки, имеющиеся в таких работах. Проведение любого углубленного статистического анализа должно обязательно включать статистическую оценку различий распределений

Следует отметить, что, формально, задача нахождения различий между контрольными и опытными группами (или несколькими группами) методически близка задаче нахождения различий между эмпирическими и теоретическими распределениями. В такой ситуации исключается только этап формирования теоретического распределения, которое заменяется данными о контрольной группе или другой наблюдаемой группе.

Точную информацию о форме полученного в результате опыта распределения и о точности его совпадения с теоретическим распределением можно получить с помощью специальных критериев нормальности, например критерия Колмогорова—Смирнова Однако наиболее наглядную картину дает визуальная проверка схожести распределений с помощью гистограммы, которая показывает частоту попаданий вариант ряда распределения в отдельные интервалы (рис. 80).

5.2.1. Нахождение нормального распределения с помощью *MS Excel*

Известно, что нормальное распределение задается центром распределения (средним арифметическим) и разбросом вариант (дисперсией или среднеквадратическим отклонением).

	A	B	C	D	E	F	G
1	Варианты	Частоты			Фактические	Теоретические	Теоретические
2	V	P	VP	V^2P	Частости	Частости	Частоты (P)
3	21	2	42	882	0,032	0,022	1,4
4	22	9	198	4356	0,145	0,079	4,9
5	23	6	138	3174	0,097	0,179	11,1
6	24	11	264	6336	0,177	0,259	16,1
7	25	18	450	11250	0,290	0,241	14,9
8	26	15	390	10140	0,242	0,144	8,9
9	27	1	27	729	0,016	0,066	3,7
10	Итого	62	1509	36867			62,0

Рис. 78

Для *примера* построим ряд теоретических частот нормального распределения (рис. 78). Для этого.

1 Введите в ячейки A1 A9 и B1 B9 исходные данные

2. Чтобы рассчитать среднее взвешенное и дисперсию, предварительно вычислим параметры VP и V^2P. С этой целью введите в ячейку C3 формулу =B3*A3. В ячейку D3 — формулу B3*A3*A3. Выделите блок ячеек C3:D3 и скопируйте в блок ячеек C4:D9. Найдите суммы блоков ячеек с помощью формул: =СУММ(B3 B9) в ячейке B10; СУММ(C3:C9) в ячейке C10 и в ячейке D10 =СУММ(D3:D9).

3. Для того чтобы перевести частоты в частости, в ячейке E3 наберите формулу =B3/B$10. Затем скопируйте ее в блок ячеек E4:E10.

4. Чтобы вычислить среднее арифметическое $M = \frac{\sum VP}{n}$, установите курсор в ячейку C12 и наберите формулу C10/B10 Дис-

персия вычисляется по формуле $D=\frac{V^2P}{n}-M^2$ В ячейке C13 введите формулу D10/B10−C12*C12. Среднеквадратическое отклонение через дисперсию $\sigma=\sqrt{D}$ в ячейке C14. =**КОРЕНЬ**(C13)

5 Теоретические частоты находим с помощью функции **НОРМРАСП.**

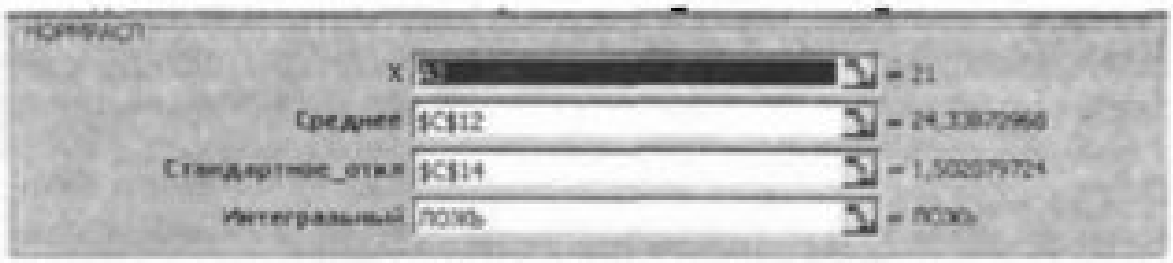


***Рис. 79.* Пример заполнения окна функции НОРМРАСП**

Установите курсор в ячейку F3, вызовите указанную функцию и заполните ее окна как показано на рис. 79. Затем скопируйте содержимое ячейки F3 в блок ячеек F4:F9.

6 Для того чтобы заменить теоретические частоты теоретическими частотами, в ячейку G3 введите формулу =B$10*F3 Затем скопируйте ее в остальные ячейки столбца G4:G9.

5.2.2. Критерий согласия *Пирсона* χ^2

Для оценивания меры соответствия (расхождения) полученных эмпирических данных и каких-либо теоретических распределений применяются различные статистические критерии Критерии, которые используют с этой целью, называются *критериями согласия.*

В наиболее общем виде **критериями согласия называют статистические критерии, предназначенные для проверки согласия опытных данных и теоретической модели.**

Среди критериев согласия большое распространение получил непараметрический критерий К. Пирсона — ***критерий*** χ^2 (хи-квадрат). Одной из причин такой популярности является

возможность его использования с различными формами распределений совокупностей. Как и любой другой статистический критерий, он не доказывает справедливость нулевой гипотезы, а лишь устанавливает с определенной вероятностью ее согласие или несогласие с данными наблюдений. Величина критерия χ^2 — есть величина случайная, так как в различных опытах она принимает неизвестные заранее значения Не вдаваясь в математическое описание причин, следует отметить, что при $n \to \infty$ распределение этой случайной величины стремится к закону распределения χ^2 с k степенями свободы. Поэтому критерий и назван по названию закона теоретического распределения χ^2 — *критерий* χ^2 (хи-квадрат).

Вычисление критерия χ^2 (хи-квадрат) на конкретных данных представлено в табл 72

Таблица 72

Пример вычисления критерия χ^2

Вес детей (кг)	Частоты (число детей)		Объединенные частоты		$P - P'$	$\frac{(P-P')^2}{P'}$
	Эмпирич P	Теоретич P'	Эмпирич P	Теоретич P'		
21	2	1,4				
22	9	4,9	11	6,3	4,7	0,36
23	6	11,1	6	11,1	−5,1	0,41
24	11	16,1	11	16,1	−5,1	0,41
25	18	14,9	18	14,9	3,1	0,15
26	15	9,9	16	13,6	2,4	0,09
27	1	3,7				
Итого	62	62	62	62		1,42

Условием применения этого метода является наличие не менее 5 наблюдений (частот) в каждой группе вариант Для того чтобы это условие выполнялось, необходимо объединить малые частоты (в таблице объединенные частоты) После этого, значения критерия вычисляются в следующей последовательности:

1 Вычисляется разность между частотами эмпирических P_i и теоретических P_i' распределений.

2. Чтобы погасить отрицательные значения этой разности она возводится в квадрат $(P_i - P_i')^2$. Затем, для уменьшения числового размера результата, делится на величину эмпирических частот

$$\chi^2 = \frac{(P_i - P_i')^2}{P_i}$$

В целом, χ^2 (хи-квадрат) тем меньше, чем меньше различаются эмпирические и теоретические частоты

3. Число степеней свободы k, которое необходимо знать для определения справедливости нулевой гипотезы, устанавливается в зависимости от вида распределения При биноминальном (Пуассоновом) распределении $k=S-2$, при нормальном $k=S-3$ Где S — число групп вариант С учетом объединения групп, $k=5-3=2$.

4 Полученное значение критерия сравнивают с табличным или найденным с помощью компьютера теоретическим значением при заданном уровне значимости. В случае, когда вычисленный критерий больше табличного (теоретического) значения критерия χ^2, нулевая гипотеза, которая предполагает соответствие эмпирического и теоретического распределений, отвергается

Способом определения достоверности различий распределений, с помощью которого можно выполнять оценку вручную и без использования таблиц, является оценка χ^2 по правилу Романовского или по формуле Б. С Ястремского.

По Романовскому нулевая гипотеза отвергается, если неравенство $(\chi^2 - k)/\sqrt{2k} > 3$ В данном примере $(1{,}42-2)/\sqrt{2\times 2} = -0{,}29$ Следовательно, опытное, эмпирическое распределение не отличается от нормального

По Б С. Ястремскому нулевая гипотеза отвергается, если выполняется неравенство

$$(\chi^2 - k)/\sqrt{2k - 4\theta} > 3,$$

где θ — поправка на число групп вариант. При $k<20$ она равна 0,6. Подставляя значения из примера, получаем $(1,42-2)/\sqrt{2\times0,6\times4}=$ $=-0,37$, что также не позволяет отвергнуть нулевую гипотезу

Следует отметить, что критерий χ^2 как критерий согласия обладает существенным недостатком. Он не позволяет иногда обнаружить реально существующие различия, поскольку некоторые «скрадывает» группировка, которая осуществляется исследователем в ходе начальной подготовки рядов распределений к анализу

В ряде случаев ориентировочную оценку расхождения или совпадений распределений позволяет давать графический метод (рис. 80)

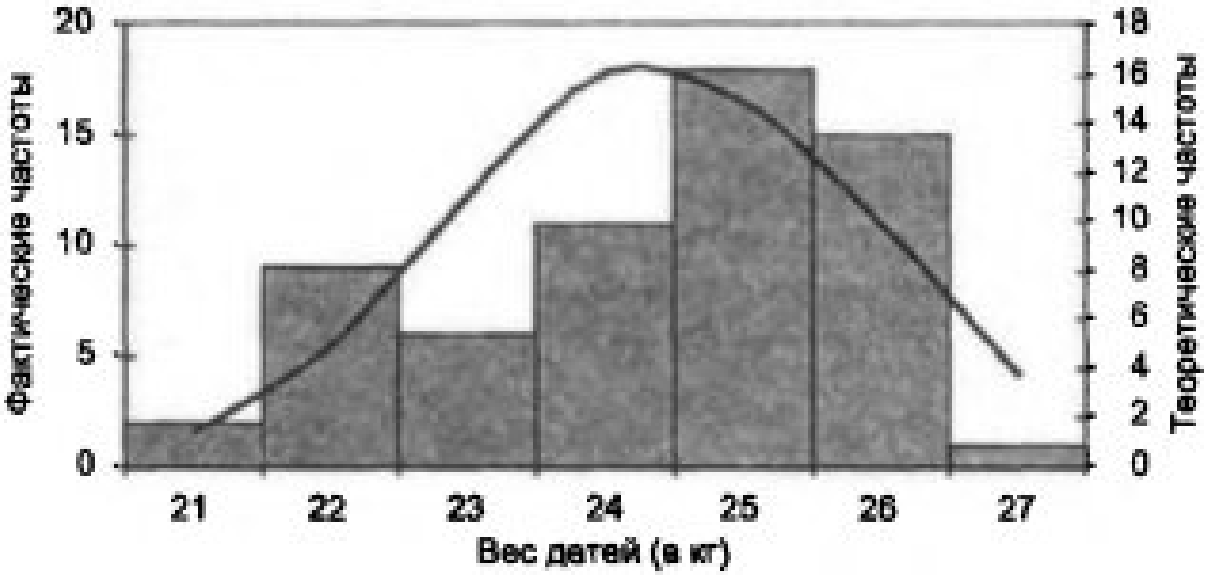


Рис. 80. **Соотношение теоретического и эмпирического распределений**

На представленной диаграмме эмпирическое распределение дано в виде гистограммы, а теоретическое — в виде кривой распределения.

Другие методики вычисления критерия χ^2, которые могут использоваться для решения задач оценки согласия распределений, представлены в разделах «Оценка различий эмпирических распределений с помощью *Excel*», «Коэффициенты сопряженности Пирсона (*C*) и *Чупрова* (*K*)», а также «Вычисление критерия сопряженности в *MS Excel*»

5.2.3. Критерий согласия *Колмогорова* $K(\lambda)$

Не сложен для практического применения и критерий Колмогорова, обозначаемый $K(\lambda)$

$$K(\lambda)=D_{max}\sqrt{N},$$

где N — число наблюдений в статистическом ряду, D_{max} — максимальная разница в накопленных эмпирических и теоретических частостях. Ограничением при его применении является возможность сравнивать с его помощью лишь взвешенные ряды. Число наблюдений при этом должно быть достаточно большим ($N \geq 50$) Не рекомендуется также объединять варианты в крупные группы Сама анализируемая совокупность должна представлять непрерывный ряд распределения Результат вычисления оценивается весьма просто, в зависимости от доверительного уровня P. Нулевая гипотеза отвергается, если $K(\lambda)>1,36$ (при $P=0,95$) или если $K(\lambda)>1,63$ (при $P=0,98$), или если $K(\lambda)>1,95$ (при $P=0,999$). Нередко название критерия включает и фамилию А. Н. Смирнова (*критерий Колмогорова—Смирнова*), который обосновал применение этого критерия для оценки различий распределения эмпирических рядов, т е двух рядов данных, полученных в результате опыта (*например* «опыт» и «контроль»)

В отличие от критерия Пирсона χ^2, критерий $K(\lambda)$ не требует расчета степеней свободы и не требует проводить объединения малых групп. Кроме того, помимо оценки значимости расхождений, критерий Колмогорова позволяет установить и величину расхождений, выявить диапазон (градацию) признака, которая играет наиболее существенную роль в расхождении распределений (табл 73)

Таблица 73

Пример вычисления критерия Колмогорова $K(\lambda)$

Частоты		Накопленные частоты		Накопленные частости		D
Эмпирические	Теоретические	Эмпирические n	Теоретические n'	Эмпирические n	Теоретические n'	
1	2	3	4	5	6	7
1	3,41	1	3,41	0,003	0,010	−0,007
42	21,63	43	25,04	0,131	0,076	0,055
65	68,42	108	93,46	0,329	0,285	0,044
77	108,10	185	201,56	0,564	0,615	−0,050
98	85,28	283	286,84	0,863	0,875	−0,012
29	33,59	312	320,43	0,951	0,977	−0,026
16	6,61	328	327,04	1,000	0,997	0,003
N=328	327,04	$K(\lambda) = 0,055 \times \sqrt{328} = 0,99$				

В табл 73 представлены частоты распределения обследованных работников предприятия по показателям форсированной жизненной емкости легких (ФЖЕЛ). Значения вариант (показателей ФЖЕЛ) в таблице не приводятся, так как они не играют роли при вычислении $K(\lambda)$. Последовательность вычисления этого критерия следующая

1 В первом столбце таблицы располагаются частоты сверху вниз в порядке возрастания вариант ряда В столбце 2 находятся соответствующие им теоретические частоты нормального распределения

2 В столбце 3 приведены накопленные эмпирические частоты. 1, 1+42=43; 43+65=108, 108+77=185 и т д Аналогичным образом определяются накопленные теоретические частоты в столбце 4 3,41+21,63=25,04, 25,04+68,42=93,46; 93,46+108,10= =201,56 и т д

3 В столбец 5 заносятся накопленные эмпирические частости Получаются они делением накопленной частоты на общее число наблюдений 1/328≈0,003; 43/328≈0,131, 108/328≈0,329 и т д Таким же образом рассчитываются накопленные теоретические частости в столбце 6 3,41/328≈0,01, 25,04/328≈0,076 и т д

4 В столбец 7 заносим разность накопленных эмпирических и теоретических частостей. 0,003−0,01=−0,007, 0,131−0,076=0,055 и т д

5 Для дальнейших расчетов нужна только одна разность из столбца 6, максимальная. При выборе числа его знак (плюс или минус) не учитывается В данном случае это 0,055 По формуле $K(\lambda)=D_{max}\sqrt{N}$ получаем критерий $K(\lambda)=0,99$ Поскольку это значение меньше чем 1,36, что обеспечивает доверительную значимость $P = 0,05$, делаем вывод о том, что нулевая гипотеза не отвергается Таким образом, предположение о соответствии распределения эмпирических частот и нормального распределения остается в силе

Построение любого теоретического распределения является отдельной и трудоемкой задачей. Ее решение для нормального распределения требует использования некоторых исходных параметров эмпирического распределения формулы

$$\varphi(x)=\frac{N}{\sqrt{2\pi\sigma}}e^{-\frac{(M-x)^2}{2\alpha^2}}.$$

Использование встроенных функций *MS Excel* значительно облегчает задачу

В заключении раздела следует отметить, что, строго говоря, критерий *Пирсона* χ^2 и критерий *Колмогорова* $K(\lambda)$ использовать как критерии проверки гипотез согласия и однородности следует весьма осторожно Это обстоятельство связано с тем, что они не позволяют обнаружить различия, которые «скрадывает» группировка *Например*·

Таблица 74

Различные варианты распределения одной и той же группы по возрасту (в % к итогам)

Возраст (лет)	I вариант группировки	Возраст (лет)	II вариант группировки
до 20	12	–	–
20–29	22	до 29	34
30–39	33	30–39	33
40 и ст	33	40 и ст	33
Итого	100	Итого	100

Анализируя представленные данные (табл 74), нетрудно заметить, что при втором варианте группировки можно сделать ошибочное заключение об отсутствии различий в возрастных группах. В то же время, критерий Колмогорова позволяет выделить наиболее существенную разность распределений в отдельном диапазоне признака В приведенном примере (см табл 73) — это диапазон признака, т е. величина ФЖЕЛ (не приведенная в таблице), которая встретилась 42 раза

6. Статистическая (корреляционная) связь между признаками. Основные виды связи

Целью большинства медико-биологических исследований, в крайнем случае одной из задач таких исследований, является выявление взаимосвязи одного или нескольких явлений. Формальная постановка задачи в такого рода исследованиях обычно выглядит следующим образом «Определить наличие и силу статистической связи какого-либо признака, от одного или нескольких других признаков» или «Определить наличие и силу статистической связи какой-либо величины, характеризующей результативный признак, от одного или нескольких факторных признаков» Знание взаимосвязи отдельных признаков дает возможность решать одну из кардинальных задач любого научного исследования возможность предвидеть, прогнозировать развитие ситуации при изменении тех или иных известных характеристик объекта исследования. Строго говоря, термин *зависимость* при статистической обработке материалов медико-биологических исследований должен использоваться весьма осторожно. Это связано с природой статистического анализа, который сам по себе не может вскрыть истинных причинно-следственных от-

ношений между факторами, взаимосвязь которых нередко опосредована через третьи факторы Причем, эти третьи факторы могут лежать вообще вне поля зрения исследователя С помощью статистических критериев можно дать только формальную оценку взаимосвязей ***Попытки механически перенести данные статистических расчетов в объективную реальность могут привести к ошибочным выводам*** Например, утверждение: «Чем громче утром кричат воробьи, тем выше встает солнце», несмотря на явную несуразность, с точки зрения формальной статистики, это утверждение вполне правомерно Таким образом, термин «*зависимость*» в статистическом анализе подразумевает только оценку соответствующих статистических критериев

Предварительно здесь следует остановиться на двух ключевых принципах статистической оценки связи, обычно оставляемых большинством исследователей без внимания, — *принципе ковариаций* и *принципе взаимной сопряженности*

Если основанием для заключения о наличии связи служит одновременное и параллельное изменение количественных характеристик, то подобное заключение основано на принципе *ковариации* В математическом отношении задача сводится к определению меры взаимных численных изменений взаимосвязанных признаков В качестве таких мер на практике особенно широко используют коэффициент корреляции, коэффициенты ранговой корреляции Спирмена, Кендала и др

Принцип взаимной сопряженности предполагает установление связи между двумя событиями в тех случаях, когда с появлением одного события происходит другое событие Например: свет в окне может означать (с той или иной вероятностью), что хозяева находятся дома, кашель с мокротой может означать заболевание хроническим бронхитом Если в серии повторяющихся наблюдений один из признаков (или его часть, градация) появляется одновременно с другим чаще, чем можно объяснить случайным стечением обстоятельств, то это служит основанием говорить о взаимосвязи, сопряженности появления этих признаков Свидетельством присутствия *сопряженности* (взаимосвязи) является определенный уровень статистической значимости, отличия от нулевой гипотезы, отрицающей такую взаимосвязь

Коэффициенты, основанные на принципе взаимной сопряженности, составляют довольно многочисленную группу непараметрических критериев ***Q*** (коэффициент ассоциации Юла), ***Ф*** (коэффициент сопряженности, контингенции), ***C*** (коэффициент взаимной сопряженности Пирсона) ***K*** (коэффициент Чупрова) и др

Здесь важно помнить, что различные коэффициенты, основанные на принципах оценки сопряженности, как и коэффициенты, основанные на принципах ковариации, дают, каждый по-своему, оценку одного явления — корреляционной связи (коэффициент корреляции — один из многих среди них).

Любые явления в окружающем нас мире могут быть связаны прямой или обратной связью Эта характеристика называется **направленностью связи** По направленности связь может быть прямой или обратной

Прямая связь характеризует зависимость, при которой увеличение или уменьшение значения одного признака ведет, соответственно, к увеличению или уменьшению — второго *Например:* при увеличение температуры возрастает давление газа (при сохранении неизменным его объема) При уменьшении температуры — снижается и давление.

Обратная связь характеризуется такой зависимостью, когда при увеличении одного признака, второй — уменьшается Или наоборот, при уменьшении одного, второй — увеличивается Обратная зависимость или обратная связь является основой нормального регулирования почти всех процессов жизнедеятельности любого организма

Всякая из этих зависимостей **по характеру связи** может быть *функциональной* или *статистической (корреляционной)*

Функциональная зависимость — такой вид зависимости, когда каждому значению одного признака соответствует точное значение другого *Например* взаимосвязь площади круга *(S)* и длины окружности *(L)* Известно, что площадь круга и длина окружности связаны вполне определенным отношением $S=\frac{r}{2}L$, где r — радиус круга Умножив длину окружности на половину радиуса

круга, можно точно определить площадь круга Такую зависимость можно считать полной (исчерпывающей) Она полностью объясняет изменение одного признака изменением другого Этот вид связи характерен для объектов, являющихся точкой приложения точных наук В медико-биологических исследованиях сталкиваться с функциональной связью приходится крайне редко, поскольку обьекты этих исследований имеют большую индивидуальную вариабельность (изменчивость) С другой стороны, характеристики биологических объектов зависят, как правило, от комплекса большого числа сложных взаимосвязей и не могут быть сведены к отношению двух или трех факторов

Иной вид зависимости — **статистическая (корреляционная) зависимость.** В этом случае при изменении величины одного признака изменяется тенденция (характер) распределения значений другого признака и, соответственно, характеристики этого распределения Например средние значения изучаемых признаков Если величины X и Y находятся в статистической связи, то это не означает, что при изменении величины X величина Y обязательно будет изменяться определенным образом Это означает только то, что при достаточно большом числе наблюдений изменение величины X сопровождается, как правило, изменением величины Y. Такая тенденция существует только в общих чертах *Например.* при изменении роста человека меняется и масса тела. Однако эта зависимость не является полной, т е функциональной. У людей с одинаковым ростом может быть разная масса тела, поскольку на нее влияют и многие другие факторы (питание, здоровье и т п) При оценке статистических связей можно говорить только о тенденции, когда возрастание одного признака вызывает тенденцию возрастания или уменьшения другого признака

Следует отметить, что в случае биологических факторов тот или иной характер связи сохраняется, как правило, только в определенном интервале изменений признаков За пределами этого интервала связь может ослабнуть, стать прямо противоположной по направлению либо совсем исчезнуть *Например*, при увеличении возраста ребенка сила скелетной мускулатуры уве-

личивается В зрелом возрасте такой связи уже нет А в старших возрастных группах тенденция становится обратной

Статистическая (корреляционная) связь вскрывается и описывается с помощью различных статистических характеристик, получаемых различными методами. Выбор метода для определения взаимосвязей обусловлен видом самих признаков, способами их группировки и предполагаемым характером связи. Подчас для выявления реально существующих взаимосвязей достаточно правильно составить статистическую таблицу распределения или построить наглядный график этого распределения.

В том случае, если для определения взаимосвязей используются характеристики нормального распределения (например, средние величины), то такие критерии называют критериями **корреляционной связи.** Строго говоря, корреляционная связь является *частным случаем статистической связи* Вместе с тем, термин корреляционная связь достаточно широко распространен и повсеместно употребляется для обозначения любого вида статистической связи

В ходе корреляционного анализа или анализа корреляционной связи решается целая группа взаимосвязанных задач.

1 Установление направления (прямая или обратная) и формы (линейная или нелинейная) корреляционной связи

2. Оценка тесноты (силы, плотности) корреляционной связи

3. Оценка репрезентативности статистических оценок взаимосвязей, полученных по выборочным данным (величина ошибки, доверительный интервал, уровень значимости).

4. Установление величины детерминации (доли взаимовлияния) коррелируемых факторов.

Обычно используют следующие численные критерии (коэффициенты) корреляционной связи

1 Для оценки взаимосвязи парных количественных признаков применяют коэффициенты линейной или нелинейной регрессии и коэффициенты (линейной или нелинейной) корреляции

2. Для оценки взаимосвязи нескольких количественных признаков применяют коэффициенты множественной регрессии и коэффициенты множественной или частной, парциальной корреляции

3 Для признаков, сформированных в порядковой (ранговой, балльной) шкале, можно применять ранговые коэффициенты линейной корреляции Пирсона или Кендэла.

4 Для оценки взаимосвязи качественных признаков, можно применять рассчитываемые на основе таблиц сопряженности коэффициент ассоциации Юла (Q), коэффициент контингенции (Φ), взаимной сопряженности (C), Чупрова (K) и др.

6.1.1. Регрессия

При анализе количественных данных показателями формы связи служат линия регрессии и коэффициент регрессии Уравнения линии регрессии относительно осей координат называют **уравнениями регрессии**. Иногда эти уравнения называют *уравнениями корреляционной связи* или *уравнениями корреляции* Тип уравнений зависит от формы связи, которая определяется по форме корреляционного поля. **Корреляционное поле** представляет собой график, отображающий распределение значений Y и X в виде точек с соответствующими абсциссами $0x$ и ординатами $0y$ Разброс точек на графике визуально представляет тесноту (плотность) связи

Если связь слабая или отсутствует, то точки распределяются бессистемно по всей площади графика в пределах значений, которые могут принимать Y и X (рис. 81).

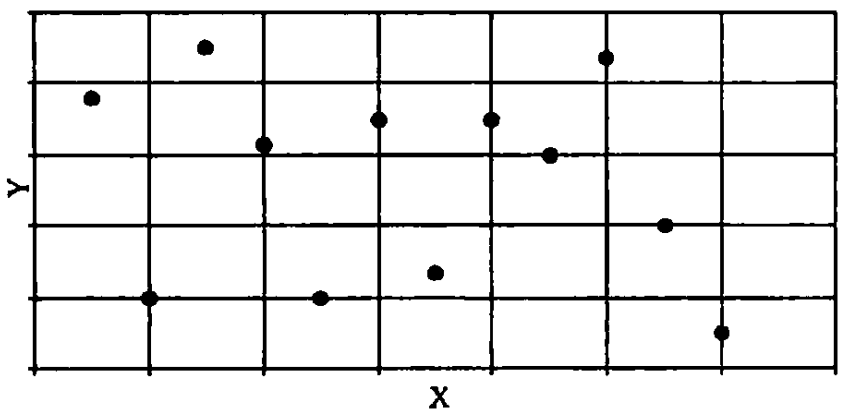


Рис. 81. **Корреляционное поле с малой плотностью связи**

Если связь сильная (плотная), то точки располагаются плотно, вдоль некоторой результирующей линии, которая называется **линией регрессии.**

Чем более тесна (плотна) корреляционная связь, тем более тесно около линии регрессии располагаются точки корреляционного поля На графике можно увидеть и направление связи прямая или обратная (рис. 82, 83)

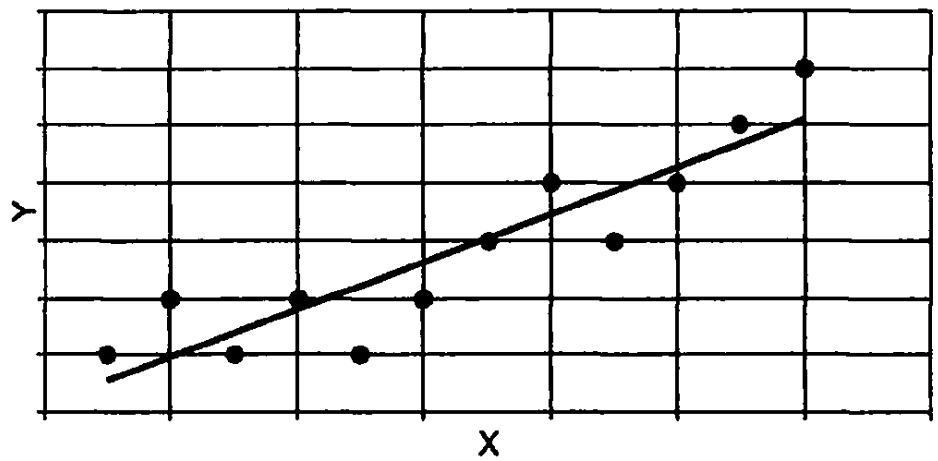


Рис 82. **Корреляционное поле сильной линейной прямой связи**

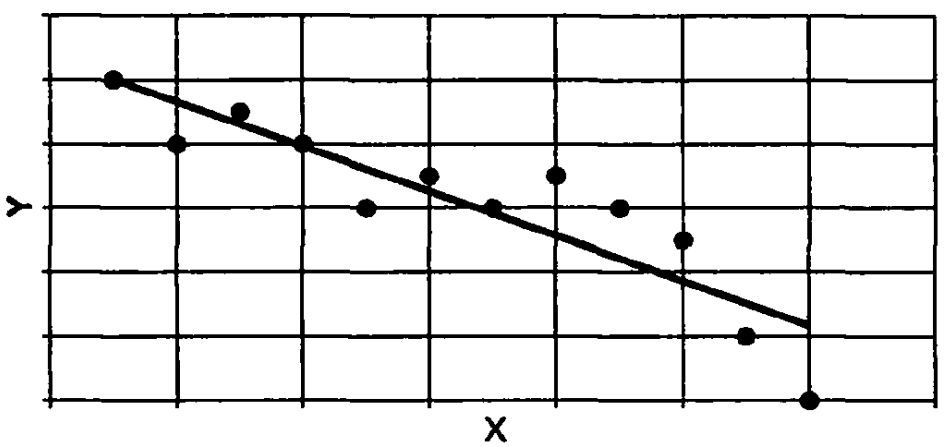


Рис 83. **Корреляционное поле сильной линейной обратной связи**

В случае линейной зависимости y от x уравнением регрессии является уравнение прямой $y = a + bx$, где y — значение результативного признака (зависимая переменная), x — значения факторного признака (независимая переменная), a и b — коэффициенты

Простейшим примером линейного уравнения регрессии может служить индекс Брока, который используется как росто-весовой индекс для исчисления нормального веса. из роста вычитают 100 и получают нормальный вес, соответствующий этому росту. Математически этот индекс записывается в виде уже приводившегося уравнения линейной регрессии $y=a+bx$, где y — вес, x — рост, $a=-100$, b — поправочный коэффициент, который изменяется для разных возрастных групп Наиболее известные антропометрические индексы представлены в разделе «Физическое развитие» настоящего издания (см табл 14)

Иногда при измерении расстояний на местности прибегают к счету шагами Длина шага человека описывастся уравнением регрессии $L=37+h/4$, где h — рост человека в см, L — длина его шага.

Полная оценка взаимосвязи признаков требует нахождения уравнения регрессии не только для зависимости y от x, но и для зависимости x от y В силу вероятностного характера статистических взаимосвязей результаты вычислений по этим уравнениям не будут зеркально похожими. Поскольку методика и порядок вычислений в обоих случаях во многом аналогичны, ограничимся рассмотрением основ обработки уравнения $y=a+bx$ (зависимость y от x)

В уравнении $y=a+bx$ коэффициент b равен тангенсу угла наклона линии регрессии Этот коэффициент, называемый **«коэффициент регрессии»**, имеет большой статистический смысл Он показывает, насколько изменяется значение одной величины (зависимой, результативной переменной) при изменении второй (независимой, факторной) на единицу *Например* при увеличении температуры тела человека на 1° С, частота пульса увеличивается в среднем на 10 ударов в минуту

Статистический анализ подразумевает решение уравнения регрессии, т е. отыскание параметров этого уравнения на основе исходных данных. Математически решение уравнения линейной регрессии сводится к вычислению параметров a и b таким образом, чтобы точки исходных данных корреляционного поля как можно ближе лежали к прямой регрессии. Для этого вычисляют параметры по формулам, которые обеспечивают наимень-

ший квадрат отклонений этих точек от линии регрессии (метод наименьших квадратов):

$$a=\frac{n\sum xy-\sum x\sum y}{n\sum x^2-\left(\sum x\right)^2} \text{ и } b=\frac{n\sum x^2\sum y-\sum x\sum xy}{n\sum x^2-\left(\sum x\right)^2}.$$

Пример: Найти выборочное уравнение регрессии по данным пяти наблюдений (n=5) зависимой и независимой переменных (Y и X) (табл 75).

Таблица 75

Расчетная таблица параметров уравнения регрессии

i	X_i	Y_i	X_i^2	X_iY_i
1	2,0	2,6	4,0	5,2
2	4,0	1,8	16,0	7,2
3	7,0	1,3	49,0	9,1
4	5,0	1,1	25,0	5,5
5	3,0	2,4	9,0	7,2
$\sum$	21,0	9,2	103,0	34,2

Согласно уравнению параметр $a=\frac{5\times34{,}2-21\times9{,}2}{5\times103{,}0-21^2}=3{,}1$, параметр $b=\frac{103{,}0\times9{,}2-21\times34{,}2}{5\times103{,}0-21{,}2^2}=-0{,}3$. Искомое уравнение регрессии $y=3{,}1+(-0{,}3)x$. Коэффициент регрессии, соответственно, равен −0,3 Т. е при изменении независимой переменной (x) на 1, зависимая переменная (y) будет уменьшаться в среднем на 0,3

Насколько близки расчетные и фактические данные по зависимому фактору y, демонстрирует табл 76, где $Y_{рас}$ для первого наблюдения (i=1) Y_1=3,1+(−0,3)×2,0=2,5 и т. д Нетрудно заметить, что между фактическими и расчетными значениями ($Y_{рас}$ и Y_i) существует определенная разница Эта разница может объясняться малым числом наблюдений и точностью самого метода

Таблица 76

Разность фактического (Y_i) и вычисленного ($Y_{рас}$) параметров

i	X_i	Y_i	$Y_{рас}$	$Y_i - Y_{рас}$
1	2,0	2,6	2,5	0,1
2	4,0	1,8	1,9	−0,1
3	7,0	1,3	1,0	0,3
4	5,0	1,1	1,6	−0,5
5	3,0	2,4	2,2	0,2

Параметры уравнения регрессии, как и любые выборочные статистические характеристики, оцениваются в определенных интервалах В том случае, если уравнение регрессии имеет вид $y = a + bx$, выборочные значения коэффициентов a и b являются оценкой соответствующих генеральных коэффициентов и отличаются от них в среднем на величину соответствующих им ошибок Ошибка коэффициента $a \left[m_a = \frac{\sigma_x}{\sqrt{n}} \right]$, где σ_x — среднеквадратическое (стандартное) отклонение по ряду x, n — число наблюдений

Ошибка коэффициента b характеризует разброс значений угла наклона линии регрессии Полная ошибка для результатов отдельных измерений y

$$m_{Y(X)} = \sqrt{\left(\frac{1}{n(n-2)}\right)\left[n\sum y^2 - \left(\sum y\right)^2 - \frac{\left[n\sum xy - \left(\sum x\right)\left(\sum y\right)\right]^2}{n\sum x^2 - \left(\sum x\right)^2}\right]}$$

Рассмотренный пример касается так называемой двухмерной зависимости В этом случае рассматривается вариант, при котором взаимодействуют два признака — зависимый (результативный) и независимый (факторный) В реальной ситуации чаще приходится сталкиваться с многофакторными зависимостями Соответственно, если рассматривается большее число независимых признаков, то расчеты проводятся по другим формулам, с учетом трехмерного, четырехмерного и т п пространства распределения С математической точки зрения, число пространственных распределений, в принципе, не ограничено Обязатель-

ным условием такого подхода является независящее друг от друга распределение факторных признаков.

В общем виде формула для расчета коэффициента множественной регрессии для результативного показателя

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ + \beta_n X_n,$$

где $\beta_0, \beta_1, \beta_2 \ . \beta_n$ — коэффициенты регрессии *Например* должные (стандартные) величины показателей ЖЕЛ — жизненной емкости легких (Р Ф Клемент и др) вычисляются для мужчин в возрасте 18—25 лет по уравнению регрессии $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, где β_0 — константа, равная −6,908, β_1 — коэффициент по росту, равный 5,8, β_2 — коэффициент по возрасту 0,085 С помощью этого уравнения, опираясь на фактические данные о конкретном человеке, путем несложных вычислений можно определить должную (стандартную) величину ЖЕЛ этого человека Так, для мужчины в возрасте 19 лет, имеющему рост 1,8 метра, должная ЖЕЛ=−6,908+5,8×1,8+0,085×19=5,2 С точки зрения клинической практики, снижение фактической ЖЕЛ по сравнению с должной ЖЕЛ может говорить о рестриктивных нарушениях вентиляционной способности легких, являющихся следствием нарушения процесса расправления легких при вдохе.

Относительная простота применения уравнений регрессии обеспечила их большое распространение: для нахождения должных величин при оценке различных физиологических параметров, в гигиенических исследованиях для прогнозирования результатов воздействия различных факторов окружающей среды и т п Вместе с тем, получение точных исходных параметров уравнений регрессии требует большой и кропотливой работы

Одной из причин, снижающих точность параметров уравнения регрессии, является несоответствие теоретического распределения, взятого за основу расчетов, и фактического распределения точек корреляционного поля Например, линия регрессии может представлять собой не прямую, а какую-либо кривую Соответственно, форма уравнения регрессии должна соответствовать криволинейной зависимости (рис. 84)

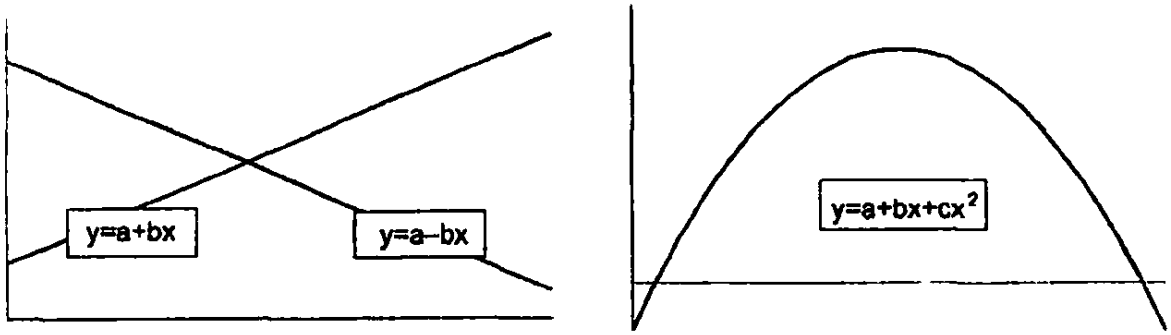


Рис. 84. **Различные линии регрессии**

Криволинейная зависимость может принимать различное математическое выражение в виде парабол 2-го и 3-го порядка *Например* для того чтобы найти параметры a, b и c в уравнении параболы второго порядка, нужно решить систему уравнений

$$\left\{\begin{array}{c} na+b\sum x+c\sum x^2=\sum y \\ a\sum x+b\sum x^2+c\sum x^3=\sum xy \\ a\sum x^2+b\sum x^3+c\sum x^2y \end{array}\right\}$$

и найти следующие промежуточные величины

$$\sum x,\ \sum x^2,\ \sum x^3,\ \sum x^4,\ \sum y,\ \sum xy,\ \sum x^2y.$$

В целом, вычисление и практическое использование этих параметров аналогичны операциям с параметрами прямой линии регрессии. Однако, в связи с громоздкостью расчетов, рекомендуется их находить с помощью специальных программ статистической обработки данных.

Метод группировок и построение статистических таблиц, а также регрессионный анализ позволяют установить наличие или отсутствие связи между факторными и результативными признаками, описать обнаруженные связи и определить некоторые количественные характеристики Различные **коэффициенты корреляции** позволяют выявить форму и силу (плотность, тесноту) этой связи

6.1.2. Коэффициент ковариации

В основу исчисления коэффициентов корреляции берется оценка совпадений колебаний значений признаков Если объективно существующие колебания (вариации) этих значений совпадают, то можно говорить о наличии корреляции (рис 85)

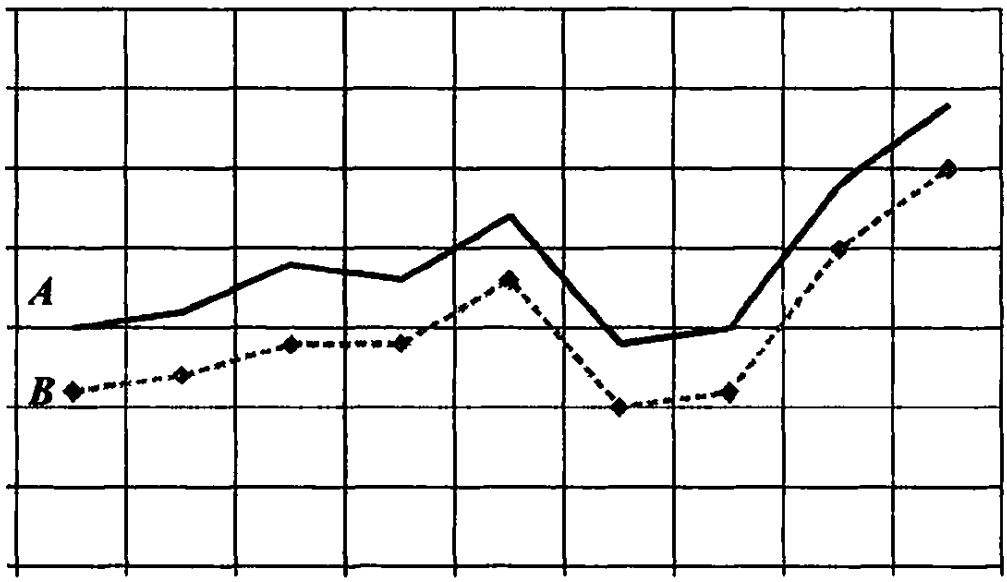


Рис. 85. **Колебания значений признаков совпадают**

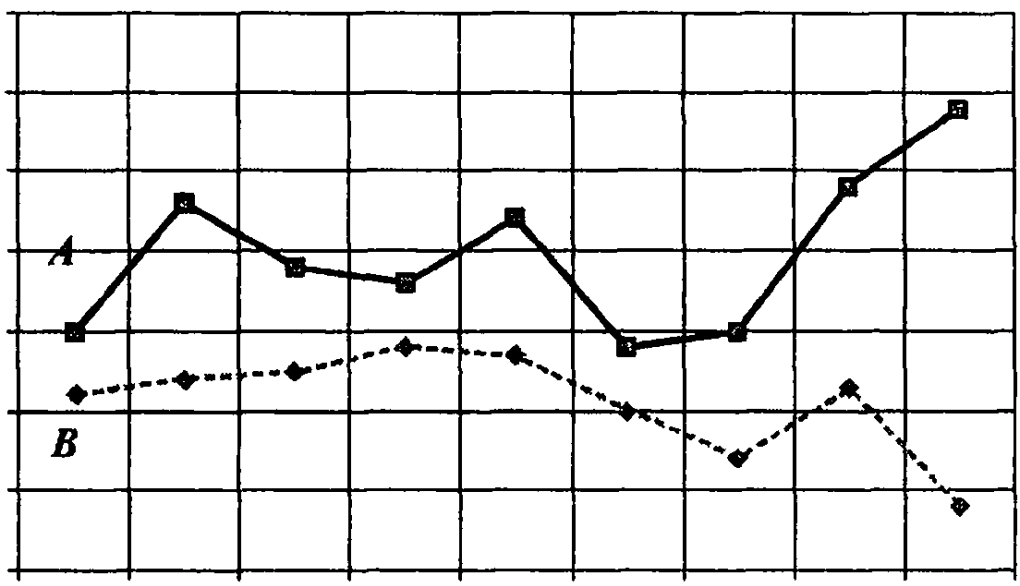


Рис 86 **Колебания значений признаков не совпадают**

Если колебания не совпадают — корреляции нет (рис 86) Рассмотрим пример оценки корреляционной связи длительности курения и частоты заболеваний органов дыхания за год (табл 77)

Таблица 77

Оценка взаимосвязи длительности курения и частоты заболеваний

Фамилии обследованных	Стаж курения (лет)		Число заболеваний		$(x-M_x)(y-M_y)$
	x	$x-M_x$	y	$y-M_y$	
Васильев	2	−1	5	1	−1
Сидоров	4	1	6	2	2
Петрова	5	2	4	0	0
Иванов	1	−2	1	−3	6
В среднем (M)	3	–	4	–	1,75

В столбцах таблицы расположены значения учетных признаков x (длительность курения) и y (число заболеваний) по каждому из группы обследованных. В соседних столбцах находятся отклонения этих признаков от своих групповых средних.

Для Васильева стаж курения $x=2$, отклонение от среднего значения M_x (2−3)=−1 Число заболеваний в году $y=5$, отклонение от среднего значения M_y (5−4)=1 В качестве общей меры обоих отклонений использовано произведений отклонений $(x-M_x)(y-M_y)=1\times(-1)=-1$ При прямой зависимости будет полное совпадение знаков всех произведений. При обратной — полное несовпадение. Для суммарной оценки всех отклонений находим среднее арифметическое всех произведений (1,75) Эта величина называется **коэффициентом ковариации** (*Co*), т. е. коэффициентом совместной вариации

Проведенные вычисления можно записать в виде формулы

$$Co=\frac{\sum P_{xy}(x-M_x)(y-M_y)}{n},$$

где P_{xy} — частоты пары признаков (в простом ряду $P_{xy}=1$), M_x — среднее ряда x, M_y — среднее ряда y, n — число наблюдений (число пар признаков)

6.1.3. Коэффициент линейной корреляции (Пирсона)

Использование *ковариации* в качестве меры связи весьма ограничено. Это связано с тем, что числовое значение коэффициента ковариации зависит от размерности и характера анализируемых признаков Поэтому в качестве меры корреляции используют не сами отклонения $(x-M_x)$ и $(y-M_y)$, а безразмерное отношение отклонений к их дисперсиям Отсюда следует, что значение коэффициента корреляции не зависит от единиц измерения. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в дюймах и футах или в сантиметрах и килограммах

После алгебраических преобразований формула для расчета **коэффициента корреляции** (Пирсона) выглядит следующим образом

$$r=\frac{\sum(x-M_x)(y-M_y)}{\sum(x-M_x)^2(y-M_y)^2} \text{ или } \hat{e}=\frac{\sum\alpha\beta}{\sqrt{\sum\alpha^2\sum\beta^2}},$$

где α — $(x-M_x)$, β — $(y-B_y)$. Для удобства «ручных» вычислений может применяться алгебраический аналог приведенных формул:

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{n}\right]\left[\sum y^2-\frac{(\sum y)^2}{n}\right]}}$$

Все возможные значения коэффициента r находятся в пределах от -1 до $+1$. Если связь между признаками прямая, то коэффициент корреляции положителен (положительное число) (табл. 78).

Таблица 78

Распределение значений коэффициента линейной корреляции Пирсона

Характеристики связи	Прямая	Обратная
Связи нет	0	0
Слабая	от 0 до 0,3	от 0 до −0,3
Средняя	от 0,3 до 0,7	от −0,3 до −0,7
Сильная	от 0,7 до 1	от −0,7 до −1
Полная (функциональная)	+1	−1

Например требуется определить зависимость числа ошибок (x), допускаемых операторами в корректурных пробах, от длительности работы на компьютере (y) (табл 79).

Таблица 79

Расчет коэффициента линейной корреляции Пирсона (n=10)

x	y	x^2	y^2	xy
6	1	36	1	6
8	1	64	1	8
7	2	49	4	14
6	3	36	9	18
6	3	36	9	18
5	4	25	16	20
8	5	64	25	40
6	5	36	25	30
9	6	81	36	54
9	6	81	36	54
$\sum x = 70$	$\sum y = 36$	$\sum x^2 = 508$	$\sum y^2 = 162$	$\sum xy = 262$

Согласно формуле, получаем

$$r = \frac{262 - \frac{70 \times 36}{10}}{\sqrt{\left(508 - \frac{70^2}{10}\right) \times \left(162 - \frac{36^2}{10}\right)}} = 0{,}41.$$

Как видно из полученного результата, между числом ошибок и длительностью работы есть прямая корреляционная связь средней силы В данном случае, по всей вероятности, можно говорить о причинно-следственной связи между этими факторами

Вместе с тем, утверждать, что показатели утомляемости связаны только с длительностью работы нельзя С помощью **коэффициента детерминации** (r^2) можно определить долю влияния анализируемого факторного признака на результативный признак Если принять во внимание, что длительность рабочего дня является не единственным фактором, способствующим развитию усталости, и если принять во внимание, что на развитие усталости влияют и другие факторы (день недели, характер выполняемой работы и т п), то r^2 показывает долю тех изменений, которые обусловлены анализируемым фактором В данном случае $r^2 = 0{,}41^2 = 0{,}17$ Таким образом, доля влияния продолжительности рабочего дня на развитие усталости у операторов 17%. Следовательно, на долю других факторов развития усталости приходится 83% влияния

Оценка значимости коэффициента корреляции r производится с помощью преобразования

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{(n-2)},$$

где n — число наблюдений В приведенном примере.

$$t = \frac{0{,}41}{\sqrt{1-0{,}41^2}}\sqrt{(10-2)} = 1{,}27 \approx 1{,}3$$

Для того чтобы результат в приведенном примере (см табл 79) можно было признать статистически достоверным с уровнем значимости $P = 0{,}05$, t должно быть не менее 1,9 для случая односторонней связи и не менее 2,3 для двусторонней связи (см. критические значения t в Приложении, табл 116)

Для оценки достоверности корреляционной связи можно воспользоваться и другим методом получения значений коэффициента: $t = \frac{|r|}{m_r}$, где $|r|$ — абсолютная величина коэффициента корреляции, а m_r — ошибка коэффициента корреляции

Ошибка вычисляется по формуле:

$$m_r = \frac{1-r^2}{\sqrt{n-2}}$$ В данном случае $m_r = \frac{1-0{,}41^2}{\sqrt{10-2}} = 0{,}29$

В приведенном примере (см табл 79) $t = 0{,}41/0{,}29 = 1{,}4$, т е результат почти не отличается от результата полученного предыдущим способом.

Приближенная оценка статистической достоверности коэффициента корреляции осуществляется исходя из того, что абсолютное значение коэффициента r должно превышать ошибку не менее чем в два раза

Определение доверительного интервала коэффициента r можно проводить в двух аспектах. В первом, более грубом приближении, можно удостовериться, что знак коэффициента определен правильно Во втором — определить непосредственные границы доверительного интервала.

Таблица 80

Число наблюдений, необходимое для подтверждения знака коэффициента корреляции (по М. Б Славину, 1989)

r	P		r	P	
	0,05	0,01		0,05	0,01
0,10	383	661	0,30	43	73
0,14	196	337	0,35	32	53
0,16	151	259	0,40	24	40
0,18	119	204	0,45	19	31
0,20	97	165	0,50	16	25
0,22	80	136	0,60	11	17
0,24	68	114	0,70	8	12
0,26	57	97	0,80	6	9
0,28	49	83	0,90	5	6

Рассматривая приведенный в примере коэффициент корреляции ($r = 0{,}41$), нетрудно заметить, что необходимое минимальное число наблюдений для того, чтобы быть уверенным в правильности знака коэффициента, составляет около 24 человек.

Для решения вопроса о доверительном интервале можно использовать величину Z. Эта величина связана с коэффициентом r соотношением $Z = \frac{1}{2}\ln\frac{1+r}{1-r}$ В приведенном примере

$$Z = 0{,}5\ln\frac{1{,}41}{0{,}59} = 0{,}43$$

Задаваясь значением P, вычисляем верхнюю и нижнюю границы величины Z:

$$Z_{max} = Z + \alpha\sigma,\ Z_{min} = Z - \beta\sigma$$

Значение σ находится по формуле $\sigma = \frac{1}{\sqrt{n-3}}$. В данном случае

$\sigma = \frac{1}{\sqrt{10-3}} = 0{,}38.$

Соотношение доверительной вероятности (P) и уровней значимости (α и β) представлены в табл. 81.

Таблица 81

Соотношение коэффициентов P, α и β

P	α	β
0,05	0,95	1,96
0,01	0,99	2,58

Принимая уровень значимости $P = 0{,}05$, находим $Z_{max} = 0{,}43 + 0{,}95 \times 0{,}38 = 0{,}79$ и $Z_{min} = 0{,}43 - 1{,}96 \times 0{,}38 = 0{,}31$.

Границы коэффициента корреляции r, соответствующие полученным границам Z, можно установить используя данные, представленные в приложении (табл. 114)

В таблице представлены значения Z. Значения соответствующих коэффициентов корреляции (r) складываются из данных столбца r (первый столбец), в котором находятся значения первого знака после запятой для r. В первой строке таблицы — значения для второго знака коэффициента r

Например. полученное значение $Z_{max} = 0{,}79$ соответствует коэффициенту $r = 0{,}65$, а значение $Z_{min} = 0{,}31$ соответствует $r = 0{,}30$.

Таким образом, при уровне значимости $P = 0{,}05$ можно считать, что величина коэффициента корреляции находится в диапазоне $0{,}30 \le r \le 0{,}65$.

Следует отметить, что надежность коэффициента корреляции увеличивается с увеличением его абсолютного значения

Общность и взаимосвязь, с точки зрения формальной логики, между различными статистическими показателями связи позво-

ляет в процессе анализа данных переходить от одного коэффициента к другому или по числовым значениям одних вычислять другие. *Например*, зная коэффициент корреляции, можно вычислить соответствующие коэффициенты регрессии:

$$R_{x/y} = r\frac{\sigma_x}{\sigma_y}, \quad R_{y/x} = r\frac{\sigma_y}{\sigma_x},$$

где $R_{x/y}$ — коэффициент регрессии y от x, $R_{y/x}$ — коэффициент регрессии x от y, r — коэффициент корреляции, σ — среднеквадратическое отклонение, соответственно по ряду x или y. Зная коэффициенты регрессии, путем простейших алгебраических преобразований по этим формулам можно вычислить коэффициент корреляции

6.1.4. Корреляционное отношение. Криволинейная корреляция

Если значение коэффициента корреляции оказалось не столь высоким, как ожидалось, это не означает, что связь действительно слабая. Возможно, что между факторным и результативным признаками связь носит криволинейный характер, которая не улавливается коэффициентом линейной корреляции Пирсона Одним из способов оценки криволинейной связи является применение **корреляционного отношения** (η). Дополнительным отличием коэффициента η от коэффициента корреляции является возможность раздельной оценки влияния фактора x на y и y на x, т. е возможность выявить неравнозначность воздействия фактора x на y и y на x Таким образом, обычно вычисляются два варианта: $\eta_{x/y}$ и $\eta_{y/x}$ В общем виде формула корреляционного отношения выглядит следующим образом

$$\eta_{y/x} = r\frac{\sigma_{\bar{y}_x}}{\sigma_y} \text{ или } \eta_{x/y} = r\frac{\sigma_{x_y}}{\sigma_x},$$

где σ_{y_x} — среднеквадратическое отклонение, представляющее изменчивость y под влиянием только x, σ_{x_y} — изменчивость x под влиянием только y, σ_x и σ_y — среднеквадратические отклонения, характеризующие изменчивость признаков x и y в целом В практических вычислениях удобнее пользоваться следующими формулами

$$\eta_{x/y} = \sqrt{\frac{\sum(x - M_x)^2 - \sum(x - M_{xk})^2}{\sum(x - M_x)^2}}$$

и

$$\eta_{y/x} = \sqrt{\frac{\sum(y - M_y)^2 - \sum(y - M_{yk})^2}{\sum(y - M_y)^2}},$$

где M_{xk} и M_{yk} — групповые средние арифметические

Например при использовании коэффициента линейной корреляции Пирсона для оценки зависимости развития усталости (корректурный тест) от продолжительности рабочего времени получено невысокое значение коэффициента корреляции (0,41) Одной из причин получения такой оценки могла быть криволинейная зависимость этих факторов работоспособность сначала повышается (происходит *врабатывание)*, затем некоторое время держится на высоком уровне (плато), затем начинает прогрессивно снижаться Для проверки этой рабочей гипотезы было проведено вычисление корреляционного отношения времени работы (y) и развития усталости (x). Корреляционное отношение y на x не вычислялось из-за невозможности такой зависимости в объективной реальности Признак «Число ошибок» группировался исходя из одинаковых значении признака «Время работы» (табл 82)

Таблица 82

Расчет корреляционного отношения

Время работы y	Число ошибок x	M_{xk}	$X-M_{xk}$	$(x-M_{xk})^2$	$x-M_x$	$(x-M_x)^2$
1	6		−1	1	−1	1
1	8	7	1	1	1	1
2	7	7	0	0	0	0
3	6		0	0	−1	1
3	6	6	0	0	−1	1
4	5	5	0	0	−2	4
5	6		−1	1	−1	1
5	8	7	1	1	1	1
6	9		0	0	2	4
6	9	9	0	0	2	4
	M_x=7,0			$\sum(x-M_{xk})^2=4$		$\sum(x-M_x)^2=18$

$$\eta_{x/y}=\sqrt{\frac{18-4}{4}}=0{,}88$$

Нетрудно заметить, что полученное значение корреляционного отношения почти в два раза больше коэффициента линейной корреляции Пирсона, вычисленного на этих же данных. Расчет коэффициента детерминации на основе полученного корреляционного отношения (r^2=0,88^2=0,78) свидетельствует о значительной роли фактора рабочего времени (78%) среди всех факторов развития усталости Рассчитанный на основе коэффициента линейной корреляции r, показатель r^2 был в три раза меньше (17%)

Это соотношение является подтверждением того, что существует криволинейная связь В случае равенства корреляционного отношения и коэффициента корреляции r можно говорить о линейном характере связи.

Коэффициент корреляции имеет одинаковое значение, рассматривается ли влияние x на y или, наоборот, y на x Корреляционные же отношения $\eta_{x/y}$ и $\eta_{y/x}$ не равны, так как выявляют неравнозначность воздействия x на y и y на x

6.1.5. Частная (парциальная) корреляция

При оценке взаимосвязи факторных и результативных признаков наиболее существенным является анализ взаимодействия не двух, а, гораздо чаще встречающегося в объективной реальности, взаимодействия нескольких факторов. Существует несколько способов решения этой проблемы В лабораторных условиях, в условиях «чистого» опыта, она решается путем последовательного анализа выделяемых «в чистом виде» отдельных факторов или их специально подобранных комбинаций При исследованиях, проводимых в реальной ситуации (группы больных, коллективы рабочих, контингенты населения), этот способ малопригоден Главная причина — практическая невозможность подобрать в натурных условиях лиц с одинаковыми исходными характеристиками (возраст, пол, стаж и т. п) и подвергающихся воздействию «в чистом виде» только одного или максимум двух факторов. Единственный выход — это комбинационная группировка собранного материала, которая, однако, требует очень большого числа наблюдений, или использование специальных статистических методов С помощью этих методов производится последовательная элиминация влияния одних факторов и выделение результатов влияния других факторов К таким методам относится метод частной корреляции

В простейшем случае, в ходе вычисления коэффициентов частной корреляции для трех признаков последовательно элиминируется влияние одного из признаков Сначала X_3, затем X_2, и, наконец, X_1. Таким образом, последовательно выявляется взаимосвязь «в чистом виде» сначала X_1 и X_2 , затем X_1 и X_3 и, наконец, X_3 и X_2. Реализуются эти расчеты следующим образом·

1. Элиминирование влияния третьего признака и выявление связи между первым и вторым производится по формуле

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}.$$

2 Аналогичным образом производится элиминирование влияния второго признака и выявление связи между первым и третьим:

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}}$$

3 Затем проводится определение взаимосвязи третьего и второго признака по формуле

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{12}^2\right)}}$$

Например требуется оценить взаимосвязь фактора длительности рабочего времени, усталости (число ошибок при корректурном тесте) и производительности труда (количество страниц текста, набираемого операторами) на персональном компьютере

Таблица 83

Исходные данные для расчета частной корреляции

Время работы	Число ошибок	Число страниц текста
1	*2*	*3*
4	5	4
1	6	6
3	6	2
3	6	6
5	6	4
2	7	3
1	8	1
5	8	3
6	9	1
6	9	1

С помощью коэффициентов парной корреляции Пирсона получены результаты, на основании которых можно сделать вывод о влиянии на появление усталости длительности рабочего времени ($r_{12} = 0{,}4$ — связь прямая, средней силы) и снижении производительности труда по мере увеличения времени работы

(r_{13}=−0,7 — связь обратная, сильная) Между увеличением усталости и снижением производительности труда также обнаружена статистическая связь (r_{23}=−0,4 — обратная, средней силы).

На основе рассчитанных коэффициентов парных корреляций можно вычислить соответствующие коэффициенты частной корреляции:

$$r_{12,3}=\frac{0,4-(-0,7)\times(-0,4)}{\sqrt{\left(1-(-0,7)^2\right)\times\left(1-(-0,4)^2\right)}}=0,2;$$

$$r_{23,1}=\frac{-0,4-0,4\times(-0,7)}{\sqrt{\left(1-(-0,7)^2\right)\times\left(1-0,4^2\right)}}=-0,1;$$

$$r_{13,2}=\frac{-0,7-0,4\times(-0,4)}{\sqrt{\left(1-(-0,4)^2\right)\times\left(1-0,4^2\right)}}=-0,7$$

Таблица 84

Итоговые значения коэффициентов корреляции

r_{12}	0,4	$r_{12,3}$	0,2
r_{13}	−0,7	$r_{23,1}$	−0,1
r_{23}	−0,4	$r_{13,2}$	−0,7

Анализ этих коэффициентов частной корреляции показывает, что при устранении влияния фактора продолжительности рабочего времени, произошел существенный сдвиг показателя $r_{23,1}$, т е связь между усталостью и производительностью труда обследованных работников почти полностью исчезла ($r_{23,1}$=−0,1) Таким образом, снижение выработки продукции (количество набранного текста) к концу рабочего дня, вероятно, связано в первую очередь не с нарастанием усталости, а какими-то другими (возможно, организационными) причинами

6.1.6. Понятие о множественной корреляции

Метод множественной корреляции обычно применяется для характеристики совместного, совокупного влияния всего комплекса факторов на результативный признак. Изменяется величина коэффициента множественной корреляции по абсолютному значению в таких же пределах (от 0 до 1), как и коэффициент парной или частной корреляции. Вычислить **коэффициент множественной корреляции** можно при помощи коэффициентов частной линейной корреляции по формуле (для трех признаков см. табл. 83).

$$R_{1.23}=\sqrt{1-\left(1-r_{12}^2\right)\left(1-r_{13.2}^2\right)}=\sqrt{1-\left(1-0{,}4^2\right)\left(1-(0{,}7)^2\right)}=0{,}75$$

На основании коэффициента множественной корреляции определяется *коэффициент детерминации* ($0{,}75^2=0{,}6$), который показывает долю (60%) совместного влияния 2-го и 3-го признака.

6.1.7. Вычисление коэффициентов корреляции и уравнений регрессии в *MS Excel*

В *MS Excel* имеется несколько встроенных функций предназначенных для корреляционного и регрессионного анализа. Наиболее простейшая из них — **КОРРЕЛ**. Эта функция вычисляет коэффициент парной линейной корреляции. *Пример:* Введите исходные данные в таблицу (рис. 87).

С помощью пиктограммы или из меню <Вставка> выполните команду «функция». В появившемся окне мастера функций выберите **КОРРЕЛ** (рис 88).

Нажмите **[ОК]** и в появившемся следующем окне укажите координаты ячеек, в которых расположены числовые массивы исходных данных (рис 89). После завершения ввода координат появится значение коэффициента корреляции.

	A	B	C
1	X	Y	Z
2	14,7	1,4	20
3	13,4	1,4	25
4	9,6	2,3	30
5	8,1	2,1	32
6	5,5	4,2	27
7	5,2	4,9	26
8	4,4	6,6	22
9	4	6,8	28
10	3,5	7	31

***Рис 87.* Расположение исходных данных для вычислений корреляции и регрессии**

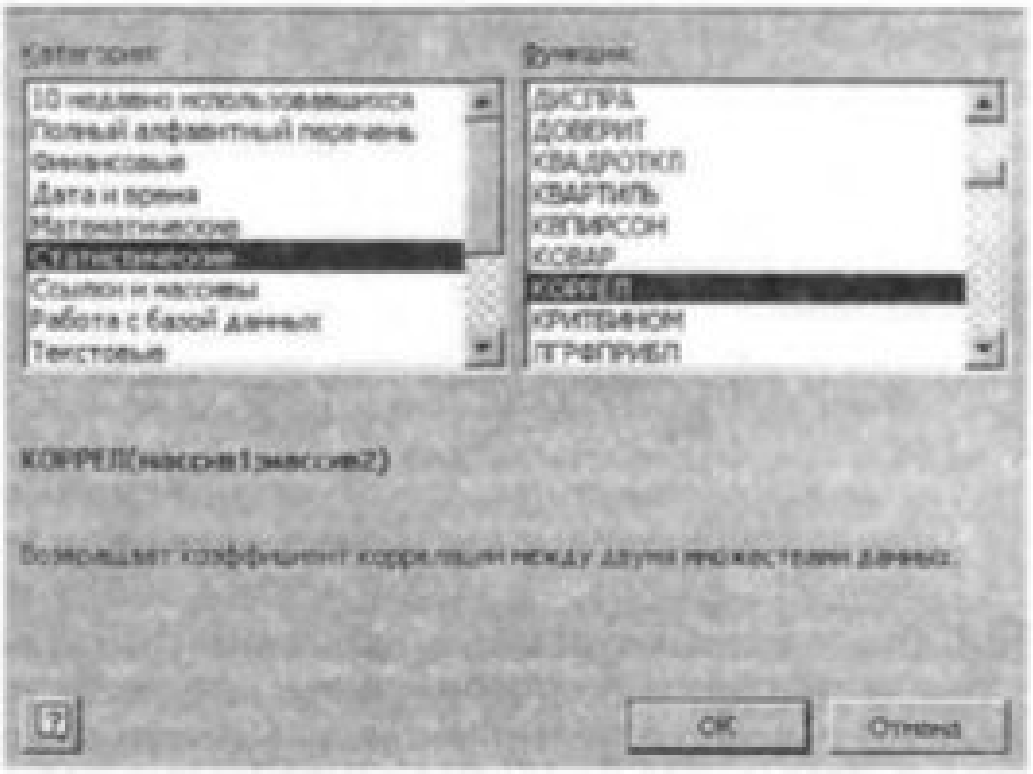


Рис. 88 **Окно мастера функций**

Рис. 89. **Заполнение полей окна функции КОРРЕЛ**

В пакете «Анализ данных» также имеется функция, позволяющая вычислять коэффициент корреляции (рис 90)

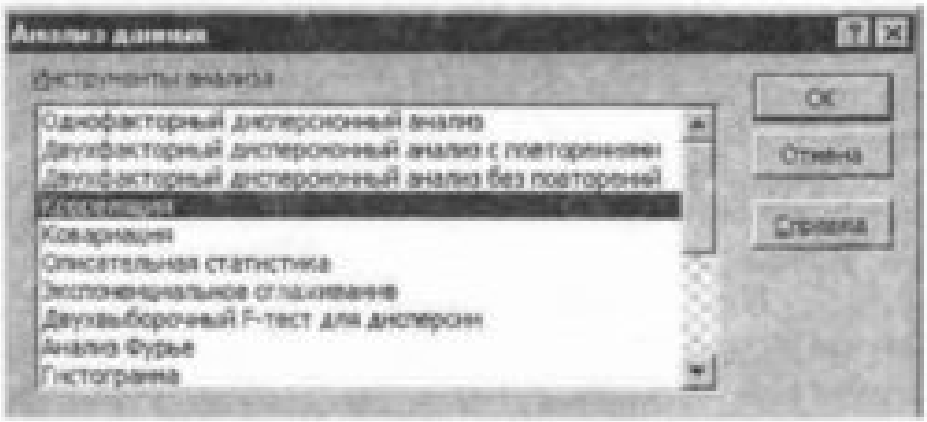


Рис. 90. **Пакет «Анализ данных»**

С помощью этой функции можно вычислять парные коэффициенты корреляции последовательно из нескольких рядов числовых данных Для *примера* используем имеющиеся исходные данные. После ввода интервала ячеек (окно «**Входной интервал**»), в которых расположены исходные данные, указывается способ группировки исходных данных (по строкам или столбцам таблицы), наличие меток (заголовки столбцов) и место вывода готовых результатов (Параметры вывода). Нажатие на клавишу **[ОК]** приведет к выводу матрицы результатов парного анализа содержимого столбцов в различных сочетаниях (рис. 91)

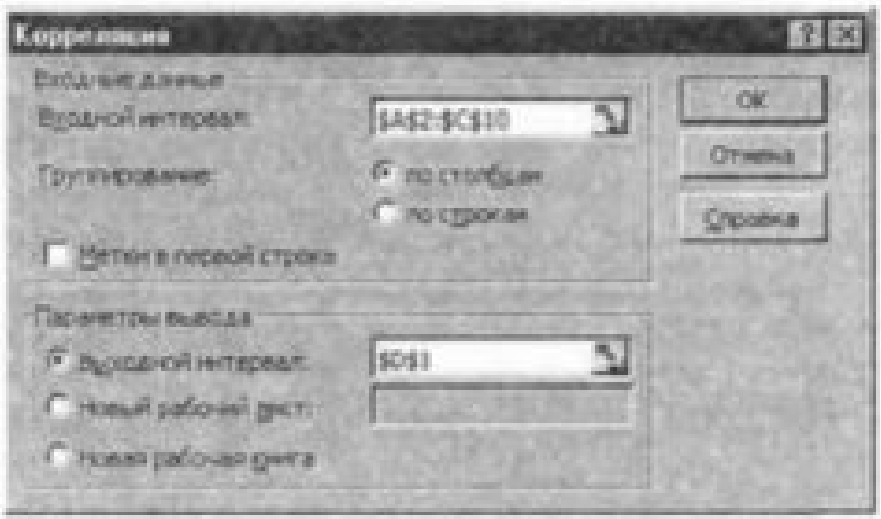


Рис. 91. **Заполнение окна «Корреляция»**

D	E	F	G
	Столбец 1	Столбец 2	Столбец 3
Столбец 1	1		
Столбец 2	-0,90748	1	
Столбец 3	-0,40614	0,129923	1

Рис. 92. **Матрица результатов (коэффициентов корреляции)**

На полученной матрице результатов видно, что корреляция числовых рядов, расположенных в столбцах 1 и 2 $r_{12} = -0{,}907$ (связь сильная, обратная), столбцов 1 и 3 $r_{13} = -0{,}406$ (связь средняя, обратная), столбцов 2 и 3 $r_{23} = 0{,}13$ (слабая, прямая) (рис. 92)

Примечание. К сожалению, в *MS Excel* отсутствуют функции прямого определения статистических ошибок и оценки достоверности различий коэффициентов корреляций При желании, эти параметры могут быть относительно просто определены вручную. (См. разделы «Коэффициент линейной корреляции» и «Оценка различий коэффициентов корреляции»).

Одним из самых удобных и практичных способов корреляционного и регрессионного анализа в *Excel* — использование пакета построения графических изображений («Вставка», «Диаграмма»). Для *примера* используем имеющиеся исходные данные (см. рис. 87)

В окне мастера диаграмм необходимо выбрать точечный тип диаграммы и ее вид без соединений точек (рис. 93) Затем нажмите кнопку **[Далее]**.

В следующем окне мастера диаграмм установите диапазон ячеек с исходными данными (см стр 87, 88). После этого необходимо проверить в окне «ряд» какие данные попадают в ряд *X*, а какие в ряд *Y*. Здесь необходимо быть внимательным, так как **регрессия X от Y и Y от X — это не одно и то же!** После проверки нажмите кнопку **[Готово]**.

С помощью редактора диаграмм откорректируйте внешний вид диаграммы (см. раздел «Графические изображения»)

На появившейся диаграмме выделите щелчком ***левой клавиши*** мыши ряд полученных точек, а затем щелчком ***правой клавиши*** мыши вызовите список команд и выберите из них: «**Добавить линии тренда**» (рис. 94).

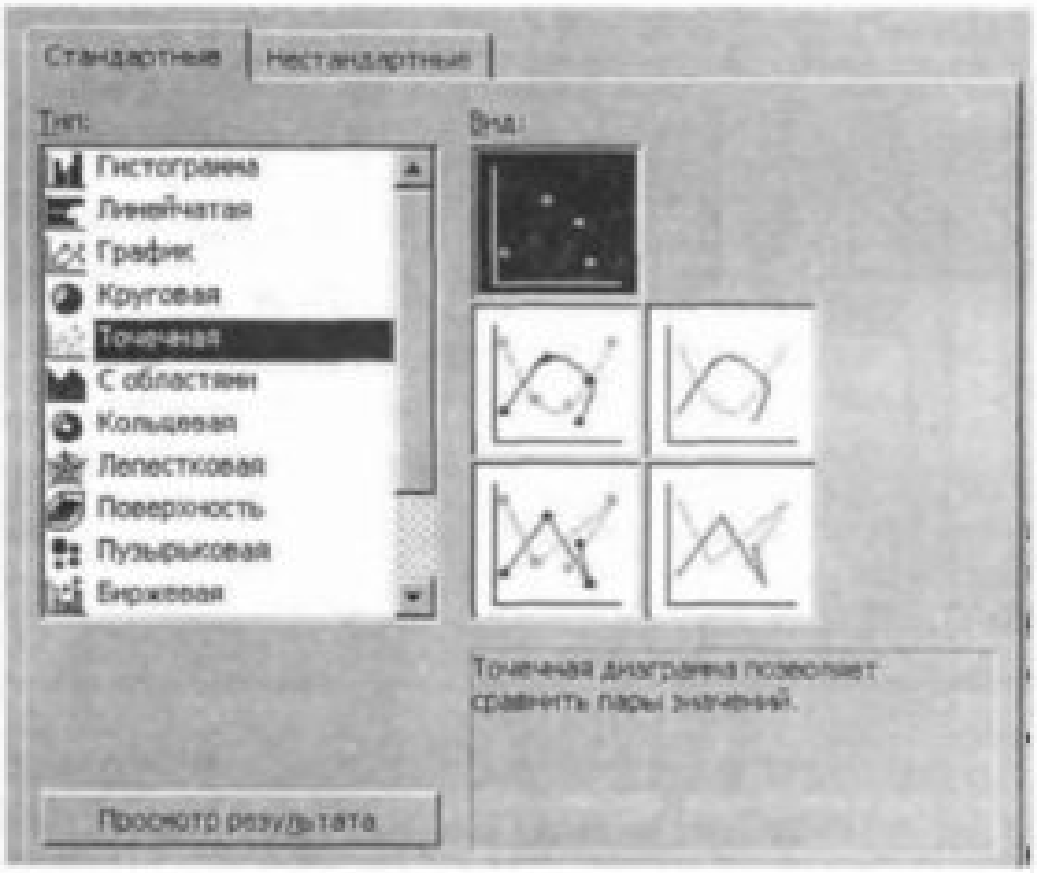


Рис. 93. **Выбор параметров диаграммы для регрессионного анализа**

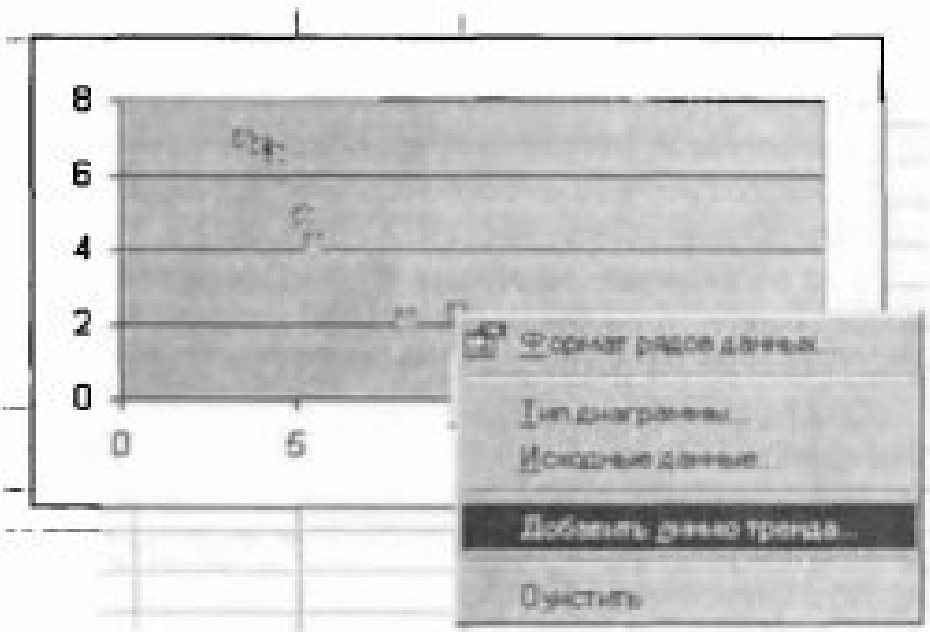


Рис. 94. **Вызов команды установки линии тренда**

В окне «Линии тренда» установите «Тип» «Линейная» (рис. 95)

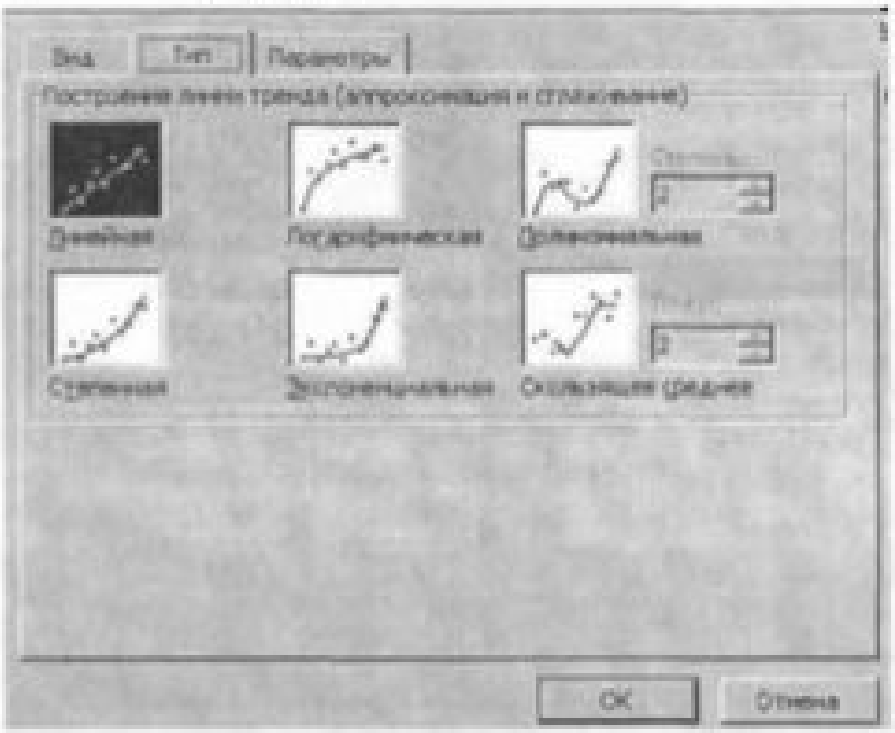


***Рис 95.* Установка типа линий тренда**

Затем выделите в окне «**Параметры**» поля «**показывать уравнение регрессии**» и «**поместить на диаграмму величину достоверности аппроксимации**»

На диаграмме появится окно с автоматически вычисленным уравнением линейной регрессии, а также значение R^2, которое показывает, насколько точно соответствует вычисленное уравнение регрессии истинной взаимосвязи параметров Y и X (рис 96). Максимально возможное значение $R^2 = 1,0$ или 100% В данном примере, в случае использования аппроксимации (сглаживания) с помощью линейной регрессии, $R^2 = 0,8235$ или 82,35% (см. в разделе «Коэффициент линейной корреляции» «коэффициент детерминации»)

С помощью этого коэффициента можно путем подбора выбрать уравнение регрессии наиболее полно аппроксимирующее ту или иную взаимосвязь Например, при анализе этих же исходных данных полином 3-й степени дает более полную аппроксимацию, поскольку $R^2 = 0,9764$ (рис. 96).

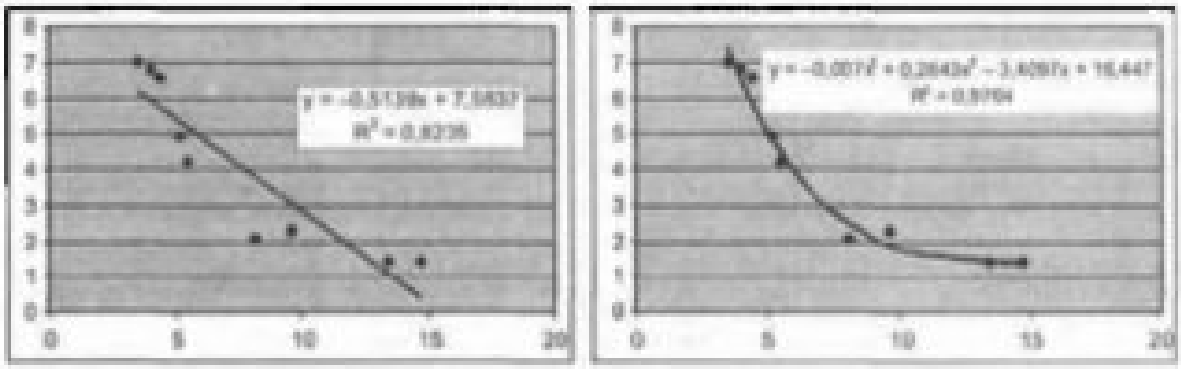


Рис. 96. **Различные варианты регрессионного сглаживания**

Кроме того, поскольку R^2 — есть квадрат коэффициента корреляции, то отсюда легко найти коэффициент корреляции Для этого достаточно извлечь из R^2 квадратный корень В данном примере, в случае линейной корреляции $\sqrt{0,8235}=0,907$ В случае криволинейной корреляции $\sqrt{0,9764}=0,989$ Вид коэффициента корреляции (линейный или не линейный) устанавливается на основе использованного при аппроксимации уравнения регрессии Знаки коэффициентов корреляции устанавливаются соответственно положению на графике линии регрессии. В данном примере угол наклона свидетельствует о том, что связь обратная Таким образом, коэффициенты корреляции отрицательны (−0,907 и −0,989).

С помощью рассмотренной функции мастера диаграмм можно производить прогнозирование значений Y Для этого при подготовке к аппроксимации данных нужно в окне **«параметры» «линия тренда»** установить величину *X* (единиц интервала ряда *X* вперед или назад).

Ошибку уравнения регрессии, точнее вычисляемой с помощью уравнения регрессии величины *Y*, можно определить с помощью функции СТОШХУ (рис 97)

Более глубоко регрессионный анализ позволяет проводить надстройка *Excel* **«Пакет анализа»**. Главным качеством этого пакета является всесторонняя оценка достоверности полученных результатов

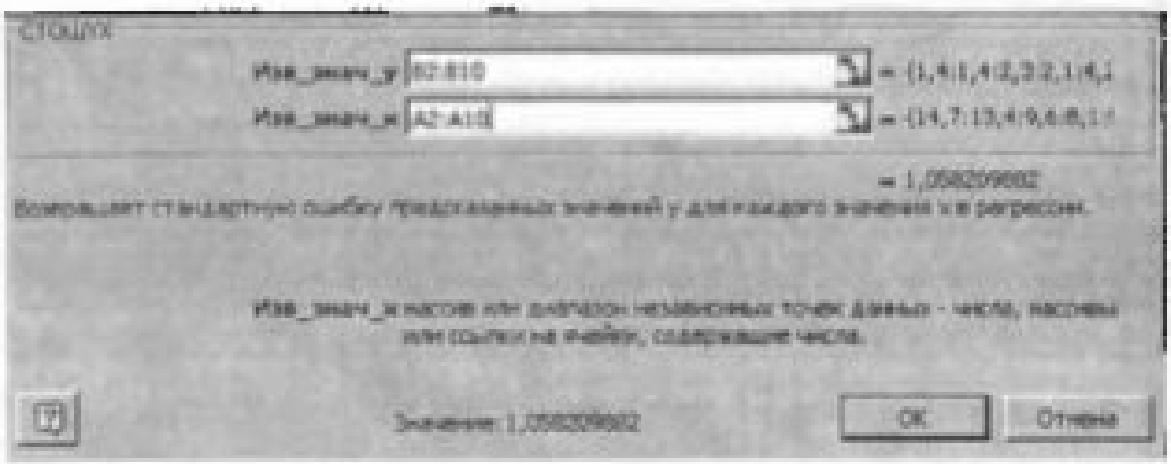


Рис. 97. **Вычисление ошибки уравнения регрессии**

Для выполнения расчетов с помощью надстройки **«Пакет анализа»**.

1 Выполните команду <**Пакет анализа**> из меню <**Сервис**> Должно появиться диалоговое окно **«Анализ данных»**.

2 Выберите метод **«Регрессия»** и нажмите кнопку **[ОК]**.

3 В появившемся окне **«Регрессия»** введите входной интервал Y, т е. ссылку на диапазон анализируемых зависимых данных: $B$2:$B$10.

Примечание: Этот диапазон должен состоять из одного столбца.

Затем введите входной интервал X, т е ссылку на диапазон независимых данных, подлежащих анализу $A$2:$A$10.

Примечание: Независимые переменные должны состоять из одного или нескольких столбцов, расположенных слева направо. Максимальное число независимых переменных равно 16

Не устанавливайте флажок **«Метки»**, т. к первая строка входного интервала не содержит заголовки.

Не устанавливайте флажок **«Уровень надежности»**, т. к уже заданная величина 95% нас устраивает. Этот флажок устанавливается, чтобы включить в выходной диапазон уровень надежности, который будет использован дополнительно к уровню 95%, применяемому по умолчанию.

Не устанавливайте флажок **«Константа-ноль»**, т. к линия регрессии, которую мы анализируем, не проходит через начало координат.

Установите переключатель **«Выходной интервал»**, чтобы вставить результаты анализа, начиная с ячейки A16. Остальные ячейки для вывода данных будут установлены автоматически.

Примечание: Для получения результатов на другом листе установите переключатель **«Новый рабочий лист»**

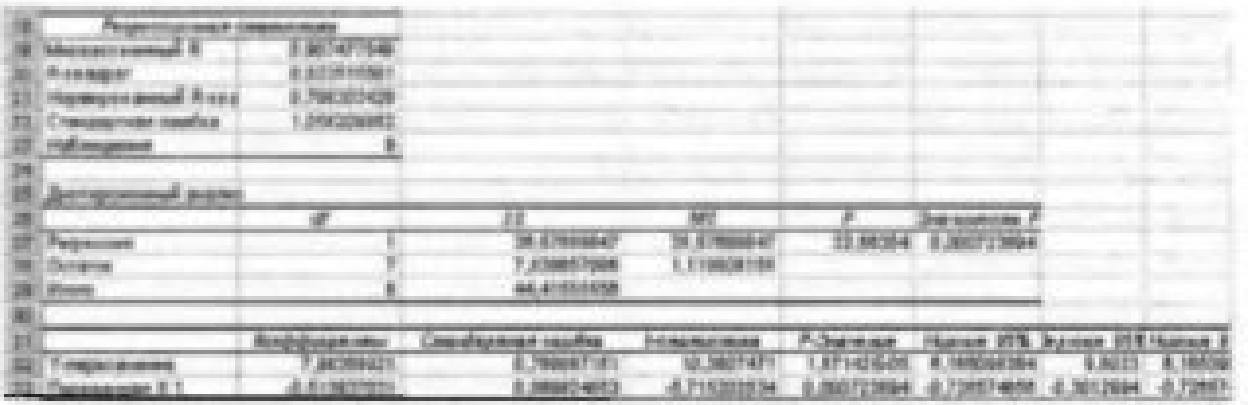


Рис. 98. **Выходные данные пакета регрессионного анализа**

Кроме того, в окне **«Регрессия»** можно установить флажки

«График подбора», чтобы построить диаграммы наблюдаемых и предсказанных значений для каждой независимой переменной.

«Остатки», чтобы включить остатки в выходной диапазон.

«Стандартизированные остатки», чтобы включить стандартизированные остатки в выходной диапазон

«График остатков», чтобы построить диаграмму остатков для каждой независимой переменной.

«График нормальной вероятности», чтобы построить диаграмму нормальной вероятности.

4. После завершения настройки параметров нажмите кнопку **[ОК]**

Результаты регрессионного анализа будут состоять из трех таблиц (рис. 98) и одного графика.

Первая таблица содержит значение корреляции и коэффициента детерминации (квадрата выборочного коэффициента кор-

реляции Пирсона), являющегося показателем качества подобранной регрессионной модели. В нашем примере R^2=0,82, т. е около 82% вариации *Y* объясняется полученным уравнением регрессии

Стандартная ошибка (ячейка А22) показывает стандартную ошибку вычисляемых значений *Y*.

Вторая таблица содержит результаты дисперсионного анализа, с помощью которого проверяется нулевая (H_0) статистическая гипотеза о равенстве нулю всех вычисленных коэффициентов регрессии Эта таблица включает следующие параметры.

- **df** — число степеней свободы,
- **SS** — сумма квадратов,
- **MS** — средний квадрат (дисперсия),
- **F** — F-статистика Фишера,
- **Значимость F** — значимость критерия Фишера

В данном случае, основной результат дисперсионного анализа состоит в том, что уравнение линейной регрессии является значимым, т. к полученная значимость *P* =0,00074 меньше 0,05. Таким образом, отвергается (H_0) гипотеза о равенстве нулю всех коэффициентов уравнения регрессии

В третьей таблице представлены результаты вычисления коэффициентов уравнения регрессии *Y'*=a+bX коэффициент a=7,984 и коэффициент b=−0,514. Таким образом, оценка уравнения регрессии имеет вид *Y'*=7,984−0,514X

Остальные результаты позволяют проверить значимость полученных коэффициентов a и b уравнения регрессии, т е. проверить нулевые гипотезы: H_a=0, H_b=0, используя t-критерий

Коэффициенты a и b значимы, т. к абсолютные значения критерия t $|t_a|$=10,38 и $|t_b|$=5,715 больше $t_{кр}$=2,36, найденного при значимости P=0,05 и числе степеней свободы V=9−1−1=7

То, что нулевые гипотезы отвергаются, подтверждается и величинами значимости, равными 0,0000167 для коэффициента a и 0,0007239 для коэффициента b

Следовательно, доказано наличие сильной зависимости между переменными *X* и *Y*.

6.1.8. Оценки взаимосвязи качественных признаков с помощью коэффициента ранговой корреляции Спирмена

В некоторых исследованиях приходится сталкиваться с ситуацией, когда учетные признаки характеризуются не точными числовыми значениями, а приближенными оценками большего или меньшего количества какого-либо свойства или явления Если отдельные единицы наблюдения расположить в порядке возрастания или убывания этих оценок, иначе *ранжировать*, то порядковый номер каждой единицы наблюдения будет номером ее ранга. Если анализируемые признаки взаимосвязаны, изменения их числовых значений в большую или меньшую сторону будут совпадать, соответственно разность их рангов будет минимальна, и наоборот.

Одно из главных достоинств коэффициента корреляции рангов заключается в простоте вычислений Применение этого коэффициента корреляции может быть рекомендовано в случаях

- когда необходимо быстро ориентировочно определить связь между какими-то признаками;
- если необходимо оценить связь между качественными и количественными признаками или только между качественными признаками,
- когда распределение значений учетных признаков (в том числе и количественных) не соответствует нормальному распределению или распределение неизвестно

Существует несколько вариантов вычисления коэффициентов ранговой корреляции. *Коэффициент корреляции рангов Спирмена* вычисляется по формуле: $\rho = 1 - \frac{6\sum d^2}{n^3 - n}$, где d — разности между рангами (порядковыми номерами), n — число сопоставляемых пар *Например* требуется оценить зависимость интенсивности запаха воды (в баллах) открытого водоема (Y) и удаленности места забора проб (в км) от возможного источника загрязнения (X)

Таблица 85

Вычисление коэффициента корреляции рангов Спирмена

X (км)	*Y (баллы)*	*Ранги X*	*Ранги Y*	*d*	d^2
1	*2*	*3*	*4*	*5*	*6*
7	1	1	6,5	−5,5	30,25
6	1	2	6,5	−4,5	20,25
5	2	3	4	−1	1
4	3	4	3	1	1
3	2	5	4	1	1
1	3	6,5	3	3,5	12,25
1	4	6,5	1	5,5	30,25
				$\sum d = 0$	$\sum d^2 = 96$

Вычисления выполняются в следующей последовательности:

1. Располагаем значения *X*, ранжируя их в первом столбце таблицы по мере убывания (табл. 85).

2 В соседнем столбце располагаем соответствующие признаку *X* значения *Y*. Специального ранжирования в этом столбце уже не проводится. Все значения во втором столбце располагаются только соответственно парным значениям *X*.

В третьем столбце выставляются ранги значений *X*. Самому большому значению *X*=7 присваивается ранг 1 Следующему значению *X*=6 присваивается ранг 2, и т д В конце ряда, на 6-м и 7-м месте имеется два одинаковых значения *X* Этим значениям *X* присваивается средний ранг (6+7)/2=6,5

В четвертом столбце таблицы размещаются ранги значений *Y*. Номера рангов выставляются начиная от самых больших значений *Y* (так же, как и по ряду *X*) Самое большое значение *Y*=4 (внизу столбца) получает ранг 1. Затем идут значения *Y*=3 Они занимают 2-е и 4-е место (считая от самого большого значения *Y*=4). Соответственно, им присваивается усредненное значение ранга (2+4)/2=3, и т. д.

В пятом столбце записывается разность рангов ряда *X* и *Y*. В шестом — квадрат этой разности. Полученные значения квадратов суммируются и подставляются в формулу:

$$\rho = 1 - \frac{6 \times 96}{7^3 - 7} = -0{,}71$$

Таким образом, связь между интенсивностью запаха воды и удаленностью от возможного источника загрязнения можно считать доказанной

6.2. Оценки взаимосвязи качественных признаков на принципе взаимной сопряженности

Задачей статистического анализа взаимосвязи между качественными признаками является количественная характеристика этой связи, которая называется *взаимной сопряженностью* Оценка сопряженности данных, сгруппированных по качественному признаку, базируется на статистической обработке частот распределения Табличным представлением взаимного распределения группировок двух качественных признаков является **таблица взаимной сопряженности**.

Визуально с помощью таблицы сопряженности или на диаграмме можно видеть взаимосвязь в распределении частот, однако количественную характеристику этой связи могут дать только различные коэффициенты сопряженности Выбор того или иного коэффициента диктуется характером исходных данных и необходимой точностью анализа

Достоинствами коэффициентов сопряженности являются

- относительная простота их вычисления,
- возможность с их помощью обрабатывать не только качественные, по своей природе, признаки, но и количественные признаки, которым после группировки можно придать характер качественных

6.2.1. Коэффициенты *Q* и *Ф*

С помощью простейшей четырехпольной таблицы сопряженности, состоящей всего из двух строк и двух столбцов

(2×2), можно установить взаимосвязь двух любых признаков и дать некоторую количественную оценку этой взаимосвязи Таблица сопряженности при альтернативном распределении признака имеет следующий вид (табл 86)

Таблица 86

Макет четырехклеточной таблицы взаимной сопряженности (2×2)

Признаки	*A*	не *A*	*B*
B	*a*	*b*	*a+b*
не *B*	*c*	*d*	*c+d*
$\sum A$	*a+c*	*b+d*	*n*

В случае, когда данные представлены такой таблицей, можно использовать **коэффициент ассоциации** (коэффициент Юла), определяемый по формуле

$$Q=\frac{ad-cb}{ad+cb}$$

Например (табл. 87):

Таблица 87

Распределение рабочих по случаям заболеваний в году

Профессии	Болели	Не болели	Итого
Станочники	78	83	161
Прочие	30	92	122
Всего	108	175	283

Требуется определить насколько сильная связь между заболеваемостью рабочих и их профессией

Подставив данные из таблицы в формулу, получим значение коэффициента ассоциации

$$Q=\frac{78\times92-30\times83}{78\times92+30\times83}=0,49\approx0,5$$

Связь считается установленной, если абсолютные значения коэффициента ассоциации находятся в пределах от 0,5 до 1 В приведенном примере связь можно считать установленной и достаточно ярко выраженной

Более осторожную оценку связи дает **коэффициент контингенции** Φ, числовое значение которого всегда меньше, чем значение коэффициента ассоциации из-за того, что Φ, в отличие от Q, дает двустороннюю оценку связи

$$\Phi = \frac{ad - cb}{\sqrt{(a+b)(c+d)(b+d)(a+c)}} =$$

$$= \frac{78 \times 92 - 30 \times 83}{\sqrt{(78+83)(30+92)(78+30)(83+92)}} = 0{,}24$$

В приведенном примере связь между двумя признаками нельзя считать плотной, но все же она заметна.

При использовании коэффициентов Q *и* Φ следует осторожно относиться к оценкам направленности связи (прямая или обратная) Это вызвано тем, что некоторые качественные признаки (пол, диагноз заболевания и т п) не имеют упорядоченного ранжирования Наличие формальной прямой или обратной связи зависит в этой ситуации только от произвольного расположения исходных данных в таблице Но сам факт наличия или отсутствия связи в целом, бесспорно, устанавливается на основании абсолютного значения коэффициентов Q или Φ

Другой особенностью коэффициента Φ является его существенно заниженные значения при анализе взаимосвязей признаков в случае резко асимметричного распределения данных в клетках таблицы

Таким образом, в случае альтернативного распределения качественных признаков (да или нет, болен или здоров и т п) максимально возможная величина коэффициента Φ зависит от распределения частот в клетках таблицы Соответственно, верхняя и нижняя границы коэффициента Φ могут быть меньше +1 и больше −1 Для определения верхнего максимального значения коэффициента можно воспользоваться формулой $\sqrt{(\pi_1/(1-\pi_1))(1-\pi_2)/\pi_2}$, где π_1 и π_2 соответственно минимальная и максимальная частоты по строкам и столбцам таблицы [В В Власов, 1988] Нижнее, минимальное отрицательное значение коэффициента Φ, при

сумме частот $\pi_1+\pi_2 \le 1$ составляет $-\sqrt{(\pi_1/(1-\pi_1))(\pi_2/(1-\pi_2))}$, а при сумме частот $\pi_1+\pi_2 > 1$ — $-\sqrt{((1-\pi_1)/\pi_1)((1-\pi_2)/\pi_2)}$

Таким образом, максимально возможные значения коэффициента Φ могут в отдельно взятом случае значительно отличаться по абсолютному значению от 1 По данным рассмотренного выше примера (табл 87), максимально возможное значение

$$\Phi=\sqrt{(0{,}22/(1-0{,}22))(1-0{,}76)/0{,}76}=0{,}35$$

Фактическое значение $\Phi=0{,}24$, что ненамного отличается от максимально возможного для этой таблицы Таким образом, несмотря на малое числовое значение коэффициента Φ сила (плотность) статистической связи между профессией и заболеваемостью в данном примере весьма существенна В нормированном значении $\Phi_{норм}=0{,}24/0{,}35=0{,}7$

Оценку статистической достоверности коэффициентов Q или Φ целесообразно проводить с помощью коэффициента согласия Пирсона χ^2 (хи-квадрат) (См разделы «Критерий согласия Пирсона χ^2» и «Коэффициенты сопряженности Пирсона и Чупрова»)

6.2.2. Коэффициенты сопряженности Пирсона *(C)* и *Чупрова (K)*

Показатели взаимосвязи C и K предназначены для измерения связи между количественными признаками, располагающимися в таблицах размером $s \times t \ge 2$, где s и t соответственно число строк и столбцов (без итогов).

Вычисление этих коэффициентов основывается на измерении отклонений наблюдаемых частот в клетках таблицы от теоретических (ожидаемых) частот, которые задаются в предположении о независимости признаков Математическая логика применения показателей C и K базируется на том, что если теоретическое распределение говорит о независимости (не связанности) признаков, то отклонение фактического от теоретического распре-

деления будет тем больше, чем больше признаки взаимосвязаны Это отклонение оценивается с помощью критерия хи-квадрат (χ^2), который включен в формулы в качестве основного элемента Сам по себе критерий χ^2 может свидетельствовать о наличии или отсутствии взаимосвязи, однако его значение во многом зависит от числа наблюдений (n) Таким образом, его значения напрямую не могут использоваться для сравнения корреляционных связей в совокупностях с разным числом наблюдений Этот недостаток устраняется при использовании коэффициентов C и K, имеющих сходные формулы

$$C=\sqrt{\frac{x^2}{x^2+n}} \text{ и } K=\sqrt{\frac{x^2}{n\sqrt{(s-1)(t-1)}}}$$

Критерий Чупрова K дает более точные результаты, когда таблица сопряженности имеет квадратную форму и число строк и столбцов не превышает 5×5 (без итогов)

Оценку статистической достоверности (значимости) коэффициентов C и K производят на основе критерия χ^2, значения которого использовались для получения этих коэффициентов

Этим критериям, как правило, не приписывают никакого знака (+ или —), так как, с математической точки зрения, они являются результатом вычисления квадратного корня Направление связи устанавливается из содержательного смысла исходных данных В том случае, когда хотя бы один из признаков не может быть ранжирован (пол, национальность и т п), выяснение направления связи вообще теряет всякий смысл.

Так как расчет критериев C и K производится на основе χ^2, то все ограничения χ^2 распространяются и на них Ограничения эти следующие

1. Частоты в клетках не должны быть меньше 5 Если частоты меньше, то необходимо объединять соседние строки или столбцы

2 Общее число наблюдений в совокупности должно быть не менее 50—100

Таблица 88

Взаимосвязь заболеваемости и профессиональной принадлежности работников предприятия

Проф группы	Кратность заболеваний			Всего
	Не болели	1–2 раза	3 и более	
Станочники	11	12	23	46
Слесари	17	16	24	57
Прочие	72	11	26	109
Итого	100	39	73	212

Критерий χ^2 вычисляется по формуле

$$\chi^2 = \frac{(p_i - p_i')^2}{p_i'},$$

где p_i — фактические частоты, p_i' — теоретические частоты; или по более удобной в практическом применении формуле

$$\chi^2 = n\left(\sum_{i=1}^{i}\sum_{j=1}^{j}\frac{f_{ij}^2}{f_j f_i} - 1\right).$$

Например

$$\chi^2 = 212\times\left(\frac{11^2}{100\times 46} + \frac{17^2}{100\times 57} + \frac{72^2}{100\times 109} + \frac{12^2}{39\times 46} + \frac{16^2}{39\times 57} + \frac{11^2}{39\times 109} + \right.$$

$$\left. + \frac{23^2}{73\times 46} + \frac{24^2}{73\times 57} + \frac{26^2}{73\times 109} - 1\right) = 33{,}4$$

Отсюда: критерий $C = \sqrt{\frac{33{,}4}{33{,}4+212}} = 0{,}37$

Критерий Чупрова $K = \sqrt{\frac{33{,}4}{n212\sqrt{(3-1)\times(3-1)}}} = 0{,}28$

Статистические оценки значений критериев C и K производятся в тех же границах, что и обычных коэффициентов корреляции: от 0 до 1.

Несмотря на функциональное сходство существуют определенные отличия в интерпретации силы связи по результатам, полученным с помощью коэффициента C Предел значений коэффициента сопряженности C также равен 1,0. К этому значению он стремится при увеличении силы связи между признаками Но не достигает этого значения даже в случае, когда связь становится полной. Это связано с тем, что величина предельного значения C зависит и от распределения частот в таблице, и от размеров самой таблицы Знание максимально возможного значения коэффициента сопряженности для каждого конкретного случая (C_{max}), позволяет получить, как и в описанном выше случае коэффициента Φ, нормированное значение:

$$C_{норм} = \frac{C}{C_{max}},$$

которое также является более точной характеристикой для каждого конкретного случая

Для симметричной таблицы, где число строк и число столбцов равно между собой:

$$C_{max} = \sqrt{\frac{t-1}{t}},$$

где t — число строк. В приведенном примере:

$$C_{max} = \sqrt{\frac{3-1}{3}} = 0{,}81$$

Отсюда нормированное значение в приведенном примере

$$C_{норм} = \frac{0{,}37}{0{,}81} = 0{,}46$$

Несмотря на то что разница между нормированным и ненормированным значениями коэффициента сопряженности в данном случае оказалась незначительной, численное значение $C_{норм}$ заметно повышает убедительность доказательства связи

Следует отметить, что при увеличении размера таблиц C_{max} тоже увеличивается, стремясь к предельному значению равному 1, обычно и принимаемому как верхняя возможная граница оценки корреляционной связи Таким образом, необходимость в

корректировке значений *C* по мере роста формата таблиц снижается Дж. Юл и М. Кендэл (1960) считали, что в таблицах 5×5 необходимость в корректировке вообще отпадает Однако, учитывая, что разбиение качественных признаков на такое количество диапазонов встречается достаточно редко из-за относительно малого числа наблюдений в клинических исследованиях, применение нормированных значений оценок сопряженности представляется весьма целесообразным.

6.2.3. Вычисление критерия сопряженности в *MS Excel*

Для удобства работы возьмем использованные в предыдущем примере данные (см. табл 88) и расположим их в таблице *Excel* следующим образом (рис. 99):

	A	B	C	D
1	11	12	23	46
2	17	16	24	57
3	72	11	26	109
4	100	39	73	212

Рис. 99. Размещение исходных данных

1. Для того что бы вычислить «теоретические» частоты, в клетке с координатами A6 наберите формулу A$4*$D1/$D$4 / D В этой формуле A4 — итог по столбцу A; D1 — итоги по первой строке; D4 — итоговая сумма. Знак $ используется для фиксации адреса при копировании **Напоминаем**: все адреса указываются латинским шрифтом!

2. Скопируйте набранную формулу в область A6 C8 После копирования в этой области будут расположены «ожидаемые» частоты (рис. 100).

21,69811321	8,46226	15,84
26,88679245	10,4858	19,627
51,41509434	20,0519	37,533

Рис. 100 **Размещение ожидаемых частот**

3. Установите курсор в ячейку E11 и занесите в нее функцию **ХИ2ТЕСТ** из меню <**Вставка**> и <**Функция**> В открывшемся окне укажите фактический интервал A1:C3, ожидаемый интервал A6:C8 Нажмите клавишу OK.

4. Для того что бы получить значение коэффициента хи-квадрат согласно формуле $C=\sqrt{\frac{x^2}{x^2+n}}$, установите курсор в позицию E12. Затем вызовите функцию **ХИ2ОБР** и укажите в соответствующих окнах адрес вычисленной вероятности E11 и число степеней свободы (3−1)*(3−1)=4.

5 Значение критерия сопряженности вычисляется через функцию **КОРЕНЬ** В окно введите выражение E12/(E12+D4), где E12 — адрес значения хи-квадрат, D4 — адрес общего числа наблюдений.

Результаты, после введения поясняющих надписей, будут представлены в следующем виде (рис 101)·

Значимость критерия X^2	1,00252E-06
Величина критерия X^2	33,37860107
Критерий сопряженности	0,368821069

Рис. 101. **Размещение поясняющих надписей**

7. Статистические критерии различия

Статистический анализ различий является одной из коренных задач медико-биологических исследований *Например* характеристика эффективности фармакологического препарата или оценка неблагоприятного гигиенического фактора складывается из статистических оценок различий каких-либо физиологических параметров подопытных и контрольных групп лабораторных животных, различий всевозможных клинических вариантов течения заболевания и его исходов, различий показателей здоровья разных профессиональных групп, разных групп населения и т п

Для того чтобы получить исчерпывающий ответ при решении такого рода задач, необходимо ответить на вопросы, которые на первый взгляд не являются напрямую связанными с оценкой различий

1 Относится та или иная варианта к данной статистической совокупности?

2 Соответствует ли полученное распределение тому или иному теоретическому распределению?

3 Соответствует ли полученное распределение другому эмпирическому (опытному) распределению?

4 Являются ли конкретные выборочные группы наблюдения выборками из одной и той же генеральной совокупности?

Необходимость решения первого вопроса диктуется сильной вариабельностью данных медико-биологических исследований Существенные, экстремальные отклонения вариант от среднего значения нередко вызывают сомнения, действительно ли та или иная варианта относится к конкретной группе или она включена в группу ошибочно Более того, нередко встречается ситуация, когда именно из-за этих, так называемых «выскакивающих вариант», создается впечатление существенных различий сравниваемых групп

Ответ на второй вопрос позволяет обосновать выбор того или иного метода статистического анализа данных, в том числе и выбор метода оценки различий

Решение третьего вопроса дает ответ о различиях исследуемых групп Следует отметить, что средние величины, ошибки средних, дисперсии и т п чаще всего не являются конечной целью статистического исследования, а являются лишь статистическими параметрами, с помощью которых изучается вопрос о различиях групп

Ответ на четвертый вопрос решает ключевую проблему всех выборочных наблюдений, которые представляются основными в медико-биологических исследованиях действительно ли различаются выборочные группы или различия, которые на первый взгляд есть, объясняются просто обычной вариабельностью исходных данных одной генеральной совокупности?

Нетрудно заметить, что, по существу, все три вопроса сводятся к одному — являются ли наблюдаемые различия отражением объективно существующей реальности Указанный вопрос решается проверкой соответствующих статистических гипотез

7.1. Принадлежность варианты к совокупности

Принадлежность варианты к конкретной совокупности решается довольно просто, если объем наблюдения достаточно велик В этом случае крайние варианты (очень большие или очень маленькие, часто называемые экстремальными вариантами или «выбросами»), а именно они вызывают сомнения относительно своей принадлежности к статистической совокупности, просто отбрасывают Это вполне допустимо, ибо в большой совокупности роль («относительный вес») нескольких крайних вариант при вычислении средних величин, как правило, весьма незначителен Выборка, оставшаяся после этой операции, называется **«цензурированная выборка»**

Если же выборка мала, то потеря даже одной варианты может исказить результат всех последующих вычислений При такой ситуации решение вопроса о принадлежности варианты к сово-

купности не представляет сложности, если распределение в этой совокупности является нормальным Для этого достаточно использовать правило трех сигм Согласно этому правилу, в пределах $M \pm 3\sigma$ находится 99,7% всех вариант. Поэтому если варианта попадает в этот интервал, то она считается принадлежащей к данной совокупности Если не попадает, то ее нужно отбросить. *Например:* требуется определить правомерность включения в нормально распределенный вариационный ряд варианты 58 кг, если среднее арифметическое этого ряда $M=40$ кг, а $\sigma=2,5$ кг Согласно правилу трех сигм, в пределах $40 \pm 3 \times 2,5$ кг или от 47,5 до 32,5 кг должно находиться 99,7% всех вариант Варианта 58 кг явно не попадает в этот интервал Таким образом, с высокой степенью вероятности можно утверждать, что данная варианта не относится к указанному ряду

Рассмотренная выше методика определения «выскакивающих» вариант неудобна для практического использования Во-первых, она требует предварительных расчетов среднего арифметического и сигмы (среднеквадратического отклонения) Во-вторых, требует пересчета этих показателей после отброса «выскакивающих» вариант Более простой способ решения вопроса о принадлежности вариант связан с использованием величин:

$$\tau' = \frac{V_n - V_{n-1}}{V_n - V_2}$$ для наименьшей варианты;

$$\tau'' = \frac{V_2 - V_1}{V_{n-1} - V_1}$$ для наибольшей варианты

Критические значения τ с учетом уровней значимости приводятся в специальных таблицах *Например* задана совокупность $V_1=45$, $V_2=55$, $V_3=58$, $V_4=62$, $V_5=64$ Вызывает сомнение правомерность включения в этот ряд варианты $V_1=45$. По приведенной формуле имеем

$$\tau = \frac{55-45}{62-45} = \frac{10}{17} = 0,59.$$

По таблице, при числе наблюдений $n=5$ и уровне значимости $P=0,01$, $\tau=0,92$. Поскольку вычисленное значение показателя

меньше табличного (теоретического), варианту V_i =45 отбросить нельзя

При использовании приведенных здесь методик, следует помнить, что они основаны на нормальности распределения В ситуации, когда распределение вариант отличается от нормального, например, если асимметрия и эксцесс больше 0,5, эти оценки применять нельзя.

7.1.1. Определение «выскакивающей» варианты с помощью *MS Excel*

В *MS Excel* имеется функция «**ZTECT**», с помощью которой можно определить принадлежность варианты к совокупности.

Рассмотрим пример Имеется некая статистическая совокупность, образованная следующим рядом числовых значений 14, 15, 17, 18, 22, 24, 19. Необходимо определить, соответствует ли варианта 24 этому ряду?

Порядок решения задачи в *MS Excel:*

1. Введите в ячейки A1 A7 исходный ряд чисел

2. Установите курсор на ячейку (в позицию A9)

3 Выполните команду <**Функции**> из меню <**Вставка**> или выберите на панели инструментов пиктограмму f_x

4 Выберите в появившемся диалоговом окне «**Мастер функций**» категорию «**Статистические**», а затем функцию «**СТАНДОТКЛОНП**» и нажмите кнопку **[OK]**

6. Введите в окошко «Число» координаты числового ряда A1:A7 и нажмите кнопку **[OK]**

Таким образом, в клетке A9 Вы получили значение среднеквадратического отклонения (сигму).

7. Установите курсор в позицию A10.

8 Выберите в диалоговом окне ««**Мастер функций**» категорию «**Статистические**», а затем функцию «**ZTECT**» и нажмите кнопку **[OK]**.

9 В открывшемся окне установите следующие параметры.

- Массив — (A1 A7)
- X — число, сомнительную принадлежность которого к совокупности вы желаете оценить, в данном случае 24, или координаты клетки, в которой это число расположено (A6)
- Сигма — значение стандартного (среднеквадратического отклонения), вычисленного вами ранее (A9).

10 После ввода исходных данных нажмите кнопку [OK]

В результате, в клетке A10 будет получена вероятность того, что сомнительное число принадлежит к данному ряду чисел (0,9999).

Если числовое значение полученной вероятности будет ≤ 0,05, то рекомендуется провести повторный анализ ситуации, исключив предварительно сомнительное число из числового ряда. Если полученные после повторного анализа результаты будут все равно оставаться на уровне критических значений вероятностей, то выбор решения о включении сомнительного числа в совокупность или его исключении из совокупности остается за исследователем

С помощью встроенной функции *MS Excel* **УРЕЗСРЕДНЕЕ** можно исключать из анализируемого ряда «выскакивающие варианты». При вычислении этой функции необходимо указать

- Анализируемый числовой ряд «**МАССИВ**»
- Долю вариант исключаемых из выборки· «**ДОЛЯ**» Эта доля указывается в относительных единицах *Например.* если нужно отбросить 10% вариант, то указывают 0,1, если 20% — указывают 0,2

Результатом вычислений является урезанное среднее арифметическое, т. е. среднее, вычисленное из урезанного, цензурированного числового ряда. Если после указанной операции урезанное среднее и исходное среднее не отличаются друг от друга или отличаются несущественно, то варианту (группу вариант) можно оставить в составе анализируемого ряда

7.2. Критерии различий эмпирических распределений

В ряде случаев нет необходимости сравнивать полученные распределения с теоретическими, поскольку достаточно ответить на вопрос о различиях двух или нескольких распределений полученных в опыте (эмпирических распределений) Для этих целей используются методические подходы во многом аналогичные тем, которые применяются в случае оценок различий эмпирических и теоретических распределений В таком случае вместо теоретического распределения берется второе эмпирическое *Например:* критерий Колмогорова—Смирнова Однако наиболее часто в этой ситуации используется критерий χ^2 (хи-квадрат) Его использование удобно, во-первых, потому, что не требует громоздких вычислений. Во-вторых, не требует никаких параметрических характеристик (средних, дисперсии и т п.) Ограничения, существующие при использовании этого критерия, описаны в разделе «Критерий согласия *Пирсона* χ^2»

Таблица 89

Распределение рабочих по кратности заболеваний (исходные данные)

Пол	Кратность заболеваний			Итого
	Не болели	1—2 раза	3 раза и более	
Мужчины	29	36	15	80
Женщины	14	24	12	50
Оба пола	43	60	27	130

Необходимо определить· зависит ли заболеваемость обследованных рабочих от пола? Или зависит ли распределение показателей заболеваемости (распределение по кратности заболеваний) от пола? (табл 89)

Порядок решения поставленной задачи может быть следующим.

1 Выдвигается нулевая гипотеза H_0 — *влияния пола на кратность заболеваний рабочих нет*

2. Вычисляются ожидаемые численности наблюдений для каждой клетки по следующему отношению (сумма по строке)× ×(общий итог)/(сумма в столбце). В данной таблице ожидаемые числа показывают, какое было бы распределение болевших и не болевших, если бы нулевая гипотеза была бы верна Для не болевших мужчин (80×43)/130=26,5 Для не болевших женщин (50×43)/130=16,5 Аналогично получают числа для мужчин и женщин, болевших 1–2 раза, 3 раза и более (табл 90)

Таблица 90

Распределение рабочих по кратности заболеваний (ожидаемые числа)

Пол	Кратность заболеваний		
	Не болели	1–2 раза	3 раза и более
Мужчины	26,5	36,9	16,6
Женщины	16,5	23,1	10,4
Оба пола	43,0	60,0	27,0

Таблица 91

Распределение рабочих по кратности заболеваний
(величина отклонения фактических от ожидаемых чисел)

Пол	Кратность заболеваний		
	Не болели	1–2 раза	3 раза и более
Мужчины	2,5	−0,9	−1,6
Женщины	−2,5	+0,9	1,6

3 Вычисляются отклонения каждой фактической величины от ожидаемой (табл 91)

4 Вычисляем значение χ^2, подставляя полученные значения:

$$\chi^2 = \frac{2,5^2}{26,5} + \frac{-0,9^2}{36,9} + \frac{-1,6^2}{16,6} + \frac{-2,5^2}{16,5} + \frac{0,9^2}{23,1} + \frac{1,6^2}{10,4} = 1,1$$

5. Число степеней свободы = (a−1)×(b−1)= (3−1)×(2−1)=2, где a — число столбцов (без итогов), b — число строк (без итогов).

6 При уровне значимости 0,05 и числе степеней свободы равном 2, критическое значение $\chi^2=5,99$ (табл 115 Приложения). Поскольку вычисленное значение $\chi^2=1,1<5,99$, нулевая гипотеза не отвергается Различия в распределении рабочих по группам заболеваний не подтверждены Упростить вычисления можно применив формулу.

$$\chi^2 = n\left(\sum_{i=1}^{i}\sum_{j=1}^{j}\frac{f_{ij}^2}{f_j f_i} - 1\right),$$

где f_{ij} — число наблюдений в ij клетке, f_j — число наблюдений в j столбце, f_i — число наблюдений в i строке таблицы, n — общее число наблюдений в таблице.

Несмотря на кажущуюся сложность формулы, вычисления достаточно просты *Например*: по данным табл. 89 имеем

$$\chi^2=130\times\left(\frac{29^2}{43\times80}+\frac{36^2}{60\times80}+\frac{15^2}{27\times80}+\frac{14^2}{43\times50}+\frac{24^2}{60\times50}+\frac{12^2}{27\times50}-1\right)=1,1$$

В тех случаях, когда данные представлены в виде четырехпольной таблицы, целесообразно использовать упрощенный метод вычисления критерия χ^2, используя формулу.

$$\chi^2=\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)},$$

где a, b, c, d — условные обозначения клеток таблицы *Например*

Таблица 92

Результаты проведения профилактической вакцинации

Отношение к вакцинации	Число обследованных		
	Заболели	Не заболели	Итого
Вакцинированы	a	b	$a+b$
	20	140	160
Не вакцинированы	c	d	$c+d$
	48	120	168
Всего	$a+c$	$b+d$	$a+b+c+d$
	68	260	328

Подставив данные (табл 92), получим:

$$\chi^2 = \frac{(20 \times 120 - 48 \times 140)^2 \times 328}{68 \times 260 \times 160 \times 168} = 12{,}8$$

Поскольку число степеней свободы в случае четырехпольной таблицы всегда равно 1, критические значения χ^2 при уровне значимости $P = 0{,}05$ всегда — 3,84 и при $P = 0{,}01$ — 6,63 Так как вычисленное значение $\chi^2 = 12{,}8 > 6{,}63$, нулевая гипотеза отвергается. Различия в распределении по частоте заболеваний в группах привитых и не привитых статистически подтверждены

Весьма удобным и простым в применении методом оценки различий в двух распределениях является оценка различий с помощью *t*-критерия для зависимых выборок

Условия применения *t*-критерия для зависимых выборок аналогичны условиям применения *t*-критерия для независимых выборок. Это означает, что попарные разности наблюдений в сравниваемых совокупностях должны быть нормально распределены. Если это условие не выполняется, то можно воспользоваться упомянутым выше непараметрическим критерием χ^2. Более подробно эта методика рассмотрена ниже (см раздел «Оценка различий между двумя долями, интенсивными величинами и средними арифметическими с помощью *t*-критерия»)

7.2.1. Оценка различий эмпирических распределений с помощью *MS Excel*

В *MS Excel* имеется функция **ХИ2ТЕСТ**, с помощью которой можно оценить различия эмпирических распределений. Ее использование уже рассматривалось выше при решении близкой по смыслу задачи (см. раздел 6.2.3).

Рассмотрим приводившийся выше пример (см табл 89) Необходимо определить, зависит ли заболеваемость от пола рабочих?

Порядок решения задачи в *MS Excel*

1. Введите в ячейки A1:E5 исходный ряд чисел (рис 102).

	A	B	C	D	E
1		Кратность заболеваний			
2	Пол	Не болели	1-2 раза	3 раза и более	Итого
3	Мужчины	29	36	15	80
4	женщины	14	24	12	50
5	Оба пола	43	60	27	130

Рис. 102. **Исходный ряд чисел**

2 Для получения «ожидаемых» распределений наберите в ячейке B9 команду =B$5*$E3/$E$5

3 Скопируйте это выражение в ячейки B10, C9, C10, D9 и D10. Для того чтобы провести копирование, сделайте активной клетку B9 (установите на нее курсор), затем подведите к правому нижнему углу выделенной клетки указатель мыши и нажав на правую клавишу перетащите содержимое этой клетки в нужное вам место

7		Ожидаемое распределение		
8	Пол	Не болели	1-2 раза	3 раза и бол
9	Мужчины	26	37	17
10	Женщины	17	23	10

Рис. 103. **Ожидаемое распределение**

4 Установите курсор на свободное место таблицы Выполните команду <**Функции**> из меню <**Вставка**> или выберите на панели инструментов пиктограмму f_x Выберите в появившемся диалоговом окне «**Мастер функций**» категорию «**Статистические**», а затем функцию «**ХИ2ТЕСТ**» и нажмите кнопку **[ОК]**.

В результате будет получена вероятность нулевой гипотезы, т. е что различий в распределении нет (0,576) (рис. 104)

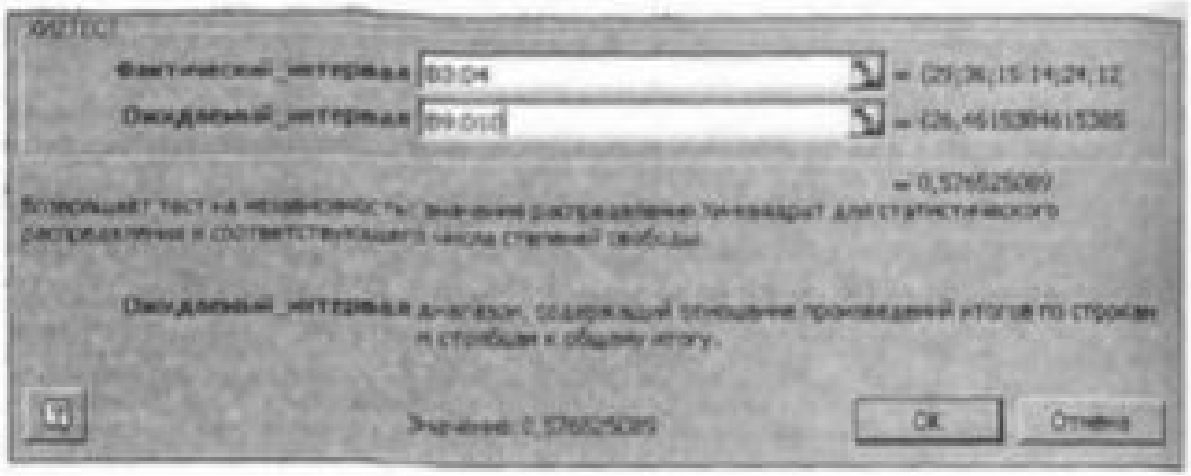


Рис. 104 **Пример заполнения функции ХИ2ТЕСТ**

7.3. Оценка различий между двумя долями, интенсивными величинами и средними арифметическими с помощью t- критерия

Вопрос оценки различий между статистическими совокупностями (распределениями) с помощью таких параметров выборки, как средние величины или доли (%), является одним из самых важных в статистике медико-биологических исследований Многие исследования, в большей части совершенно не обоснованно, заканчиваются ответом именно на этот вопрос *Например:* при оценке токсичности какого-либо вещества обычно берутся две группы лабораторных животных Подбираются животные одинакового возраста, пола, одинакового содержания и т. п. Т е. делается все, чтобы эти группы животных представляли собой единую, как можно более однородную статистическую совокупность, с тем, чтобы максимально снизить исходную вариабельность анализируемых данных Оптимальным с этой точки зрения считается ситуация, когда отличия сравниваемых групп заключаются только в том, что одна из групп (опытная) подвергается воздействию токсического вещества, а другая (контрольная) — нет. В любом случае, произошли ли после воз-

действия токсического вещества изменения в опытной группе или нет, различия средних показателей в обоих группах обязательно будут. Вопрос состоит в том, являются ли эти различия только следствием различий, обычно существующих в любых выборках, или разница возникла из-за того, что произошли существенные сдвиги физиологических функций животных опытной группы. Иначе, принадлежат ли животные опытной и контрольной групп к той же самой генеральной совокупности или опытная группа принадлежит к другой генеральной совокупности (совокупности с измененными физиологическими параметрами)?

Для сравнения долей в случае, когда выборки, из которых эти доли вычислены, достаточно велики, используется формула

$$t=\frac{P_1-P_2}{\sqrt{m_{P_1}^2+m_{P_2}^2}},$$

где P_1 и P_2 — сравниваемые доли (%) Ошибки долей m могут быть вычислены по формуле:

$$m=\sqrt{\frac{Pq}{n}},$$

где P — показатель в %, а $q=100-P$ Одной из весьма распространенных ошибок является использование для определения различий интенсивных показателей этой формулы, предназначенной только для альтернативного варьирования, т е для случаев, когда возможен один вариант «болел» или «не болел». Однако для оценки показателей общей заболеваемости (по обращаемости), заболеваемости с временной утратой трудоспособности, а также заболеваемости теми болезнями, которые могут встречаться по несколько раз в исследуемый период (год, квартал и т п.) эта формула не пригодна В А Мозглякова (1964) предложила использовать в этих случаях формулу $m=\sqrt{\frac{M}{n}}$, где M — среднее число заболеваний, n — число наблюдений Эта формула обычно используется при распределении Пуассона или асимметричном биноминальном распределении, которому под-

чиняются, как правило, распределения случаев заболеваний при относительно невысоких показателях и небольшом числе наблюдений (≈100) При больших числах наблюдений или высоких показателях заболеваемости (>150 сл. на 100 рабочих и т п) определять статистические ошибки следует только через построение вариационных рядов распределения *Например*

Кратность заболеваний в году *V*	Число рабочих *P*
0	40
1	55
2	15
И т д	

И далее см в разделах «Ряды распределений. Вариационные ряды Дисперсия (D)»

При статистической оценке различий выборочных средних могут рассматриваться следующие варианты, накладывающие отпечаток на способ вычисления *t*-критерия: малые выборки ($n<30$) или большие выборки ($n>30$)

Кроме того, выбор конкретной методики оценки различий требует также и учета следующих аспектов.

- различия определяются для средних арифметических в независимых (несвязанных) выборках, их состав не зависит друг от друга, формируется случайным отбором;
- различия вычисляются для средних в парных, связанных выборках, их состав тесно связан или тот же самый (до опыта и после);
- различия находятся между выборочными и генеральными средними (теоретическими стандартами)

Наиболее широко для решения этих задач в практике медико-биологических исследований используется параметрический критерий Стьюдента (*t*). Применение этого критерия не допустимо в следующих случаях:

- не соответствие распределения вариант в сравниваемых группах закону нормального распределения,

- не подтвержденное равенство несмещенных дисперсий (разброса вариант в выборке) сравниваемых групп;
- дискретный (балльный) характер сравниваемых числовых признаков

Рассмотрим несколько примеров, в которых распределение исходных данных будем считать соответствующим нормальному распределению.

Пример 1: статистическая проверка гипотезы разности двух средних Независимые выборки, небольшой объем выборки, число наблюдений одинаково (табл 93)

Принимаем нулевую гипотезу — различий между средними нет H_0 $M_1=M_2$ (различия случайны, различия не существенны), при уровне значимости нулевой гипотезы $P=0,05$

1 Среднее первой выборки $M_1=124/8=15,5$ дня Второй выборки $M_2=135/8=16,9$ дня.

2 Находим несмещенные дисперсии по формуле

$$D=\frac{\sum V_j^2-M^2n}{n-1}$$

$$D_1=\frac{1936-15,5^2\times 8}{8-1}=2,0 \text{ дня};\quad D_2=\frac{2293-16,9^2\times 8}{8-1}=2,1 \text{ дня}$$

Таблица 93

Длительность пребывания в стационарах больных астматическим бронхитом

Номер больного j	Число дней госпитализации V			
	Первая больиица		Вторая больница	
	V_1	V_1^2	V_2	V_2^2
1	14	196	15	225
2	16	256	17	289
3	15	225	16	256
4	17	289	17	289
5	18	324	19	361
6	15	225	19	361
7	14	196	16	256
8	15	225	16	256
Итого	124	1936	135	2293

3 С помощью критерия Фишера проверяем равенство дисперсий для уровня значимости $P=0,05$ и для степеней свободы, соответственно равным для первого числового ряда $k_1=8-1=7$ и второго ряда $k_2=8-1=7$. Наблюдаемое значение критерия Фишера $F=2,1/2,0=1,05$ Критическое значение (определяется либо по специальной таблице, либо с помощью *Excel*) $F_{кр}=3,78$ Поскольку фактическое значение критерия (1,1) не превышает его критического значения (3,78), нулевая гипотеза, говорящая о равенстве дисперсий, не отвергается

4 Вычисляем ошибки средних:

$$m=\sqrt{\frac{D}{n}};\ m_1=\sqrt{\frac{2,0}{8}}=0,53 \text{ и } m_2=\sqrt{\frac{2,1}{8}}=0,55 \text{ дня}$$

5 Вычисляем критерий Стьюдента:

$$t=\frac{M_1-M_2}{\sqrt{m_1^2+m_2^2}}=\frac{16,9-15,5}{\sqrt{0,55^2+0,53^2}}=1,79$$

Примечание: *строго говоря, значение критерия Стьюдента в приведенном примере равно −1,79 Однако при вычислении этого критерия принимается только его абсолютное значение*

6 Находим число степеней свободы по формуле $k=(n_1-1)+(n_2-1)=n_1+n_2-2$, где n_1 — численность первой выборки; n_2 — второй выборки $k=8+8-2=14$

7. Критическое значение $t=2,15$, при уровне значимости $P=0,05$ и числе степеней свободы $k=14$ (Приложения, табл 116) Таким образом, фактическое, вычисленное значение критерия *Стьюдента* (1,79) меньше критического (2,15), т е находится в области принятия нулевой гипотезы. Различия длительности стационарного лечения в первой и второй больницах можно считать статистически не существенными

Пример 2: зависимые выборки· результаты обследования одних и тех же лабораторных животных (до опыта и после опыта), одним и тем же способом (табл 94) Дисперсии в этом случае могут быть не известны Более того, в этом случае, формально, нет необходимости вычислять и средние. Несмотря на го что при вычислениях *t*-критерия таким способом не используются параметры нормального распределения (средние и дисперсии),

этот метод считается параметрическим, поскольку попарная разность d должна подчиняться закону нормального распределения.

Таблица 94

Сравнение результатов обследования

Номер животного	Результаты обследования		Разность d	Квадрат разности d^2
	До опыта	После опыта		
1	14	15	−1	1
2	16	17	−1	1
3	15	16	−1	1
4	17	17	0	0
5	18	19	−1	1
6	15	19	−4	16
7	14	16	−2	4
8	15	16	−1	1
Итого			−11	25

1. Принимаем нулевую гипотезу — средняя разность не существенна H_0. $d_{ср}=0$ и рассчитываем фактическое t.

$$t=\sum d / \sqrt{\left(N\sum d^2-\left(\sum d\right)^2\right)/(N-1)};$$

$$t=-11 / \sqrt{\left(8\times 25-11^2\right)/(8-1)}=3{,}27$$

2 Число степеней свободы $k=n-1=8-1=7$ дает возможность найти при доверительной вероятности 0,05 критическое значение $t=2{,}4$ Поскольку фактическое значение критерия Стьюдента превосходит критическое (Приложения, табл 116), нулевая гипотеза отвергается. Таким образом различия «до» и «после» следует признать статистически значимыми

Если вычислить значение критерия t, используя методику, применяемую для оценки различий в несвязанных выборках, то получим значение $t=2{,}08$. Т. е различия должны были бы быть признаны не существенными

Из результатов *примера* видно, что использование вместо одного другого метода определения достоверности различий может кардинально менять результаты. Объясняется это тем, что, вмес-

то исследования каждой группы наблюдения отдельно и анализа исходных значений, рассматриваются просто разности между двумя измерениями («до опыта» и «после опыта») для каждого объекта наблюдения Вычитая первые значения из вторых (для каждого объекта) и анализируя затем только парные разности, исключают ту часть вариации, которая является результатом объективных различий, не связанных непосредственно с опытом Именно так и проводятся вычисления *t*-критерия для зависимых выборок В сравнении с *t*-критерием для несвязанных (независимых) выборок, такой подход дает всегда более объективный результат, т к критерий становится более чувствительным

Иногда выборочное значение среднего нужно сравнить с генеральным средним или с каким-нибудь теоретическим стандартом

Пример 3: требуется оценить различия между опытными (выборочными) данными и теоретическим стандартом. При определении *pH* раствора в 8 пробах были получены следующие результаты: 6,9; 7,7; 7,3; 7,1, 7,3; 7,2; 7,6; 6,9 Можно ли считать с доверительной вероятностью 0,01 реакцию раствора щелочной, если M=7,25±0,1 , а стандартное M_c=7,0 В этом случае вычисления проводятся по формуле.

$$t=\frac{M-M_c}{m}=\frac{7,25-7,0}{0,1}=2,41$$

При числе степеней свободы $k=n-1=6-1=5$ и уровне значимости $P=0,05$ теоретическое значение t=3,5 Поскольку фактическое значение критерия *Стьюдента* (2,41) меньше критического (2,57), статистическую значимость щелочной реакции раствора можно считать сомнительной Однако в данном примере можно определять одностороннее различие показателей, поскольку здесь интересует только наличие существенных различий от значений *pH* в сторону меньшую 7,0 Критическое значение одностороннего критерия $t=2,07$ (при $P=0,1$) Фактическое значение t в этой ситуации превышает критическое Таким образом, различия можно считать статистически подтвержденными

Если сравниваемые средние получены при большом объеме наблюдений и при соблюдении нормального распределения, то определение достоверности различий средних можно произвести упрощенным способом

$$(M_1 - M_2)^2 / (m_1^2 + m_2^2) \geq 9$$

Если левая часть неравенства будет больше 9, то различия средних арифметических можно считать статистически достоверными. Если меньше — то различия статистически не достоверны

7.3.1. Использование *MS Excel* при статистической проверке различий

В *MS Excel* имеются статистические функции, предназначенные для решения задач статистической проверки различий средних.

Функция **ФТЕСТ** представляет одностороннюю вероятность того, что дисперсии числового массива 1 и массива 2 различаются несущественно. Окно функции имеет поля ввода исходной информации **Массив1** и **Массив2**. В эти окна вводятся числовые последовательности или ссылки на ячейки, содержащие числа *Например,* исходная информация представлена двумя числовыми последовательностями (рис. 105)

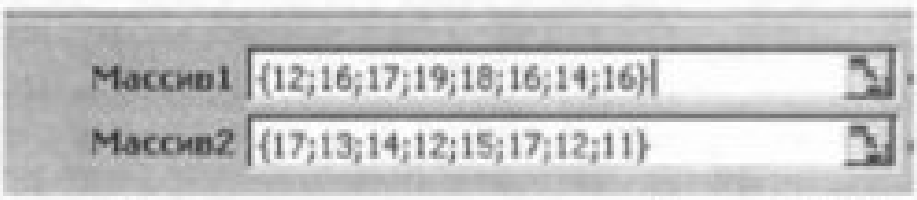


Рис. 105. **Пример заполнения окон массивов функции ФТЕСТ**

Ответ. односторонняя вероятность сходства =0,91747, т е нулевая гипотеза (H_0) об отсутствии различий дисперсий указанных числовых рядов не отвергается (P>0,05) Таким образом, имеющиеся различия можно признать не существенными

Функция **ТТЕСТ** представляет результаты (*t*) теста в виде односторонней или двусторонней вероятности того, что средние арифметические числового массива 1 и массива 2 различаются несущественно Окно функции имеет поля ввода исходной информации **Массив1** и **Массив2**. В эти окна могут вводиться числовые последовательности или ссылки на ячейки, содержащие числа. *Например*

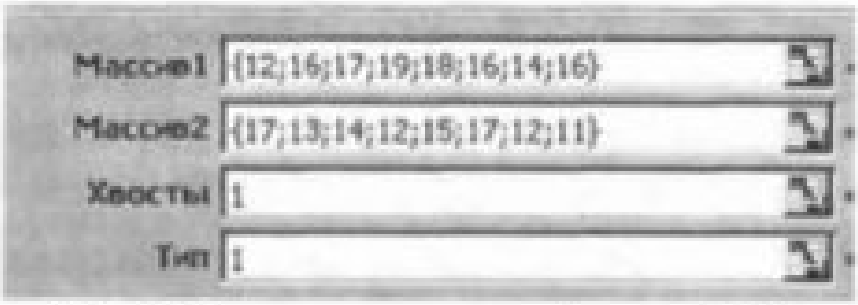


Рис. 106. **Пример заполнения окон функции ТТЕСТ**

В окно **Хвосты** вводится цифра 1 или 2, в зависимости от того, какую вероятность нужно получить: двустороннюю (2) или одностороннюю (1)

В окне **Тип** указывается тип исполняемого теста (1 — парный, 2 — двухвыборочный с одинаковыми дисперсиями, 3 — двухвыборочный с неравными дисперсиями)

При анализе данных, указанных для примера (рис. 106) получается ответ: односторонняя вероятность сходства средних парной выборки $P=0{,}073$ Таким образом, нулевая гипотеза H_0 не отвергается, поскольку $P>0{,}05$. Следовательно, имеющиеся различия можно признать не существенными.

Более глубокий статистический анализ позволяет проводить пакет анализа. Здесь предусмотрены три метода статистического анализа различий с помощью *t*-критерия. парный двухвыборочный тест, двухвыборочный тест с одинаковыми и разными дисперсиями.

Окно **парного двухвыборочного теста** содержит следующие поля ввода исходной информации (рис. 107):

1. В разделе **Входные данные**, в окнах **Интервал переменной 1 и 2** указываются числовые ряды или адреса клеток, где расположе-

ны исходные числовые данные. В окне **Гипотетическая разность** указывается (необязательно) ожидаемая разность среднего Если предполагается отсутствие различий, то здесь указывается 0 **Метки** — показывают наличие в числовых рядах текстовых заголовков, которые нужно исключить из математической обработки Параметр **Альфа** задает точность статистического анализа

2 В разделе **Параметры вывода** указывается место вывода информации:

Выходной интервал — место вывода информации на том же листе *Excel* . Причем для вывода информации достаточно указать только первую клетку, относительно которой будет располагаться вся выводимая информация. Окна **Новый рабочий лист** и **Новая рабочая книга** указывают, соответственно, что вывод будет осуществлен на новый лист или в новую книгу (новый файл) *Excel*.

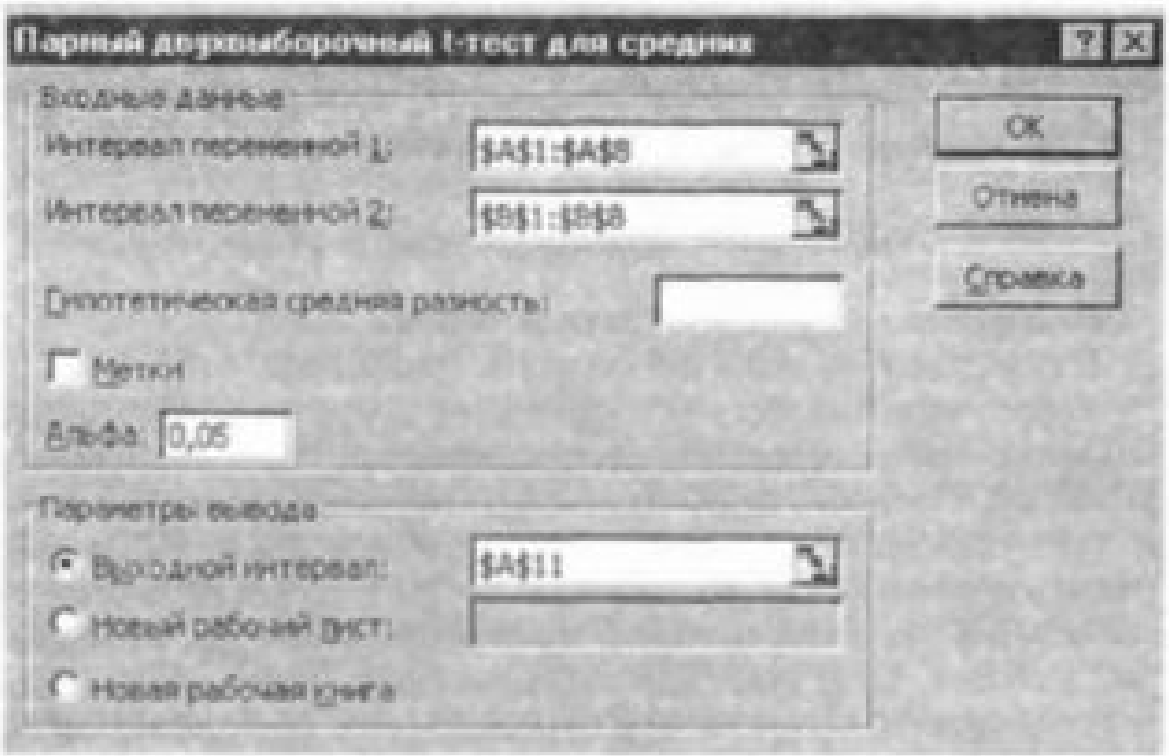


Рис. 107. **Окно ввода данных теста для средних**

Результаты анализа двух числовых рядов (рис 108) представлены следующими характеристиками:

Парный двухвыборочный t-тест для средних		
	Переменная 1	*Переменная 2*
Среднее	16	13,875
Дисперсия	4,857142857	5,267857143
Наблюдения	8	8
Корреляция Пироона	-0,338903281	
Гипотетическая разность сред	0	
df	7	
t-статистика	1,632589705	
P(T<=t) одностороннее	0,073287092	
t критическое одностороннее	1,894577508	
P(T<=t) двухстороннее	0,146574184	
t критическое двухстороннее	2,36462256	

Рис. 108. **Результаты анализа различий с помощью парного *t*-критерия**

Среднее — средние арифметические обоих рядов распределений признаков

Дисперсия — дисперсии двух выборок.

Наблюдения — числа наблюдений в выборках

Корреляция Пирсона — коэффициент корреляции Пирсона, с помощью которого оценивается направление (прямая или обратная) и плотность связи двух анализируемых признаков (от 0 до 0,3 — слабая; от 0,3 до 0,7 — средняя, более 0,7 — сильная) Эта характеристика необходима для доказательства правомерности оценки различий методом, предназначенным для несвязанных выборок Если будет установлена сильная связь анализируемых выборок, то полученные с помощью этого метода результаты будут представляться весьма сомнительными

Df — число степеней свободы.

Далее выводятся фактические и критические значения критерия Стьюдента, а также вычисленные с помощью этих харак-

теристик статистические значимости (P) различий выборок Поскольку значимыми считаются различия при $P<0,05$, в данном примере различия можно признать статистически не подтвержденными.

Аналогично осуществляется ввод и вывод информации при анализе существенности различий с помощью других методов исчисления t-критерия, имеющихся в пакете анализа *Excel*

7.4. Критерии различия между двумя средними тенденциями

Нередко в медико-биологических исследованиях применение критерия Стьюдента ограничено из-за требований, обусловленных параметрическим характером этого критерия. Непараметрические критерии не имеют таких ограничений Из большого числа таких критериев, в частности при проверке средних тенденций, можно использовать отличающиеся простотой определения критерий знаков и критерий U Вилкоксона

7.4.1. Критерий знаков

Критерий знаков — самый простой Основан на подсчете числа пар однонаправленных эффектов в парных выборках При его определении учитываются не числовые значения, а только лишь знак различия попарно связанных вариант *Например*: при пульмонологическом обследовании в группе промышленных рабочих, состоящей из 20 человек, у пяти после приема бронхолитиков был выявлен скрытый бронхоспазм Требуется установить, является ли появление бронхоспазма в этой группе статистически значимым или выявленные эпизоды можно считать случайными Находим число знаков (+ или −, «да» или «нет»), которые меньше всего встречались В приведенном примере таких случаев — 5. Критическое значение критерия знаков

(из табл. приложения) при доверительной вероятности 0,05 и числе наблюдений 20 равно 6 Таким образом, различия можно было бы признать статистически существенными (отвергнуть нулевую гипотезу — «разницы в среднем нет»), если бы число эпизодов было не менее 6 Поскольку это условие не выполнено (фактическое число различий 5), можно считать появление случаев бронхоспазма в испытуемой группе статистически случайным

При использовании критерия знаков учитывается не вся информация, содержащаяся в экспериментальных данных, поскольку используются только знаки данных, но не их величины Это обстоятельство снижает *мощность* критерия, т е повышает вероятность признания нулевой гипотезы (разницы нет), при том, что справедлива конкурирующая гипотеза (разница есть) Среди более мощных непараметрических критериев, которые используют не только знак разности, но и величину разности рангов, относительно часто используется критерий Вилкоксона

7.4.2. Критерий Вилкоксона

Критерий Вилкоксона *U* (Вилкоксона—Манна—Уитни) — один из исторически первых критериев такого рода Позволяет сравнивать центральные тенденции в разных совокупностях на основе ранговых распределений при независимых выборках с разным числом наблюдений. Для этого критерия существенны не сами числовые значения наблюдений, а их ранговое расположение. *Например*

При оценке показателей частоты пульса у мужчин и женщин в конце рабочего дня были получены следующие результаты у мужчин — 72, 70, 74, 73, 80, 79, 76; у женщин — 75, 78, 71, 77 уд/мин

В примере (табл 95) представлен ранжированный ряд этих значений. Различия между двумя группами можно считать наиболее существенными, если после всех значений одного ряда будут располагаться значения второго ряда Однако в реальном

распределении существуют некоторые отклонения — *инверсии* Число инверсий подсчитывают через *U*. В приведенных распределениях перед значениями пульса у мужчин 72, 73, 74 стоит одна инверсия (71) Перед значением 76 стоят две инверсии (71 и 75) и т. д Общая сумма инверсий — 13. По таблице приложения (табл. 119) определяем, что для числа наблюдений 7 и 4 максимальное значение *U*, при котором различия еще достоверны, равно 3 при доверительной вероятности 0,05 При доверительной вероятности 0,01 значение *U* не должно превышать 1 Следовательно, признавать различия достоверными оснований нет

Таблица 95

Распределение показателей частоты пульса

Ранжированные значения частоты пульса		Инверсии
Мужчины	Женщины	
70		
	71	
72		1
73		1
74		1
	75	
76		2
	77	
	78	
79		4
80		4
		13

7.5. Оценка различия между несколькими средними. Дисперсионный анализ

Одним из способов определения различий, или наоборот схожести нескольких групп наблюдения, является **дисперсионный анализ** (См раздел «Дисперсия».)

Если этот анализ показывает, что не только дисперсии, но и средние в исследуемых группах тоже одинаковы, то можно считать эти группы статистически схожими по анализируемому признаку Уверенность в схожести позволяет в ходе обработки результатов объединять группы и тем самым получать значительно более полную и статистически более достоверную информацию

Более сложной задачей, решаемой с помощью дисперсионного анализа, является **факторный анализ** Такой анализ позволяет оценить существенность влияния некоторого качественного фактора на изучаемую величину, точнее на ее результирующие показатели В зависимости от глубины исследования оценивается влияние одного или нескольких факторов (многофакторный анализ)

Условия применения дисперсионного анализа

1 Нормальность распределения анализируемых групп или соответствие выборочных групп генеральным совокупностям с нормальным распределением

2 Независимость (не связанность) распрсделения наблюдений в группах

В основе дисперсионного анализа лежит анализ отклонений всех единиц исследуемой совокупности от среднего арифметического При этом основной идеей является сравнение отклонений, вызываемых воздействием факторного признака (фактора), с величиной отклонений, вызываемых случайными обстоятельствами Если отклонения, вызываемые факторным признаком, более существенны, чем случайные отклонения, то считается, что фактор оказывает существенное влияние на результирующий признак В качестве меры отклонений берется *дисперсия — средний квадрат отклонений* Чтобы ее вычислить, значения отклонений каждой варианты (каждого зарегистрированного числового значения) от среднего арифметического возводят в квадрат Тем самым избавляются от отрицательных знаков Затем эти разности суммируют и делят на число наблюдений, т. е усредняют отклонения Таким образом, получают значения дисперсий

При выполнении всех условий применения дисперсионного анализа, разложение общей дисперсии математически выглядит следующим образом

$$D_{общ}=D_{факт}+D_{ост},$$

где $D_{общ}$ — общая дисперсия, наблюдаемых значений (вариант), характеризуется разбросом вариант от общего среднего;

$D_{факт}$ — факторная (межгрупповая) дисперсия, характеризуется разбросом групповых средних от общего среднего,

$D_{ост}$ — остаточная (внутригрупповая) дисперсия, которая характеризует рассеяние вариант внутри групп Иногда ее называют дисперсией ошибки, поскольку она обычно не может быть предсказана или точно объяснена

При этом виде статистического анализа группы рассматриваются качественные группы вариант, т е группы, сформированные по качественному признаку (полу, профессии, диагнозам заболеваний и т. п) При непрерывно меняющихся значениях вариант группировка совокупности производится по интервалам значений факторов *Например* распределение обследованных по полу *мужчины и женщины* (дискретные группы), распределение обследованных по возрасту. 20—29 лет, 30—39 лет и т д (сгруппированный непрерывный ряд) Варианты, в свою очередь, могут быть дискретными (прерывными) или непрерывными

Не вдаваясь в детали математического обоснования методики вычисления различных данных в ходе дисперсионного анализа, можно выделить следующие последовательные этапы

1 Вычисление средних квадратов отклонений

2 Вычисление дисперсий

3 Сравнение факторной и остаточной дисперсий Оценка результатов с помощью теоретических значений распределения Фишера—Снедекора

Вычисление средних квадратов отклонений включает

1. Вычисление общей суммы квадратов отклонений наблюдаемых вариант от общего среднего:

$$S_{общ}=\sum_{i=1}^{p}P_i-\frac{\left[\sum_{i=1}^{p}R_i\right]^2}{pq},$$

где p — число повторностей, q — число групп факторного признака, $P_i = \sum_{i=1}^{q} V_{ij}^2$ — суммы квадратов наблюдаемых значений (вариант V) по группам, $R_i = \sum_{i=1}^{q} V_{ij}$ — суммы наблюдаемых значений по группам

2. Вычисление факторной суммы квадратов отклонений групповых средних от общего среднего, которая характеризует разброс между группами (межгрупповая сумма квадратов):

$$S_{факт} = \frac{\sum_{i=1}^{p} R_i^2}{q} - \frac{\left[\sum_{i=1}^{p} R_i\right]^2}{pq}$$

3. Остаточную (внутригрупповую) сумму квадратов отклонений вычисляют по формуле: $S_{ост} = S_{общ} - S_{факт}$

Примечание. В специальной литературе и пособиях можно встретить обозначение средних квадратов отклонений в виде символов *SS*, от английского *Sum of Squares* (сумма квадратов)

На основании значений сумм квадратов отклонений определяются несмещенные оценки факторной и остаточной дисперсии

$$D_{факт} = S_{факт}/(p-1) \text{ и } D_{ост} = S_{ост}/p(q-1),$$

где q — число групп факторного признака, p — число испытаний (повторностей) Для более подробного рассмотрения метода ограничимся относительно простым случаем однофакторного анализа, когда на результирующий признак воздействует один фактор Главным условием использования однофакторного дисперсионного анализа является повторность опытов (наблюдений).

Например: требуется методом дисперсионного анализа проверить нулевую гипотезу о равенстве средней продолжительности заболеваний хроническим бронхитом, повлекшим временную нетрудоспособность рабочих основных профессиональных групп станочники, слесари, прочие. Выборки сделаны на предприятии из 4 цехов с близкими условиями и характером труда работников, или с 4 повторностями (табл 96)

Таблица 96

Длительность заболевания хроническим бронхитом (исходные данные)

Номер цеха (повторности)	Факторный признак (профессия)		
	Уровни фактора		
	Станочники	Слесари	Прочие
1	10	5	4
2	14	11	6
3	9	7	11
4	12	8	4
В среднем M_{zpi}	11	8	6

Общее среднее M=8,42

Для упрощения расчетов вычтем из каждой варианты исходных данных число равное общему среднему (8,42) (табл 97) На конечный результат такое преобразование не окажет влияния, но значительно облегчит «ручные» вычисления

Таблица 97

Длительность заболевания хроническим бронхитом (исходные данные уменьшены на величину среднего арифметического =8,42)

Номер цеха (повторности)	Факторный признак (профессия)		
	Уроони фактора		
	Станочникн	Слесари	Прочие
1	1,6	−3,4	−4,4
2	5,6	2,6	−2,4
3	0,6	−1,4	2,6
4	3,6	−0,4	−4,4
R_i	11,4	−2,7	−8,7
R_i^2	128,4	7,1	75,1

$\sum R_i^2$=210,7

Все значения, находящиеся в клетках таблицы, возведем в квадрат.

Таблица 98

Длительность заболевания хроническим бронхитом (предыдущие данные возведены в квадрвт)

Номер цеха (повторности)	Факторный признак (профессия)		
	Уровни фактора		
	Станочники	Слесари	Прочие
1	2,5	11,7	19,5
2	31,2	6,7	5,8
3	0,3	2,0	6,7
4	12,8	0,2	19,5
$\sum P_i^2$	46,9	20,5	51,5

$\sum P^2 = 118{,}9$

Используя полученные в табл 98 промежуточные данные и учитывая, что число уровней фактора 3, число повторов 4, получаем:

1. Общую сумму квадратов отклонений $S_{общ}=118{,}9-0=118{,}9$

2 Факторную сумму квадратов $S_{факт}=210{,}7/4-0=52{,}7$

3 Находим остаточную сумму квадратов отклонений $S_{ост}=S_{общ}-S_{факт}=118{,}9-52{,}7=66{,}3$.

4 Находим факторную и остаточную дисперсии $D_{факт}=52{,}7/(3-1)=26{,}3$; $D_{ост}=66{,}3/3(4-1)=7{,}4$

5 Сравниваем факторную и остаточную дисперсии по критерию Фишера·

$$F=\frac{D_{факт}}{D_{ост}}=\frac{26{,}3}{7{,}4}=3{,}6$$

Учитывая, что число степеней свободы числителя равно 2 (число групп — 1), а знаменателя — 9 (число всех наблюдений — 1) и уровень значимости 0,05, с помощью встроенной функции *Microsoft Exel* или по таблице находим критическую точку $F = 4{,}3$ Так как полученные данные не превышают критического значения, нулевую гипотезу об отсутствии различий групповых средних (о равенстве средних) отвергать нет основания. Другими словами, группы, в целом, статистически не различаются По-

скольку значимых различий длительности заболевания в профессиональных группах не выявлено, говорить о статистически существенном влиянии факторного признака (профессионального фактора) на величину вариант (длительность заболевания хроническим бронхитом) также нет оснований

Если такой итог не устраивает исследователя, то можно попробовать сравнить отдельные пары, наиболее «ярко» различающихся групповых средних с помощью критерия Стьюдента При большом объеме наблюдений и большом количестве групп, на которые разделен факторный признак, этот прием может дать положительный результат.

При оценке результатов дисперсионного анализа необходимо помнить, что при его корректном использовании должно выполняться равенство дисперсий $D_{общ}=D_{факт}+D_{ост}$. В приведенном примере $D_{общ}=118,9/(12-1)=10,8$

Следует обратить внимание на весьма распространенную ошибку, которая часто встречается при оценке полученных в результате дисперсионного анализа результатов *Например* соотношение вычисленного значения $D_{общ}=10,8$ и суммы ранее вычисленных дисперсий вместо равенства, казалось бы, образует неравенство $10,8\neq26,3+7,4$ Можно предположить, что причиной такого несоответствия фактических данных и теоретических предпосылок является нарушение одного или сразу обоих условий применения дисперсионного анализа: *независимость групповых распределений и их соответствие нормальному распределению* Однако выявленное несоответствие объясняется довольно просто: при суммировании остаточной и факторной дисперсий не были учтены их весовые значения (весовые значения дисперсий, как известно, определяются числом наблюдений, на основании которых они получаются).

Таким образом, результаты приведенного в качестве примера анализа следует признать корректными.

В практике медико-биологических исследований нередко встречается необходимость анализировать большее число взаимодействующих факторов. Решение таких задач вызывает резкое увеличение громоздкости дисперсионного анализа *Например* если при однофакторном комплексе взаимоотношение диспер-

сий устанавливается из равенства $D_{общ}=D_{ост}+D_{факт}$, то при двухфакторном комплексе (с повторениями) используется равенство $D_{общ}=D_{ост}+D_A+D_B+D_{AB}$, где D_A — дисперсия фактора A, D_B — дисперсия фактора B, D_{AB} — дисперсия эффекта взаимодействия. При трехфакторном — величина общей дисперсии равна $D_{общ}=D_{ост}+D_A+D_B+D_C+D_{AB}+D_{BC}+D_{AC}$, где D_{AB}, D_{BC}, D_{AC} — дисперсии эффектов соответствующих взаимодействий.

Другой причиной, которая затрудняет проведение многофакторного дисперсионного анализа, является необходимость располагать исходные данные в виде комбинационных таблиц, которые требуют большого числа наблюдений. Выполнить это требование в условиях обычного медико-биологического исследования крайне трудно. Поэтому обычно используются только однофакторные, максимум двухфакторные дисперсионные комплексы. Многофакторные комплексы можно обрабатывать путем последовательного анализа выделяемых из всей анализируемой совокупности одно- или двухфакторных комплексов

7.6. Дисперсионный анализ в *MS Excel*

В программе *MS Excel* для статистического анализа данных имеется надстройка **«Пакет анализа»**, которая позволяет проводить дисперсионный анализ следующих видов:

- однофакторный,
- двухфакторный без повторений,
- двухфакторный с повторениями

7.6.1. Однофакторный дисперсионный анализ

Условием применения такого анализа является повторность данных.

Например, необходимо оценить влияние условий и характера труда (в данном случае профиля цеха) на заболеваемость рабо-

чих острым и хроническим гастритом. Повторность данных обеспечена наблюдением за работниками предприятий, схожими по профилю Исходные данные представлены в таблице (рис. 109)

Согласно таблице, исследуемый фактор имеет 3 уровня (доменный, мартеновский и прокатный цеха) В каждой группе имеется только по 4 наблюдения (повторности).

	A	B	C	D
1	Заболеваемость гастритом на 100 рабочих			
2	Заводы	Цех (фактор)		
3	(повторность)	Доменный	Мартен	Прокатный
4	1	23,4	26,4	43,1
5	2	13,8	45,5	48,5
6	3	26,9	35,6	21,9
7	4	21,5	29,6	38,1

Рис. 109. **Исходные данные однофакторного комплекса**

Для решения задачи в *MS Excel*:

1. Сформируйте таблицу с исходными данными (см. рис. 109).
2. Выберите <**Пакет анализа**> из меню <**Сервис**>.
3. В соответствии с условиями задачи выберите в появившемся диалоговом окне метод «**Однофакторный дисперсионный анализ**» и нажмите кнопку [OK].
4. В окне «**Однофакторный дисперсионный анализ**» установите для входных данных следующие параметры.
   - входной интервал ($B$4:$D$7),
   - метки (выбранный нами входной диапазон не содержит метки, т. е названий строк и столбцов),
   - альфа (уровень значимости =0,05).
5. Для параметров вывода установите переключатель в положение «**Выходной интервал**» и укажите клетку с координатой (E1).
6. После завершения настройки параметров нажмите кнопку [OK]

Диалоговое окно с заполненными исходными параметрами должно выглядеть следующим образом (рис. 110):

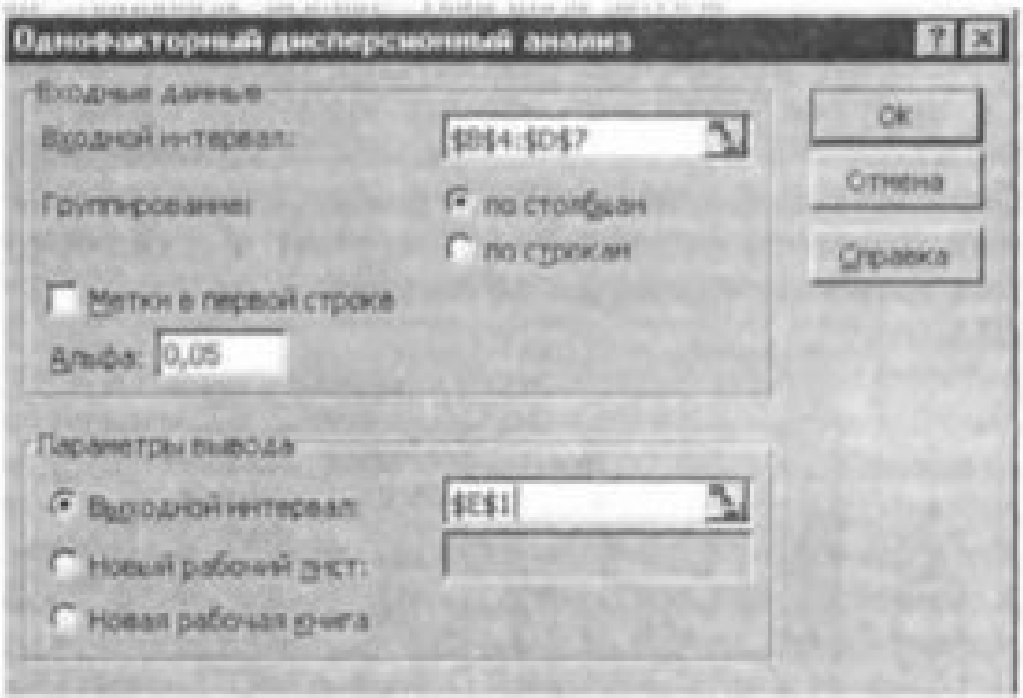


Рис. 110. **Окно исходных параметров однофакторного дисперсионного комплекса**

Результаты дисперсионного анализа будут состоять из двух таблиц В первой таблице для каждого столбца исходной таблицы, в которых располагаются анализируемые группы, приведены числовые параметры количество чисел (счет), суммы по столбцам, средние дисперсии по столбцам (рис 111)

Во второй части результатов *MS Excel* использует следующие обозначения:

- *SS* — сумма квадратов,
- *df* — степени свободы,
- *MS* — средний квадрат (дисперсия),
- *F* — F-статистика Фишера (фактическое значение),
- *P-значение* — значимость критерия Фишера,
- *F критическое* — критическое значение F-статистики при *P* =0,05.

Таким образом, сумма квадратов, обусловленная влиянием исследуемого фактора (межгрупповая сумма), равна 601,5. Остаточная сумма квадратов (внутригрупповая) равна 699,09 Соответствующие дисперсии: межгрупповая (для исследуемого фактора) — 300,8; остаточная (внутригрупповая) — 77,7

	B	F	G	H	I	J	K
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	Столбец 1	4	85,8	21,4	30,8733		
6	Столбец 2	4	137,1	34,275	70,9425		
7	Столбец 3	4	151,8	37,9	131,813		
8							
9							
10	Дисперсионный анализ						
11	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
12	Между группами	601,542	2	300,771	3,8721	0,0611929	4,25649205
13	Внутри групп	699,088	9	77,6764			
14							
15	Итого	1300,63	11				

Рис. 111 **Результаты анализа однофакторного комплекса**

Основной вывод из полученных результатов заключается в следующем

Есть основания не отвергать нулевую гипотезу об отсутствии влияния рассмотренного фактора (условия и характер работы в разных цехах) на заболеваемость рабочих гастритом: поскольку не выполняется неравенство $F>F_{кр}$, величина значимости $P=0,06$ Для отрицания нулевой гипотезы она должна быть не более 0,05

7.6.2. Двухфакторный анализ с неповторяющимися данными

Для удобства работы возьмем уже использованные в предыдущем примере данные Однако рассмотрим их несколько иначе Будем считать, что мы оцениваем влияние на заболеваемость рабочих гастритом уже двух факторов условий труда на рабочих местах (цех) и заводских условий (№ предприятия)

Согласно таблице (рис 112), первый фактор имеет три уровня (три столбца): доменный, мартеновский и прокатный цеха Второй фактор имеет четыре уровня (четыре строки) завод 1, завод 2; завод 3 и завод 4

	A	B	C	D
1	Заболеваемость гастритом на 100 рабочих			
2	Заводы	Цех (фактор 1)		
3	(фактор 2)	Доменный	Мартен.	Прокатный
4	Завод 1	23,4	26,4	43,1
5	Завод 2	13,8	45,5	48,5
6	Завод 3	26,9	35,6	21,9
7	Завод 4	21,5	29,6	38,1

Рис. 112. **Исходные данные двухфакторного комплекса**

Для решения задачи в *MS Excel*

1 Сформируйте таблицу с исходными данными (см рис 112)

2 Выполните команду <Пакет анализа> из меню <Сервис>

3. В соответствии с условиями задачи выберите в появившемся диалоговом окне метод «Двухфакторный дисперсионный анализ без повторений» и нажмите кнопку [ОК]

4 В окне «Двухфакторный дисперсионный анализ без повторений» установите для входных данных следующие параметры.

- входной интервал ($B$4:$D$7),
- метки (выбранный нами входной диапазон не содержит метки, т. е названий строк и столбцов),
- альфа (уровень значимости =0,05)

5 Для параметров вывода установите переключатель в положение «Выходной интервал» и укажите клетку с координатой (E1)

6. После завершения настройки параметров нажмите кнопку [ОК].

Диалоговое окно с заполненными исходными параметрами должно выглядеть следующим образом (рис 113).

Результаты дисперсионного анализа будут состоять из двух частей. В первой части для каждой строки и каждого столбца исходной таблицы приведены числовые параметры количество чисел (счет), суммы по столбцам и строкам Средние и дисперсии по столбцам и строкам выводятся во второй части таблицы результатов (рис. 114)

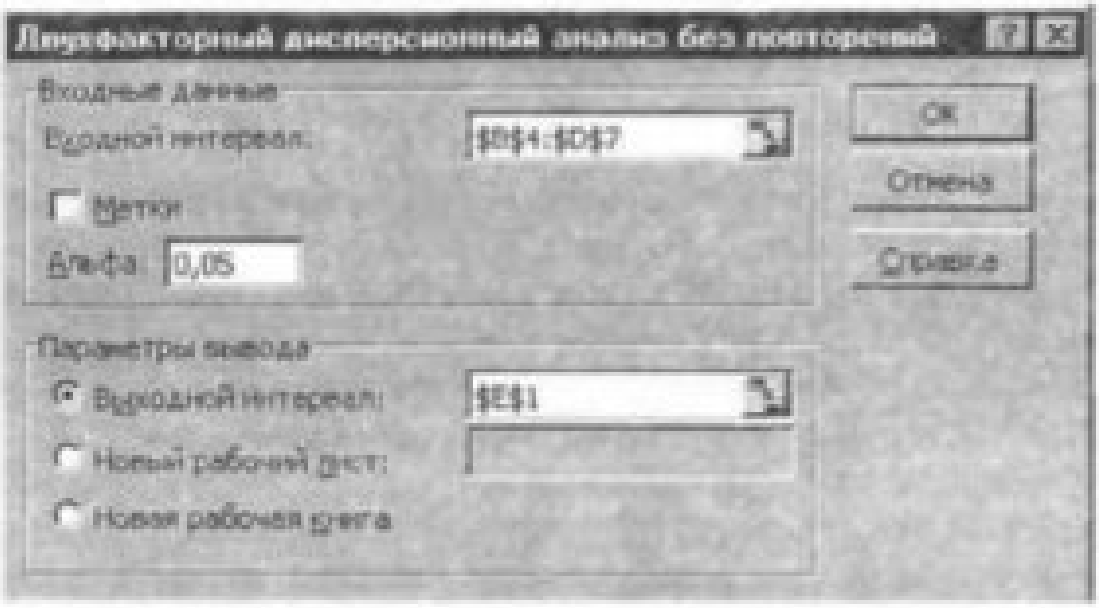


Рис. 113. **Окно исходных параметров двухфакторного дисперсионного комплекса без повторений**

	E	F	G	H	I	J	K
1	Двухфакторный дисперсионный анализ без повторений						
2							
3	Итоги	Счет	Сумма	Среднее	Дисперсия		
4	Строка 1	3	92,5	30,56667	112,963333		
5	Строка 2	3	107,8	35,93333	369,663333		
6	Строка 3	3	84,4	28,13333	48,0633333		
7	Строка 4	3	89,2	29,73333	68,9033333		
8							
9	Столбец 1	4	96,8	21,4	30,6733333		
10	Столбец 2	4	137,1	34,275	70,5425		
11	Столбец 3	4	151,6	37,9	131,013333		
12							
13							
14	Дисперсионный анализ						
15	Источник вариации	SS	df	MS	F	P-Значение	F критическое
16	Строки	100,0426	3	34,01417	0,341835161	0,796536819	4,757063
17	Столбцы	601,5417	2	300,7708	3,022594461	0,123599406	5,143249
18	Погрешность	597,046	6	99,5075			
19							
20	Итого	1300,629	11				

Рис. 114. **Результаты анализа двухфакторного комплекса без повторений**

Во второй части результатов *MS Excel* использует следующие обозначения:

SS — сумма квадратов,

df — степени свободы,

MS — средний квадрат (дисперсия),

F — F-статистика Фишера (фактическое значение),

P-значение — значимость критерия Фишера,

F критическое — критическое значение F-статистики при *P*=0,05.

Сумма квадратов, обусловленная влиянием первого фактора (столбцы — наименование цеха), равна 601,5 Сумма квадратов, обусловленная влиянием второго фактора (строки — номер завода), равна 601,5, а остаточная, внутригрупповая сумма квадратов равна 102,04 Остаточная, внутригрупповая дисперсия, которая в таблице результатов названа «погрешность», равна 597,05

Из полученных результатов можно сделать следующий **вывод**. Есть основания не отвергать нулевую гипотезу об отсутствии влияния двух рассмотренных факторов (условий труда на рабочих местах и заводских условий) на заболеваемость рабочих гастритом. так как не выполняется неравенство *F*-статистики и критических значений *F* (*F>Fкр*), величина значимости *P* превышает в обоих случаях 0,05.

7.6.3. Двухфакторный анализ с повторяющимися данными

Двухфакторный анализ с повторяющимися данными позволяет учесть как влияние отдельных факторов, так и их совместное действие на результативный признак Рассмотрим пример применения двухфакторного дисперсионного анализа с повторениями

Исследуем влияние на заболеваемость рабочих гастритом уже двух факторов: условий труда на рабочих местах (цех) — фактор 1, и влияние заводских условий (№ предприятия) — фактор 2 В качестве повторностей наблюдения будем рассматривать повторения наблюдений в течение трех лет

Исходный статистический комплекс предс гавлен в *Excel*-таблице (рис. 115)

	A	B	C	D	E
1	Заболеваемость гастритом (на 100 рабочих)				
2	Завод	Год	Цех (фактор 1)		
3	(фактор 2)	(повторности)	Доменный	Мартен	Прокатный
4	Завод1	1998	23,1	25,4	43,1
5		1999	22,0	26,5	47,0
6		2000	25,0	27,2	39,2
7	Завод2	1998	11,1	45,5	52,2
8		1999	18,3	48,7	54,6
9		2000	12,0	42,2	38,7
10	Завод3	1998	24,5	38,3	18,5
11		1999	28,9	35,4	22,1
12		2000	27,4	33,1	25,2
13	Завод4	1998	18,9	33,5	38,8
14		1999	22,2	25,5	41,1
15		2000	23,3	29,9	34,5

Рис. 115. **Исходная таблица для двухфакторного дисперсионного комплекса с повторениями**

Для решения задачи в *MS Excel*

1 Выполните команду <Пакет анализа> из меню <Сервис>

2. В появившемся диалоговом окне нужно выбрать метод «Двухфакторный дисперсионный анализ с повторениями» и нажать кнопку [OK].

3. В окне «Двухфакторный дисперсионный анализ с повторениями» установите для входных данных следующие параметры

- входной интервал ($B$3 $E$15),

Примечание: Обязательно в этот интервал должны быть включены названия факторов в виде одного столбца слева и одной строки сверху.

- число строк для выборки — 3,
- метки (входной диапазон не содержит метки),
- альфа (уровень значимости =0,05).

4 Для параметров вывода установите переключатель в положение «Новый рабочий лист» Выходной интервал можно задать в виде одной координаты $F$1

Заполненные исходные параметры в диалоговом окне должны выглядеть следующим образом (рис. 116)

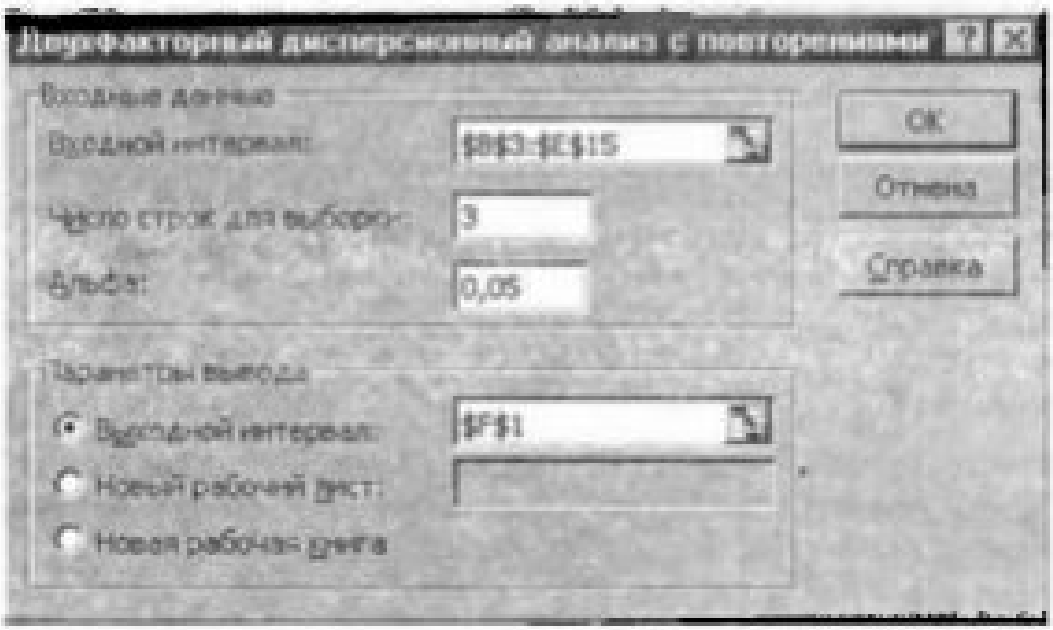


Рис. 116. **Окно исходных параметров двухфакторного дисперсионного анализа с повторениями**

5 После завершения настройки параметров нажмите кнопку **[ОК]**.

Результаты дисперсионного анализа будут состоять из нескольких частей В первых частях для каждой строки и каждого столбца исходной таблицы приведены числовые параметры количество чисел, сумма, среднее и дисперсия по второму фактору (в этом примере — по № завода) и итогам Вторая часть результатов представлена на рис. 117

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	303,76528	3	101,25509	7,010642	0,00150731	3,008786109
Столбцы	1809,655	2	904,3275	62,61331	2,9949E-10	3,402831794
Взаимодействие	1769,4339	6	294,23899	20,6493	2,194E-08	2,508187436
Внутри	346,63333	24	14,443056			

Рис. 117. **Результаты двухфакторного дисперсионного анализа с повторениями**

Основной вывод из полученных результатов заключается в следующем Конкурирующая гипотеза может быть принята для

фактора 1 и для фактора 2, а также их взаимного влияния, т. к во всех случаях выполняется неравенство $F > F_{кр}$ Принятие конкурирующей гипотезы подтверждается и величинами значимостей, удовлетворяющими условию $P<0,05$

Примечание. Уровни значимости P в строках итоговой таблицы «*Столбцы*» и «*Взаимодействие*» приведены в компактной, экспоненциальной форме *Например* число 2,194E-08 выглядит в обычной записи как 0,00000002194 E-08 обозначает, что первая значащая цифра стоит на восьмом месте после запятой

Таким образом, можно считать доказанным, с высокой степенью статистической достоверности, влияние условий и характера труда на показатели заболеваемости хроническим гастритом

7.7. Оценка различий коэффициентов корреляции

Различия коэффициентов корреляции, используемых при анализе количественных признаков, определяются исходя из общих принципов статистического анализа различий В качестве нулевой гипотезы рассматривается утверждение о том, что статистически значимых различий между коэффициентами корреляции нет (H_0: $r_1 = r_2$). Если исходные выборки, которые использовались для вычислений этих коэффициентов, состоят из однородных величин, то в случае принятия нулевой гипотезы, можно предполагать, что обе выборки принадлежали к одной генеральной совокупности. *Например.* требуется оценить статистическую достоверность различий следующих коэффициентов r_1=0,45 и r_2=0,58 Числа наблюдений в первой и второй группах составили, соответственно, N_1=74 и N_2=50

Для ответа на этот вопрос требуется выполнить следующие операции:

1 По таблице значений Z (Приложения, стр 394) значения коэффициентов корреляции переводятся в соответствующие им величины Z_1=0,48 и Z_2=0,66 (Методика перевода на стр 280)

2. Вычисляется критерий Стьюдента по формуле

$$t=|Z_2-Z_1|/\sqrt{(N_1+N_2)/[(N_1-3)(N_2-3)]},$$

$$t=|0,66-0,48|/\sqrt{(74+50)/[(74-3)(50-3)]}=1,09$$

3. Число степеней свободы = $N_1+N_2-4=74+50-4=120$

4. При уровне значимости P=0,05 критическая величина критерия Стьюдента составляет 1,98 (табл. 116), что намного больше вычисленного Таким образом, оснований отвергнуть нулевую гипотезу — нет Различия коэффициентов корреляции следует признать статистически не достоверными

Вообще, при анализе различий коэффициентов корреляции действует общее правило чем больше коэффициенты, тем меньшие различия становятся значимыми *Например*, разница 0,1 между коэффициентами корреляции, которые соответственно равны 0,14 и 0,24 может быть статистически не значимой В той же выборке разность 0,1 может оказаться значимой для коэффициентов 0,80 и 0,90.

7.8. Оценка достоверности различий коэффициентов вариации

Необходимость оценки достоверности различий коэффициентов вариации чаще всего возникает в двух случаях

1. При сравнительной оценке изменений, произошедших в ряду разнородных показателей под воздействием какого-либо фактора или до наступления какого-либо события и после наступления этого события (до опыта и после, в разные временны́е промежутки и т п) *Например* необходимо получить ответ, какие показатели развития ребенка больше всего изменяются под воздействием неблагоприятных факторов городской среды?

2 При сравнительной оценке разнородных показателей (учетных признаков) одной и той же совокупности, одного и того же явления, в одно и то же время. *Например* сравнительная оценка однородности коллектива по стажу работы, уровню образования, и т п

В первом случае оценка различий коэффициентов вариации часто заменяется оценкой различий средних или тенденций распределения в рядах однородных величин В принципе, это может приводить к потере достаточно ценной информации о сравнительной вариабельности разнородных показателей под воздействием какого-либо фактора, т е. косвенной характеристике чувствительности тех или иных показателей

Определение статистической достоверности различий производится на основании следующего алгебраического выражения неравенства

$$|C_1 - C_2| / \sqrt{m_{C_1}^2 + m_{C_2}^2} > 3 + 6/(N-4),$$

где $|C_1 - C_2|$ — абсолютное значение разности коэффициентов вариации;

N — число наблюдений в меньшей из сравниваемых выборок;

m_{C_1} и m_{C_2} — ошибки коэффициентов вариации

Если неравенство выполняется, т е левая часть больше правой, то различия считаются статистически достоверными

Ошибки коэффициентов вариации вычисляются по формуле

$$m_{C_v} = \frac{C_v}{\sqrt{n}} \sqrt{\frac{1}{2} + \left(\frac{C_v}{100}\right)^2}$$

Пример: при изучении влияния повышенной запыленности воздуха рабочей зоны на органы дыхания работников одного из предприятий установлено, что изменение показателей частоты заболеваний органов дыхания рабочих (112 человек) в зависимости от стажа работы оценивалось коэффициентом вариации C_1=24%, а изменение показателей функции внешнего дыхания (у 77 обследованных рабочих) C_2=31%

$$m_1 = \frac{24}{\sqrt{112}}\sqrt{\frac{1}{2}+\left(\frac{24}{100}\right)^2} = 1{,}69\%; \quad m_2 = \frac{31}{\sqrt{77}}\sqrt{\frac{1}{2}+\left(\frac{31}{100}\right)^2} = 2{,}73\%;$$

$$31-24/\sqrt{1{,}69^2+2{,}73^2} > 3+6/(77-4), \text{ или } 2{,}18 > 3{,}06$$

Поскольку это неравенство не выполняется, можно считать различия разброса показателей заболеваемости органов дыхания и функции внешнего дыхания не существенными.

8. Динамические (временны́е) ряды

Динамический ряд — ряд однородных статистических величин, показывающий изменение какого-либо явления во времени. С помощью статистического анализа динамических рядов решаются следующие задачи·

- выявление и описание характерных тенденций изменения явления во времени,
- подбор статистической модели, описывающей эти изменения;
- отыскание отсутствующих промежуточных значений (**интерполяция**) на основе имеющихся показателей;
- предсказание на основе имеющихся результатов будущих значений (**экстраполяция**) анализируемого ряда.

Различают следующие **виды динамических рядов**.

Простой — ряд, составленный из абсолютных величин, характеризующих динамику одного явления

Сложный — динамический ряд, отражающий изменение во времени параллельно нескольких явлений.

Производный — ряд, составленный из средних или относительных величин.

Моментный — динамический ряд, состоящий из величин, характеризующих явление на какой-либо определенный момент времени *Например* численность населения на конец года

Интервальный — ряд, характеризующий изменение явления в течение какого-либо периода (интервала) *Например* число рождений, заболеваний за год, месяц, квартал и т. п

8.1. Показатели динамического ряда. Вычисление основных показателей динамического ряда

На первом этапе статистической обработки динамических рядов анализируются основные тенденции **(тренд)** изменения явления во времени Для этого, во-первых, используются графические изображения, которые часто дают самую исчерпывающую информацию. Во-вторых, вычисляется комплекс специальных показателей, позволяющих дать количественную оценку динамики анализируемого явления При этом, если полученные показатели дают достаточно ясную и наглядную картину тенденций, то на этом этапе нередко и заканчивается весь анализ динамического ряда

Абсолютный прирост или убыль характеризует изменение явления в единицу времени (за интервал времени) Получается путем вычитания из данных последующего периода данных предыдущего Если ряд возрастает, то прирост положителен Если убывает — отрицателен (убыль). Этот показатель не может использоваться при сравнении динамики разнородных данных (вес в кг, рост в см) Кроме того, на его значение оказывает влияние и абсолютный размер анализируемой характеристики *Например* рост в см — трехзначное число, окружность бедра в см — двухзначное

Темп роста или снижения показывает соотношение в процентах последующего уровня и предыдущего, поэтому может использоваться при сравнительном анализе динамики разнородных

величин. Получается путем деления последующего уровня на предыдущий и умножения на 100 Если прирост положителен, то показатель больше 100%, если отрицателен — меньше 100%

Темп прироста показывает, на сколько процентов увеличился или уменьшился уровень явления. По существу отражает относительную скорость изменения явления от одного отрезка времени к другому. Вычисляется путем деления абсолютного прироста на предыдущий уровень, либо вычитанием из показателя темпа роста 100 Если прирост положителен — показатель больше 0. Если отрицателен — меньше

Абсолютное значение 1% прироста характеризует значение (стоимость) 1% прироста изучаемого явления Этот показатель может вычисляться делением абсолютного прироста на темп прироста или делением показателя предыдущего уровня на 100 Является одним из самых существенных, поскольку «стоимость» одного процента темпа роста и прироста в различных совокупностях разная *Например число районов города «N» с высоким уровнем загрязнения атмосферного воздуха в 1995 году было 4, в 1996 стало — 8. Темп роста — 200% В городе «NN» таких районов в 1995 году было 10, стало — 15 Темп роста — 50% Однако в первом случае число неблагополучных районов увеличилось на 4, а во втором — на 5* Даже в одном динамическом ряду значение одного процента роста и темпа прироста может существенно различаться на разных отрезках времени.

Показатели наглядности характеризуют динамику явления в процентах относительно исходного уровня Исходный уровень принимается за 100% В отличие от предыдущих показателей на всем протяжении временно́го ряда «стоимость» одного процента этого показателя остается неизменной Однако динамика изменения исходных данных от одного промежутка времени к другому становится менее выразительной

Существуют различные варианты вычисления показателей динамики Эти способы отличаются набором исходных данных и трудоемкостью вычислений. *Например* (табл 99)

Таблица 99

Динамика случаев заболеваний с временной утратой трудоспособности (ЗВУТ)

Год	Уровень ЗВУТ	Абсолютный прирост	Темп роста	Темп прироста	1% прироста	Показатели наглядности
	У	А	Т	Р	П	Н
1985	65,8					100,0
1986	90,2	24,4	137,1	37,1	0,7	137,1
1987	67,4	−22,8	74,7	−25,3	0,9	102,4
1988	94,3	26,9	139,9	39,9	0,7	143,3
1989	55,4	38,9	58,7	−41,3	0,9	84,2
1990	45,1	10,3	81,4	−18,6	0,6	68,5
1991	48,2	3,1	106,9	6,9	0,5	73,3

Абсолютный прирост или снижение заболеваемости в 1986 и 1987 годах.

$$А_{1986}=У_{1986}-У_{1985}=90,2-65,8=24,4;$$

$$А_{1987}=У_{1987}-У_{1986}=67,4-90,2=-22,8$$

Темп роста или снижения заболеваемости в 1986 и 1987 годах

$$Т_{1986}=(У_{1986}/У_{1985})\times 100=90,2/65,8\times 100=137,1,$$

$$Т_{1987}=(У_{1987}/У_{1986})\times 100=67,4/90,2\times 100=74,7$$

Темп прироста заболеваемости в 1986 и 1987 годах

1-й способ расчета:

$$Р_{1986}=(А_{1986}/У_{1985})\times 100=24,4/65,8\times 100=37,1,$$

$$Р_{1987}=(А_{1987}/У_{1986})\times 100=-22,8/90,2\times 100=-25,3.$$

2-й способ расчета.

$$Р_{1986}=Т_{1986}-100=137,1-100=37,1$$

$$Р_{1987}=Т_{1987}-100=74,7-100=-25,3$$

Абсолютное значение 1% прироста заболеваемости в 1986 и 1987 годах

1-й способ расчета:

$$П_{1986}=У_{1985}/100=65,8/100\ П1986\approx 0,66\approx 0,7;$$

$$П_{1987}=У_{1986}/100=90,2/100\approx 0,90\approx 0,9.$$

2-й способ расчета

$$П_{1986}=А_{1986}/Р_{1986}=24,4/37,1\approx0,7;$$

$$П_{1987}=А_{1987}/Р_{1986}=-22,8/-25,3\approx0,9$$

Показатели наглядности прироста заболеваемости в 1986 и 1987 годах по сравнению с 1985 годом

$$Н_{1986}=(У_{1986}/У_{1985})\times100=90,2/65,8\times100=137,1,$$

$$Н_{1987}=(У_{1987}/Н_{1985})\times100=67,4/65,8\times100=102,4$$

Вычисление средних: в моментном ряду с равными промежутками между датами = $(1/2\ У_{1985}+У_{1986}+У_{1987}+.\ \ +1/2\ У_{1991})/n$, где n — число анализируемых наблюдений

Средний уровень в моментном ряду с неравными промежутками между датами = $(1/2\ У_{1985}\times t_{1985}+У_{1986}\times t_{1986}+.\ \ +1/2\ У_{1991}\times\times\times t_{1991})/(t_{1985}+t_{1986}+..\ ....+t_{1991})$, где t — число дней в году.

Средний уровень в интервальном ряду = $(У_{1985}+У_{1986}+\ У_{1987}+.....+У_{1991})/n$

Средний абсолютный прирост = $(А_{1985}+А_{1986}+А_{1987}+\ \ +А_{1991})/n$

Средний темп роста (среднее хронологическое) вычисляется в виде среднего геометрического = $n\times(Р_{1985}\times Р_{1986}\times Р_{1987}\times\ \ \times Р_{1991})$

Средний темп прироста = средний темп роста — 100

Изменение уровня исходных показателей может происходить с различной интенсивностью Несмотря на кажущуюся однонаправленность всех используемых показателей динамического ряда характер их изменения может принимать самые разнообразные формы *Например*, абсолютные приросты могут быть стабильными, а темпы роста (прироста) при этом увеличиваться или уменьшаться.

8.2. Углубленный анализ динамических рядов

Углубленный статистический анализ временны́х (динамических) рядов направлен на выявление следующих составляющих:

- систематической (регулярной), которая может включать одну или несколько компонент;
- случайной (шум, ошибку), которая затрудняет обнаружение регулярных тенденций Выделение (фильтрация) шума, позволяет увидеть систематическую составляющую более отчетливо

Большинство систематических составляющих временны́х рядов задаются либо трендом, либо сезонной составляющей

Сезонная составляющая (сезонность) — это периодически повторяющаяся компонента *Например* постоянный «подскок» инфекционной заболеваемости в определенные месяцы года

Тренд представляет собой общую изменяющуюся во времени линейную или нелинейную составляющую *Например.* постоянный рост числа дорожно-транспортных происшествий за последнее десятилетие

Тренд и сезонность могут присутствовать одновременно *Например*: детский травматизм в результате дорожно-транспортных происшествий (ДТП) возрастает из года в год, но он также содержит сезонную составляющую (как правило, детский травматизм из-за ДТП всегда меньше в летний период)

8.2.1. Показатели сезонности

При изучении динамики явлений большое значение имеет учет так называемой сезонной компоненты динамического ряда. Сезонная компонента отражает циклическую повторяемость каких-либо процессов в течение определенного периода времени (года, месяца и т п) Учет сезонных явлений позволяет, с одной стороны, выявить и оценить влияние сезонных циклов на уровни показателей, с другой — исключить их влияние на более тонкие процессы относительного подъема и спада (циклические компоненты), происходящие в более длительные периоды времени.

Одной из важных методических особенностей анализа сезонных изменений является необходимость сравнения значений

показателей через определенные промежутки времени Так, при изучении динамики сезонной заболеваемости по годам необходимо сравнивать показатели в соответствующих интервалах времени (декабрь одного года с декабрем другого, январь — с январем и т. д., а не с данными других месяцев) Кроме того, специальные показатели сезонности, вычисляемые при анализе сезонных изменений, позволяют исключить влияние на уровни показателей разного числа дней в месяцах (январь — 31, февраль — 28 дней и т п) *Например*, для определения сезонности заболеваемости следует пользоваться отношением среднедневного числа заболеваний в каждом месяце к среднедневному числу заболеваний за год. Это соотношение целесообразно выражать в процентах (табл 100).

Таблица 100

Распределение острых кишечных инфекций (ОКИ) за год

Месяц года	Дней в месяце	Всего ОКИ	В среднем за 1 день	Сезонность	
				Абс	%
январь	31	26	0,84	0,939	93,9
февраль	28	24	0,86	0,960	96,0
март	31	25	0,81	0,903	90,3
апрель	30	21	0,70	0,784	78,4
май	31	25	0,81	0,903	90,3
июнь	30	31	1,03	1,157	115,7
июль	31	29	0,94	1,047	104,7
август	31	34	1,10	1,228	122,8
сентябрь	30	33	1,10	1,232	123,2
октябрь	31	28	0,90	1,011	101,1
ноябрь	30	25	0,83	0,933	93,3
декабрь	31	25	0,81	0,903	90,3
Итого	365	326	0,89	1,000	100,0

Вычисление показателей сезонности

1. Определяется среднедневное месячное число заболеваний

в январе	26/31=0,84
в феврале	24/28=0,86
в марте	25/31=0,81 и т д

2 Определяется среднедневное число заболеваний за год

326/365=0,89

3 Вычисляются показатели сезонности в абсолютных числах

в январе	0,84/0,89=0,939
в феврале	0,86/0,89=0,96
в марте	0,81/0,89=0,903 и т д

4 Вычисляются показатели сезонности в процентах

в январе	0,939100=93,9%
в феврале	0,96100=96,0%
в марте	0,903100=90,3% и т д

8.2.2. Вычисление показателей сезонности в *MS Excel*

В качестве примера создадим в *Excel* таблицу «Распределение острых кишечных инфекций (ОКИ) за год».

	A	B	C	D	E	F
1	Месяц	Дней в	Всего	В среднем	Сезонность	Сезонность
2	года	месяце	ОКИ	за день	Абс	%
3	январь	31	26			
4	февраль	28	24			
5	март	31	25			
6	апрель	30	21			
7	май	31	25			
8	июнь	30	31			
9	июль	31	29			
10	август	31	34			
11	сетябрь	30	33			
12	октябрь	31	28			
13	ноябрь	30	25			
14	декабрь	31	25			
15	Итого	365				

Рис 118 Пример таблицы в *Excel*

1 После запуска *Excel* сформируйте таблицу, содержащую названия колонок и исходные данные (рис 118)

2 Для изменения ширины столбцов выделите все столбцы и воспользуйтесь командой <**Столбец/Автоподбор ширины**> из меню <**Формат**> Эта команда установит оптимальную ширину для каждого столбца в зависимости от размера содержимого ячеек

3. Для определения общего числа заболеваний за год введите в ячейку C15 формулу =**СУММ**(C3 C14). Для определения среднедневного месячного числа заболеваний введите в ячейку D3 формулу =C3/B3

4. Заполнение клеток D4 D15 выполните с помощью процедуры копирования формул. Для этого установите указатель ячейки в ячейку D3, выберите команду <**Копировать**> из меню <**Правка**>, передвиньте указатель ячейки в ячейку D4 и маркируйте ячейки D4:D15, нажмите клавишу **[Enter]**

5 Заполните графу «Показатель сезонности в абсолютных числах» Для этого введите в ячейку E3 формулу =D3/D$15 Чтобы распространить действие введенной формулы на весь столбец установите указатель ячейки в ячейку E3, выберите команду <**Копировать**> из меню <**Правка**>, передвиньте указатель ячейки в ячейку E4 и маркируйте ячейки E4 E15, нажмите клавишу **[Enter]**.

6 Для заполнения графы «Показатель сезонности в процентах» в ячейку F3 введите формулу =E3*100 Скопируйте формулу в ячейки F4 F15.

7 Для задания в столбцах таблицы нового числового формата выделите соответствующий блок ячеек, выполните команду <**Ячейки**> из меню <**Формат**>, в панели **Число** выберите из списка «Числовые форматы» категорию «Числовой» и установите необходимое число десятичных знаков, активизируйте кнопку **[OK]**.

8.2.3. Повышение наглядности тенденций динамических рядов. Прогноз динамики

Анализ динамических рядов может строиться на относительных величинах, получаемых на этапе сводки и группировки первичного материала статистического исследования Вместе с тем, для углубленного анализа временны́х рядов используются более сложные методики математической статистики В первую очередь, применение таких методик связывается с необходимостью анализа неясных тенденций и прогнозирования динамики изучаемого явления

Если динамические ряды содержат значительную случайную ошибку (шум), то можно применить один из двух наиболее простых приемов **сглаживания** или **выравнивания** динамических рядов·

- *укрупнение интервалов* путем суммирования исходных уровней по нескольким интервалам *Например.* суммируются числа рождений за 1982, 1983 и 1984 годы (84+94+92=270), затем за 1985, 1986 и 1987 годы и т д (табл 101),
- *вычисление групповых средних,* которые определяются на основе данных по укрупненным интервалам (270/3=90, 262/3=87,3 и т. д);

Таблица 101

Сглаживание динамического ряда укрупнением интервалов и скользящим средним

Учетный год	Число рождений	Суммы по 3 годам	Средние по 3 годам	Скользящие средние
1982	84			
1983	94	270	90,0	90,0
1984	92			89,7
1985	83			88,7
1986	91	262	87,3	87,3
1987	88			87,0
1988	82			86,7
1989	90	249	83,0	83,0
1990	77			82,3
1991	80			
1992	90	248	82,7	82,3
1993	78			82,7

Укрупнение интервалов или расчет группового среднего внутри этих интервалов позволяет относительно легко повысить наглядность ряда, особенно если большинство «шумовых» составляющих находятся именно внутри этих интервалов. Но в случае если шум не согласуется с этой периодичностью, распределение уровней показателей становится грубым, что ограничивает возможности детального анализа изменения явления во времени Более точные характеристики получаются, когда используются **скользящие средние.** Этот метод — один из самых широко применяемых методов сглаживания показателей временно́го ряда

Он основан на переходе от начальных значений ряда к значениям, усредненным в определенном интервале времени В этом случае интервал времени при вычислении каждого последующего показателя как бы скользит по временно́му ряду

Применение скользящего среднего особенно полезно при неясных тенденциях динамического ряда или в ситуациях, когда на показатели сильно воздействуют циклически повторяющиеся выбросы (резко выделяющиеся данные, так называемые интервенции).

Таблица 102

Вычисление скользящего среднего

Годы	Арифметические операции
1982	(84+94+92)/3=90,0
1983	(94+92+83)/3=89,7
1984	(92+83+91)/3=88,7
1985	(83+91+88)/3=87,3 и т д

В приведенном примере временно́й интервал для вычисления скользящего среднего принят равным 3 годам В результате проведенного сглаживания основная тенденция динамического ряда стала более наглядной

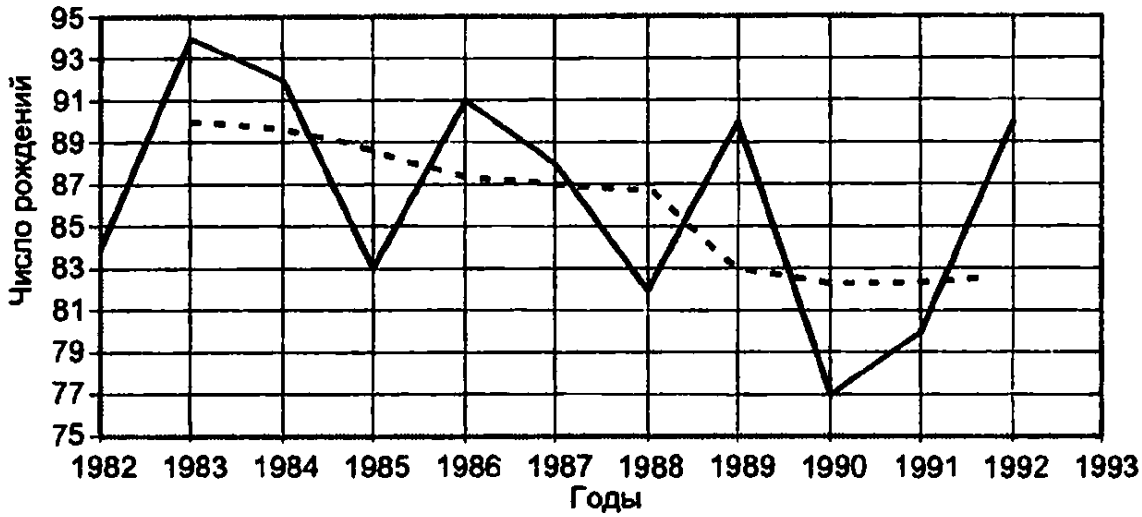


Рис. 119. **Результаты сглаживания методом скользящего среднего**

В целом, чем больше интервал сглаживания, тем более плавный вид имеет диаграмма скользящих средних. При выборе величины интервала сглаживания необходимо исходить из величины динамического ряда и содержательного смысла отражаемой динамики. Большая величина динамического ряда с большим числом исходных точек позволяет использовать более крупные временны́е интервалы сглаживания (5, 7, 10 и т д). Если процедура скользящего среднего используется для сглаживания не сезонного ряда, то чаще всего величину интервала сглаживания принимают равной 3 или 5.

Весьма результативным методом, хотя и более сложным, является сглаживание (выравнивание) рядов динамики с **помощью различных функций аппроксимации.**

При помощи этих функций формируется плавный уровень общей тенденции и основная ось динамики, около которой на протяжении определенного периода времени происходят колебания вверх и вниз

Одним из самых эффективных методов сглаживания с помощью математических функций является **простое экспоненциальное сглаживание**. Не вдаваясь в детальное математическое описание этого метода, следует отметить, что, в отличие от скользящего среднего или группового среднего, методикой простого экспоненциального сглаживания учитываются все предшествующие

наблюдения ряда, а не те, что попали в определенное интервальное окно. Точная формула простого экспоненциального сглаживания имеет следующий вид

$$S_t = \alpha X_t + (1-\alpha) S_{t-1},$$

где S_t — каждое новое сглаженное значение в момент времени t; S_{t-1} — сглаженное значение в предыдущий момент времени $t-1$, X_t — фактическое значение ряда в момент времени t; α — параметр сглаживания Если $\alpha = 1$, то предыдущие наблюдения полностью игнорируются Если $\alpha=0$, то игнорируются текущие наблюдения Значения α между 0 и 1 дают промежуточные результаты Изменяя значения этого параметра можно подобрать наиболее приемлемый вариант выравнивания Выбор наиболее оптимального значения α осуществляется путем анализа полученных графических изображений исходной и выровненной кривых, либо на основе учета суммы квадратов ошибок (погрешностей) вычисленных точек Более полно практическое использование этого метода представлено далее, в разделе «Обработка динамических рядов и прогноз динамики в *MS Excel*».

Одним из самых эффективных считается выравнивание по способу наименьших квадратов Согласно ему из бесконечного числа линий, которые могли бы быть теоретически проведены между точками, изображающими исходный ряд, выбирается только одна прямая, которая имела бы наименьшую сумму квадратов отклонений исходных (эмпирических) точек от этой теоретической прямой. Практически выравнивание производят либо по уравнению прямой $y = a + bt$, либо по уравнению параболы Уравнение параболы второго порядка выглядит следующим образом $y = a + bt + ct^2$. (В основе выбора параболы для выравнивания лежит предположение о том, что не скорость динамики, а ускорение является постоянной величиной) Где a, b и c — постоянные величины, t — порядковый номер какого-либо периода или момента времени (года и т п). С помощью этого уравнения вычисляются необходимые для построения соответствующие данные (табл 103) Подставляя из таблицы в соответствующую систему уравнений итоговые суммы, получим решение для прямой.

$$y^1 = 1{,}832 \times t + 17{,}748$$

Таблица 103

Выравнивание динамического ряда по способу панменьших квадратов

Номер года	Фактический уровень	Отклонение от центра	Расчетные параметры уравнений		
t	y	d	d^2	yd	y^1
1	16,5	−7	49	−115,5	20,6
2	14,3	−6	36	−85,8	22,4
3	44,0	−5	25	−220	24,3
4	35,6	−4	16	−142,4	26,1
5	30,4	−3	9	−91,2	27,9
6	32,4	−2	4	−64,8	29,8
7	22,5	−1	1	−22,5	31,6
8	28,8	0	0	0	33,5
9	15,2	1	1	15,2	35,3
10	42,0	2	4	84	37,1
11	26,6	3	9	79,8	39,0
12	42,6	4	16	170,4	40,8
13	51,3	5	25	256,5	42,6
14	46,2	6	36	277,2	44,5
15	53,4	7	49	373,8	46,3
Итого	501,8		280,0	514,7	

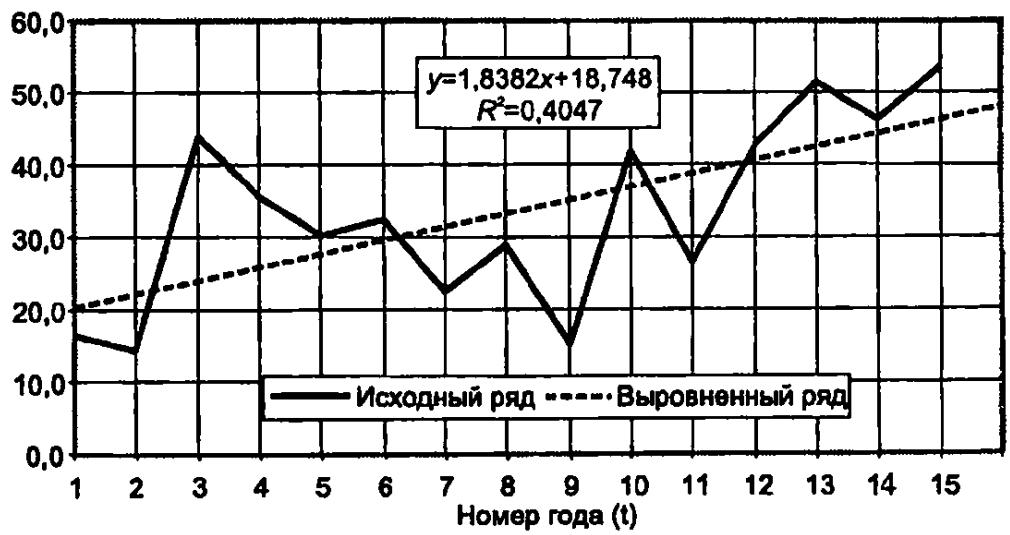


Рис 120. **Фактические и выровненные ряды**

Показателем правильности выбора того или иного уравнения служит коэффициент R^2. Чем больше его значение приближается к единице, тем больше соответствие фактического и выровненного распределений Максимальное значение, которое R^2 может принимать в предельном случае, равно 1

Следует отметить, что при решении проблемы выбора типа прямой или кривой, нельзя исходить из формальных соображений: та линия лучше, которая дает меньшую сумму отклонений эмпирического ряда от теоретического распределения Выбор кривой может быть обоснованным только на основе глубокого знания сути исследуемого явления

Современные программы статистической обработки данных позволяют получать различные теоретические кривые в автоматическом режиме, без каких-либо усилий со стороны исследователя. Имея эти результаты можно проводить математическую **экстраполяцию**, т е. давать прогноз показателей в продолжение проанализированного периода, или проводить **интерполяцию** рядов, т е. определять показатели в любой точке середины интервала проанализированного временно́го ряда *Например*, прогноз данных на последующий год. По номеру этот год будет 16 (см табл 103) С помощью имеющегося уравнения нетрудно провести линейную экстраполяцию путем следующих вычислений

$$y_{16} = 1{,}832 \times 16 + 17{,}748 = 47{,}1$$

Говорить о достоверности статистических прогнозов динамики каких-либо явлений можно лишь при сохранении общих тенденций, т е при наличии определенной степени инерционности явлений. Здесь имеется в виду инерционность взаимосвязей, которая обеспечивает сохранение в общих чертах механизма формирования явления, и инерционность характера динамики процесса (темп, направление, устойчивость) на протяжении достаточно длительных отрезков времени При этом существует закономерность чем на больший период времени вперед (или назад) производится экстраполяция данных, тем ниже точность прогноза Особенно резко снижается точность прогноза при значениях $R^2 < 0{,}6$

8.3. Обработка динамических рядов и прогноз динамики в *MS Excel*

В программе *MS Excel* имеется целый ряд возможностей статистической обработки динамических рядов

В качестве примера создадим в Excel таблицу «Динамика случаев заболеваний с временной утратой трудоспособности».

1. После запуска Excel введите в ячейку А2 слово: **Год** и нажмите клавишу **[Enter]**

2 В ячейки А3–А9 введите года **1985...1991.** Для ускорения ввода чисел достаточно ввести только одно число 1985 в ячейку А3, затем маркировать (выделить) блок ячеек А3:А9 и воспользоваться командой <**Заполнить** .>< **Прогрессия**> из меню <**Правка**> (рис 121)

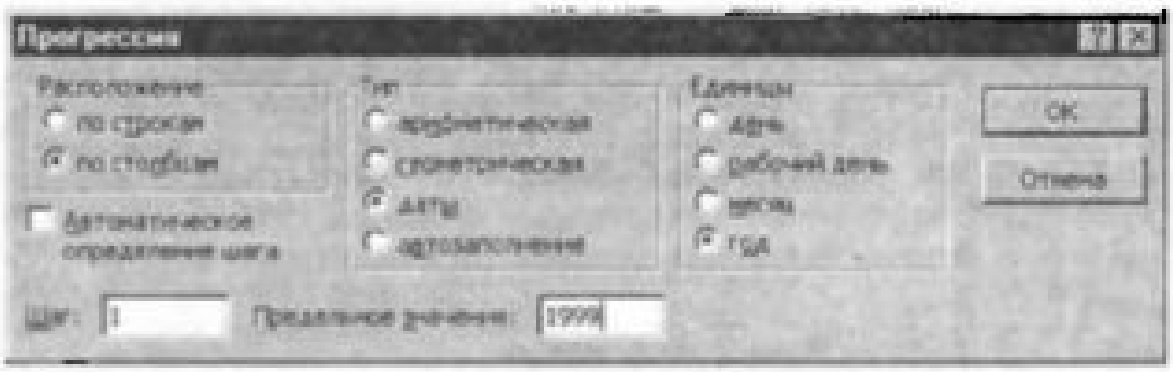
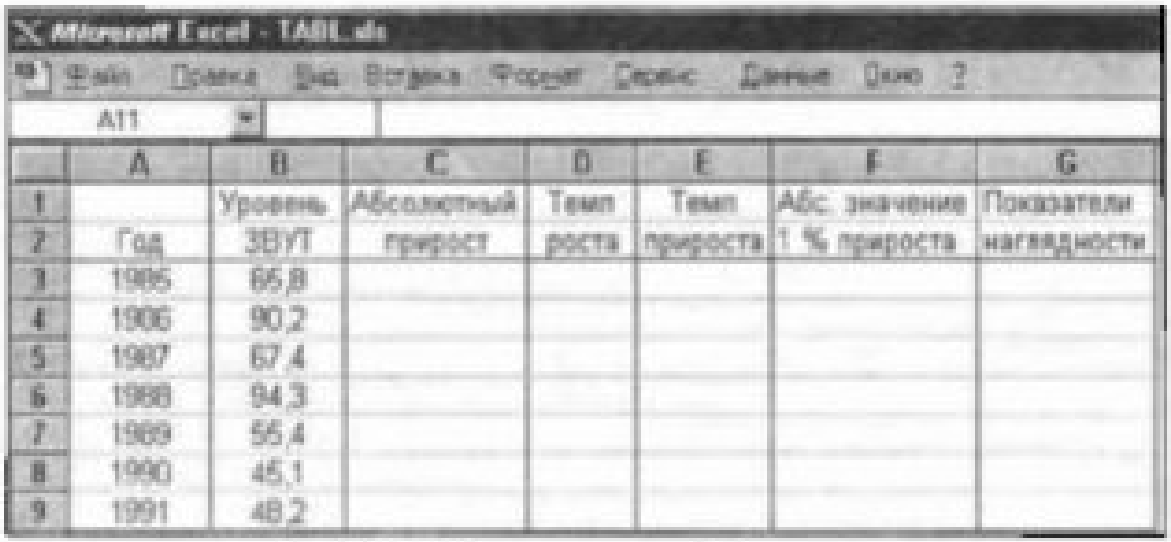


Рис 121. **Опция «Прогрессия» из команды «Заполнить»**

3. Чтобы завершить создание таблицы, запишите в ячейки В1–G2 названия столбцов, а также введите в ячейки В2–В9 числовые данные (рис. 122)

4. Заполните пустые столбцы таблицы В графу «Абсолютный прирост» занесите разность между последующим и предыдущим уровнями Для этого введите в ячейку С4 формулу =В4–В3



Microsoft Excel - TABL.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно ?

A11

	A	B	C	D	E	F	G
1		Уровень	Абсолютный	Темп	Темп	Абс. значение	Показатели
2	Год	ЗВУТ	прирост	роста	прироста	1 % прироста	наглядности
3	1985	65,8					
4	1986	90,2					
5	1987	67,4					
6	1988	94,3					
7	1989	55,4					
8	1990	45,1					
9	1991	48,2					

Рис. 122. **Таблица динамики заболеваемости с временной утратой трудоспособности (ЗВУТ)**

5. В графу «Темп роста» заносится отношение (в %) каждого последующего уровня к предыдущему Для этого введите в ячейку D4 формулу: =B4/B3*100

6. В графу «Темп прироста» заносится формула. =D4−100

7 Заполните графу «Абс значение 1% прироста» Для этого введите в ячейку F4 формулу =B3/100.

8. В графу «Показатель наглядности» заносится отношение (в %) каждого уровня к исходному уровню на 1985 г. Для этого введите в ячейку G3 формулу: =B4/65,8*100 или B4/$B$3*100. Знак $ включается в формулу, чтобы адрес ячейки B3 не изменялся, как обычно при копировании.

9 Скопируйте формулы из ячеек C3:G3 в область ячеек C4 G9 с помощью команды <**Копировать**> из меню <**Правка**> Для этого

- установите указатель ячейки в ячейку C3,
- промаркируйте ячейки с исходными данными. С этой целью: нажмите правую клавишу мыши и не отпуская ее переместитесь в ячейку G3. В результате этой операции область ячеек C3:G3, в которой располагаются формулы, окажется выделенной;
- выберите команду <**Копировать**> из меню <**Правка**>;
- передвиньте указатель ячейки в ячейку C5 и маркируйте ячейки C5:C9,

• нажмите клавишу **[Enter]**

Подготовленный ранее макет таблицы окажется заполненным результатами вычислений

Математическое выражение закономерностей динамики данных можно получить с помощью **функции экспоненциального сглаживания.**

Например: необходимо провести сглаживание числового ряда, отражающего динамику рождаемости за 1982—1993 годы.

Предварительно в таблице *MS Excel* необходимо разместить исходные данные (рис 123).

	A	B
1	1982	84
2	1983	94
3	1984	92
4	1985	83
5	1986	91
6	1987	88
7	1988	82
8	1989	90
9	1990	77
10	1991	80
11	1992	90
12	1993	78

***Рис. 123* Исходные данные для простого экспоненциального выравнивания**

Для проведения расчетов нужно вызвать из **«Сервис»** пакет **«Анализ данных»** В открывшемся окне выберите **«Экспоненциальное сглаживание»** и нажмите клавишу [OK] В открывшемся окне функции экспоненциального сглаживания (рис. 124) необходимо заполнить следующие поля

• **Входной интервал** — в нем указывается место размещения на листе *MS Excel* исходных данных

• **Фактор затухания** — в этом поле указывается значение коэффициента, показывающего глубину использования предыдущих данных для вычисления каждого последующего значения Этот коэффициент может принимать значения от 0 до 1 (0,1 или 0,2 или 0,3 и т д) Если он равен 1, то предыдущие наблюдения в ряду полностью игнорируются По умолчанию значение этого параметра принимается равным 0,3

• **Метки** — указываются, если были использованы заглавия столбцов.

• **Выходной интервал** — указывает место вывода (блок ячеек) готовых результатов. Можно указывать только одну клетку, ко-

торая в этом случае будет использована как левая верхняя точка диапазона вывода данных

• **Вывод графика** — используется, когда необходимо получить график, отражающий данные до и после выравнивания.

• **Стандартные погрешности** — демонстрируют значения ошибок для каждой точки выровненного ряда

Сумму этих погрешностей можно использовать как показатель точности подбора функции экспоненциального выравнивания. Чем меньше сумма этих погрешностей, тем больше соответствие между найденным и исходным распределением, т е. тем точнее выполнено выравнивание (аппроксимация) **Примечание**: обычно используют сумму квадратов погрешностей, т к формально они имеют знаки + и −.

После нажатия клавиши [OK], на экран будет выведен результат вычислений и график функции выравнивания

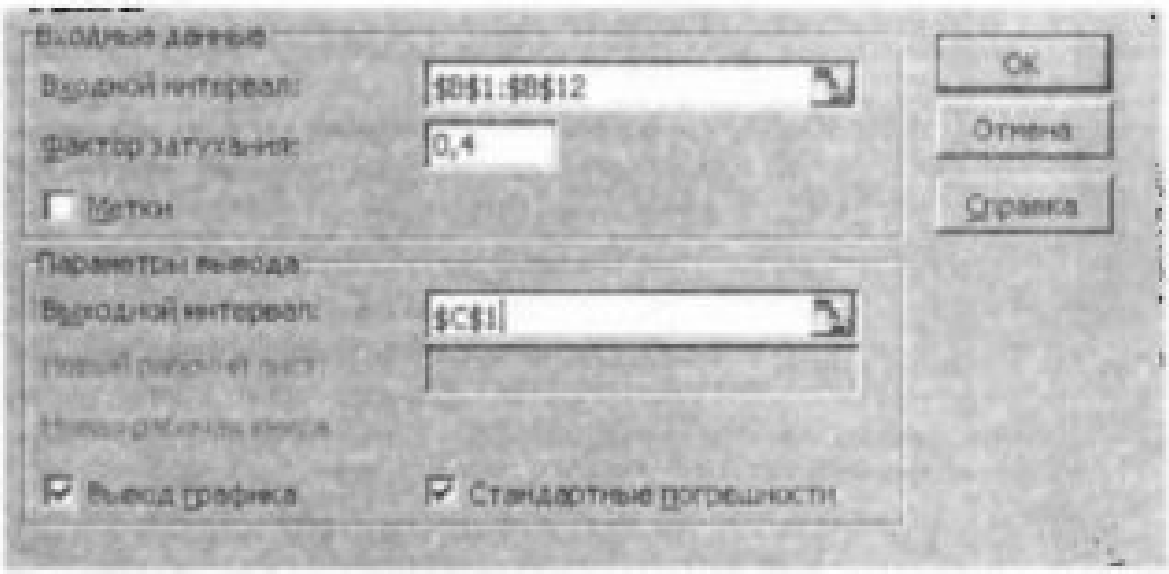


Рис. 124. **Окно функции экспоненциального сглаживания**

Для углубленной математико-статистической обработки динамического ряда весьма удобно использовать возможности пакета графического анализа *MS Excel*

Пример: необходимо с помощью статистических методов выравнивания выявить основную тенденцию следующего ряда распределения за 1984–1998 годы:

	A	B
1	Год	Данные ряда
2	84	18,5
3	85	14,3
4	86	44,0
5	87	35,8
6	88	30,4
7	89	32,4
8	90	22,5
9	91	28,8
10	92	15,2
11	93	42,0
12	94	26,8
13	95	42,8
14	96	51,3
15	97	46,2
16	98	53,4
17	99	

Рис. 125. **Начальное расположение информации для проведения анализа динамического ряда**

На первом этапе решения этой задачи в *MS Excel* необходимо построить диаграмму распределения данных Для этого·

1 Сформируйте таблицу с исходными данными (рис 125)

2 Вызовите мастер диаграмм Для этого с помощью пиктограммы на панели инструментов вызовите мастер диаграмм

Если пиктограмма не установлена, последовательно выполните команды <**Панели инструментов**> и <**Диаграммы**> из меню <**Вид**>

3. Выберите в появившемся диалоговом окне мастера диаграмм «1 из 4» график, отображающий развитие процесса во времени или по категориям и нажмите кнопку **[Далее>]**

4 В следующем окне («2 из 4») мастера диаграмм установите диапазон $B$2 $B1$6. Поставьте отметку **«Ряды в столбцах»** Переключитесь в окно **«Ряд»** и заполните его (см стр 89)

- в окне **Значения** установите адреса ячеек с исходными данными $B$2:$B$16
- в окне **Подписи оси X** — $A$2 $A$17

С помощью редактора диаграмм откорректируйте внешний вид диаграммы. (См раздел «2.6.3 Построение диаграмм в MS Excel»).

Затем, на появившейся диаграмме, выделите щелчком **левой клавиши** мыши ряд полученных точек, а затем щелчком **правой клавиши** мыши вызовите список команд и выберите **Добавить линию тренда**. В окне **Тип** и **Линии тренда** выберите **Полиномиальная** и **Степень 2**. Затем переключитесь в окно **Параметры** и установите там флажки

V **— Показывать уравнение диаграммы**

***V* — Поместить на диаграмму величину достоверности аппроксимации**

После этого нажмите на клавишу **[OK]** На экране появится изображение графика исходного ряда данных, аппроксимирующей линии (линии сглаживания) и уравнения аппроксимирующей функции (рис. 126). По своему виду эта функция представляет собой уравнение регрессии.

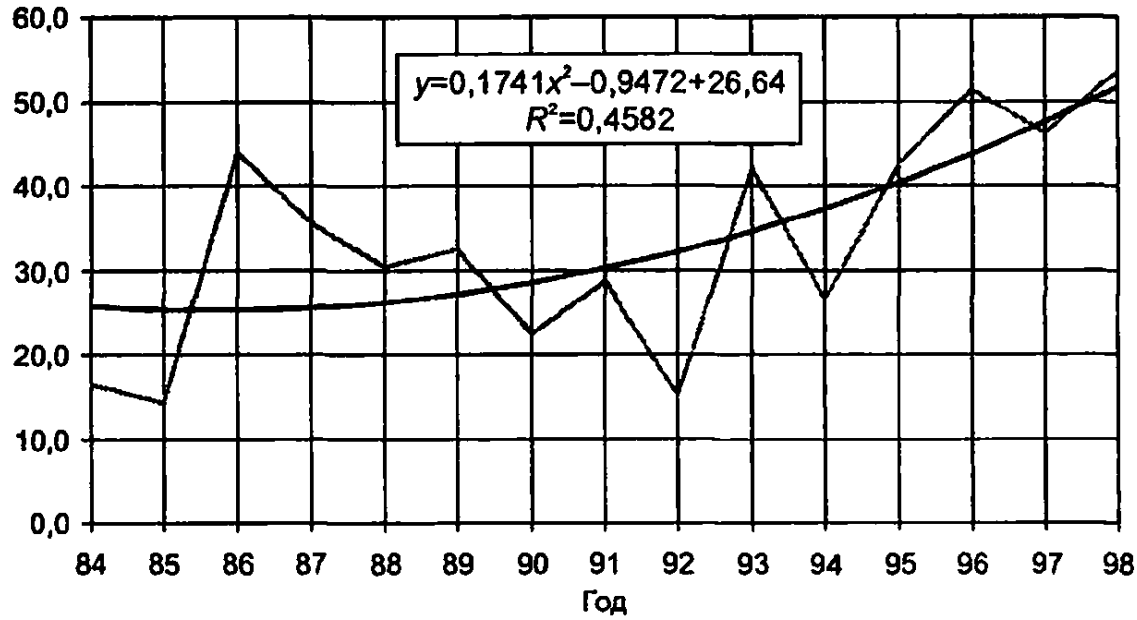


Рис. 126. **Диаграмма выравнивания дивамического ряда с помощью иолинома второй степени**

Параметр R^2 показывает насколько точно соответствует вычисленное уравнение регрессии истинной тенденции динамического ряда. Максимально возможное значение R^2=1,0 или 100% В данном примере, в случае использования аппроксимации (сглаживания) с помощью полинома 2-й степени, R^2=0,45 или 45%. (См. раздел 6.1 3 «Коэффициент линейной корреляции (Пирсона)»)

С помощью этого коэффициента можно подбирать функцию наиболее полно аппроксимирующую ту или иную тенденцию Например, при анализе этих же исходных данных полином 3-й степени дает более полную аппроксимацию, поскольку R^2 будет равен 0,56

С помощью рассмотренной функции мастера диаграмм можно производить прогнозирование значений *Y*. Для этого при подготовке к аппроксимации данных нужно в окне **«линия тренда»** установить величину *X* (единиц интервала ряда *X* вперед или назад). После нажатия на клавишу **[OK]** на экране появится изображение диаграммы с прогнозируемой точкой.

9. Оценка различий показателей заболеваемости

Одной из самых распространенных задач медицинской статистики является анализ различий показателей заболеваемости Этот анализ обычно проводится с целью выяснения роли конкретных факторов формирования уровней и структуры заболеваемости, вероятности возникновения тех или иных форм и исходов различных заболеваний. Организационно-методическим проблемам изучения заболеваемости посвящено большое количество специальных изданий по санитарной и медицинской статистике. Не останавливаясь на вопросах теории и практики решения этих проблем, более целесообразным представляется знакомство со способами оценки достоверности показателей заболеваемости. Этот выбор объясняется большим числом такого рода задач и большим количеством ошибок, допускаемых в этой области применения прикладной медицинской статистики.

9.1. Типичные ошибки, допускаемые при анализе показателей заболеваемости

Предпосылки к возникновению ошибок нередко возникают уже на первых этапах статистического исследования при планировании и сборе данных. В частности, когда из-за боль-

шой трудоемкости и высокой стоимости получения информации о состоянии здоровья тех или иных контингентов путем специально организуемого сбора информации предпринимаются попытки использовать более доступные данные из официальных источников Такого рода попытки почти всегда приводят исследователей к существенным просчетам Наиболес грубые просчеты возникают при использовании данных о заболеваемости по обращаемости (общей заболеваемости, инфекционной заболеваемости, травматизме, и т п), полученных из официальных источников

Ошибки здесь чаще всего возникают по двум причинам

1 Наличие механических, систематических ошибок в массиве данных Эти ошибки обычно возникают из-за неполной регистрации наблюдений по причине плохо налаженной работы соответствующих управленческих и организационных структур, из-за желания приукрасить истинное положение вещей и т д

2 Перенос данных, полученных сплошным наблюдением, но на ограниченной территории (в пределах небольшого населенного пункта, городского или сельского района и т п), на более обширные территории и контингенты населения

Неправомерность использования таких неполных данных объясняется методологией выборочного метода Выборочным способом, как известно, получают характеристики генеральной совокупности в пределах ошибок, которые возникают при перенесении результатов изучения выборки на всю генеральную совокупность Статистическая оценка достоверности различий базируется на вычислении этих ошибок, называемых *ошибками репрезентативности* Ошибки репрезентативности по своей природе не тождественны систематическим (механическим) ошибкам *Систематические ошибки* наблюдения, в отличие от ошибок репрезентативности, искажают результат наблюдения в одном направлении Поэтому распространять методологию выборочного метода для анализа совокупности, заведомо содержашей систематические ошибки, недопустимо Систематические ошибки, как правило, не могут быть устранены увеличением объема выборки Таким образом, попытки оправдать использование неполных данных их большой численностью несостоятельны

Проведение сплошного исследования ограниченной совокупности с целью последующего переноса данных на большую (генеральную) совокупность требует специальной организации несплошного исследования с соблюдением всех методических требований, которые предъявляются к выборочным исследованиям (см раздел «Когортный метод, монографический и др методы несплошного исследования») В любом случае при использовании данных официальной статистики необходимо иметь в виду реальное положение дел на местах с регистрацией всех случаев заболеваемости, особенно если имеется в виду заболеваемость по обращаемости Здесь большое значение имеет учет таких местных особенностей, как доступность медицинской помощи (радиус обслуживания медучреждений, наличие врачей-специалистов), близость крупных мегаполисов с развитой сетью специализированных медучреждений и т п

Вторая группа ошибок связана с неправомерным применением методик оценок различий, которые даются без учета особенностей тех или иных показателей заболеваемости и соответствующих этим особенностям методик статистического анализа

Учитывая наибольшую распространенность работ, посвященных такому виду заболеваемости, как заболеваемость с временной утратой трудоспособности, основные приемы статистической оценки различий показателей заболеваемости рассмотрим на этих показателях При соблюдении остальных общеизвестных требований к выборкам, описываемые методики оценки различий показателей заболеваемости с временной утраты трудоспособности можно перенести на оценки различий показателей большинства других видов заболеваемости

Известно, что для определения степени достоверности результатов выборочного исследования с помощью параметрических методов необходимо для относительных величин и для средних вычислять их средние ошибки (m) Наиболее просто вычислять такие ошибки в случае, если показатели отражают так называемое альтернативное распределение

9.2. Определение различий альтернативных показателей заболеваемости

Показатели заболеваемости, как интенсивные, так и экстенсивные, могут быть представлены в форме так называемых альтернативных распределений Например. показатели частоты болевших и не болевших лиц, показатели летальности (умер или остался жив) Онкологическая заболеваемость· альтернатива (заболел раком или не заболел и т п) В отличие от этих ситуаций можно несколько раз болеть острыми респираторными инфекциями. Поэтому при углубленном статистическом анализе заболеваемости целесообразно проводить сводку первичных данных таким образом, чтобы потом можно было использовать именно альтернативные показатели *Например*, при анализе заболеваемости с временной утратой трудоспособности можно использовать показатели болевших или не болевших лиц, которые обычно рассчитывают на основе численности круглогодовых контингентов (т е отработавших на момент исследования не менее 1 года). При таком подходе предусматривается возможность только двух исходов. работник мог болеть или не болеть в течение года (альтернативное варьирование) Средняя ошибка для альтернативных показателей рассчитывается по формуле:

$$m=\sqrt{\frac{Pq}{n}},$$

где P — % болевших (не болевших) лиц, исчисленный на постоянно (круглый год) работавших, $q=100-P$; n — число наблюдений (численность круглогодовых лиц, заболеваемость которых изучается)

Например В течение 1999 года на предприятии круглый год работали 1447 человек, из которых 827 человек имели случаи утраты трудоспособности из-за заболеваний Показатель болевших лиц:

$$P_б=\frac{827\times100}{1447}=57{,}1\%.$$

Показатель не болевших лиц.

$$P_н = \frac{(1447-827)\times 100}{1447} = 42{,}9\%$$ Или проще $P_н = 100-57{,}1 = 42{,}9$

Величины соответствующих ошибок:

$$m_б = \sqrt{\frac{57{,}1\times(100-57{,}1)}{1447}} = 1{,}3; \quad m_н = \sqrt{\frac{42{,}9\times(100-42{,}9)}{1447}} = 1{,}3$$

Обратите внимание, ошибки равны, т к сомножители в числителе формулы при таком распределении одинаковы

Величина средней ошибки, как известно, показывает интервал значений, в котором находится истинное (генеральное) значение показателя, т е. в каких пределах может колебаться размер показателя в зависимости от случайных причин, которые могут быть и не учтены в данном исследовании

В нашем примере показатель болевших лиц равен 57,1%, а его средняя ошибка ±1,3%. Понимать значение этих данных следует так Можно утверждать с вероятностью равной 68%, что во всех аналогичных производственно-профессиональных группах рабочих (при одинаковом возрастно-половом составе), показатели не болевших будут в пределах 57,1%±1,3%. То есть от 55,8% до 58,4% среди всех круглогодовых рабочих.

Если же необходимо получить выводы, степень вероятности которых должна составлять 95,5% , то тогда показатели не болевших будут в пределах интервала 57,1%±1,3%×2, т. е от 54,5% до 59,7%. (Подробнее о доверительном интервале см в разделе «Доверительная значимость, доверительная вероятность, доверительный интервал».)

Для того чтобы оценить достоверность разности между показателями, можно использовать критерий Стьюдента (*t*)

$$t = \frac{P_1 - P_2}{\sqrt{m_1^2 + m_2^2}}$$

(Подробнее см раздел «Критерии различия между долями, интенсивными величинами и средними арифметическими»)

Другим способом оценки достоверности разности показателей, полученных из совокупностей с альтернативным варьиро-

ванием, является оценка с помощью критерия Фишера по методу «фи», иногда называемому **точным методом Фишера** Его рекомендуется применять в тех случаях, когда показатель в % будет менее 20% или более 80% В указанных границах этот метод позволяет получать более точные оценки, чем критерий Стьюдента (*t*) Метод не требует больших расчетов, но для получения угловых значений «фи» и стандартных значений преобразованного критерия Фишера необходимо иметь две таблицы (см Приложение), одну из которых можно заменить функцией *MS Excel* FPACП. Вычисления производятся по формуле:

$$F=(\varphi_1-\varphi_2)^2\frac{n_1n_2}{n_1+n_2},$$

где F — критерий Фишера; φ_1 и φ_2 — углы «фи» в радианах, которые находят по специальной таблице (см. Приложение); n_1 и n_2 — число наблюдений в группах.

Например. В обследованной профессиональной группе было 150 круглогодовых рабочих, в контрольной группе — 600 Лиц, у которых был выявлен хронический бронхит, в первой группе было 4, а во второй 7. Требуется определить достоверность различий частоты заболеваний хроническим бронхитом, т е влияет ли на заболеваемость хроническим бронхитом профессиональный фактор (для данной профессии)

Показатели болевших лиц равны:

$$P_1=\frac{4\times100}{150}=2{,}7\%\ \text{и}\ P_2=\frac{7\times100}{600}=1{,}2\%$$

Затем следует перевести показатели в радианы Для первого показателя 2,7% угол φ_1 будет равен 0,33 Для второго показателя 1,2% угол φ_2 будет равен 0,22. Подставляем исходные значения в формулу:

$$F=(0{,}33-0{,}22)^2\times\frac{150\times600}{150+600}=1{,}45$$

Определяем число степеней свободы Первая степень свободы при альтернативном распределении всегда равна 1 Вторая = $n_1+n_2-2=150+600-2=748$ С помощью функции *MS Excel* FPACП находим вероятность нулевой гипотезы H_0 ($P_1=P_2$), которая

утверждает, что статистически значимых различий между двумя показателями нет. После подстановки значений в окно функции FPACП, получаем вероятность нулевой гипотезы равную 0,23 Для того чтобы нулевую гипотезу о равенстве показателей отвергнуть, ее вероятность не должна превышать 0,05. Таким образом, различия показателей частоты заболеваний хроническим бронхитом можно считать статистически несущественными

9.3. Определение различий интенсивных показателей заболеваемости при неальтернативном распределении

Особой проблемой в медицинской (санитарной) статистике является определение различий показателей заболеваемости, когда эти показатели построены на базе данных, не имеющих альтернативного распределения К этим данным относятся, прежде всего, интенсивные показатели частоты заболеваемости. Например, один и тот же человек может иметь в течение года несколько случаев заболеваний или несколько случаев временной утраты трудоспособности в связи с возникновением острых или обострений хронических заболеваний В связи с этим среднюю ошибку таких показателей нельзя вычислять по формуле, используемой в случае альтернативного распределения. Правомерным в этом случае может являться использование формулы $m=\frac{\sigma}{\sqrt{n}}$, где σ — среднеквадратическое отклонение (стандартное отклонение), n — численность наблюдений

Расчеты по этой формуле, обоснованные с точки зрения теории статистики, невозможно осуществить опираясь на большинство стандартных, официальных статистических учетных форм по заболеваемости, т. к. необходимо строить вариационные ряды. Эта операция возможна только при полицевом учете всего наблюдаемого контингента. На практике осуществить такой учет можно только при проведении специальных исследований Так, при изучении заболеваемости с временной утратой трудо-

способности вариантами (V) в вариационных рядах будут случаи временной нетрудоспособности (0, 1, 2, 3 и т д) в связи с определенным заболеванием или по всем болезням, вместе взятым, которые возникли в коллективе в течение года Частотами (P) для вариант будут числа лиц, утративших трудоспособность по этим причинам в течение года определенное число раз (0, 1, 2, 3 и т. д) *Например* (табл 104)·

Таблица 104

Распределение случаев заболеваний с временной утратой трудоспособности

Кратность заболеваний в году (V)	Число рабочих (P)	VP	V^2P
0	29	0	0
1	22	22	22
2	14	28	56
3	12	36	108
4	3	12	48
Итого	80	98	234

Отсюда, заболеваемость в среднем (среднее арифметическое) на 1 рабочего = 98/80=1,225 случая. На 100 рабочих 1,225×100= =122,5 случая за год. Дисперсия:

$$D=\frac{\sum V_i^2 P_j}{\sum P_i}-M^2, \text{ или } D=\frac{234}{80}-1{,}5=1{,}424.$$

Среднеквадратическое отклонение $\sigma=\sqrt{D}=1{,}19$ Ошибка показателя m=0,133 на 1 рабочего или 13,3 на 100 рабочих за год

С точки зрения формальной статистики, случаи заболеваний принято считать независимыми и случайными событиями с одинаковой вероятностью их возникновения Проанализировав характер распределения случаев заболеваний с временной утратой трудоспособности, В А Мозглякова (1964) предложила в небольших выборках (порядка 100 единиц наблюдения), где распределение данных относительно соответствует распределению Пуассона, для приближенных расчетов ошибки интенсивных показателей случаев заболеваний использовать формулу $m=\sqrt{\frac{M}{n}}$, где M —

среднее число случаев заболеваний на одного человека в год Расчеты могут упроститься, если вместо M взять интенсивный показатель P Тогда $m_P = \sqrt{\frac{P}{n}}$ В Ю Урбах (1967) считает, что эти формулы пригодны для оценки различий, когда вычисленное значение доверительного коэффициента (t) будет значительно отличаться в любую сторону от критического значения, равного 1,96

Указанный способ расчета средней ошибки показателя частоты случаев заболеваний является наиболее простым и может применяться не только для анализа заболеваемости с временной утратой трудоспособности, но и других видов заболеваемости по обращаемости. Однако здесь нужно помнить, что использовать доверительный критерий Стьюдента t следует весьма осторожно. Это связано с тем, что использование t-критерия возможно только в случае нормального распределения сравниваемых совокупностей Поэтому проводить сравнение можно только после проверки полученных (эмпирических) распределений на их соответствие нормальному распределению Такое уточнение особенно важно в тех случаях, когда доверительный критерий (t) оказывается близким к критическому (1,96).

Оригинальную методику оценки уровней заболеваемости в виде доверительного интервала с заданной доверительной вероятностью предложил М Б Славин (1989) Условием ее применения является использование малого временно́го интервала выборки, иначе минимальной плотности инцидентности Например, 1 день В этом случае в дальнейший расчет будет приниматься число случаев заболеваний намного меньшее численности контингента, заболеваемость которого анализируется Например в поселке АА за год было зарегистрировано 3650 случаев заболеваний (по обращаемости в местную поликлинику). Численность взрослого населения в поселке 2500 человек Таким образом, интенсивность заболеваемости составила 3650/2500×1000=1460 случаев за год на тысячу взрослого населения поселка АА При пересчете на 1 день (в абсолютных числах) получаем 3650/365=10 случаев заболеваний на 1 день; 10/2500=0,004 случая на 1 человека за 1 день

При такой ситуации закон распределения числа случаев заболеваний может быть описан распределением Пуассона Тогда при заданной доверительной вероятности α истинное значение заболеваемости будет располагаться внутри интервала, нижняя граница которого будет равна:

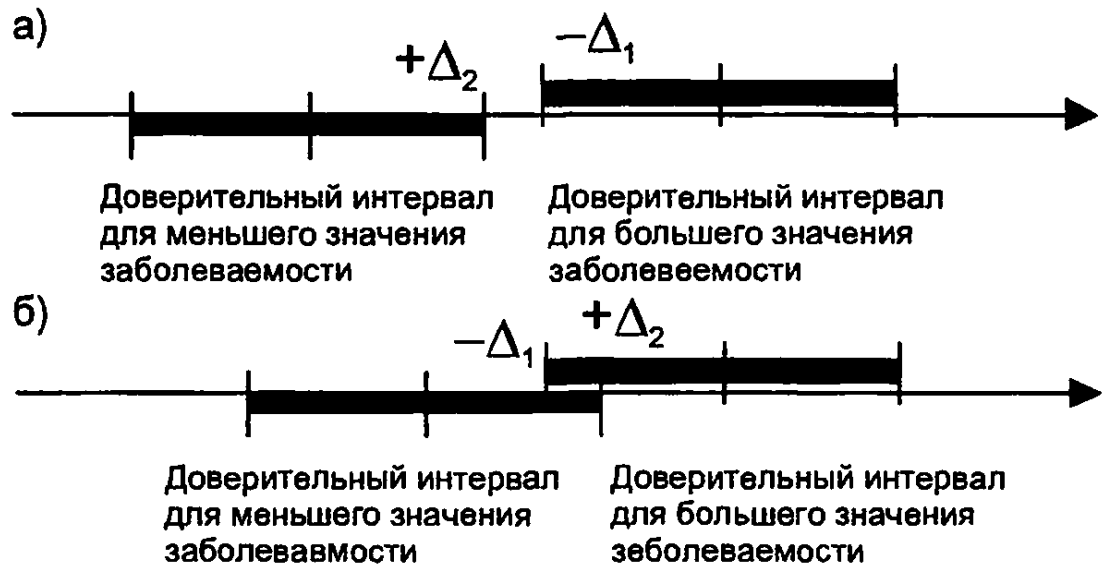
$$-\Delta=\frac{1}{2n}\chi^2\left(2m,\frac{1+\alpha}{2}\right).$$

Верхняя граница:

$$+\Delta=\frac{1}{2n}\chi^2\left(2(m+1);\frac{1-\alpha}{2}\right),$$

где n — численность контингента (объем выборки), m — число случаев заболеваний (на один день).

Не вдаваясь в детали математического обоснования этого метода, следует остановиться на принципах расчетов Их смысл сводится к тому, что если разница достоверна, то при сравнении двух показателей заболеваемости верхняя граница меньшего показателя должна быть меньше нижней границы большего показателя Указанное соотношение будет более понятным, если рассмотреть рисунок, на котором изображены доверительные интервалы для двух выборочных оценок заболеваемости



В случае а) различие выборочных оценок уровней заболеваемости достоверно, в случае б) — недостоверно

9.3.1. Расчет доверительных интервалов для показателей заболеваемости в программе *Excel*

После ввода в ячейки A1 C5 исходных данных и расчета выборочных оценок заболеваемости следует выполнить

- ввести в ячейку B6 формулу:
  =1/(2*B2)*ХИ2ОБР((1+B5)/2;2*B3)),
- ввести в ячейку B7 формулу:
  =1/(2*B2)*ХИ2ОБР((1−B5)/2;2*(B3+1)).

Далее эти формулы можно скопировать в ячейки, соответственно, C6 и C7. Для проверки гипотезы $F_1 > F_2$ необходимо сравнить нижнюю границу для большей заболеваемости с верхней границей для меньшей заболеваемости

B7 = =1/(2*B2)*ХИ2ОБР((1-B5)/2;2*(B3+1))

	A	B	C
1		Выборка 1	Выборка 2
2	Объем выборки	2500	3100
3	Число случаев заболевания	28	13
4	Выборочная оценка заболеваемости	0,0112	0,0042
5	Доверительная вероятность	0,95	0,95
6	Нижняя граница заболеваемости	0,0074	0,0022
7	Верхняя граница заболеваемости	0,0162	0,0072
8			
9	0,0074 > 0,0072 - заболеваемость в первой группе достоверно		
10	выше, чем во второй группе		

Рис 127. **Вывод результатов анализа различий показателей заболеваемости**

9.4. Непараметрические критерии оценки различий показателей заболеваемости

Непараметрические критерии позволяют обнаружить различие между сравниваемыми совокупностями, в чем бы оно ни заключалось: в различии средних значений или в колеблемости (разбросе) значений вариант Сравнивая ряды наблюдений, состоящие из связанных попарно или из независимых вариант, выявляют различия между этими рядами, но по какому именно параметру (среднему уровню, вариабельности и др.) они различаются, установить с помощью этих критериев нельзя

При оценке заболеваемости с помощью непараметрических критериев используют абсолютные или производные характеристики, которые в данном случае рассматривают как отдельные варианты, составляющие выборочные совокупности

Существует большое число непараметрических критериев, каждый из которых показано применять при решении определенных задач и наличии необходимого числа наблюдений (пар) Так, с помощью критерия знаков, парного критерия Вилкоксона оценивают различия двух взаимосвязанных (сопряженных) совокупностей При их расчете принимают во внимание величины вариантов, составляющих сравниваемые совокупности Это позволяет расположить варианты в определенной последовательности или даже ранжировать их с тем, чтобы получить соответствующие критерии, которые оценивают затем по специальным таблицам, содержащим граничные значения этих критериев. С помощью непараметрических коэффициентов корреляции рангов Спирмена и Кендэла измеряют силу связи между двумя признаками В данной главе освещается методика расчета лишь некоторых критериев, которые чаще других могут быть использованы при оценке данных о заболеваемости

Критерий знаков (Z) С помощью этого критерия оценивают направленность изменений показателей заболеваемости в сравниваемых взаимосвязанных парах вариант. Достоверность определяют при подсчете числа однонаправленных эффектов. В экс-

периментальных и клинических исследованиях этот критерий обычно применяют в случаях, когда анализу подвергают сопряженные между собой варианты, полученные на одной совокупности Если число наблюдений менее 25 и критерий знаков не выявил различий, то тогда можно попробовать провести анализ с более чувствительным парным критерием Вилкоксона Техника расчета критерия знаков подробно описана в разделе 7 4.1 «Критерий знаков» настоящего издания

В качестве *примера*, с помощью критерия знаков определим достоверность различий в показателях заболеваемости работников одного из предприятий хроническим бронхитом в течение года до санатория-профилактория и после профилактория (табл 105)

Из 16 групп в 12 отмечено снижение частоты случаев временной утраты трудоспособности из-за хронического бронхита Меньшее число раз встречался рост заболеваемости — 4 раза Число степеней свободы 16−1=15. По таблице критических значений (см Приложение) находим, что для числа степеней свободы 15 меньшее число знаков должно составлять не более 3 при $P=0,05$ и не более 2 при $P=0,01$. Таким образом, положительное влияние пребывания в санатории-профилактории на снижение заболеваемости бронхитом статистически не подтверждено. **Обратите внимание!** Критерий знаков устанавливает статистическую значимость различий, но не величину этих различий

Сравнение двух и более независимых рядов показателей заболеваемости по **Неймени** Этот критерий относится к числу «быстрых» ранговых критериев. Может использоваться как для анализа данных о заболеваемости, представленных в абсолютных числах, так и для анализа относительных величин (интенсивных, экстенсивных и т п). Может использоваться для сравнительного анализа показателей заболеваемости из официальных источников. Условие применения такого критерия — число наблюдений в сравниваемых рядах должно быть одинаковым. *Например.* Необходимо оценить различия показателей заболеваемости трех районов области *Н* за период с 1992 по 1998 год

Таблица 105

Заболеваемость хроническим бронхитом работников, находившихся в санатории-профилактории
(случаев на 100 работников за год)

Номера групп, проходивших оздоровление в профилактории	До профилактория	После профилактория	Направленность изменений
1	9,2	6,0	–
2	12,3	11,5	–
3	8,4	6,8	–
4	7,5	5,6	–
5	10,2	15,4	+
6	15,3	10,6	–
7	10,4	12,7	+
8	14,9	11,4	–
9	12,2	9,1	–
10	15,2	12,2	–
11	10,8	9,3	–
12	11,7	8,1	–
13	8,4	4,9	–
14	6,7	10,2	+
15	8,0	11,0	+
16	15 9	12,7	–

Таблица 106

Сравнительная заболеваемость взрослого населения районов A, B и C
(случаев на 1000 населения за год)

Год	Район A		Район B		Район C	
	Заболеваемость	Ранг R_1	Заболеваемость	Ранг R_2	Заболеваемость	Ранг R_3
1992	1061,3	11	1071,3	12	1008,3	2
1993	1076,2	14	1082,2	17	1015,2	3
1994	1022,9	5	1102,9	21	1024,7	6
1995	1038,2	7	1073,4	13	1043,6	9
1996	1077,4	16	1001,9	1	1076,7	14
1997	1086,3	18,5	1016,1	4	1038,9	8
1998	1049,8	10	1099,3	20	1086,3	18,5
	n=7	$\sum R_1$=81,5	n=7	$\sum R_2$=88,0	n=7	$\sum R_3$=60,5

Каждому значению исследуемого признака (в данном случае заболеваемости), независимо от принадлежности к тому или иному ряду распределения, присваивается ранг Равным значениям придается средний ранг *Например* на 18-м и 19-м местах (табл 106) расположены одинаковые показатели заболеваемости (в районе А — 1086,3 и районе С также 1086,3) Им присваивается ранг =(18+19)/2=18,5 Затем, составляется таблица разностей рангов (табл. 107)

Таблица 107

Разность сумм рангов показателей заболеваемости районов А, В и С

	Район А $\sum R_1$=81,5	Район В $\sum R_2$=88,0	Район С $\sum R_3$=60,5
I группа	–	6,5	21,0
II группа	–	–	27,5
III группа	–	–	–

Полученную разность сравнивают с табличными критическими значениями (см Приложение) Если величина разности при данном числе наблюдений и количестве выборок (групп) превышает табличное значение, то различия следует признать достоверными. При n=7 и k=3 (в каждой из 3 групп по 7 единиц наблюдения) критическая величина разности должна быть не менее 47,6. Таким образом, самая большая разность не перекрывает критического значения Различия в этом случае считаются статистически не подтвержденными

Критерий Уайта (К) наиболее простой порядковый критерий Может использоваться для определения статистической достоверности различий групп попарно не связанных величин Другим достоинством этого критерия является возможность анализировать ряды с разным числом данных Критерий Уайта дает обобщенную оценку средних тенденций двух различных групп Техника вычислений критерия предельно проста В качестве примера рассмотрим данные уже использованные в предыдущем примере Для анализа будем использовать только данные рядов I и II (см. табл 106)

Таблица 108

Сравнительная заболеваемость взрослого населения районов А и В (случаев на 1000 населения за год)

Год	Район А		Район В	
	Заболеваемость	Ранг	Заболеваемость	Ранг
1992	1061,3	6	1071,3	7
1993	1076,2	9	1082,2	11
1994	1022,9	3	1102,9	14
1995	1038,2	4	1073,4	8
1996	1077,4	10	1001,9	1
1997	1086,3	12	1016,1	2
1998	1049,8	5	1099,3	13
	$n_1=7$	$\sum R_1=49$	$n_2=7$	$\sum R_2=56$

Так же, как и в предыдущем примере, расставим единые для обоих рядов номера рангов После того, как будут определены суммы рангов, проведем оценку полученных результатов с помощью таблицы критических значений критерия Уайта из Приложения По I ряду (район А) сумма рангов $\sum R_1=49$ и по II ряду (район В) $\sum R_2=56$. Находим в Приложении значение критерия Уайта при $n_1=7$ и $n_2=7$ При $P=0,05$ критическое значение критерия 36, при $P=0,01$ критическое значение 32. Меньшая из найденных сумм равна 49, что превышает критическое значение критерия Уайта Следовательно, имеющиеся различия в распределении показателей заболеваемости нельзя признать статистически достоверными.

Критерий соответствия (согласия) Пирсона χ^2 является одним из самых распространенных непараметрических критериев Поскольку он основан на оценке распределений, то целесообразно его применять для сравнительной оценки уровней заболеваемости, распределенных по группам (по возрасту, по времени и т п)

Известно, что относительные величины (статистические показатели) подразделяются на общие и частные. При изучении влияния возраста на уровень заболеваемости анализируется распределение показателей заболеваемости в различных возрастных группах. В основе описываемой методики применения критерия

χ^2 лежит анализ соответствия фактического и ожидаемого распределения показателей заболеваемости При этом условии общий показатель (суммарный для всех возрастных групп) становится как бы основой для нахождения ожидаемого распределения показателя заболеваемости в каждой возрастной группе.

Таблица 109

Оценка повозрастных различий показателей заболеваемости с помощью критерия χ^2

Вычисляемые данные	Возраст (лет)				
	до 30	30–39	40–49	50 и ст	Все
Число работников	55	60	120	47	282
Фактические числа случаев заболеваний (P)	11	17	33	6	67
Частота случаев заболеваний (на 100 раб)	20,0	28,3	27,5	12,8	23,8
Ожидаемые числа случаев заболеваний (P_1)	13	14	29	11	70
$(P-P_1)$	−2	3	4	−5	–
$(P-P_1)^2$	4	7	20	26	–
$(P-P_1)^2$ /P1	0,327	0,528	0,707	2,391	3,953

В данном примере (табл 109) порядок вычисления следующий:

1 Определяются «Ожидаемые» числа P_1 Для возрастной группы «до 30 лет» число обследованных работников — 55 человек. Если бы заболеваемость в этой группе была такая, как и в целом по всем обследованным (см группу Все), то численность заболевших (P_1) составила бы 55×23,8/100=13 человек

2. Затем получают разность фактических и «ожидаемых» чисел $(P-P_1)$ или для возрастной группы «до 30 лет» 11−13=−2 Каждую такую разность возводят в квадрат и делят на P_1 $(-2^2/13)$= =0,307.

3. Полученные таким образом частные в других возрастных группах суммируют. Эта сумма и будет вычисленным значением критерия χ^2.

4. С помощью функций *Microsoft Excel* или по таблице критических значений χ^2 определяют значимость результата В дан-

ном случае число степеней свободы можно определить как число столбцов (без итогов) минус 1, т е. 4−1=3. При уровне значимости 0,05 значение критерия должно быть не меньше 9,2 При уровне значимости 0,01 — 6,0. Поскольку вычисленное значение критерия меньше указанных, следует принять нулевую гипотезу, говорящую об отсутствии различий в распределении анализируемых показателей.

Аналогичные вычисления проводятся, если заболеваемость анализируется на основе других показателей заболеваемости *Например*, при расчете «ожидаемого» числа лиц, не болевших, используют общий процент не болевших и т. п

Применение критерия χ^2 не требует никаких предположений, касающихся характера распределения сравниваемых совокупностей Вместе с тем, как справедливо указывал В Ю Урбах (1967), необходимо иметь в виду, что если критерий χ^2 обнаруживает значимое различие между двумя распределениями, то остается все же неизвестным, обусловлено ли оно различиями в средних величинах изучаемых признаков или же несовпадением других параметров двух сравниваемых распределений (дисперсии, асимметрии и др)

В ряде случаев можно применять упрощенные методики оценок различий показателей заболеваемости с помощью χ^2 *Например*: требуется оценить достоверность различий в частоте обращений за медицинской помощью в связи с заболеванием ишемической болезнью сердца (ИБС) в зависимости от месяца заболевания (в первом полугодии) в населенном пункте *N* (табл. 110).

Поскольку вычисленное значение критерия χ^2 при числе степеней свободы 5−1=4 значительно меньше критического (определяется с помощью таблицы критических значений χ^2 или с помощью функции *Microsoft Excel*), статистическая значимость различий обращаемости по поводу ИБС не подтверждена. **Обратите внимание!** Данные о заболеваемости приводятся в абсолютных числах Поэтому о равенстве частоты заболеваний можно говорить только принимая численность населения в пункте N одинаковой в каждом месяце

Таблица 110

Оценка различий обращаемости по поводу ИБС в населенном пункте *N* за первое полугодие

Месяц	Число обращений	$P-P_1$	$(P-P_1)^2$	$(P-P_1)^2/P_1$
1	17	1	1	0,06
2	15	−1	1	0,06
3	18	2	4	0,25
4	14	−2	4	0,25
5	16	0	0	0,00
Всего	80			**0,63**
P_1	80/5=16			

Л. Е. Поляков, Д. М. Малинский, М. В Дубовик (1981) предложили использовать критерий χ^2 для оценки показателей структуры заболеваемости. С этой целью использовалась методика определения, рекомендованная В Ю. Урбахом, которую эти авторы незначительно модифицировали Указанную методику целесообразно использовать для сопоставления двух рядов данных о заболеваемости (для двух коллективов) При этом заболеваемость в сравниваемых группах может существенно отличаться как по общему числу случаев, так и по числу случаев по отдельным группам болезней Для расчетов не требуются данные о численности коллективов, заболеваемость которых изучается Для вычислений используется формула

$$\chi^2 = N_A N_B \frac{\sum \frac{d^2}{A+B}}{10000},$$

где N_A и N_B — итоговые числа случаев для групп А и В; d^2 — квадрат разности удельных весов случаев заболеваний по отдельным классам болезней; $A+B$ — суммарное число случаев заболеваний по отдельным классам болезней. Последовательность расчетов проследим на примере (табл 111)

Таблица 111

Сравнение структуры заболеваемости двух групп обследованных (А и Б)

Сокращенное наименование классов болезней	Случаев заболеваний					d=А−Б %	d^2	$d^2/(А+Б)$
	А	Б	А+Б	А%	Б%			
Инфекционные заболевания	8	5	13	2,0	1,1	0,861	0,74	0,06
Психические заболевания	8	4	12	2,0	0,9	1,087	1,18	0,10
Болезни нервной системы	37	29	66	9,2	6,5	2,658	7,06	0,11
Болезни системы кровообращения	26	39	65	6,5	8,8	−2,336	5,46	0,08
Болезни органов дыхания	199	224	423	49,5	50,6	−1,062	1,13	0,00
Болезни мочеполовых органов	12	19	31	3,0	4,3	−1,304	1,70	0 05
Болезни кожи и п к клетчатки	26	32	58	6,5	7,2	−0,756	0,57	0,01
Прочие	86	91	177	21,4	20,5	0,851	0,72	0,00
Итого	402	443	845	100	100		18,57	0,42

Подставляя в формулу данные из таблицы, получим: $\chi^2 = 402 \times \times 443 \times \times \frac{0,42}{10000} = 7,44$ Число степеней свободы (8−1)×(2−1)=7

По таблице критических значений критерия χ^2 находим (см. выше), что при $P=0,05$ критическое значение критерия равно 14,1 Таким образом, можно отметить отсутствие статистически значимых различий в структуре заболеваемости двух изучаемых групп

В определенном смысле любой из рассмотренных выше критериев различий показателей заболеваемости является отражением вероятности возникновения заболеваний, т е величина различий этих показателей является во многих случаях отражением риска возникновения заболеваний или, наоборот, отражением положительного эффекта воздействия. Таким образом, термин «риск» является во многом условным, и его толкование может изменяться в зависимости от характера изучаемого явления

В практике зарубежных эпидемиологических исследований широко применяются специальные показатели риска Особен-

ностью большинства из этих показателей является отсутствие оценок репрезентативности получаемых с помощью них результатов (иногда предлагаются косвенные оценки, получаемые через другие критерии)

Абсолютный риск (R), связанный с каким-либо потенциальным фактором риска (R_e), отражает вероятность изучаемого исхода (заболевания, смерти и т. п) у лиц, подвергающихся или подвергавшихся воздействию данного фактора (экспонированных к данному фактору). Наиболее часто R_e определяется как отношение количества заболевших (умерших и т п) из числа подвергшихся воздействию фактора к числу всех подвергшихся воздействию этого фактора Абсолютный риск возникновения изучаемого исхода у лиц, не подвергавшихся воздействию данного фактора (R_{ne}) вычисляется как показатель кумулятивной инцидентности в группе лиц, не подвергавшихся воздействию данного фактора (См. раздел «Статистические коэффициенты»)

Наиболее доступными и понятными критериями, которые применяются в эпидемиологии для оценки силы воздействия причинного фактора на изучаемый исход, являются показатели относительного риска

Относительный риск (RR) — отношение абсолютных рисков при наличии и отсутствии воздействия изучаемого фактора $RR=\frac{R_e}{R_{ne}}$ Иногда в отечественной литературе встречается другое обозначение этого, в принципе простого, критерия: OP (относительный риск). Показатель RR характеризует силу связи между воздействием и заболеванием. Если относительный риск >1 , то возникновение болезни может быть связано с действием данного фактора. Чем больше значение RR, тем важнее этиологическая роль фактора Если RR=1, то фактор не оказывает воздействия, а RR<1 означает превентивное действие изучаемого фактора

Для дальнейшего анализа данные аналитических исследований обычно сводятся в четырехпольную таблицу, иначе часто называемую таблицей сопряженности (См раздел «Коэффициенты Q и $Ф$»)

По данным четырехпольной таблицы сопряженности можно рассчитать еще ряд полезных показателей, интерпретация которых дает более глубокое представление о силе связи между изучаемым воздействием и заболеваемостью

Разность рисков (*RD*) — разность абсолютных рисков при наличии и отсутствии воздействия изучаемого фактора Показывает, к какому абсолютному повышению заболеваемости приводит воздействие фактора

$$RD = R_e - R_{ne} = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

Атрибутивная фракция (*AF*) — отношение разности рисков к абсолютному риску у подвергшихся воздействию изучаемого фактора, выраженное в процентах Представляет собой долю всех случаев заболеваний, обусловленную данным фактором Показатель не имеет смысла, если причинная связь на самом деле отсутствует

$$AF = \frac{RD}{R_e} \div 100 = \frac{R_e - R_{ne}}{R_e} \times \times 100(\%).$$

Для оценки относительного риска в исследованиях случай-контроль используется показатель *отношения преобладаний OR* (*odds ratio*, отношение шансов)

Отношение преобладаний (*OR*) или *отношение шансов* (ОШ) — специальный показатель, используемый для оценки относительного риска путем сравнения относительных частот воздействия фактора риска среди групп с различным исходом (случаев и контролей и т п) Следует обратить внимание на то, что иногда для обозначения этого критерия используют термин «вероятность» или «относительная вероятность» С точки зрения формальной математической статистики, использование такого термина является неправомерным (частота и вероятность не одно и то же!) Поэтому более правильно использовать упомянутые выше синонимы названий этого критерия

Оценка *OR* не отличается от таковой для *RR*. Если изучаемый фактор действительно является фактором риска, то *OR* должен быть больше 1 Чем больше значение *OR*, тем более существенно влияние данного фактора Если *OR*=1, то влияния фактора риска нет.

Таблица 112

Критические значения критерия разности для сравнения независимых выборок (по Неймени) для P =0,05

n	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
1	3,3	4,7	6,1	7,5	9,0	10,5	12,0	13,5
2	8,8	12,6	16,5	20,5	24,7	28,9	33,1	37,4
3	15,7	22,7	29,9	37,3	44,0	52,5	60,3	68,2
4	23,9	34,6	45,6	57,0	68,6	80,4	92,4	104,6
3	33,1	48,1	63,5	79,3	95,5	112,0	128,8	145,8
6	43,3	62,9	83,2	104,0	123,3	147,0	169,1	191,4
7	54,4	79,1	104,5	130,8	137,6	184,9	212,8	240,9
8	66,3	96,4	127,6	159,6	192,4	225,7	259,7	294,1
9	78,9	114,3	132,3	190,2	229,3	269,1	309,6	330,6
10	92,3	134,3	177,8	222,6	268,4	315,0	362,4	410,5
11	106,3	154,8	205,0	256,6	309,4	363,2	417,9	473,7
12	120,9	176,2	233,4	292,2	352,4	413,6	476,0	539,1
13	136,2	198,5	263,0	329,3	397,1	466,2	536,5	607,7
14	152,1	221,7	293,8	367,8	443,6	520,8	599,4	679,0
13	168,6	245,7	325,7	407,8	491,9	577,4	664,6	752,8
16	183,6	270,6	358,6	449,1	541,7	635,9	732,0	829,2
17	203,1	296,2	392,6	491,7	593,1	696,3	801,5	907,9
18	221,2	322,6	427,6	535,5	646,1	758,5	873,1	989,0
19	239,3	349,7	463,6	560,6	700,5	822,4	946,7	1072,4
20	258,8	377,6	500,5	626,9	756,4	883,1	1022,3	1158,1
21	278,4	406,1	538,4	674,4	813,7	955,4	1099,8	1243,9
22	298,4	435,3	577,2	723,0	872,3	1024,3	1179,1	1335,7
23	318,9	465,2	616,9	772,7	932,4	1094,8	1260,3	1427,7
24	339,3	495,8	657,4	823,3	993,7	1166,8	1343,2	1521,7
25	359,9	527,0	698,8	873,4	1036,3	1240,4	1427,9	1617,6

Таблица 113

Критические значения критерия Уайта (*K*)

n/n	4	5		6		7		8		9		10	
	0,05	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
4	10	11		12	10	13	10	14	11	15	11	15	12
5		17	15	18	16	20	17	21	17	22	18	23	19
6				26	23	27	24	29	25	31	26	32	27
7						36	32	38	34	40	35	42	37
8								49	43	51	45	53	47
9										63	56	65	58
10												78	71

Таблица 114

Таблица значений Z ($r = f(Z)$)

r	0	1	2	3	4	5	6	7	8	9
0	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1	0,10	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,20
2	0,20	0,21	0,22	0,23	0,24	0,26	0,27	0,28	0,29	0,30
3	0,31	0,32	0,33	0,34	0,35	0,37	0,38	0,39	0,40	0,41
4	0,42	0,44	0,44	0,46	0,47	0,48	0,50	0,51	0,52	0,54
5	0,55	0,56	0,56	0,59	0,60	0,62	0,63	0,65	0,66	0,68
6	0,69	0,71	0,71	0,74	0,76	0,78	0,80	0,81	0,83	0,85
7	0,87	0,89	0,89	0,93	0,95	0,97	1,00	1,02	1,05	1,07
8	1,10	1,13	1,13	1,19	1,22	1,26	1,29	1,33	1,38	1,42
9	1,47	1,53	1,53	1,66	1,74	1,83	1,95	2,10	2,30	2,65

Таблица 115

Критические значения одностороннего критерия χ^2 (хи-квадрат)

Число степеней свободы	Уровни значимости			Число степеней свободы	Уровни значимости		
	$P=0,05$	$P=0,01$	$P=0,001$		$P=0,05$	$P=0,01$	$P=0,001$
1	3,84	6,63	10,83	21	32,67	38,93	46,80
2	5,99	9,21	13,82	22	33,92	40,29	48,27
3	7,81	11,34	16,27	23	35,17	41,64	49,73
4	9,49	13,28	18,47	24	36,42	42,98	51,18
5	11,07	15,09	20,51	25	37,65	44,31	52,62
6	12,59	16,81	22,46	26	38,89	45,64	54,05
7	14,07	18,48	24,32	27	40,11	46,96	55,48
8	15,51	20,09	26,12	28	41,34	48,28	56,89
9	16,92	21,67	27,88	29	42,56	49,59	58,30
10	18,31	23,21	29,59	30	43,77	50,89	59,70
11	19,68	24,73	31,26	31	44,99	52,19	61,10
12	21,03	26,22	32,91	32	46,19	53,49	62,49
13	22,36	27,69	34,53	33	47,40	54,78	63,87
14	23,68	29,14	36,12	34	48,60	56,06	65,25
15	25,00	30,58	37,70	35	49,80	57,34	66,62
16	26,30	32,00	39,25	36	51,00	58,62	67,98
17	27,59	33,41	40,79	37	52,19	59,89	69,35
18	28,87	34,81	42,31	38	53,38	61,16	70,70
19	30,14	36,19	43,82	39	54,57	62,43	72,06
20	31,41	37,57	45,31	40	55,76	63,69	73,40

Таблица 116

Критические значения двустороннего *t*-критерия Стьюдента

Число степеней свободы	Уровни значимости			Число степеней свободы	Уровни значимости		
	$P=0,1$	$P=0,05$	$P=0,01$		$P=0,1$	$P=0,05$	$P=0,01$
1	6,31	12,71	63,66	31	1,70	2,04	2,74
2	2,92	4,30	9,92	32	1,69	2,04	2,74
3	2,35	3,18	5,84	33	1,69	2,03	2,73
4	2,13	2,78	4,60	34	1,69	2,03	2,73
5	2,02	2,57	4,03	35	1,69	2,03	2,72
6	1,94	2,45	3,71	36	1,69	2,03	2,72
7	1,89	2,36	3,50	37	1,69	2,03	2,72
8	1,86	2,31	3,36	38	1,69	2,02	2,71
9	1,83	2,26	3,25	39	1,68	2,02	2,71
10	1,81	2,23	3,17	40	1,68	2,02	2,70
11	1,80	2,20	3,11	41	1,68	2,02	2,70
12	1,78	2,18	3,05	42	1,68	2,02	2,70
13	1,77	2,16	3,01	43	1,68	2,02	2,70
14	1,76	2,14	2,98	44	1,68	2,02	2,69
15	1,75	2,13	2,95	45	1,68	2,01	2,69
16	1,75	2,12	2,92	46	1,68	2,01	2,69
17	1,74	2,11	2,90	47	1,68	2,01	2,68
18	1,73	2,10	2,88	48	1,68	2,01	2,68
19	1,73	2,09	2,86	49	1,68	2,01	2,68
20	1,72	2,09	2,85	50	1,68	2,01	2,68
21	1,72	2,08	2,83	51	1,68	2,01	2,68
22	1,72	2,07	2,82	52	1,67	2,01	2,67
23	1,71	2,07	2,81	53	1,67	2,01	2,67
24	1,71	2,06	2,80	54	1,67	2,00	2,67
25	1,71	2,06	2,79	55	1,67	2,00	2,67
26	1,71	2,06	2,78	56	1,67	2,00	2,67
27	1,70	2,05	2,77	57	1,67	2,00	2,66
28	1,70	2,05	2,76	58	1,67	2,00	2,66
29	1,70	2,05	2,76	59	1,67	2,00	2,66
30	1,70	2,04	2,75	60	1,67	2,00	2,66

Окончание таблицы 116

Число степеней свободы	Уровни значимости			Число степеней свободы	Уровни значимости		
	$P=0,1$	$P=0,05$	$P=0,01$		$P=0,1$	$P=0,05$	$P=0,01$
61	1,67	2,00	2,66	91	1,66	1,99	2,63
62	1,67	2,00	2,66	92	1,66	1,99	2,63
63	1,67	2,00	2,66	93	1,66	1,99	2,63
64	1,67	2,00	2,65	94	1,66	1,99	2,63
65	1,67	2,00	2,65	95	1,66	1,99	2,63
66	1,67	2,00	2,65	96	1,66	1,98	2,63
67	1,67	2,00	2,65	97	1,66	1,98	2,63
68	1,67	2,00	2,65	98	1,66	1,98	2,63
69	1,67	1,99	2,65	99	1,66	1,98	2,63
70	1,67	1,99	2,65	100	1,66	1,98	2,63
71	1,67	1,99	2,65	101	1,66	1,98	2,63
72	1,67	1,99	2,65	102	1,66	1,98	2,62
73	1,67	1,99	2,64	103	1,66	1,98	2,62
74	1,67	1,99	2,64	104	1,66	1,98	2,62
75	1,67	1,99	2,64	105	1,66	1,98	2,62
76	1,67	1,99	2,64	106	1,66	1,98	2,62
77	1,66	1,99	2,64	107	1,66	1,98	2,62
78	1,66	1,99	2,64	108	1,66	1,98	2,62
79	1,66	1,99	2,64	109	1,66	1,98	2,62
80	1,66	1,99	2,64	110	1,66	1,98	2,62
81	1,66	1,99	2,64	111	1,66	1,98	2,62
82	1,66	1,99	2,64	112	1,66	1,98	2,62
83	1,66	1,99	2,64	113	1,66	1,98	2,62
84	1,66	1,99	2,64	114	1,66	1,98	2,62
85	1,66	1,99	2,63	115	1,66	1,98	2,62
86	1,66	1,99	2,63	116	1,66	1,98	2,62
87	1,66	1,99	2,63	117	1,66	1,98	2,62
88	1,66	1,99	2,63	118	1,66	1,98	2,62
89	1,66	1,99	2,63	119	1,66	1,98	2,62
90	1,66	1,99	2,63	120	1,66	1,98	2,62

Таблица 117

Значения F' при $P=0,05$

Число степеней свободы	1	2	3	4	5	8	12	24
1	161,4	18,5	10,1	7,7	6,6	5,3	4,7	4,3
2	199,5	19,0	9,6	6,9	5,8	4,5	3,9	3,4
3	215,7	19,2	9,3	6,6	5,4	4,1	3,5	3,0
4	224,6	19,2	9,1	6,4	5,2	3,8	3,3	2,8
5	230,2	19,3	9,0	6,3	5,1	3,7	3,1	2,6
6	234,0	19,3	8,9	6,2	5,0	3,6	3,0	2,5
7	236,8	19,4	8,9	6,1	4,9	3,5	2,9	2,4
8	238,9	19,4	8,8	6,0	4,8	3,4	2,8	2,4
9	240,5	19,4	8,8	6,0	4,8	3,4	2,8	2,3
10	241,9	19,4	8,8	6,0	4,7	3,3	2,8	2,3
11	243,0	19,4	8,8	5,9	4,7	3,3	2,7	2,2
12	243,9	19,4	8,7	5,9	4,7	3,3	2,7	2,2
13	244,7	19,4	8,7	5,9	4,7	3,3	2,7	2,2
14	245,4	19,4	8,7	5,9	4,6	3,2	2,6	2,1
15	245,9	19,4	8,7	5,9	4,6	3,2	2,6	2,1
16	246,5	19,4	8,7	5,8	4,6	3,2	2,6	2,1
17	246,9	19,4	8,7	5,8	4,6	3,2	2,6	2,1
18	247,3	19,4	8,7	5,8	4,6	3,2	2,6	2,1
19	247,7	19,4	8,7	5,8	4,6	3,2	2,6	2,0
20	248,0	19,4	8,7	5,8	4,6	3,2	2,5	2,0

Таблица 118

Значения F' при P =0,01

Число степеней свободы	1	2	3	4	5	8	12	24
1	4052,2	98,5	34,1	21,2	16,3	11,3	9,3	7,8
2	4999,3	99,0	30,8	18,0	13,3	8,6	6,9	5,6
3	5403,5	99,2	29,5	16,7	12,1	7,6	6,0	4,7
4	5624,3	99,3	28,7	16,0	11,4	7,0	5,4	4,2
5	5764,0	99,3	28,2	15,5	11,0	6,6	5,1	3,9
6	5859,0	99,3	27,9	15,2	10,7	6,4	4,8	3,7
7	5928,3	99,4	27,7	15,0	10,5	6,2	4,6	3,5
8	5981,0	99,4	27,5	14,8	10,3	6,0	4,5	3,4
9	6022,4	99,4	27,3	14,7	10,2	5,9	4,4	3,3
10	6055,9	99,4	27,2	14,5	10,1	5,8	4,3	3,2
11	6083,4	99,4	27,1	14,5	10,0	5,7	4,2	3,1
12	6106,7	99,4	27,1	14,4	9,9	5,7	4,2	3,0
13	6125,8	99,4	27,0	14,3	9,8	5,6	4,1	3,0
14	6143,0	99,4	26,9	14,2	9,8	5,6	4,1	2,9
15	6157,0	99,4	26,9	14,2	9,7	5,5	4,0	2,9
16	6170,0	99,4	26,8	14,2	9,7	5,5	4,0	2,9
17	6181,2	99,4	26,8	14,1	9,6	5,4	3,9	2,8
18	6191,4	99,4	26,8	14,1	9,6	5,4	3,9	2,8
19	6200,7	99,4	26,7	14,0	9,6	5,4	3,9	2,8
20	6208,7	99,4	26,7	14,0	9,6	5,4	3,9	2,7

Таблица 119

Односторонний критерий U (Вилкоксона–Манна–Уитни)
Максимальное число инверсий, при которых различия между группами наблюдений можно считать значимыми
(по Е. В. Гублеру, 1978)

P=0,05

n_1/n_2	2	3	4	5	6	7	8
3	–	0					
4	–	0	1				
5	0	1	2	4			
6	0	2	3	5	7		
7	0	2	4	6	8	11	
8	1	3	5	8	10	13	15
9	1	4	6	9	12	15	18
10	1	4	7	11	14	17	20
11	1	5	8	12	16	19	23
12	2	5	9	13	17	21	26
13	2	6	10	15	19	24	28
14	3	7	11	16	21	26	31
15	3	7	12	18	23	28	33
16	3	8	14	19	25	30	36
17	3	9	15	20	26	33	39
18	4	9	16	22	28	35	41
19	4	10	17	23	30	37	44

P=0,01

n_1/n_2	2	3	4	5	6	7	8
5	–	–	0	1			
6	–	–	1	2	3		
7	–	0	1	3	4	6	
8	–	0	2	4	6	7	9
9	–	1	3	5	7	9	11
10	–	1	3	6	8	11	13
11	–	1	4	7	9	12	15
12	–	2	5	8	11	14	17
13	0	2	5	9	12	16	20
14	0	2	6	10	13	17	22
15	0	3	7	11	15	19	24
16	0	3	7	12	16	21	26
17	0	4	8	13	18	23	28
18	0	4	9	14	19	24	30
19	1	4	9	15	20	26	32

Таблица 120

Значение угла $\varphi = 2\arcsin\sqrt{P}$

% P	0	1	2	3	4	5	6	7	8	9
0,1	0,063	0,066	0,069	0,072	0,075	0,077	0,080	0,082	0,085	0,087
0,2	0,089	0,092	0,094	0 096	0,098	0,100	0,102	0,104	0 106	0,108
0,3	0,110	0,111	0,113	0,115	0,117	0,118	0,120	0,122	0,123	0,125
0,4	0,127	0,128	0,130	0 131	0,133	0,134	0,136	0,137	0 139	0,140
0,5	0,142	0,143	0,144	0,146	0,147	0,148	0,150	0,151	0 152	0,154
0,6	0,155	0,156	0 158	0,159	0,160	0,161	0 163	0,164	0,165	0,166
0,7	0,168	0,169	0,170	0,171	0,172	0,173	0,175	0,176	0,177	0,178
0,8	0,179	0 180	0,181	0 182	0,184	0 185	0,186	0,187	0 188	0,189
0,9	0,190	0,191	0,192	0,193	0,194	0,195	0 196	0,197	0,198	0,199
1	0,200	0,201	0,202	0,203	0,204	0,205	0,206	0,207	0,208	0,209
2	0,284	0,285	0,285	0,286	0,287	0,287	0,288	0,289	0,289	0,290
3	0,348	0,349	0,349	0,350	0,351	0,351	0,352	0,352	0 353	0,353
4	0 403	0,403	0 404	0 404	0,405	0 405	0,406	0,406	0 407	0,407
5	0,451	0,451	0,452	0,452	0 453	0,453	0,454	0,454	0 455	0,455
6	0,495	0,495	0 496	0,496	0,497	0,497	0,497	0,498	0 498	0,499
7	0,536	0,536	0,536	0,537	0,537	0 537	0,538	0,538	0,539	0,539
8	0,574	0,574	0,574	0,575	0,575	0,575	0,576	0 576	0 576	0,577
9	0,609	0,610	0,610	0 610	0,611	0,611	0,611	0,612	0,612	0,613
10	0,644	0,644	0,644	0,645	0,645	0,645	0,645	0,646	0,646	0 646
11	0,676	0,676	0,677	0,677	0,677	0,678	0,678	0,678	0,679	0,679
12	0,707	0,708	0,708	0,708	0,709	0,709	0,709	0,710	0,710	0,710
13	0,738	0 738	0,738	0,739	0 739	0,739	0 740	0,740	0 740	0,740
14	0,767	0,767	0,768	0,768	0,768	0,768	0,769	0,769	0,769	0,770
17	0,850	0 850	0,851	0,851	0,851	0,851	0,852	0,852	0,852	0,852
18	0,876	0,877	0,877	0,877	0,877	0,878	0,878	0,878	0,878	0,879
19	0,902	0,902	0,903	0,903	0,903	0,903	0 904	0 904	0,904	0,904
20	0,927	0,928	0,928	0,928	0,928	0,929	0,929	0,929	0,929	0,930
21	0,952	0,952	0,953	0,953	0 953	0,953	0,954	0,954	0,954	0,954
22	0,976	0,977	0 977	0,977	0 977	0,978	0,978	0 978	0,978	0,979
24	1,024	1,024	1 024	1,025	1,025	1,025	1,025	1 026	1,026	1,026

Продолжение таблицы 120

% P	0	1	2	3	4	5	6	7	8	9
25	1,047	1,047	1,048	1,048	1,048	1,048	1,049	1,049	1,049	1,049
27	1,093	1,093	1,093	1,093	1,094	1,094	1,094	1,094	1 095	1,095
28	1,115	1,115	1,116	1,116	1,116	1,116	1,117	1,117	1,117	1,117
29	1,137	1,138	1,138	1,138	1 138	1,138	1,139	1,139	1,139	1,139
30	1,159	1,159	1,160	1,160	1,160	1,160	1,161	1,161	1,161	1,161
31	1,181	1,181	1,181	1,182	1,182	1,182	1,182	1,183	1,183	1,183
32	1,203	1,203	1,203	1,203	1,203	1,204	1,204	1,204	1,204	1,204
34	1,245	1,245	1,245	1,246	1,246	1,246	1,246	1,247	1,247	1,247
35	1,266	1,266	1,267	1,267	1,267	1,267	1,267	1,268	1,268	1,268
37	1,308	1,308	1,308	1,308	1,309	1,309	1,309	1,309	1,309	1,310
38	1,328	1,329	1,329	1,329	1,329	1,329	1,330	1,330	1,330	1,330
39	1,349	1,349	1,349	1,350	1,350	1,350	1,350	1,350	1,351	1,351
41	1,390	1,390	1,390	1,390	1,391	1,391	1,391	1,391	1,391	1,392
42	1,410	1,410	1,411	1,411	1,411	1,411	1,411	1,412	1,412	1,412
43	1,430	1,431	1,431	1,431	1,431	1,431	1,432	1,432	1,432	1,432
44	1,451	1,451	1,451	1,451	1,451	1,452	1,452	1,452	1,452	1 452
45	1,471	1,471	1,471	1,471	1,471	1,472	1,472	1,472	1,472	1,472
46	1,491	1,491	1,491	1,491	1,492	1,492	1,492	1,492	1,492	1,493
47	1,511	1,511	1,511	1,511	1,512	1,512	1,512	1 512	1,512	1,513
49	1,551	1,551	1,551	1,551	1,552	1,552	1,552	1 552	1,552	1,553
50	1,571	1,571	1,571	1,571	1,572	1,572	1,572	1,572	1,572	1,573
52	1,611	1,611	1,611	1,611	1,612	1,612	1 612	1,612	1,612	1,613
53	1,631	1,631	1,631	1,631	1,632	1,632	1,632	1,632	1,632	1,633
54	1,651	1,651	1,651	1,651	1,652	1,652	1,652	1,652	1,652	1,653
55	1,671	1,671	1,671	1,672	1,672	1,672	1,672	1,672	1,673	1 673
56	1,691	1,691	1,691	1,692	1,692	1,692	1,692	1,692	1,693	1,693
57	1,711	1,711	1,712	1,712	1,712	1,712	1,712	1,713	1,713	1,713
58	1,731	1,732	1,732	1,732	1,732	1,733	1,733	1,733	1,733	1,733
59	1,752	1,752	1,752	1,752	1,753	1,753	1,753	1,753	1,753	1,754
60	1,772	1,772	1,773	1,773	1,773	1,773	1,773	1,774	1,774	1 774
61	1,793	1,793	1 793	1,793	1,793	1,794	1,794	1,794	1,794	1,794
63	1,834	1,834	1 834	1,834	1,835	1,835	1,835	1,835	1,835	1,836

Окончание таблицы 120

% P	0	1	2	3	4	5	6	7	8	9
64	1,855	1,855	1,855	1,855	1,855	1,856	1 856	1,856	1,856	1,856
65	1,875	1,876	1,876	1,876	1,876	1,877	1,877	1,877	1,877	1,877
67	1,918	1,918	1,918	1,918	1,919	1,919	1,919	1,919	1,919	1,920
68	1,939	1,939	1,939	1,940	1,940	1,940	1,940	1,941	1,941	1,941
69	1,961	1,961	1,961	1,961	1,961	1,962	1,962	1,962	1,962	1,963
70	1,982	1,983	1,983	1,983	1,983	1,983	1,984	1,984	1,984	1,984
71	2,004	2,004	2,005	2,005	2,005	2,005	2,006	2,006	2 006	2,006
72	2,026	2,027	2,027	2,027	2,027	2,028	2,028	2,028	2,028	2,028
73	2,049	2,049	2,049	2,049	2,050	2,050	2 050	2,050	2,051	2,051
74	2,071	2,072	2,072	2,072	2,072	2,073	2,073	2 073	2,073	2 074
75	2,094	2,095	2,095	2,095	2,095	2,096	2,096	2,096	2,096	2,096
76	2,118	2,118	2,118	2,118	2,119	2,119	2,119	2,119	2,120	2,120
77	2,141	2,141	2,142	2,142	2,142	2,142	2 143	2,143	2,143	2,143
78	2,165	2,165	2,166	2,166	2,166	2,166	2 167	2,167	2,167	2,167
79	2,190	2,190	2,190	2,190	2,191	2,191	2,191	2,191	2,191	2 192
80	2,214	2,215	2,215	2,215	2,215	2,216	2,216	2,216	2,216	2,217
81	2,240	2,240	2,240	2,240	2,241	2,241	2,241	2 241	2 242	2 242
82	2,265	2,266	2,266	2,266	2,266	2,267	2,267	2,267	2,267	2,268
83	2,292	2,292	2,292	2,292	2,293	2,293	2,293	2,293	2,294	2,294
84	2,319	2,319	2,319	2,319	2,320	2,320	2,320	2,320	2,321	2,321
86	2,375	2,375	2,375	2,375	2,376	2,376	2,376	2,377	2 377	2,377
88	2,434	2,434	2,435	2,435	2,435	2,436	2,436	2,436	2,437	2,437
89	2,465	2,466	2,466	2,466	2,467	2,467	2,467	2,468	2,468	2,468
90	2,498	2,498	2,499	2,499	2,499	2,500	2,500	2 500	2,501	2,501
91	2,532	2,533	2,533	2,533	2,534	2,534	2,534	2,535	2,535	2,535
92	2,568	2,568	2,569	2,569	2,570	2,570	2,570	2,571	2 571	2,571
95	2,691	2,691	2,691	2,692	2,692	2,693	2,693	2,694	2,694	2,695
96	2,739	2,739	2,740	2,740	2,741	2,741	2,742	2,742	2,743	2,743
97	2,793	2,794	2,795	2,795	2,796	2,796	2,797	2,798	2,798	2,799
98	2,858	2,859	2,859	2,860	2,861	2,861	2,862	2 863	2,864	2,864
99	2,941	2,942	2,943	2,944	2,945	2 946	2 947	2 948	2,949	2,951

Таблица 121

Граничные значения $\tau = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$ **и** $\tau = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$ **при** $P = 0{,}01$

n	τ	n	τ	n	τ
4	0,991	13	0,520	22	0,414
5	0,916	14	0,502	23	0,407
6	0,805	15	0,486	24	0,400
7	0,740	16	0,472	25	0,394
8	0,683	17	0,460	26	0,389
9	0,635	18	0,449	27	0,383
10	0,597	19	0,439	28	0,378
11	0,566	20	0,430	29	0,374
12	0,541	21	0,421	30	0,369

Словарь терминов

Перечень, а иногда и трактовка используемых в различных разделах статистики и информатики терминов, являются весьма обширными Исчерпать их полностью в данном небольшом словаре невозможно Цель этого словаря — дать в краткой форме пояснения к основным терминам, которые использованы в издании. Кроме того, приводятся термины и не упомянутые в издании, но часто встречающиеся в медицинской статистике и информатике.

Абсолютная величина (модуль) — числовое значение признака (переменной), взятое без знака + или — В *Microsoft Excel* абсолютная величина может быть получена с помощью встроенной функции **ABS**

Абсолютный адрес — адрес ячейки электронной таблицы, не изменяющийся при переносе формулы в другую ячейку Используется в том случае, когда при копировании формулы адрес ячейки не должен изменяться Признаком абсолютного адреса в *Ms Excel* является знак $ (доллара) Например, $D$4 — абсолютный адрес ячейки, в которой зафиксированы координаты ячейки D4.

Адрес ячейки в MS Excel — координаты места положения ячейки в электронной таблице Адрес задается группой символов, включающей буквенное наименование столбца и номер строки, на пересечении которых находится ячейка Например, ячейка с адресом D5 находится на пересечении столбца D и 5-й строки таблицы. Кроме того, адрес ячейки может содержать название файла, листа и др. атрибуты

Активная ячейка в MS Excel — ячейка электронной таблицы, выделенная указателем ячейки В активную ячейку можно вводить данные или выполнять над ее содержимым операции.

Алгебра булева — раздел математики, оперирующий логическими (булевыми) переменными «ИЛИ», «НЕ», «И» Является основой при составлении различных алгоритмов

Альтернативная (конкурирующая) статистическая гипотеза — предположение (H_1), являющееся логическим отрицанием проверяемого предположения При этом проверяемую статистическую гипотезу (H_0) называют нулевой

Альтернативный признак — признак, принимающий только одно конкретное значение. (Пол. или мужской или женский)

Амплитуда (размах вариации) — один из статистических критериев разнообразия признака (характеристика вариабельности, изменчивости признака) Представляет собой разность между наибольшим и наименьшим значениями признака

Анализ — метод исследования, основанный на расчленении целого на составные части. Логическая последовательность этого метода от общего — к частному

Анамнестическое обследование — статистическое обследование, при котором исходные данные собираются путем опроса о прошлых событиях

Аппроксимация — приближенное выражение математических данных через другие, как правило, более простые

Априорно — до получения фактических результатов опыта, наблюдения

Артефакты (выбросы, резко выделяющиеся наблюдения) — варианты выборки, сильно отличающиеся от основной массы наблюдений Обычно это — ошибочные данные, результаты грубых просчетов, которые способны значительно исказить итоги всего анализа Располагаются обычно в самом конце или в самом начале ранжированного ряда наблюдений Исключить артефакты можно либо отбрасывая крайние значения ранжированного ряда, либо с помощью специальных статистических

критериев, позволяющих выявлять «выскакивающие» варианты в выборке.

Асимметрия — в статистике — мера скошенности распределения случайной величины Количественное описание асимметрии дается с помощью коэффициента *асимметрии* (*Ас*) Если $Ac = 0$, то распределение считается асимметричным Используется при числе наблюдений $n > 50$

Байт — наименьшая единица информации (равна 8 битам) Имеет собственный адрес в памяти ЭВМ. В наиболее общем виде представляет собой один печатный знак (буква, цифра, точка, пробел)

Библиотека программ — организованная по какому-либо принципу совокупность программ (программных средств) для ЭВМ

Бимодальное распределение — распределение, имеющее две моды (два пика) Является доказательством того, что распределение отличается от нормального Встречается, как правило, в ситуациях, когда две разнородных выборки объединены в одну *Например* общее распределение показателей физического развития для мужчин и женщин.

Бит — единица энтропии. Обозначает одно из двух возможных состояний 0 или 1 (в двоичной системе), да — нет, включено — выключено и т п Используется как наименьшая единица количества информации

Блок (диапазон) ячеек — прямоугольная область в электронной таблице, которая задается адресами левой верхней и правой нижней ячеек Например, блок ячеек A2:B3 включает в себя ячейки A2, A3, B2 и B3

Верификация — установление подлинности Проверка истинности

Вероятность — числовая мера возможности наступления случайного события Описывается числом в интервале от 0 до 1

Вероятность доверительная — вероятность, оценивающая достоверность характеристик, полученных выборочным путем

Версия (в информатике) — вариант программного продукта (алгоритма). Информация, являющаяся модификацией другого информационного продукта.

Вес — коэффициент, на который умножается каждый элемент ряда, например варианты вариационного ряда. Для вариационного ряда этот коэффициент определяется частотой встречаемости той или иной варианты в исследуемой совокупности.

Возрастная пирамида — графическое изображение распределения группы людей по полу и возрасту.

Воспроизводимость (reproducibility) наблюдений — качество наблюдений, отражающее близость друг к другу результатов измерений. Зависит не только от точности метода наблюдений, но и от стабильности характеристик объекта.

Выборка — часть генеральной совокупности, подвергающаяся статистическому обследованию. Обязательное требование к выборке — ее репрезентативность (представительность).

Выборка бесповторная — статистическая выборка, в которой отдельные единицы наблюдения из генеральной совокупности могут встретиться только один раз.

Выборка повторная — выборка, в которой отдельные единицы наблюдения из генеральной совокупности могут встретиться несколько раз.

Выборочная характеристика — приближенное значение какой-либо статистической характеристики генеральной совокупности, полученное на выборочной группе.

Выбросы — нетипичные или редкие значения, которые существенно отличаются от остальных данных числового ряда. Чаще всего (но не всегда!) их появление связано с ошибками измерения или аномальными явлениями.

Выделение — способ указания диапазона ячеек в электронной таблице. После выделения диапазон ячеек готов к выполнению операций над ним. Как правило, выделенный диапазон ячеек на экране отображается в инверсном виде.

Выравнивание — метод статистического анализа динамических рядов. Применяется для выявления основных тенденций

изменения показателей в случае отсутствия выраженных закономерностей

Гигабайт (Гбайт) — единица объема памяти Равна 2^{30} или 10 байт

Гипотеза нулевая (H_0) — статистическая гипотеза, в основу которой закладывается отрицание (отсутствие статистической взаимосвязи, не принадлежность выборки к данной совокупности, не соответствие выборки нормальному распределению и т п) Справедливость этой гипотезы подтверждается или не подтверждается статистической проверкой

Гипотеза статистическая — в наиболее общем виде — научное предположение о закономерностях случайных явлений или предположение о распределении статистических совокупностей Если гипотеза формулирует однозначное предположение, то она считается простой В противоположном случае — сложной

Гистограмма — графический способ представления интервальных распределений в виде ступенчатых фигур, состоящих из сдвинутых прямоугольников, площади которых пропорциональны частотам (частостям).

Графа — вертикальная составляющая статистической таблицы

Дефиниция — синоним «определение понятия»

Диаграмма — графическое изображение числовых величин, наглядно показывающее соотношение между ними

Директорий (директория) — в информатике — аналог терминов «каталог», «каталог файлов», «папка»

Дискретная величина — величина, значения которой меняются прерывно Является результатом счета Например число ударов пульса, число эритроцитов и т д

Дискриминантный анализ используется для принятия решения о том, какие признаки (факторы, переменные) дискриминируют или разделяют объекты на две или более естественно возникающих групп.

Дисперсионный анализ — статистический анализ, позволяющий оценивать связь между факторными и результативными признаками.

Дисперсия (D) — статистический критерий разнообразия признака (Синонимы характеристика вариабельности, изменчивости признака, разброса вариант, однородности распределений случайной величины.) В наиболее простом случае вычисляется как усредненный квадрат отклонений всех значений признака от среднего арифметического

Доверительные границы — границы интервальной статистической оценки

Доверительный коэффициент (надежность доверительной области) — см *Вероятность доверительная*

Достоверность информационная — свойство информации правильно отображать описываемые ею явления.

Достоверность статистическая — степень соответствия статистических характеристик выборки и генеральной совокупности, к которой эта выборка относится

Дуаль-карта — носитель информации, сочетающий функции первичного документа регистрации данных и машинного носителя информации Например, анкета опроса, адаптированная для непосредственного сканирования данных с нее в ЭВМ

Заболеваемость (болезненность) — число впервые зарегистрированных случаев заболеваний, а также длительно текущих хронических заболеваний, по поводу которых было обращение в течение года Существенным недостатком показателя является неполная регистрация хронических заболеваний

Загрузка — пересылка данных из носителя данных (дискета, жесткий диск или винчестер и т п) в оперативную память ЭВМ

Закон больших чисел — ряд положений теории вероятности, устанавливающих устойчивость средних характеристик при большом числе наблюдений.

Закрытие файла — процедура завершения работы программы с файлом При этом все изменения могут быть или сохране-

ны или нет Выделенные файлу буферы памяти освобождаются и файл становится недоступным.

Земская статистика — результаты исследований, проведенных на Европейской части дореволюционной России земскими органами (земствами) В основном содержали сведения о социально-экономическом состоянии деревни в период Столыпинских реформ

Значимость — вероятность ошибочно отвергнуть статистическую гипотезу или статистическая существенность Подтверждается отклонением нулевой гипотезы

Идентичность статистическая — тождество объектов по какому-либо признаку, достигнутое различными статистическими методами. Например, идентичные результаты, полученные на двух различных по объему выборках

Индекс *(лат index — показатель, список)* — статистический относительный показатель, характеризующий в сопоставимых величинах соотношение во времени (динамический индекс) или в пространстве (территориальный индекс) социально-экономических явлений цены, объемы продукции, себестоимость и т п

Интенсивный коэффициент — статистический показатель уровня, распространенности, частоты явления в своей среде Например: частота заболеваний гриппом на 1000 человек взрослого городского населения в 1998 году

Интервальная шкала — шкала измерений, которая позволяет упорядочить наблюдения и количественно выразить расстояния между ними

Интерполяция — в статистике — отыскание неизвестных значений признака по его известным значениям, если неизвестные значения находятся между известными значениями признака Например, нахождение показателей заболеваемости за 1995 год, при известных показателях за 1990–1994 и 1996–1998 годы Наиболее простой способ интерполяции — по графику кривой (прямой), отображающей динамику показателей

Инцидентность — частота выявления (развития) новых случаев заболеваний в группе за определенный срок Чаще всего за год

Каталог (директорий, папка) — справочник файлов с указанием их расположения Система каталогов обычно состоит из корневого (главного) каталога и подкаталогов. Подкаталоги могут иметь иерархическую структуру, при которой в подкаталоги более высокого уровня включаются подкаталоги более низкого уровня.

Кило — приставка в единицах измерения Соответствует 10^3 Единица объема памяти — 1 килобит = =1024 бит. 1 килобайт (1 К) = 1024 байт Единица производительности 1 килобайт в секунду.

Книга (синоним *Файлы*) — документ *Microsoft Excel,* состоящий из листов По умолчанию каждая книга *Excel* состоит из шестнадцати рабочих листов Количество листов в книге может быть увеличено или уменьшено Все листы книги сохраняются в одном файле с расширением XSL

Книга страшного суда — свод первой статистической поземельной переписи в Европе (Англия, 1086 год).

Колебания сезонные — повторяющиеся из года в год циклические изменения статистического показателя в одни и те же периоды времени (квартал, месяц)

Колеблемость (числового ряда, ряда распределения, вариационного ряда, динамического ряда) — случайные отклонения отдельных наблюдений ряда от их средних значений В статистике в качестве меры колеблемости чаще всего используют дисперсию, среднеквадратическое отклонение и коэффициент вариации

Конкурирующая гипотеза — статистическая гипотеза обратная (альтернативная) нулевой гипотезе

Конфигурация базовая — минимальный набор технических средств ЭВМ, позволяющий решать конкретные задачи

Конфигурация ЭВМ — совокупность аппаратных средств и соединений между ними, включаемых в данный комплект ЭВМ.

Копирование — процесс воспроизведения данных с сохранением исходной информации

Корреляция — статистическая зависимость, когда при изменении среднего значения одного признака, изменяется среднее значение другого

Коэффициент ассоциации — показатель статистической связи между двумя альтернативными признаками Чаще всего используется для оценки взаимосвязи качественных признаков.

Коэффициент вариации — один из статистических критериев разнообразия (колеблемости, разброса вариант) признака Вычисляется как процентное отношение среднеквадратического отклонения к среднему арифметическому Поскольку выражается в процентах, используется для сравнения разброса разнородных величин (килограммы, сантиметры и т п)

Коэффициент детерминации — квадрат коэффициента корреляции (R^2 или r^2). Характеризует долю влияния изучаемого факторного признака среди всех других, в том числе и не включенных в исследование, на результативный признак

Коэффициент контингенции — статистический показатель сходства Используется при оценке силы корреляционной связи качественных альтернативных величин По своему абсолютному значению всегда меньше аналогичного коэффициента — коэффициента ассоциации

Коэффициент корреляции множественный — мера тесноты статистической связи между одним признаком и линейной комбинацией набора нескольких других признаков.

Коэффициент корреляции частный — мера взаимосвязи двух статистических признаков при исключении влияния других признаков этой же совокупности

Коэффициент парной корреляции — мера линейной взаимосвязи двух статистических признаков

Коэффициент связи качественных признаков — числовая величина, характеризующая статистическую взаимосвязь качественных признаков. Сюда относятся коэффициенты корреляции Спирмена и Кендалла, вычисляемые на основе ранжированных значений исследуемых признаков Для признаков, измеренных в номинальной шкале, применяются коэффициенты Чупрова и

Пирсона (сопряженности) На основе таблиц сопряженности вычисляются коэффициенты ассоциации и детерминации

Критерий — характеристика объекта, которая при сравнении (оценке) рассматривается как наиболее существенная

Критерий Вилкоксона — один из первых непараметрических статистических критериев, позволяющих проверить гипотезу об одинаковости (однородности) распределения двух совокупностей

Критерий значимости различий — критерий, с помощью которого оценивается вероятность того, что сравниваемые выборки принадлежат к одной совокупности

Критерий непараметрический — статистический критерий, используемый в ситуации, когда закон распределения анализируемой совокупности неизвестен либо заведомо отличается от нормального закона

Критерий параметрический — статистический критерий, используемый в ситуации, когда распределение анализируемой совокупности подчиняется закону нормального распределения Например, критерий Стьюдента *t*

Критерий Пирсона (критерий согласия) — непараметрический критерий, который используется для статистической оценки соответствия распределений статистических совокупностей. Например соответствие распределения генеральной совокупности и выборочной, соответствие двух выборочных совокупностей и т п.

Критерии статистический — критерий, на основе которого принимается решение о статистической достоверности статистических гипотез

Курсор — метка на экране дисплея ЭВМ, указывающая место, куда будет выведен очередной символ

Лаг статистический (временной, запаздывания) — промежуток времени, за который изменение факторного признака приведет к изменению результативного Наличие лага означает, что результат влияния одного признака на другой обозначится только через определенный промежуток времени

Летальность — процентное отношение числа умерших от заболеваний к числу заболевших

Лист (синоним — *таблица)* — в *Microsoft Excel* ввод данных и вычисления выполняются в листах книги Лист (таблица) разделен на строки и столбцы

Логика вероятностная — логика, в которой истинные значения задаются вероятностями

Макет таблицы — статистическая таблица, не содержащая числовых данных Проект таблицы Формирование макетов таблиц — ключевой этап составления программы разработки статистических данных.

Марковский процесс — случайный процесс, вероятностные характеристики которого для будущего не зависят от того, когда и как этот процесс пришел в состояние на данный момент времени

Мастер — специальные программы *Microsoft Excel* для выполнения некоторых операций Например мастер диаграмм, мастер формул и др Эти программы значительно упрощают работу с электронной таблицей

Математическое ожидание — часто используемая в математической статистике величина, характеризующая среднее теоретическое значение, генеральное среднее

Матрица — система чисел, расположенных в прямоугольной таблице Замена строк на столбцы называется транспонированием матрицы. Математические действия над матрицами производятся согласно специальных правил

Мега — приставка в единицах измерения Соответствует 10^6 Единица объема памяти — 1 мегабит (Мбит)=1048576 бит; 1 мегабайт (1 М)=1,048576 байт Единица производительности ЭВМ — 1 мегабайт в секунду

Медиана — варианта, делящая вариационный ряд на две равные части Один из статистических показателей центра распределения В случае нормального распределения медиана равна среднему арифметическому и моде

Мера точности — показатель точности измерений. Используется наряду со среднеквадратическим отклонением (стандартным отклонением).

Метод группировок — расчленение изучаемой совокупности на группы, обладающие существенными общими признаками. Один из основополагающих методов статистического анализа.

Метод основного массива — метод выборочного статистического наблюдения, при котором из генеральной совокупности отбираются наиболее крупные, существенные единицы наблюдения.

Мода — наиболее часто встречающаяся варианта в вариационном ряду. Наиболее типичное значение случайной величины.

Модуль — см. *Абсолютная величина.*

Моменты — числовые характеристики распределения вероятностей случайных величин. Вычисляются как математическое ожидание *k* степени отклонения случайной величины от некоторой заданной точки (С). Если С=0, то момент считается начальным. Если точка С равна центру распределения, то момент называется центральным. Например: центральный момент второго порядка равен дисперсии.

Мощность статистического критерия — вероятность правильного отклонения (исключения) нулевой гипотезы (H_0), т. е. вероятность не совершить ошибку второго рода.

Наблюдение выборочное — обследование некоторого числа единиц наблюдения, отобранных из генеральной совокупности, с целью получения обобщающих статистических характеристик этой генеральной совокупности. При выборочном наблюдении необходимо соблюдать специальные правила, гарантирующие получение репрезентативных (представительных) данных.

Наблюдение единовременное — статистическое наблюдение, при котором фиксируются данные, имеющиеся на определенный момент времени. Например: численность населения района на конец 1997 года.

Наблюдение непрерывное (текущее) — статистическое наблюдение, при котором данные регистрируются по мере их поступле-

ния в течение какого-либо отрезка времени Например случаи инфекционных заболеваний за первое полугодие 1997 года.

Наблюдение сплошное — наблюдение всех без исключения единиц обследуемой совокупности

Наглядности коэффициент (относительная величина сравнения) — соотношение величин одноименных показателей, относящихся к разным промежуткам времени, территориям и т п Вычисляется, как правило, в процентах Например если число студентов, принятых на первый курс вуза, принять равным 100%, то на втором курсе их, по отношению к численности принятых в вуз, — 94%

Независимость случайных величин — одна из важнейших предпосылок теории вероятности и математической статистики. Случайные величины считаются независимыми, если значение одной из них не меняется от исхода другой

Нормализующее преобразование — преобразование числовых характеристик, после которого эти характеристики подчиняются закону нормального распределения

Объем выборки — число единиц статистического наблюдения в выборочной совокупности Чем больше объем выборки, тем более точные (репрезентативные) результаты могут быть получены Необходимый объем выборки определяется по специальным методикам с учетом видов выборки, способа отбора единиц наблюдения и изменчивости исследуемого признака

Отбор бесповторный — способ отбора, когда конкретная единица наблюдения может встретиться в выборке только один раз То есть эта единица не возвращается назад в генеральную совокупность после того, как она попала в выборочную совокупность

Отклонение от среднего (d) — разность между отдельными значениями признака (вариантами — *V)* и средним арифметическим *(M)* данной совокупности Служит основой для вычисления статистических критериев разнообразия признака Поскольку сумма всех отклонений от среднего всегда равна нулю, то перед суммированием указанные отклонения обычно возво-

дят в квадрат (d^2) или суммируют по модулю (без учета знака + или −).

Отклонение среднеквадратическое, стандартное отклонение (σ): статистический критерий разнообразия признака, характеристика вариабельности, изменчивости признака, разброса вариант, однородности распределения случайной величины.

Относительная величина (статистический коэффициент) — величина, выражающая соотношение статистических показателей изучаемых явлений Наиболее распространены в практике медико-биологических исследований интенсивные коэффициенты, экстенсивные, наглядности, соотношения, коэффициенты относительной интенсивности. При вычислении и использовании статистических коэффициентов необходимо обязательно учитывать единство места и времени анализируемых данных

Относительный адрес — адрес, используемый в *Microsoft Excel* в формулах для указания положения ячейки, на которые в формулах есть ссылки. При копировании или перемещении формулы относительные адреса модифицируются, т е изменяются в соответствии с изменением адреса, по которому записывается формула.

Оценка интервальная (доверительная) — оценка параметра или статистической величины в виде интервала Интервальное оценивание при выборочных исследованиях сводится к отысканию доверительного интервала и доверительной вероятности оцениваемого статистического параметра или отдельной величины.

Оценка несмещенная — точечная статистическая оценка, лишенная систематической ошибки выборки, совпадающая с оцениваемой величиной генеральной совокупности Например дисперсия выборки считается смещенной оценкой, чтобы ее можно было принять в качестве несмещенной оценки, обычно берут исправленное значение выборочной дисперсии

Оценка состоятельная — точечная статистическая оценка, которая при увеличении объема выборки максимально приближается к оцениваемому параметру генеральной совокупности

Оценка статистическая — приближенная оценка неизвестного статистического параметра распределения или самого распределения генеральной совокупности, проводимая на основании выборочных данных Статистические оценки бывают точечными и интервальными.

Оценка точечная — приближенное значение неизвестного статистического параметра генеральной совокупности, получаемое по данным выборочной совокупности Оценка получается вычислением одного значения (точки) этого параметра Например: выборочная средняя

Ошибка — см *Погрешность.*

Ошибка абсолютная — абсолютная разность между вычисленной статистически величиной признака и его действительным значением.

Ошибка второго рода — ошибка, при которой неверная статистическая гипотеза принимается как верная.

Ошибка выборки, репрезентативности (представительности) — расхождение между показателями (характеристиками) выборочной и генеральной статистической совокупности, к которой эта выборка относится. Основой возникновения такой ошибки является суждение по части (выборочным данным) о целом (генеральной совокупности).

Ошибка относительная — процентное отношение абсолютной ошибки к фактической величине изучаемого признака

Ошибка первого рода — ошибка, при которой верная статистическая гипотеза отвергается как неверная

Ошибка стандартная — среднеквадратическое отклонение

Панель инструментов — набор изображений кнопок, предназначенных для вызова с помощью мыши наиболее часто используемых команд *Microsoft Excel*

Папка — см *Каталог*

Параметр статистический — 1 Постоянная величина, характерная для какого-либо объекта 2 Вспомогательная переменная, от которой зависят другие переменные величины

Патологическая пораженность — показатель распространенности заболеваний, получаемый по данным медицинских осмотров При правильно организованном медицинском обследовании — наиболее точно характеризует распространенность хронических заболеваний и преморбидных состояний Недостатком показателя является неполная регистрация острых заболеваний.

Переменная — величина, значение которой в условиях данной задачи может меняться Величина, которая меняется в зависимости от значений аргумента (независимой переменной), называется зависимой переменной

Плотность инцидентности — частный случай показателя инцидентности (см). Используется если речь идет о различном по продолжительности пребывании в месте риска заболевания, заражения и т. п

Плотность распределения — статистическая характеристика распределения непрерывной случайной величины Производная функции распределения случайной величины. Кривая, изображающая плотность распределения, называется кривой распределения.

Погрешность — разность между истинным и приближенным значениями измеряемой величины

Погрешность абсолютная — погрешность, выраженная в единицах измеряемой величины

Погрешность грубая — погрешность, существенно превышающая фактическое значение погрешности

Погрешность измерения — погрешность, происходящая от несовершенства метода измерения

Погрешность относительная — отношение абсолютной погрешности к истинному значению измеряемой величины. Может быть выражена в процентах

Погрешность систематическая — погрешность, остающаяся постоянной или закономерно изменяющаяся при каждом повторном измерении

Погрешность случайная — погрешность, изменяющаяся случайным образом при каждом повторном измерении

Подкаталог (субдиректорий) — каталог, который является элементом другого каталога, называемого каталогом более высокого уровня. По аналогии каталог и подкаталог.

Показатели вариации — статистические показатели размаха колебаний значений признака К числу этих показателей обычно относят дисперсию, среднеквадратическое отклонение, коэффициент вариации

Полигон распределения — один из вариантов графического изображения статистических рядов распределения.

Полином — см. *Многочлен.*

Последовательный статистический анализ — метод, при котором число наблюдений может изменяться по ходу получения данных (метод Вальда) В медицинской информатике применяется при разработке диагностических программ

Превалентность — аналог широко применяющегося в отечественной статистике интенсивного показателя **патологическая пораженность** С помощью этого показателя оценивается уровень распространенности болезни на определенный момент времени (момент медицинского осмотра и т п)

Признак альтернативный — см *Альтернативный признак*

Признак атрибутивный (качественный) — статистический признак, отражающий атрибуты объекта наблюдения и выраженный словесно. Например. пол, диагноз заболевания, место жительства и т п

Признак количественный — статистический признак, выражаемый числом. Является результатом счета или измерения Например рост (см), вес (кг), частота пульса и т п.

Признак результативный — статистический признак, значение которого изменяется под воздействием фактора (факторного, факториального признака) Его значение является результатом изменения факторного признака

Признак статистический — отличительная черта, свойство, качество, присущее единицам наблюдения в статистической совокупности

Признак факторный (фактор) — признак, определяющий изменение результативного признака

Принцип практической невозможности маловероятных событий — принцип теории вероятности, утверждающий невозможность появления в отдельном наблюдении (испытании) события, вероятность которого в аналогичных случаях была близка к нулю

Прогноз интервальный — статистический прогноз, охватывающий «вилку» прогнозируемой величины Чем шире «вилка» значений, тем выше вероятность осуществления прогноза и ниже его точность

Прогноз точечный — прогноз в виде единственной статистической величины (точки)

Программа статистического наблюдения — перечень учетных признаков, подлежащих регистрации в ходе статистического наблюдения В зависимости от способа сбора материала, представляет собой регистрационную карту, опросный лист, анкету и т п

Программа статистической разработки — система группировок учетных признаков и статистических показателей, вычисляемых в ходе реализации статистического наблюдения Практически представляет собой перечень статистических таблиц, подлежащих заполнению и статистической обработке

Прогрессия арифметическая — числовая последовательность, каждый член которой, начиная со второго, равен предыдущему, сложенному с одним и тем же постоянным числом

Прогрессия геометрическая — числовая последовательность, первый член которой отличен от нуля, а каждый последующий, начиная со второго, равен предыдущему, умноженному на одно и то же постоянное число

Продецимилле — десятитысячная часть числа Обозначается ‰

Продольный анализ (когортный метод, метод реального поколения) — статистический метод, при котором наблюдения проводятся в когортах, т е в группах, одновременно вступивших в какое-либо состояние Например изучение демографических процессов в группах лиц одного года рождения

Производная величина (в статистике) — величина, получаемая путем преобразования других. Например темп прироста доли несовершеннолетних среди заболевших венерическими заболеваниями. Темп прироста — величина, измеряемая в процентах Доля заболевших — также измеряется в процентах

Промилле — тысячная часть числа Обозначается ‰

Проспективное наблюдение (когортное) — статистическое наблюдение, проводимое от факторного признака Т е факторный признак принимается за статистическое подлежащее. Например при изучении влияния возраста матери на вес новорожденных детей, все обследованные матери разбиваются на группы (когорты) в возрасте до 30 лет и 30 лет и старше. Затем проводится сравнительная оценка веса детей у матерей этих групп

Разведочный анализ данных — применяется для нахождения статистических связей между переменными в ситуациях, когда отсутствуют представления о возможной природе этих связей

Размах (амплитуда) числового ряда — разность между максимальным и минимальным значениями членов статистического ряда

Рандомизация — метод случайного отбора, по специально подобранному закону распределения

Распознавание образов — раздел информатики, посвященный проблемам формальной классификации и автоматической идентификации объектов на основе набора признаков

Распределение вероятностей — совокупность всех возможных значений случайной величины и соответствующих им вероятностей

Регрессионный анализ — статистический анализ, с помощью которого дается формализованная оценка зависимости среднего значения какого-либо статистического показателя (результатив-

ного признака) от нескольких других величин (факторных признаков)

Респондент (опрашиваемый) — лицо, сообщающее сведения при статистическом наблюдении, проводимом методом опроса.

Ретроспективное наблюдение — получение сведений о событиях в прошлом.

Референтное наблюдение (референтная группа) — эталонное, стандартное наблюдение или группа наблюдений Этот термин используют в случае, когда необходимо указать на величину или группу наблюдений, относительно которых производится оценка результатов конкретного исследования

Сводка — этап статистического наблюдения, предусматривающий систематизацию, подсчет групповых итогов, расчет производных величин

Сводка первичная — сводка статистического материала непосредственно в процессе наблюдения

Сезонности показатели — статистические характеристики проявлений сезонных колебаний.

Сезонность — периодически повторяющиеся во времени изменения, как правило, связанные со временем года

Система счисления — система изображения информации с помощью ограниченного числа знаков (цифр и букв) В двоичной системе используются цифры 0 и 1 В шестнадцатеричной — 1, 2, 3, 15 и т д

Систематические ошибки — ошибки измерения, наблюдения, исчисления, искажающие результат в одном направлении Ошибки такого рода не могут быть устранены увеличением объема выборки или наблюдения

Скользящее среднее — ***средняя*** величина, используемая для выявления основных тенденций рядов динамики При использовании скользящих средних фактические уровни динамического ряда заменяются средними, полученными из нескольких смежных значений динамического ряда

Случайная величина — величина, принимающая какое-либо значение в зависимости от случайных обстоятельств. Статисти-

ческие признаки в математической статистике обычно рассматриваются как случайные величины В зависимости от типа, случайные величины считаются дискретными (прерывными) и непрерывными, количественными и качественными, и т п

Случайное событие — событие, которое при определенных обстоятельствах может произойти, а может и не произойти, что задается вероятностью события

Совокупность статистическая — множество объектов, явлений, являющихся единицами статистического наблюдения, объединенных общими свойствами (признаками сходства) Отдельные единицы статистического наблюдения различаются между собой признаками различия

Средняя величина — обобщенная количественная характеристика статистической совокупности. В статистической совокупности при достаточно большом числе наблюдений влияние случайных факторов взаимопогашается и средняя величина дает характеристику типичного уровня признака. образовавшегося в данных условиях места и времени. Наиболее часто в статистике используют: моду, медиану, среднее арифметическое. В зависимости от характера статистической совокупности применяют среднюю гармоническую, среднюю логарифмическую, среднюю геометрическую, среднюю хронологическую и др

Стандартизация (коэффициентов) — математическое преобразование статистических коэффициентов с целью исключения неоднородности состава сравниваемых групп (по возрасту, полу и т п.)

Статистика здоровья населения — один из разделов санитарной статистики Оценка здоровья групп населения проводится на основе показателей заболеваемости, инвалидности, демографических показателей естественного движения населения (рождаемость, смертность и т п), показателей физического развития

Статистика здравоохранения — один из разделов санитарной статистики, посвященный изучению деятельности системы здравоохранения, разработке отраслевых норм и нормативов

(учреждения здравоохранения, основные фонды, штаты, подготовка специалистов и т. п.)

Статистическая модель — аналог явления или процесса, отображающий в главных чертах наиболее существенные статистические закономерности определенного объекта

Статистическое описание — характеристика сложных явлений с помощью статистических показателей

Степени свободы — числовая характеристика такого варьирования элементов совокупности, которое не приводит к изменению итоговых, суммарных показателей

Строка формул — предназначена для ввода данных или формулы в ячейку таблицы *Microsoft Excel* Строка формул отображает содержимое активной ячейки полностью, даже если его не видно в самой ячейке. Если активная ячейка содержит число, то в строке формул отображается формула, с помощью которой это число было получено

Счет — 1 Операция, позволяющая установить, сколько элементов содержит данное конечное множество 2 Совокупность первых четырех действий над рациональными числами сложение, вычитание, умножение, деление

Сходимость наблюдений (precision, repeatability) — качество наблюдений (измерений), выполненных в одинаковых условиях, отражающих близость к нулю случайных ошибок

Таблица — способ формализованного двухмерного представления данных

Таблица групповая — статистическая таблица, в которой представлено взаимное распределение двух признаков

Таблица комбинационная — статистическая таблица, в которой представлено взаимное распределение трех и более признаков

Таблица простая — статистическая таблица, в которой представлено распределение одного признака

Таблица сопряженности — табличное представление взаимного распределения двух качественных статистических признаков

Теорема сложения вероятностей — теорема, согласно которой вероятность нескольких несовместных событий, т е событий, совместное появление которых исключено, равна сумме вероятностей этих событий

Точность — 1. Мера возможности различать близкие объекты 2 Качество измерений, соответствующее близости измерений к истинному значению измеряемой величины

Тренд — общее направление развития, общая тенденция показателей динамического ряда

Указатель ячейки — рамка, с помощью которой в *Microsoft Excel* выделяется активная ячейка Указатель ячейки можно перемещать по таблице с помощью клавиш управления курсором или мыши.

Уровень доверительный — задается при статистической проверке гипотез величиной $(1-\alpha)$, где α — уровень значимости Обычно в медико-биологических исследованиях значение доверительного уровня принимается не менее 0,95, иногда 0,99

Уровень значимости (α *или* P) — величина вероятности того, что отвергаемая статистическая гипотеза окажется правильной. В медико-биологических исследованиях нулевая гипотеза отвергается при уровне значимости $P>0,05$ или в случае необходимости получения более точных результатов при $P>0,01$

Уровень ряда динамики — исходное значение, на основе которого вычисляются показатели динамического ряда

Устойчивость — свойство объекта сохранять определенные черты при малом изменении статистических параметров, от которых он зависит

Утилита — сервисная программа Например программа копирования данных

Файл — фрагмент данных, произвольного размера и типа, имеющий собственное имя (реквизиты) и хранящийся в памяти ЭВМ как единое целое

Фактор (в статистике) — учетный признак, находящийся в логической взаимосвязи с другим признаком или группой признаков

Флоппи-диск — то же, что гибкий магнитный диск (ГМД)

Формат данных — способ представления данных вне и в памяти ЭВМ с помощью специальных символов, соглашений и правил

Формат документа — порядок размещения реквизитов, размеры документа.

Форматирование таблицы — формирование внешнего вида и структуры электронной таблицы вид шрифта, цвет текста и фона, ширина столбцов, количество десятичных знаков в числах и т. д.

Функция — переменная величина, значение которой зависит от других переменных

Функция встроенная — функция, выполнение которой в данной программе возможно в автоматическом (полуавтоматическом) режиме Например функция *ABS* в *Microsoft Excel* позволяет получить абсолютное значение (модуль) числа

Характеристика статистическая — параметр, описывающий одно из свойств единицы наблюдения или статистической совокупности.

Цифры арабские — символы 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, с помощью которых можно написать любое натуральное число

Цифры римские — символы I, V, X, L, C, D, M, соответствующие числам 1, 5, 10, 50, 100, 500, 1000 С их помощью, используя повторения и определенные правила размещения, записывают натуральные числа в римской нумерации

Частость (в статистике) — относительная величина, показывающая долю частот отдельных вариант или группы вариант в общей сумме частот (в общем числе наблюдений) Сумма всех частостей равна 1,0. Если частости выражены в процентах, то их сумма равна 100%

Частота P — 1 Абсолютное число, показывающее сколько раз (как часто) встречается то или иное значение признака 2 Отношение числа испытаний, в которых случайное событие произошло, к числу всех испытаний При больших числах испытаний частота случайного события равна его вероятности.

Число π — отношение длины окружности к длине ее диаметра Приближенно равно 3,141592653590

Число e — является основанием натуральных логарифмов Определяется как предел выражения $(1+1/n)^{11}$ При $n\to\infty$ приближенно равно 2,718281828459

Шаг вариационного ряда (интервал) — величина приращения вариационного ряда при переходе от одной варианты к другой

Шкала — совокупность делений и их обозначений, зависящих от одной переменной Например, на логарифмической шкале расстояния между делениями пропорциональны разности логарифмов чисел, соответствующих этим делениям

Штрих — знак ('), помещаемый справа или слева вверху от буквы или выражения С его помощью отличают близкие объекты

Экстенсивный коэффициент (относительная величина доли, относительная величина структуры) — статистический показатель, представляющий долю, удельный вес, часть от целого Выражается в процентах

Экстраполяция — распространение результатов, полученных из наблюдений над одной частью явления, на другую его часть Например, если известна статистическая закономерность динамики показателей заболеваемости за период 5 лет, то можно определить возможный уровень заболеваемости в последующем. 6-м году

Эксцесс — статистическая мера островершинности распределения случайной величины Количественное описание эксцесса дается с помощью коэффициента *эксцесса (E)* При $E>0$ распределение принято считать островершинным, при $E<0$ — туповершинным. Распределение близко к нормальному, если $E=0$

Ячейка — в электронной таблице — минимальный структурный элемент, имеющий свой адрес В *Microsoft Excel* ячейка может содержать данные в виде текста, чисел, дат, формул или параметров форматирования

В М Зайцев
В Г Лифляндский
В И Маринкин

ПРИКЛАДНАЯ МЕДИЦИНСКАЯ СТАТИСТИКА

Учебное пособие

Лицензия ИД № 01081 от 28 02 2000
ООО «Издательство ФОЛИАНТ»

Редактор *Ю Н Пахомов*
Корректор *Н Д Пылева*
Компьютарная варстка *Н Н Сергиевской*

АДРЕС ИЗДАТЕЛЬСТВА
198020, Санкт-Петербург, Нгревский пр., 18, офис 501
Тел /фгкс (812) 325-39-86; 186-72-36
e-mail: foliant@peterlink.ru

Подписано к печати 25 12 2002
Формат 60×88 $^{1}/_{16}$ Печ л 27
Гарнитура Таймс Печать офсетная
Доп тираж 1000 экз Закгз № 4339

Отпечатано с готовых диапозитивов
в Академической типографии «Наука» РАН
199034, Санкт-Петербург, 9 линия, 12