

В.М. ЗАЙЦЕВ, С.И. САВЕЛЬЕВ

**ПРАКТИЧЕСКАЯ МЕДИЦИНСКАЯ
СТАТИСТИКА**

В.М. ЗАЙЦЕВ, С.И. САВЕЛЬЕВ

ПРАКТИЧЕСКАЯ МЕДИЦИНСКАЯ СТАТИСТИКА

УЧЕБНОЕ ПОСОБИЕ

Под редакцией
академика РАМН, профессора, д.м.н.,
заслуженного деятеля науки Российской Федерации
А.И. Поталова
и профессора, д.м.н. О.Г. Хурцилава

2013

УДК 614.1.
ББК 51.1.(2)
З-17

Зайцев В.М., Савельев С.И. Практическая медицинская статистика: Учебное пособие/ под редакцией академика РАМН, профессора, д.м.н., заслуженного деятеля науки России А.И. Потапова и профессора, д.м.н. О.Г. Хурцилава. – Тамбов: ООО «Цифра», 2013. – 580 с.

Составители: *Тихомирова А.А., Ланько С.В., Сидоров А.А., Гвоздарев С.И., Бондарев В.А., Короткова Т.С., Котова Г.Н., Маринкин В.И., Двоеглазова С.В., Дубровский В.А., Дрожжин Ю.Б., Белов Д.Ю., Камнев Д.А., Зайцев В.В.*

Рецензенты:

И.И. Вишняков, зав. кафедрой общественного здоровья и здравоохранения ГОУ ВПО Санкт-Петербургского государственного медицинского университета им. И.П. Павлова, д.м.н., профессор, заслуженный деятель науки РФ;

В.Я. Гельман, зав. кафедрой информационных процессов и технологий НОУ ВПО БАТИП, д.т.н., профессор.

Рекомендовано Учебно-методическим объединением РАЕ по классическому университетскому и техническому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальностям: 060000 – «Здравоохранение», 060101 – «Лечебное дело», 060104 – «Медико-профилактическое дело».

Настоящее издание является пособием по практическому использованию статистики в медицине. В пособии рассматриваются методики сбора и разработки статистической информации на основе современных технологий обработки больших и малых выборок в электронных таблицах. Детально, на наглядных примерах представлены основные этапы статистического анализа в научных и научно-практических исследованиях, дается оценка информативности источников данных об общественном здоровье, рассмотрены практические алгоритмы расчета статистических показателей в рамках современных возможностей широко распространенных офисных пакетов автоматизированной обработки информации.

Издание предназначено для студентов медицинских вузов, клинических ординаторов, аспирантов, организаторов и исполнителей практических исследований в медицине и здравоохранении.

Табл.151. Ил.136. Библиогр.: 67 назв.

ISBN 978-5-9903085-9-6

© В.М. Зайцев, 2013
© С.И. Савельев, 2013

ОГЛАВЛЕНИЕ

От ответственного редактора.....	7
1. Некоторые вопросы прикладной статистики как элемента современных информационных систем (Тихомирова А.А.).....	9
2. Цель, задачи, план и программа статистического исследования.....	18
2.1. Статистическая совокупность, единица наблюдения, учетные признаки.....	21
2.2. Программа сбора статистической информации. Текущее и одновременное наблюдение. Генеральная и выборочная совокупности. Репрезентативность данных.....	30
2.3. Пути повышения надежности выборок.....	51
2.4. Программа разработки (систематизации) статистических данных.....	58
2.4.1. Классификация статистических группировок.....	62
2.4.2. Статистические таблицы. Правила оформления.....	67
2.4.3. Диаграммы. Правила построения.....	77
2.5. Основы применения MS Excel при разработке статистических данных (Ланько С.В.).....	98
2.5.1. Ввод и редактирование данных.....	100
2.5.2. Выделение блока ячеек.....	101
2.5.3. Ввод математических формул.....	102
2.5.4. Копирование и дублирование данных.....	105
2.6. Базы и хранилища данных. Формирование и обработка таблиц данных (Тихомирова А.А.).....	106
2.6.1. Таблицы баз данных. Пополнение, редактирование, фильтрация и сортировка информации.....	107
2.6.2. Сводные таблицы в MS Excel.....	114
3. Общественное здоровье, как предмет статистического анализа.....	120
3.1. Принципы статистики физического развития.....	122
3.1.1. Центильный метод оценки физического развития.....	130
3.2. Заболеваемость. Источники получения информации (Сидоров А.А., Гвоздарев С.И.).....	133
3.2.1. Данные о причинах смерти.....	138
3.2.2. Обращаемость за медицинской помощью.....	142
3.2.3. Результаты профилактических осмотров.....	150
3.2.4. Использование результатов тестовых опросов.....	157
3.2.5. Заболеваемость с временной утратой трудоспособности.....	166
3.2.6. Стойкая утрата трудоспособности, инвалидность.....	187
3.2.7. Статистика обязательного и добровольного медицинского страхования.....	191
3.2.8. Наиболее важные для общественного здоровья болезни (Бондарев В.А., Короткова Т.С.).....	209
3.3. Медицинская демография (Котова Г.Н.).....	216
3.3.1. Показатели статистики.....	229

3.3.2. Показатели динамики народонаселения.....	238
3.3.3. Смертность населения, средняя продолжительность предстоящей жизни.....	248
4. Основы математико-статистической обработки данных (Маринкин В.И.).....	256
4.1. Относительные величины. Статистические коэффициенты.....	257
4.2. Стандартизация.....	267
4.2.1. Прямой метод.....	269
4.2.2. Косвенный метод.....	272
4.2.3. Обратный метод.....	273
5. Показатели описательной статистики (Двоглазова С.В.).....	276
5.1. Ряды распределения. Вариационные ряды.....	277
5.1.1. Построение вариационных рядов в MS Excel.....	287
5.2. Показатели центра распределения. Средние величины.....	290
5.3. Степенные средние.....	291
5.3.1. Среднее арифметическое. Статистическое взвешивание. Суммирование вариационных рядов и средних.....	294
5.3.2. Статистическое взвешивание. Вычисление некоторых индексов.....	303
5.3.3. Упрощенные способы вычисления среднего арифметического.....	306
5.3.4. Вычисление степенных средних в MS Excel.....	309
5.4. Структурные средние. Мода и медиана.....	313
5.5. Показатели разброса значений рядов наблюдений.....	318
5.5.1. Дисперсия.....	321
5.5.2. Среднеквадратическое отклонение.....	324
5.5.3. Коэффициент вариации.....	327
5.5.4. Квантили.....	328
5.5.5. Использование MS Excel для нахождения квантилей.....	330
5.6. Асимметрия и эксцесс. Статистические моменты.....	334
5.6.1. Описательная статистика в MS Excel.....	336
6. Статистическая проверка статистических гипотез (Маринкин В.И.).....	341
6.1. Оценка статистических параметров выборочных данных.....	351
6.1.1. Точечные оценки. Состоятельность, несмещенность и эффективность точечных оценок.....	351
6.1.2. Интервальные оценки. Доверительная значимость, доверительная вероятность, доверительный интервал, доверительный предел.....	354
7. Статистические распределения.....	360
7.1. Биноминальное распределение и распределение Пуассона.....	361
7.2. Нормальное распределение.....	363

7.2.1. Нахождение нормального распределения с помощью MS Excel.....	367
7.3. Преобразования эмпирических распределений. Логит- и пробит-анализ.....	369
7.3.1. Выполнение пробит-анализа в Ms Excel.....	376
8. Сравнение статистических распределений, нахождение различий (Дубровский В.А.).....	382
8.1. Критерии согласия эмпирических данных и теоретических распределений.....	382
8.1.1. Критерий Пирсона χ^2 (хи-квадрат).....	384
8.2. Критерии согласия эмпирических распределений.....	387
8.2.1. Критерий Мостеллера.....	387
8.2.2. Критерий Розенбаума.....	388
8.2.3. Критерий знаков (Z).....	388
8.2.4. Критерий Вилкоксона U (Вилкоксона – Манна – Уитни).....	389
8.2.5. Критерий гомогенности Брандта-Снедекора.....	390
8.2.6. Модифицированный критерий гомогенности хи-квадрат.....	392
8.2.7. Критерий Колмогорова.....	393
9. Сравнение параметров распределений, нахождение различий.....	396
9.1. Принадлежность варианты к совокупности. «Выскакивающие» варианты.....	397
9.1.1. Определение выскакивающей варианты с помощью MS Excel.....	400
9.2. Сравнение средних арифметических и относительных показателей.....	402
9.2.1. Статистическая проверка различий средних арифметических в MS Excel.....	413
9.3. Сравнение нескольких средних. Модифицированный критерий Стьюдента (Q).....	417
10. Статистика динамики изменений явлений. Динамические ряды.....	418
10.1. Показатели динамического ряда (Дрожжин Ю.Б.).....	419
10.2. Обработка динамических рядов в MS Excel.....	425
10.3. Углубленный анализ динамических рядов.....	426
10.3.1. Показатели сезонности.....	427
10.3.2. Вычисление показателей сезонности в MS Excel.....	429
10.4. Повышение наглядности динамических рядов. Прогноз динамики.....	431
10.4.1. Обработка динамических рядов и прогноз динамики в MS Excel.....	436
11. Статистические оценки связей.....	439
11.1. Оценки статистических связей на принципе ковариации.....	444
11.1.1. Коэффициент ковариации.....	444
11.1.2. Коэффициент линейной корреляции Пирсона.....	446
11.1.3. Коэффициент детерминации.....	451
11.1.4. Коэффициент Фехнера.....	452

11.1.5. Коэффициент ранговой корреляции Спирмена.....	453
11.2. Частная (парциальная) корреляция.....	456
11.3. Множественная корреляция. Корреляционная матрица.....	458
11.4. Нелинейная связь. Корреляционное отношение.....	461
11.5. Оценка взаимосвязи динамических рядов. Автокорреляция.....	465
11.6. Регрессия (Белов Д.Ю).....	468
11.7. Оценки взаимосвязи качественных признаков на принципе взаимной сопряженности.....	480
11.7.1. Коэффициенты ассоциации и контингенции.....	483
11.7.2. Коэффициенты сопряженности.....	485
11.8. Вычисление показателей статистической связи в MS Excel.....	490
11.8.1. Показатели парной корреляции и параметры регрессии.....	490
11.8.2. Параметры многофакторной линейной регрессии.....	503
11.8.3. Критерии сопряженности.....	509
11.9. Дисперсионный анализ (Камнев Д.А.).....	511
11.9.1. Общие принципы.....	511
11.9.2. Дисперсионный анализ в MS Excel.....	520
11.9.3. Методики дисперсионного анализа, основанные на применении функции MS Excel «Пакет анализа».....	520
11.9.4. Методики дисперсионного анализа с использованием сводных таблиц и «Пакета анализа» MS Excel.....	532
12. Оценка различий показателей заболеваемости.....	541
12.1. Типичные ошибки, допускаемые при анализе показателей заболеваемости.....	541
12.2. Параметрические оценки различий показателей заболеваемости.....	544
12.2.1. Альтернативное распределение.....	544
12.2.2. Неальтернативное распределение.....	550
12.2.3. Анализ различий показателей заболеваемости в MS Excel с помощью метода доверительных интервалов.....	554
12.3. Непараметрические критерии оценки различий показателей заболеваемости (Зайцев В.В.).....	554
12.3.1. Простейшие критерии быстрой оценки различий (критерий Мостеллера, Розенбаума, критерий знаков).....	555
12.3.2. Критерий соответствия (согласия) Пирсона χ^2	559
Приложения.....	564
Список литературы.....	575

От ответственного редактора

Существенно расширившиеся за последнее десятилетие возможности информационных технологий создали ситуацию, когда в Российской Федерации использование тех или иных статистических данных становится обязательным элементом повседневной профессиональной деятельности не только математиков, но и политиков, специалистов государственного управления, а также медиков. Повсеместное распространение персональных компьютеров, их современное программное обеспечение, имеющиеся в распоряжении практически каждого специалиста, создало предпосылки значительного повышения научно-методического уровня анализа любой информации. Однако сегодня практическая медицинская статистика развивается в большей степени в ширь, чем вглубь. Во многих случаях уровень статистической обработки данных остается дилетантским. Итогом такого положения является регулярное появление в разного рода, учебных пособиях и руководствах, а также в СМИ информации, отличающейся слабой обоснованностью с точки зрения теории и практики общей статистики. Во многих научно-практических работах, особенно регионального уровня, тоже отмечается явная недостаточность подготовки современных специалистов-медиков в области статистики.

На фоне происходящей в настоящее время бурной информатизации медицины, обеспечивающей оперативное получение, переработку и доступ к детальным данным о здоровье отдельного человека и различных групп населения; информации о достижениях в области клинической и профилактической медицины существует реальная опасность механического, бездумного применения численных характеристик тех или иных процессов, происходящих в общественном здоровье, применения тех или иных средств профилактики и лечения болезней. В связи с этим проблема грамотного

повседневной работе не только научных работников, но и специалистов практического здравоохранения.

Предлагаемая авторами книга посвящена описанию наиболее часто используемых в медицинской практике методов сбора и обработки первичного статистического материала. Книгу следует рассматривать как пособие для работников практического здравоохранения, медико-социальных служб, отраслевых научных учреждений, не имеющих специальной математической подготовки, что может быть чрезвычайно полезным на первом этапе знакомства с применением методов углубленного статистического анализа медицинской информации.

**Академик РАМН, профессор, д.м.н.,
Заслуженный деятель науки
Российской Федерации**



А.И. Потапов

1. Некоторые вопросы прикладной статистики как элемента современных информационных систем

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

В наиболее широком понимании статистика как отрасль знаний представляет собой специальную научную дисциплину. Предмет этой дисциплины – явления, оценка которых производится только в массе наблюдений. Такая ключевая особенность статистики особенно характерна для медицины, поскольку изучаемые медицинской статистикой явления не имеют постоянных, всегда одних и тех же исходов (говоря терминами статистики – весьма вариабельны). Например: масса тела даже у одного и того же человека постоянно меняется, а 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. Френк Йетс (1937) писал по этому поводу: «Большинству биологических объектов свойственна вариабельность, и прелесть простоты и воспроизводимости физических или химических экспериментов утрачивается. А значит, на передний план начинают выдвигаться статистические проблемы».

В общем, по своей природе термин «*статистика*» имеет несколько толкований. Наиболее примитивное из них (и, увы, наиболее распространенное) 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 существование современной статистики обусловлено необходимостью реализации двух потребностей цивилизованного общества. Во-первых, информационного обеспечения системы государственного управления. Во-вторых, возможностью научного прогнозирования развития различных процессов, происходящих в объективной реальности и определенно наблюдаемых только в массе явлений.

Таким образом, в настоящее время под общим термином «статистика» подразумеваются два разных, хотя и в определенной степени взаимосвязанных раздела научно-практической деятельности:

1. статистика как атрибут информационной системы, обеспечивающей государственное управление;
2. статистика вероятностей или математическая статистика, обеспечивающая на основе теории вероятности научное познание и прогнозирование объективной реальности, проявляющейся в массе наблюдений¹.

Как информационная система обеспечения государственного управления *статистика* имеет свои определенные корни. В этом понимании слово «*status*» обозначает состояние (*лат.*). Несомненна связь и с итальянским «*stato*» – государство. Собираение 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 эпоху Возрождения в Италии были чиновники, которых называли «*statisto*» – знаток государства. Как синоним терминов «политическая арифметика» и

¹ Под объективной реальностью здесь подразумевается весь объективно существующий мир, включая и его субъективные продукты.

«государствоведение» (родиной этих терминов считается Англия и Германия) термин «*статистика*» стал употребляться с середины XVII века. В это же время началось отделение из системы государственной или, как ее иногда именовали, «политической статистики» самостоятельной науки, названной «демография». В настоящее время статистика как система гос. учета представляет собой сложную, разветвленную систему взаимосвязанных разделов и дисциплин. Все они имеют в большей или меньшей степени самостоятельное 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 отдельных отраслей промышленности и т.п. Государственная медицинская статистика также является одной из многих отраслевых статистик, как раньше говорили «статистик отраслей народного хозяйства». Отраслевая медицинская статистика, называвшаяся в прошлом «санитарная статистика», имеет свои специфические разделы. В первую очередь это статистика системы здравоохранения (обеспеченность населения 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 заболеваемости и др.

Становление современной математической статистики во многом определялось (и определяется в настоящее время) необходимостью решения самых разнообразных практических задач познания человеком объективной реальности. Отсюда разнообразие профессиональной деятельности авторов, разнообразие широко применяемых, в том числе и в медицине, статистических методов. В качестве примера:

Даниил Бернулли (1700-1782 гг.), один из членов знаменитого в науке швейцарского рода Бернулли. Занимался физиологией, медициной,

математикой и механикой. В 1725-1733 гг. работал в Петербургской академии наук сначала на кафедре физиологии, а затем механики. Автор работ по теории вероятности с приложением к статистике народонаселения России. В ряду знаковых работ стоит и его этапный для медицинской статистики труд «О прививках против оспы. О смерти и теории вероятности».

Томас Байес (Бейес, *Reverend Thomas Bayes*, 1702-1761 гг.), английский математик, пресвитерианский священник. Сформулировал и решил одну из основных задач теории вероятности (теорема Байеса), играющую ключевую роль в современной математической статистике.

Фрэнсис Гальтон (*Francis Galton*, 1822-1911 гг.). Много сделал для развития теории эволюции, предложенной своим кузеном (Ч. Дарвином), доказав, в том числе и с помощью статистических выкладок, что она не противоречит теории божественного происхождения мира. К числу его научных интересов можно отнести антропологию, биологию, генетику. Стал основателем евгеники. Одно из самых ярких его творений – дактилоскопия, базируется на статистических выкладках о вероятностях совпадений узоров отпечатков пальцев. Ф. Гальтон внес большой вклад в развитие статистики корреляционных отношений. Считается одним из авторов регрессионного моделирования, как и самого термина «регрессия».

Чарльз Эдвард Спирмен (*Charles Edward Spearman*, 1863-1945 гг.), психолог, доктор философии. Один из разработчиков общей методологии статистического корреляционного и факторного анализа.

Карл Пирсон (*Karl (Carl) Pearson*, 1857-1936 гг.), один из основателей математической статистики (кривые Пирсона, распределение Пирсона, критерий согласия Пирсона (*хи-квадрат*), корреляционный анализ, коэффициент линейной корреляции Пирсона и др.), биолог и философ. Много усилий приложил для прикладного применения статистических методов в биологии и медицине.

Уильям Сили Госсет (1876-1937 гг.), более известный под псевдонимом *Стьюдент*. Математик, работник лаборатории пивоваренного завода, автор

широко применяемой при оценке различий t статистики (*t*-Стьюдента). С 1935 г. занимал должность главного пивовара в одной из пивоварен Северо-Западной части Лондона.

Фишер (*Ronald Aylmer Fisher*, 1890-1962 гг.), известный биолог, генетик. Один из основоположников современной математической статистики. Основатель фундаментального метода статистики – дисперсионного анализа.

Фрэнк Вилкоксон (*Frank Wilcoxon*, 1892-1965 гг.), доктор философии, химик, разработал ряд широко применяющихся для решения прикладных задач в медицине и биологии статистических критериев (критерий Манна – Уитни – Вилкоксона, критерий Вилкоксона).

Адольф Кетле (*Lambert-Adolph-Jacques Quetelet*, 1796-1874 гг.), известный астроном, метеоролог, социолог. По праву считаюсь одним из родоначальников прикладной медицинской и психосоциальной статистики, рассматривал наблюдаемые явления в жизни людей как проявления статистических законов. Первым показал, что человеческие деяния, подобно явлениям физического мира, подчинены известным закономерностям. Увлечшись математико-статистическими методами в работе «О человеке и развитии его способностей», он вывел тип среднего человека, наделенного наряду со средними показателями физического развития (рост, вес) средними умственными способностями и средними моральными качествами.

В современной медицинской статистике нередко выделяют статистику клиническую, онкологическую, статистику инфекционной заболеваемости, заболеваемости особо опасными инфекциями и т.д. Это многообразие определяется многообразием медицины как науки и разнообразием видов конкретной практической деятельности медиков. Здесь явно прослеживается некоторое чрезмерное увлечение развитием такого рода «статистической номенклатуры», поскольку все разделы медицинской статистики тесно между собой взаимосвязаны, имеют единую методическую основу и их деление в подавляющем числе случаев является весьма условным.

Говоря о современной статистике, специалисты обычно опираются на то обстоятельство, что многие хаотичные, на первый взгляд, явления зачастую вполне упорядочены и поэтому могут иметь вполне конкретные численные оценки. Главное условие – регулярность изучаемых явлений, так называемая «статистическая устойчивость» явлений, то есть существование строго определенных закономерностей, пусть даже скрытых на первый взгляд, которые можно описать методами вероятностной (математической) статистики. Именно поэтому теория вероятности заняла прочное место в медицинской статистике, делая сегодня невозможным статистическое исследование объектов без применения основных концепций этой теории.

Необходимо помнить, что большинство методов статистического анализа являются универсальными и могут применяться в самых разнообразных отраслях человеческой деятельности. *Например*, с точки зрения формальной логики, статистический прогноз динамики инфекционной заболеваемости и прогноз динамики курса доллара – во многом одна и та же задача. В настоящее время, для того чтобы подчеркнуть особую медико-биологическую направленность тех или иных методов математической статистики, часто используют термин – «**биометрия**».

Окончательный выбор конкретных методик и средств их реализации при решении тех или иных научных и практических задач зависит от многих обстоятельств, не последним из которых является уровень подготовки специалиста-исследователя. Возможные области практического применения и доступность методов глубокого статистического анализа особенно резко возросли за последние 2 десятилетия в связи с бурным ростом парка персональных компьютеров и его насыщением программными средствами статистической обработки данных.

Все программные средства, которые могут использоваться для статистической обработки данных на персональных компьютерах, можно разделить на:

- узкоспециализированные пакеты. Применяются для решения узкого круга задач с использованием специальных методов статистического анализа. Эксплуатация этих программ подразумевает высокий уровень подготовки пользователя в области определенных разделов знаний (*например*: ядерная физика);

- статистические программные пакеты общего назначения (*например*: *statistica*). Они более доступны и могут использоваться широким кругом специалистов различного профиля, имеющих подготовку в области статистических методов. Их недостатком можно считать немалую стоимость, которая плохо окупается при нечастом использовании (результаты наблюдений появляются далеко не каждый день), и необходимость специальной подготовки пользователей этих программ, что тоже связано с определенными тратами времени и средств;

- табличные процессоры и электронные таблицы (*например*: *Microsoft Excel, OpenOffice* и др.). Имеют возможности обработки данных распространенными методами описательной статистики, сглаживания динамических рядов, прогнозирования, регрессионного анализа и т.п. Повсеместно распространены среди пользователей персональных компьютеров. Чрезвычайно широко используются при повседневном решении самых различных прикладных задач. Недостатки этих программных продуктов (менее обширная информация, описывающая результаты статической обработки, отсутствие многомерной статистики, относительно малый набор методов непараметрической статистики) во многом окупаются возможностью накопления обширных баз (хранилищ) данных, возможностью повсеместного оперативного применения методов математической статистики и достаточностью результатов при умелом использовании приемов статистического анализа.

Особенностью любого пакета статистических программ является выдача большого количества информации, описывающей результаты стат. анализа. В этой ситуации пользователь, не имеющий достаточной подготовки, зачастую

неправильно сформулировавший задачу, не способен получить адекватный результат, правильно воспринять полученную информацию. Попытки решать такого рода проблемы с помощью различных сопроводительных описаний пакетов прикладных программ, как правило, оказываются безуспешными, так как эти описания, обычно создаются для людей, уже имеющих существенную подготовку в математической статистике. При этом ситуация усугубляется особенностями не всегда удачного перевода (чаще всего с английского языка) с применением малораспространенных в отечественной статистике терминов и т.п.

В завершение следует подчеркнуть, что любая прикладная статистика является лишь методом, инструментом познания объективной реальности. Успешность использования этого инструмента может быть гарантирована только неразрывным сосуществованием двух факторов. Во-первых, глубоким знанием и пониманием исследователем сути изучаемой проблемы. Во-вторых, владением методикой получения и использования численных статистических оценок в качестве доказательной базы. Умением выбирать из обширного арсенала современных средств обработки данных те методы статистики, которые необходимы именно для конкретного исследования.

В медицине, как не в одной другой отрасли творческой деятельности человека, недопустимыми ошибками являются и отказ от статистических методов обработки данных наблюдений, и неправомерное, бездумное использование полученной статистики, в том числе в качестве единственного элемента доказательной базы.

Владение основами практической статистики, изложенными в издании, может служить первым этапом детального изучения статистики и овладения сложными методиками глубокого статистического анализа результатов медико-биологических исследований.

Настоящее пособие дает основы знаний, необходимых для понимания и практического использования элементарных методов статистического анализа, широко применяющихся в медицине. При этом разъясняются основные

термины и постулаты общей теории статистики. Рассматриваются примеры применения для статистического анализа возможностей *Microsoft Excel*, который является составной частью одного из самых распространенных в цивилизованном мире пользовательских пакетов ПК – *Microsoft Office*. В издании рассмотрены методики организации сбора, сводки, первичной статистической обработки и группировки данных, способы повышения наглядности статистического материала, наиболее широко применяющиеся в медицинской науке и практике. Разъясняются доступные для лиц, не имеющих глубокой специализированной подготовки, основы практической математико-статистической обработки данных.

При подготовке издания использовался практический опыт разработки учебно-методических пособий по медицинской статистике, накопленный сотрудниками кафедры медицинской информатики и физики Северо-Западного государственного медицинского университета им. И.И. Мечникова, Федерального научного центра гигиены им. Ф.Ф. Эрисмана, Управления Роспотребнадзора по Липецкой области и ФБУЗ «Центр гигиены и эпидемиологии в Липецкой области».

2. Цель, задачи, план и программа статистического исследования

Традиционно считается, что выполнение статистических исследований производится строго по определенным этапам. Собственно говоря, любое серьезное исследование всегда имеет этапный характер. Такая последовательность действия крайне необходима из-за того, что этапы статистического исследования тесно взаимосвязаны. Ошибки и просчеты на любом из этапов неизбежно ведут к возникновению проблем на других этапах (если не провалу) исследования. Причиной ошибок в самом начале исследования чаще всего бывают два обстоятельства. Во-первых, во всяком научном исследовании имеется элемент неопределенности, который трудно предугадать и спланировать. Во-вторых, на практике детальное планирование хода статистического анализа (статистического исследования) данных начинается, как правило, тогда, когда уже есть некоторые результаты наблюдений и вырисовываются «в первом приближении» определенные рабочие гипотезы. И создается впечатление, что нужно только вписать эти результаты в существующие методики статистического анализа, снисходительно относясь к требованиям репрезентативности исследования и т.п. деталям, иногда ставящим перед необходимостью начинать все сначала.

Следует отметить, что существуют исследования, которые по своей сути носят ярко выраженный статистический характер (медико-демографические исследования, изучение заболеваемости, оценка работы медицинских и социальных учреждений и т.п.), начало которых без планирования невозможно в принципе. В этих случаях рекомендуется проводить предварительные, пробные, «пилотажные» исследования, позволяющие на практике уточнить конкретные технологии сбора и анализа информации.

Классическая схема статистического исследования включает следующие этапы:

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

2-й этап – наблюдение (сбор информации), сводка и группировка полученных данных. Вычисление первичных итогов;

3-й этап – углубленная математико-статистическая обработка данных;

4-й этап – анализ полученных результатов, выводы.

На I-м этапе статистического исследования определяют:

- что делать (т.е. цели и задачи исследования). Последующий выбор конкретных технологий стат. анализа и его ход напрямую определяются целью и задачами исследования;

- как делать (программа исследования);

- кто, когда и за чей счет будет все делать (план исследования, иногда говорят «обеспечение исследования»).

Цель исследования обуславливается конечным результатом, на достижение которого направлено исследование, то, ради чего оно проводится. Цель большинства статистических исследований в медицине – раскрытие взаимосвязи и оценка влияния на здоровье человека тех или иных *факторов*. Под ними здесь могут пониматься методы лечения, фармакологические препараты, средства специфической и неспецифической профилактики, условия труда, быта, образ жизни и т.п. Необходимо помнить, что главным методологическим ядром любого статистического исследования является **статистическая оценка распределения** численных характеристик наблюдаемых явлений, объектов и т.п. При этом подразумевается, что статистическое исследование – инструмент научного познания, применяемый в случаях, когда добраться до истины (до конечной цели исследования) невозможно иначе, как изучая явления **в массе наблюдений**.

Задачи исследования отражают частные вопросы, которые необходимо последовательно решить, чтобы достигнуть конечной цели исследования. Например, если целью исследования является изучение (оценка) влияния неблагоприятных производственных факторов на здоровье работников предприятия, то целесообразно предусмотреть решение (как минимум) следующих задач:

- изучить социально-демографический состав работников;
- изучить санитарно-гигиенические оценки условий и характера труда на предприятии;
- по данным целевых и периодических медицинских осмотров выявить уровень патологической пораженности работников;
- выявить и оценить взаимосвязь неблагоприятных факторов производства и патологической пораженности работников основных профессий.

При определении конкретных задач обязательно учитывается главный принцип любого статистического исследования – достоверность (репрезентативность) исходной информации, в том числе и элементарная объективность отражения в регистрируемых статистических характеристиках 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 исследовательских кадров, подготовка необходимых аппаратных и программных средств, регистрационных бланков и т.п. Для детальной проработки организационных и методических вопросов целесообразно предусмотреть проведение пробных, ограниченных по численности (но не по объему регистрируемой на каждую единицу наблюдения информации), так наз. «пилотажных» исследований.

Программа исследования включает *программу сбора* и *программу разработки* материалов исследования. Программа сбора определяет объект исследования, единицы наблюдения, а также перечень учетных признаков, подлежащих регистрации в ходе исследования. Здесь следует обратить внимание, что, когда говорят о программе статистического исследования, имеют в виду не те или иные компьютерные программы, а только алгоритм регистрации (включая способ регистрации и перечень учетных признаков), т.е. – *программу сбора*, а также последовательность и методику статистического анализа – *программу разработки*.

2.1. Статистическая совокупность, единица наблюдения, учетные признаки

Статистическая совокупность, объект наблюдения – совокупность предметов и явлений, избранная для статистического наблюдения.

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*.

Другой вариант: при изучении эффективности гипотензивного препарата общий объект наблюдения – группа пациентов, которая состоит из двух подгрупп: основной (пациенты, принимающие препарат) и контрольной (его не принимающие). Эти подгруппы могут различаться только в одном – пациенты принимают препарат или нет.

Учетные признаки – частные характеристики отдельных единиц наблюдения, подлежащие регистрации в ходе статистического исследования (в приведенных примерах – пол, возраст, профессия работника, уровень артериального давления, рост, масса тела пациента и т.п.). Особенностью большинства учетных признаков в медико-биологических исследованиях является их вероятностный, случайный характер. Случайный характер учетных признаков определен индивидуальными особенностями отдельной единицы наблюдения или наблюдаемой группы. *Например,* уровень артериального давления даже у одного здорового человека может колебаться в определенных пределах, поэтому данные его регистрации носят в определенной степени случайный характер. Величина роста и масса тела человека изменяются 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, хотя и в заранее определенных границах.

Основой объективности статистического анализа является шкалирование учетных признаков. В основе шкалирования лежит учет **вида учетного признака**. Учетные признаки по виду могут быть качественными или количественными.

Качественные признаки, иногда называемые «атрибутивные» или «номинальные», характеризуют учетный признак с помощью слова (словесных образований) или условных знаков. *Например*: пол мужской или женский; образование начальное, среднее, высшее; диагноз заболевания и т.п.

Количественные признаки – признаки, значения которых имеют числовое выражение. *Например*: рост – см, масса тела – кг, частота пульса – число уд/мин и т.д. Весьма существенной особенностью количественных признаков является их деление на *непрерывные* и *прерывные (дискретные)*. В наиболее простом виде различия между ними выглядят следующим образом. *Прерывные (дискретные)* величины являются результатом прямого счета (число ударов пульса, число эритроцитов, число врачебных посещений и т.п.). Выражаются только целыми числами. *Непрерывные* – являются результатом измерения. Их величина может выражаться дробным числом и зависит от точности измерения. Необходимо избегать чрезмерной точности измерения, так как это приводит не столько к повышению точности результатов, сколько к росту числа разного рода ошибок и отсюда – к снижению итоговой точности. *Например*: точность измерения массы тела взрослого человека более чем в 0,5 кг для характеристики его физического развития не имеет существенного значения, даже если приурочить измерение массы тела к определенному моменту времени суток (что, кстати, практически не делается). Кроме того,

последующая группировка исходных данных неизбежно уменьшит исходную точность замеров. В общем, здесь действует известный принцип: «Чрезмерная точность – признак безграмотности».

Некоторые признаки можно представить 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 такого рода данных необходимо учитывать цели конкретного исследования, а также программу дальнейшей статистической обработки, поскольку для каждого вида признаков существуют свои правила обработки. *Например*, даже простейшая арифметическая обработка качественных данных не допускает расчета средних величин (средний пол, средняя стадия заболевания). Здесь допустимы либо арифметические операции сложения эквивалентных единиц (сложение всех случаев одинаковых заболеваний), либо расчет относительных величин (показатели структуры, частоты, распространенности, соотношения и т.д.). Аналитическая оценка совокупности качественных и количественных признаков проводится только после перевода количественных признаков в качественные. *При этом следует помнить, что перевод количественных признаков в качественные возможен. Обратный перевод качественных признаков в количественные нет!*²

Забвение основных различий между качественными и количественными учетными признаками приводит к существенным ошибкам. Например, 1-я, 2-я или 3-я группы диспансеризации имеют цифровое выражение, однако это не меняет их качественной сути. Средняя группа диспансерного наблюдения не имеет здравого смысла. Вместе с тем, попытки обозначить среднюю группу диспансерного наблюдения неоднократно встречались в печати. Более

² Иногда, чтобы ввести качественные переменные в различные статистические модели, необходимо качественные переменные преобразовать в количественные. Такие сконструированные переменные принято называть *фиктивными переменными*. Например, фактор «пол» в виде фиктивной переменной представляется в виде: мужчины – 1, женщины – 2. Однако допустимость применения в стат. анализе фиктивных переменных является сомнительной.

привычное, почти повсеместное некорректное использование арифметических операций с качественными признаками встречается при применении балльных оценок – выставление среднего балла, если речь идет о разных единицах наблюдения. Но сумма баллов двух школьников-двоечников не свидетельствует о равенстве суммы их знаний знаниям одного четверочника. С точки зрения формальной статистики, здесь можно говорить только о распределении тех или иных баллов в группах (частота встречаемости тех или иных оценок, % отличников, троечников в группе и т.п.).

Для повышения информативности представления учетных признаков могут использоваться более детальные характеристики шкал. Различают следующие типы шкал: номинальная, порядковая, интервальная (относительная) и абсолютная шкалы.

Качественной шкалой самого низкого уровня является *номинальная шкала*, в рамках которой единицам наблюдений дается некоторое название (название признака). С формальной точки зрения разные объекты по этим названиям либо совпадают, либо различаются по качеству. Никакие другие, более тонкие соотношения при статистической обработке между ними не учитываются. Примерами номинальных признаков могут служить диагнозы заболеваний, их коды, имена лиц, их пол и т.п. Единственная цель такого шкалирования – учет различий между объектами разных классов. *Например*: или объект $A=B$, или $A \neq B$.

В номинальных качественных шкалах могут применяться для условных обозначений символы 1, 2, 3,... и т.п. Но служат они лишь для обозначения качественных различий объектов. Статистическая обработка результатов наблюдений, сгруппированных в номинальной шкале, должна учитывать эти особенности. В противном случае могут быть сделаны ошибочные оценки, не соответствующие действительности. С помощью подсчета зафиксированных по номинальной шкале признаков можно установить:

- число различных классов: число больных с тем или иным диагнозом; численность мужчин и женщин, проживающих в населенном пункте; число лиц, принимающих тот или иной препарат, и т.п.;

- частоту встречаемости объектов и их сравнительное распределение (заболевания, встречающиеся наиболее часто; вид препарата, производством которого занимается наибольшее число предприятий, и т.п.).

К числу качественных шкал относится и *порядковая, ординальная или ранговая шкала*. Она указывает помимо различий и на сравнительную степень выраженности качественного признака у разных единиц наблюдения, что позволяет получать упорядоченные распределения данных.

Например, больных с одним диагнозом можно ранжировать по стадии заболевания или школьников – по баллам успеваемости. В последнем случае порядковая шкала имеет величины от 1 до 5, и учащимся присваивается тот или иной балл в зависимости от способности правильно выполнять задания. Интервалы между баллами, разделяющими школьников, условны. Примером порядковых данных являются и стадии заболевания. Их хоть и можно выразить в числовой шкале (например: 1, 2 или 3), но это шкалирование носит качественный характер. Нельзя считать, что разница в последствиях между 2 и 3 стадией болезни такая же, как и между 1 и 2.

По этой причине нецелесообразно складывать, вычитать, умножать и делить баллы, обозначающие лишь порядковые места. Порядковые данные можно ранжировать и использовать результаты этого ранжирования только там, где это определено самой сутью конкретного метода, *например*, при вычислении ранговых коэффициентов корреляции.

При использовании порядковых шкал в рамках одного исследования (программы одного исследования) целесообразно соблюдать единый подход к порядку нумерации рангов. Т.е. либо возрастание балльных оценок соответствует росту качества любого из регистрируемых учетных признаков. Либо наоборот – возрастание балльных оценок соответствует ухудшению качества. *Например*, можно обозначать состояние здоровья обследованных лиц

в виде балльных оценок 1, 2, 3 и т.д., где больший номер оценки соответствует ухудшению показателей здоровья. То есть 1 группа – здоровые лица, 2 группа – практически здоровые, 3 группа – лица, нуждающиеся в консультации специалиста, и т.д. В этом случае и оценка других учетных признаков, регистрируемых в данном исследовании в виде балльных шкал, должна соответствовать ухудшению качества. *Например*, лица, не употребляющие алкоголь, – 1 группа; употребляющие эпизодически – 2, часто – 3 и т.д.

Соблюдение такого правила позволит избежать ошибок в трактовке статистических оценок взаимной сопряженности изучаемых явлений, а также избежать излишних сложностей и противоречий при моделировании различных вариантов проявлений выявленных закономерностей.

Более информативная шкала, относящаяся в отличие от предыдущих к количественным шкалам, – *шкала интервалов*, которая, кроме упорядочивания обозначений, позволяет оценить интервал между ними и выполнять некоторые математические действия над этими интервалами. Разновидностью шкалы интервалов является шкала разностей, в которой можно определять только разность. Например, температура у одного больного (37 C^0) на 1 градус меньше температуры другого (38 C^0). Но сумма температур больных или средняя температура больных в стационаре – величины несуразные.

Самая информативная шкала – *абсолютная* (число столов, стульев, коек, больных и т.п.). В ней можно выполнять любые математические действия без каких-либо ограничений.

Существование перечисленных различий в шкалировании обуславливает невозможность единой статистической разработки признаков, значения которых фиксированы в разных шкалах. Здесь требуется приводить все шкалы к единому виду, что достигается понижением уровня более высоких шкал до уровня самых низких. При этом приведение всех шкал, применяемых в конкретном исследовании, к единому виду путем повышения уровня шкал низкого уровня невозможно в принципе.

По роли в статистической совокупности учетные признаки можно подразделить на факторные (факториальные) и результативные (результатирующие). *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 органов дыхания у сварщиков – результативный признак. Рольевая значимость факторных и результативных признаков зачастую бывает весьма сложной. Кроме того, бесспорная, на первый взгляд, взаимосвязь каких-либо факторов и результатов на самом деле может быть обусловлена взаимодействием других, не учитывавшихся в данном исследовании явлений.

Здесь следует отметить одно принципиально важное обстоятельство: только исследователь, глубоко знающий суть изучаемой проблемы, определяет истинные причинно-следственные отношения между связанными явлениями. Статистика дает лишь формальное описание связей и только в пределах известных современной науке характеристик и видов связей!

Все отдельно взятые единицы наблюдения, относящиеся к одной статистической совокупности, имеют некоторое число общих учетных признаков (характеристик), свидетельствующих о принадлежности конкретной единицы наблюдения к этой совокупности. Такие признаки называются *признаками сходства* (место работы, время работы на предприятии, место жительства, диагноз заболевания, характер лечения и т.п.).

Признаки различия представляют индивидуальные особенности каждой единицы наблюдения. В медицинских исследованиях это могут быть: диагноз заболевания, тот или иной биохимический показатель крови, уровень артериального давления и т.п. Строго говоря, признаки различия и являются конечным объектом статистического исследования. Отсутствие признаков различия в отдельных единицах наблюдения делает бессмысленным

проведение любого исследования. Понятно, что один и тот же признак может в определенной ситуации быть признаком сходства, в другой – признаком различия.

Выбор единицы наблюдения и учетных признаков, а также правила их регистрации определяют весь ход и итоги статистического исследования, начиная с формирования статистической совокупности и кончая выбором метода статистической обработки и логики анализа. *Например*, при изучении заболеваемости, по данным амбулаторно-поликлинических учреждений, за единицу наблюдения обычно принимают *случай заболевания*, выявленный при обращении за медицинской помощью в поликлинику или по результатам профилактических медицинских осмотров. Несмотря на то, что учетные признаки при этом обычно схожи (диагноз, пол, возраст заболевшего и т.п.), результаты наблюдения в первом и во втором вариантах получения информации будут принципиально разными. Большинство причин обращений за медицинской помощью в поликлинику обусловлены острыми заболеваниями. Соответственно, в итогах будут превалировать именно эти болезни. При профилактических осмотрах выявляются, в основном, хронические болезни (больной с острой респираторной инфекцией не идет на медосмотр). Кроме того, зависят результаты этих осмотров и от наличия тех или иных врачей-специалистов в составе медицинских комиссий, и от их квалификации, и от уровня доступности сложного диагностического оборудования и т.п.

В настоящее время имеют место разные подходы к классификации типов медико-биологических исследований в зависимости от характера формирования статистической совокупности данных. Это, в свою очередь, породило существование различных классификаций и разных, неоднозначных толкований ключевых терминов в таких классификациях. При этом следует помнить, что здесь само существование различных классификаций не носит схоластического, беспредметного характера, поскольку результаты исследования одного класса исходных данных могут отличаться от исследования данных другого класса. Эти

результаты, оставаясь, по большому счету, одинаковыми, объективно могут существенно отличаться друг от друга в деталях.

Согласно одной из существующих ныне классификаций, в соответствии с тем, как учитывается фактор времени при сборе информации, исследования подразделяются на две группы: *продольные и поперечные*³. Поперечные исследования – сбор информации происходит на определенный момент времени (моментные наблюдения), что не исключает логическую реконструкцию развития причинно-следственных отношений во времени. Продольные исследования (наблюдения), часто называемые в отечественной статистике интервальными наблюдениями, разнесены во времени, т.е. осуществляются на основании данных, накопленных за определенный промежуток времени. Кроме того, нередко клинические исследования, в зависимости от способа формирования статистической совокупности, могут классифицироваться как *ретроспективные или проспективные*. В ретроспективных исследованиях ситуация изучается после того, как заболевание возникло, т.е. на заранее отобранных группах больных. Проспективными, как правило, считаются исследования, в которых анализируется ситуация до того, как возникло заболевание. Например, в проспективном исследовании сравнивается профилактическая эффективность гипотензивного препарата среди лиц, не имеющих гипертонической болезни, но относящихся к группам повышенного риска. В ретроспективном сравнивается эффективность гипотензивного препарата отдельно по группам уже сформировавшейся патологии: 1 группа – больные с 3 стадией гипертонической болезни, 2 группа – лица со 2 стадией гипертонической болезни и т.п. Понятно, что в первом случае речь идет только о лицах с той или иной вероятностью возникновения заболевания. Во втором рассматриваются уже состоявшиеся больные. Соответственно, в первом случае на результат оценки профилактической эффективности препарата будет оказывать

³ В.В. Власов «Эффективность диагностических исследований». – М.: Медицина, 1988. – С. 256.

существенное влияние точность определения вероятности возникновения болезни (неизбежно снижая в большей или меньшей степени видимый эффект применения профилактического препарата). Во втором случае этот фактор не будет играть никакой роли.

2.2. Программа сбора статистической информации.

Текущее и единовременное наблюдение. Генеральная и выборочная совокупности. Репрезентативность данных

В обобщенном виде программа сбора статистической информации – это перечень учетных признаков наблюдения, которые позволяют достаточно полно характеризовать каждую единицу наблюдения и факторы изучаемых явлений. Конкретное воплощение этой программы (не путать с компьютерной программой!) – перечень учетных признаков, содержащихся в формуляре статистического наблюдения. Таким формуляром могут быть анкета опроса, регистрационная карта наблюдения и т.п. Формуляр статистического наблюдения является ключом всего исследования. В правильно составленном формуляре есть все признаки, существенные для данного исследования. В нем не должно быть громоздких, малосодержательных, неинформативных 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 даже при автоматизированном вводе данных. Малая вероятность ошибок ввода данных в компьютер дает возможность быстрой обработки и интерпретации данных. Для снижения числа ошибок и повышения скорости обработки данных в настоящее время все чаще используются специальные аппаратные средства автоматизации регистрации информации (автоматические регистраторы физиологических параметров, сканеры ввода анкетных данных и т.п.). Следует отметить, что отказ от «твердых» формуляров путем прямой регистрации информации только в электронных базах данных более оправдан в условиях обработки больших массивов информации, когда её сохранность

гарантируется специальными приемами. При проведении относительно небольших по объему наблюдений, что обычно для медико-биологических исследований, особенно клинических, когда объем наблюдений определяется несколькими десятком или сотней единиц, громоздкая система защиты информации на одном электронном носителе не всегда экономически оправдана или затруднена по разным причинам. В любом случае хранение исходных данных только на одном носителе информации представляется весьма рискованным.

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

Непосредственное наблюдение, точнее – *непосредственная регистрация*, предполагает регистрацию данных «в натуре», в том числе и вручную на специальные бланки, и путем записи с помощью технических средств. *Например*, компьютерная регистрация результатов измерения жизненной емкости легких, форсированного выдоха, параметров кардиограмм и т.п. *Выкопировка данных* из отчетно-учетной документации предполагает использование в виде источника информации различных документов или баз данных (истории болезни, истории развития ребенка, больничные листы, базы данных страховых компаний). Этот способ получения информации требует обязательной предварительной экспертной оценки наличия информации в полном объеме, правильности заполнения и полноты записей. При этом имеется в виду, что, несмотря на обычно декларируемую полную регистрацию данных, информация из официальных источников нередко имеет серьезные дефекты, появление которых связано с недоброкачественной работой регистраторов, особенностями местных условий, традициями, экономической ситуацией и т.п. Кроме того, некоторые источники в силу своей специфики постоянно обновляются (*например*, базы данных о составе больных в больницах). А это обновление

сопряжено с удалением устаревшей информации и заменой ее на новую. Понятно, что такая особенность хранения информации накладывает определенный негативный отпечаток на полноту данных при продольных исследованиях.

Опрос обеспечивает получение информации со слов опрашиваемого (респондента) методом интервью или заочным путем (почтовые, телефонные, прессовые опросы). Методам опроса посвящено немало специальных исследований, методической литературы и пособий (Никандров В.В., 2002; Ильин Е.П., 2002 и др.). Следует отметить, что в РФ к настоящему времени действует ряд законов, прямо или косвенно касающихся проведения опросов общественного мнения, референдумов и т.п., а также об ответственности за сохранность информации, затрагивающей личные интересы граждан. Знакомство с этими законами и обязательность их соблюдения – обязательное требование сегодняшнего дня.

Что касается основных аспектов организации опроса, то с точки зрения статистической обработки данных опроса следует остановиться на 2-х главных аспектах. Во-первых, регистрация информации, полученной путем опроса, обязательно производится на специально разработанные, отдельно для каждого исследования, опросные листы или анкеты. Во-вторых, для качественного проведения опроса рекомендуется привлекать специалистов по опросам. Выбор метода опроса диктуется организационными и финансовыми возможностями. При этом любому из существующих методов присущи определенные недостатки. В частности, интерактивные телевизионные опросы и т.п. опросы, которыми решаются в основном журналистские задачи, к способам получения статистически достоверной (качественно репрезентативной) информации отнесены быть не могут.

Наиболее точный, но дорогостоящий вариант – опрос методом интервью. Для такого опроса, когда объем наблюдений исчисляется сотнями человек, требуется иметь специально подготовленных специалистов-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 формальных и неформальных лидеров его окружения (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, соответствие программы статистической обработки целям, финансовым и организационным возможностям исследователей и т.п.

Для получения оптимальных результатов посредством любого вида опроса необходимо придерживаться следующих требований к содержанию опросного листа (анкеты):

- *простота и доступность изложения вопросов и вариантов ответов (если они предусмотрены) с учетом 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 должны быть предусмотрены контрольные вопросы;
- идентификационные данные (возраст, место жительства и т.п.) желательно относить в конец опроса;

- оптимальность объема опроса (анкеты). С одной стороны, размер анкеты должен обеспечивать получение максимума информации. С другой – анкета не должна быть утомительной для заполнения, т.к. усталость респондента и опрашивающего резко повышает число неустраняемых ошибок и затрудняет 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 вопроснике вся информация, записанная в большом количестве пунктов, очень часто может быть заменена крайне ограниченным набором тривиальных вопросов и 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 (механических) неточностей в ответах респондентов и систематических ошибок регистраторов. В итоге надежность такой программы сбора данных, как правило, оказывается за пределами допустимого уровня.

Здесь действует общая для любого статистического исследования закономерность: всякое необдуманное увеличение объема регистрируемой информации, помимо удорожания и увеличения времени обработки, неизбежно ведет к росту числа погрешностей, резко снижающих ценность исследования.

Частично уменьшить число этих погрешностей, в статистике называемых систематическими ошибками, 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. Однако даже при использовании высококвалифицированного персонала (что на практике выглядит весьма фантастично) при регистрации больших объемов информации все равно неизбежно накапливается масса систематических ошибок.

Вообще чрезмерное увеличение объема исходной информации для статистического исследования, 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, называемых почему-то систематическими, ошибок регистрации и ошибок, связанных с *вариабельностью* исходной информации, которая обусловлена большой изменчивостью, случайностью значений регистрируемых характеристик медико-биологических процессов, точностью методик сбора информации (измерительной аппаратуры) и др.

Достигая известного предела, такой «информационный шум» просто подавляет искомую исследователем информацию, делает ошибочными результаты применения любых, в том числе и самых изощренных методов статистической обработки. При этом использование высокоточных и совершенных приборов и аппаратуры не является панацеей от упомянутого «шума». Считается, что именно это обстоятельство неустойчивости, неточности результатов является одним из главных факторов недоверия к статистике. Известно высказывание, связанное с недоверием к статистике, приписываемое или М. Твену, или У. Черчиллю, которое выглядит в несколько перефразированном виде следующим образом: «Ложь бывает большой, маленькой и статистической».

Существует несколько методов повышения устойчивости результатов статистических данных (*робастности*⁴ статистических оценок, критериев и т.п.). Среди них наиболее технически простой – *винзорирование* исходных

⁴ Основы методологии получения робастных оценок были разработаны отечественными учеными А.Н. Кодмогоровым, Н.В. Смирновым и Б.С. Ястремским.

данных, т.е. отбрасывание 1-2 крайних измерений, или замена этих значений максимальным значением из оставшихся измерений. Сторонники этой спорной методики основывают правомерность ее применения тем, что обычно наибольшее число систематических ошибок (более 90%) скапливается в районе крайних значений.

Строго говоря, неустойчивость (вариабельность) исходных данных складывается из нескольких составляющих. Одной из самых существенных, с точки зрения возможной утраты информационной ценности итогов статистической обработки, является *аналитическая вариабельность*. Она возникает из-за расхождения между результатами измерений даже одной единицы наблюдения. 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, но отличаться. Близка к ней по своему смысловому значению *внутрииндивидуальная вариабельность*, которая характеризует объективное расхождение значений какого-либо параметра у одного и того же объекта наблюдения. Например: масса тела у одного и того же человека в течение суток изменяется.

Межиндивидуальная, или внутригрупповая вариабельность характеризует неизбежное расхождение результатов замеров какого-либо параметра у разных людей – например, массы тела, причем и у лиц, имеющих другие одинаковые параметры физического развития.

Межгрупповая вариабельность в статистике обычно является результатом целевой группировки (например, различия контрольной и основной групп наблюдения). Изучение такой вариабельности, собственно говоря, и является целью статистического исследования.

Статистическая оценка соотношения разнообразных вариаций лежит в основе специального метода статистического анализа – **дисперсионного анализа**.

Принято считать, что высокую вариабельность медико-биологических данных можно снижать, используя специальные, порой весьма громоздкие,

методики регистрации измеряемых параметров (стандартизация условий регистрации, многократная регистрация замеров, более «тонкие» средства измерения и т.д.), а также употребляя специальные приемы статистической обработки. Однако результативность таких операций остается во многих случаях невысокой. В лучшем случае удастся уменьшить один из компонентов общей вариабельности (чаще всего обусловленный механическими, систематическими ошибками измерения).

Наиболее эффективным методом снижения «излишней» вариабельности обследуемых групп (выборок) является использование парносопряженного метода. *Например*, в эксперименте сравниваются наблюдения одной и той же группы до и после воздействия изучаемого фактора. Таким способом исключается часть внутригрупповой вариации, связанной с индивидуальными различиями субъектов, поскольку рассматриваются просто разности между двумя измерениями («до» и «после») отдельно для каждой единицы наблюдения. Группы, формируемые и обследуемые таким способом, принято обозначать как *зависимые выборки*. Именно для таких случаев и проводится вычисление разного рода статистических критериев оценивания различий 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. Однако значительное число клинико-физиологических параметров не укладывается 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. Другой пример: вариации параметров форсированного выдоха в течение дня у здоровых лиц составляют около 25% от межиндивидуальных. У лиц с хроническими заболеваниями легких это соотношение достигает 50% и более (Власов В. В., 1988 и др.).

По времени статистическое наблюдение может быть текущим или единовременным. В зарубежной статистике для обозначения этих двух видов наблюдений нередко используют термины «*продольное*» и «*поперечное*» наблюдение.

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 и т.д. Считается, что отдельные результаты, полученные таким путем, можно суммировать (помесячную заболеваемость можно суммировать по кварталам, за год). Однако на практике необходимо учитывать одно обстоятельство, порождаемое ограниченностью во времени текущего наблюдения, – полноту регистрации данных. Например, число зарегистрированных случаев заболеваний с временной утратой трудоспособности в течение месяца, как правило, ниже числа действительно состоявшихся случаев утраты трудоспособности на 15-20%. Связано это с тем, что больничные листы, на основании которых ведется статистика этой заболеваемости, учитываются с некоторым опозданием. Кроме того, часть заболеваний к концу месяца еще не завершилась, соответственно больничные листы еще не закрыты. Размер недоучета больничных листов за полгода достигает 7-8%. Отсюда следует, что, если вести полугодовой подсчет путем простого суммирования больничных листов отдельно по дискретным данным каждого месяца, получится итог, существенно заниженный по сравнению с итогом непрерывной регистрации больничных листов за полгода. Такая же закономерность может прослеживаться и при регистрации других видов заболеваемости.

Единовременное (прерывное) наблюдение предусматривает регистрацию данных в один момент времени или по состоянию на один момент времени. Например, данные о численности населения обычно приводятся на конец или начало года, численность студентов – на начало учебного года. Иногда при

вычислении значений годовых показателей, особенно если значения показателей в течение года существенно менялись, эти данные приводят в среднегодовом исчислении. Т.е. эти показатели, формально представленные на определенный момент времени (*например*, на 2012 год), являются средними величинами, несущими отпечаток влияния динамики изменения явления «внутри точки отсчета» во времени (в течение 2012 года). Следует помнить, что расчет такого рода величин не всегда сводится к вычислению простого среднего арифметического из помесечных значений или из полсуммы данных на начало и конец года. Так, в ситуации, когда значения наблюдаемого признака (*например*, размер когорты родившихся и умерших за год детей) изменялись по экспоненте, используются другие, более сложные методики определения среднегодового уровня показателей⁵.

При переписи населения момент времени, когда производится регистрация данных, называется *критический момент наблюдения*. Обычно критический момент устанавливается на полночь, предшествующую первому официально объявленному дню переписи. Это объясняется тем, что большинство населения в ночное время находится по месту постоянного или временного жительства. Данные Всесоюзной переписи населения 1989 года, *например*, регистрировались по состоянию на 12 ч. ночи с 11 на 12 января. Такое начало позволило свести к минимуму влияние фактора миграции населения (в январе она обычно минимальна). Всероссийская перепись населения 2010 года проводилась с ночь с 14 на 15 октября⁶. Несмотря на то, что переписчик, проводивший регистрацию, мог получить данные в любой из последующих дней переписи, данные регистрировались строго на критический момент времени, и изменения, произошедшие после этого момента, во внимание не принимались. Т.е. если за эти дни кто-либо умирал, то он регистрировался как живой. И наоборот, если за это время рождался ребенок, то он не учитывался.

⁵ См. раздел настоящего издания «Смертность населения».

⁶ На отдалённых и труднодоступных территориях, транспортное сообщение с которыми было затруднено, эта перепись проводилась с 1 апреля по 20 декабря 2010 года.

На определенный момент времени (на конец года) учитывается численность населения, численность медицинского персонала, количество учреждений медицинской помощи. Заболеваемость по данным профилактических осмотров (патологическая пораженность) также регистрируется на определенный момент времени. Особенность данных, полученных таким путем при разных по времени наблюдениях, нельзя просто суммировать. *Например*, численность населения района в 1998 году составила 2,5 млн. человек, в 1997 году – 2,6 млн. человек. За 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*.

Комбинацией моментного (единовременного, поперечного) и текущего (продольного) наблюдения является *ретроспективное* наблюдение. В этом случае на первом этапе сбора информации ведется регистрация данных на определенный момент времени – например, на момент углубленного обследования больного или группы больных. Затем регистрируется информация, отражающая динамику состояния этого больного или группы больных за определенный период времени.

Необходимо помнить, что от способа регистрации зависит содержательная часть информации. Например, как уже отмечалось ранее, данные о заболеваемости, полученные путем регистрации обращаемости за медицинской помощью (текущее наблюдение), существенно отличаются от данных, полученных при медицинских осмотрах (единовременное наблюдение). В первом случае среди зарегистрированных заболеваний будут преобладать острые заболевания, во втором – хронические.

Еще одним чрезвычайно важным аспектом программы статистического наблюдения, способного в корне менять всю технологию стат. анализа, является полнота охвата статистической совокупности. По охвату статистической совокупности исследование может быть *сплошное* или

несплошное, выборочное. Выбор охвата объекта наблюдения определяет весь дальнейший ход статистического анализа.

При *сплошном* статистическом исследовании группа наблюдения формируется путем полного охвата всех единиц изучаемого объекта наблюдения. Множество всех единиц наблюдения, охватываемых таким сплошным наблюдением, называется **генеральной совокупностью**. На практике сплошное исследование проводится крайне редко, поскольку его осуществить организационно очень трудно, часто физически невозможно из-за больших размеров или из-за отсутствия определенных границ генеральной совокупности. В ряде ситуаций, даже если генеральная совокупность ограничена в своих размерах, её сплошное исследование ведет к уничтожению объекта наблюдения. *Например*, анализ качества промышленных партий вакцин, сывороток и многих медикаментов связан с частичным или полным разрушением объекта анализа. К тому же сплошные исследования из-за их объема во много раз дороже несплошных. Существенным недостатком сплошных исследований являются и большие затраты времени и средств на обработку данных даже при использовании современных средств телекоммуникаций, что приводит к задержке получения данных⁷. Такие наблюдения используют, как правило, только для решения общегосударственных задач переписи населения, сбора информации об инфекционной заболеваемости и в некоторых других случаях.

Этих недостатков лишено **несплошное, выборочное наблюдение**. К методам несплошного 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

⁷ Например, окончательное подведение итогов переписи населения РФ 2010 года предусмотрено в I квартале 2013 года. Получение итогов переписей 70-80 гг. растягивалось на десятилетия.

исследований, относятся либо только к конкретному объекту исследования, либо могут быть распространены на 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 при таком исследовании, могут иметь значение только для решения частных вопросов. *Например,* вопросов организации госпитального лечения той или иной патологии. Судить о распространенности и структуре 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. Во-первых, оно дает значительную экономию средств и требует существенно меньше времени, чем любое из перечисленных выше. Во-вторых, при выборочном исследовании может быть достигнута большая глубина и детальность изучения вопроса. В-третьих, при меньшем числе наблюдений уменьшаются вероятности *механических, систематических ошибок наблюдения*, возникающих, когда объектам приписываются 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 (репрезентативность)* выборки. Каждое выборочное исследование несет в себе некоторую погрешность, вытекающую из самого факта применения выборки, когда результаты, полученные на выборке, переносят на всю генеральную совокупность, т.е. *по части судят о целом*. Условно можно выделить два тесно взаимосвязанных компонента репрезентативности: **репрезентативность количественную и качественную (структурную)**. *Количественная репрезентативность* определяется таким

статистическим параметром выборки, как количество наблюдений, гарантирующих 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. В достаточно большой массе наблюдений 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 могут возникать как при сборе, так и при сводке информации, искажают результат наблюдения в одном направлении. Они могут возникать, например, из-за халатности регистратора, 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 других статистических параметров выборки. *Например*, изменчивости изучаемого явления. Если бы все единицы совокупности были одинаковы, то результаты, полученные на одной единице наблюдения, можно было бы распространить на все остальные. Например, группа обследуемых больных имеет одну массу тела. Понятно, что в этой ситуации для оценки показателя физического развития у всей группы достаточно узнать массу тела одного больного. Однако реально всегда имеется какой-то разброс значений изучаемых признаков. И чем он больше, тем больше вероятность ошибочного заключения, полученного на основании изучения данных выборки. Именно поэтому при анализе статистических данных крайне необходимы характеристики изменчивости (разброса) значений. 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. Отсюда произрастает и другой, широко распространенный в статистике термин: «случайная величина», который свидетельствует о том, что та или иная величина, характеризующая исследуемую совокупность, попала в выборку с соблюдением принципа случайного отбора. Соответственно и сама выборка в таком случае обоснованно называется «случайной выборкой». Только соблюдение условия случайного отбора позволяет определенно утверждать, что объекты выборки правильно представляют генеральную совокупность. Разумеется, при выполнении других определенных условий, среди которых одно из главных – достаточное число наблюдений.

Случайность, гарантирующая качественную (структурную) репрезентативность статистических исследований, достигается выполнением ряда условий отбора, иначе говоря, условий формирования репрезентативных выборочных групп:

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 лиц с разными фамилиями будет разная, т.к. частота встречаемости различных букв алфавита в началах фамилий разная. Или данные опроса граждан об отношении к представителям муниципальной власти

по интернету не будут репрезентативны, так как не все граждане пользуются этим средством коммуникации.

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:

1. целенаправленно сформировать группы обследуемых в зависимости от их отношения к курению (не курят, курят мало, курят много и т.п.). В этом случае независимо должно формироваться распределение показателей здоровья респондентов в этих группах;

2. 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 (не курят, курят мало, много и т.п.).

В первом случае доказательством связи курения и здоровья будут разные уровни показателей здоровья в группах респондентов, различающихся по отношению к курению. 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 сравнительной 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 предварительного разделения генеральной совокупности на части (собственно случайная, повторная или бесповторная выборка).

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

Собственно случайная выборка формируется случайным отбором – на удачу. Технологически в основе случайного отбора – перемешивание. *Например*, выбор шара в спортлото после периодического 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 данных 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 гипотензивного препарата исследователь, как правило, не знает точного числа людей, которые в объективной реальности принимают этот препарат или нуждаются в нем.

Поэтому в большинстве статистических исследований разница между повторной и бесповторной выборками практически отсутствует и априорно принимается условие, что выборка повторная.

Поскольку на репрезентативность результатов исследования существенно влияют вышперечисленные условия отбора, вариабельность исходных данных и необходимая точность расчетов, возникает вопрос: не объясняются ли те или иные итоговые результаты статистического анализа перечисленными свойствами выборочного исследования, а не объективной реальностью изучаемого явления? Для того, чтобы устранить эти сомнения, необходимо обосновать размер использованной в исследовании выборки. При неизвестной величине генеральной совокупности размер повторной выборки (n), гарантирующей репрезентативные результаты, если они отображаются в виде относительной величины, например, в %, определяется по формуле:

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

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% вероятности безошибочного прогноза); Δ – предельная ошибка показателя.

Например, одним из показателей, характеризующих здоровье рабочих предприятий, является процент работников, не болевших в течение года. При определении массива выборки, необходимой для того, чтобы определить репрезентативную величину этого показателя на предприятии, выполняем следующие действия. Предположим, что для промышленной отрасли, к которой относится обследуемое предприятие, процент неболевших обычно равен (или близок) 25%. Предельная ошибка, которую можно допустить, чтобы разброс значений показателя, т.е. ошибка показателя, не превышал разумные границы, –

5%. Таким образом, предполагаем, что показатель может принимать значения $25\% \pm 5\%$ или от 20% до 30%. Допуская значение доверительного коэффициента $t=2$, получаем:

$$n = \frac{2^2 \times 25 \times 75}{5^2} = 300(\text{чел.})$$

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:

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

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:

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

Нетрудно заметить, что необходимая численность выборки при бесповторном отборе меньше, чем при повторном (в приведенном примере соответственно, 188 и 300 рабочих). В целом число наблюдений, необходимое

для получения репрезентативных данных, изменяется обратно пропорционально квадрату предельной ошибки показателя.

Следует помнить, что весьма распространенной оплошностью, допускаемой при определении необходимого числа исследований, является недоучет того обстоятельства, что, как правило, обработка результатов исследований проводится по нескольким результирующим признакам (нескольким показателям). *Например*, анализ динамики физического развития детей производится с учетом роста, массы тела и т.п. Таким образом, расчет необходимого числа наблюдений нужно проводить отдельно для роста, массы тела и т.п., а общее число наблюдений выбирается по наибольшему из нескольких рассчитанных чисел наблюдений.

Механическая выборка – выборка, когда из обследуемой совокупности единицы наблюдения для выборочной совокупности 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 обращаемость в поликлинику взрослого населения может зависеть от состояния здоровья (здоровые почти не обращаются). В результате амбулаторные карты имеются, как правило, только на больных. Кроме того, на момент отбора определенная часть амбулаторных карт может находиться на руках у пациентов или у врачей.

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 (продольный) отбор относится к целенаправленным отборам. Этот метод отбора информации и соответственно исследование, основанное на такого рода информации, в зарубежной демографии иногда называют *АСР-метод* (от аббревиатуры англоязычных терминов «Age» – возраст, «Cohort» – когорты, «Period» – за который производится наблюдение). При этом способе из генеральной совокупности отбираются лица, объединенные моментом какого-либо события, играющего существенную роль в исследовании (год рождения, начало болезни и т.п.). Примером когорты является возрастная группа, т.е. совокупность ровесников, объединенных общими границами дат рождений, за которыми устанавливается определенное по длительности наблюдение или проводится ретроспективный анализ за время, предшествующее точке отсчета.

Поперечный (поперечный) отбор – отбор на основе данных, полученных в определенные моменты времени или событий, т.е. данных моментного наблюдения. Например, отбор пациентов, чье состояние регистрировалось в определенные моменты времени суток, до и после дозированной нагрузки, приема препаратов; выбор параметров состояния атмосферного воздуха города, зарегистрированных с учетом времени, дня недели, времени года и т.п.

В отличие от других областей науки выборочные исследования в медицине, особенно при изучении общественного здоровья, 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. Однако вплоть до сегодняшнего дня

отсутствуют единые методические подходы проведения выборочных санитарно-статистических исследований. В частности, нет четких, однозначных 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. Сегодня при социологических опросах населения России численностью более 143 млн. человек объем выборки обычно составляет около 2 тыс. человек. Если опираться на рекомендации В.В. Паевского, эта выборка должна бы составлять 14,3 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. Пути повышения надежности выборок

Среди факторов, кардинально снижающих репрезентативность выборочных исследований, можно выделить два наиболее главных. Первый – недостаточность объема выборки, второй – несоблюдение принципа случайного отбора, гарантирующего полное структурное соответствие выборки генеральной совокупности хотя бы по основным параметрам.

Отказ от применения классических приемов случайного отбора как гарантов соблюдения таких условий – довольно распространенное явление в практике медико-биологических исследований. Такое явление часто вызвано объективными обстоятельствами. Среди них одно из главных – неопределенность размеров и структуры генеральной совокупности (незнание всех больных данным заболеванием, всех, у кого есть тот или иной симптом, и т.п.). Кроме того, из-за финансовых, организационных или других причин углубленные клинические и лабораторные исследования обычно отличаются малым числом наблюдений, что требует особенно тщательного соблюдения

принципа случайности отбора. Правильная организация и планирование исследований, обеспечивающих получение достоверных, репрезентативных исходных данных, являются ключевыми проблемами статистики. Все вышесказанное в первую очередь относится и к малым выборкам ($n < 30$), и продольным выборкам, которые формируются путем длительного наблюдения (многие месяцы, годы и т.п.).

Необходимость особенно четкой организации таких наблюдений объясняется и тем, что возможность «потом что-то поправить» (отбросить сомнительные, «выскакивающие» варианты, перегруппировать их и т.п.) в этих выборках весьма ограничена. Ограничено в малых выборках и взаимное погашение случайных, непредвиденных факторов (так называемых *неуправляемых* факторов), обычно проявляющееся при достаточно большом объеме наблюдений.

В этой ситуации существенное значение для повышения репрезентативности статистического исследования имеет возможность регистрации только такой вариабельности наблюдаемого явления, которая закономерно вызвана факторами, представляющими непосредственный интерес для исследования. Но в объективной реальности отделить закономерные изменения от изменений, вызванных сторонними, неуправляемыми факторами, можно либо с помощью специальных, достаточно сложных методов статистической обработки, либо с помощью специальных методов формирования выборок.

Один из таких методов – *метод «копи-пара»* или *метод стратификации*. А выборка, полученная таким методом, называется *парно-сопряженная выборка*. Метод «копи-пара» – это распределение (*стратификация*) экспериментальных единиц в относительно однородные группы (блоки, слои), что позволяет минимизировать эффект неуправляемых источников вариабельности. Теоретически внутри каждой такой группы ошибки, вызванные неуправляемой вариабельностью, предполагаются существенно меньшими, чем при классическом случайном отборе такого же числа единиц

наблюдения. *Например*, при исследовании лекарственного препарата формируются две абсолютно однородные группы наблюдения, различающиеся только одним: «получающие препарат» и «получающие плацебо». Другой *пример*: 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 (обычно токсического вещества), а контрольная группа – нет.

Таким образом, *метод «копи-пара»* предусматривает образование собственно двух групп: группы наблюдения и контрольной группы. Эти группы уравниваются структурно по основным признакам, способным оказывать существенное влияние на результат. Неуравновешенным остается главный фактор, действие которого изучается, что позволяет обходиться малым числом наблюдений. Однако в практике исследований, которые проводятся среди людей, подбор копий-пар трудно осуществим.

Другой метод снижения независимой вариабельности, по принципу весьма схожий с приведенным выше, – использование *парно-сопряженной выборки*. Т.е. использование одних и тех же объектов наблюдения до и после эксперимента. Часто такая выборка называется более кратко – *парная выборка*. Такие выборки считаются наиболее эффективными с точки зрения материальных затрат на проведение наблюдения и репрезентативности итогов.

Применение этих методов устраняет или сводит к минимуму независимую вариабельность объектов, играющую определенную роль в появлении *систематических* (механических) ошибок наблюдения. Понятно, что индивидуальная вариабельность, обусловленная свойством всего живого – индивидуальной изменчивостью, остается при любом способе формирования выборки⁸.

⁸ В настоящем издании не рассматривается такой важный источник ошибок (погрешностей) исследования, как качество исследования, описываемое специальными характеристиками (точность, правильность, сходимость, воспроизводимость результатов наблюдения и т.п.), имеющими особенно существенное значение при использовании разного рода лабораторного оборудования.

Р.А. Фишер (1950) дал знаменитое толкование правил отбора, гарантирующих репрезентативность выборки. В качестве примера он рассмотрел гипотетический случай, когда некой английской леди было предложено провести (очень важные для англичан) оценки: что было ранее налито в чашку – чай или молоко? По мнению Фишера, такое исследование должно строиться с соблюдением следующих правил (цит. По Э. Ллойд и У. Ледерману, 1989):

- *повторяемость (репликация)*. 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 насыщением чаем и т.п.

Следование указанным правилам в определенной степени позволяет решить проблемы формирования репрезентативных выборочных групп при неизвестных генеральных совокупностях. И наоборот, забвение этих правил ставит под сомнение результаты любого исследования. В качестве негативного примера можно привести информацию, обнаруженную в *Internet*.

В последнее время в зарубежной и отечественной статистике часто используется как показатель эффекта воздействия (отрицательного или положительного) так называемый «относительный риск» (*relative risk*), обозначаемый в русскоязычной транскрипции – *ОР* или в латинице – *RR*. По своей сути этот показатель характеризует отношение каких-либо показателей здоровья лиц, подверженных воздействию некоего фактора (экспонированных лиц), отрицательно влияющего на здоровье, и показателей здоровья лиц, не подверженных этому воздействию (неэкспонированных лиц).

Например: рак легкого встречается в 10 раз чаще среди курильщиков, чем среди некурящих ($OP=10$).

Обозначим численные выражения риска среди экспонированных, как P_1 , и среди неэкспонированных, как P_2 , $OP=P_1/P_2$

В качестве негативного примера: по утверждению одного из авторов, в проведенном им исследовании было обнаружено, что состояние менопаузы является фактором риска повышенной заболеваемости женщин ишемической болезнью сердца (ИБС). В качестве доказательной базы этого утверждения автор привел следующие исходные данные (табл.1), расчеты и логику заключения.

Таблица 1

Статус	Число человеко-лет под риском	Число случаев ИБС	Показатель на 1000 человеко-лет
Постменопауза	6848	26	3,8
Предменопауза	8384	6	0,7

Следуя таблице, принимая состояние менопаузы как воздействующий фактор, *ОР* ИБС, связанный с ним, составит: $OP = 3,8/0,7 = 5,43$.

Этот результат, по мнению автора, можно интерпретировать таким образом, что у женщин в постменопаузе ИБС встречается почти в 5,5 раза чаще, чем у женщин в предменопаузе.

Однако! Во-первых, менопауза может быть ранней, т.е. наступать у женщин в возрасте 40-45 лет, или поздней, т.е. наступать в возрасте 60 лет и

старше. А показатель «человеко-лет» ничего не говорит о том, сколько на самом деле было женщин и какого они возраста в группах «до...» и «постменопауза». Например, производная величина 1000 человеко-лет может быть получена путем перемножения 20 человек на 50-летний возраст, а может быть и наоборот: 50 человек на 20 лет от роду⁹. Во-вторых, известно, что возраст решающим образом влияет на вероятность появления ИБС. Отсюда вытекает второе обстоятельство, делающее утверждение автора весьма сомнительным, поскольку средний возраст женщин группы «постменопауза» обычно всегда больше, поэтому и заболеваемость ИБС там всегда больше.

Для предотвращения возможного влияния на конечный результат неоднородности выборочной совокупности можно попытаться целенаправленно сформировать выборочную группу путем уравнивания факторов, которые явно исказят результат статистического анализа. Например, при углубленном обследовании работников предприятия проведено углубленное медицинское освидетельствование только 187 рабочих. Здесь одним из самых существенных факторов, способных оказать влияние на итоги, может быть неоднородная возрастно-половая структура обследованной выборочной группы. Для того, чтобы выполнить проверку однородности (гомогенности) выборочной совокупности с помощью специальных статистических критериев (см. в настоящем издании разделы «Критерии гомогенности»), необходимо выполнить следующие действия:

1. Получить таблицу фактического распределения обследованных работников предприятия по возрасту и полу в процентах к общему итогу (табл. 2).

⁹ В статистике показатель «человеко-лет» используется в демографических таблицах дожития, но там обязательно указывается определенное число людей (человек), поэтому двусмысленное толкование показателя полностью исключается.

Таблица 2

Распределение обследованных работников предприятия по возрасту и полу (в % от числа всех обследованных)

Пол	Возраст (лет)			Итого
	До 30	30-39	40 и ст.	
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 в относительных частотах (табл.3).

Таблица 3

Распределение обследованных работников предприятия по возрасту и полу (в % от числа всех обследованных)

Пол	Возраст (лет)			Итого
	До 30	30-39	40 и ст.	
Мужской	0,059	0,099	0,099	0,257
Женский	0,109	0,119	0,515	0,743
Оба пола	0,168	0,218	0,614	1,000

3. Проверить с помощью модифицированного критерия Пирсона χ^2 распределение частот подгрупп в выборочной группе (см. раздел «Критерии гомогенности») с помощью *Excel*.

В данном случае различия возрастно-половой структуры выборочной совокупности можно признать несущественными ($P=0,401$)¹⁰.

Аналогичным образом можно проверить гомогенность выборочной совокупности 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 известными, нужными учетными признаками или, наоборот, отбросить. Такого рода операции называются *цензурированием* выборки. Относиться к ним нужно

¹⁰ Обратить внимание! Здесь речь идет только об однородности, гомогенности структуры исследуемой совокупности!

с большой осторожностью, поскольку требования к чистоте эксперимента и, соответственно, правомерность тех или иных операций с выборками могут иметь в каждом конкретном случае свои особенности.

2.4. Программа разработки (систематизации) статистических данных

Программа разработки предусматривает *сводку и группировку* данных. Эти операции, осуществляемые на основе статистических таблиц, позволяют систематизировать полученные в ходе наблюдения данные, провести 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 первичного материала. Играет исключительную роль в статистике. Группировка позволяет облегчить регистрацию или понизить излишнюю точность измерений, допущенную на этапе сбора исходных данных, без снижения результативности исследования, а также представить в единообразной форме разнородные, качественные и количественные признаки, что на практике осуществляется путем перевода количественных признаков в качественные. Группировка исходной информации определяет весь ход статистического анализа. Ошибки, допущенные на этом этапе, неизбежно сказываются на итогах всего исследования¹¹. Необходимо помнить, что возврат, т.е. разгруппировка, уже скомпонованных групп на практике, как правило, невозможен. Поэтому рекомендуется сохранять исходную информацию (до группировки) отдельно, в месте, исключающем случайный доступ.

¹¹ Необходимо помнить, что обратная разгруппировка уже скомпонованных групп, как правило, не возможна. Поэтому рекомендуется сохранять исходную информацию (до группировки) отдельно, там, где исключается к ней случайный доступ.

В практике статистической обработки данных используются следующие варианты группировок:

- *разделение (разбиение) анализируемой статистической совокупности* на группы по тем или иным, фактически существующим признакам отдельных единиц наблюдения (по полу, возрасту, диагнозу заболеваний и т.п.). С такого рода группировкой приходится сталкиваться уже при подготовке программы сбора и в ходе разработки данных любого исследования;

- *объединение мелких однородных групп или их частных характеристик (признаков) в более крупные.* Этот вариант группировки применяется, если выясняется несостоятельность мелких групп (малое число наблюдений, нечетко выраженный характер распределений и т.п.). Возможность такой группировки целесообразно предусмотреть уже на этапе подготовки программы сбора данных, т.е. предусмотреть возможность укрупнения групп в соответствии с общепринятыми границами группировок. При этом необходимо учитывать возможность изменения однородности состава объединенных групп или распределения отдельных единиц наблюдения в укрупненных группах;

- *комплексная группировка* обеспечивается на основе формирования объединенных (комплексных) учетных признаков из многих отдельных признаков, даже если они разнородны. Эта группировка часто делается с помощью специально разрабатываемых алгоритмов или экспертных оценок (*например*, 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 объемом, глубиной исследования и видом учетных признаков. Для группировки качественных признаков обычно используется альтернативная шкала.

Альтернативная шкала, шкала рангов, шкала номиналов, шкала категорий характеризуется 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 (даже если он числовой) или его качественную характеристику относительно другого признака (мужской пол лучше, женский – хуже). При такой статистической сводке, для удобства дальнейшей обработки, все значения качественных учетных признаков часто кодируются (мужской пол – 1, женский – 2 и т.п.). Таким образом, характеристики приобретают некое количественное выражение, оставаясь в принципе качественными. Такое количественное выражение нередко побуждает исследователя механически использовать их как количественные характеристики, что является грубой ошибкой.

Следует отметить, что в научных исследованиях последнего времени придание качественным признакам числового выражения стало уже малооправданным, поскольку современное программное и аппаратное обеспечение исследований позволяет свободно обрабатывать массивы с качественными (атрибутивными) признаками.

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 смысловое 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 и т.п.) при этом остаются качественными. Механическая 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 ($4=2+2$).

Достоинством балльных шкал является возможность получения интегрированных оценок (табл. 4).

Таблица 4

**Сравнительная оценка распространенности курения
среди 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

Особенно удобно использовать балльные шкалы 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 балльная оценка этих показателей, а затем – суммирование баллов. Если рост ребенка оценен в 2 балла, вес – в 3 балла, то его комплексная оценка по росту и весу будет $(2+3)=5$ баллов. (Более подробно – в разделе «Квантили»).

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

Оптимальное число рангов при распределении медико-биологической информации обычно 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 подсозна-

тельно 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) в определенных интервалах и упорядочивать объекты исследования. *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. 5).

Таблица 5

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

2.4.1. Классификация статистических группировок

Аналитическая группировка позволяет выявлять взаимосвязи между явлениями. Точнее – признаками, их характеризующими. При этом признаки 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 масса тела 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 изменения структурной группировки во времени дает представление о структурных сдвигах, происходящих под воздействием какого-либо фактора.

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 (возрастно-половые группы, группы категорий тяжести труда и т.п.) и границ групп. Несоблюдение этих правил ведет к потере ценности данных. В частности, необходимо учитывать сопоставимость 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 в этом случае результаты будут просто несопоставимыми с результатами других исследований.

Стандартные группировки обусловлены, как правило, определенными мотивами, игнорирование которых может лишить научной содержательности все исследование. Например, возрастная группировка 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):

- дети до 3 лет. Эта группа находится под наблюдением детских консультаций и обслуживается детскими яслями;

- дошкольники – дети от 3 до 7 лет. Обслуживаются детскими садами;
- дети и подростки школьного возраста – 7-13 лет и 13-16 лет;
- подростки 16-18 лет;
- трудоспособный контингент – мужчины 16-60 лет, женщины 16-55 лет;
- лица пенсионного возраста – мужчины 60 лет и старше, женщины 55 лет и старше;
- женщины репродуктивного возраста – от 15 до 45 лет.

Для группировки по возрасту обычно используют одногодичные или (для взрослых) пятилетние, реже десятилетние интервалы. Группировка с пятилетним возрастным 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 этом, например, формально в возрастную группу 20-24 года относят лиц, имеющих возраст не моложе 20 лет, и лиц, не превышающих точной границы возраста: до 24 лет, 11 месяцев, 30 дней и 23 часов, 59 минут. При изучении заболеваемости в связи с производственными факторами обычно используются 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, т.е. использовать группировки в неравных интервалах. Например, до 20 лет, 20-29, 30-39, 40-44, 45-49, 50-59, 60 и старше.

При медико-демографических исследованиях рекомендуется применять комплексную группировку, которая позволяет учитывать наряду с возрастом и другие социально-экономические признаки, пол, потенциальную трудоспособность населения и т.п. Такую группировку называют группировкой контингентов, когортной группировкой (табл. 6).

Классификация возрастных контингентов (когорт)

Новорожденные	1-10 дней
Грудной возраст	от 10 дней до 1 года
Раннее детство	1-3 года
Первое детство	4-7 лет
Второе детство	8-12 лет, мальчики
	8-11 лет, девочки
Подростковый возраст (моложе трудоспособного)	13-16 лет, мальчики
	12-15 лет, девочки
Юношеский, трудоспособный возраст (1-й период)	17-21 год, юноши
	16-20 лет, девушки
Зрелый, трудоспособный возраст (1-й период)	22-35 лет, мужчины
	21-35 лет, женщины
Зрелый, трудоспособный возраст (2-й период)	36-60 лет, мужчины
	36-55 лет, женщины
Пожилой возраст, старше трудоспособного	61-74 года, мужчины
	56-74 года, женщины
Старческий возраст	75-89 лет, мужчины и женщины
Долгожители	90 лет и старше

В зависимости от определенности границ в крайних группах группировки могут быть открытыми или закрытыми. Открытые – начинаются и заканчиваются вне определенных точек отсчета. Например, может отсутствовать конкретное значение минимального возраста, который служит основанием для включения в начальную группу «до 20 лет». Может отсутствовать и значение максимального возраста, служащего основанием для включения в группу «60 лет и старше». Преимуществом таких группировок является автоматическое устранение систематических ошибок регистрации данных, которые, как показывает практика, в 90-95% случаев допускаются именно в конечных значениях. Недостатком является то, что в рядах данных, образованных в рамках таких группировок, нельзя вычислять некоторые статистические величины. Например, средний возраст. Этого недостатка лишены закрытые ряды, т.е. ряды с конкретными минимальными и максимальными значениями наблюдаемых признаков. Например, 16-20 лет,, 60-70 лет.

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, т.е. закона. Примеры действующих в настоящее время классификаторов:

- ОК 029 (ОКВЭД) «Общероссийский классификатор видов экономической деятельности»;
- ОК 019 (ОКАТО) «Общероссийский классификатор объектов административно-территориального деления»;
- ОК 011 (ОКУД) «Общероссийский классификатор управленческой документации»;
- ОК 001 (ОКС) «Общероссийский классификатор стандартов»;
- ОК 005 (ОКП) «Общероссийский классификатор продукции»;
- ОК 028 (ОКОПФ) «Общероссийский классификатор организационно-правовых форм»;
- ОК 027 (ОКФС) «Общероссийский классификатор форм собственности».

Основные классификаторы рассчитаны на длительное использование. Однако с течением времени они пересматриваются, дополняются, в них вносятся необходимые коррективы. Так, с 1983 до 1999 годы в СССР, а затем в России общепринятой считалась *Международная статистическая классификация болезней, травм и причин смерти 9-го пересмотра*. С 1999 года – классификация 10-го пересмотра (с 01.01.1999 г. по Приказу Минздрава России № 3 от 12.01.1998 г.)¹².

¹² Полное название этого классификатора «Международная статистическая классификация болезней и проблем, связанных со здоровьем, 10 пересмотр».

Другим примером международной классификации может служить классификатор ООН, позволяющий оценивать степень демографической старости населения (табл. 7).

Таблица 7

Вариант классификации населения в зависимости от степени старения

Группы	% лиц в возрасте 60 лет и старше к общей численности населения	Характеристики отдельных групп
I	Менее 8	Демографическая молодость
II	8-10	Преддверие старения
III	10-12	Собственно старение
IV	12 и более	Демографическая старость

В медико-биологических научных исследованиях использование государственных и международных классификаторов не является строго обязательным. Однако только их прямое применение или возможность путем вторичной группировки привести данные исследований в рамки, определяемые этими классификаторами, гарантирует сопоставимость статистических материалов, полученных на различных территориях страны и за рубежом.

2.4.2. Статистические таблицы. Правила оформления

Для сводки (систематизации) данных, в зависимости от их целевого назначения, используют различные типы таблиц:

- перечневые, описательные таблицы;
- таблицы электронных баз (хранилищ) данных;
- специальные таблицы математико-статистического моделирования;
- статистические таблицы.

Перечневые или описательно-информационные таблицы представляют собой простой перечень данных. *Например*, таблицы умножения, таблицы обратных величин, логарифмов, расписание железнодорожных перевозок и т.д. Порядок систематизации данных в таких таблицах определяется требованиями дизайна, наглядности отображаемой информации, в т.ч. и местными традициями.

Таблицы электронных баз данных (БД) – совокупность логически связанных и систематизированных данных, обеспечивающая функционирование систем управления базами данных (СУБД). Форма размещения данных в таблицах БД определяется их основной задачей – накоплением информации в виде, пригодном для дальнейшей обработки средствами СУБД. Практически во всех базах данных стандартным является использование только двух элементов: записей (располагаются по горизонтали, в строках) и полей (располагаются по вертикали, в столбцах). Набор полей, количество записей и формат данных в этих таблицах определяются функциональными возможностями и целевым назначением конкретных СУБД. Эти таблицы не содержат итогов по строкам (записям) и столбцам (полям). Информация, размещаемая в таблицах баз данных, постоянно динамично изменяется (дополняется, корректируется, удаляется). Для накопления всей «первозданной» информации за какой-либо достаточно длительный период используются **таблицы хранилищ данных**. Эти типы таблиц имеют аналогичную, относительно несложную формализованную структуру, также состоящую только из двух элементов: *записей* – информации, размещаемой по строкам, и *полей* – частных характеристик каждой записи, размещаемых по столбцам таблицы. Основное отличие от таблиц баз данных – неизменность хранимых данных. Такое структурное единообразие таблиц баз и хранилищ данных позволяет физически или виртуально¹³ объединять множество таблиц в сложные, взаимосвязанные совокупности информационных систем.

Специальные таблицы математико-статистического моделирования представляют собой совокупность расположенных в табличных формах и функционально взаимосвязанных числовых данных, характеризующих развитие какого-либо процесса. Примером таких таблиц могут быть таблицы

¹³ *Виртуальное объединение* (лат. «*virtualis*» – возможный) – условное, т.е. реально не существующее объединение.

динамических рядов, демографические таблицы вымирания (дожития)¹⁴. В таких таблицах имеется только один-два столбца с исходной числовой информацией, которая, как правило, располагается в первых столбцах и представляет собой данные фактических наблюдений, а остальные столбцы таблиц заполняются результатами специальных расчетов.

Статистические таблицы принципиально отличаются от перечисленных выше таблиц тем, что в них вся информация фиксируется (размещается) таким образом, чтобы отражать главную статистическую характеристику любой исследуемой совокупности данных – статистическое распределение. Т.е., приводимые в статистической таблице данные располагаются таким образом, чтобы дать возможность не просто их систематизировать, но и проводить анализ распределения данных. Фактически вся программа разработки любого статистического исследования воплощается в статистических таблицах. Умение составлять и читать эти таблицы, как правило, приобретается в процессе практической деятельности.

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. 8).

Таблица 8

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 (боковые внутренние заголовки)		Графо-клетки			
Всего					

¹⁴ См. разделы «Динамические ряды» и «Медицинская демография».

Статистическое подлежащее – это те группы или отдельные единицы статистической совокупности, которые характеризуются в таблице. В подлежащем обычно размещают показатели результативных признаков. Как правило, подлежащее располагается по столбцам таблицы.

Статистическое сказуемое – это признаки, которыми характеризуются группы или единицы совокупности. В сказуемом размещают числовые показатели подлежащего.

Расположение и форму сочетания конструктивных элементов представляет *макет* статистической таблицы. В целом таблица должна быть по возможности небольшой по размеру, так как краткую таблицу легче анализировать. Иногда целесообразней 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, 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 обязательно указывается размерность или (и) наименование единиц («%», «на 1000 населения» и т.п.). Соответственно, в клетках таблицы приводятся показатели без указания этих атрибутов.

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 условные обозначения. Во-вторых, это дает возможность показать 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. 8).

Не следует строить очень громоздких статистических 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, соответствующему понятию *синтез* (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 распределении рассматриваемого явления. Этот порядок соответствует методологии анализа: от *общего к частному*. Затем анализируют данные отдельных граф и строк, но их следует читать не подряд, а выбрать сначала наиболее характерные данные, а затем – 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, то в соответствующих строках или столбцах указывается 100%. Отсутствие итоговых показателей является методической ошибкой (табл. 13), поскольку не дает возможности 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.

Следует различать «итого» и «всего». «Итого» является итогом для некоторой части совокупности. «Всего» – для всей представленной в таблице совокупности. К сожалению, эта разница не учитывается в системах автоматизированных расчетов, например, в *Excel*.

Цифровые данные целесообразно округлять. При этом округление должно быть единообразным для всех данных в графе (столбце). Размерность приводимых показателей и точность округления указываются 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 так же единообразно, на основе общепринятых правил или стандартов. Если

возможно, расшифровка используемых сокращений приводится в общем заголовке таблицы, *например*, «Динамика снижения 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, недостаточно велика, чтобы сравнивать показатели накопленной за этот период жизни детей 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 распределение одного признака (табл. 9)¹⁵. Часто это распределение приводится вместе с одной или несколькими количественными характеристиками этого распределения. Как правило, эти характеристики приводятся в виде производных величин (интенсивных, экстенсивных, коэффициентов соотношения и т.п.).

Таблица 9

Распределение валового выброса антропогенных загрязнителей по обследованным в 2005 г. 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. 10). 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, столбцам или «уголком». Порядок расположения диктуется логикой проводимого анализа. Например: на основании показателей структуры, вычисленных по строкам (табл. 10), можно говорить о составе антропогенных загрязнителей (специфические или неспецифические) в том или ином обследованном районе промышленного города.

По показателям структуры, вычисленным по столбцам (табл. 11), можно анализировать распределение общей суммы обоих видов загрязнителей по обследованным районам.

¹⁵ Пример условный

Таблица 10

**Распределение антропогенных загрязнителей 2005 г.
(в % к валовому выбросу в каждом обследованном районе)**

Номер района	Загрязнители				Валовой выброс	
	Специфические		Неспецифические		тонн/год	%
	тонн/год	%	тонн/год	%		
1	89568	79,0	23809	21,0	113377	100,0
2	136422	89,2	16540	10,8	152962	100,0
3	8812	86,7	1352	13,3	10164	100,0
4	10256	98,2	188	1,8	10444	100,0
5	60124	86,6	9276	13,4	69400	100,0
Итого	305182	85,9	50165	14,1	355347	100,0

Таблица 11

**Распределение антропогенных загрязнителей
по обследованным районам по данным 2005 года
(в % к итогу)**

Номер района	Загрязнители				Валовый выброс	
	Специфические		Неспецифические		тонн/год	%
	тонн/год	%	тонн/год	%		
1	89568	29,3	23809	47,5	113377	31,9
2	136422	44,7	16540	33,0	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	9276	18,5	69400	19,3
Итого	305182	100,0	50165	100,0	355347	100,0

Допускается вычисление показателей структуры «уголком» (табл. 12).

Таблица 12

**Распределение антропогенных загрязнителей
по данным 2005 года (в % к общей сумме валовых выбросов)**

Номер района	Загрязнители				Валовый выброс	
	Специфические		Неспецифические		тонн/год	%
	тонн/год	%	тонн/год	%		
1	89568	25,2	23809	6,7	113377	31,9
2	136422	38,4	16540	4,7	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	9276	2,6	69400	19,3
Итого	305182	85,9	50165	14,1	355347	100,0

Нередко простые и групповые статистические таблицы дают возможность проводить только ориентировочный, предварительный анализ. В них 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 используются *комбинационные таблицы*, в которых оценивается распределение трех и более признаков (табл. 13, табл. 14).

Существенным недостатком комбинационных таблиц является большое число наблюдений, которое требуется для получения с их помощью статистически репрезентативных данных. Чтобы вычислить показатели, приведенные в табл. 14, потребовалось более 1000 единиц наблюдения.

Таблица 13

Неполная форма комбинационной таблицы 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.	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

Для повышения наглядности отчетных данных целесообразно использовать графические изображения (диаграммы). С их помощью легче уяснить закономерности развития, распределения и размещения явлений.

Необходимо помнить, что абсолютное дублирование данных отчета в графических изображениях и статистических таблицах является некорректным. При решении вопроса, какие данные представлять в виде статистических таблиц, а какие выносить в диаграммы, следует помнить, что повышение

наглядности анализируемой информации является целевым назначением графических изображений. Если эта задача с помощью диаграммы не решается, тогда нет и необходимости в ее использовании.

Таблица 14

Полная форма комбинационной таблицы распределения 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	79,4	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

2.4.3. Диаграммы. Правила построения

Исторически сложилось так, что все первые диаграммы, начавшие применяться в статистике, назывались одним термином – «гистограммы». По мере развития статистики происходило и развитие технологий графического анализа, накапливались и становились стандартными различные формы представления результатов обработки статистических данных. Возникла и стала развиваться соответствующая номенклатура диаграмм. Особенно резкий скачок произошел с началом широкого использования для построения разного рода диаграмм средств вычислительной техники. Сегодня термин «гистограмма» применим только к узкому кругу графических отображений, играющих существенную роль в математической статистике для графической интерпретации параметров распределения. Однако «ренессанс» расширенного толкования этого термина встречается еще довольно часто.

В настоящее время, когда практически каждое рабочее место исследователя оснащено аппаратными и программными средствами, позволяющими без особого труда проводить графический анализ статистической информации, главной проблемой стало нахождение из множества возможных одного, правильного графического решения, образа диаграммы и ее экспликации.

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).

Вне зависимости от типа используемого графического изображения целесообразно придерживаться следующих **правил построения диаграмм**:

- общая структура диаграммы должна предполагать чтение слева направо;
- следует избегать попыток отображения линейных величин с помощью площадей и объемов, как не соответствующих сути показателей. Кроме того, из-за обмана зрения могут возникать ошибки сравнительного восприятия величин, отображаемых в виде площадей и объемов;
- вертикальную шкалу на диаграмме следует выбрать так, чтобы на ней оказалась нулевая отметка. Если это невозможно, например, из-за большого значения показателей, то точку отсчета целесообразно определять, либо от круглого числа, либо от уровня, имеющего какое-либо смысловое значение (стандарт, среднее и т.п.). Правильный выбор точки начала отсчета и точки максимального значения существенно повышает наглядность отображаемой статистики (рис. 1 и рис. 2);
- Для шкал, изображающих проценты, промилле и т.п., каким-то образом выделять, соответственно, 100, 1000, 10000 и т.д. Целесообразно выделять величины, обозначающие норму, стандарт или средний уровень показателей. Когда шкалы относятся к датам, лучше не выделять первые и последние ординаты, т.к. подобные диаграммы, как правило, не отражают начало и конец времени;



Рис. 1. Удельный вес работников основных профессиональных групп, не имевших случаев ЗВУТ в отчетном году



Рис. 2. Удельный вес работников основных профессиональных групп, не имевших случаев ЗВУТ в отчетном году (за точку отсчета показателя принято 50%)

- 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;
- горизонтальную шкалу для кривых следует читать, как правило, слева направо, а вертикальную – снизу вверх. Если отображаемые с помощью кривой данные резко отличаются друг от друга по своей величине, рекомендуется делать разрыв масштабной шкалы. Этот же прием применяется, если данных за какой-либо отрезок анализируемого периода нет. 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 (рис. 3);

- когда шкалы относятся к датам, лучше не выделять первые ординаты, т.к. подобные диаграммы, как правило, не отражают начало времени;

- при использовании условных обозначений необходимо давать пояснения к ним. При использовании в нескольких последовательно расположенных диаграммах одних и тех же учетных признаков применяются обязательно одни и те же условные обозначения для этих признаков;

- желательно включать в диаграмму используемые формулы;

- наименования графических изображений в книгах, журналах и т.п. обычно указывают снизу от рисунка. Названия таблиц – сверху. В диаграммах, не предусмотренных для печати, например, настенных диаграммах и слайдах, целесообразно писать заголовки сверху;

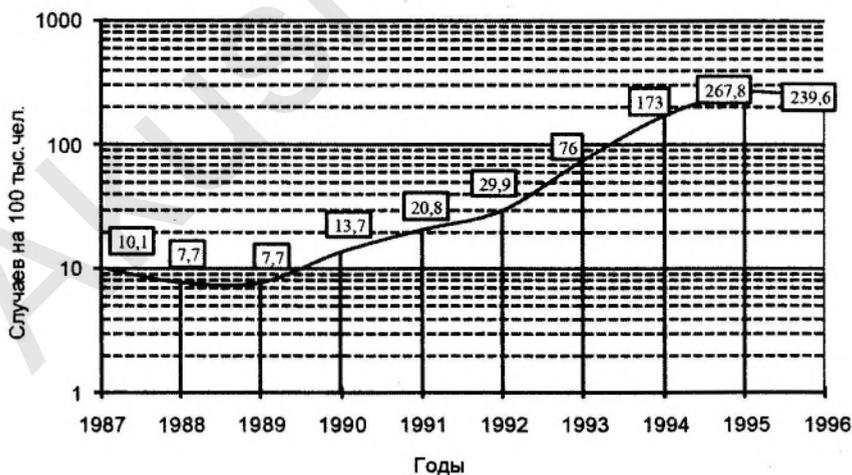


Рис. 3. Рост первичной заболеваемости населения Санкт-Петербурга сифилисом (логарифмическая шкала ординат)

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 обобщенном виде это соотношение равно 3 к 2 (Рис. 4);

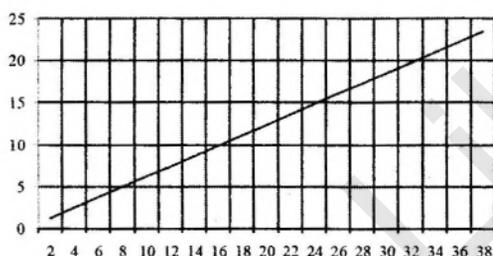


Рис. 4. Соотношение осей X и Y при соблюдении пропорций «золотого сечения»

- при построении секторной диаграммы начало отсчета производится от верхней точки («12 часов») и по ходу часовой стрелки (). Не рекомендуется использовать эту диаграмму для отображения более 5-7 показателей. Если такая необходимость существует, то нужно использовать другой тип диаграмм;
- на секторной диаграмме целесообразно откладывать числовые значения признака от большего к меньшему. Если этот порядок противоречит логической последовательности данных, то он может быть нарушен.

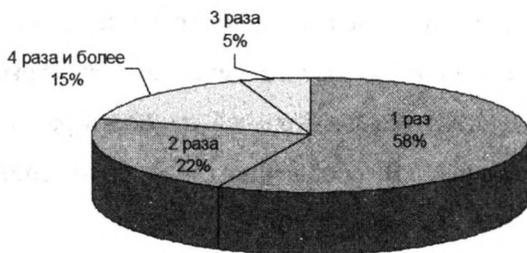


Рис. 5. Распределение работников предприятия по кратности случаев ЗВУТ в отчетном году (секторная диаграмма)

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.

Одним из простейших вариантов построения графических изображений являются различные *трафаретные рисунки и фигурные диаграммы*, соответствующие определенному содержанию статистических данных. Правильно выбранные такого рода изображения делают информацию легко доступной для восприятия массовым зрителем, не имеющим подготовки в области статистики. Например, сравнительное потребление чая мужчинами и женщинами. Каждая фигура – определенное количество чая (рис. 6).



Рис. 6. Фигурная диаграмма.
Сравнительное потребление чая мужчинами и женщинами

Линейные диаграммы – наиболее распространенный вид диаграмм. Применяются для отображения практически любых статистических величин. Вместе с тем, следует помнить, что по горизонтальной оси откладываются только непрерывные величины (данные измерения). 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.

Если графическое изображение плоскостное, то применяется двухмерная система координат. Если объемное – трехмерная. Вариант линейного графика в виде объемной ленточной диаграммы трехмерной диаграммой не является, поскольку характеризует распределение явления в двухмерной системе координат (рис. 7).

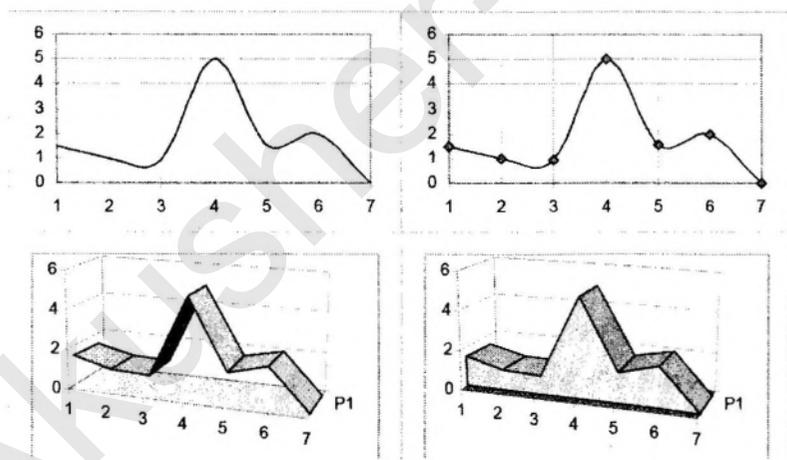


Рис. 7. Примеры компоновки (экспликация) линейных диаграмм

Достоинством плоскостных линейных диаграмм является возможность наглядно отражать совместное распределение большого количества рядов, в том числе рядов, состоящих из несопоставимых показателей. Это могут быть разнородные или однородные, но сильно отличающиеся по своим значениям,

так наз. разномерные показатели. Если на диаграмме требуется представить динамику двух разнородных или разномерных показателей, то используют не одну, а две масштабные шкалы. Одну из них размещают справа, а другую, вспомогательную, – слева. При этом каждой шкале может соответствовать свой тип диаграммы (рис. 8).

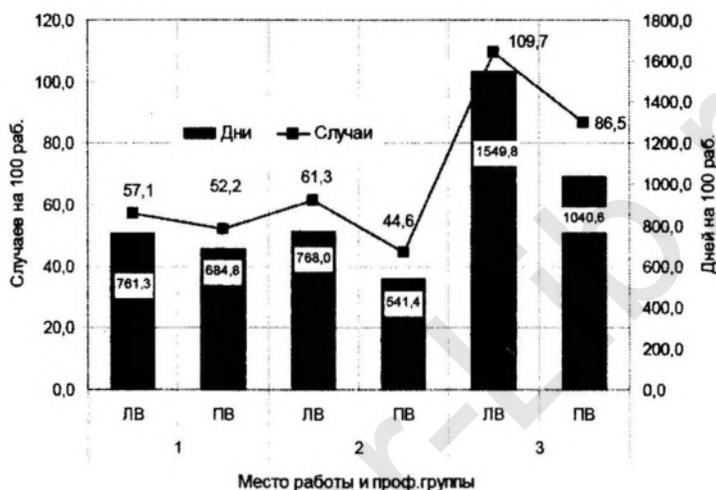


Рис. 8. Распределение показателей ЗВУТ среди работников 1-го и 2-го филиалов предприятия для разных профессиональных групп

Подобная экспликация не всегда дает достаточно яркой картины динамики явления (явлений). Кроме того, применение этого метода невозможно в случаях, когда необходим одновременный сравнительный анализ многих разнородных и разномерных показателей (табл.15, рис. 9).

Таблица 15

Динамика первичной заболеваемости жителей Санкт-Петербурга (на 100 тыс. населения)

Заболевания	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Злокачественные	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. 9. Динамика первичной заболеваемости жителей Санкт-Петербурга (случаев на 100 тыс. населения)

В этой ситуации диаграмму наиболее целесообразно строить на основе использования одного приведенного масштаба. Приведение к одному масштабу достигается путем преобразования исходных данных в относительные величины.

В примере (рис. 10 и табл. 16) исходные данные о заболеваемости переведены в показатели наглядности. Т.е. за исходный, нулевой, уровень приняты показатели 1987 года. Остальные уровни по каждой нозологической форме рассчитаны в процентах по каждому году относительно своего исходного значения в 1987 году.

Таблица 16

Динамика первичной заболеваемости жителей Санкт-Петербурга 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. 10. Динамика первичной заболеваемости жителей Санкт-Петербурга в показателях наглядности (в % к 1987 году)

Столбиковые диаграммы представляют собой изображения различных величин в виде расположенных в высоту прямоугольников одинаковой толщины и разной высоты. Построение столбиковой диаграммы требует только одной масштабной 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 внутрестолбиковые или секторные диаграммы. Столбиковые диаграммы могут быть представлены в плоскостном варианте (рис. 11) или объемном (рис. 12, рис. 13 и рис. 14).

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

Поскольку главным критерием выбора той или иной диаграммы является наглядность и удобство анализа результатов, то в данном случае более целесообразно представить попарно сгруппированные показатели мужчин и женщин.

Объемные диаграммы могут давать более насыщенное отображение показателей. Однако использование трехмерного распределения не гарантирует само по себе увеличение наглядности и ограничивает число данных, которые можно выводить на диаграмму. На примере рис.13 видно, что наглядность представленных на объемной диаграмме данных не возросла.

Для отображения экстенсивных показателей, характеризующих часть от целого (долю, удельный вес) в %, используются **внутристолбиковые** диаграммы. При этом весь объем столбика принимается за 100%. Затем из этого объема выделяются части, размеры которых пропорциональны величинам показателей (рис. 14).

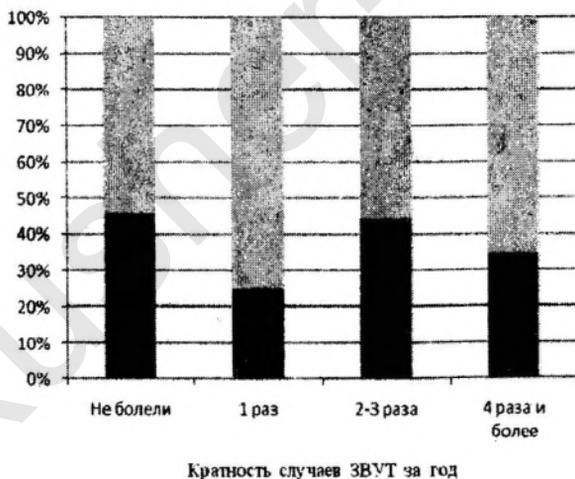


Рис. 11. Сравнительное распределение мужчин и женщин в группах разной кратности случаев ЗВУТ в году (в %)



Рис. 12. Сравнительное распределение мужчин и женщин в группах разной кратности случаев ЗВУТ в году (в %)



Рис. 13. Сравнительное распределение мужчин и женщин по кратности случаев ЗВУТ в году (в % к общему числу каждой половой группы)

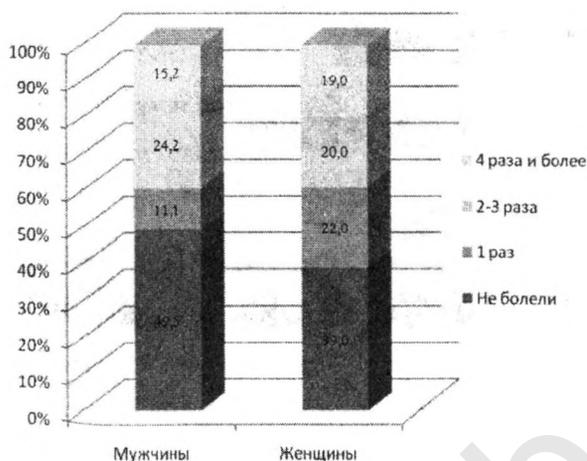


Рис. 14. Распределение мужчин и женщин по кратности случаев ЗВУТ в году (в %)

В ряде случаев для отображения структуры может использоваться комбинация внутрисклонной и столбчатой диаграммы (рис. 15).

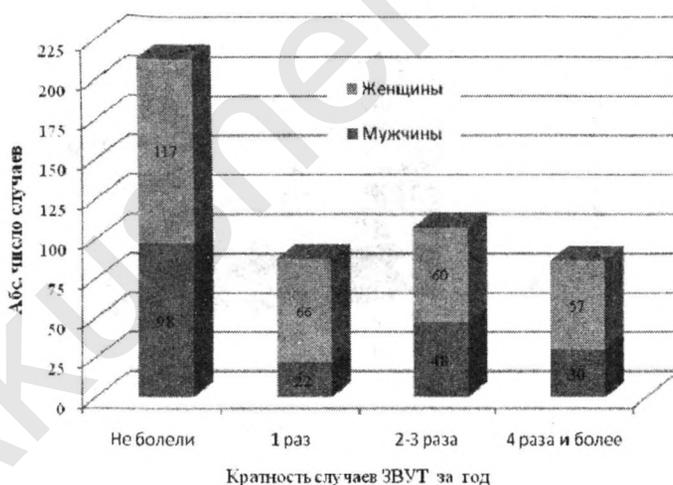


Рис. 15. Сравнительное распределение мужчин и женщин в группах разной кратности случаев ЗВУТ за год (в %)

Одним из вариантов столбчатых диаграмм являются ленточные диаграммы, которые в MS Excel называются «линейчатые». Практически они представляют столбчатую диаграмму, уложенную на бок (рис. 16).

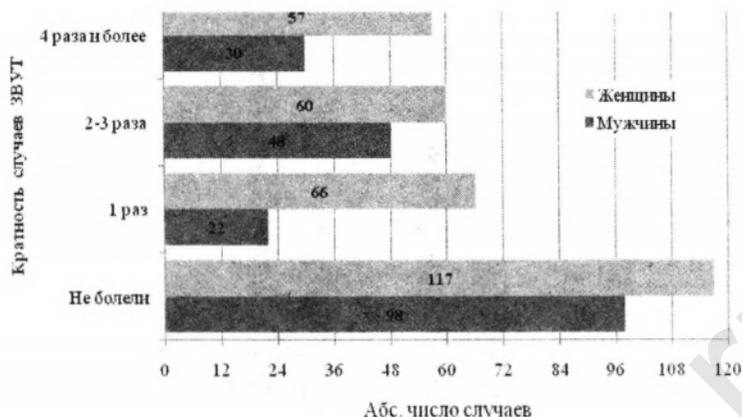


Рис. 16. Распределение мужчин и женщин по кратности заболеваний в году

Для отображения экстенсивных показателей, по аналогии с внутрисклонковыми, могут использоваться внутриленточные диаграммы, которые могут отображать как производные величины, так и абсолютные числа (рис. 17).

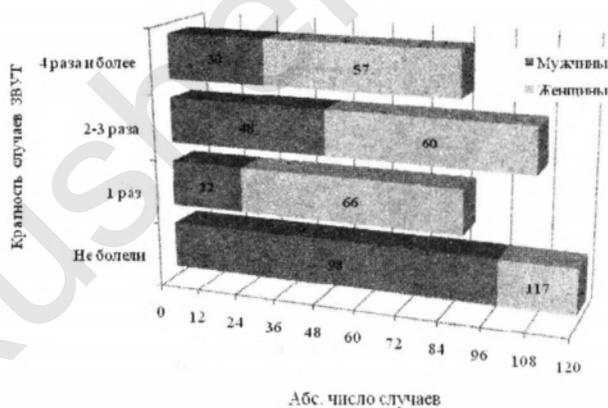


Рис. 17. Сравнительное распределение мужчин и женщин в группах разной кратности случаев ЗВУТ за год

Наибольшая компактность обеспечивается одной из разновидностей ленточной диаграммы – пирамидальной диаграммой. С ее помощью удастся отобразить большое количество сложнораспределенных данных. Наиболее

часто она используется для представления демографических данных (см. раздел «Медицинская демография», рис. 39 и 40).

Специальные диаграммы. Приемы их построения и принципы анализа отображаемых результатов имеют свои жесткие правила, отступление от которых может привести к серьезным погрешностям. К таким диаграммам относятся: *радиальная диаграмма сезонности, гистограмма, полигон распределения, огива, кумулята (кумулярующая кривая), диаграмма Парето.*

Построение *радиальной диаграммы сезонности* начинается с формирования окружности, длина радиуса которой принимается за среднегодовой уровень. Если показатели сезонности выражаются в процентах, то длина радиуса окружности принимается равной 100%. Далее по длине каждого из 12-ти радиусов, от центра к периферии, по числу месяцев в году последовательно по ходу часовой стрелки, начиная от «12 часов», откладываются значения сезонных показателей. (Подробнее о статистике показателей сезонности смотри в разделе «Динамические ряды»). При превышении среднегодовых показателей кривая выходит за внешние пределы окружности. В случае более низких уровней помесечных показателей кривая оказывается внутри пределов окружности. При правильном построении диаграммы площадь поверхности, которая выделена кривой, должна быть равна площади круга (табл. 17 и рис. 18).

Таблица 17

Показатели частоты кишечных инфекций в году

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

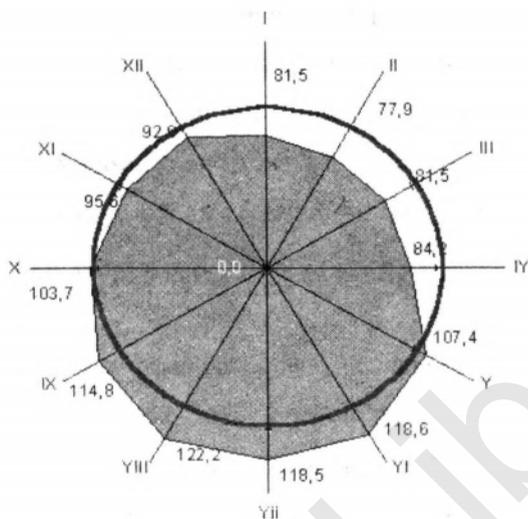


Рис. 18. Диаграмма сезонности острых кишечных инфекций

Гистограмма применяется для графического изображения интервальных рядов распределения. Позволяет качественно оценить различные распределения. 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 уровнем (частотой) характеристики группы.

При построении графического изображения гистограмм между отдельными четырехугольниками, образующими гистограмму, не должно быть интервалов (рис. 19). Необходимо помнить, что в гистограмме, в отличие от столбиковой диаграммы, общая площадь четырехугольников отражает один из статистических параметров распределения.

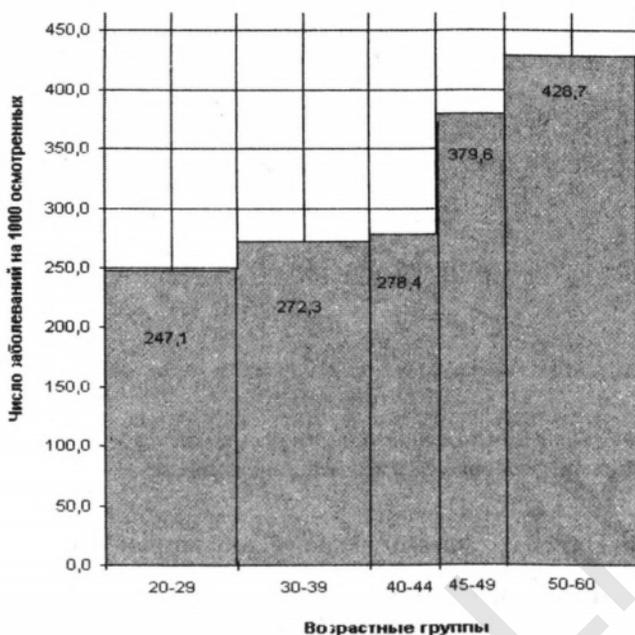


Рис. 19. Гистограмма. Уровень патологической пораженности работников предприятия в различных возрастных группах

В ряде случаев при статистическом анализе распределений целесообразно использовать ряд накопленных частот – кумулятивный ряд, который отображается специальной кривой, называемой «кумулятой». Она позволяет наглядно показать число случаев ниже или выше определенного уровня. Т.е. позволяет определить, какая часть совокупности обладает значениями, не превышающими заданного предела, а какая, наоборот, превышает этот предел. Наглядность кумуляты повышается, если представить ее в виде накопленных частот (частостей), то есть долей от общего числа наблюдений. Приведенный пример сравнительного кумулятивного распределения посещений врачей-специалистов свидетельствует, что работники в возрасте до 40 лет составляли около 15% всех посетителей офтальмолога. Эта же возрастная группа составляла почти 30% от всех посетителей терапевта (рис. 20).

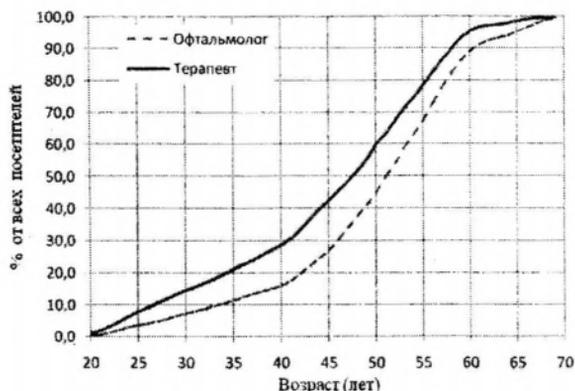


Рис. 20. Сравнительное изображение кумулят распределения возрастного состава посетителей врачей-специалистов

Диаграмма Парето – элемент одного из методов системного анализа (метод Парето), позволяющего выделить основные факторы, на которые надо влиять для достижения максимального успеха при минимальной затрате усилий. С помощью этого метода можно анализировать широкий круг проблем организации эффективной медицинской помощи.

Истоки метода Парето заложены трудами экономистов, занимавшихся проблемами земельной реформы в США в конце XIX столетия. В ходе исследования этой и связанных с ней других социально-экономических проблем в 2005 году Лоренц Макс Отто (Lorenz Max Otto) предложил использовать графические изображения, отображающие кумулятивные величины. В 1906 году итальянский экономист Вильфредо Парето (Vilfredo Pareto) использовал этот метод для научного обоснования выявленной им закономерности, согласно которой 80% благосостояния итальянского общества были обусловлены 20% общественного капитала. Соответственно метод распределения кумулятивных величин и соответствующее графическое отображение называются методом Парето и кривой Лоренца.

Обоснованная этими учеными закономерность неравномерного распределения затрат и результатов (доходов) в приложении к проблемам управления, в том числе управления здравоохранения, в общем виде формулируется следующим образом: в процессе работы 80% результата

достигается при контроле за 20% факторов, влияющих на этот результат. Или иначе: небольшое число наиболее существенных факторов регулирует самый большой процент целого (Верховодова О.В., 2007). В другом варианте:

- 80% потраченного времени (многочисленные второстепенные проблемы) дают 20% всего конечного результата;

- 20% потраченного времени (несколько жизненно важных проблем) = 80% конечного результата.

В качестве *примера* рассмотрим построение диаграммы Парето по условным данным о распределении услуг, оказанных по программе добровольного медицинского страхования (ДМС) работникам промышленного предприятия. Для удобства построения диаграммы необходимо сформировать таблицу исходных данных (табл. 18).

Таблица 18

Распределение медицинских услуг, оказанных врачами-специалистами по программе ДМС работникам предприятия

Код врача-специалиста	Номенклатура врачей-специалистов	Число услуг по ДМС	Доля услуг (в % к итогу)	Накопленный (суммарный) %
1	Терапевт	543	26,1	26,1
2	Невропатолог	229	11,0	37,1
3	Отоларинголог	158	7,6	44,7
4	Хирург	151	7,3	51,9
5	Эндокринолог	140	6,7	58,6
6	Гинеколог	125	6,0	64,6
7	Травматолог	118	5,7	70,3
8	Офтальмолог	86	4,1	74,4
9	Уролог	60	2,9	77,3
10	Педиатр	59	2,8	80,2
11	Кардиолог	50	2,4	82,6
12	Гастроэнтеролог	47	2,3	84,8
13	Дерматолог	45	2,2	87,0
14	Стоматолог	43	2,1	89,0
15	Аллерголог	36	1,7	90,8
16	Ревматолог	35	1,7	92,5
17	Ортопед	32	1,5	94,0
18	Онколог	30	1,4	95,4
19	Физиотерапевт	16	0,8	96,2
20	Проктолог	16	0,8	97,0
21	Инфекционист	10	0,5	97,5
22	Пульмонолог	9	0,4	97,9
23	Гематолог	8	0,4	98,3

24	Нейрохирург	7	0,3	98,6
25	Челюстно-лицевой хирург	7	0,3	98,9
26	Нефролог	6	0,3	99,2
27	Хирург сосудистый	6	0,3	99,6
28	Прочие	4	0,2	99,8
29	Маммолог	2	0,1	99,9
	Итого	–	100,0	100,0

В графах этой таблицы последовательно располагаются следующие данные:

– код (порядковый номер) учетного признака – специальность врачей, оказывавших медицинские услуги работникам предприятия в рамках программы ДМС;

– наименование учетного признака – полное наименование врачебной специальности;

– численные характеристики учетного признака – абсолютное число услуг по ДМС и доля услуг (в %), оказанных отдельными специалистами, к общей сумме услуг по ДМС;

– накопленный (кумулятивный, суммарный) процент.

При построении графика Парето по оси абсцисс откладывают учетные признаки (код врача-специалиста). По одной оси ординат количественные характеристики в абсолютных числах (число медицинских услуг по ДМС), по другой – накопленный процент (рис. 21).

Как явствует из представленных данных, около 80% всех медицинских услуг по ДМС были выполнены врачами 10 специальностей из 29 имеющихся в штате специалистов.

Такое распределение может служить основой для углубленного анализа причин, вызвавших относительное преобладание этих специалистов при оказании услуг по ДМС. Соответственно, рассмотреть возможности организации приема работников предприятия квалифицированными специалистами этого профиля на базе собственного ЛПУ, а оказание каких услуг экономически более целесообразно организовать во внешних медицинских учреждениях.

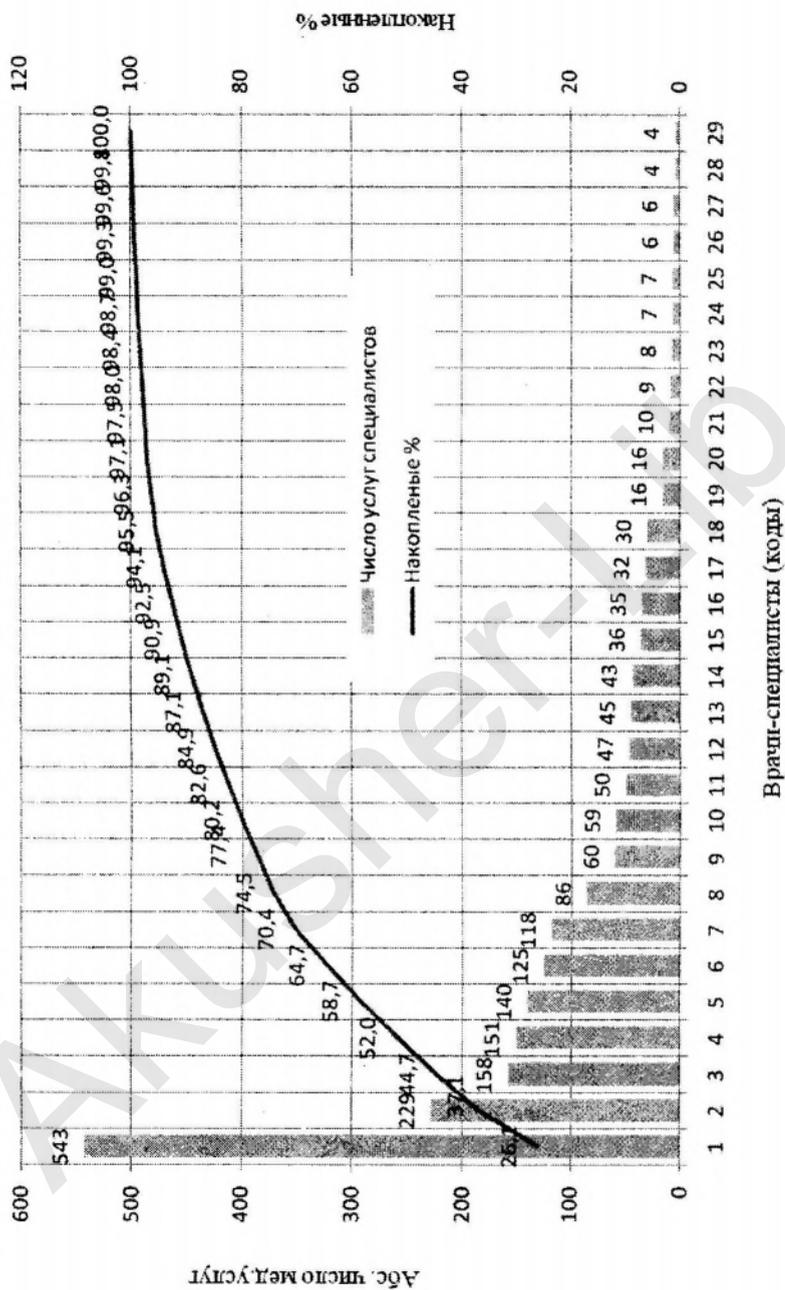


Рис. 21. Диаграмма Парето: Вклад отдельных врачей-специалистов в общую сумму услуг, оказанных по ДМС

2.5. Основы применения MS Excel при разработке статистических данных

Microsoft Excel – это весьма распространенная в РФ программа управления электронными таблицами, позволяющая обрабатывать числовые данные, создавать графики и анализировать информацию таблиц баз данных. К настоящему времени выпущено более 10 версий этой программы, и интерфейс, при помощи которого пользователь взаимодействует с *Excel*, в каждой из версий имеет свои особенности. В пособии будут представлены упрощенные варианты действий пользователя и элементы интерфейса, которые применимы к подавляющему большинству версий.

Для запуска *Excel* нажмите виртуальную 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 программы на экране появится окно *Excel* (рис. 22), центральную часть которого занимает так называемая *рабочая область*.

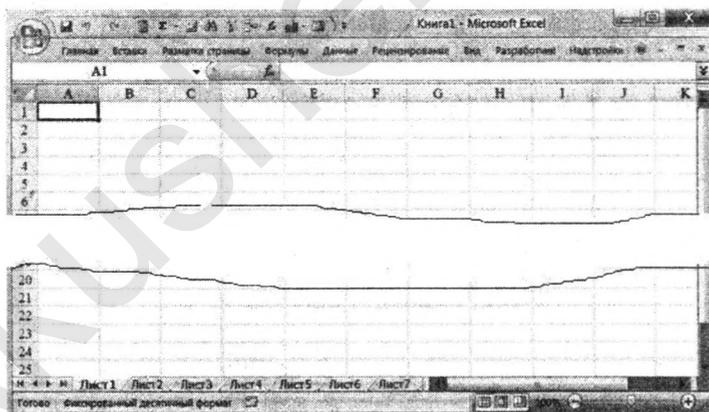


Рис. 22. Окно программы *Excel* 2007

В заголовке окна *Excel*, кроме названия программы, отражаются имя редактируемого файла и настраиваемая пользователем панель быстрого доступа. Вторая строка сверху в окне – это лента вкладок (лента пользовательского интерфейса), отражающая логические группы команд (рис. 23). Каждая вкладка отражает группы команд, связанные с видом

выполняемого действия (например, набором текста или компоновкой страницы). Чтобы увеличить рабочую область, некоторые вкладки можно выводить только при необходимости. Когда лента свернута, видны только вкладки.

Чуть ниже 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 вводимая информация.

Основную часть экрана занимает непосредственно электронная таблица, которая, по аналогии с шахматной доской, делится на строки и столбцы. Столбцы обозначаются буквами латинского алфавита, начиная от А, строки – цифрами, начиная от 1. Чем более высокая версия *Excel*, тем больше строк и столбцов может размещаться в таблице. Место пересечения строки и столбца называется ячейкой. Адресом ячейки служат буква и число, соответствующие пересекающимся в ячейке столбцу и строке. *Например*: В5 обозначает ячейку на пересечении столбца В и строки 5.

Один файл *Excel* может содержать множество 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: **Лист 1**, **Лист 2** и т. д. Щелкнув мышью на нужном ярлычке, вы вызовете на экран соответствующий рабочий лист.

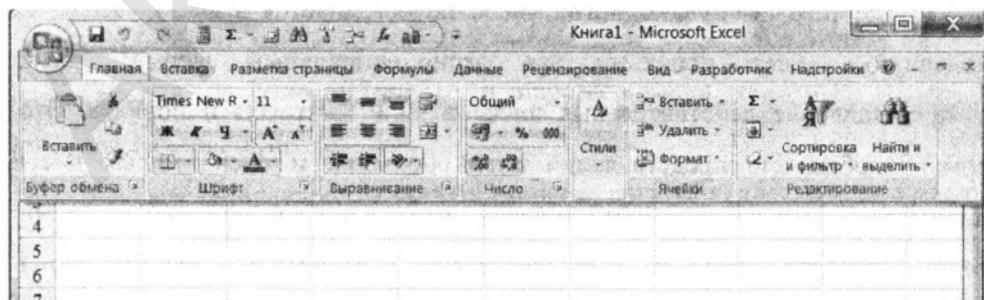


Рис. 23. Окно программы *Excel* 2007

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 виртуальной кнопки [Ввод], которая обычно 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 *диз* (#), при этом содержимое ячейки полностью будет представлено в строке формул.

Кроме того, в случаях, когда число не помещается в ячейку *Excel*, по умолчанию может использоваться экспоненциальный формат числа, т.е. представление действительных чисел в виде мантиссы и порядка. Этот формат удобен при представлении очень больших и очень малых чисел. Например: вместо числа 0,00012 в ячейке оказывается выражение 1,23E-04. Чтобы перевести число из экспоненциальной в обычную форму записи, необходимо, во-первых, расширить ячейку таблицы, во-вторых, изменить формат числа, *например*, через вкладку «Формат».

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 поместить 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

Уже на стадии предварительной статистической разработки данных могут быть полезны различные алгоритмы, позволяющие пополнять массивы результатов наблюдений вычисляемыми значениями регистрируемых признаков. *Например:* при исследовании функции дыхания целесообразно наряду с введением значений фактических и должных величин этой функции сразу обозначить и процентное отношение фактической к должной величин.

Ввод любой формулы должен всегда начинаться со знака = (равно) или со знака +. Формула может содержать обычные арифметические операторы: + (плюс), - (минус), * (умножить), / (разделить). *Например,* для получения в ячейке C1 суммы двух чисел, находящихся в ячейках A1 и B1, следует в ячейку C1 ввести формулу: =A1+B1.

Для задания формулы можно использовать различные технические приемы. Формула может быть задана путем ввода с клавиатуры. Однако здесь существует «подводный камень»: символы, используемые для обозначения строк, где содержится исходная информация (аргументы формулы), могут быть только латинскими (A, B, C, D и т.д.). Но многие из них внешне не отличаются от русских (А, В, С и т.п.), которые применяются в названии формул. Из-за этого часто возникают ошибки ввода адресов ячеек. Чтобы избежать таких ошибок, целесообразно использовать другой способ задания адресов: *например,* после ввода знака равенства в ячейку C1 следует выполнить щелчок мышью на ячейке, которая должна быть указана в аргументе формулы первой (A1). Адрес данной ячейки появится в итоговой ячейке (C1). Далее следует ввести оператор сложения, а затем выполнить щелчок на следующей ячейке (B1). Использование этого способа значительно упрощает ввод адресов ячеек.

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 указываются адрес левой верхней ячейки и через разделитель (точку или двоеточие) адрес правой нижней ячейки. Например: = СУММ (A1:C3) – сумма чисел, расположенных в 9 ячейках (A1, A2, A3, B1, B2, B3, C1, C2, C3).

Если в формуле указываются несмежные ячейки, то их адреса следует разделить точкой с запятой. Например: = СРЗНАЧ (A1;B3:C5) – среднее арифметическое чисел, расположенных в ячейке A1 и блоке ячеек B3:C5.

Название функции можно ввести в ячейку с клавиатуры. Аргументы функции должны быть указаны после ее названия в круглых скобках. Для этого после ввода открывающей круглой скобки следует выделить с помощью мыши блок ячеек, содержимое которого 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].

Кроме того, при вводе сложных функций удобно работать с «Мастером функций». Его вызов осуществляется через расположенную на ленте вкладок вкладку «Формулы». Например, вызов статистических функций, выполняющих операции по вычислению различных величин или параметров их распределений (рис 24). При этом сначала указывается категория функции (например, ДИСПР). Затем аргументы функции – диапазон ячеек (рис. 25).

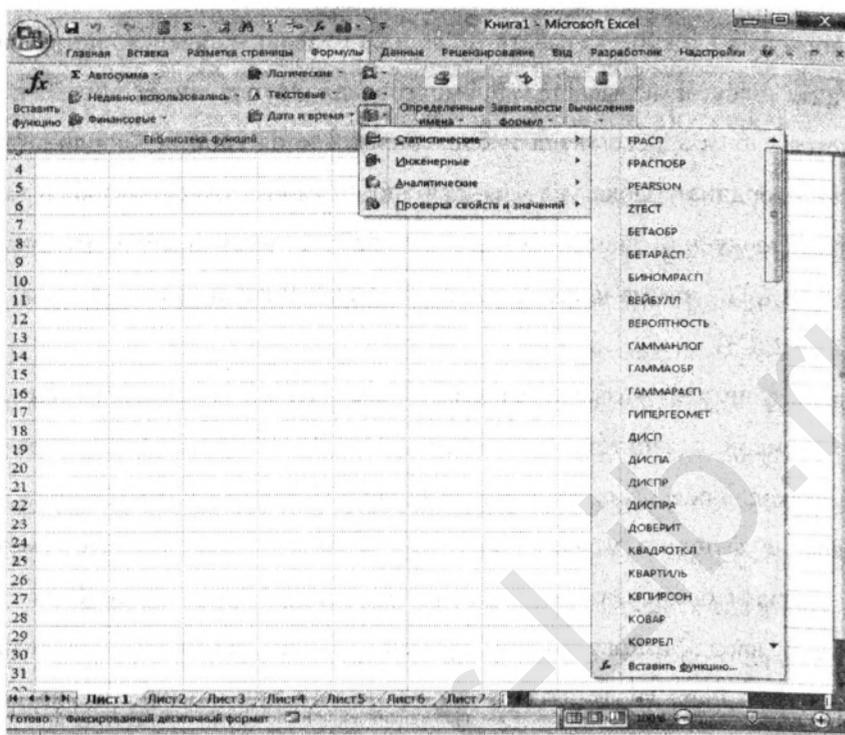


Рис. 24. Окно формул (статистические функции)

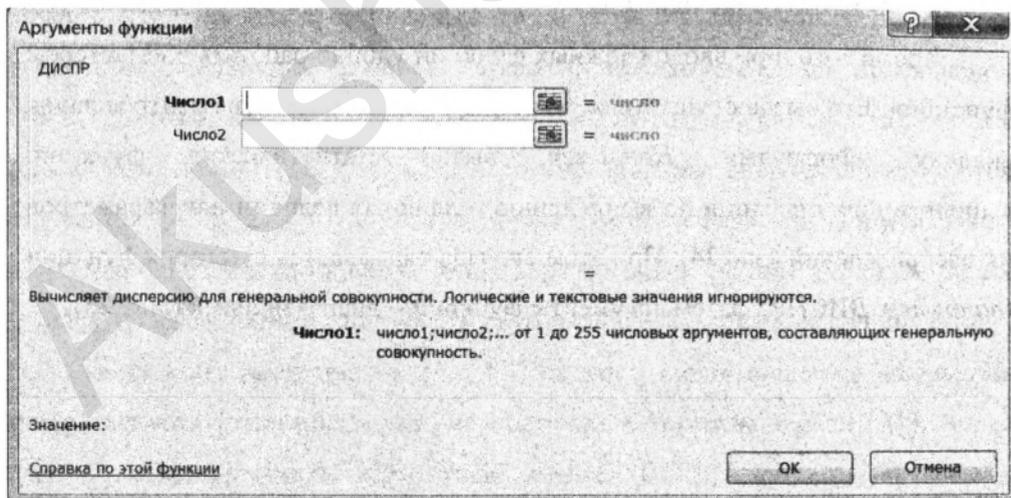


Рис. 25. Аргументы функции ДИСПР

2.5.4. Копирование и дублирование 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:

- установить указатель ячейки в ту ячейку, которую надо скопировать;
- выбрать команду <Копировать> из меню <Главная>;
- выделить интервал ячеек, в который нужно скопировать данные, и нажать клавишу [Enter].

При копировании формул 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 \$.

Для дублирования (копирования) формулы в соседние ячейки можно использовать также команду [Заполнение]. Для этого:

- выделите ячейку с исходной формулой;
- поместите курсор мыши на маркер заполнения (маленький квадратик) в правом нижнем углу копируемой ячейки. При этом курсор должен приобрести вид маленького черного крестика;
- перетащите мышью маркер заполнения по ячейкам, которые требуется заполнить.

2.6. Базы и хранилища данных.

Формирование и обработка таблиц данных

В программе *MS Excel* имеется специальный пакет функций, который позволяет создавать обширные массивы любых по содержанию данных в виде *таблиц данных*, производить из них отбор информации и вычислять производные величины, необходимые для статистического анализа. Обычно при обработке больших массивов данных, представляющих сложные структурные образования, требуются специальные логически структурированные объекты хранения и обработки информации: *базы данных* и *хранилища данных*. В настоящее время существует множество определений этих понятий, которые различаются в основном целевым предназначением баз и хранилищ информации, организацией хранения и используемым программным обеспечением. В наиболее общем виде можно считать, что *базы данных* – сформированные по определённым правилам совокупности данных, обеспечивающие удовлетворение информационных потребностей пользователей. Функционирование электронных баз данных (БД) обеспечивается с помощью специальных программных средств *систем управления базами данных* (СУБД).

Хранилища данных во многом схожие с БД понятия (тем более что формируются они, как правило, на основе последних), однако хранилища имеют свои особенности, среди которых, в первую очередь, можно выделить следующие:

- хранилища обычно имеют более узкую предметную ориентацию на накопление и обработку определенной информации;
- информация, заносимая в БД, как правило, мало пригодна для прямой статистической обработки большого числа одновременных наблюдений. Информация из хранилищ данных более пригодна для такого рода целей;
- информация, накапливаемая в хранилищах, не предназначена для текущего редактирования. Возможно только ее пополнение.

На первый взгляд эти различия несущественны, однако на практике статистическая обработка больших объемов динамически меняющихся данных возможна только на основе хранилищ данных. Особенно значимо это обстоятельство при попытках прямого использования корпоративных (производственных, учрежденческих и т.п.) баз данных для углубленного статистического анализа. Их прямое использование оправдано только для получения ограниченной по объему информации и может применяться только для предварительного, разведочного стат. анализа. Именно поэтому в СУБД, как правило, не предусматриваются функции глубокого стат. анализа.

Создание баз и хранилищ данных – сложные инженерно-информационные задачи, требующие участия специалистов, имеющих соответствующую подготовку. В настоящем издании будут рассмотрены приемы работы, доступные широкому кругу пользователей *MS Excel*, позволяющих без применения сложных СУБД и других программных средств проводить формирование, редактирование и получение информации из таблиц данных. Эти таблицы, являющиеся по своей сути основами баз или хранилищ данных, часто вполне достаточны для прямого решения текущих задач статистической разработки материалов медико-биологических и медико-социальных исследований.

2.6.1. Таблицы баз данных. Пополнение, редактирование, фильтрация и сортировка информации

Первичной структурированной совокупностью информации в любой базе или хранилище данных является **таблица данных**. Все учетные признаки в таблице данных всегда располагаются строго в определенном порядке. При формировании такой таблицы в первой строке обязательно указываются наименования столбцов (полей), куда будут заноситься значения соответствующих учетных признаков. Наименования столбцов следует всегда указывать в одну строку. Отдельная клетка этой строки, которая является заголовком конкретного столбца, называется **Поле** базы данных. В данном

примере (рис. 26), где база данных состоит из одной таблицы, организованной в *MS Excel*, используются следующие поля: возраст, пол, профессия, ЖЕЛ факт. (фактическая жизненная емкость легких), ЖЕЛ долж. (должная жизненная емкость легких) и Ф/Д (отношение фактической к должной жизненной емкости легких).

	A	B	C	D	E	F	G	H
1	ФИО	Возраст	Пол	Профессия	ЖЕЛ факт	ЖЕЛ долж.	Ф/Д	
2	Севастьянов С.Н.	54	М	Машинист насос.уст.	3,14	3,01	104,3%	
3	Рогов Ю.Н.	33	М	Мастер участка	2,52	3,11	81,0%	
4	Сапегин А.А.	26	М	Слесарь	3,57	3,21	111,2%	
5	Дедуля И.А.	50	М	Мастер участка	2,76	2,93	94,2%	
6	Ушреев Е.В.	34	М	Машинист насос.уст.	3,81	3,25	111,1%	
7	Трофимов Г.И.	36	М	Водитель	3,21	3,51	91,5%	
8	Алексеев А.А.	50	М	Инженер	3,52	3,06	115,0%	
9	Абанин С.Н.	38	М	Оператор котельной	2,99	3,26	91,7%	
10	Александрова И.С.	29	Ж	Оператор котельной	2,56	3,21	79,8%	
11	Деева А.И.	45	Ж	Инженер	2,59	3,02	85,8%	
12	Зацепина В.М.	53	Ж	Кочегар котельной	2,09	2,53	82,6%	
13	Юринова Г.Н.	21	Ж	Оператор котельной	2,55	3,25	78,5%	
14	Абаканович Т.Л.	50	Ж	Уборщик	3,78	3,06	123,5%	
15	Богомолова Л.М.	35	Ж	Машинист насос.уст.	2,05	3,24	63,3%	
16	Панина А.Г.	42	Ж	Оператор котельной	3,17	3,08	102,9%	
17	Принцева К.С.	40	Ж	Слесарь	2,78	3,26	85,3%	
18	Тарабукина И.Г.	22	Ж	Инженер	3,28	3,67	89,4%	
19								

Рис. 26. Пример таблицы базы данных

Все учетные признаки по каждой отдельной единице наблюдения записываются в отдельную строку. Содержимое этой строки называется *запись*. В данном примере простейшей базы данных, состоящей из одной таблицы, представлено 16 записей, расположенных в строках, начиная со 2 по 17. В строке 2 записаны данные первой единицы наблюдения: Севастьянов С.Н., в полях этой записи размещены учетные признаки: возраст – 49 лет, пол – мужской и т.д. В строке 3 – вторая единица наблюдения: Рогов Ю.Н., 33 года и т.д.

База данных может постоянно пополняться или модифицироваться. *Например:* в качестве исходной информации необязательно 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 $D2/E2*100$ или в эту клетку ввести сокращенный вариант формулы $D2/E2$ и обозначить через вкладку <Главная> формат клетки % и точность отображения информации (разряд числа) на уровне десятичного знака;

- скопировать содержимое клетки F2 в блок ячеек F3:F17.

После того, как исходная информация размещена на первом листе *MS Excel* и в нижней части листа установлен переключатель <Лист1>, можно приступить к работе с этой таблицей базы данных. Обратите внимание, что перед началом обработки данных целесообразно сохранить сформированную таблицу в отдельном файле долговременной памяти (диск С:, флеш-память и т.п.) с затрудненным для пользователя доступом (обычно удаленном). Это позволит уберечь исходную информацию при любом неблагоприятном развитии событий в процессе обработки данных (случайные удаления фрагментов баз данных, аварийные отключения компьютера и т.п.).

Значительно упростить формирование и редактирование таблицы базы данных в *MS Excel* можно, пользуясь окном формы, в котором данные представлены как картотека (в каждый момент отображается запись только одной единицы наблюдения). В этом окне можно вести поиск записей по заданному критерию, дополнять список новыми записями, удалять ненужные, а также редактировать записи. Кнопки [Форма] нет на ленте вкладок, однако эту функцию можно использовать в *Excel*, добавив кнопку [Форма] на панель быстрого доступа. Для этого:

- щелкните стрелку на панели быстрого доступа и выберите элемент <Другие команды>;

- в поле <Выбрать команды из> щелкните элемент [Все команды];

- в списке выберите кнопку [Форма] и щелкните элемент [Добавить].

В окне *Формы данных* указываются общее количество записей в списке и позиция текущей записи (рис. 27). 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 сразу восстановить с помощью кнопки [Вернуть].

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ФИО	Возраст	Пол	Профессия	ЖЕЛ факт	ЖЕЛ дозе	Ф/Д							
2	Севастьянов С.И.	54	м	Машинист насос.уст.	3,14	3,01	104,3							
3	Рогов Ю.Н.	33	м	Мастер участка	2,52	3,11	81,0							
4	Сатепин А.А.	26	м	Слесарь	3,57	3,21	111,2							
5	Делуза И.А.	50	м	Мастер участка	2,76	2,93	94,2							
6	Ушрев В.В.	34	м	Машинист насос.уст.	3,61	3,25	111,1							
7	Трофимов Г.И.	36	м	Водитель	3,21	3,51	91,5							
8	Алексеев А.А.	50	м	Инженер	3,52	3,06	115,0							
9	Абанин С.Н.	68	м	Оператор котельной	2,99	3,26	91,7							
10	Александрова И.С.	29	ж	Оператор котельной	2,56	3,21	79,8							
11	Дева А.И.	45	ж	Инженер	2,59	3,02	85,8							
12	Зацепина В.М.	53	ж	Кочегар котельной	2,09	2,52	82,9							
13	Юркина Г.Н.	21	ж	Оператор котельной	3,55	3,25	78,5							
14	Абакамович Т.Л.	50	ж	Уборщик	3,78	3,06	123,5							
15	Богомолова Л.М.	35	ж	Мастер участка	2,05	3,24	63,3							
16	Павина А.Г.	42	ж	Машинист насос.уст.	3,17	3,08	102,9							
17	Привцева К.С.	40	ж	Оператор котельной	2,78	3,26	85,3							
18	Гарбукина И.Г.	22	ж	Инженер	3,28	3,67	89,4							
19														
20														
21														
22														
23														

Рис.27. Окно формы таблицы (базы данных)

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. 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 <. Для решения поставленной задачи в качестве критерия поиска в поле Пол введите Ж, а в поле Возраст – <40.

Сортировка данных в диапазоне или таблице позволяет выполнять сортировку данных по тексту (от А к Я или от Я к А), числам (от наименьших к наибольшему или от наибольших к наименьшему), а также датам и времени (от старых к новым или от новых к старым) в нескольких столбцах. Можно также выполнять сортировку по цвету ячеек, цвету шрифта, а также по значкам. Большинство сортировок применяется к столбцам, но возможно также применить сортировку к строкам.

Сортировка текстовых значений:

выберите столбец с алфавитно-цифровыми данными в диапазоне (блоке) ячеек или убедитесь, что активная ячейка находится в столбце, который содержит эти данные;

на вкладке <Данные> выберите пункт <Сортировка> (рис. 28).

Для сортировки ФИО по возрастанию выберите вариант <Сортировка значения> и <Порядок от А до Я>. Для сортировки по убыванию – вариант <Порядок от Я до А>. После выбора необходимого варианта щелкните правой клавишей мыши.

	A	B	C	D	E	F	G	H	I	J	K	L
1	ФИО	Возраст	Пол	Профессия	Жел факт	Жел долж	Ф/Д					
2	Севастьянов С.Н.	54	М	Машинист насос. уст.	3,14	3,01	104,3%					
3	Рогов Ю.Н.	33	М	Мастер участка	2,52	3,11	81,0%					
4	Александрова И.С.	29	Ж	Оператор котельной	2,56	3,21	79,8%					
5	Салегин А.А.	26	М	Слесарь	3,57	3,21	111,2%					
6	Деева А.И.	45	Ж	Инженер	2,59	3,02	85,8%					
7	Зацепина В.М.	53	Ж	Кочегар котельной	2,09	2,53	82,6%					
8	Дедуля И.А.	50	М	Мастер участка	2,76	2,93	94,2%					
9	Ушрев Е.В.	34	М	Машинист насос. уст.	3,81	3,25	111,1%					
10	Юринова Г.Н.	21	Ж	Оператор котельной	2,55	3,25	78,5%					
11	Трофимов Г.И.	36	М	Водитель	3,21	3,51	91,5%					
12	Абакамович Т.Л.	27	Ж	Уборщик	3,78	3,06	123,5%					
13	Алексеев А.А.	21	М	Инженер	3,52	3,06	115,0%					
14	Богомолова Л.М.											
15	Абанин С.Н.											
16	Панина А.Г.											
17	Принцева К.С.											
18	Теребукина И.Г.											
19												
20												
21												
22												
23												
24												
25												
26												
27												
28												
29												
30												

Рис. 28. Пример сортировки всех записей по полю ФИО

Сортировка чисел:

- выберите столбец с цифровыми данными в диапазоне ячеек или убедитесь, что активная ячейка находится в столбце, который содержит цифровые данные;

- на вкладке <Данные> выберите пункт <Сортировка>;
- для сортировки чисел в столбце Возраст по возрастанию выберите вариант <Сортировка значения> и <Порядок по возрастанию>.

Сортировка одного столбца в диапазоне ячеек без связи с другими столбцами требует соблюдения осторожности, т.к. может привести к нежелательным результатам, например, перемещению ячеек в этом столбце независимо от других ячеек строк и столбцов. Таким образом, происходит практическое изменение всей структуры (перемешивание) исходной базы данных. Поэтому при попытке произвести такого рода сортировку появляется сообщение «Обнаружены данные вне указанного диапазона», содержащее варианты предполагаемого действия: либо автоматически расширить диапазон, т.е. провести обычную сортировку всего массива ячеек (всех записей); либо сортировать в пределах указанного диапазона, т.е. только одного или нескольких выделенных столбцов (рис. 29).

	A	B	C	D	E	F	G	H	I	J
1	ФИО	Возраст	Пол	Профессия	ЖЕЛ факт	ЖЕЛ долж	Ф/Д			
2	Севастьянов С.Н.	54	М	Машинист насос.уст.	3,14	3,01	104,3%			
3	Рогов Ю.Н.	33	М	Мастер участка	2,52	3,11	81,0%			
4	Сапегин А.А.	26	М	Слесарь	3,57	3,21	111,2%			
5	Дедуля И.А.	50	М	Мастер участка	2,76	2,93	94,2%			
6	Ушреев Е.В.	34	М	Машинист насос.уст.	3,61	3,25	111,1%			
7	Трофимов Г.И.	36	М	Водитель	3,21	3,51	91,5%			
8	Алексеев А.А.	50	М	Инженер	3,52	3,06	115,0%			
9	Абанин С.Н.	38	М	Оператор котельной						
10	Александрова И.С.	29	Ж	Оператор котельной						
11	Деева А.И.	45	Ж	Инженер						
12	Зацепина В.М.	53	Ж	Кочегар котельной						
13	Юринова Г.Н.	21	Ж	Оператор котельной						
14	Абаканович Т.Л.	50	Ж	Уборщик						
15	Богомолова Л.М.	35	Ж	Машинист насос.уст.						
16	Панина А.Г.	42	Ж	Оператор котельной						
17	Принцева К.С.	40	Ж	Слесарь						
18	Тарабукина И.Г.	22	Ж	Инженер						
19										
20										

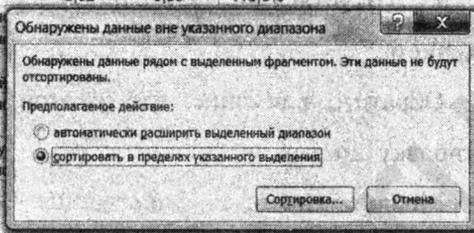


Рис.29. Пример сортировки столбца Пол

Для того, чтобы провести сортировку ячеек одного столбца без влияния на другие столбцы:

- выберите столбец в базе данных, содержащий два или более столбца;
- для выбора необходимого столбца щелкните его заголовок;
- на вкладке <Сортировка> и сообщения <Обнаружены данные вне указанного диапазона> выберите вариант сортировать в пределах указанного выделения;
- в диалоговом окне <Сортировка> задайте остальные параметры сортировки, а затем нажмите кнопку [ОК];
- если результат не соответствует желаемому, нажмите кнопку [Отменить].

Фильтрация. С помощью фильтрации можно производить отбор записей, удовлетворяющих заданному критерию.

Чтобы отфильтровать список, необходимо выполнить следующие действия:

- выбрать команду <Фильтр> 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 в этой вкладке обозначены команды и на сортировку данных.

Можно выполнять фильтрацию данных одновременно по нескольким полям, последовательно задавая команды на фильтрацию по разным столбцам. В том числе выполнять отбор записей, формируя сложные логические условия.

2.6.2. Сводные таблицы в MS Excel

Одним из наиболее удобных инструментов статистической разработки данных в *MS Excel* являются *Сводные таблицы*. Эти таблицы запускаются через вкладку основного меню <Данные>. Для размещения сводной таблицы лучше всего использовать новый лист. С этой целью переключите указатель в нижней части листа *MS Excel* в положение Лист 2 (при условии, что на Лист 1 размещена исходная таблица данных). Затем установите курсор в верхний левый угол Листа 2 (не выше позиции А 2). После нужно вызвать пакет *Сводные таблицы*. На экране появится первое окно мастера создания сводной таблицы (рис. 30).

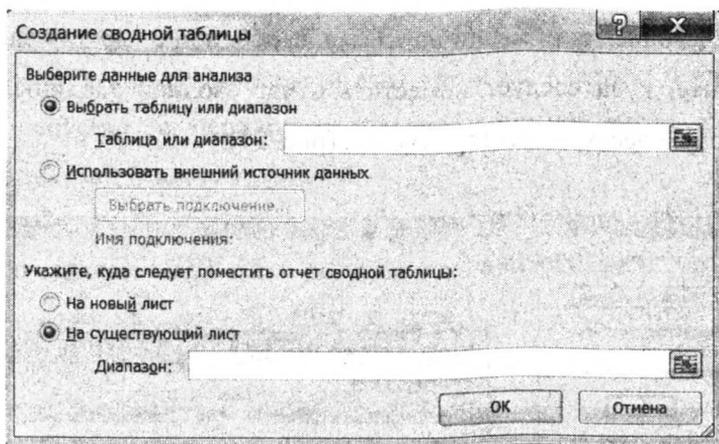


Рис. 30. Первое окно мастера создания сводных таблиц

Окно мастера сводных таблиц может располагаться непосредственно на листе, с левого или правого края, снизу или сверху листа. Для того чтобы переместить его в наиболее удобное для работы место, нужно установить указатель мыши на поверхность окна мастера и, нажав левую клавишу мыши, перетащить его в нужное место. Затем последовательно выполните следующие действия.

В открывшемся окне мастера сводных таблиц укажите источник информации, откуда будут поступать исходные данные. В данном примере необходимо указать: **в таблице или диапазоне MS Excel.**

Затем укажите диапазон (блок ячеек) исходных данных, куда обязательно должна попасть первая строка с именами полей базы данных. Если на листе, где располагается база данных, нет другой информации, то можно не выделять блок ячеек, содержащий таблицу данных, а, щелкнув мышкой в левый верхний угол листа, выделить весь лист как источник информации. В этом случае необходимо иметь в виду, что если таблица данных велика, т.е. содержит очень большое число полей (столбцов) и строк (записей), то вывод информации в сводную таблицу будет занимать несколько большее время.

После чего щелчком мыши по клетке, свободной от других записей части листа, укажите, куда следует поместить отчет сводной таблицы. На экране появится следующее окно мастера таблиц (рис. 31).

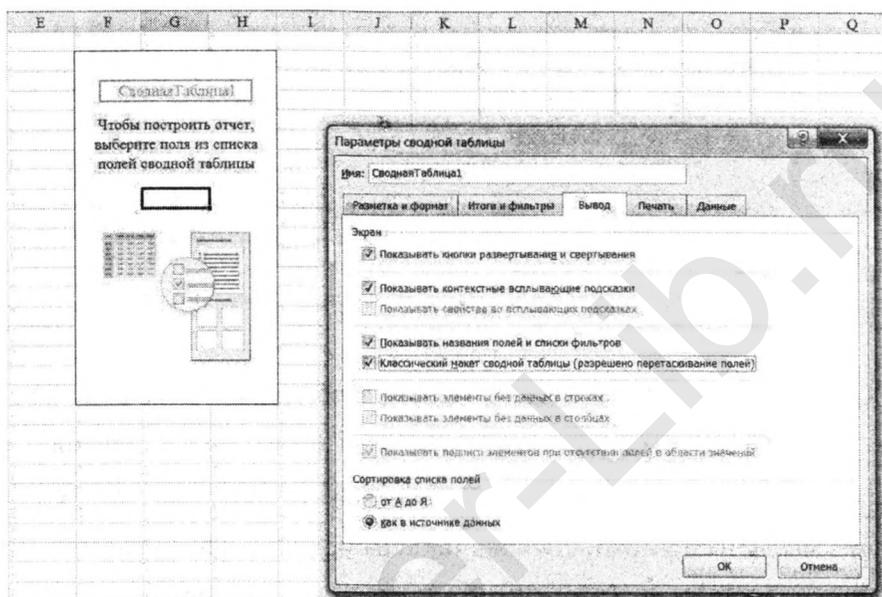


Рис. 31. Переключение отчета сводной таблицы на классическую версию

В этом окне, справа, указан перечень полей, которые можно включать в таблицу (список полей сводной таблицы). В версиях 2003, 2007 и 2010 гг. это окно мастера таблиц имеет не это, а другое оформление (интерфейс). Существенно отличаясь от «классического» интерфейса предыдущих версий по существу выводимой на экран информации, новые варианты отчета сводных таблиц мало чем отличаются от старых. Таким образом, чтобы сохранить единообразие выводимой информации, рекомендуется пользоваться старой, классической версией оформления отчета. Для того чтобы вызвать эту версию, установите указатель мышки на окно отчета таблицы и, щелкнув правой клавишей, вызовите закладку «Параметры сводной таблицы», «Вид». Затем отметьте опцию «Классический отчет сводной таблицы» и нажмите клавишу <OK> ().

После того, как откроется очередное окно мастера сводных таблиц, 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. 32.

Признак Пол нужно расположить в области столбцов, признак Профессия — в области строк, признак ЖЕЛ факт. располагается в центральной области данных. Обратите внимание: в сводной таблице (рис. 33) предусматривается вывод информации в виде средних значений по полю ЖЕЛ факт.

Если выводится другая информация (сумма и т. д.), тогда установите указатель мыши в область сводной таблицы, где расположены числовые результаты, щелкните правой клавишей и в открывшемся окне <Параметры поля значений> выберите операцию <Среднее> и нажмите клавишу [OK.]. На экране появится групповая таблица распределения в средних величинах. Необходимый числовой формат этих величин устанавливается через <Числовой формат>. Аналогичным образом можно использовать для представления данных и другие производные величины (дисперсии, среднеквадратические отклонения и т.п.).

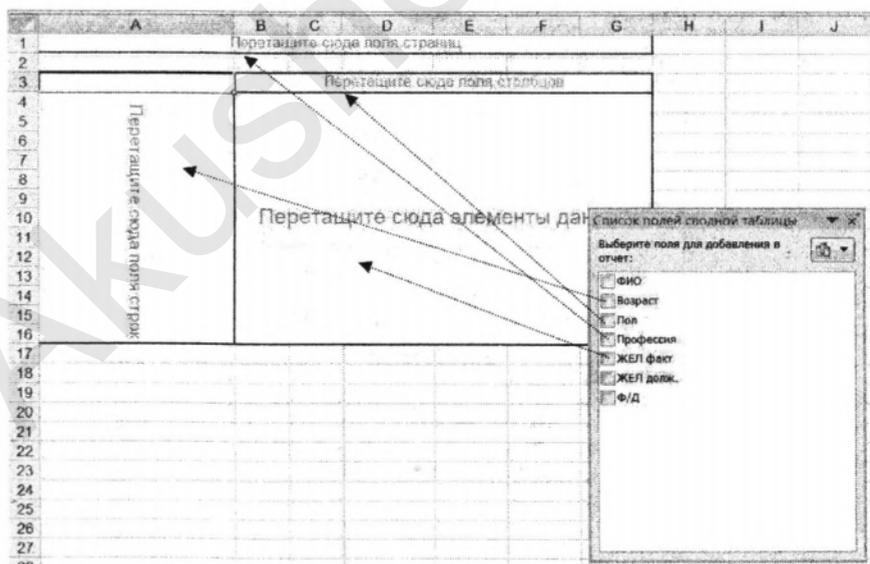


Рис.32. Второе окно мастера сводных таблиц

Вычисление некоторых производных величины устанавливается через опцию «Дополнительные вычисления» в окне «Параметры поля значений». 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. Полученная таблица считается групповой потому, что в ней приводится распределение по двум учетным признакам: Пол и Профессия.

Если установить признак Возраст выше полученной сводной таблицы (поле страницы), то можно рассматривать результаты распределения с учетом того или иного возраста. Таким образом, распределение будет уже учитывать три группировочных признака. Таблица в этом случае становится комбинационной. Для того чтобы выбрать с помощью переключателя тот или иной диапазон признака, установите указатель мыши на значок ▼, располагающийся рядом с названием строки, столбца или поля страницы, и щелкните правой клавишей. Затем выберите из появившегося перечня нужные вам диапазоны признака.

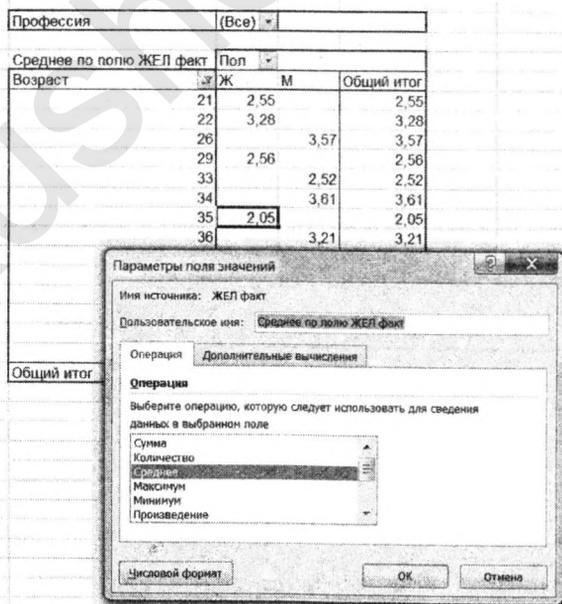


Рис. 33. Третье окно мастера создания сводных таблиц

Появившиеся в столбце таблицы значения признака **Возраст** показывают распределение по достаточно мелким возрастным интервалам, т.е. практически по каждому отдельно взятому возрасту. Для объединения всех возрастных интервалов в 3 группы (до 30 лет, от 30 до 39 лет и от 40 лет и старше) необходимо выполнить следующие операции: выделить мышкой клетки столбца **Возраст**, в которых указан возраст 21, затем, нажав правую клавишу мыши, не отпуская ее, выделить диапазоны возраста: 21, 22, 26, 29.

После этого, оставив указатель мыши в выделенном диапазоне, нажмите на правую клавишу мыши, затем на клавишу закладки с командой <Сгруппировать>. В таблице появятся еще один столбик **Возраст 2** и в нем **Группа 1**. Аналогичным образом проведите группировку остальной части возрастных интервалов.

После этого в таблице будет оставлен только столбик сгруппированных данных (рис. 34). Названия **Группа 1**, **Группа 2** и **Группа 3** можно изменить. Для этого выделите клетку таблицы с названием, нажмите на клавишу компьютера F2 и введите нужное вам название, удалив все ненужное.

Профессия	(Все) ▾		
Среднее по полю ЖЕЛ факт	Пол ▾		
Возраст2	<input checked="" type="checkbox"/> Ж	<input type="checkbox"/> М	Общий итог
Группа1	2,80	3,57	2,99
Группа2	2,05	3,08	2,88
Группа3	2,88	3,14	2,96
Общий итог	2,76	3,17	2,94

Рис.34. Сводная таблица сгруппированных по возрасту данных

3. Общественное здоровье как предмет статистического анализа

Оценка здоровья человека и закономерностей его трансформации является одной из фундаментальных проблем медицины. Медицинская статистика, обладая своей уникальной методологией, служит особым инструментом решения этой проблемы, позволяющим взглянуть на объективную реальность с помощью своих показателей, называемых показателями *общественного здоровья*.

В статистике общественного здоровья довольно часто встречаются выкладки, получение и интерпретация которых определены рядом особенностей, обусловленных спецификой медицины как отрасли деятельности человека и отрасли знаний. На эти особенности следует обратить внимание, поскольку они имеют большое значение для правильного понимания информационной значимости показателей общественного здоровья. Учитывая, что неумелое, а подчас и просто безграмотное, пользование данными автоматизированного статистического наблюдения имеет определенное распространение в отечественных и переводных публикациях, следует более детально рассмотреть особенности получения и интерпретации статистической информации о здоровье населения.

Отечественная статистика имеет богатый научно-практический опыт проведения широких, комплексных социально-гигиенических, эпидемиологических и т.п. исследований здоровья больших групп населения. Немалый опыт организации, правда, не всегда успешной, массового динамического наблюдения за здоровьем определенных групп населения. Нужно особо отметить, что такого рода масштабные работы за рубежом или вообще не проводились, или проводились в весьма ограниченных объемах. Поэтому настоятельно рекомендуется прежде, чем обращаться к зарубежным авторам, детально ознакомиться с имеющимся опытом статистических исследований общественного здоровья в России или в прошлом СССР.

Это избавит от повторения ошибок, которые давно пройдены отечественной статистикой, и облегчит понимание реальных преимуществ и недостатков множества статистических методик, предлагаемых зарубежными авторами.

Несмотря на запутанную терминологию, осложненную не всегда качественным переводом с одного европейского языка на другой (или на русский язык), эти методики часто не несут ничего принципиально нового или декларируют весьма спорные вещи. *Например:* в зарубежной статистике почти повсеместно при описании распространенности тех или иных заболеваний используются термины *эпидемиологическое изучение, популяция* и т.п.

Сущность чрезвычайно распространившегося **эпидемиологического подхода** при решении проблем неинфекционной патологии состоит в том, что делаются попытки решать проблемы распространения неинфекционных болезней таким же образом, как были решены проблемы инфекционной заболеваемости. Поскольку весьма успешная борьба с массовыми неинфекционными заболеваниями в XX веке строилась на основе учета ясных и конкретных закономерностей распространения патологии, а также конкретных методов специфической профилактики, априори предполагалось, что эти методы достаточно универсальны, чтобы решать и остальные проблемы общественного здоровья. Кроме того, среди причин, способствовавших распространению такого рода эпидемиологического подхода, нельзя не отметить слабое знание энтузиастами этого термина социальной гигиены и эпидемиологии (недаром специалисты-эпидемиологи практически не применяют термин «эпидемиология» вне рамок инфекционных болезней). Определенное влияние здесь оказали и публикации исследователей из стран Западной Европы и Северной Америки, где социально-гигиеническим исследованиям никогда не уделялось столько внимания, как в России и в СССР. «В угаре» такого рода перестройки «эпидемиологическими» нередко стали называться любые исследования распространенности болезней (и не только болезней), методики изучения роли сложнейших и до конца не ясных

причинно-следственных отношений неинфекционных заболеваний, отдельных компонентов образа жизни людей, возникновение и распространение буквально всех напастей, обрушивающихся на современного человека. Например: «Эпидемиология алкоголизма», «Эпидемиология травматизма... у теннисистов» (Abrams GD, Renstrom PA, Safran MR, 2012) и даже «Эпидемиология... беременности»!!! (сайт «Родитель.РУ», Сабурин А, 2012).

Однако, по мнению специалистов, занимающихся проблемами общественного здоровья, массовые неинфекционные заболевания – суть проявления глубинных процессов, которые не поддаются профилактическим мероприятиям, построенным на эпидемиологическом подходе.

Что касается другого термина «**популяция**», впервые он стал применяться в биологии, где популяцией называли и называют поныне совокупность особей одного вида, занимающую одну территорию и воспроизводящую себя в нескольких поколениях. Таким образом, перенос этого термина на сообщество людей явно не корректен.

В классической медицинской статистике основными источниками получения информации о здоровье тех или иных контингентов населения являются:

- показатели физического развития;
- заболеваемость;
- медико-демографические данные естественного движения населения (смертность, рождаемость, естественный прирост).

В последние десятилетия этот перечень дополнен таким источником информации об общественном здоровье, как очный или заочный *опрос*.

3.1. Принципы статистики физического развития

Физическое развитие – комплекс морфологических и функциональных показателей, которые определяют физическую работоспособность и биологическое состояние человека. Характеристика физического развития в зависимости от ее целевого предназначения может включать множество

элементов: физическую силу, телосложение, общую и мышечную массу тела, уровень аэробной и анаэробной мощности, силу скелетной мускулатуры, мышечную выносливость, координацию движений, двигательную активность, ее мотивацию и др. Физическое развитие отражает не только фиксированный на определенный момент времени статус, но и динамику процессов роста или, наоборот, деструкцию организма, так как практически любое заболевание тормозит процесс физического развития или наделяет его патологическими чертами, что особенно заметно у детей и подростков. С другой стороны, течение и исход болезни, способность противостоять неблагоприятным факторам среды во многом зависят от уровня физического развития индивидуумов.

В целом физическое развитие является одним из самых информативных критериев здоровья. Групповые показатели физического развития употребляются для обобщенных оценок здоровья различных категорий населения. Широко используются и индивидуальные оценки физического развития, например, в клинической практике для оценки здоровья отдельно взятого пациента или тяжести отдельных патологических состояний. Здесь, а также в спортивной, авиационной и космической медицине все большее развитие получают психофизиологические характеристики развития (выносливость, способность противостоять перегрузкам и т.п.), с помощью которых оцениваются резервные возможности человека или выявляются скрытые патологические состояния.

Физическое развитие человека динамически изменяется на каждом возрастном этапе. Эти изменения существенно зависят от социальных, экономических, санитарно-гигиенических и других условий, определяющих здоровье человека. Влияние факторов внешней среды на показатели физического развития наиболее выражено в так называемые сензитивные возрастные периоды, например:

- в грудном и подростковом возрасте, когда особенно интенсивно протекают процессы роста и развития;

- в пожилом и старческом возрасте, для которого характерны процессы деструкции, инволюции.

На основе более детального учета сензитивных периодов детей устанавливаются *этикризные сроки* – это промежутки времени, через которые в РФ проводится обязательная комплексная оценка состояния здоровья. Например, для детей 1-3 года жизни:

- на 1-м году жизни – раз в месяц;
- на 2-м году – раз в 3 месяца;
- на 3-м году – раз в 6 месяцев.

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 в виде комплекса морфофункциональных показателей. Морфофункциональные показатели могут включать целый комплекс разнообразных характеристик: *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 и повсеместно применяющихся на практике комплексных характеристик физического развития является такая характеристика, как *конституция*. Она отражает не только особенности телосложения, но также и нервно-психическую деятельность, метаболизм, вегетативные реакции, адаптационные, в том числе и патологические реакции индивидуума. В настоящее время не существует единой, общепринятой схемы конституциональной типологии. В РФ наиболее широко применяется схема 3-х конституциональных типов по М.В. Черноруцкому (1927):

- *нормостенический* тип характеризуется пропорциональными размерами тела и гармоничным развитием костно-мышечной системы. Нормостенической форме грудной клетки соответствует пропорциональное соотношение между переднезадними и поперечными ее размерами; умеренная

выраженность над- и подключичных пространств; плотно прилегающие к грудной клетке лопатки. При этом межреберные пространства выражены не резко; надчревный угол приближается к прямому и равен приблизительно 90° ;

- *астенический* тип отличается стройным телом, слабым развитием мышечной системы, преобладанием длины конечностей над длиной туловища, преобладанием над размерами живота продольных размеров тела и размеров грудной клетки, передне-задний размер которой уменьшен по отношению к поперечному. При этом над- и подключичные пространства западают, лопатки отстоят от грудной клетки, край X ребра свободен и легко определяется при пальпации; надчревный угол острый – менее 90° ;

- *гиперстенический* тип отличается хорошей упитанностью, длинным туловищем и короткими конечностями, относительным преобладанием поперечных размеров тела и размеров живота над размерами грудной клетки; поперечный разрез приближается к кругу; межреберные промежутки узкие, над- и подключичные пространства слабо выражены. Надчревный угол тупой – больше 90° .

Одним из неоспоримых доказательств влияния внешних факторов на физическое развитие человека является наблюдавшаяся в последнее столетие в развитых странах акселерация.

Акселерация – ускорение по сравнению с предыдущими поколениями темпов роста и развития, в первую очередь, детей. Увеличение размеров тела ребенка и наступление полового созревания в более ранние сроки. Большинство специалистов считает, что это явление обусловлено комплексом генетических и внешних факторов, среди которых ведущая роль принадлежит социально-экономическим условиям жизни.

Ретардация – явление, обратное акселерации, – замедление биологического развития организма ребенка. Проявляется в виде снижения темпов роста показателей физического развития. К концу пубертатного возраста ретарданты, как правило, догоняют своих сверстников по календарному возрасту.

В последнее десятилетие XX века в России начало отмечаться заметное ухудшение физического развития подрастающего поколения. В 1,5 раза увеличилось число школьников с дефицитом массы тела, более чем у 25% школьников-подростков стала отмечаться задержка полового созревания. До 40% призывников 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 детей и подростков считаются:

- зрелость, оцениваемая по развитию вторичных половых признаков;
- скелетная зрелость (костный возраст) оценивается с учетом стадий и сроков оссификации скелета;
- зубная зрелость (сроки прорезывания молочных и постоянных зубов).

При оценке степени развития вторичных половых признаков учитываются следующие характеристики: развитие волос на лобке (P) и в подмышечных впадинах (Ax); развитие молочных желез (Ma) и наступление менархе (Me) у девочек; пубертатное набухание сосков (Q) 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 сроков их прорезывания. 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 информативны, чем сроки оксификации скелета и появления вторичных половых признаков.

Процессы роста проявляются увеличением линейных размеров и веса тела. Прекращение роста и накопление массы тела только за счет отложения жира свидетельствует о наступлении зрелости.

Для индивидуальной оценки физического развития многие десятилетия использовался метод сигмальных отклонений (сигмальных шкал). В основу этого метода положены статистические таблицы стандартных антропометрических показателей для различных возрастно-половых и этнических групп. Стандарты в этих таблицах представлены в виде интервалов $M \pm \delta$, где M – стандартная величина показателя (вес, рост и т.п.), δ – среднеквадратическое (стандартное) отклонение от этой величины. Физическое развитие конкретного индивидуума оценивалось на основании того, в каком интервале находилось числовое значение его антропометрических данных. За норму принимался интервал в пределах $M \pm 1\delta$. Если оцениваемый антропометрический параметр ребенка оказывался в пределах от $\pm 1\delta$ до $\pm 2\delta$, то такого ребенка считали практически здоровым. При попадании оцениваемого параметра в интервал от $\pm 2\delta$ до $\pm 3\delta$ и более отклонение параметра считалось высоким. Ребенок, получивший такую оценку, считался нуждающимся в специальном углубленном обследовании из-за высокой вероятности наличия патологии. Эти среднестатистические нормативы не позволяют получать комплексные оценки физического развития.

Одним из комплексных методов оценки физического развития считается метод регрессионного анализа (по шкалам регрессии). Данный метод с помощью 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; возраст, пол, рост и величина жизненной емкости легких.

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 Полем Броком (индекс Брока): идеальный вес (кг) = рост (см) – 100. В эту формулу впоследствии были внесены уточнения. В частности, уточнения по полу: для мужчин идеальный вес (кг) = 0,9 (рост (см) – 100). Для женщин идеальный вес (кг) = 0,85 (рост (см) – 100).

$$M=L-100 \text{ (кг) при росте 155-165 см,}$$

$$M=L-105 \text{ (кг) при росте 166-175 см,}$$

$$M=L-110 \text{ (кг) при росте более 175 см.}$$

Считается, что для точности качественных оценок (табл. 19) из полученных расчетных величин следует вычесть около 8%: рост – 100-8%.

Таблица 19

**Качественные оценки весо-ростового индекса
(Дубровский В.И., Дубровская А.В., 2011)**

Граммов на сантиметр роста	Показатель упитанности
Больше 540	Ожирение
451-540	Чрезмерный вес
416-450	Излишний вес
401-415	Хорошая
400	Наилучшая для мужчин
390	Наилучшая для женщин
360-389	Средняя
320-359	Плохая
300-319	Очень плохая
200-299	Истощение

Для оценки физического развития детей и подростков достаточно часто используют индекс Рорера (ИНр) = W/H^3 кг/см³, где W – масса тела (кг), H – рост тела (м). При ИНр < 10,7 кг/м³ физическое развитие считается низким, а при значении индекса > 13,7 кг/м³ диагностируют высокое физическое развитие. Величина ИНр от 10,7 до 13,7 кг/м³ соответствует гармоничному развитию.

Достаточно надежным показателем гармоничности развития, используемым во многих странах мира при заключении договора страхования, является так называемый индекс массы, или индекс Кетле. В расчет его вводятся все те же величины, и формула выглядит следующим образом: индекс Кетле = вес (кг)/рост² (см). Считается, что индекс Кетле для женщин в норме должен быть равен 19-24, а для мужчин – 20-25. Если значение индекса равно 26 или превышает этот показатель, то речь идет о вредном для организма избытке веса. Об ожирении I ст. говорит индекс Кетле равный 26-30; II ст. – если он равен 30-40; III ст. – если он более 40. Если индекс Кетле ниже указанных нормативов, то это указывает на дефицит веса.

Зная длину тела стоя и сидя, можно найти коэффициент пропорциональности (КП) тела. $КП = ((L_1 - L_2)/2) \times 100$, где: L_1 – длина тела стоя, L_2 – длина тела сидя. В норме КП = 87-92%, у женщин он несколько ниже, чем у мужчин.

Одним из важных показателей физического развития считают площадь поверхности тела, которая определяется формулой Issakson (1958). Для лиц с суммой веса и длины тела больше 160 единиц: $S = [100 + W + (H - 160)] / 100$, где: S – площадь поверхности тела (м²), W – вес тела (г), H – длина тела (см).

Для низкорослых людей с суммой веса и длины тела меньше 160 единиц используют формулу Бойда (Boyd, 1935): $S = 3,207 \times H^{0,3} \times W^{0,7285} - 0,0188 \log W$, где: S – площадь тела (см²), H – длина тела (см), W – вес тела в граммах.

Для сопоставления индивидуальных особенностей телосложения, рассчитывают относительную силу мышц (относительно веса тела):

$F_{отн.} = F_{абс.} / W$, где $F_{отн.}$ – относительная сила (кг), $F_{абс.}$ – абсолютная сила (кг), W – вес тела (кг).

3.1.1. Центильный метод оценки физического развития

В последние десятилетия в практике широко используется центильный метод оценки индивидуального физического развития, который широко применяется и за рубежом. В основу метода положено процентное (центильное) распределение величины данного параметра физического развития (табл. 20 и 21). 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 – низкие величины, 3-й интервал – сниженные показатели. Соответственно распределяются величины, превышающие средние значения: 6-й интервал включает повышенные показатели, 7-й – высокие показатели, а 8-й включает очень высокие величины. Таким образом, зоны ниже 10-го и выше 90-го центиля свидетельствуют о выраженном снижении или, соответственно, повышении измеряемого показателя. Таким образом можно оценить любой показатель физического развития. Несмотря на кажущуюся громоздкость метода, на самом деле он прост.

Таблица 20

Центильная таблица роста (см) для мальчиков

Возраст	Длина роста						
	Центильный интервал						
	3%	10%	25%	50%	75%	90%	97%
0	46,5	48,0	49,8	51,3	52,3	53,3	55,0
1 мес.	49,5	51,2	52,7	54,5	55,6	56,5	57,3
2 мес.	52,6	53,8	55,3	57,3	58,2	59,4	60,9
3 мес.	55,3	56,5	58,1	60,0	60,9	62,0	63,8
4 мес.	57,5	58,7	60,6	62,0	63,1	64,5	66,3
5 мес.	59,9	61,1	62,3	64,3	65,6	67,0	68,9
6 мес.	61,7	63,0	64,8	66,1	67,7	69,0	71,2
7 мес.	63,8	65,1	66,3	68,0	69,8	71,1	73,5
8 мес.	65,5	66,8	68,1	70,0	71,3	73,1	75,3
9 мес.	67,3	68,2	69,8	71,3	73,2	75,1	78,8
10 мес.	68,8	69,1	71,2	73,0	75,1	76,9	78,8
11 мес.	70,1	71,3	72,6	74,3	76,2	78,0	80,3
1 год	71,2	72,3	74,0	75,5	77,3	79,7	81,7
15 мес.	74,8	75,9	77,1	79,0	81,0	83,0	85,3
18 мес.	76,9	78,4	79,8	81,7	83,9	85,9	89,4
21 мес.	79,3	80,8	82,3	84,3	86,5	88,3	91,2
2 года	81,3	83,0	84,5	86,8	89,0	90,8	94,0
27 мес.	83,0	84,9	86,8	88,7	91,3	93,9	96,8
30 мес.	84,5	87,0	89,0	91,3	93,7	95,3	99,0
33 мес.	86,3	88,8	91,3	93,5	96,0	98,1	101,2
3 года	88,0	90,0	92,3	96,0	99,8	102,0	104,3
3,5 года	90,3	92,6	95,0	99,1	102,5	105,0	107,3
4 года	93,2	95,5	98,3	102,0	105,3	108,0	110,6
4,5 года	96,0	98,3	101,2	105,1	108,6	111,0	113,6
5 лет	98,9	101,5	104,4	108,3	112,0	114,5	117,0
5,5 лет	101,8	104,7	107,8	111,5	115,1	118,0	120,6
6 лет	105,0	107,7	110,9	115,0	118,7	121,1	123,8
6,5 лет	108,0	110,8	113,8	118,2	121,8	124,6	127,2
7 лет	111,0	113,6	116,8	121,2	125,0	128,0	130,6
8 лет	116,3	119,0	122,1	126,9	130,8	134,5	137,0
9 лет	121,5	124,7	125,6	133,4	136,3	140,3	143,0
10 лет	126,3	129,4	133,0	137,8	142,0	146,7	149,2
11 лет	131,3	134,5	138,5	143,2	148,3	152,9	156,2
12 лет	136,2	140,0	143,6	149,2	154,3	159,5	163,5
13 лет	141,8	145,7	149,8	154,8	160,6	166,0	170,7
14 лет	148,3	152,3	156,2	161,2	167,7	172,0	176,7
15 лет	154,6	158,6	162,5	166,8	173,5	177,6	181,6
16 лет	158,8	163,2	166,8	173,3	177,8	182,0	186,3
17 лет	162,8	166,6	171,6	177,3	181,6	186,0	188,3

Центильная таблица роста (см) для девочек

Возраст	Длина роста						
	Центильный интервал						
	3%	10%	25%	50%	75%	90%	97%
0	46,5	48,0	49,8	51,3	52,3	53,3	55,0
1 мес.	49,5	51,2	52,7	54,5	55,6	56,5	57,3
2 мес.	52,6	53,8	55,3	57,3	58,2	59,4	60,9
3 мес.	55,3	56,5	58,1	60,0	60,9	62,0	63,8
4 мес.	57,5	58,7	60,6	62,0	63,1	64,5	66,3
5 мес.	59,9	61,1	62,3	64,3	65,6	67,0	68,9
6 мес.	61,7	63,0	64,8	66,1	67,7	69,0	71,2
7 мес.	63,8	65,1	66,3	68,0	69,8	71,1	73,5
8 мес.	65,5	66,8	68,1	70,0	71,3	73,1	75,3
9 мес.	67,3	68,2	69,8	71,3	73,2	75,1	78,8
10 мес.	68,8	69,1	71,2	73,0	75,1	76,9	78,8
11 мес.	70,1	71,3	72,6	74,3	76,2	78,0	80,3
1 год	71,2	72,3	74,0	75,5	77,3	79,7	81,7
15 мес.	74,8	75,9	77,1	79,0	81,0	83,0	85,3
18 мес.	76,9	78,4	79,8	81,7	83,9	85,9	89,4
21 мес.	79,3	80,8	82,3	84,3	86,5	88,3	91,2
2 года	81,3	83,0	84,5	86,8	89,0	90,8	94,0
27 мес.	83,0	84,9	86,8	88,7	91,3	93,9	96,8
30 мес.	84,5	87,0	89,0	91,3	93,7	95,3	99,0
33 мес.	86,3	88,8	91,3	93,5	96,0	98,1	101,2
3 года	88,0	90,0	92,3	96,0	99,8	102,0	104,5
3,5 года	90,3	92,6	95,0	99,1	102,5	105,0	107,5
4 года	93,2	95,5	98,3	102,0	105,5	108,0	110,6
4,5 года	96,0	98,3	101,2	105,1	108,6	111,0	115,6
5 лет	98,9	101,5	104,4	108,3	112,0	114,5	117,0
5,5 лет	101,8	104,7	107,8	111,5	115,1	118,0	120,6
6 лет	105,0	107,7	110,9	115,0	118,7	121,1	123,8
6,5 лет	108,0	110,8	113,8	118,2	121,8	124,6	127,2
7 лет	111,0	113,6	116,8	121,2	125,0	128,0	130,6
8 лет	116,3	119,0	122,1	126,9	130,8	134,5	137,0
9 лет	121,5	124,7	125,6	133,4	136,3	140,3	143,0
10 лет	126,5	129,4	133,0	137,8	142,0	146,7	149,2
11 лет	131,3	134,5	138,5	143,2	148,3	152,9	156,2
12 лет	136,2	140,0	143,6	149,2	154,5	159,5	163,3
13 лет	141,8	145,7	149,8	154,8	160,6	166,0	170,7
14 лет	148,3	152,3	156,2	161,2	167,7	172,0	176,7
15 лет	154,6	158,6	162,5	166,8	173,5	177,6	181,6
16 лет	158,8	163,2	166,8	173,3	177,8	182,0	186,3
17 лет	162,8	166,6	171,6	177,3	181,6	186,0	188,5

В настоящее время существуют разные методики комплексного оценивания на основе центильных таблиц гармоничности физического развития. Отличаясь друг от друга в несущественных деталях, все они

одинаково просты и результативны. Например, с помощью методики, представленной сотрудниками Оренбургского государственного университета, гармоничность физического развития на основе роста (длины тела) и массы тела можно оценить, используя двухмерный квадрат гармоничности, в котором имеются шкалы длины и массы тела, необходимо лишь найти место пересечения центильных рядов (табл. 22).

Физическое развитие считается гармоничным, если все исследуемые антропометрические показатели соответствуют одному и тому же центильному ряду либо разница их между собой – в пределах соседнего центиля. Большая разница свидетельствует о негармоничном развитии.

Таблица 22

Квадрат гармоничности
(Вспомогательная таблица для оценки физического развития)

		Процентные (центильные) ряды						
		3%	10%	25%	50%	75%	90%	97%
Масса тела по возрасту	97%						Гармоничное развитие, опережающее возраст	
	90%							
	75%			Гармоничное развитие, соответствующее возрасту				
	50%							
	25%							
	10%	Гармоничное развитие ниже возрастных норм						
	3%							
		Длина тела по возрасту						

3.2. Заболеваемость. Источники получения информации

Среди статистических показателей здоровья населения, нередко называемых показателями общественного здоровья, особое место занимает *заболеваемость*. Статистические показатели заболеваемости (*health indicator* – англ.) нередко называют отрицательными показателями здоровья. Вместе с тем следует отметить, что в медицинской статистике определенной, четко очерченной группы положительных показателей здоровья вообще не

существует. Как правило, в качестве положительных используют показатели заболеваемости, когда отмечается их достаточно позитивная динамика.

Таким образом, показатели заболеваемости являются основой оценки своей противоположности – общественного здоровья. На основе таких оценок строится медико-социальная политика государства, планируется и осуществляется контроль над деятельностью учреждений здравоохранения, результативностью лечебно-профилактических мероприятий и т.п. Характеристики заболеваемости косвенно отражают уровень социально-экономического благополучия населения в целом.

В государственной статистике России для регистрации первичных данных о заболеваемости используют большое количество различных 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.

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 10 тысяч ($^{0}/_{000}$), 1 тысячу ($^{0}/_{00}$) или 100 человек. Вычисляется также показатель структуры заболеваемости: доля в $^{0}/_{0}$ (син. удельный вес) того или иного заболевания среди всех заболеваний.

Специальные показатели характеризуют уровень и структуру заболеваемости по определенным нозологическим формам, а также по

различным возрастным, половым или социальным группам. Такой подход позволяет установить причины повышенной заболеваемости в той или иной группе.

Общая методология статистики заболеваемости в здравоохранении предусматривает 4 подхода изучения заболеваемости.

Изучение по случаям возникновения болезней на основе статистики относительных показателей заболеваемости (распространенность, частота и структура). Недостатком этого подхода является усредненность, обезличенность показателей, поскольку они вычисляются относительно достаточно крупных контингентов населения.

Изучение заболеваемости по данным полицейского учета позволяет выделять группы лиц, не имевших случаев заболеваний (за год, 2 года и т.п.), практически здоровых, больных и т.п. Такой подход позволяет более детально оценивать здоровье контингентов, контролировать динамику состояния здоровья, более целенаправленно разрабатывать и осуществлять лечебно-профилактические мероприятия. Полицейской учет служит организационно-методической основой применения диспансерного метода в здравоохранении. Система полицейского учета заболеваемости весьма трудоемка, поэтому требует обязательного использования современных информационных технологий сбора, хранения и переработки информации.

Изучение заболеваемости с учетом хронизации заболеваемости. В основе такого подхода лежит отдельный учет острых и хронических заболеваний. Считается, что такой подход позволяет оценивать частоту перехода острых форм заболеваний в хронические, вести целенаправленный учет наиболее тяжелой хронической патологии и на этой основе более целенаправленно планировать и осуществлять лечебно-профилактические мероприятия. Применение этого метода требует решения (ныне нерешенной) проблемы однозначной классификации острых и хронических форм болезней.

Валеологический подход в изучении заболеваемости. В основе такого подхода, периодически «всплывающего» и угасающего в здравоохранении,

лежит признание необходимости специального изучения «здоровья здоровых», физиологических резервов организма здорового человека, в том числе и донозологических стадий развития патологических процессов. Конкретное оформление идеи валеологии (лат. *valeo* – «здоровствовать», «быть здоровым») принадлежит выпускнику Военно-морской медицинской академии И.И. Брехману (1987). В советской медицине в 70-80 годах прошлого (20-го) столетия появилось целое направление, в основе которого лежала неопределенная и обозначенная только в общих чертах категория – «здоровье». При этом, не дождавшись получения конкретных результатов в рамках этого научного направления, опираясь только на логическое обоснование проблемы и рассуждения о возможных перспективах их решения, был организован предмет обучения практических врачей на кафедрах нескольких вузов.

Поспешность такого подхода стала ясна уже через несколько лет. В итоге, несмотря на протесты энтузиастов, в 2000 году в России валеология как учебная дисциплина Министерством образования РФ была ликвидирована. Неопределенность базовых понятий, трудная и неоднозначная судьба этого направления науки предопределила неспособность валеологии решать практические проблемы. Проблемы, во многом обусловившие рождение валеологии, по мнению ее поклонников, способной эффективно решать задачи сохранения здоровья человека, в том числе и общественного здоровья (Савельев С.И., Зайцев В.М. и соавт., 2007, Сидоров А.А., 2011).

В настоящее время статистика здоровья, основанная на валеологическом подходе, применяется только среди контингентов, в составе которых здоровые лица составляют основной массив наблюдения. *Например*, среди спортсменов. В широкой практике здравоохранения эти методы статистики не применяются.

Существуют различные варианты классификаций заболеваемости. Эти классификации (виды, типы) строятся либо на особенностях структуры заболеваемости, либо по степени утраты трудоспособности в результате

болезней, либо по источникам получения статистических данных о заболеваемости.

В зависимости от структуры 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 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 зависимости от утраты трудоспособности в статистике заболеваемости выделяют:

- заболеваемость без утраты трудоспособности. На практике этот показатель применяется крайне редко;
- заболеваемость с временной утратой трудоспособности (ЗВУТ). Показатели этой группы широко применялись в практике «дореформенного»

здравоохранения РФ. В настоящее время этому виду заболеваемости уделяется существенно меньшее внимание;

- заболеваемость со стойкой, ограниченной или полной утратой трудоспособности (инвалидность).

В зависимости от метода получения исходной информации, большинство показателей заболеваемости можно условно разбить на следующие группы:

- полученные путем регистрации обращений за медицинской помощью;
- полученные путем активного выявления патологии или «донозологических» состояний на медицинских осмотрах;
- вычисляемые на основе данных о причинах смерти.

Показатели, получаемые на основании данных опроса населения (пациентов), в официальной статистике не применяются, а используются только как дополнение к другим оценкам здоровья в ходе специальных исследований.

3.2.1. Данные о причинах смерти

Заболеваемость по данным о причинах смерти – совокупность заболеваний, зарегистрированных по данным о причинах смерти. Может быть представлена двумя группами показателей: показателями смертности и (или) летальности.

Источник информации в России: врачебное свидетельство о смерти или фельдшерская справка о смерти. В практике этот показатель чаще всего трактуется как показатель причин смертности населения (*proportionate mortality* – англ.). Так как причиной смерти человека почти всегда являются заболевания (за исключением травм, отравлений и убийств), считается, что показатель достаточно полно отражает состояние здоровья населения. Этот показатель является одним из самых распространенных в государственной и международной статистике общественного здоровья, поскольку на него в

меньшей степени, чем на статистику заболеваемости по обращаемости, оказывает влияние уровень развития (доступности) системы здравоохранения.

По данным Федеральной службы статистики России за период с 1995 по 2006 годы, показатель смертности трудоспособного населения в стране колебался в пределах от 488,8 случая на 100 тыс. населения в 1995 году до 826,5 случая на 100 тыс. населения в 2005 году. В 2006 году этот показатель несколько снизился: до 746,1 случая на 100 тыс. населения.

Общие показатели смертности существенно отличаются по разным территориям России. Вместе с тем, структура общей смертности изменяется мало: более 75% причин приходится на болезни системы кровообращения, пищеварения и новообразования (табл.23).

Таблица 23

**Смертность по основным классам причин смерти по субъектам РФ
(умерших на 100 тыс. чел. населения)**

Причины смерти	1990	1995	2000	2005	2010
Болезни системы кровообращения	618,7	790,7	846,1	908,	805,9
Новообразования	194,4	203,	204,7	201,2	205,1
Внешние причины	134,	236,8	219,0	220,7	151,7
Болезни органов пищеварения	28,7	46,1	44,4	65,5	64,4
Болезни органов дыхания	59,4	73,9	70,2	66,2	52,3
Некоторые инфекционные и паразитарные болезни	12,1	20,7	24,9	27,2	23,5
Самоубийства	26,5	41,4	39,1	32,2	23,4
Случайные отравления алкоголем	10,9	29,5	25,6	28,6	13,4
Убийства	14,3	30,8	28,2	24,9	13,3
Всего умерших от всех причин	1 119,1	1 497,7	1 529,0	1 609,9	1 419,2

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 у 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.

Летальность – частота смерти от конкретного заболевания в конкретных условиях. Например, больничная летальность, летальность в амбулаторно-поликлинических учреждениях и т.п. Летальность заболевания существенно влияет на показатели распространенности какого-либо заболевания, полученные на основе изучения причин смерти. 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 соотношению истинной распространенности этих болезней.

Показатели заболеваемости по данным о причинах смерти играют незначительную роль в статистике деятельности конкретных амбулаторно-поликлинических или стационарных учреждений. Причин этому несколько. Одна из главных – невозможность определить «виновность» стационаров или амбулаторно-поликлинических учреждений в смерти того или иного человека. Кроме того, из-за относительно небольшого числа наблюдений (смертей) по отдельному территориальному образованию или предприятию статистика смертности и летальности имеет только ориентировочное значение даже при суммарном учете показателей за несколько лет. Вместе с тем, каждый случай летального исхода строго контролируется и служит причиной детального клинического анализа.

К числу достоинств учета заболеваемости по данным о причинах смерти следует отнести достаточно полное выявление наиболее тяжелых болезней и патологических состояний, приводящих к летальному исходу. При этом регистрируются и те заболевания, по которым не было обращений за

медицинской помощью. Вследствие этого в последнее время появляются рекомендации механически суммировать заболеваемость, установленную по данным о причинах смерти, и показатели заболеваемости, полученные из других источников. Однако такая статистика требует обязательного полицейского учета (суммарного учета на одно лицо) всех случаев заболеваний по обращаемости, по данным профилактических осмотров и смерти в независимости от места регистрации. В противном случае неизбежно будет происходить искажение итогов, например, из-за дублирования учета одних и тех же случаев заболеваний, зарегистрированных по разным источникам. Кроме того, суммирование показателей общей заболеваемости и показателей по причинам смерти недопустимо с точки зрения методологии статистического исследования. Имеется в виду отсутствие равновероятной возможности стать объектом регистрации (статистического наблюдения) любого человека, т.к. данные о заболеваемости по причинам смерти можно получить только по умершим людям. По живым лицам естественно таких данных получено быть не может. Таким образом, нарушается единство объекта наблюдения.

Показатели смертности, особенно младенческой и детской, являются существенными характеристиками общественного здоровья, которые дают врачу-организатору любого лечебно-профилактического учреждения общий фон общественного здоровья жителей прикрепленных территорий. Кроме того, показатели детской смертности прямо используются для оценки эффективности деятельности роддомов и детских мед.учреждений.

Как уже отмечалось, в отличие от показателей смертности, которые по своей сути являются медико-демографическими показателями здоровья, непосредственно в практике отдельных амбулаторно-поликлинических учреждений используются специальные показатели, вычисляемые на основе данных о причинах смерти не всего населения, а только амбулаторных больных, – показатели *внебольничной летальности* (лат. *letal* – «смертельный»). Эти интенсивные показатели характеризуют частоту смертельных исходов заболеваний только среди амбулаторных больных или, что встречается

достаточно редко, больных, нигде определенно не лечившихся, но умерших на дому. Вычисляются в % на 100 больных (показатели смертности относительно средней численности населения).

Соответственно, при анализе деятельности стационаров используются специальные показатели, вычисляемые на основе данных о причинах смерти только стационарных больных, – показатели *больничной летальности*. Эти интенсивные показатели характеризуют частоту смертельных исходов заболеваний среди стационарных больных. Вычисляются, как правило, за год в % на 100 больных, пребывавших в стационаре. Здесь, на достоверность информации влияет своевременная выписка из стационара больных с заведомо ясным летальным исходом заболевания.

Как и в случаях других статистических показателей общественного здоровья, при анализе летальности вычисляются общие и специальные показатели. *Общие* вычисляются на общую численность лечившегося (пролеченного за год) в амбулаторно-поликлиническом или стационарном учреждении контингента больных. *Специальные* вычисляются относительно контингента с конкретной патологией, послужившей причиной смерти, места смерти (на дому, в поликлинике, в стационаре) и т.п.

3.2.2. Обрацаемость за медицинской помощью

Заболеваемость по обрацаемости – 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 обрацаемости формируются при обращении за медицинской помощью. Соответственно, их полнота и достоверность зависят от доступности медицинской помощи, принципов ее организации (радиус обслуживания, наличие врачей-специалистов, диагностическое оборудование и т.п.). Отсюда вытекает важное обстоятельство: проводить сравнение показателей заболеваемости по обрацаемости, полученных из разных стран

или разных регионов одной страны, имеющих существенные различия по этим факторам, 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 местных управленческих и организационных структур здравоохранения, из-за желания приукрасить истинное положение вещей и т.д. (Иногда говорят, что заболеваемость по обращаемости формируется самими врачами). 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 заболеваемость жителей села, устанавливаемая углубленными медицинскими осмотрами, выше, чем в городах.

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

сомнительную достоверность¹⁶. 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 заболеваемости по обращаемости сохраняет в условиях РФ своё значение информативного фактора оценки здоровья населения и решения конкретных задач планирования, а также оценки результатов деятельности системы медико-социальной помощи.

Группа статистических показателей по обращаемости представлена следующими основными вариантами: показатели, в которых единицей наблюдения служат случаи заболеваний, и показатели, в которых единицей наблюдения служат непосредственно случаи обращения населения, за определенный период времени проживающего на определенной территории или в районе обслуживания иным медицинским учреждением (группой учреждений).

Показатели, в которых единицей наблюдения служат случаи заболеваний контингентов населения:

первичная заболеваемость (*инцидентность – incidence rate*) – частота ранее не зарегистрированных и впервые в данном году выявленных заболеваний среди населения, обратившегося за медицинской помощью. Регистрируется при установлении большого диагноза впервые в жизни.

Общая заболеваемость населения характеризуется общим числом случаев заболеваний, зарегистрированных в течение года. При этом учитываются больные, которым диагноз установлен как впервые, так и при повторном (многократном) обращении по поводу данного заболевания (табл. 24).

¹⁶ Следует отметить, что в последние 2 десятилетия и в государственных учреждениях появился фактор, подталкивающий недобросовестных работников мед. учреждений к завышению реальной обращаемости, – желание получить больше средств из общегосударственной системы ОМС.

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. При этом хронические заболевания могут быть зарегистрированы только один раз, а острые регистрируются при каждом их новом возникновении (регистрируется только первое, первичное обращение в случае заболевания). 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.

Показатели, в которых единицей наблюдения служат случаи обращений.

Обращаемость за медицинской помощью – число обращений больных, впервые обратившихся за медицинской помощью по данному заболеванию. 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. Несмотря на то, что в связи с внедрением страховой медицины в систему здравоохранения РФ, появились новые показатели (число, структура и стоимость медицинских услуг), посещаемость остается основным плановым показателем деятельности любого амбулаторно-поликлинического учреждения. Вычисляется, как число

посещений, приходящихся на 1 человека (жителя региона, города, района, прикрепленного контингента) в год. Следует отметить, что Минздравсоцразвития России в рамках целевой программы «Обеспечение государственного контроля предоставляемой населению медицинской помощи» на 2008-2010 годы в качестве контрольного уровня достаточности объема амбулаторно-поликлинической помощи взрослому населению предусматривает 15000 посещений на тыс. населения.

Среди показателей заболеваемости по обращаемости выделяют специальные показатели. В их числе *заболеваемость по данным обращаемости в поликлинику* – основная компонента общей заболеваемости по обращаемости. Учитывая, что 70-80% больных начинают и заканчивают лечение в амбулаторно-поликлинических учреждениях, при высокой доступности этого вида медицинской помощи показатель может давать достаточно объективную информацию о распространенности болезней. Следует отметить, что с финансированием наблюдается прямо противоположная ситуация: на амбулаторно-поликлиническую помощь в России к 2000 году тратилось 20-26% всех средств, выделяемых на здравоохранение.

В последние десятилетия, в связи с развитием страховой медицины¹⁷, в практической статистике все чаще стали появляться такие характеристики заболеваемости населения и деятельности медицинских учреждений, как медицинская услуга и ее производная величина – случай медицинского обслуживания. Эти показатели вычисляются как в абсолютной численности услуг, так и в стоимости этих услуг. Согласно ныне действующему Федеральному закону РФ от 21 ноября 2011 года (№ 323-ФЗ), *медицинская услуга* – медицинское вмешательство или комплекс медицинских вмешательств, направленных на профилактику, диагностику и лечение

¹⁷ Подробнее основные методические аспекты статистики добровольного и общего медицинского страхования рассмотрены в соответствующем разделе настоящего издания.

заболеваний, медицинскую реабилитацию и имеющих самостоятельное законченное значение. В настоящее время в РФ действует смешанная система бесплатного и платного здравоохранения. Платные медицинские услуги – это медицинские услуги, предоставляемые на возмездной основе за счет личных средств граждан, средств юридических лиц и иных средств на основании договоров, в том числе договоров добровольного медицинского страхования. Платные медицинские услуги могут предоставляться в полном объеме стандарта медицинской помощи, утвержденного Минздравом России, либо по просьбе потребителя в виде осуществления отдельных консультаций или медицинских вмешательств, в том числе в объеме, превышающем объем выполняемого стандарта медицинской помощи. В настоящее время номенклатура, стоимость и порядок предоставления разных услуг регулируются правилами, утвержденными Постановлениями Правительства РФ¹⁸. Из-за существующих ныне различий в востребованности, доступности и стоимости услуг в отдельных регионах РФ в практической статистике здоровья и здравоохранения медицинские услуги используются только в статистике здравоохранения на местном уровне.

¹⁸ ↑ Постановление Правительства Российской Федерации от 21 октября 2011 г. № 856 «О Программе государственных гарантий оказания гражданам Российской Федерации бесплатной медицинской помощи на 2012 год» и Постановление Правительства РФ от 4 октября 2012 г. № 1006 «Об утверждении Правил предоставления медицинскими организациями платных медицинских услуг».

Таблица 24

Заболееваемость населения РФ по основным классам болезней в 2000, 2005-2006 гг. на 100 тыс. чел. (РОССТАТ, 2007)

МКБ 10	Зарегистрировано больных					в т.ч. с диагнозом, установленным впервые в жизни		
	2000	2005	2006	2000	2005	2006	2000	2006
Все болезни:	131439,5	146352,0	151735,1	73054,6	74587,8	76387,1		
Некоторые инфекционные и паразитарные болезни	6173,2	5352,6	5342,4	4430,2	3742,0	3738,8		
Новообразования	3061,3	3548,8	3594,2	842,7	955,8	994,9		
Болезни крови, кровотоковых органов и отдельные нарушения, во- влекающие иммунный механизм	878,4	1097,7	1236,2	378,5	455,9	536,7		
из них — анемии	757,6	954,9	1094,5	335,6	407,1	484,7		
Болезни эндокринной системы, нарушения обмена веществ	3813,1	4816,3	5312,4	848,2	958,4	1173,8		
из них — сахарный диабет	1416,6	1759,7	1883,8	111,3	174,4	202,8		
Болезни нервной системы	4482,4	4969,5	5292,0	1529,8	1534,2	1626,8		
Болезни глаза и его придаточного аппарата	9309,4	10413,6	11044,3	3186,3	3365,9	3584,0		
Болезни уха и сосцевидного отростка	3225,6	3688,5	3746,6	2192,7	2412,8	2457,7		
Болезни системы кровообращения	13802,4	19314,6	20923,0	1705,8	2290,3	2657,5		
из них — характеризующиеся повышенным кровяным давлением	3459,8	6230,0	7120,8	298,7	539,8	707,0		
— ишемическая болезнь сердца ¹⁾	4857,4	4731,4	4962,7	421,7	414,8	456,1		
— стенокардия ¹⁾	1894,2	2018,6	2158,6	125,6	137,3	157,2		
— острый инфаркт миокарда	143,3	114,1	114,1	143,3	114,1	114,1		
Болезни органов дыхания	36767,7	35388,8	35791,7	31721,9	29525,3	29713,7		
Болезни органов пищеварения	10570,7	11333,4	11372,0	3228,0	3546,1	3525,9		
из них — язва желудка и двенадцатиперстной кишки	1282,6	1205,6	1201,2	131,4	106,6	112,7		
Болезни кожи и подкожной клетчатки	5513,2	6234,1	6383,8	4402,3	4982,3	5080,2		
Болезни костно-мышечной системы и соединительной ткани	8750,3	11038,6	11857,7	3059,2	3342,9	3536,8		
Болезни мочеполовой системы	7890,7	9881,9	10463,8	3758,0	4621,1	4889,2		
Осложнения беременности, родов и послеродового периода	6327,7	7828,6	2237,3	5288,2	6289,2	1767,9		
Врожденные аномалии деформации и хромосомные нарушения	520,8	616,9	640,8	146,9	171,0	180,5		
Травмы, отравления и другие последствия воздействия внешних причин	8750,6	9099,9	8774,6	8618,3	8949,3	8477,0		

Госпитализированная (госпитальная) заболеваемость характеризует совокупность болезней, послуживших причиной пребывания в стационаре. 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 чел. проходили стационарное лечение, в 2011 году – 219 чел. Средний срок госпитализации в 1999 году составлял 18 дней, к 2011 году этот показатель существенно сократился (12,4 дня).

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.

Госпитализированная заболеваемость как критерий оценки общественного здоровья по сравнению с другими видами заболеваемости представляет наименьший интерес по нескольким причинам. Одна из них – неполная регистрация случаев заболеваний, так как большинство больных (до 80%) начинают и заканчивают своё лечение в амбулаторно-поликлинических учреждениях. Кроме того, этот вид заболеваемости существенно зависит от доступности госпитальной помощи, поскольку обеспеченность населения больничными койками на различных территориях неодинакова (табл. 35).

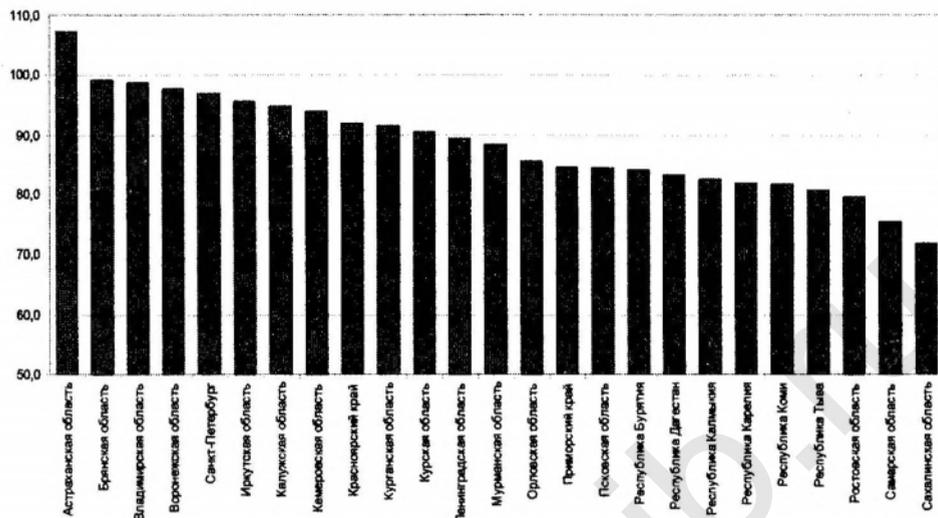


Рис.35. Обеспеченность отдельных территорий РФ больничными койками в 2011 году на 10 тыс. населения (по данным Единой межведомственной информационно-статистической системы РФ 2013 г.)

3.2.3. Результаты профилактических осмотров

Активное выявление заболеваний путем медицинских осмотров при надлежащей их организации дает достаточно объективную картину распространенности болезней. Поэтому считается, что профилактические осмотры являются одним из информативных показателей, характеризующих лечебно-профилактическую активность амбулаторно-поликлинического учреждения. Существенным достоинством этого вида источника информации о здоровье населения является массовость профилактических осмотров в РФ. По данным Министерства здравоохранения и социального развития РФ, только в 2010 году прошли медицинские периодические осмотры почти 17 млн. человек (табл. 25).

Некоторое снижение абсолютного числа лиц, прошедших периодические осмотры в 2010 году по сравнению с 1990 годом, может объясняться динамикой общей численности населения в РФ за указанный период времени.

Динамика профилактических осмотров населения

Осмотрено подлежащих периодическим осмотрам	Годы				
	1990	1995	2000	2005	2010
Всего тыс. человек (взрослых)	22307,3	19215,9	17954,9	16781,5	16930,2
% от числа подлежащих осмотрам:	88,8	87,0	90,8	91,8	95,0

Показатели заболеваемости по данным медицинских осмотров или патологическая пораженность представляют собой совокупность болезней и патологических состояний, выявленных при профилактических медицинских осмотрах. Существенным достоинством информации о здоровье, полученной путем профилактических осмотров, является возможность учета заболеваний в ранних стадиях и преморбидных состояний, что гарантирует высокую эффективность лечения и динамического наблюдения. Профилактические осмотры представляют собой классическое моментное статистическое наблюдение, поскольку при медосмотрах, как правило, регистрируются только те заболевания и патологические состояния, которые были у пациента на момент осмотра. Эта особенность является источником основного недостатка показателей о заболеваемости, полученных путем активного выявления на медосмотрах: недоучет острых заболеваний, поскольку во время медицинских осмотров регистрируются в основном хронические заболевания, имеющиеся на момент освидетельствования.

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.). Таким образом, проводить сравнительные оценки статистических показателей патологической пораженности различных контингентов, обследованных по разным или неизвестным программам, следует весьма осторожно.

К числу немаловажных недостатков профилактических осмотров относятся крупные материальные затраты на их проведение. Особенно резко

эти затраты возрастают при использовании труда высококвалифицированных врачей-специалистов, дорогостоящего лабораторно-диагностического оборудования и т.п.

Основным статистическим показателем итогов профилактических осмотров является показатель патологической пораженности, который может вычисляться отдельно в различных возрастно-половых, профессиональных и т.п. группах наблюдения:

$$\frac{\text{Число заболеваний, выявленных при проф.осмотрах} \times 1000}{\text{Число осмотренных}}$$

Медицинская эффективность профилактических осмотров оценивается частотой впервые выявленных заболеваний и патологических состояний:

$$\frac{\text{Число заболеваний, впервые выявленных при проф.осмотрах} \times 1000}{\text{Число осмотренных}}$$

Полнота охвата медицинскими осмотрами характеризует качество организации профилактической работы, в том числе качество диспансерного наблюдения и наблюдения за группами «проф. вредников»:

$$\frac{\text{Число осмотренных лиц} \times 100}{\text{Число, подлежащих осмотрам}}$$

Медицинские осмотры подразделяют на: предварительные осмотры при поступлении на работу, периодические углубленные и целевые осмотры.

Предварительные (входные) медицинские осмотры. Согласно ныне действующему Трудовому кодексу РФ, обязательному предварительному медицинскому осмотру (обследованию) при заключении трудового договора подлежат лица, не достигшие возраста 18 лет. В соответствии с этим кодексом медицинскому осмотру при приеме на работу, а также периодическим

медицинским осмотрам подлежат работники, занятые на тяжелых работах и на работах с вредными или опасными условиями труда; работники организаций пищевой промышленности, общественного питания и торговли, водопроводных сооружений, лечебно-профилактических и детских учреждений.

Работники, связанные с движением транспорта, проходят обязательные предварительные (при поступлении на работу) и периодические (медицинские) осмотры для определения пригодности этих работников для выполнения поручаемой работы и для предупреждения профессиональных заболеваний. Эти осмотры производятся за счет средств работодателя. Если гражданин работает в качестве индивидуального предпринимателя, это не отменяет для него обязанности самому проходить обязательный медицинский осмотр.

В соответствии с медицинскими рекомендациями и в связи с другими обстоятельствами отдельные лица могут привлекаться и для внеочередных медицинских осмотров.

Перечень вредных или опасных производственных факторов, при наличии которых проводятся предварительные и периодические медицинские осмотры, а также и перечень работ, при выполнении которых проводятся предварительные и периодические медицинские осмотры, приведены в Приказе Минздравсоцразвития РФ № 83 от 16.08.2004.

В случае необходимости по решению органов местного самоуправления у отдельных работодателей могут вводиться дополнительные условия и показания к проведению обязательных медицинских осмотров. Работники, осуществляющие отдельные виды деятельности, в том числе связанной с источниками повышенной опасности (с влиянием вредных веществ и неблагоприятных производственных факторов), а также работающие в условиях повышенной опасности, проходят обязательное психиатрическое освидетельствование не реже одного раза в пять лет.

Периодические углубленные медицинские осмотры (УМО) – регулярные, через равные промежутки времени медицинские осмотры (для лиц в возрасте до 21 года – ежегодные), проводимые с целью оценки динамики состояния

здоровья, раннего выявления заболеваний или их признаков, как этиологически связанных, так и не связанных с профессией, но при которых данная профессиональная деятельность представляет опасность для здоровья работника. Проводятся эти осмотры на базах ЛПУ, имеющих лицензии на этот вид деятельности. Работникам, занятым на вредных работах и на работах с вредными или опасными производственными факторами в течение 5 и более лет, периодические медицинские осмотры проводятся в центрах профессиональной патологии и других медицинских организациях, имеющих лицензии на экспертизу профессиональной пригодности и экспертизу связи заболевания с профессией, один раз в 5 лет.

Целевые осмотры – осмотры проводятся, как правило, в рамках научно-практических исследований с целью выявления конкретной патологии (гинекологической, онкологической и т.п.) или с целью освидетельствования относительно небольших групп с высокой вероятностью (групп повышенного риска) возникновения одного или нескольких этиологически связанных заболеваний.

Кроме того, в специальной литературе и в результатах научных исследований встречаются термины «*массовые научно-медицинские осмотры*» и «*комплексные профилактические осмотры*».

Массовые научно-медицинские осмотры – разовые комплексные медицинские осмотры, главной целью которых является не проведение лечебно-оздоровительных мероприятий, а получение статистики распространенности определенных заболеваний, создание нормативной базы, оценка воздействия неблагоприятных факторов среды на здоровье определенных групп населения. *Комплексные профилактические осмотры* – медицинские осмотры, предусматривающие использование комплекса лабораторно-диагностических исследований, участие в осмотрах «бригад» врачей нескольких специальностей. В настоящее время термин «комплексный осмотр» в условиях промышленного предприятия утратил практическое

значение, поскольку все периодические углубленные медицинские осмотры и осмотры при приеме на работу являются комплексными.

Следует отметить, что единого понимания и обозначения тех или иных видов медицинских профилактических осмотров в РФ и в зарубежных странах нет. Не существует и единых подходов к организации профилактических осмотров и клинической трактовке их результатов. Основная причина этого – различия организации системы здравоохранения и отличия в принципах финансирования этой системы в разных странах. *Например*, стремление снизить величину финансовых затрат на массовые профилактические обследования, сохранив их высокую медицинскую эффективность (выявляемость патологии), привело к широкому внедрению в практику зарубежного здравоохранения различных тестов и организации на этой основе этапной системы профилактических обследований. Относительно простые тесты, выполняемые на первом этапе обследования (как правило, без участия врачей) позволяют разделить весь обследуемый контингент на несколько групп. *Например*, здоровые, без признаков патологических состояний; здоровые лица с отдельными признаками заболеваний (группа риска); лица, имеющие явную патологию, и т.п. Такое предварительное разделение позволяет, во-первых, сократить нагрузку на врачей-специалистов узкого профиля за счет сокращения числа пациентов, которые с достаточной степенью вероятности не нуждаются в их услугах. Во-вторых, существенно снизить затраты на дорогостоящие методы углубленных лабораторно-диагностических обследований.

Использование разнообразных подходов к организации профилактических медицинских осмотров приводило и приводит к разным неоднозначным результатам. Наглядным *примером* здесь могут быть итоги специального исследования эффективности профилактических осмотров, опубликованные еще в 1981 году И.Н. Тюльпиным, Г.З. Демченковой и М.П. Полонским (табл. 26).

Таблица 26

Сравнительные итоги разных медицинских осмотров (Тюльпин И.Н. и соавт., 1981)

Возраст осмотренного населения (годы)	Выявлено заболеваний и состояний				При осмотре отдельными специалистами
	При комплексных осмотрах	При осмотре терапевтов, остальными специалистами по показаниям	При осмотре бригадой врачей, определенных спецификой профессиональной деятельности	При осмотре остальными специалистами	
0-4	748,6	518,2	-	427,6	
5-9	903,8	611,7	-	503,9	
10-14	928,4	641,4	-	511,2	
15-19	1010,6	729,8	637,1	517,4	
20-24	1281,6	823,6	702,2	549,6	
25-29	1341,6	897,9	734,7	592,4	
30-39	1679,6	1124,5	839,7	678,2	
40-49	1944,8	1311,2	1004,5	813,3	
50-59	2635,8	1979,8	1271,2	907,9	
60-69	3564,7	2317,1	1513,6	1028,8	
70 и старше	4071,8	2874,2	-	1742,4	
Всего	1810,1	1104,8	782,4	598,2	

3.2.4. Использование результатов тестовых опросов

Считается, что использование высокоэффективных тестов позволяет решать целый ряд проблем медицинской статистики:

- получать стандартные характеристики здоровья при дистанционных наблюдениях (телемедицина), а также стандартизованные в известных границах регистрируемых параметров показатели здоровья при массовых обследованиях;
- формировать статистику эффективности лечебно-профилактических и реабилитационных мероприятий на основе периодической оценки параметров отдельных функций организма и общей жизнедеятельности пациентов;
- решать проблему повышения эффективности массовых профилактических осмотров на основе использования результатов скрининг-обследований.

Эти причины являются фактором широкого распространения в практике научных исследований, а также в современном отечественном и зарубежном здравоохранении различных видов тестирования. Вместе с тем, статистика непосредственных результатов тестирования не нашла практического применения в системе учета и оценки уровня общественного здоровья.

За последние несколько десятилетий предложены сотни различных вариантов тестирования. Все они независимо от практической ценности и направленности могут быть разделены на две группы:

- тесты, с помощью которых оцениваются возможности и состояние отдельных систем организма человека или устанавливается факт наличия или риска появления конкретных синдромов патологических состояний;
- тесты обобщенного, всестороннего диагностического обследования пациентов.

К числу тестов первой группы относятся повсеместно применяемые в практической медицине функциональные тесты и тесты работы отдельных органов и систем, так наз. анализы. Современные аппаратные технологии позволяют быстро и в большом объеме получать широкие многоаспектные

характеристики биохимического и клеточного состава крови, характеристики состояния сердечно-сосудистой системы человека и т.п. К числу недостатков таких систем, за редким исключением, относятся дороговизна таких методов, низкая мобильность аппаратных комплексов, недостаточная специфичность результатов.

В последние десятилетия внимание исследователей и практиков здравоохранения привлекают различные тест-опросники, которые значительно удешевляют процесс углубленного диагностического обследования. Вариантом, позволяющим существенно улучшить результативность тестирования, является комплексное тестирование, т.е. сочетание аппаратных методов обследования и тест-опросников.

В обоих вариантах тестирования могут использоваться пакеты различных оценочных шкал, вопросников, многоаспектные тесты и т.п. В ходе целевых медицинских обследований обычно применяются специальные целевые тесты-опросники (Случевский Ф.И., 1977; Белова А.Н., 2002 и др.). С точки зрения повышения эффективности профилактических осмотров, наиболее предпочтительным представляется комплексное тестирование.

Согласно общепринятой точки зрения, в опросной части комплексных тестов должны быть отражены, по крайней мере, медико-демографические характеристики личности, субъективные характеристики здоровья, оценки уровня повседневной активности и экономического состояния.

Характеристики здоровья, регистрируемые обычно в опросниках, как правило, включают: самооценку здоровья в целом, наличие симптомов наиболее распространенных заболеваний, в том числе хронических заболеваний, уровень медицинской (профилактической) активности, в т.ч. кратность обращения за медицинской помощью, самостоятельное лечение, обращение к конкретным врачам-специалистам (Савельев С.И., Зайцев В.М. и соавт., 2004).

Основной проблемой тестирования здоровья на основе опросников является отсутствие каких-либо определенных конкретных норм оценивания

отдельных характеристик здоровья, включаемых в состав опросников. Кроме того, результативность таких тестов во многом зависит от образа жизни человека, уровня его притязаний и ожиданий, эмоционально-личностных реакций человека и т.п.

Разработка и внедрение в широкую практику тестов-опросников, несмотря на кажущуюся внешнюю простоту их содержания, являются сложнейшей научно-практической задачей. За последние десятилетия за рубежом и в России в интересах практического здравоохранения проведены многие сотни разработок различных тестов. Однако примеров успешного решения этих задач единицы.

В качестве *примера* теста-опросника, широкоприменяемого в отечественной и зарубежной практике специальных клинических исследований, можно рассмотреть тест *MOS SF-36 (MOS-36-Item Short-Form Health Survey)*.

Этот опросник (авторы Ware J., Sherburne C., 1992 и J., 1993) был создан в США на основе других опросников, которые разрабатывались в рамках реализации программы *MOS (Medical Outcome Study)*. Считается, что этот опросник может широко применяться в практике клинических проспективных обследований. Опросник MOS SF-36 содержит 11 разделов, которые имеют различное количество подразделов. В итоге общее число вопросов равно 36, что и отражено в названии опросника. Перечень вопросов представляет собой комплекс характеристик здоровья, включающих функцию и дисфункцию, стресс и благополучие, объективные и субъективные оценки, положительные и отрицательные самооценки общего состояния здоровья. На основании опросника *MOS-36-Item Short-Form Health Survey* были созданы различные варианты более кратких опросников. Например: короткая версия опросника *Здоровье -12 (Item Short-Form Health Survey)* или SF-12 (Белова А.Н., 2002).

Российскими исследователями Межнационального центра исследования качества жизни (МЦИКЖ, г. Санкт-Петербург) в 1998 г. была создана русскоязычная версия опросника SF-36, которая была использована для

изучения качества жизни 2114 жителей Санкт-Петербурга. Результаты исследования показали высокую согласованность с характеристиками исследований, проведенных в других странах. Форма для заполнения опросника имеет следующий вид (Кудряшова И.В., 2003):

ОПРОСНИК SF-36

(русскоязычная версия, созданная и рекомендованная МЦИКЖ).

Ф. И. О. _____

Дата заполнения _____

1. В целом как Вы бы оценили состояние Вашего здоровья (обведите одну цифру):

- | | |
|----------------|---|
| Отличное | 1 |
| Очень хорошее | 2 |
| Хорошее | 3 |
| Посредственное | 4 |
| Плохое | 5 |

2. Как бы Вы в целом оценили свое здоровье *сейчас* по сравнению с тем, что было *год назад* (обведите одну цифру):

- | | |
|----------------------------------|---|
| Значительно лучше, чем год назад | 1 |
| Несколько лучше, чем год назад | 2 |
| Примерно так же, как год назад | 3 |
| Несколько хуже, чем год назад | 4 |
| Гораздо хуже, чем год назад | 5 |

3. Следующие вопросы касаются физических нагрузок, с которыми Вы, возможно, сталкиваетесь в течение своего обычного дня. Ограничивает ли Вас состояние Вашего здоровья в настоящее время в выполнении перечисленных ниже физических нагрузок? Если да, то в какой степени? (табл. 27).

(обведите одну цифру в каждой строке)

	Да, значи- тельно огра- ничивает	Да, немного ограничивает	Нет, со- всем не ограничи- вает
А. Тяжелые физические нагрузки, такие, как бег, поднятие тяжестей, занятие силовыми видами спорта	1	2	3
Б. Умеренные физические нагрузки, такие, как передвинуть стол, поработать с пылесосом, собирать грибы или ягоды	1	2	3
В. Поднять или нести сумку с продуктами	1	2	3
Г. Подняться пешком по лестнице на несколько пролетов	1	2	3
Д. Подняться пешком по лестнице на один пролет	1	2	3
Е. Наклониться, встать на колени, присесть на корточки	1	2	3
Ж. Пройти расстояние более одного километра	1	2	3
З. Пройти расстояние в несколько кварталов	1	2	3
И. Пройти расстояние в один квартал	1	2	3
К. Самостоятельно вымыться, одеться	1	2	3

4. Бывало ли за последние 4 недели, что Ваше физическое состояние вызывало затруднения в Вашей работе или другой обычной повседневной деятельности, вследствие чего (табл. 28):

Таблица 28

(обведите одну цифру в каждой строке)

	Да	Нет
А. Пришлось сократить количество времени, затрачиваемого на работу или другие дела	1	2
Б. Выполнили меньше, чем хотели	1	2
В. Вы были ограничены в выполнении какого-либо определенного вида работ или другой деятельности	1	2
Г. Были трудности при выполнении своей работы или других дел (например, они потребовали дополнительных усилий)	1	2

5. Бывало ли за последние 4 недели, что Ваше эмоциональное состояние вызывало затруднения в Вашей работе или другой обычной повседневной деятельности, вследствие чего (табл. 29):

Таблица 29

(обведите одну цифру в каждой строке)

	Да	Нет
А. Пришлось сократить количество времени, затрачиваемого на работу или другие дела	1	2
Б. Выполнили меньше, чем хотели	1	2
В. Выполняли свою работу или другие. Дела не так аккуратно, как обычно	1	2

6. Насколько Ваше физическое и эмоциональное состояние в течение последних 4 недель мешало Вам проводить время с семьей, друзьями, соседями или в коллективе? (Обведите одну цифру)

Совсем не мешало 1

Немного 2

Умеренно 3

Сильно 4

Очень сильно 5

7. Насколько сильную физическую боль Вы испытывали за последние 4 недели? (Обведите одну цифру)

Совсем не испытывал(а) 1

Очень слабую 2

Слабую 3

Умеренную 4

Сильную 5

Очень сильную 6

8. В какой степени боль в течение последних 4 недель мешала Вам заниматься Вашей нормальной работой (включая работу вне дома или по дому)? (Обведите одну цифру)

- Совсем не мешала 1
 Немного 2
 Умеренно 3
 Сильно 4
 Очень сильно 5

9. Следующие вопросы касаются того, как Вы себя чувствовали и каким было Ваше настроение в течение последних 4 недель. Пожалуйста, на каждый вопрос дайте один ответ, который наиболее соответствует Вашим ощущениям (табл. 30):

Таблица 30

(обведите одну цифру)

	Все время	Большую часть времени	Часто	Иногда	Редко	Ни разу
А. Вы чувствовали себя бодрым(ой)?	1	2	3	4	5	6
Б. Вы сильно нервничали?	1	2	3	4	5	6
В. Вы чувствовали себя таким(ой) подавленным(ой), что ничто не могло Вас взбодрить?	1	2	3	4	5	6
Г. Вы чувствовали себя спокойным(ой) и умиротворенным(ой)?	1	2	3	4	5	6
Д. Вы чувствовали себя полным (ой) сил и энергии?	1	2	3	4	5	6
Е. Вы чувствовали себя упавшим(ой) духом и печальным(ой)?	1	2	3	4	5	6
Ж. Вы чувствовали себя измученным(ой)?	1	2	3	4	5	6
З. Вы чувствовали себя счастливым(ой)?	1	2	3	4	5	6
И. Вы чувствовали себя уставшим(ей)?	1	2	3	4	5	6

10. Как часто за последние 4 недели Ваше физическое или эмоциональное состояние мешало Вам активно общаться с людьми (навещать друзей, родственников и т.п.)? (Обведите одну цифру)

- Все время 1
 Большую часть времени 2
 Иногда 3
 Редко 4
 Ни разу 5

Насколько ВЕРНЫМ или НЕВЕРНЫМ представляется по отношению к Вам каждое из ниже перечисленных утверждений? (табл.31)

Таблица 31

(обведите одну цифру в каждой строке)

	Определенно верно	В основном верно	Не знаю	В основном неверно	Определенно неверно
а. Мне кажется, что я более склонен к болезням, чем другие	1	2	3	4	5
б. Мое здоровье не хуже, чем у большинства моих знакомых	1	2	3	4	5
в. Я ожидаю, что мое здоровье ухудшится	1	2	3	4	5
г. У меня отличное здоровье	1	2	3	4	5

Все шкалы опросника объединены (табл. 32) в 2 суммарных измерения – физический компонент здоровья (1-4 шкалы) и психический (5-8 шкалы). Каждая из этих шкал имеет свой набор вопросов (учетных признаков) и особенности вычисления показателей (табл. 33).

Шкалы опросника

1	Физическое функционирование (PF)
2	Рольное (физическое) функционирование (RP)
3	Боль (P)
4	Общее здоровье (GH)
5	Жизнеспособность (VT)
6	Социальное функционирование (SF)
7	Эмоциональное функционирование (RE)
8	Психологическое здоровье (MH)

Методика вычисления основных показателей по опроснику SF-36

Показатели	Вопросы	Минимальное и максимальное значения	Возможный диапазон значений
Физическое функционирование (PF)	3а, 3б, 3в, 3г, 3д, 3е, 3ж, 3з, 3и, 3к	10 - 30	20
Рольное (физическое) функционирование (RP)	4а, 4б, 4в, 4г	4 - 8	4
Боль (P)	7, 8	2 - 12	10
Общее здоровье (GH)	1, 11а, 11б, 11в, 11г	5 - 25	20
Жизнеспособность (VT)	9а, 9д, 9ж, 9и	4 - 24	20
Социальное функционирование (SF)	6, 10	2 - 10	8
Эмоциональное функционирование (RE)	5а, 5б, 5в	3 - 6	3
Психологическое здоровье (MH)	9б, 9в, 9г, 9е, 9з	5 - 30	25

В пунктах 6, 9а, 9д, 9г, 9з, 10, 11 производится обратный счет значений. Формула вычисления значений: [(реальное значение показателя) – (минимально возможное значение показателя)] : (возможный диапазон значений) * 100.

Обобщающие показатели качества жизни, полученные по опроснику SF-36, обычно приводятся по отдельным группам (см. табл. 33) в виде средних величин и соответствующих им значений средних ошибок репрезентативности.

3.2.5. Заболеваемость с временной утратой трудоспособности

Несмотря на то, что изучение заболеваемости уже многие десятилетия привлекает к себе самое пристальное внимание статистиков, оценка здоровья населения связана с существованием ряда нерешенных методологических проблем. В том числе нерешенной проблемой остаются объективные оценки тяжести болезни и степени ущерба для здоровья, нанесенного развитием патологического процесса.

Один из основоположников отечественной санитарной статистики А.М. Мерков (1899-1971гг.) писал по этому поводу еще в 1937 году: *«Принимая всех больных с данным диагнозом за равные единицы, мы в значительной мере лишаемся возможности учитывать эффективность оздоровительных мероприятий и изучать динамику заболеваемости... по тяжести заболеваний»*. Эта проблема во многом обострилась в связи с бурным развитием в последние десятилетия эффективной клинической диагностики. Точная клиническая диагностика породила массовое появление объемных, многословных диагнозов, конкретизирующих течение заболеваний у отдельных больных. Резко увеличилась выявляемость состояний, которые можно расценивать как состояния угрозы возникновения заболевания, что вызвало появление известного афоризма: «Нет здорового человека, есть плохо обследованный человек». В этих условиях невозможность статистического учета тяжести заболеваний все больше сказывается на эффективности и результативности анализа общественного здоровья.

Одним из способов решения этой проблемы является изучение заболеваемости по данным *временной нетрудоспособности*. Как статистический показатель общественного здоровья заболеваемость с временной утратой трудоспособности (ЗВУТ) опирается на оценку трудоспособности. *Трудоспособность* – совокупность духовных и физических возможностей человека, зависящих от состояния его здоровья, позволяющих заниматься ему трудовой деятельностью. (Энциклопедический словарь медицинских терминов, 1984). 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 ЗВУТ является то, что основанием для выдачи больничного листа 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. Тем не менее, статистика заболеваемости с временной утратой трудоспособности занимала особое место в санитарной статистике СССР и продолжает занимать заметное место в номенклатуре современной гос. статистики РФ, используя при планировании и оценке эффективности медико-социальных мероприятий среди работающего населения, при изучении влияния на здоровье работающих неблагоприятных производственных факторов. Среди причин высокой информативности статистики ЗВУТ следует отметить: жесткую регламентацию условий и порядка регистрации случаев ЗВУТ, относительную простоту сбора информации и обработки показателей, тройной государственный контроль (финансовый, административно-производственный, медицинский) за выдачей больничных листов.

Учетными документами ЗВУТ служат «Книга регистрации листов нетрудоспособности» (форма № 036/у), «Талон на законченный случай с

временной утратой трудоспособности» (форма 025 - 9/у), «Талон амбулаторного пациента» (форма 025-12/у).

К началу 90-х годов этот вид статистики заболеваемости потерял своё значение из-за крупных преобразований, произошедших в стране. Регистрация случаев временной нетрудоспособности перестала быть полной и достоверной 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, недекларируемой заработной платы работника, что сплошь и рядом стало приводить к отказу от «ухода на больничный лист». По мере улучшения экономической ситуации, упорядочения взаимоотношений между работодателями и работниками на производстве происходит рост полноты и достоверности статистики ЗВУТ.

Немаловажным обстоятельством, сыгравшим заметную роль в снижении информативной ценности ЗВУТ как источника статистического изучения распространенности той или иной патологии, стало то, что с конца 80-х гг. с целью соблюдения врачебной тайны в больничных листках диагноз заболевания стал указываться только с согласия пациента. Отказ от обязательной формулировки диагноза на практике полностью исключил возможность проводить статистический анализ структуры заболеваемости и распространенности отдельных групп заболеваний с временной нетрудоспособностью.

Первичный учетный документ, фиксирующий факт заболевания, повлекшего утрату трудоспособности, – листок временной утраты трудоспособности, а 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, поскольку служит основанием для выплаты пособия по болезни.

Основанием для выдачи больничного листа является экспертиза трудоспособности, которая устанавливает наличие, тяжесть заболевания и его прогноз, определяет возможность человека выполнять профессиональные обязанности.

Годовая форма отчета (форма 16-ВН) является официальным статистическим отчетным документом федеральной статистики РФ, в котором регистрируются накопленные за год данные о случаях ЗВУТ. Следует помнить, что данные ЗВУТ, накопленные за месяц, квартал и полугодие, всегда являются неполными, так как часть больничных листов не успевает быть зарегистрированной в указанные периоды из-за того, что ещё находится на руках у болеющих. В квартальный отчет может не попадать до 30% больничных листов. Особенно сильно этот фактор сказывается при малом числе работающих на предприятии или учреждении.

Официальный отчет о ЗВУТ 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 статистическую стандартизацию показателей.

Методика углубленного изучения заболеваемости с выделением физических лиц, так называемый полицейский учет ЗВУТ, позволяет проводить детальный статистический анализ с учетом основных социально-демографических, производственно-профессиональных и других факторов, способствующих возникновению заболеваний. Только полицейский учет позволяет анализировать распределение работников по кратности заболеваний

в году, определять группу ни разу не болевших в году (индекс здоровья), группу часто и длительно болеющих (ЧДБ). Подобная группировка исходной информации по ЗВУТ позволяет дифференцированно планировать и проводить необходимый комплекс лечебно-профилактических и оздоровительных мероприятий.

Обычно для углубленного анализа ЗВУТ отбираются лица, которые проработали на данном производстве не менее 1 года. Так называемые «круглогодные работники». Ранее такие исследования проводились на основе карт полицейского учета, которые должны были заполняться 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 трудоёмкости учета этой информации и обработки таких карт в широкой практике полицейской учет распространения не получил. По мере создания автоматизированных информационных баз данных, в том числе баз данных системы медицинского страхования, возникает возможность возрождения статистики полицейского учета.

В последнее десятилетие появились новые организационные и методические аспекты статистики ЗВУТ, связанные с коренными изменениями социально-экономического уклада жизни в стране и бурным развитием информационных технологий. Одним из существенных социально-экономических факторов этого периода времени, повлиявших на ЗВУТ как критерий общественного здоровья, явился глубокий экономический кризис в стране. Резкое падение жизненного уровня в 90-х гг., сокращение размеров пособия по утрате трудоспособности, уход частных работодателей от проблем охраны здоровья своих работников существенно снизили информативность официальной статистики ЗВУТ. В то же время бурное развитие в эти годы информационных технологий играло и играет позитивную роль, предоставляя возможность сосредотачивать большие объемы наблюдений, систематизировать их в разнообразные базы данных и на этой основе

осуществлять глубокий и всесторонний анализ ЗВУТ, применяя сложнейшие алгоритмы многофакторной статистической обработки данных.

Традиционно в качестве главных статистических характеристик ЗВУТ используются:

- частота (распространенность) случаев заболеваний;
- число дней временной нетрудоспособности;
- структура распределения случаев и дней ЗВУТ с учетом диагноза заболевания;
- средняя продолжительность одного случая заболевания.

Эти показатели, вычисляемые с учетом отдельных социально-демографических, производственно-профессиональных и других исследуемых характеристик, способных влиять на заболеваемость как медико-социальное явление, являются обязательной основой любой статистики ЗВУТ, будь то уровень здравпункта, региональный или общегосударственный уровень. Как уже отмечалось, дополнительной статистической характеристикой, получаемой на основе полицейского учета заболеваемости, являются показатели кратности ЗВУТ.

Общие показатели ЗВУТ за год вычисляются на основе учета суммарных показателей и учета среднегодовой численности работников предприятия. Среднегодовая численность работающих в наиболее простом варианте, в случае относительной стабильности численности работников, определяется, как полусумма численности работников предприятия на конец и начало года. Соответственно, если показатели ЗВУТ вычисляются за квартал или полугодие, то средняя численность работников вычисляется, как полусумма из численности на начало и конец квартала или полугодия.

Показатели ЗВУТ за квартал или полугодие могут вычисляться при достаточно большой численности работающих (не менее 1000 человек). При меньшей численности работников вычисление этих показателей теряет практический смысл. Одной из причин этого обстоятельства является то, что при учете ЗВУТ за квартал к концу текущего квартала остаются

незарегистрированными до 25% больничных листов. За полугодие, соответственно, 10-15%. Помесячный учет больных листков ведет к еще большему недоучету заболеваемости.

Такая неполная регистрация связана с тем, что часть заболеваний на момент подсчета итогов еще продолжается, т.е. случаи заболеваний еще не закончены или больничные листы еще не успели закрыть или, если уже закрыты, но не успели зарегистрировать. При учете больничных листов, особенно за неполный год, необходимо учитывать, что при длительном заболевании по поводу одного случая ЗВУТ может выдаваться последовательно несколько больничных листков. Из-за этого случай заболевания может оказаться зарегистрированным, а все дни нетрудоспособности по этому случаю ЗВУТ будут учтены полностью только после регистрации последнего больничного листка.

Частота случаев (синонимы: распространенность, уровень) заболеваемости с ВУТ за год на 100 работников:

$$\frac{\text{Число случаев заболеваний с ВУТ} \times 100}{\text{Среднегодовая численность работников}}$$

Среднегодовая численность работников

Число дней заболеваемости с ВУТ за год на 100 работников:

$$\frac{\text{Число случаев заболеваний с ВУТ} \times 100}{\text{Среднегодовая численность работников}}$$

Среднегодовая численность работников

При качественной оценке частоты зарегистрированных случаев ЗВУТ следует помнить, что эти показатели во многом определяются социально-демографическим составом контингентов работников, достаточностью медицинской помощи, характером экономических отношений на производстве, доступностью и особенностями организации медико-профилактической помощи и т.п. Поэтому реальную ценность, как источник информации, эти показатели представляют при сравнительной оценке динамики показателей по

годам в одних и тех же конкретных контингентах работников или при сравнении распространенности ЗВУТ на одних и тех же территориях, в конкретных отраслях промышленности или в разных подразделениях предприятия.

Следует отметить, что никаких нормативов по уровню показателей ЗВУТ не существует. Существуют лишь рекомендации по длительности случаев ЗВУТ, в днях временной нетрудоспособности: «Ориентировочные сроки временной нетрудоспособности при наиболее распространенных заболеваниях и травмах (в соответствии с МКБ-10)», утвержденные Министерством здравоохранения РФ и Фондом социального страхования РФ в 2000 году. Приведенные в этом документе сроки временной утраты трудоспособности носят ориентировочный, рекомендательный характер. Вместе с тем, в документе подчеркивается, что увеличение или сокращение ориентировочных сроков ВУТ (на 30 и более процентов) должно служить поводом для проведения экспертизы ВУТ заведующим отделением, клинико-экспертной комиссией (КЭК) с оценкой объемов, качества и эффективности медицинской помощи, применения современных медицинских технологий, своевременного привлечения к лечебному процессу других специалистов или медицинских учреждений с целью определения объективных или субъективных факторов и принятия адекватных мер.

Объективно обоснованная продолжительность нетрудоспособности в днях ЗВУТ характеризует необходимую длительность освобождения больных от работы для проведения диагностических, лечебных и реабилитационных мероприятий для восстановления утраченной трудоспособности. При неблагоприятном трудовом и клиническом прогнозе требуется время, необходимое для медико-социальной экспертизы по поводу признания лица инвалидом. Кроме того, показатель длительности трудопотери необходим для расчета экономических потерь в связи с ЗВУТ.

Длительность 1 случая является стат. характеристикой тяжести ЗВУТ:

Число случаев ЗВУТ/число дней ЗВУТ

Средний уровень этого показателя ЗВУТ обычно находится в пределах 10-14 дней временной нетрудоспособности на 1 случай заболевания. Особенностью этого показателя является его способность выступать в качестве индикатора административных форм борьбы с ЗВУТ. При попытке насильственного давления на ситуацию с целью снижения частоты случаев ЗВУТ неизбежно возрастает средняя продолжительность одного случая.

Необходимо помнить, что при статистическом анализе дней ЗВУТ учитываются календарные дни, а при анализе экономических потерь более целесообразным является учет рабочих дней. Однако из-за того, что официальная статистика ЗВУТ строится на регистрации календарных дней, на практике приходится проводить расчет потерь на основе календарных дней ЗВУТ. При использовании автоматизированных баз данных, когда появляется возможность на основе использования специальных, «машинных» алгоритмов обработать данные, целесообразно в этих расчетах опираться на учет рабочих дней. К числу экономических характеристик ЗВУТ можно отнести (табл.34):

Процент нетрудоспособности показывает процент лиц, условно не работавших в году:

Число календарных дней ВУТ в связи с заболеванием x 100

Среднегодовая численность работников x 365 дней (366 в високосном году)

Потери предприятием дней в отчетном году. Показатель демонстрирует, сколько дней в году из-за заболеваемости не работает предприятие и т.п.:

Число календарных дней ВУТ в связи с заболеванием

Среднегодовая численность работников

Если соотнести случаи трудопотерь и число дней в году, то получаем показатель, демонстрирующий среднее число *случаев невыхода на работу ежедневно в отчетном году*:

Число случаев ВУТ в связи с заболеванием

365 дней (366 в високосном году)

Коэффициент трудоспособности:

(Число раб. x 365 – число календарных дней нетрудоспособности) x 100

Число работающих x 365

Таблица 34

**Некоторые экономические показатели последствий ЗВУТ в 2007 году
одном из предприятий Санкт-Петербурга**

Экономические показатели	Номера филиалов предприятия						Всего
	1	2	3	4	5	6	
% нетрудоспособности	4	4	5	5	5	4	4,4
Потеряно предприятием дней	9	9	10	7	9	4	9
Среднее число ежедневных невыходов на работу	5	5	2	1	0	0	12,1
Коэффициент трудоспособности предприятия (%)	98	98	97	98	98	99	97,5

На протяжении многих десятилетий в России одними из основных показателей в статистике ЗВУТ считались показатели структуры заболеваемости или, другими словами, удельного веса тех или иных заболеваний среди ЗВУТ отдельно в случаях и днях нетрудоспособности (табл. 35). Эта структура заболеваемости позволяет выявлять ведущие группы заболеваний, «узкие» места в организации труда врачей-специалистов, обосновывать приоритетность отдельных форм профилактической работы.

**Сравнительная структура ЗВУТ в днях
нетрудоспособности работников одного из предприятий Санкт-Петербурга**

Сокращенные наименования классов МКБ 10	2006 год	2007 год	Прирост или снижение
Болезни органов дыхания	27,2	31,8	4,5
Болезни костно-мышечной системы и соединительной ткани	16,5	13,1	-3,4
Травмы	14,5	15,1	0,6
Болезни системы кровообращения	13,2	14,3	1,1
Болезни органов пищеварения	6,7	5,9	-0,8
Факторы, влияющие на состояние здоровья	2,1	4,7	2,6
Болезни мочеполовой системы	4,1	3,3	-0,9
Болезни кожи и подкожной клетчатки	2,8	2,2	-0,6
Другие	1,6	1,8	0,2
Некоторые инфекционные и паразитарные болезни	1,8	1,4	-0,4
Болезни нервной системы	3,0	2,0	-1,0
Болезни глаза и его придаточного аппарата	1,5	1,2	-0,4
Беременность, роды и послеродовой период	0,5	1,1	0,6
Болезни уха и сосцевидного отростка	1,0	0,5	-0,5
Болезни эндокринной системы и нарушения обмена веществ	1,6	0,7	-0,9
Новообразования	1,7	0,9	-0,9
Общий итог	100,0	100,0	100,0

Как уже отмечалось, в настоящее время в больничных листках регистрация диагнозов по каждому случаю ЗВУТ не производится. Более того, попытки на отдельных предприятиях административным путем ввести такую регистрацию считаются незаконными. Таким образом, вычисление показателей структуры заболеваемости с ВУТ на практике остается возможным только в пределах учреждений, выдавших больничные листы. Другим источником получения информации по конкретным выставленным диагнозам могут

являться территориальные фонды обязательного медицинского страхования, которые обладают базами данных, где, как это принято в ряде регионов, отражаются диагнозы заболеваний, служившие причиной обращения за медицинской помощью.

При анализе структуры вычисляются два основных показателя заболеваемости по классам МКБ 10 (в ряде отчетных форм представлены не полное наименование отдельных классов, а более мелкая рубрикация МКБ): структура заболеваемости в случаях и днях ВУТ.

Структура ЗВУТ в случаях временной нетрудоспособности:

Число случаев заболеваний с ВУТ по данному классу заболеваний x 100

Число случаев заболеваний с ВУТ по всем классам заболеваний

Структура ЗВУТ в днях временной нетрудоспособности:

Число календарных дней ВУТ в связи с заболеваниями данного класса x 100

Число календарных дней ВУТ в связи со всеми заболеваниями с ВУТ

В структуре заболеваемости с ВУТ почти половину всех случаев и до 60% всех дней нетрудоспособности составляют заболевания, в просторечье называемые простудными: острые респираторные заболевания и грипп, ангина, фарингит, тонзиллит и т.п. Вторую по значимости группу заболеваний среди работников промышленных предприятий обычно занимают заболевания нервной системы, травмы, болезни костно-мышечной системы, заболевания сердечно-сосудистой системы. Эта последовательность не является стандартной, поскольку сильно зависит от условий, характера труда и возраста работников.

Учитывая значительный удельный вес, который занимают простудные заболевания, рекомендуется дополнительно вычислять итоговые показатели ЗВУТ без учета перечисленных простудных заболеваний. Таким образом, элиминируются значительные сезонные колебания частоты случаев этой

группы болезней, что важно при месячном и поквартальном анализе основных тенденций ЗВУТ. Кроме того, в ряде случаев и годовые колебания показателей ЗВУТ могут быть следствием эпидемических подъемов частоты респираторных инфекций в отдельные годы.

Пример вычисления показателей ЗВУТ: за 2007 год было зарегистрировано 4770 случаев и 49806 календарных дней ЗВУТ. В том числе 580 случаев ЗВУТ в связи с сердечно-сосудистыми заболеваниями и 7720 дней нетрудоспособности по этой причине.

Всего на предприятии на начало 2007 года числилось 8000 постоянных работников. К концу года в связи с сокращением штатов на предприятии числилось уже 7100 человек. Среднегодовая численность работников составила $(8000 + 7100)/2 = 7550$ человек. Более детально практическое вычисление, ориентировочные границы и стандартное наименование основных показателей ЗВУТ представлены ниже (табл. 36).

Примеры расчета показателей ЗВУТ

Номенклатура показателей	Порядок вычисления	Результат
Частота ЗВУТ за год	$\frac{4770 \text{ случаев} \times 100}{7550 \text{ работников}}$	63,2 сл. на 100 раб. за год
Число дней ЗВУТ за год	$49806 \text{ дней} \times 100$ 7550 работников	659,7 дня на 100 раб. за год
Средняя продолжительность случая ЗВУТ в году	$\frac{49806 \text{ дней ЗВУТ}}{4770 \text{ случаев ЗВУТ}}$	Продолжительность одного случая ЗВУТ 10,4 дня
Доля отдельных классов ЗВУТ в общем числе случаев нетрудоспособности (при наличии данных по отдельным классам заболеваний)	$\frac{580 \text{ случаев сердечно-сосудистых заболеваний (09 класс МКБ10)} \times 100}{4770 \text{ всех случаев ЗВУТ}}$	Удельный вес сердечно-сосудистых (9 класс МКБ10) заболеваний среди всех случаев ЗВУТ составил 12,2%
Доля отдельных классов ЗВУТ в днях нетрудоспособности (при наличии данных по отдельным классам заболеваний)	$\frac{7720 \text{ дней ВУТ в связи с сердечно-сосудистыми заболеваниями (09 класс МКБ10)} \times 100}{49806 \text{ всех дней ЗВУТ}}$	Удельный вес сердечно-сосудистых (9 класс МКБ10) заболеваний среди всех дней ЗВУТ составил 15,5%
Средняя продолжительность случая ЗВУТ по определенному классу заболеваний в году	$\frac{7720 \text{ дней ВУТ в связи с сердечно-сосудистыми заболеваниями}}{580 \text{ случаев сердечно-сосудистых заболеваний}}$	13,3 дня – средняя продолжительность 1 случая сердечно-сосудистых заболеваний
Процент нетрудоспособности в связи с ВУТ в году	$\frac{49806 \text{ дней ЗВУТ} \times 100}{4770 \text{ случ.} \times 365 \text{ дней}}$	2,9% всех раб. предприятия не работали в году из-за случаев ЗВУТ
Потери предприятия дней в текущем году из-за ЗВУТ	49806 дней ЗВУТ 7550 работников	Теоретически из-за ЗВУТ предприятие не работало в году 6,6 дня
Среднее число случаев невыхода на работу ежедневно в связи с ЗВУТ	$\frac{7720 \text{ случаев ЗВУТ}}{365 \text{ дней}}$	Среднее число случаев ЗВУТ на 1 календарный день в году – 13,1 дня
Коэффициент трудоспособности коллектива предприятия	$\frac{7550 \text{ раб.} \times 365 \text{ дней}}{7550 \text{ раб.} \times 365 \text{ дней}}$	Коэффициент трудоспособности – 98%

Приближенная качественная оценка перечисленных показателей ЗВУТ может быть осуществлена по шкале, предложенной Е.Л. Ноткиным еще в 1979 году. Несмотря на существенные социально-экономические изменения, произошедшие в России с того времени, эти границы оценок могут приниматься во внимание при статистическом анализе годовых показателей ЗВУТ. При этом следует учитывать, что практически для каждого предприятия характерны свои условия и характер труда, которые, даже оставаясь в пределах существующих санитарных норм, неизбежно откладывают отпечаток на уровни показателей ЗВУТ.

Кроме того, на каждом предприятии существуют определенные особенности организации труда, принципы оплаты и стимулирования трудовой деятельности и т.п. факторы, способные оказывать существенное влияние на уровень и структуру трудопотерь работников. Таким образом, представленные данные (табл. 37) могут использоваться только для приблизительной, ориентировочной информации.

Таблица 37

**Шкала оценки основных показателей заболеваемости
с временной утратой трудоспособности по Е.Л. Ноткину (1979)**

Уровень показателей	Процент болевших лиц	Временная утрата трудоспособности на 100 раб. за год	
		Число случаев	Число дней
Очень высокий	80 и более	150 и более	1500 и более
Высокий	70-79	120-149	1200-1499
Выше среднего	60-69	100-119	1000-1199
Средний	50-59	80-99	800-999
Ниже среднего	40-49	60-79	600-799
Низкий	35-39	50-59	500-599
Очень низкий	Менее 35	Менее 50	Менее 500

Более информативным представляется анализ показателей ЗВУТ по конкретному предприятию или производственным участкам предприятия в динамике за 2-3 года и более.

Для анализа изменения показателей ЗВУТ в динамике определяют отношение уровня показателя заболеваемости за анализируемый период к исходному показателю, принятому за 100%. Аналогичным образом производят сравнительную оценку показателей, полученных в разных группах наблюдения (например: предприятие А и предприятие Б). В аналитических таблицах данные такого рода сравнения обычно представляются в графе «% изменений». В статистике можно встретить еще одно, более точное название этого показателя – «темп роста» или «темп убыли», в зависимости от направленности изменений. Для оценки, например, динамики случаев общего уровня ЗВУТ в случаях нетрудоспособности показатель вычисляется следующим образом:

$$\frac{\text{Число случаев ЗВУТ работников предприятия за 2007 год} \times 100}{\text{Число случаев ЗВУТ за предыдущий год (2006 г. или 2005 г. и т.д.)}}$$

Аналогичным образом вычисляется показатель темпа прироста и для сравнения дней ЗВУТ, и для оценки динамики средней продолжительности одного случая заболевания с ВУТ. Этот же показатель темпа роста может применяться и для сравнительной оценки динамики показателей заболеваемости по отдельным нозологическим формам.

Например: в 2006 году на 100 работников предприятия было зарегистрировано 42 случая ВУТ в связи с заболеваниями органов дыхания, в 2007 – 45 случаев ВУТ по этим же заболеваниям. Соответственно, в 2006 году на 100 работающих было зарегистрировано 11 случаев ВУТ в связи с болезнями органов кровообращения, а в 2007 – 14 случаев.

$$\frac{45 \text{ случаев ЗВУТ органов дыхания на } 100 \text{ раб. за } 2007 \text{ год} \times 100}{41 \text{ случай ЗВУТ органов дыхания на } 100 \text{ раб. за } 2006 \text{ год}}$$

$$\frac{10 \text{ случаев ЗВУТ органов кровообращения на } 100 \text{ раб. за } 2007 \text{ год} \times 100}{14 \text{ случаев ЗВУТ органов кровообращения на } 100 \text{ раб. за } 2006 \text{ год}}$$

К числу достоинств показателя темпа роста относится его наглядность. Вместе с тем, требуется обращать внимание на то, что временной период (расчетный интервал времени) сравниваемых показателей должен быть абсолютно одинаков. Кроме того, при оценке значимости изменений необходимо помнить, что показатель темпа роста в процентах имеет свою «стоимость», зависящую от размеров сравниваемых величин. Так, в приведенном выше примере по заболеваниям органов кровообращения отмечался более высокий темп роста, чем по заболеваниям органов дыхания (соответственно, 127% и 110%). Однако оба этих класса заболеваний дали одинаковый финальный итог – по 4 случая ЗВУТ на 100 работников за год.

Общие показатели ЗВУТ малопригодны для выявления роли конкретных социально-демографических, производственно-профессиональных и т.п. факторов утраты или, наоборот, укрепления здоровья работающего населения. Вместе с тем, единая для всей России методика сбора, регистрации и учета данных, тесная связь этих показателей с факторами внешней среды, индивидуальными особенностями личности и общества делают показатели ЗВУТ весьма информативными.

Углубленный анализ ЗВУТ даже на уровне отдельного амбулаторно-поликлинического учреждения, оказывающего медико-профилактическую помощь работникам конкретного предприятия, играет весьма существенную роль в оценке медицинской, экономической и социальной эффективности деятельности этого учреждения. С точки зрения статистики, при углубленном анализе ЗВУТ используются две категории показателей.

Показатели стандартного учета (частота, структура и средняя продолжительность одного случая ВУТ). В этом случае единицей учета является больничный листок, но с целью получения углубленной информации статистика ЗВУТ ведется не на весь контингент работающих, а относительно отдельных возрастно-половых, стажевых, профессиональных и других групп.

Показатели полцевого учета ЗВУТ. Используются только при углубленном изучении заболеваемости. Единицей наблюдения в этом случае

является каждый отдельно взятый работник и соответствующие ему учетные признаки: пол, возраст, профессия, характер, условия труда и т.п. Несмотря на кажущуюся схожесть этих двух категорий показателей, только последняя дает возможность получать детальные характеристики распределения работающих по группам различной кратности временной утраты трудоспособности (неболевшие в году, болевшие 1 раз, 2 раза, 3 раза, 4 раза и более), получать детальные группировки возрастно-половых, профессиональных и т.п. характеристик, не предусмотренных стандартными формами отчетности, или предусмотренных, но только в очень общем виде.

Возрастно-половые характеристики – один из обязательных компонентов, который должен учитываться при статистическом анализе показателей ЗВУТ, как при общем, так и углубленном анализе. Это требование обусловлено существенной зависимостью уровней и структуры ЗВУТ от возрастно-полового состава работников.

Например, специалистам хорошо известно, что заболеваемость гипертонической болезнью, ангиной, болезнью почек и желчевыводящих путей среди женщин выше, чем среди мужчин. Язвенная болезнь желудка и 12-перстной кишки, грыжи, болезни костно-мышечной системы, травмы выше среди мужчин. Кроме того, при анализе ЗВУТ необходимо учитывать, что опосредованной причиной различий показателей этого вида заболеваемости почти всегда бывает неодинаковое профессиональное распределение мужчин и женщин, так же как и неодинаковое распределение различных возрастных групп среди работников разных профессий.

Стаж работы имеет также весьма существенное значение при углубленном анализе ЗВУТ. При этом необходимо помнить, что, согласно действующим положениям, выделяют как минимум несколько видов стажа, каждый из которых может играть собственную роль в распределении показателей заболеваемости: общий трудовой стаж, стаж работы на данном предприятии, стаж работы по данной профессии, стаж работы на данном месте и т.п. Кроме того, существенное влияние на уровни заболеваемости оказывают

лица, прервавшие в течение года свою работу на предприятии по данной профессии, на данном месте и т.п. По данным специальных исследований, считается, что заболеваемость среди уволенных работников всегда выше, чем среди отработавших не менее 1 года. С другой стороны, среди лиц, принятых на работу в текущем году, уровень заболеваемости всегда ниже. Таким образом, на показатели ЗВУТ существенное влияние может оказывать текучесть кадров. Для того, чтобы избавиться от влияния этого фактора, при углубленном анализе и в ходе специальных исследований показатели заболеваемости принято рассчитывать только для лиц, отработавших не менее 1 года, так называемых «круглогодových» работников. Соответственно, могут вычисляться показатели ЗВУТ отдельно для уволенных лиц и для принятых на работу в течение года.

В ситуации, когда основной контингент на предприятии – временные работники или существует большая текучесть кадров, а также изменение численного состава по другим причинам (реструктуризация предприятия, расширение производства и т.п.), показатели ЗВУТ вычисляются относительно «круглогодových» работников. При расчете обычных показателей ЗВУТ в числителе ставится число случаев (дней) ЗВУТ «круглогодových» работников за год (полгода, месяц), а в знаменателе – число календарных дней в году.

Считается, что показатели ЗВУТ «круглогодových» работников по итогам 3-летнего наблюдения наиболее точно отражают влияние производства, социальных факторов и качество медико-профилактического обслуживания. В данном случае необходимо учитывать, что после выделения работников, которые отработали не менее 3-х лет на производстве, объем статистического наблюдения, как правило, существенно сокращается. В свою очередь, такое сокращение может вести к снижению качественной и количественной репрезентативности полученных результатов. Кроме того, оценка ЗВУТ по «круглогодovým» работникам с целью более точного отслеживания динамики развития патологических процессов может обесцениваться другим фактором – фактором естественного отбора, хорошо известным специалистам

промышленной медицины. Здесь подразумевается действие системы контроля за состоянием здоровья работающих на производствах с вредными и опасными условиями труда. Одним из результатов такого контроля является административное удаление с производства лиц, чье здоровье не соответствует требованиям, обусловленным неблагоприятными условиями и характером труда. При этом многие из этих больных, согласно действующему законодательству, трудоустраиваются на работу в более благоприятных условиях. Итогом нередко становится повышенная «концентрация» хронически больных там, где условия труда благоприятные, и пониженная концентрация там, где условия труда неблагоприятные. Следует отметить, что эта тенденция отбора может проявляться и естественным путем, без административного давления.

В результате на тех участках производства, где имеются неблагоприятные факторы, ЗВУТ гораздо меньше, чем на участках с более комфортными гигиеническими условиями труда.

Одним из самых информативных показателей, использующихся в углубленном анализе ЗВУТ, является показатель доли (удельного веса) группы работников, не имевших случаев ЗВУТ в году. Нередко для обозначения этого показателя используется термин «индекс здоровья». Величина индекса здоровья во многом зависит от возраста работников. Наиболее высокий уровень показателя отмечается в молодых возрастных группах. По мере увеличения возраста работников его величина снижается.

Удельный вес лиц, не имевших ЗВУТ («индекс здоровья»):

Число лиц, не имевших случаи ЗВУТ в году x 100

Среднегодовая численность работников

Показатель, имеющий обратное смысловое значение, т.е. удельный вес лиц, имевших в году случаи ЗВУТ:

Число лиц, имевших случаи ЗВУТ в году x 100

Среднегодовая численность работников

Или

«Индекс здоровья» – 100%

Распределение работников по кратности ЗВУТ может вычисляться в двух вариантах. В одном – рассматривается распределение всех работающих, в том числе и неболевших:

Число лиц, имевших в году 1 (2 или 3) случ. ЗВУТ x 100

Среднегодовая численность работников

В другом варианте учитывается распределение только болевших работников:

Число лиц, имевших в году 1 (2 или 3) случ. ЗВУТ x 100

Число работников, имевших в году случаи ЗВУТ

В ряду показателей углубленного анализа ЗВУТ особое место занимает характеристика группы часто и длительно болеющих – ЧДБ. Иногда без изменения смыслового содержания употребляется другая последовательность букв в этой аббревиатуре – ДЧБ. На сегодняшний день отсутствуют единые критерии формирования этих групп. Наиболее целесообразным представляется формирование этой группы на основе учета лиц, имевших 4 и более случая (и) или 40 дней и более ЗВУТ за год. При низких уровнях ЗВУТ рекомендуется формировать группу ЧДБ из числа лиц, имевших 3 и более случая или 30 и более дней ЗВУТ за год. В Руководстве по оценке профессионального риска для здоровья работников, утвержденном Главным государственным санитарным врачом РФ 24.06.2003 г., в качестве критерия длительного расстройства здоровья рекомендуется 60-дневная длительность временной утраты трудоспособности.

Особенностью группы ЧДБ (4 случая и 40 дней ЗВУТ за год) является то, что обычно она составляет от 3% до 8% от числа всех работников предприятия, но на долю этих работников может приходиться свыше 40% всех дней ВУТ в связи с заболеваниями. Группа часто и длительно болевших (ЧДБ) определяется по формуле:

$$\frac{\text{Число лиц, имевших в году} \\ \text{4 и более случая и (или) 40 и более дней ЗВУТ} \times 100}{\text{Среднегодовая численность работников}}$$

3.2.6. Стойкая утрата трудоспособности, инвалидность

Инвалидность – понятие, основа которого – постоянная или длительная, полная или частичная потеря трудоспособности. Согласно Федеральному закону № 181-ФЗ РФ от 24 ноября 1995 года (с изменениями и дополнениями, внесенными до 12.12.2011 г.), *инвалид* (лат. *Invalidus* – «бессильный», «слабый») — лицо, которое имеет нарушение здоровья со стойким расстройством функций организма, обусловленное заболеваниями, последствиями травм или дефектами, приводящее к ограничению жизнедеятельности и вызывающее необходимость его социальной защиты. Ограничение жизнедеятельности – полная или частичная утрата лицом способности или возможности осуществлять самообслуживание, самостоятельно передвигаться, ориентироваться, общаться, контролировать свое поведение, обучаться и заниматься трудовой деятельностью.

Статистически в общую численность инвалидов включаются лица, состоящие на учете и получающие пенсию в системе Пенсионного фонда Российской Федерации (до 2002 г. – по данным бывшего Минтруда России и бывшего Минздрава России), Министерстве обороны Российской Федерации, Министерстве внутренних дел Российской Федерации, Федеральной службе безопасности Российской Федерации.

Первичный выход на инвалидность – установление инвалидности впервые в отчетном году.

По данным ВОЗ, к концу XX столетия 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 большинство инвалидов (37-50%) становились инвалидами в трудоспособном возрасте. Общая численность инвалидов колебалась от 10648 тыс. чел. в 2000 году до 13134 тыс. чел – в 2010 году. В РФ причинами инвалидности в настоящее время наиболее часто являются: общее заболевание; болезни системы кровообращения, злокачественные новообразования, болезни костно-мышечной системы, травматизм (табл. 38). На распределение причин инвалидности по частоте встречаемости существенно влияет система социального обеспечения, общая социально-экономическая ситуация и правовая база, существующие в стране. Эти обстоятельства делают практически невозможным использование официальной статистики для объективного сравнения инвалидности в различных странах или даже в одной стране в период ее глубокого политического и экономического реформирования.

**Распределение численности лиц, впервые признанных инвалидами,
по причинам инвалидности в РФ, на 10 тыс. населения
(по данным Росстат, 2011)**

Заболевания	2000	2005	2006	2007	2008	2009
Туберкулез	3,7	3,2	3,0	2,6	2,4	2,5
Злокачественные новообразования	12,7	17,5	16,6	14,9	14,7	15,7
Болезни эндокринной системы, расстройства питания и нарушения обмена веществ	3,1	5,5	4,3	2,8	2,4	2,3
Психические расстройства и расстройства поведения	4,3	4,8	4,2	3,2	3,0	3,1
Болезни нервной системы	3,1	4,0	3,5	2,7	2,5	2,7
Болезни глаза и его придаточного аппарата	3,6	5,8	4,1	2,9	2,5	2,4
Болезни уха и сосцевидного отростка	0,5	1,0	1,2	1,0	0,9	0,9
Болезни системы кровообращения	45,7	85,9	65,6	46,2	37,2	33,3
Болезни органов дыхания	3,6	5,2	3,9	2,7	2,2	2,1
Болезни органов пищеварения	1,8	2,1	2,0	1,6	1,5	1,5
Болезни костно-мышечной системы и соединительной ткани	6,7	11,5	10,3	7,8	7,0	7,1
Последствия травм, отравлений и других воздействий внешних причин	6,4	6,7	6,0	4,9	4,6	4,4
Последствия производственных травм	1,0	0,7	0,6	0,5	0,5	0,4
Профессиональные болезни	0,4	0,3	0,3	0,2	0,2	0,2
Все болезни	3,7	3,2	3,0	2,6	2,4	2,5

В настоящее время 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 нарушения тяжесть инвалидности может быть различной – от ограничения трудоспособности в основной профессии до полной утраты трудоспособности в любом виде деятельности. С учетом этого устанавливается I, II или III группы инвалидности.

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 и 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, 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.

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

3.2.7. Статистика обязательного и добровольного медицинского страхования

Как уже отмечалось, при существующих методических подходах к сбору и статистическому анализу данных о заболеваемости с ВУТ добиться получения достоверной информации об истинной распространенности хронической патологии среди работников промышленных предприятий невозможно. Связано это с несколькими обстоятельствами. Во-первых, хронические заболевания могут иметь ремиссии более 3-лет, в этой ситуации традиционный анализ по данным обращаемости за год, анализ заболеваемости «круглогодичных» или стажированных работников становится неполным. Во-вторых, даже единичное обострение хронического заболевания или возникновение нового нередко ведет к переходу работника на новое место

труда, в другую профессиональную группу на данном предприятии или способствует увольнению работника с предприятия вообще. На существование такой тенденции указывал еще Н.А. Вигдорчик в 1928 году. Переход на другую работу или увольнение работника с хроническим заболеванием ведет к «выпадению» его как единицы статистического наблюдения. Таким образом, теряется одно из главных условий углубленного математико-статистического анализа традиционных показателей заболеваемости – независимость проявления факта хронического заболевания.

Существенным недостатком результатов традиционной статистики общей заболеваемости по данным обращаемости является возможность обращения за амбулаторно-поликлинической помощью работников одного и того же предприятия не только в медицинские учреждения, закрепленные за конкретным предприятием, но и в другие амбулаторно-поликлинические учреждения: по месту жительства, областные, краевые учреждения, медицинские учреждения негосударственной системы и т.п.

В определенной степени этих недостатков лишена статистика ЗВУТ, поскольку выплата социальных пособий по больничным листкам, осуществляемая через кассу предприятия, позволяет вести на предприятии полный учет всех случаев ЗВУТ. Однако регистрация случаев этих заболеваний происходит только в случае утраты трудоспособности. Если болезнь не сопровождается утратой трудоспособности или заболевание не является препятствием для труда по конкретной специальности (сохраняется так называемая «остаточная трудоспособность»), то регистрации случая заболевания не будет.

В свою очередь, к числу существенных недостатков статистики ЗВУТ следует отнести невозможность использовать эту статистику для прямого учета нагрузки медицинского персонала конкретного учреждения, осуществляющего медицинскую помощь работникам предприятия. Связано это с тем, что начало лечения и выдача больничного листа могут состояться не только в прикрепленном к данному предприятию медицинском учреждении.

Кроме того, показатели ЗВУТ дают возможность лишь косвенно судить о потребности работников в различных видах медицинской помощи, поскольку не учитывают всю посещаемость и обращаемость. Этих недостатков во многом лишен учет посещаемости медицинских учреждений, организованный на основе баз данных страховой медицины.

Еще одно преимущество статистики медицинской помощи по данным страховых компаний связано с другим обстоятельством. Как известно, методической основой, позволяющей наиболее полно решать задачи медицинской статистики, является так называемый «полицевой» учет, т.е. регистрация на конкретное лицо всех случаев заболеваний, обращений, посещений медицинских учреждений. Организация такого учета – весьма сложная задача, решение которой так и не состоялось в рамках государственной системы здравоохранения СССР, несмотря на многочисленные попытки в течение десятилетий ввести такой учет в промышленной медицине. На основе баз данных по ОМС, ведущихся в страховых компаниях, организация полицевого учета не представляет труда. По ДМС полицевой учет вообще является основой деятельности страховых компаний.

Как уже отмечалось, основной причиной, затрудняющей организацию полного учета заболеваемости по обращаемости, является широкая доступность (пусть даже и относительная) медицинской помощи, которая обеспечивается сетью разнообразных медицинских учреждений. В результате этой доступности трудоспособное население любого крупного населенного пункта может обратиться за медицинской помощью в медико-санитарную часть или цеховое отделение, прикрепленные к его предприятию, территориальную поликлинику, поликлинику или стационар по месту жительства, а также в любое специализированное учреждение в своем или другом городе. Суммарная регистрация таких обращений вне страховой медицины на практике просто невозможна.

На сегодняшний день гражданам России гарантировано получение медицинской помощи при обращении в государственные и негосударственные медицинские учреждения, имеющие лицензию на право заниматься лечебно-профилактической деятельностью. Эта гарантия обеспечивается за счет обязательного медицинского страхования (ОМС), которое может дополняться добровольным медицинским страхованием (ДМС). Условиями медицинского страхования предусматривается получение медицинской помощи, как правило, в разных учреждениях. В острых ситуациях ограничений в выборе места обращения по ОМС не существует вообще. Сплошной учет этих обращений, ведущийся страховыми компаниями в пределах города или сельского района (при определенных условиях даже если обращения были за пределами территорий проживания), создает предпосылки накопления объективной и полной информации, позволяющей достоверно судить о здоровье работающего населения, нагрузке и эффективности использования медицинского персонала медучреждений. Т.е. решается задача, решение которой было невозможно при существовавшем ранее учете заболеваемости, – задача полного учета всех обращений за медицинской помощью, т.е. задача получения данных об исчерпанной обращаемости, исключающей возможность массовой неполной регистрации случаев обращений.

В условиях существующей системы медицинского страхования в России накопление информации в единых базах данных обеспечивается двумя особенностями страховой медицины в России:

во-первых, данные об оказании медицинских услуг населению поступают из всех страховых компаний, осуществляющих страховые услуги населению, в Единые информационные центры (ЕИЦ) ОМС административных территорий. Во-вторых, обязательное медицинское страхование граждан организуется по территориально-производственному принципу. Это означает, что страхователем работающего населения выступают работодатели по договору с какой-либо страховой компанией. Таким образом, все постоянные работники предприятия, как правило, страхуются одной

страховой компанией, что значительно облегчает накопление данных об обращаемости за медицинской помощью не только в ЕИЦ ОМС, но и в отдельные, прикрепленные к предприятию страховые компании, в независимости от места проживания работников предприятия. Это обстоятельство особенно важно при изучении заболеваемости работников предприятия в условиях крупных населенных пунктов, где медицинская помощь осуществляется развитой сетью медицинских учреждений и получение медицинской помощи всегда рассредоточено по разным учреждениям.

При всех известных дефектах (запаздывание актуализации, ограниченность доступа к информации для других организаций), как уже отмечалось, система оказания и оплаты медицинских услуг через ОМС и ДМС позволяет накапливать объемные базы данных. Таким образом, создаются предпосылки решения ключевой проблемы медицинской статистики – проблемы сведения всей информации на одно лицо в одном месте.

Система ОМС и ДМС, значительно расширяя возможности достоверной статистики заболеваемости, порождает не только новые возможности, но и новые проблемы. Главной из этих проблем является то, что основной единицей учета в системе ОМС считается услуга, оказанная специалистом различного профиля в конкретном медицинском учреждении. Для медицинской статистики важен учет случаев заболеваний или обращений, а оказание услуги может быть не связано с наличием заболевания. «Медицинская услуга» - это товарная единица, которая необязательно является характеристикой заболевания или оказания медицинской помощи в связи с заболеванием, она может быть фактом оказания парамедицинской услуги (выдача справки, направления и т.п.). Поэтому очень важно в базе данных ОМС предусмотреть возможность выделения из всей совокупности оплачиваемых по страхованию услуг услуги, являющей фактом оказания медицинской помощи.

Существенным достоинством страховой медицины является возможность более реально подчитать материальные средства, затраченные на оказание медицинских услуг. Таким образом, оценка заболеваемости может

производиться в рублях не только на основе учета невыпущенной продукции и величины выплаченного пособия, но и на основе учета (пусть и неполного) конкретной стоимости медицинских услуг.

Не решенным к настоящему времени организационно-методическим вопросом является обеспечение единого, строго обоснованного набора статистических учетных признаков при регистрации случаев обращений за медицинской помощью во всех административных территориях страны. Вместе с тем, сравнительное изучение заболеваемости на отдельных предприятиях, предприятиях родственных отраслей промышленности и различных территориях через систему ОМС и ДМС требует жесткой стандартизации перечня учетных признаков, процедуры сбора и обработки информации страховыми компаниями.

Важным условием комплексного изучения заболеваемости по данным ОМС и ДМС является равная доступность для всех лиц, обследуемого контингента, медицинской помощи по одинаковым программам ДМС и ОМС.

В качестве примеров статистики ОМС и ДМС в настоящем издании рассмотрены показатели посещаемости работников одного из предприятий Санкт-Петербурга медицинских учреждений города в 2006-2009 гг. по данным ОМС и ДМС, реализованным в одной из страховых компаний, а также методика использования этих показателей для сравнительной оценки здоровья работников двух филиалов предприятия (табл. 39 – табл. 46).

Как уже отмечалось, при изучении заболеваемости по данным ОМС и ДМС с целью оценки влияния производственных факторов на здоровье работающих необходимым условием (по аналогии с данными ЗВУТ) является учет только тех посещений, которые прямо связаны с получением лечебно-профилактической помощи.

Результаты проведенного обследования свидетельствуют, что подавляющее большинство врачебных посещений по ОМС и ДМС работниками предприятия ЖКХ составляли посещения амбулаторно-

поликлинических учреждений: 78,8% – по программе ДМС и 93,4% – по программе ОМС.

Среди обследованного контингента на 1000 работающих за год, в среднем регистрировалось 696,8 случая посещения с лечебно-профилактической целью медицинских учреждений Санкт-Петербурга, а по программе ДМС 4977,8 случая посещения. Таким образом, на одно посещение по программе ДМС приходилось 5,1 врачебное посещение по программе ОМС.

Существенно различалась структура причин посещений с целью лечения при ОМС и ДМС. В структуре причин посещений медицинских учреждений по программе ОМС на первом месте стояли болезни органов пищеварения (11 класс по МКБ 10). На долю этих заболеваний пришлось в среднем 24,8% всех зарегистрированных по ОМС заболеваний, или 1048,7 случая на 1000 застрахованных по ОМС за год. Следует отметить, что подавляющее большинство этих посещений составляли посещения стоматолога. Без учета посещений к стоматологу доля заболеваний органов пищеварения снижалась почти в 5 раз (до 5,1%), или в случаях обращений с 1048,5 до 45,0 за год на 1000 застрахованных по ОМС.

В структуре посещений по ДМС болезни органов пищеварения стояли лишь на 5 месте: 63,2 случая на 1000 застрахованных по ДМС. Следует отметить, что наличие или отсутствие обращений к стоматологу не вызывало существенных изменений в структуре посещаемости по ДМС. Этот факт, по всей видимости, связан с существующей на предприятии программой медико-социальной поддержки, согласно которой каждый работник получает на стоматологическое лечение и протезирование 22500 рублей в виде безвозмездного разового пособия.

Результаты углубленного статистического анализа свидетельствуют, что рост показателей заболеваемости органов пищеварения по ОМС при учете посещений стоматолога был связан в основном с возникновением острой боли.

Случаев посещения стоматолога с острой болью по ДМС в обследованной совокупности вообще не было зарегистрировано.

На втором месте в структуре посещения ОМС стояли болезни органов дыхания (10 класс МКБ 10). На их долю пришлось 16,4% всех посещений, или 693,4 случая на 1000 застрахованных по ОМС за год. По ДМС этот класс заболеваний стоял на 6 месте: 6,4% от всех заболеваний, или 44,4 случая на 1000 застрахованных по ДМС.

На третьем месте по ОМС – болезни системы кровообращения (9 класс МКБ 10). Их доля составила 11,0% от всех обращений по ОМС, или 464,6 случая на 1000 застрахованных по ОМС за год. В структуре ДМС этот класс заболеваний стоял на том же, третьем, месте: 10,0% от всех зарегистрированных при посещении заболеваниях, или 96,2 случая на 1000 застрахованных по ДМС.

На четвертом и пятом местах в посещениях ОМС – болезни костно-мышечной системы и соединительной ткани (13 класс МКБ 10) и травмы, отравления и т.п. (19 класс МКБ 10), соответственно, 10,2% и 9,2%, или в случаях обращений 431,3 и 390,3 случая за год на 1000 застрахованных по ОМС за год. По ДМС эти классы заболеваний занимали по частоте встречаемости диаметрально противоположные места. Болезни костно-мышечной системы и соединительной ткани (13 класс МКБ 10) среди причин посещения медицинских учреждений занимали 1 место. На их долю приходилось 21,3% всех посещений по ДМС, или 150,4 случая на 1000 застрахованных по ДМС. На долю травм и отравлений пришлось только 6,4% всех посещений, или 32,5 случая на 1000 застрахованных по ДМС.

В то же время болезни мочеполовой системы (14 класс МКБ 10) в структуре посещения по ОМС занимали предпоследнее, 12, место, а в структуре ДМС эти заболевания стояли на втором месте, составляя 93,1 случая посещения за год на 1000 застрахованных по ДМС.

Такое соотношение показателей посещаемости по ОМС и ДМС в основном сохранялось во всех группах наблюдения. Различия показателей

посещаемости по ОМС и ДМС объясняются комплексом взаимосвязанных факторов, обусловленных специфическими различиями условий страхования и реализации услуг по ОМС и ДМС: разным набором медицинских услуг, предусмотренных программами ОМС и ДМС; экономическими различиями программ медицинского страхования; организационными особенностями получения услуг и т.п.

По этим причинам углубленная групповая статистическая оценка показателей заболеваемости по обращаемости, выполнявшихся по ОМС, ДМС и ЗВУТ, путем механического сопоставления или сложения случаев тех и других посещений является некорректной. Однако результаты, полученные на основе учета тенденций интенсивности ЗВУТ, посещаемости по ОМС и ДМС, несмотря на различные уровни и структуру последних, имели статистически устойчивую, однонаправленную тенденцию превышения показателей среди обследованных работников 1-го филиала по отношению к работникам 2-го. При этом взаимосвязь тенденций распределения частот посещений по ОМС, ДМС и случаев ЗВУТ описывалась в следующих значениях коэффициента корреляции Пирсона: в первом случае, т.е. между показателями частоты посещаемости по ОМС и частотой случаев ЗВУТ, значение коэффициента составило: $r=0,35$ при $P<0,05$, во втором случае $r=0,64$ при $P<0,05$.

Однонаправленная тенденция разнородных показателей здоровья свидетельствует о неблагоприятном влиянии производственной среды на здоровье работающих 1-го филиала предприятия. В частности, выявлены следующие статистически существенные различия показателей посещаемости по ОМС и ДМС обследованных работников 1-го и 2-го филиалов.

Уровень показателей посещаемости по ОМС работников 1-го филиала составил 4869,0 случая, 2-го – 3456,6 случая на 1000 застрахованных за год. Темп прироста указанных показателей 1-го филиала по сравнению с показателями 2-го филиала составил 40,9%. По ДМС уровни показателей посещаемости по обоим филиалам были значительно ниже, соответственно,

867,4 случая и 484,9 случая за год на 1000 застрахованных по ДМС, но темп их прироста был существенно (почти в 2 раза) выше: 78,9% .

В целом общая тенденция сравнительного расхождения частот посещений по ОМС и ДМС с учетом филиалов (1-го и 2-го) предприятия соответствовала тенденциям распределений показателей ЗВУТ. Однако темп прироста показателей на 1-м филиале по сравнению со 2-м филиалом по ОМС и ДМС (особенно по ДМС) был существенно выше, чем темп прироста показателей ЗВУТ.

При рассмотрении причин различной посещаемости установлено, что общая структура распределения зарегистрированных классов заболеваний и по ОМС и по ДМС среди работников 1-го и 2-го филиалов существенно (статистически значимо) не отличалась. Но частота посещений по основным классам заболеваний имела существенные отличия, которые и обеспечили основной рост посещаемости по ОМС и ДМС работников 1-го филиала в сравнении с работниками 2-го филиала.

При анализе посещаемости по ОМС установлено, что основной прирост обращаемости среди работников 1-го филиала по сравнению со 2-м филиалом происходил за счет болезней нервной системы (06 класс МКБ 10), темп прироста – 73,1%.

Темп прироста болезней крови и кроветворных органов (03 класс МКБ 10) – 65,4%; новообразований (02 класс МКБ 10) – 57,0%; болезней системы кровообращения (09 класс МКБ 10) – 52,8%; болезней органов дыхания (10 класс МКБ 10) – 52,8%; пищеварения (11 класс МКБ 10) – 39,5% и болезней кожи и подкожной клетчатки (12 класс МКБ 10) – 39,8%.

Из перечисленного ранжированного перечня классов заболевания наибольший интерес представляет 02 класс заболеваний – новообразования. Несмотря на то, что прирост болезней нервной системы и болезней крови и кроветворных органов был выше, чем прирост новообразований, но стоимость одного процента прироста по новообразованиям была существенно больше, соответственно, 0,9 и 0,6 против 1,4 посещения по новообразованиям. Кроме

того, прогноз большинства болезней нервной системы и болезней крови (если это не онкологические заболевания) значительно более благоприятен, чем прогноз новообразований.

При рассмотрении сравнительной структуры посещаемости по ДМС установлено, что основной прирост посещаемости работников 1-го филиала по сравнению с работниками 2-го филиала происходил за счет новообразований (02 класс МКБ 10) – 179,0%. Следует отметить, что по болезням кожи и подкожной клетчатки (12 класс МКБ 10) показатель темпа прироста составлял 412%. Но это различие не имело практического значения из-за относительно малой «стоимости» – 1% прироста, который по классу «Болезни кожи и подкожной клетчатки» был ниже более чем в 2 раза, чем по классу «Онкологические заболевания».

Таблица 39

Частота зарегистрированных диагнозов при посещениях медицинских учреждений Санкт-Петербурга с учетом места работы обследованных работников одного из предприятий города по ОМС (на 1000 застрахованных)

Диагноз МКБ-10	Место работы		Темп прироста 1-го ко 2-му, в %
	1 филиал	2 филиал	
01 класс. Некоторые инфекционные и паразитарные заболевания	101,4	71,7	41,4
02 класс. Новообразования	204,4	130,2	57,0
03 класс. Болезни крови	15,6	9,4	65,4
04 класс. Болезни эндокринной системы	65,5	52,8	24,0
06 класс. Болезни нервной системы	104,5	60,4	73,1
07 класс. Болезни глаза	107,6	94,3	14,1
08 класс. Болезни уха	60,8	39,6	53,6
09 класс. Болезни системы кровообращения	550,7	360,4	52,8
10 класс. Болезни органов дыхания	815,9	545,3	49,6
11 класс. Болезни органов пищеварения	1202,8	862,3	39,5
12 класс. Болезни кожи и п.кожной клетчатки	145,1	103,8	39,8
13 класс. Болезни костно-мышечной системы	493,0	356,6	38,2
14 класс. Болезни мочеполовой системы	218,4	175,5	24,5
15 класс. Беременность, роды и послеродовый период	45,2	50,9	-11,2
19 класс. Травмы, отравления	455,5	311,3	46,3
21 класс. Факторы, влияющие на здоровье населения	282,4	232,1	21,7
Общий итог	4869,0	3456,6	40,9

Примечание: 17 класс «Врожденные аномалии и пороки развития». 18 класс «Отклонения от нормы, выявленные при исследовании» и 20 класс «Внешние причины» не включены в совокупность из-за малого числа наблюдений (5,6 и 2, случая или 0,14% от общего числа посещений)

**Структура зарегистрированных диагнозов при посещениях
медицинских учреждений Санкт-Петербурга с учетом места работы
обследованных работников по ОМС, в % к итогу**

Диагноз МКБ-10	Место работы		Оба филиала
	1 филиал	2 филиал	
01 класс. Некоторые инфекционные и паразитарные заболевания	2,1	2,1	2,1
02 класс. Новообразования	4,2	3,8	4,0
03 класс. Болезни крови	0,3	0,3	0,3
04 класс. Болезни эндокринной системы	1,3	1,5	1,4
06 класс. Болезни нервной системы	2,1	1,7	2,0
07 класс. Болезни глаза	2,2	2,7	2,4
08 класс. Болезни уха	1,2	1,1	1,2
09 класс. Болезни системы кровообращения	11,3	10,4	11,0
10 класс. Болезни органов дыхания	16,8	15,8	16,4
11 класс. Болезни органов пищеварения	24,7	24,9	24,8
12 класс. Болезни кожи и п.кожной клетчатки	3,0	3,0	3,0
13 класс. Болезни костно-мышечной системы	10,1	10,3	10,2
14 класс. Болезни мочеполовой системы	4,5	5,1	4,7
15 класс. Беременность, роды и послерод. период	0,9	1,5	1,1
19 класс. Травмы, отравления	9,4	9,0	9,2
21 класс. Факторы, влияющие на здоровье населения	5,8	6,7	6,1
Общий итог	100,0	100,0	100,0

Примечание: 17 класс «Врожденные аномалии и пороки развития», 18 класс «Отклонения от нормы, выявленные при исследовании» и 20 класс «Внешние причины» не включены в совокупность из-за малого числа наблюдений (5,6 и 2 случая, или 0,14% от общего числа посещений)

**Частота зарегистрированных диагнозов при посещениях
медицинских учреждений Санкт-Петербурга с учетом места работы
обследованных работников по ДМС, на 1000 застрахованных**

Диагноз МКБ-10	Место работы		Темп прироста 1 ко 2 филиалу, в %
	1 филиал	2 филиал	
01 класс. Некоторые инфекционные и паразитарные болезни	9,4	5,7	65,4
02 класс. Новообразования	42,1	15,1	179,1
03 класс. Болезни крови	1,6	1,9	-17,3
04 класс. Болезни эндокринной системы	79,6	43,4	83,3
06 класс. Болезни нервной системы	32,8	17,0	92,9
07 класс. Болезни глаза и его придаточного аппарата	39,0	24,5	59,0
08 класс. Болезни уха и сосцевидного отростка	25,0	17,0	47,0
09 класс. Болезни системы кровообращения	87,4	47,2	85,2
10 класс. Болезни органов дыхания	59,3	26,4	124,4
11 класс. Болезни органов пищеварения	76,4	47,2	62,1
12 класс. Болезни кожи и подкожной клетчатки	48,4	9,4	412,6
13 класс. Болезни костно-мышечной системы и соединительной ткани	179,4	115,1	55,9
14 класс. Болезни мочеполовой системы	107,6	75,5	42,6
19 класс. Травмы, отравления и некоторые другие последствия воздействия внешних причин	60,8	24,5	148,0
21 класс. Факторы, влияющие на здоровье населения	18,7	15,1	24,0
Итого	867,4	484,9	78,9

Примечание: 3 класс «Болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм», 5 класс «Психические расстройства и расстройства поведения» и 18 класс «Отклонения от нормы, выявленные при исследовании» не включены в совокупность из-за малого числа наблюдений (1, 2 и 0 случаев)

**Структура зарегистрированных диагнозов при посещениях
медицинских учреждений Санкт-Петербурга с учетом места работы
обследованных работников по ДМС, в % к итогу**

Диагноз МКБ-10 (с сокращениями)	Место работы		Оба филиала
	1 филиал	2 филиал	
01 класс. Инфекционные и паразитарные болезни	1,1	1,2	1,1
02 класс. Новообразования	4,9	3,1	4,3
03 класс. Болезни крови, кроветворных органов	0,2	0,4	0,2
04 класс. Болезни эндокринной системы	9,2	8,9	9,1
06 класс. Болезни нервной системы	3,8	3,5	3,7
07 класс. Болезни глаза	4,5	5,1	4,7
08 класс. Болезни уха и сосцевидного отростка	2,9	3,5	3,1
09 класс. Болезни системы кровообращения	10,1	9,7	10,0
10 класс. Болезни органов дыхания	6,8	5,4	6,4
11 класс. Болезни органов пищеварения	8,8	9,7	9,1
12 класс. Болезни кожи и подкожной клетчатки	5,6	1,9	4,4
13 класс. Болезни костно-мышечной системы и соединительной ткани	20,7	23,7	21,6
14 класс. Болезни мочеполовой системы	12,4	15,6	13,4
19 класс. Травмы	7,0	5,1	6,4
21 класс. Факторы, влияющие на здоровье населения	2,2	3,1	2,5
Итого	100,0	100,0	100,0

Примечание: 3 класс «Болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм», 5 класс «Психические расстройства и расстройства поведения» и 18 класс «Отклонения от нормы, выявленные при исследовании» не включены в совокупность из-за малого числа наблюдений (1, 2 и 0 случаев)

Среди других причин роста посещаемости по ОМС работников Левобережного филиала – травмы и отравления (19 класс МКБ 10) – 148,0%, болезни органов дыхания (10 класс МКБ 10) – 124,4% и болезни нервной системы (06 класс МКБ 10) – 92,2%.

Таким образом, из перечисленного ранжированного перечня классов заболеваний, несмотря на некоторые различия в ранговом распределении, наибольшая динамика показателей обращаемости по ДМС отмечается практически по тем же классам заболеваний, что и по ОМС. Особое медико-социальное значение в обоих случаях имеет существенная динамика онкологической патологии.

Как видно из представленных данных, общая тенденция превышения показателей посещаемости медицинских учреждений по программам ДМС и ОМС сохранялась и при разбиении исследуемой подгруппы не только по месту работы, но и с учетом всех трех групп наблюдения.

Таблица 43

Частота зарегистрированных диагнозов при посещениях медицинских учреждений СПб с учетом места работы и группы наблюдения обследованных работников по ОМС, на 1000 застрахованных

Диагноз МКБ-10	Группы наблюдения						Р статистика Фишера
	1		2		3		
	1 филиал	2 филиал	1 филиал	2 филиал	1 филиал	2 филиал	
01 класс. Некоторые инфекционные и паразитарные заболевания	89,3	101,4	97,7	49,5	115,9	76,5	0,312
02 класс. Новообразования	244,0	144,9	188,0	117,1	193,2	135,3	0,025
03 класс. Болезни крови	11,9	14,5	15,0	4,5	19,3	11,8	0,324
04 класс. Болезни эндокринной системы	83,3	50,7	67,7	49,5	48,3	58,8	0,400
06 класс. Болезни нервной системы	107,1	50,7	94,0	27,0	115,9	111,8	0,160
07 класс. Болезни глаза	101,2	137,7	105,3	81,1	115,9	76,5	0,734
08 класс. Болезни уха	65,5	43,5	56,4	36,0	62,8	41,2	0,001
09 класс. Болезни системы кровообращения	565,5	478,3	485,0	288,3	623,2	358,8	0,071
10 класс. Болезни органов дыхания	892,9	659,4	684,2	482,0	922,7	535,3	0,041
11 класс. Болезни органов пищеварения	1238,1	898,6	1075,2	851,4	1338,2	847,1	0,045
12 класс. Болезни кожи и п.кожной клетчатки	125,0	123,2	165,4	103,6	135,3	88,2	0,178
13 класс. Болезни костно-мышечной системы	565,5	463,8	424,8	342,3	521,7	288,2	0,099
14 класс. Болезни мочеполовой системы	166,7	231,9	191,7	144,1	294,7	170,6	0,584
15 класс. Беременность, роды и послеродовый период	101,2	65,2	26,3	54,1	24,2	35,3	0,964
19 класс. Травмы, отравления	500,0	369,6	417,3	270,3	468,6	317,6	0,002
21 класс. Факторы, влияющие на здоровье населения	291,7	195,7	259,4	229,7	299,5	258,8	0,114
Общий итог	5148,8	4029,0	4353,4	3130,6	5299,5	3411,8	0,028

**Структура диагнозов, зарегистрированных при посещениях (по ОМС)
медицинских учреждений СПб с учетом места работы
и группы наблюдения обследованных работников, в % к итогу**

Диагноз МКБ-10	Группы наблюдения					
	1		2		3	
	1 филиал	2 филиал	1 филиал	2 филиал	1 филиал	2 филиал
11 класс. Болезни органов пищеварения	24,0	22,3	24,7	27,2	25,3	24,8
10 класс. Болезни органов дыхания	17,3	16,4	15,7	15,4	17,4	15,7
09 класс. Болезни системы кровообращения	11,0	11,9	11,1	9,2	11,8	10,5
13 класс. Болезни костно-мышечной системы	11,0	11,5	9,8	10,9	9,8	8,4
19 класс. Травмы, отравления	9,7	9,2	9,6	8,6	8,8	9,3
21 класс. Факторы, влияющие на здоровье	5,7	4,9	6,0	7,3	5,7	7,6
14 класс. Болезни мочеполовой системы	3,2	5,8	4,4	4,6	5,6	5,0
02 класс. Новообразования	4,7	3,6	4,3	3,7	3,6	4,0
12 класс. Болезни кожи и п. кожной клетчатки	2,4	3,1	3,8	3,3	2,6	2,6
07 класс. Болезни глаза	2,0	3,4	2,4	2,6	2,2	2,2
01 класс. Некоторые инфекционные и паразитарные заболевания	1,7	2,5	2,2	1,6	2,2	2,2
06 класс. Болезни нервной системы	2,1	1,3	2,2	0,9	2,2	3,3
04 класс. Болезни эндокринной системы	1,6	1,3	1,6	1,6	0,9	1,7
08 класс. Болезни уха	1,3	1,1	1,3	1,2	1,2	1,2
15 класс. Беременность, роды и послеродовый период	2,0	1,6	0,6	1,7	0,5	1,0
03 класс. Болезни крови	0,2	0,4	0,3	0,1	0,4	0,3
Общий итог	100,0	100,0	100,0	100,0	100,0	100,0

**Частота зарегистрированных диагнозов при посещениях
медицинских учреждений СПб с учетом места работы
и группы наблюдения обследованных работников по ДМС,
на 1000 застрахованных**

Диагноз МКБ-10	Группы наблюдения					
	1		2		3	
	1 филиал	2 филиал	1 филиал	2 филиал	1 филиал	2 филиал
01 класс. Некоторые инфекционные и паразитарные болезни	6,0	0,0	0,0	9,0	24,2	5,9
02 класс. Новообразования	71,4	43,5	56,4	9,0	0,0	0,0
03 класс. Болезни крови и кроветворных органов	0,0	7,2	0,0	0,0	4,8	0,0
04 класс. Болезни эндокринной системы	107,1	50,7	109,0	49,5	19,3	29,4
06 класс. Болезни нервной системы	41,7	29,0	41,4	18,0	14,5	5,9
07 класс. Болезни глаза	17,9	36,2	48,9	27,0	43,5	11,8
08 класс. Болезни уха и сосцевидного отростка	29,8	29,0	18,8	9,0	29,0	17,6
09 класс. Болезни системы кровообращения	53,6	29,0	94,0	67,6	106,3	35,3
10 класс. Болезни органов дыхания	35,7	29,0	56,4	36,0	82,1	11,8
11 класс. Болезни органов пищеварения	59,5	50,7	52,6	49,5	120,8	41,2
12 класс. Болезни кожи и подкожной клетчатки	41,7	29,0	45,1	0,0	58,0	5,9
13 класс. Болезни костно-мышечной системы	190,5	144,9	161,7	72,1	193,2	147,1
14 класс. Болезни мочеполовой системы	256,0	181,2	90,2	49,5	9,7	23,5
19 класс. Травмы, отравления	53,6	21,7	71,4	27,0	53,1	23,5
21 класс. Факторы, влияющие на здоровье населения	11,9	50,7	37,6	4,5	0,0	0,0
Общий итог	976,2	731,9	883,5	427,9	758,5	358,8

**Структура зарегистрированных диагнозов при посещениях
медицинских учреждений СПб с учетом места работы
и группы наблюдения обследованных работников по ДМС, в % к итогу**

Диагноз МКБ-10	Группы наблюдения					
	1		2		3	
	1 филиал	2 филиал	1 филиал	2 филиал	1 филиал	2 филиал
13 класс. Болезни костно-мышечной системы	19,5	19,8	18,3	16,8	25,5	41,0
14 класс. Болезни мочеполовой системы	26,2	24,8	10,2	11,6	1,3	6,6
09 класс. Болезни системы кровообращения	5,5	4,0	10,6	15,8	14,0	9,8
04 класс. Болезни эндокринной системы	11,0	6,9	12,3	11,6	2,5	8,2
11 класс. Болезни органов пищеварения	6,1	6,9	6,0	11,6	15,9	11,5
19 класс. Травмы, отравления	5,5	3,0	8,1	6,3	7,0	6,6
10 класс. Болезни органов дыхания	3,7	4,0	6,4	8,4	10,8	3,3
07 класс. Болезни глаза	1,8	5,0	5,5	6,3	5,7	3,3
12 класс. Болезни кожи и подкожной клетчатки	4,3	4,0	5,1	0,0	7,6	1,6
02 класс. Новообразования	7,3	5,9	6,4	2,1	0,0	0,0
06 класс. Болезни нервной системы	4,3	4,0	4,7	4,2	1,9	1,6
08 класс. Болезни уха	3,0	4,0	2,1	2,1	3,8	4,9
21 класс. Факторы, влияющие на здоровье населения	1,2	6,9	4,3	1,1	0,0	0,0
01 класс. Некоторые инфекционные и паразитарные болезни	0,6	0,0	0,0	2,1	3,2	1,6
03 класс. Болезни крови, кроветворных органов	0,0	1,0	0,0	0,0	0,6	0,0
Общий итог	100,0	100,0	100,0	100,0	100,0	100,0

Углубленный анализ распределения показателей посещаемости по классам заболеваний по МКБ 10, проведенный с учетом разбиения изучавшейся совокупности на группы наблюдения, подтвердил общую тенденцию превышения показателей среди работников 1-го филиала по сравнению с работниками 2-го. Из 96 подгрупп, сформированных из числа обследованных работников, обращавшихся по ОМС с учетом их заболеваемости, места работы и профессиональных групп, только в нескольких

подгруппах выявлена тенденция изменения показателей, противоположная общей тенденции. При этом ни в одном из этих случаев не подтверждена статистическая значимость этих единичных случаев противоречивой тенденции.

Аналогичная ситуация выявлена и при разбиении обследованной совокупности на подобные 96 подгрупп групп при анализе посещаемости по ДМС.

Проведенная детальная группировка позволила выделить классы заболеваний по МКБ 10 с большей статистической устойчивостью, т.е. способностью сохранять статистически значимую тенденцию даже при относительно малом числе наблюдений и при определенном качественном изменении подгрупп. На основе двухфакторного дисперсионного анализа были выделены следующие классы заболеваний, зарегистрированные при посещаемости по ОМС, F статистика Фишера которых неизменно подтверждала тенденцию превышения показателей 1-го филиала по сравнению со 2-м: 02 класс – новообразования, 08 класс – болезни уха, 10 класс – болезни органов дыхания, 11 класс – болезни органов пищеварения, 19 класс – травмы и отравления.

По ДМС из-за малой численности наблюдений такой углубленный анализ не проводился.

3.2.8. Наиболее важные для общественного здоровья болезни

Особый раздел медицинской статистики составляют показатели заболеваемости наиболее важными для общественного здоровья болезнями. К ним обычно относят показатели заболеваемости по данным о причинах смерти, заболеваемости важнейшими социальнозначимыми болезнями, инфекционной (эпидемической) заболеваемости, заболеваемости с временной утратой трудоспособности, заболеваемости со стойкой утратой трудоспособности (инвалидности) и др.

Заболеваемость важнейшими социальнозначимыми болезнями.

В настоящее время номенклатура социально значимых заболеваний утверждена Постановлением Правительства РФ от 1 декабря 2004 г. № 715 «Об утверждении перечня социально значимых заболеваний и перечня заболеваний, представляющих опасность для окружающих».

Согласно этому постановлению, к социально значимым заболеваниям относятся: туберкулез, инфекции, передающиеся преимущественно половым путем, гепатиты В и С, болезнь, вызванная вирусом иммунодефицита человека (ВИЧ), злокачественные новообразования, сахарный диабет, психические расстройства и расстройства поведения, болезни, характеризующиеся повышенным кровяным давлением.

К заболеваниям, представляющим опасность для окружающих, относятся: болезнь, вызванная вирусом иммунодефицита человека (ВИЧ), вирусные лихорадки, передаваемые членистоногими, и вирусные геморрагические лихорадки, гельминтозы, гепатиты В и С, дифтерия, инфекции, передающиеся преимущественно половым путем, лепра, малярия, педикулез, акариаз и другие инфекции, сепсис и мелиоидоз, сибирская язва, туберкулез, холера и чума.

Иногда в этой группе выделяют *важнейшие неинфекционные заболевания*. Сюда включаются сердечно-сосудистые заболевания и болезни системы кровообращения, злокачественные новообразования, психические расстройства и т.п. заболевания, имеющие большую медико-социальную значимость.

Туберкулез. Одной из причин, по которым туберкулез отнесен в группу важнейших заболеваний, является то, 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 туберкулёзом. Несмотря на успехи, достигнутые в России в борьбе с этим недугом к 70 годам, в последнее десятилетие XX века заболеваемость туберкулёзом опять начала расти (табл. 47). К 2011 году этот показатель снизился до 104,3 случая на 100 тыс. жителей РФ.

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 тысячи случаев туберкулеза были диагностированы после смерти больных).

Динамика заболеваемости населения РФ важнейшими социальнозначимыми болезнями в период с 2000 по 2011 гг. (по данным РОССТАТ, 2012 г.)

	2000	2005	2006	2007	2008	2009	2010	2011
Зарегистрировано больных с диагнозом, установленным впервые в жизни, тыс. человек:								
активный туберкулез	130,7	119,2	117,6	118,4	120,8	117,2	109,9	104,3
сахарный диабет	162,0	249,5	289,0	302,4	301,6	308,6	324,0	319,0
болезни, характеризующиеся повышенным кровяным давлением	434,8	772,6	1007,3	990,3	979,0	920,3	870,7	848,3
инфекции, передающиеся преимущественно половым путем:								
сифилис	239,4	97,9	93,1	89,6	85,0	75,7	63,7	53,8
гонококковая инфекция	175,9	101,8	91,0	86,4	80,1	68,2	60,6	54,5
трихомоноз	462,9	305,9	284,8	264,8	237,9	205,3	180,0	159,2
Взято на учет больных с впервые в жизни установленным диагнозом, тыс. человек:								
злокачественные новообразования	427,5	442,9	448,3	455,5	458,1	469,8	479,5	481,6
Взято под наблюдение больных с диагнозом, установленным впервые в жизни, тыс. человек:								
психические расстройства и расстройства поведения	121,0	95,9	92,3	84,8	78,5	77,5	74,3	70,3
Зарегистрировано случаев заболевания:								
острый вирусный гепатит В	62,0	12,4	10,1	7,5	5,7	3,8	3,2	2,4
острый вирусный гепатит С	30,8	6,4	5,9	5,1	4,0	3,2	3,0	2,6
Болезнь, вызванная вирусом иммунодефицита человека (ВИЧ), тыс. человек:								
зарегистрировано больных всего	78,6	234,8	237,2	267,5	301,3	332,9	372,9	422,3
зарегистрировано больных с диагнозом, установленным впервые в жизни	55,4	32,7	28,5	38,8	44,1	50,7	57,2	59,6

Зарегистрировано больных с диагнозом, установленным впервые в жизни, на 100 тыс. жителей РФ:	89,8	83,7	82,2	82,9	84,7	82,1	76,9	73,0
	111,3	175,3	202,0	211,8	211,3	216,2	226,8	223,1
	298,7	542,6	704,2	693,5	685,9	644,5	609,5	593,4
активный туберкулез								
	164,5	68,8	65,1	62,8	59,6	53,0	44,6	37,6
	120,9	71,5	63,6	60,5	56,1	47,8	42,4	38,1
сахарный диабет	318,1	214,8	199,1	185,4	166,6	143,8	125,9	111,4
болезни, характеризующиеся повышенным кровяным давлением								
инфекции, передающиеся преимущественно половым путем:								
сифилис								
гонорея								
трихомониаз								
Взято на учет больных с впервые в жизни установленным диагнозом, на 100 тыс. жителей РФ:	293,7	311,1	313,4	319,0	320,9	329,1	335,7	336,9
злокачественные новообразования								
Взято под наблюдение больных с диагнозом, установленным впервые в жизни, на 100 тыс. жителей РФ:								
психические расстройства и расстройства поведения	83,1	67,3	64,5	59,4	55,0	54,3	52,0	49,1
Зарегистрировано случаев заболевания на 100 тыс. жителей РФ:								
острый вирусный гепатит В	42,3	8,7	7,0	5,3	4,0	2,7	2,2	1,7
	21,0	4,5	4,1	3,6	2,8	2,2	2,1	1,8
острый вирусный гепатит С								
Болезнь, вызванная вирусом иммунодефицита человека (ВИЧ)								
Зарегистрировано больных всего на 100 тыс. жителей РФ:	54,0	164,9	165,8	187,3	211,0	233,2	261,0	295,4
зарегистрировано больных с диагнозом, установленным впервые в жизни, на 100 тыс. жителей РФ:	38,1	23,0	19,9	27,1	30,9	35,5	40,1	41,7

К концу XX века в мире ежегодно погибало от *сердечно-сосудистых заболеваний и болезней системы кровообращения* 10 миллионов человек. В России в 2000 году на 100 тыс. человек населения регистрировалось 13802,4 больных с этой патологией, в том числе с впервые в жизни выявленными заболеваниями – 1705,8 на 100 тыс. населения. К 2006 году эти показатели значительно возросли, достигнув, соответственно, 20923,0 и 2657,5 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 смертности. Если в России умерли от этих болезней в 1999 году 915,5 тыс. человек, то в 2006 году – 1232,2 тыс. человек.

По оценкам Всемирной организации здравоохранения, *нервно-психические заболевания* к 2000 году по распространенности заняли второе место после сердечно-сосудистой патологии. В этом году в мире нервно-психические расстройства имели более 400 миллионов человек. В России за 90 годы распространенность этих заболеваний увеличилась в 1,5 раза, а среди детей и подростков – в 2,5 раза. К концу XX столетия в России имели психические расстройства и расстройства поведения около 10 миллионов человек, то есть каждый 15 россиянин имел нарушения психики. За период с 2000 по 2006 годы число зарегистрированных случаев психических расстройств и расстройств поведения, в том числе и впервые выявленных в жизни, сократилось в 1,5-2 раза. Вместе с тем, к 2006 году в Российской Федерации из каждых 100 тысяч жителей свыше 39,4 человека страдали тяжелыми формами психозов и слабоумия, из них шизофренией – 10,8 на 100 тыс. жителей, 22,1 – умственной отсталостью. Обращает на себя внимание резкий (почти двухкратный) рост обращаемости за консультативно-лечебной помощью по поводу этих заболеваний с 159,2 больных на 100 тыс. населения в 2000 году до 311,7 на 100 тыс. населения в 2006 году.

Другой отнесенной в группу важнейших является *онкологическая заболеваемость*. О медико-социальной значимости этих болезней говорит тот факт, что по этой причине в течение последнего десятилетия ежегодно умирало 286,2-295,3 тысячи человек.

В России ежегодно регистрируется более 400000 злокачественных новообразований. За последнее десятилетие произошел рост злокачественной заболеваемости населения России. Однако основной причиной этого роста были заболевания мочеполовой системы, почек, лимфатической и кроветворной ткани. По остальным видам онкологической патологии наблюдалось снижение показателей распространенности.

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 туберкулёза, венерической болезни, трихофитии, микроспории, фавуса, трахомы, рака и других злокачественных новообразований» (учетная форма № 089 Т/У-2003 (ранее форма №281-У).

Извещение посылается в трехдневный срок в районный (городской) кожно-венерологический кабинет (диспансер). Копия извещения посылается в органы/учреждения госсанэпиднадзора по месту жительства больного, а для лиц, не имеющих определенного места жительства, – по месту выявления заболевания. При выявлении больных с бациллярной формой туберкулёза составляют дополнительное извещение (учетная форма № 58), которое в течение 24 часов посылают в органы/учреждения 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 независимо 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 погибало 16,6 миллиона человек. В России в 2006 году было зарегистрировано 7,6 млн. случаев инфекционных заболеваний (по I классу МКБ 10 «Некоторые инфекционные и паразитарные болезни»). Известно, что уровень инфекционной заболеваемости отражает социально-экономическое благополучие в стране. Российская Федерация не была исключением из этого правила: период начала экономических реформ в России сопровождался существенным ростом инфекционной заболеваемости. В 1995 году по сравнению с 1990 годом уровень инфекционной заболеваемости вырос на 22,5%, в том числе по впервые в жизни выявленным заболеваниям – на 34,8%. Соответственно, с 5143,2 до 6298,8 случая на 100 тыс. населения и с 3490,9 до 4706,4 случая на 100 тыс. населения.

По мере стабилизации социально-экономической ситуации в стране с 2000 года уровень инфекционной заболеваемости начал заметно снижаться.

3.3. Медицинская демография

Демография – наука о народонаселении. Назвать одного, конкретного основателя этой науки трудно. К числу ученых, сделавших многое для ее становления и определения круга проблем, которыми она занимается, относят Х. Бернулли (1772-1863), который предложил назвать новую для того времени науку «популяционистикой». Использовать термин «демография»¹⁹ предложил французский ученый А. Гийяр (1799-1876). Среди отечественных ученых впервые озвучил этот термин Е.Н. Анучин (1831-1905). До настоящего времени 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*).

¹⁹ «Демография» от греч. «*demos*» – народ и «*grapho*» – писать, изображать.

Среди ученых, оставивших заметный след в становлении демографии как науки, можно отметить Д. Граунта (1620-1674), опубликовавшего в 1662 году первую статистическую таблицу смертности (дожития – *life table*), служащую методической основой решения одной из главных задач демографических исследований – задачи изучения динамики народонаселения. В целом становление демографии как полноправной, самостоятельной науки, имеющей свой предмет и объект исследования, специфическую терминологию, свои специфические методы, начало обозначаться в конце XIX – начале XX веков. До этого обычным было восприятие демографии как раздела научной статистики – статистики народонаселения. Такое понимание долгое время определяло и направление ее развития. «Только к середине XX века демографы пришли к убеждению, что предметом демографии является воспроизводство населения, т.е. процесс непрерывного возобновления его численности и структур через смену поколений, через процессы рождаемости и смертности» (В.М. Медков, 2005).

Сегодня считается общепризнанным, что демография – самостоятельная наука, имеющая своим объектом исследования процессы воспроизводства народонаселения. В свою очередь, результаты научного познания основных законов и закономерностей этих процессов служат опорой развития ряда естественных и гуманитарных наук, в том числе и медицины.

Воспроизводство народонаселения – процесс постоянного замещения одних поколений людей другими, основным элементом которого является взаимодействие двух явлений: рождаемости и смертности. Их интенсивность обусловлена комплексом неразрывно связанных факторов, среди которых сегодня наиболее значимыми являются социально-экономические и этнокультурные. В различные периоды человеческой истории, включая и настоящее время, процесс воспроизводства населения и его результаты имели свои особенности, которые в странах с различным уровнем социально-

экономического развития и разным этносом имели (и имеют ныне) свои уникальные черты. Статистически интенсивность воспроизводства населения отражается в брутто- и нетто-коэффициентах воспроизводства женского населения. Брутто-коэффициент воспроизводства свидетельствует о том, сколько в среднем у одной женщины в возрасте от 15 до 49 лет родится девочек, иначе говоря, каково замещение матерей дочерью. Однако не все родившиеся девочки достигают возраста начала деторождения. Поэтому в брутто-коэффициент вносится поправка, которая определяет, сколько девочек не только родится в расчете на одну женщину репродуктивного возраста, но и доживет до среднего возраста начала рождений в своем поколении. Так получают нетто-коэффициент. В России до середины 60-х годов нетто-коэффициент, или чистый коэффициент воспроизводства населения был выше единицы (расширенное воспроизводство). Затем, вплоть до настоящего времени, этот показатель ниже единицы.

К настоящему времени опубликовано множество научных работ, посвященных проблемам демографии, методикам сбора, статистической обработки и методологии анализа демографической информации. При этом практическая медицинская статистика остается тесно связанной с демографической статистикой. Отсюда вытекают два обстоятельства, делающие необходимость детального знакомства с ее основами практиков медицинской статистики. Во-первых, должно обязательно присутствовать правильное понимание демографии, поскольку внешняя схожесть методологии и терминов медицинской статистики с аналогичными компонентами демографической статистики может порождать существенные ошибки в ходе обработки тех или иных данных. Например, при определении существенности рисков повышенной заболеваемости или смертности нередко используется показатель изменения средней продолжительности жизни. Но в демографии «средняя продолжительность жизни» – это не средняя величина, используемая обычной

статистикой, а совершенно другая производная величина, более полное название которой «ожидаемая средняя продолжительность жизни». Величина, которая рассчитывается для каждого поколения отдельно. И ее расчет – трудоемкая процедура, основанная на математико-статистическом моделировании процесса вымирания отдельных поколений. Отсюда вытекает недопустимость восприятия этого показателя как «обыкновенной» средней, о чем энтузиасты статистики рисков постоянно забывают, путая его с другим демографическим показателем – средним возрастом умерших.

Вторым важным обстоятельством является необходимость понимания границ информативности тех или иных демографических показателей. Природу возникновения и накопления в них систематических ошибок, которые ограничивают достаточность, информативность тех или иных данных. Например, уже упоминавшийся выше среднестатистический показатель смертности рассчитывается как отношение суммы умерших в течение анализируемого периода, к средней численности населения за этот же период. Т.е. если речь идет о годовых показателях, то это отношение числа умерших за год к среднегодовой численности населения. Но при такой методике вычислений не учитывается тот факт, что в течение года численность населения сокращается в результате вымирания, которое как демографический процесс, увы, непрерывно. Соответственно, при постоянстве процесса вымирания число умерших в начале и в конце года будет разным. Поэтому для построения математических моделей вместо обычного показателя смертности используют его производную величину, часто называемую «табличным показателем смертности». Несмотря на то, что в определенных ситуациях «обычный», среднестатистический и «табличный» показатели смертности могут быть близкими, их значения, как правило, различны.

Медицинская демография как источник информации о статистике общественного здоровья оперирует специальными демографическими

показателями. Эти показатели общественного здоровья неразрывно связаны с общей демографической ситуацией. Кроме того, многие общие демографические показатели (*demographic information* – англ.), не являясь напрямую показателями здоровья, служат базой для статистики здоровья населения и широко используются в организации здравоохранения. Например, без знания территориального размещения, возрастного-полового и социального состава населения невозможно практическое получение многих статистических коэффициентов заболеваемости, и в первую очередь интенсивных показателей, которые рассчитываются относительно численности населения, проживающего на определенной территории (на 1 тыс., 10 тыс., 100 тыс. жителей различного возраста, пола, социального положения и т.п.). Отсутствие такого рода демографических показателей послужило причиной того, что в трудах статистиков дореволюционной России, посвященных заболеваемости, почти все данные о распространенности заболеваний 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, т.е. причин, обуславливающих распространение того или иного заболевания, преобладание той или иной патологии. Это обусловлено тем, что рост частоты регистрируемых заболеваний может быть следствием не ухудшившейся экологической ситуации или падения уровня жизни, а следствием старения населения. Либо, наоборот, низкий уровень заболеваемости может быть обусловлен только 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 новообразования и т.п.).

Демографические характеристики определенных контингентов населения являются основой планирования медицинской помощи. Практически все нормативы по обеспечению населения различными 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 акушеров-гинекологов на 1000 женщин и т.п.).

Для получения демографической информации используются различные источники. Все их можно разделить на 2 основные группы:

1. Данные выборочных обследований, которые проводятся при научных исследованиях и реализации специальных программ углубленного обследования населения в интересах отдельных территорий или регионов.

2. Данные сплошных наблюдений, проводимых в национальных, общегосударственных масштабах.

Основные источники и характеристики сплошных наблюдений, в соответствии с рекомендациями ООН представлены в табл. 48.

Таблица 48

Краткое сравнение источников демографических данных при сплошных исследованиях (ООН, Нью-Йорк, 2005)

Элемент	Перепись населения	Регистрация актов гражданского состояния	Регистр населения
Цель	Регистрация каждого домохозяйства и человека в стране в некий момент времени	Регистрация на непрерывной основе каждого случая рождения и смерти, имеющего место в стране	Ведение непрерывно обновляемой записи по каждому человеку
Метод	Сбор информации от каждого домохозяйства в стране в конкретный момент времени	Создание и поддержание общенациональной сети местных бюро регистрации актов гражданского состояния, укомплектованных местными регистраторами, которые получают и регистрируют информацию о событиях естественного движения населения	Создание и постоянное обновление записей по каждому человеку на основе регистрации рождений, случаев смерти, иммиграции, эмиграции, внутренних перемещений и изменений в личных характеристиках
Персонал	Счетчики и инспекторы, привлеченные в достаточном количестве для проведения опроса по всей стране на период подготовки и работы по сбору данных на местах	Местные регистраторы, постоянно работающие в местных бюро регистрации актов гражданского состояния по всей стране	Персонал, способный эксплуатировать и поддерживать системы регистрации рождений, случаев смерти, иммиграции, эмиграции, изменения места жительства и изменений в личных характеристиках
Сравнительный уровень требуемых ресурсов	Требуется много сил и средств из-за крупных масштабов операций	Требуется больше сил и средств, чем перепись населения, из-за необходимости содержания по всей стране постоянно действующих учреждений, укомплектованных постоянным персоналом, требуется много сил и средств из-за крупных масштабов операций	Требуется больше сил и средств, чем регистрация актов гражданского состояния, из-за необходимости регистрировать случаи внутренней и международной миграции и изменения личных характеристик
Сроки проведения сбора данных	Как правило, один раз в 5 лет	Ведется непрерывно	Ведется непрерывно

Сроки представления отчетов	Как правило, 1-3 года для получения полных результатов переписи	Как правило, представляются ежегодные отчеты в течение одного года по окончании отчетного года	Недостаточно опыта
Временной диапазон собираемых данных	Данные за 15 или больше лет, предшествующих базисной дате (данные о смертности взрослого населения только за 1-2 года)	Гибкий (с учетом ограничений, связанных с просроченной регистрацией)	Гибкий
Географический охват	Все географические районы	Все географические районы с учетом неполноты регистрации	Все географические районы с учетом точной регистрации всех случаев миграции
Другие детали информации	Основная информация	Основная информация	Основная информация
Гибкость	Менее гибкие, чем специальные (научные) обследования; масштабами переписи обусловлено малое число и относительная простота вопросов	Менее гибкий метод, чем переписи; трудно изменить элементы регистрационных карточек	Менее гибкий метод, чем регистрация актов гражданского состояния; трудно изменять системы, которые собирают данные об изменениях в личных характеристиках

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, Месопотамии, Китае и Японии. Они проводились, как правило, с фискальными (налоговыми) целями и в целях пополнения армии, поэтому нередко охватывали только мужское население. То, что переписи играли важную роль в истории человечества, свидетельствует и Библия. Классическим примером организации переписи может служить Книга Чисел – глава Ветхого Завета Библии *«...И сказал Господь Моисею в пустыне Синайской: «Исчислите все общество сынов Израилевых по родам и по семействам их, по числу имен, всех мужского пола поголовно от двадцати лет и выше, всех годных для войны у Израиля».* Описанное в Новом Завете появление Иисуса на свет Божий связано с переписью населения, организованной царем Иродом.

В России, как это ни покажется странным патриотически настроенным гражданам, более-менее регулярный статистический учет народонаселения был впервые организован монгольскими ханами во времена татаро-монгольского ига. Целью этого учета было обложение данью населения. В качестве единицы счета использовались дома, или «дымы».

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, независимо 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 др., 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, допускаявших 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 воскрешенного 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 были проведены 8 всеобщих переписей: в 1920, 1926, 1937, 1939, 1959, 1970, 1979, 1989 и 2002 годах. После 1917 года переписи на ограниченных территориях проводились в августе 1920 г. (на территориях, не охваченных гражданской войной), в марте 1923 года (перепись городского населения). Всеобщие переписи – в декабре 1926 года, в январе 1937 года (результаты переписи признаны несостоятельными, повторная перепись – в январе 1939 года). Послевоенные – в январе 1959 года, 1970 года, 1979 года и в 1989 году. После распада СССР перепись населения РФ планировалась на 1999 год, но была отложена из-за финансовых проблем и проводилась 9 октября 2002 года. Очередная, последняя, перепись населения РФ проводилась в 2010 году.

Перепись населения в современных условиях представляет собой важную статистическую операцию общегосударственного масштаба. 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-ти 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, в 1998 году в России ежедневно рождалось за один час в среднем 147 человек, а умирало – 227. Критический момент переписи, как правило, назначался на зимнее время, в середине недели, когда миграция населения наименьшая, число отпусков незначительное, а школьные каникулы уже кончились. Например: критический момент переписи 1970 года был в 0 часов в ночь на 15 января, переписи 1989 г. – в ночь на 17 января, переписи 2002 года – в ночь на 6 октября 2002 года, переписи 2010 года – в ночь на 14 октября 2010 года.

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 переписи 1989 года, 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 для получения статистических итоговых данных. Это требование, называемое «статистической тайной», является одним из главных условий получения достоверной информации. В СССР до конца 80-х годов 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 называемые межпереписные периоды, оценка численности населения и некоторых других демографических показателей 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 учетных форм прибытия или убытия, которые заполняются в органах власти, занимающихся регистрацией приезжающих или убывающих. На практике, в условиях 2008 и 2012 гг., этот способ нередко оказывался малонадежным из-за ряда изменений законодательства РФ и недостатков в системе регистрации граждан по месту проживания.

Другой способ получения данных о составе населения в межпереписной период – специальные статистические вычисления. *Интерполяция* –

используется для получения данных в промежутке между двумя уже прошедшими переписями. *Экстраполяция (прогноз)* – получение на основании результатов последней переписи данных на последующие годы.

Во время Великой Отечественной войны в СССР для получения оперативной информации о массовых перемещениях 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 населения на местах. Всего за годы войны было проведено 100 таких переписей.

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

3.3.1. Показатели статистики

Показатели статистики народонаселения – отражают 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.п.

По данным ООН, численность населения Земли в ноябре 1999 года достигла 6 млрд. человек и продолжает расти. Считается, что численность жителей Земли достигла первого миллиарда в 1804 году, второго миллиарда – 123 года спустя, в 1927 году. До 1970 года абсолютное число жителей Земли росло по гиперболическому закону, затем этот темп роста несколько снизился. Чтобы численность населения увеличилась с 5 млрд. в 1987 году до 6 млрд. в 1999 году, потребовалось всего 12 лет, а для того чтобы она возросла до 7 млрд. чел. – всего два года (2011 год). По прогнозу ООН, дальнейший рост населения Земли будет немного замедляться: 8 млрд. – 2028 год, 9 млрд. – 2054 год.

Если в 1950 году государств с населением более 100 млн. человек было 4 (Китай, Индия, США, СССР), то в 2012 году количество таких стран достигло

12-ти. Из них самые населенные – Китай (1354,8 млн.чел) и Индия (1227,2 млн. человек). В США население к концу 2012 года составляло 314,7 млн. чел., в Индонезии – 237,6 млн. чел. Россия по численности населения занимала 9-е место (143,4 млн. чел.).

За последние 12 лет число россиян заметно сократилось (рис. 36).

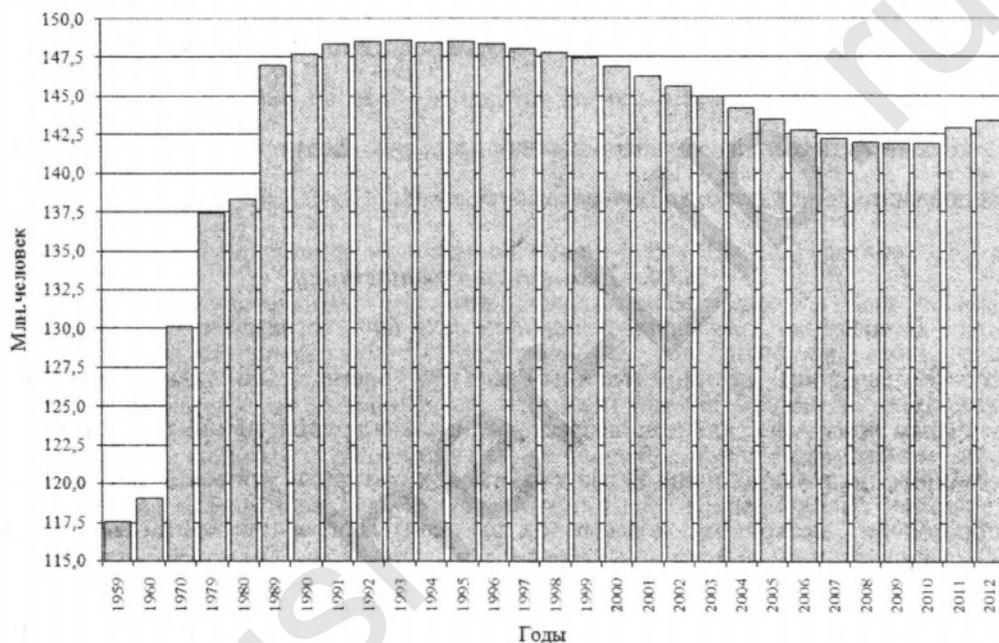


Рис. 36. Изменение численности населения России в последнем десятилетии XX века (млн. чел.)

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 особенно заметно (на 25-30%). Только население США в ближайшие 50 лет будет расти за счет миграции и, согласно упомянутым прогнозным оценкам, увеличится почти на четверть. Сокращение численности и старение населения являются одними 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 лиц пожилого возраста, рост отчислений из зарплаток в пользу социального обеспечения пожилых и т.п.

Возрастно-половой состав населения определяет распространенность и структуру патологии, уровень рождаемости, смертности населения. Уровень обращаемости за медицинской помощью населения различных демографических групп также имеет существенные отличия. *Например*, в РФ 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%.

Статистическая характеристика возрастной структуры населения является не только важнейшей медико-демографической, но и социально-экономической характеристикой. В РФ возрастная структура качественно оценивается соотношением 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.

Существуют и другие варианты возрастных группировок и оценок их соотношения. По классификации, рекомендованной ООН, «демографически

старым» считается население, когда в его составе доля лиц в возрасте 65 лет и старше превышает 12% от общей численности населения.

В 1894 г. шведский демограф А.Г. Сундберг предложил оценивать тип возрастной структуры населения в зависимости от соотношения следующих значений доли детей (0-15 лет) и стариков (60 лет и старше):

- Прогрессивный – детей 40%, стариков 10%,
- Стационарный – детей 27%, стариков 23%,
- Регрессивный – детей 20%, стариков 30%.

По оценкам ООН, 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%.

В РФ при возрастной группировке обычно учитывается социальный статус людей (пенсионер или непенсионер), поэтому в группу «старики» относят мужчин 60 лет и старше, а в группу женщин 55 лет и старше. Чтобы избавиться от влияния неоднозначных границ этих групп, варианты возрастного распределения населения делают отдельно для женщин и отдельно для мужчин.

В 1930-е годы немецкий демограф Ф. Бургдёрфер предложил для повышения наглядности использовать графические изображения возрастных пирамид: при прогрессивном типе структуры населения пирамида имеет правильную форму, в регрессивном – форму урны (рис. 37).

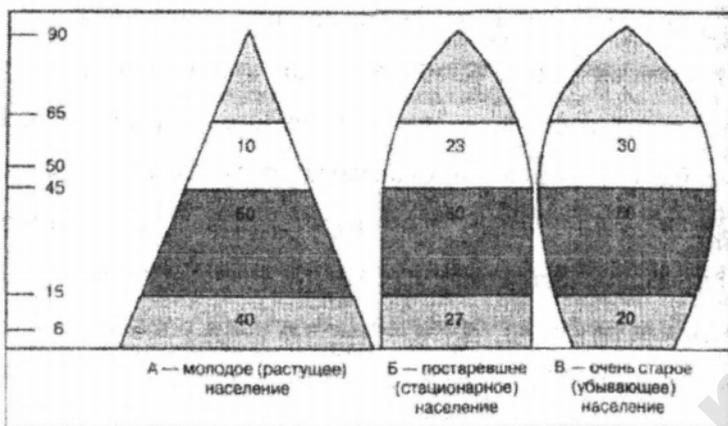


Рис. 37. Типы возрастной структуры по Ф. Бурдёрферу

Существенным недостатком такого рода характеристик является то, что при высокой смертности населения в старших возрастных группах демографическая ситуация считается «прогрессивной», а в обратной ситуации, т.е. при увеличении средней продолжительности предстоящей жизни – «регрессивной». Примером «более осторожной» и детальной классификации демографического старения является шкала Ж. Боже-Гарнье – Э. Россета (табл. 49).

Таблица 49

Шкала демографического старения Ж. Боже-Гарнье – Э. Россета

Этапы	Доля лиц в возрасте 60 лет и старше, %	Этап старения и уровня старости населения
1	< 8	Демографическая молодость
2	8-10	Первое преддверие старости
3	10-12	Собственно преддверие старости
4	12 и выше	Демографическая старость
	12-14	Начальный уровень демографической старости
	14-16	Средний уровень демографической старости
	16-18	Высокий уровень демографической старости
	18 и выше	Очень высокий уровень демографической старости

В последние 2 десятилетия в России сохраняется неблагоприятная демографическая ситуация, четко проявляющаяся в большинстве территорий. Так, если в 1990 году доля лиц пенсионного возраста и детей в Санкт-Петербурге была практически одинакова (соответственно, 20,6% и 19,8%), то в 2011 году доля лиц пенсионного возраста более чем в два раза превышала долю детей (соответственно, 25,5% и 12,1%). Аналогичная тенденция наблюдалась и в целом по РФ. В 2009 году доля россиян нетрудоспособного возраста составляла 18,7% и детей 24,5%. В 2011 году это соотношение поменялось: доля пенсионеров возросла до 22,3%, а детей – упала до 16,3% (рис. 38). Понятно, что такая тенденция не может не накладывать существенный отпечаток на статистику показателей общественного здоровья и организацию медицинской помощи населению.

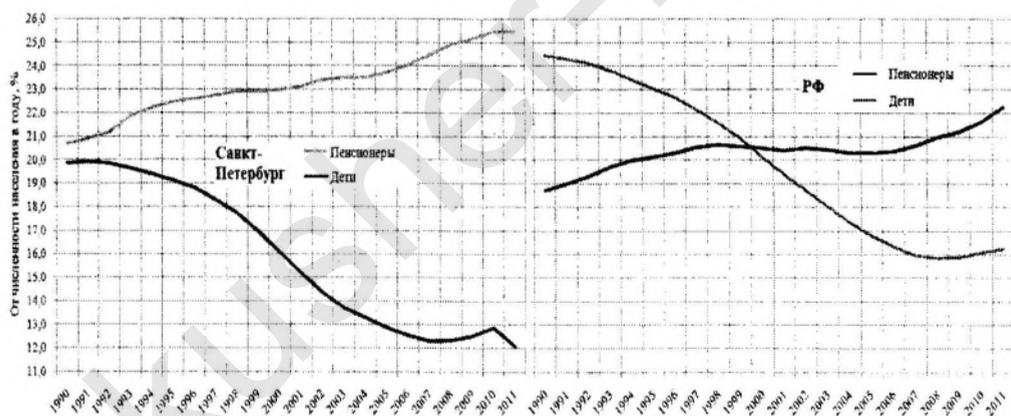


Рис.38. Распределение населения Санкт-Петербурга и Российской Федерации нетрудоспособного возраста (женщины >55 лет, мужчины > 60 лет) и детей (<15 лет) за 1990-2011 гг. (% к численности населения по каждому году наблюдения).

Для более детального, графического отражения пропорций соотношения возрастных групп строят возрастную пирамиду, используя годичные временные интервалы, что позволяет наглядно оценить динамику и перспективы воспроизводства населения.

Возрастная пирамида – наглядное графическое отображение распределения населения по возрасту. Представляет собой двустороннюю ленточную диаграмму, на которой каждая возрастная группа располагается последовательно, вертикально, по мере увеличения возраста (обычно от 0 до 100 лет), в виде горизонтальной ленты одинаковой ширины, но разной длины, зависящей от величины сравниваемых характеристик (пол, территориальное распределение и т.п.). Изменение численности населения по возрастным группам вследствие демографических процессов приводит к тому, что длина каждой последующей ленты не соответствует предыдущей. При этом общая тенденция естественного сокращения численности населения по мере его вымирания создает эффект пирамиды.

Возрастная пирамида строится с однолетним или 5-летним возрастным непрерывным интервалом отдельных групп, поэтому она представляет собой как бы моментную фотографию процесса воспроизводства населения. На практике длина отдельных лент (или групп лент) придает пирамиде волнообразный характер, определяемый волнообразностью процессов воспроизводства населения. Сокращается, когда численность населения в каком-либо возрастном интервале более интенсивно уменьшается в силу повышенной смертности из-за войн, социальных катаклизмов, миграции и т.п. И наоборот, увеличивается, отражая отдаленные последствия расширенного воспроизводства населения. Интенсивность процесса воспроизводства населения отражается шириной основания возрастной пирамиды.

Отражаемая таким способом динамика численности отдельных возрастных групп делает наглядной эволюцию процессов рождаемости и смертности на протяжении длительного периода времени в виде движения демографических волн, образовавшихся в результате резких изменений процессов воспроизводства населения.

Так, в РФ значительные нарушения интенсивности воспроизводства населения (падение рождаемости в годы войн, социальных катаклизмов и т.п.) вызывали уменьшение численности отдельных возрастных групп (сокращение

длины выступов, соответствующих этому возрасту), показывая, в какие годы и в какой мере происходило снижение процесса естественного воспроизводства населения (рис. 39).

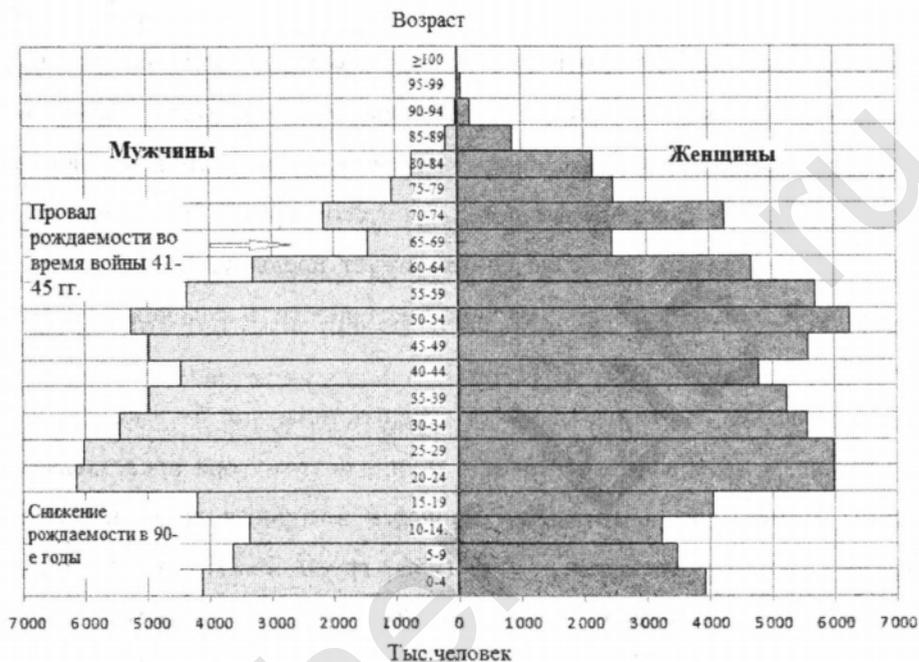


Рис. 39. Пирамида возрастного-полового распределения населения РФ в пятилетнем интервале (тыс. человек в 2011 году)

Обращает на себя внимание и перспективное падение детородного потенциала страны, поскольку особенно резкое снижение численности отмечается в поколении девочек, еще не достигших 14 лет (рис. 40). Через 10-15 лет именно это поколение будет составлять основную численность женщин наиболее репродуктивного возраста.

С точки зрения решения задач практической медицинской статистики, возрастная пирамида делает наглядной возможность нарушения качественной репрезентативности показателей, вычисляемых в укрупненных возрастных интервалах, поскольку возрастная структура женского населения РФ в 2009 и 2011 гг. имела определение отличия. Например, структура возрастной группы «от 20 до 34 лет» в 2009 г. была совершенно отличной от структуры

аналогичной группы в 2011 г. В 2011 году женщины в возрасте 30, 31, 32, 34 лет занимали в ней больший удельный вес, чем в 2009 году, что может влиять на величины и, соответственно, различия, например, интенсивных показателей заболеваемости.

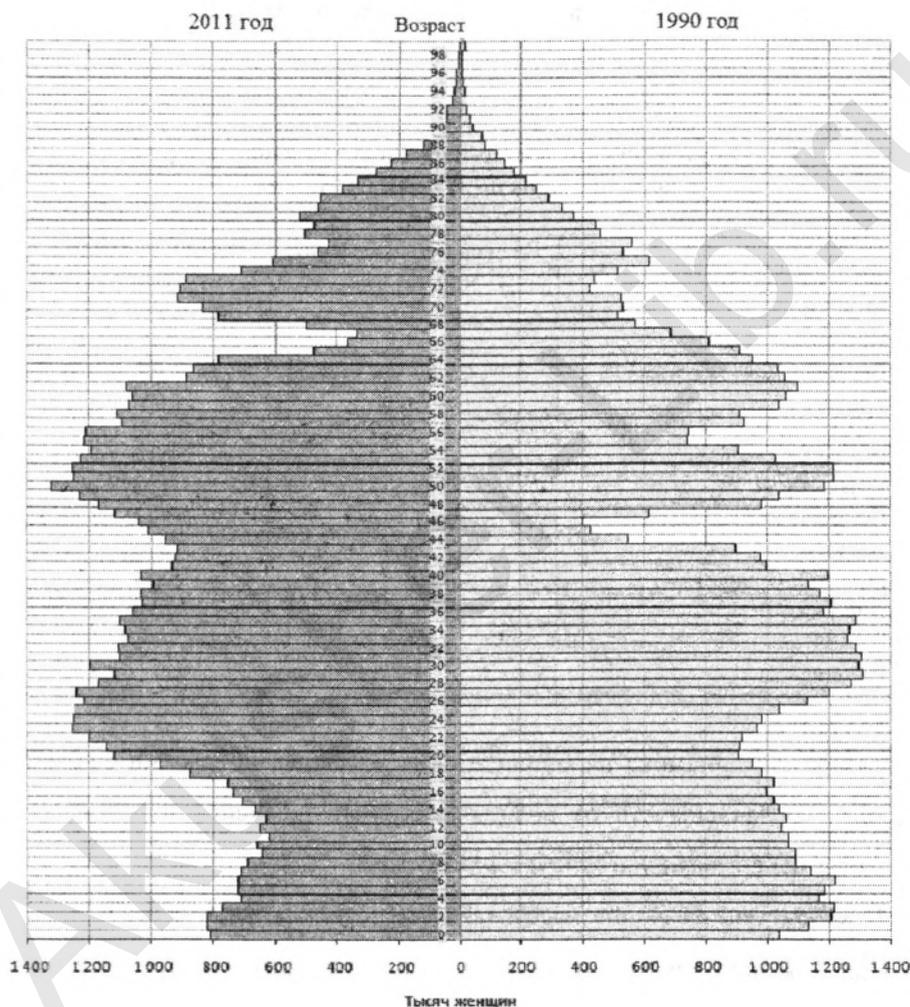


Рис.40. Возрастная пирамида распределения состава женского населения РФ в 2011 и 2009 гг. (ЕМИСС, 2012 год)²⁰

²⁰ Для определения численности женского населения РФ в 1990 и 2011 гг. к данным последних предыдущих переписей прибавлены числа родившихся и прибывших и вычтены числа умерших и выбывших. (Инструктивно-методическое пособие по демографической статистике. Утверждено 7.12.2007 г. Росстат).

3.3.2. Показатели динамики народонаселения

Динамика народонаселения – это изменение (движение) количества населения вследствие естественных биологических и социально-экономических процессов. Одним из ярких статистических показателей динамики народонаселения является период удвоения численности населения (табл. 50).

Таблица 50

Период удвоения численности населения отдельных стран и регионов мира по данным ООН, 1998²¹

Страна или регион	Численность населения, млн. человек	Период удвоения, лет
Весь мир	5,926	49
Нигерия	121,8	28
Развивающиеся страны	4,748	40
Китай	1 242,5	69
США	270,2	116
Северная Америка	301	117
Казахстан	15,8	133
Грузия	5,4	173
Япония	126,4	330
Развитые страны	1,178	584
Австрия	8,1	990
Россия	146,9	–
Швеция	8,9	–

Понятно, что динамика народонаселения обусловлена самыми различными причинами. В зависимости от вида этих причин движение народонаселения подразделяется на: *механическое и естественное*²².

Механическое движение населения происходит в результате переселения или миграции, связанной чаще всего со сменой жительства. Термин «миграция» происходит от латинского слова «*migratio*» (*migro* – перехожу, переселяюсь). Миграция подразделяется на: безвозвратную (постоянную); временную (переселение на длительный, но ограниченный срок); сезонную (перемещение в

²¹ Цит. по В.М. Медкову, 2005 г.

²² Иногда выделяют социальное, естественное и миграционное движение населения.

определенные периоды года); маятниковую (регулярное перемещение к месту работы или учебы за пределы своего населенного пункта).

Кроме того, различают *внешнюю или международную миграцию* (переселение за пределы своей страны) и *внутреннюю* (перемещение внутри страны). Вариантами 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. Так как прирост коренного населения трудоспособного возраста замедлился и почти остановился в конце прошлого десятилетия, то иммигранты играют важную роль в восполнении нехватки рабочей силы в США и ЕС. Чтобы снять проблемы нелегальной миграции, США, Израиль, Германия, Италия, Греция, Португалия, Испания, Австралия и некоторые другие повышают квоты на прием иммигрантов, главным образом, трудоспособного возраста. Это позволяет не только заполнять образовавшийся вакуум на рынках труда (в первую очередь малоквалифицированного), но и

омолаживать нацию, поддерживать на должном уровне соотношение трудоспособных и нетрудоспособных лиц.

В итоге в Швейцарии 70% (2,8 миллиона) роста населения в течение 1945-2000 годов (с 4,4 до 7,2 миллиона²³) – это прямое или косвенное влияние иммигрантов. В Бельгии 45% увеличения численности населения с 1945 года по настоящее время происходило только благодаря иммиграции. В Швеции, где сегодня проживает около 9 млн. чел., 75% всего прироста населения – миграционный прирост.

В Кувейте, например, численность населения к июлю 2000 года составляла около 2,8 млн. человек. Из них коренных жителей – арабов-кувейтцев, какими считаются только те лица, которые смогут доказать свои кувейтские генеалогические корни с 1920 года, – по разным источникам только от 33% до 45%, остальные – иммигранты. При этом в частном секторе задействованы, главным образом, иностранцы. Кувейтские граждане – в государственном. Причина – более высокие социальные гарантии государственных служащих.

Большой размах приобрела иммиграция в России. 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. В 2000 году эмигрировало из страны 145,7 тыс. человек, иммигрировало в страну – 359,3 тыс. человек (табл. 51).

Положительное сальдо миграции прослеживается на протяжении последнего десятилетия за счет приезда из стран СНГ и Балтии. Отрицательное сальдо – за счет стран ЕС, Великобритании и США.

²³ По итогам переписи-2010, на октябрь 2010 года в Москве постоянно проживало 11,6 млн. человек. В 2012 гг. – около 11,9 млн. чел.

**Динамика процессов внешней миграции в Российской Федерации
в 2000-2011 гг., по данным Росстат**

Миграция	2000 г.	2005 г.	2006 г.	2007 г.	2008 г.	2009 г.	2010 г.	2011 г.
Прибыло в РФ	359 330	177 230	186 380	286 956	281 614	279 907	191 656	356 535
Выбыло из РФ	145 720	69 798	54 061	47 013	39 508	32 458	33 578	36 774
Баланс (сальдо миграции)	213 610	107 432	132 319	239 943	242 106	247 449	158 078	319 761
В том числе:								
из стран СНГ	326 561	163 101	170 851	263 277	261 170	261 495	171 940	310 549
в страны СНГ	80 510	35 418	34 669	30 726	25 542	20 326	21 206	22 568
Сальдо	213 610	107 432	132 319	239 943	242 106	247 449	158 078	319 761
из стран ЕС	6 259	2 005	2 006	2 007	2 008	2 009	2 010	2 011
в страны ЕС	45 064	25 236	12 182	10 521	8 795	7 509	7 185	7 749
Сальдо	-38 805	-23 231	-10 176	-8 514	-6 787	-5 500	-5 175	-5 738

Еще одной из составляющих внутренней миграции населения является *урбанизация*. Урбанизация («*urbus*» – город) – это процесс миграции сельских жителей в развитые промышленные центры, который имеет место в большинстве государств мира. В результате число жителей сел сокращается²⁴, а города — стремительно растут. 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. чел., Москвы – более 11 млн. чел.

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 годы она выросла до 73,5%. С 1993 года численность городского населения стала несколько снижаться (73,0%). С 1994 года удельный вес городских жителей, опять несколько повысившись к 2002 году, фактически остается стабильным (табл. 52).

²⁴ В Кировской области РФ во время переписи 2010 года было обнаружено село с населением 1 человек.

**Динамика численности городского и сельского населения РФ
по данным переписей 2002 и 2010 гг.**

Население	2002 год		2010 год		2010 год в % к 2010
	Тыс.чел	%	Тыс.чел.	%	
Городское	106 430	73,3	105 314	73,7	99,0
Сельское	37 738	26,7	37 543	26,3	96,9
Все	145 168	100,0	142 857	100,0	98,4

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 др. могут оказывать неблагоприятное воздействие на здоровье людей. Следует отметить расхожес мнение о том, что «на селе люди здоровее», в условиях РФ не соответствует действительности. Данные официальной статистики о заболеваемости населения, вроде бы подтверждающие это мнение, не соответствуют действительности. Дело в том, что показатели заболеваемости по обращаемости за медицинской помощью (практически основной вид информации о заболеваемости населения) формируются на основании посещений медицинских учреждений. Но доступность квалифицированной и специализированной медицинской помощи на селе, как правило, ниже из-за существенного увеличения удаленности потенциальных пациентов от мед. учреждений. Специалисты называют этот фактор «увеличением радиуса обслуживания» на селе по сравнению с городом. Кроме того, учреждения, оказывающие высококвалифицированную и специализированную медицинскую помощь, по экономическим соображениям обычно располагаются в городах, в лучшем случае – в крупных

межрайонных центрах, что тоже негативно сказывается на доступности мед. помощи. Таким образом, сокращается число реальных обращений за этой помощью, а из числа состоявшихся обращений жителей села многие обращения регистрируются в городе и крупных центрах. Отсюда заниженные показатели заболеваемости по обращаемости населения, постоянно проживающего на селе.

Практически все виды миграции, будь то внутренняя (урбанизация, маятниковая и др.) или внешняя (эмиграция, иммиграция), меняют уровень и структуру заболеваемости, смертности населения, приводят к изменению нагрузки на учреждения здравоохранения, изменяют эпидемическую ситуацию. Например, маятниковая (ежедневно из дома на работу и обратно) и сезонная миграции способствуют распространению инфекций, ухудшению общей эпидемиологической ситуации, травматизму. Все это обуславливает необходимость учета такого рода факторов при организации медицинской помощи.

Естественное движение – это совокупность таких демографических явлений, как рождаемость, смертность, в том числе младенческая и материнская смертность, естественный прирост населения, средняя продолжительность предстоящей жизни и др. Расчет каждого из статистических показателей, описывающих эти явления, имеет свои методические особенности и определенные логические границы информативности. Недоучет этих особенностей и границ ведет к серьезным ошибкам анализа статистических данных.

Рождаемость – процесс деторождения в совокупности наблюдений. Этот процесс, описываемый статистическими показателями, является важнейшим критерием жизнеспособности и воспроизводства населения. Именно от рождаемости зависит замещение выбывающих поколений. В демографической статистике используется целый комплекс показателей, характеризующих рождаемость.

Следует различать термины «рождаемость» и «плодовитость». *Плодовитость* – это биологическая способность к деторождению, а *рождаемость* – реализация этой способности. Как антонимы, т.е. лексически противоположные понятия, *плодовитости* выделяют: *бездетность, бесплодие, инфертильность, стерильность*.

Бездетность – отсутствие детей у лица или брачной пары. Определяется распространенностью бесплодия, добровольного отказа от рождения детей и смертностью детей. *Бездетность* – одна из характеристик состава семьи. Статистически оценивается долей лиц или брачных пар, не имеющих детей на момент переписи или выборочного обследования. На величину этой характеристики влияют многие обстоятельства, в том числе возраст, брачное состояние, продолжительность брака.

Бесплодие (в демографии) – всякая неспособность человека или брачной пары иметь детей. Для брачных пар вычисляется как процентное отношение бесплодных пар к общему числу браков. Бесплодие может быть вызвано неспособностью одного или обоих супругов к зачатию (временная или постоянная *стерильность*) или тем, что женщина не может выносить и родить живого ребенка. В медицинской статистике учитывается только *непроизвольное бесплодие*, которое подразделяется на *первичное*, при котором женщина или брачная пара, живущая нормальной половой жизнью, никогда не имела живорождений, и *вторичное*, которое наступило после рождения одного или нескольких детей либо в результате неадекватного регулирования репродуктивного цикла. Показатели бесплодия могут относиться к лицам детородного возраста или к брачным союзам. Согласно определению ВОЗ (1993), брак считается бесплодным, если беременность у женщины не наступает в течение года регулярной половой жизни без применения каких-либо мер контрацепции. Иногда используются и другие определения бесплодия. *Например*, брак считается бесплодным, если в течение 3-х лет регулярной половой жизни и при условии неприменения

средств контрацепции и искусственных абортов в семье не происходит рождение здорового ребенка (Медков В.М., 2005).

В настоящее время существуют различные оценки распространенности бесплодия. В экономически развитых странах показатель абсолютного бесплодия (нулевые шансы рождения ребенка) колеблется в пределах 10%-15%. Еще столько же брачных пар обладают пониженной плодовитостью.

Рождаемость как статистический показатель определяется либо абсолютным числом рождений, либо относительным коэффициентом, который называется – «*общий коэффициент рождаемости*». Он вычисляется, как отношение числа родившихся в данном году на 1000 человек населения, т.е. в промилле ($^0/_{00}$). При этом за основу расчета обычно берется среднегодовая численность населения, которая рассчитывается, как полусумма численности населения на начало и конец календарного года. Иногда ее вычисляют, как численность населения на начало года плюс половина прироста численности населения за год; при наличии численности населения на начало каждого месяца рассчитывается средняя хронологическая численность населения за год: берется 1/12 от суммы половины численности на начало года, численности на начало каждого месяца и половины численности на конец года²⁵.

При оценке рождаемости применяются две разные методики вычисления показателей рождаемости. Если общий показатель (коэффициент) рождаемости – есть средняя величина числа рождений, приходящаяся на 1000 жителей в год, то специальный показатель рождаемости (плодовитости) – показатель, рассчитываемый только относительно той группы населения, которая способна рожать, т.е. применительно к женщинам репродуктивного возраста. В демографии обычно такой группой принято считать женщин в возрасте от 15 до 50 лет. В некоторых странах возрастные границы этой группы имеют свои национальные особенности.

²⁵ В ряде случаев расчет рождаемости производится только относительно живорожденных.

В РФ уровень общей рождаемости оценивается в следующих границах показателей: низкий уровень – до $19,9^0/_{00}$, средний – $20,0-29,9^0/_{00}$, высокий – $30,0^0/_{00}$ и выше.

Наиболее точную характеристику демографического процесса рождаемости дает *суммарный коэффициент рождаемости*, иначе называемый «*коэффициент фертильности*». Смысл этого коэффициента – учет повозрастных различий интенсивности деторождения, т.е. возможности влияния на статистику рождений различного возрастного состава сравниваемых групп населения. Характеризует среднее число рождений у одной женщины репродуктивного возраста (от 15 до 50 лет) в гипотетическом поколении за всю её жизнь при сохранении существующих уровней рождаемости в каждом возрасте независимо от смертности и от изменений возрастного состава. Называется «суммарный», т.к. вычисляется, как сумма возрастных коэффициентов рождаемости для женщин в возрасте 15-50 лет. В условиях низкой смертности для простого замещения поколений суммарный коэффициент рождаемости должен быть не ниже 2,15. Суммарный коэффициент рождаемости выше 4,0 считается высоким, а ниже 2,15 – низким.

Коэффициент суммарной рождаемости в России, несмотря на некоторую позитивную тенденцию, обозначившуюся в последние годы (рис. 41), начиная с 1965 года, остается низким.

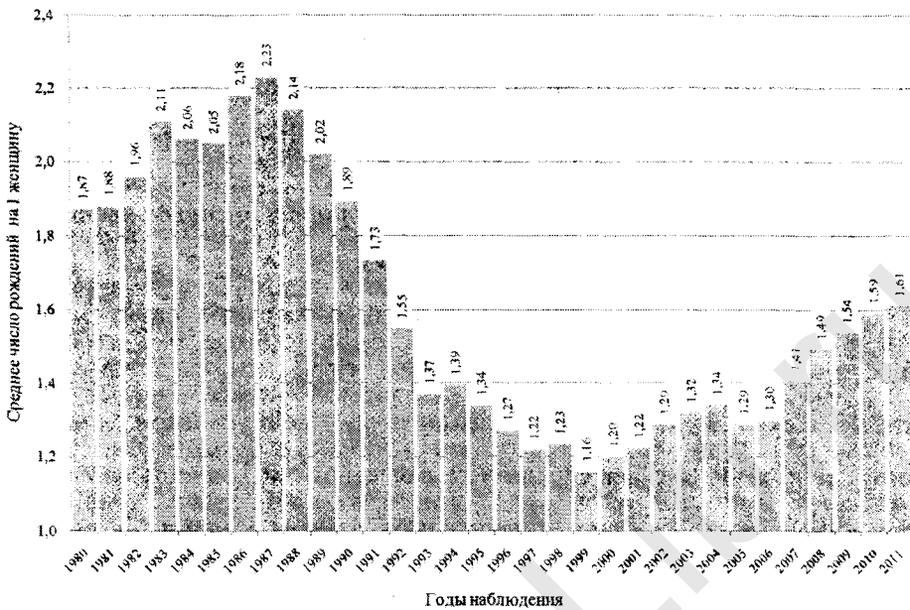


Рис.41. Динамика суммарного коэффициента рождаемости в России в 1980-2011 гг. (число рождений на 1 женщину)

Эта тенденция в целом соответствует общемировой тенденции снижения рождаемости населения. Суммарный коэффициент рождаемости в мире, например, сократился с 4,95 рождений, приходившихся в среднем на одну женщину в 1960 году, до 2,56 – в конце XX века (2000 г.).

Для более углубленного демографического анализа рождаемости используются и многие другие частные статистические коэффициенты: брачная и внебрачная рождаемость, повозрастная рождаемость и т.п. В последние годы в отечественной демографической статистике все чаще стали появляться англоязычные термины: «тайминг» и «спейсинг» рождаемости. *Тайминг* характеризует интервалы времени между рожденьями. *Спейсинг* – распределение рождений на протяжении всего репродуктивного периода.

Детальную характеристику рождаемости как существенного элемента воспроизводства населения дают специальные показатели воспроизводства: *брутто-* и *нетто-коэффициенты*. *Брутто-коэффициент* воспроизводства учитывает только родившихся у женщин детородного возраста девочек, для

этого число родившихся детей умножают на долю девочек среди новорожденных. *Нетто-коэффициент*, называемый по фамилиям авторов коэффициентов Бёка-Кучински, отличается от брутто-коэффициента тем, что учитывает только девочек, доживающих до генеративного возраста.

3.3.3. *Смертность населения, средняя продолжительность предстоящей жизни*

Статистические таблицы смертности (дожития) – один из основных видов демографических таблиц. По своей сути, они представляют собой математические модели процесса вымирания некой условной совокупности (когорты) населения с фиксированной начальной численностью, называемой «корень таблицы». Его размер принимают равным какой-либо степени 10: 10 000, 100 000, 1 000 000 чел. и т.п.

Таблицы дожития делают наглядной динамику процесса вымирания в разных поколениях (табл. 53 и табл. 54). В основе этих таблиц – шкала чисел полных лет, прошедших с момента рождения. Начальный возраст в таблице – 0 лет (момент рождения), конечный – возраст, к которому вымирает практически вся совокупность родившихся.

Исходные данные для расчетов представлены по соответствующим колонкам таблицы дожития.

Точный возраст x . Это независимая переменная, которая может измеряться в годах, днях, месяцах (при изучении младенческой смертности), но обычно измеряется в годах.

В колонке, следующей за колонкой точного возраста, указывается возрастной интервал таблицы n . В зависимости от его протяженности выделяют полные ($n=1$) и краткие ($n=5$ или $n=10$) таблицы дожития. В табл. 53 приведен пример краткой таблицы смертности (дожития) при $n=5$. Но в любой таблице первый интервал обычно принимается в границах от 0 до 1 года, т.е. в возрастных границах младенческой смертности. Такой интервал

обусловлен тем, что младенческая смертность как демографическое явление имеет уникальные закономерности, характерные только для этого возраста. Определенные особенности имеет и смертность детей в возрасте от 1 до 4 лет. Поэтому этот период нередко также представляется в виде четырех одногодичных возрастных интервалов: 1, 2, 3, 4 года. Свою особенность имеет еще один, последний, интервал таблицы дожития (W): в приведенном примере – «85 лет и более». Т.е. этот интервал открытый, без нижней границы. (Иногда выбирают и другой последний интервал. Но в любом случае он берется открытым, т.е. интервалом без нижней границы). Связано это с тем, что число смертей в этом интервале весьма незначительно, и поскольку оно ничтожно мало влияет на итоговые показатели всей таблицы дожития, то здесь не имеет практического смысла указывать точный возраст смерти.

Затем указываются уровни смертности (m_i) по соответствующим возрастным группам, где i – порядковый номер группы. В официальных источниках он обычно приводится в пересчете на 1000 или 100000 населения. В таблице дожития следует его пересчитать относительно 1.

Таблица 53

Таблица смертности (дожития) населения РФ в 2009 году

№ возраст-ного интер-вала, j	Возраст, лет x _i	Длина воз-растного интервала, n _i	Повозрастной коэффициент смертности		Поправочный коэффициент a _i	По Паев-скому I _x	q _x	d _x	L _x	T _x	Ē _x
			m _i в ‰	m _i /1000							
0	0-1	1	8,14	0,0081	0,10	10 000	0,0081	81	9 927	687 406,2	68,74
1	1	1	0,84	0,0008	0,10	9 919	0,0008	8	9 911	677 479,5	68,30
2	2	1	0,50	0,0005	0,10	9 911	0,0005	5	9 906	667 568,0	67,36
3	3	1	0,40	0,0004	0,10	9 906	0,0004	4	9 902	657 661,9	66,39
4	4	1	0,35	0,0004	0,10	9 902	0,0003	3	9 899	647 759,8	65,42
5	5-9	5	0,31	0,0003	0,50	9 898	0,0015	15	49 453	637 861,2	64,44
6	10-14	5	0,35	0,0004	0,50	9 883	0,0017	17	49 371	588 408,5	59,54
7	15-19	5	0,98	0,0010	0,50	9 866	0,0049	48	49 207	539 037,2	54,64
8	20-24	5	1,74	0,0017	0,50	9 817	0,0087	85	48 874	489 829,7	49,89
9	25-29	5	2,94	0,0029	0,50	9 732	0,0146	142	48 307	440 955,4	45,31
10	30-34	5	4,32	0,0043	0,50	9 590	0,0214	205	47 439	392 648,7	40,94
11	35-39	5	4,97	0,0050	0,50	9 385	0,0245	230	46 351	345 209,4	36,78
12	40-44	5	6,34	0,0063	0,50	9 155	0,0312	286	45 061	298 858,3	32,64
13	45-49	5	8,60	0,0086	0,50	8 869	0,0421	373	43 414	253 797,3	28,61
14	50-54	5	12,10	0,0121	0,50	8 496	0,0587	499	41 233	210 383,8	24,76
15	55-59	5	17,00	0,0170	0,50	7 997	0,0815	652	38 356	169 150,7	21,15
16	60-64	5	23,50	0,0235	0,50	7 346	0,1110	815	34 690	130 794,3	17,81
17	65-69	5	31,80	0,0318	0,50	6 531	0,1473	962	30 251	96 104,2	14,71
18	70-74	5	45,40	0,0454	0,50	5 571	0,2039	1136	25 016	65 852,7	11,82
19	75-80	5	70,40	0,0704	0,50	4 440	0,2993	1329	18 877	40 836,2	9,20
20	80-84	5	107,74	0,1077	0,50	3 122	0,4244	1325	12 299	21 959,6	7,03
21	80 и старше		188,60	0,1886	0,50	1 822	1,0000	1 822	9 660	9 660,4	5,30

Таблица смертности (дожития) населения РФ в 2009 году при исключении онкологических заболеваний

№ интервала, i	Возраст лет x_i	Длина интервала, h_i	Общий коэффициент смертности		Остаточный коэффициент смертности m_i на 1 ед.	Поправочный коэффициент a_i	По Паревскому l_x	q_x	d_x	L_x	T_x	e_x	Доп. прод. жизни
			m_i / 100	m_i / 1000									
0	0-1	1	0,057	0,008083	0,10	10,000	0,0081	81	9 927	706 947,8	70,69	1,95	
1	1	1	0,054	0,000786	0,10	9 919	0,0008	8	9 912	697 020,5	70,27	1,97	
2	2	1	0,045	0,000455	0,10	9 912	0,0005	5	9 908	687 108,0	69,32	1,96	
3	3	1	0,056	0,000344	0,10	9 907	0,0003	3	9 904	677 200,4	68,35	1,96	
4	4	1	0,049	0,000301	0,10	9 904	0,0003	3	9 901	667 296,3	67,38	1,96	
5	5-9	5	0,038	0,000272	0,50	9 901	0,0014	13	49 470	657 395,2	66,40	1,96	
6	10-14	5	0,036	0,000314	0,50	9 887	0,0016	16	49 398	607 924,8	61,49	1,95	
7	15-19	5	0,058	0,000922	0,50	9 872	0,0046	45	49 246	558 526,8	56,58	1,94	
8	20-24	5	0,066	0,001674	0,50	9 826	0,0083	82	48 927	509 281,1	51,83	1,93	
9	25-29	5	0,113	0,002827	0,50	9 745	0,0140	137	48 381	460 353,7	47,24	1,93	
10	30-34	5	0,205	0,004115	0,50	9 608	0,0204	196	47 550	411 973,0	42,88	1,94	
11	35-39	5	0,335	0,004635	0,50	9 412	0,0229	216	46 521	364 423,4	38,72	1,94	
12	40-44	5	0,635	0,005705	0,50	9 196	0,0281	259	45 336	317 901,9	34,57	1,92	
13	45-49	5	1,204	0,007396	0,50	8 938	0,0363	325	43 878	272 566,1	30,50	1,88	
14	50-54	5	0,226	0,011874	0,50	8 613	0,0577	497	41 825	228 688,1	26,55	1,79	
15	55-59	5	3,778	0,013222	0,50	8 117	0,0640	519	39 286	186 862,8	23,02	1,87	
16	60-64	5	5,554	0,017946	0,50	7 598	0,0859	652	36 357	147 577,0	19,42	1,62	
17	65-69	5	7,197	0,024603	0,50	6 946	0,1159	805	32 716	111 220,0	16,01	1,30	
18	70-74	5	8,862	0,036538	0,50	6 142	0,1674	1028	28 138	78 504,4	12,78	0,96	
19	75-80	5	9,924	0,060476	0,50	5 116	0,2627	1344	22 221	50 366,5	9,84	0,65	
20	80-84	5	10,250	0,097490	0,50	3 781	0,3919	1482	15 201	28 145,4	7,44	0,41	
21	80 и старше		9,194	0,179406	0,50	2 322	1,0000	2 322	12 945	12 944,6	5,57	0,27	

l_x – число доживших до точного возраста x из начальной численности когорты. Начальная численность когорты (поколения) или корень в табл. 53 и табл. 54 принят равным 10000 человек.

${}_n d_x$ – числа умирающих в интервале возраста от x_i до (x_i+n_i) . Сумма всех значений ${}_n d_x$ должна быть равной корню таблицы. В данном примере 10 000. Остальные показатели этого столбца при известных значениях чисел доживших до точного возраста l_{x+1} равны:

$${}_n d_x = l_{x+1} - l_x$$

${}_n q_x$ – вероятность (в долях от единицы) умереть в интервале возраста от x_i до x_i+n_i . Для всех тех, кто дожил до возраста x_i .

$${}_n q_x = {}_n d_x / l_x$$

Этот показатель иногда называют «табличной смертностью». В последней строке любой таблицы этот показатель всегда равен 1,000. Существенной методологической проблемой, во многом обусловившей появление разных технологий расчета таблиц дожития, является то, что вероятность смерти на определенном возрастном этапе не есть показатель возрастной смертности. Проблема состоит в том, что табличный показатель ${}_n q_x$ рассчитывается по отношению к среднегодовой численности населения возраста x_i , а показатель повозрастной смертности – по отношению к численности населения в начале возрастного интервала x_i .

Одним из наиболее удобных способов решения этой проблемы является использование метода В.В. Паевского (1893-1934), с помощью которого осуществляется переход от возрастного коэффициента смертности m_i к величине $l_{x_i+n_i}$ по формуле:

$$L_{x_i+n_i} = L_{x_i} * e^{-n_i m_i}, \text{ где } e=2,71828183$$

${}_n p_x$ – вероятность выжить к возрасту x_i+n_i для всех тех, кто дожил до возраста x_i ;

$${}_n p_x = 1 - {}_n q_x$$

${}_nL_x$ – среднее число человеко-лет, прожитое в интервале возраста от x_i до x_i+n_i теми, кто дожил до начала данного интервала (lx). Для первого интервала формула выглядит следующим образом:

$${}_nL_x = 2/3d_x$$

Для остальных интервалов: ${}_nL_x = n \cdot l_{x+n} + {}_n a_x \cdot d_x$,

где a_i – специальный коэффициент. Не вдаваясь в математическое обоснование выбора его величины, следует отметить, что его целесообразно устанавливать для возрастных интервалов: до 4-лет включительно он равен 1, для более старших возрастов он равен 5.

T_x – число человеко-лет жизни в возрасте x лет и старше:

$$T_x = L_x + L_{x+n} + \dots + L_w$$

e_x – ожидаемая продолжительность жизни в возрасте x лет:

$$e_x = T_x / l_x$$

Чаще всего пользуются важнейшей демографической характеристикой смертности населения – ожидаемой продолжительностью жизни при рождении. При условии, что число умирающих на первом году жизни достаточно велико, может проявиться так называемый парадокс младенческой смертности, когда $e_1 > e_0$. Она рассчитывается следующим образом:

$$e_0 = T_0 / l_0$$

В приведенной табл. 53 этот показатель равен 68,7 года, что соответствует официальным данным РОССТАТ. Следует обратить внимание на то, что средняя продолжительность предстоящей жизни для каждого возраста имеет свои значения. Например: для возрастной группы 4 года он равен 65,4 г. Отсюда вероятный средний возраст смерти этой когорты населения РФ = $4 + 65,4 = 69,4$ г. Таким образом можно рассчитать вероятную среднюю длительность предстоящей жизни для каждой возрастной группы (когорты) населения. При этом, по мере увеличения возраста когорты, будет расти средняя длительность предстоящей жизни.

Таблица смертности от всех причин при условии исключения одной причины. Демографические таблицы дожития позволяют произвести расчет показателей средней продолжительности предстоящей жизни при исключении одной или нескольких причин смерти людей. Основой логического построения этих расчетов является то, что вероятность смерти людей является суммой вероятностей смерти от отдельных причин. При этом для упрощения расчетов, как правило, рассматриваются такие причины, которые являются несовместными событиями²⁶. Например, смерть от дорожно-транспортного происшествия и смерть от онкологического заболевания можно рассматривать как независимые события.

Рассмотрим пример изменения длительности предстоящей жизни для населения РФ при исключении смертей от онкологических заболеваний, т.е. некоторой гипотетической ситуации, когда онкологическая заболеваемость исчезла²⁷.

В составе демографической таблицы дожития (табл. 54) приведены данные о по возрастной онкологической заболеваемости на 1000 населения. В следующей графе приведены показатели остаточной смертности, равной общей смертности минус смертность от онкологических болезней. Для удобства восприятия этот показатель представлен в пересчете на 1 чел. Например, показатель общей смертности для возрастной группы 20-24 года равен $1,74^{0/00}$. Показатель смертности от онкологических болезней для этой же возрастной группы равен $0,066^{0/00}$. Разность между ними, т.е. остаточная смертность $=1,74-0,066=1,674$, или в пересчета на 1 чел. $0,001674$. Все остальные показатели в этом столбце рассчитаны аналогично, а остальные столбцы пересчитаны в соответствии с алгоритмом табл. 53.

²⁶ Согласно теореме о сложении вероятностей, вероятность появления несовместных событий равна простой сумме вероятностей этих событий. Если события совместные, то алгоритм сложения несколько сложнее, однако это не отменяет возможности сложения вероятностей событий.

²⁷ О том, что такая ситуация, в принципе, возможна, свидетельствует заявление в 1980 г. ВОЗ о полной ликвидации натуральной оспы.

Из результатов расчетов (табл. 55) следует, что продолжительность предстоящей жизни в ситуации отсутствия онкологической смертности увеличилась на $70,69-67,74=1,95$ года, т.е. на 2,8%, тогда как показатель общей смертности уменьшился на 14,5%.

Отсюда становится очевидным, что избежавший смерти от одной причины может умереть от другой. Таким образом, утверждение, что при исключении онкологической смертности общая смертность не может уменьшиться на 14,5%. Т.е. показатели смертности, даже в случае рассмотрения заведомо не связанных (несовместных) случаев смерти, не обладают свойством аддитивности, т.е. свойством простого сложения. Отсюда вытекает важное обстоятельство, повсеместно упускаемое разработчиками разного рода показателей «рисков», — снижение смертности от одной причины не гарантирует пропорционального снижения общей смертности. В равной степени это можно отнести и к другим показателям общественного здоровья.

Таблица 55

Итоги расчетов средней длительности предстоящей жизни населения РФ в 2009 году в условиях гипотетической ситуации отсутствия онкологических болезней у населения

2009 год	Смертность на 100 тыс.	Длительность жизни, лет
Вся смертность	1416,8	68,74
Смертность от онкологических болезней	204,9	—
Вся смертность без онкологических болезней	1211,9	70,69
Темп роста	-14,5%	2,8%

4. Основы математико-статистической обработки данных

Математическая статистика – раздел статистики, посвященный математическим методам систематизации, обработки и анализа данных. Методы математической статистики основаны на вероятностной природе этих данных, поэтому нередко термин «статистические» используется для того, чтобы подчеркнуть вероятностный, случайный характер тех величин, которыми оперирует математическая статистика. Можно выделить два основных, базовых принципа математической статистики:

- оценка явлений только в массе (в группе) наблюдений;
- вероятностный характер всех статистических оценок.

На практике применение методов математической статистики почти всегда обусловлено необходимостью решения **одной из двух** (или **обоих сразу**) **задач**:

1. **Нахождение различий в группах наблюдений.** Так как математическая статистика оперирует, как правило, с выборочными группами наблюдений, то эти группы и называются зачастую кратко: *выборки*. Кроме того, поскольку выборки описываются параметрами распределения наблюдений (средними, дисперсиями и т.п.), то задача нахождения различий по существу сводится к оценке различий параметров *распределений выборок*.

2. **Оценка взаимосвязи разных явлений либо отдельных значений одного явления, разнесенных в пространстве и (или) времени.** И эта группа задач реализуется опять-таки через оценки взаимосвязи *распределений выборок*.

Вся числовая информация в статистике воспроизводится в виде абсолютных или производных величин. **Все величины** могут быть простыми, которые всегда представляются в именованных единицах измерения (случай посещения, число медицинских услуг, длительность утраты трудоспособности и т.п.), или сложными, которые выражаются произведениями единиц различной размерности (человеко-часы и т.п.). Кроме того, в ряде случаев используются безразмерные величины или величины, имеющие какую-либо условную размерность, – *условные величины*. Например, условные единицы трудоемкости

(УЕТ) выполненной работы, которые широко применяются в статистике стоматологической медицинской помощи.

Абсолютные величины, служа основой любого статистического анализа, обычно используются только на этапе сводки или группировки исходных данных. Вместе с тем, уже на этом этапе обработки статистических данных могут вычисляться предварительные, промежуточные итоги в виде простейших производных величин (суммы, средние величины и т.п.). Проводятся такого рода вычисления обычно с целью предварительной оценки репрезентативности исходной информации и обоснования выбора конкретных методик статистического анализа. Для получения же исчерпывающей информации применяются *производные величины*.

4.1. Относительные величины. Статистические коэффициенты

Наиболее давно и широко применяемыми во всех разделах медицинской статистики, традиционной группой статистических величин являются **относительные величины** (*синоним* – **статистические коэффициенты**). Они широко используются при изучении медико-демографической и санитарно-эпидемиологической ситуации, оценки деятельности медицинских учреждений и т.п. Эти коэффициенты являются базой анализа деятельности любого медицинского учреждения, органов управления здравоохранения, основой планирования и оценки итогов реализации различных медико-социальных программ.

Относительной статистической величиной в наиболее общем виде называется отношение двух чисел, выражающее меру каких-либо явлений. Смысл их расчета – существенное повышение информативности данных статистического наблюдения путем приведения их к общему знаменателю. *Например:* численность врачей в Санкт-Петербурге в 1990 году составляла 32671 чел., а в 1996 году – 31549. На основании этих данных, представленных в абсолютных числах, можно сделать вывод о снижении обеспеченности врачами населения города на $32671-31549=1122$ врача. Однако за указанный

период времени (с 1990 по 1996 годы) численность жителей города уменьшилась с 5035200 до 4801500 человек. Если пересчитать результаты в относительных показателях, с учетом численности населения, то окажется, что обеспеченность врачами населения не уменьшилась, а возросла с 65,3 до 66,5 врача на 10000 человек населения города.

Из относительных величин наибольшее практическое значение в медицинской статистике имеют: интенсивные, экстенсивные коэффициенты, показатели соотношения, наглядности, а также показатели относительной интенсивности. Среди них различают *общие и специальные* коэффициенты. **Общие** коэффициенты дают обобщенную (усредненную)²⁸ характеристику оцениваемых явлений. Именно поэтому в западной статистике их называют «грубыми» (*crude* – англ.) коэффициентами. **Специальные** – дают более детальные характеристики. Вычислению и анализу специальных статистических коэффициентов всегда предшествует вычисление и анализ общих коэффициентов.

Интенсивные коэффициенты показывают интенсивность развития (частоту, уровень, распространенность – *синонимы*) явления в своей среде. В среде, которая продуцирует это явление. Явление и среда, продуцирующая это явление, по своей сути, представляют единое, неразрывное целое. Т.е. если убрать среду, то не будет и явления. *Например, смертность и заболеваемость населения* представляют собой классические интенсивные показатели. Если убрать население, то не будет ни заболеваемости, ни смертности. Применяются интенсивные коэффициенты, за редким исключением, только в медицинской статистике, демографии или на стыке этих наук – медицинской демографии. Интенсивные коэффициенты, которые иногда называют *коэффициентами интенсивности*, отвечают на вопрос: как часто явление встречается в известной среде?

²⁸ Именно поэтому статистические коэффициенты раньше считали особым видом средних величин, так называемыми «изолированными средними», которые отличаются от других видов средних отсутствием внятных параметров распределений статистических совокупностей, на основе которых они рассчитаны.

Вычисление интенсивных коэффициентов производится через пропорцию. *Например*, на предприятии в 2012 году зарегистрировано 1800 случаев заболеваний с временной утратой трудоспособности (ЗВУТ), в том числе 600 случаев заболеваний органов дыхания. Среднесписочное число работников в 2012 году составляло 2000 человек. Требуется на основе исходной информации, представленной в абсолютных числах, определить интенсивность явления ЗВУТ. Для этого составляем пропорцию:

1800 случаев ЗВУТ пришлось на 2000 человек;

X случаев ЗВУТ – на 100 человек;

Отсюда интенсивность ЗВУТ (общий показатель) $= (1800 \cdot 100) / 2000 = 90,0$ случаев ЗВУТ на 100 работников предприятия за 2012 год.

В данном примере *явление* – это заболеваемость, сопровождавшаяся ВУТ. *Среда* – работники предприятия. Следует обратить внимание на то, что случаи заболевания и численность работников предприятия берутся за один и тот же период времени (в данном примере за 2012 год). Одной из самых грубых ошибок, допускаемых при вычислении статистических коэффициентов, является несоблюдение единства времени для исходных данных.

Аналогичным образом вычисляются специальные показатели. *Например*, уровень ЗВУТ органов дыхания на предприятии (специальный показатель) $= (600 \cdot 100) / 2000 = 30,0$ случаев ЗВУТ органов дыхания на 100 работников предприятия в 2012 году.

Выбор множителя для интенсивного показателя обычно обусловлен удобством пользования результатами вычислений. Если в приведенном примере интенсивность ЗВУТ считать на 100000 работников, то получим достаточно громоздкий результат: 30000 случаев ЗВУТ на 100000 работников. В данной ситуации разумнее всего в качестве множителя использовать 100, т.к. при этом получается негромоздкое число.

Следует отметить, что интенсивные показатели ЗВУТ вычисляются всегда на 100 человек. Но это исключение, обычно при вычислении интенсивных коэффициентов могут применяться любые удобные для

пользователей значения множителей (100, 1000, 10000 и т.п.). Иногда вычисляют интенсивный показатель на 1 ед. *Например*, число посещений амбулаторно-поликлинических учреждений за год, приходившихся на 1 чел. В Москве в 2011 году показатель посещений к врачам («врачебных посещений») составлял 13,3 посещения на 1 чел. за год²⁹.

При выборе точности представления данных главное, чтобы в итоговых результатах число значащих цифр в дробной части было меньше, чем в целой части. Таким образом, при выборе размерности множителей строгих ограничений не существует. Существуют только общепринятые нормы, но строго заданными стандартами обычно они не являются. Поэтому, чтобы не было путаницы, важно обязательно указывать размерность интенсивного показателя.

Иногда для обозначения размерности 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 именованной величиной (случаи заболеваний, рождений, смертей и т.п.) и величиной, вычисляемой на определенную дату или за ограниченный временной интервал (за год, полугодие, квартал и т.п.).

Таким образом, обязательными атрибутами любого интенсивного показателя являются его наименование, размерность, период времени или момент времени, на который он вычислялся.

При расчетах, связанных с численностью отдельных контингентов или населения территорий, следует учитывать характер рассчитываемых производных величин. *Например*, обеспеченность населения какого-либо населенного пункта врачами-специалистами обычно рассчитывается на конец года (так наз. *моментное наблюдение*). Поэтому здесь и численность населения, и число врачей учитываются по состоянию на конец года.

²⁹ Учитывая врачебные посещения в поликлиники и на дому.

Если показатели представлены *интервальным наблюдением*, например, если заболеваемость представлена числом случаев ЗВУТ за год, то при расчетах число заболеваний берется в виде накопленной за год суммы всех случаев заболеваний, а число работников берется в виде среднегодовой численности работающих. В упрощенном варианте расчетов среднегодовая численность контингентов определяется, как полусумма численности контингента на начало и конец года. Более точно эти данные, особенно если в году происходили существенные изменения количества работников, исчисляются путем суммирования среднесписочной численности работников за каждый месяц соответствующего года и деления полученной суммы на 12. При этом среднесписочная численность работников за месяц определяется делением суммы численности работников списочного состава за каждый календарный день на число календарных дней месяца.

Одной из особенностей интенсивных коэффициентов является недопустимость их прямого сложения. Такое сложение интенсивных коэффициентов возможно только в особых случаях, о которых здесь не будет упомянуто.

Например, среди жителей района была проведена профилактическая вакцинация. Прошли вакцинацию 11420 человек, не прошли – 1053. Заболеваемость среди жителей, прошедших вакцинацию (в случаях на 1000 чел.), составила $42/11420 \cdot 1000 = 3,68\text{‰}$. Среди жителей, не прошедших вакцинацию, – $168/1053 \cdot 1000 = 159,54\text{‰}$. Требуется определить суммарный уровень заболеваемости.

Если определять итоговый уровень заболеваемости путем простого сложения интенсивных показателей ($3,68\text{‰} + 159,54\text{‰} = 163,22\text{‰}$), то результат будет меньше истинного суммарного показателя заболеваемости (табл. 56).

Правильное вычисление сумм интенсивных показателей производится следующим образом: сначала суммируются исходные абсолютные числа ($11420 + 1053 = 12473$ и $42 + 168 = 210$), затем на основании этих сумм вычисляется итоговый показатель ($210/124730 \times 1000 \approx 168,4\text{‰}$).

Суммирование интенсивных коэффициентов

Отношение к вакцинации	Число жителей	Заболеваний	
		абс.	на 1000 чел.
Вакцинированные	11420	42	3,7
Невакцинированные	1053	168	159,5
Итого	12473	210	168,4

Экстенсивные коэффициенты отражают структуру, распределение, отношение части статистической совокупности к целой совокупности. То есть отношение (долю, удельный вес) отдельного элемента к итогу. Выражаются только в процентах. *Например:* в структуре инфекционной заболеваемости жителей Владимирского района в 2012 году доля инфекционного гепатита среди всех инфекционных заболеваний определялась как:

$$\frac{\text{число случаев инфекционного гепатита} * 100}{\text{число всех случаев инфекционных заболеваний}}$$

Другой пример: на рис. 42 представлена динамика показателей распределения больных хроническими неспецифическими заболеваниями легких (ХНЗЛ), выявленных по результатам целевого профилактического осмотра. Группа А – здоровые, группа В – лица с начальными, обратимыми признаками заболевания, С – группа сформировавшейся хронической патологии. Из представленных данных наглядно видно, что доля (удельный вес) здоровых мало менялась в разных возрастных группах. В то время как доля группы хронически больных росла в основном за счет группы лиц с начальными, обратимыми формами болезни.

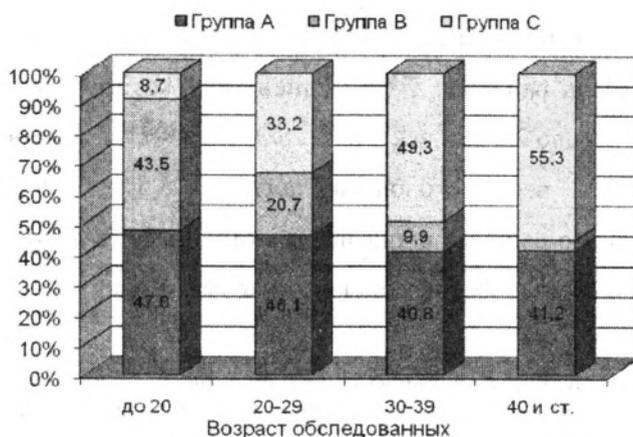


Рис.42. Распределение обследованных работников по группам заболеваемости ХНЗЛ (в % к численности обследованных каждого возраста)

Одной из самых распространенных ошибок, встречающихся в практике статистического анализа, является неправильное использование экстенсивных коэффициентов, когда по ним пытаются судить о величине явления. Например, на основании снижения доли средств консолидированных бюджетов субъектов РФ в структуре расходов на систему здравоохранения за 2001, 2002, 2003, 2004, 2005 и 2006 гг. (54,9%; 51,2%; 50,8%; 50,4%; 43,9% и 40,1%, соответственно) нельзя утверждать, что в этот период расходы на здравоохранение за счет средств консолидированных бюджетов субъектов РФ снижались. На самом деле эти расходы увеличились в 2,5 раза: с 139,9 млрд. руб. до 348,3 млрд. руб., причем инфляция в данном случае не покрывала всего прироста. Снизилась только доля расходов на здравоохранение из консолидированных бюджетов субъектов РФ из-за более интенсивного роста расходов за счет других источников.

Знание реальной ситуации и умение пользоваться интенсивными и экстенсивными относительными величинами – обязательное требование при анализе статистических данных.

Рассмотрим использование интенсивных и экстенсивных коэффициентов на условном примере анализа госпитальной летальности (табл. 57).

Летальность – интенсивный коэффициент³⁰, в котором среда – это раненые, явление – гибель раненых. Из 35 раненых, у которых ранение осложнилось столбняком, умерло 25 человек. Или $(25/35)*100=71,4\%$. Из всех 400 умерших в стационаре *МН* из-за столбняка погибло 6,5% человек (экстенсивный коэффициент). Но раненых с диагнозом «столбняк» в данном стационаре было всего $35/5435*100=0,64\%$ (экстенсивный коэффициент).

Таблица 57

Распределение больных с раневой инфекцией в стационаре *МН*

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

Из представленных данных видно, что чаще всего (интенсивность явления) раненые погибали от столбняка. Но среди всех умерших в стационаре *МН* доля лиц со столбняком не превышала 6,5%.

Таким образом, эти показатели характеризуют близкие, но не однозначные явления: летальность от столбняка, долю столбняка среди причин смерти лечившихся. Среди перечисленных показателей только интенсивный коэффициент летальности свидетельствует об интенсивности, распространенности случаев гибели больных столбняком. И этот показатель можно использовать для сравнения эффективности работы схожих медицинских учреждений с разными потоками больных. Экстенсивные показатели в приведенном примере имеют только местное значение,

³⁰ Иногда возникают трудности с идентификацией вида статистического коэффициента. Следуя логике определения, летальность — интенсивный коэффициент (частота явления в своей среде). С другой стороны, летальность — экстенсивный показатель (доля умерших больных от числа всех больных). Но разница между трактовками этих коэффициентов весьма существенна, т.к. интенсивные коэффициенты дают характеристику интенсивности, частоты, распространенности явления, а экстенсивные — нет. В проблемных ситуациях статистики руководствуются «фельдшерским» правилом: все интенсивные коэффициенты кончатся на ... ость (летальность, заболеваемость, рождаемость и т.п.).

т.к. зависят от состава больных в стационаре *NN*, а этот состав мог определяться не учтенными в примере причинами³¹.

Коэффициенты наглядности используются для облегчения сравнения и повышения наглядности данных. Не меняя по существу отношения между числами, они дают более отчетливое представление о характере изменения явления во времени и пространстве. Выражаются коэффициенты наглядности, как правило, в процентах, которые вычисляют от исходного уровня, принимаемого за 100%. Поэтому эти показатели относятся к группе базовых показателей динамики. За исходный уровень можно выбрать любой уровень, который имеет наибольшее смысловое значение (табл. 16).

При анализе параллельных изменений нескольких явлений исходный уровень может быть принят за 0%. Поскольку коэффициенты наглядности являются неименованными величинами, их удобно использовать для сравнения числовых рядов, которые состоят из разнородных величин (табл. 58 и рис. 10).

Таблица 58

Некоторые характеристики деятельности учреждений здравоохранения Санкт-Петербурга в показателях наглядности (100% – уровень 1992 года)

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

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

³¹ Следует отметить, что при сводке приведенных в примере данных допущена существенная ошибка, способная повлиять на объективность статистической информации: у одного и того же больного могло быть несколько осложнений. Т.е. на практике число лечившихся больных и число осложнений, как правило, не совпадают.

используются в статистике здравоохранения при отсутствии 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 обращений и полового состава жителей.

Из приведенных данных можно косвенным путем сделать заключение о более высоком травматизме женщин, хотя их доля среди всех обратившихся за медицинской помощью по этому поводу была несколько меньше (табл. 59).

Таблица 59

Относительная частота обращений по поводу травм

Пол	Обратились по поводу травм (% к итогу)	Состав населения (% к итогу)	Коэффициенты относительной интенсивности
Женщины	49	40	1,23
Мужчины	51	60	0,85
Итого	100	100	–

Известно, что расширенное употребление алкогольных напитков в СССР было одной из трудноразрешимых социальных проблем. Прямую статистику по этой проблеме не было принято широко публиковать. Коэффициенты относительной интенсивности, рассчитанные по данным открытой печати, позволяли обойти это препятствие.

Результаты статистических расчетов, представленные в табл. 60, характеризуют опережающий рост потребления алкоголя над ростом численности населения СССР в 50-60-е годы прошлого столетия.

Таблица 60

Динамика продажи алкогольных изделий в СССР (в % к 1940 году)

Годы	1958	1964	1965	1966	1967
Объем продажи, %	166,0	261,0	279,0	306,0	351,0
Численность населения, %	105,5	116,5	118,1	119,4	120,7
Коэффициент относительной интенсивности	1,5	2,2	2,3	2,5	2,9

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 1, на 100, на 1000, на 10000 (табл. 61).

Таблица 61

Динамика показателей здравоохранения РФ (РОССТАТ, 2012)

Показатели соотношения	2006	2007	2008	2009	2010	2011
Обеспеченность койками (на 10 тыс. населения)	109,2	107,2	98,6	96,8	87,5	85,6
Обеспеченность врачами всех специальностей (на 10 тыс. населения)	49,4	49,8	49,6	50,1	43,8	44,0
Численность среднего медицинского персонала всех специальностей (на 10 тыс. населения)	108,0	108,6	108,6	106,5	106,9	105,9
Численность среднего медицинского персонала, приходящаяся на 1 врача	2,2	2,2	2,2	2,1	2,5	2,4

4.2. Стандартизация

Существенной проблемой, возникающей при анализе различий общих интенсивных коэффициентов, является необходимость убедиться, что выявленные различия (или их отсутствие) определяются не разным составом наблюдаемых совокупностей. Один из методов статистического анализа, позволяющих ответить на этот вопрос, – стандартизация показателей. С помощью этого метода исследуется гипотетическая ситуация: как изменилась

бы направленность различий сравниваемых величин, если бы исследуемые совокупности наблюдений были бы одинаковы по своему составу, по своей структуре. *Например*, различия коэффициентов смертности населения двух районов области дают возможность судить о последствиях воздействия на здоровье населения неблагоприятной экологической ситуации (если в одном районе этот фактор, безусловно, присутствует, а в другом – нет). Вместе с тем, в данной ситуации использование различий уровней показателей смертности в качестве доказательной базы суждения о существенности фактора экологического неблагополучия требует учета того обстоятельства, что разница показателей смертности может быть обусловлена не наличием или отсутствием экологического фактора, а неодинаковым возрастным составом населения в этих районах.

Еще одним фактором, обуславливающим применение метода стандартизации, является повсеместное отсутствие возможности статистической оценки репрезентативности различий общих интенсивных показателей с помощью традиционных методов параметрической статистики (t критерий Стьюдента и т.п.). В таком случае подтверждение статистической устойчивости соотношения интенсивных показателей после их стандартизации, может служить косвенным свидетельством репрезентативности выявленных различий.

При использовании любых из представленных методов стандартизации необходимо помнить:

- для получения точных результатов число подгрупп в сравниваемых совокупностях должно быть как можно большим;
- стандартизованные коэффициенты могут использоваться весьма ограниченно, т.к. в основе их расчетов лежит гипотетическая ситуация («если бы»);
- необходимо помнить, что многие из исследуемых медицинской статистикой факторов взаимосвязаны между собой опосредованно. *Например*,

известно, что уровень заболеваемости связан со стажем работы. Но стаж работы, в свою очередь, тесно взаимосвязан с возрастом работника. С возрастным фактором тесно связана и длительность проживания в данном месте мигрантов, и т.п.

Поэтому не рекомендуется принимать результаты стандартизации как абсолютное доказательство существенности тех или иных зависимостей, поскольку эти результаты носят условный характер и могут быть итогом действия других, не учтенных в исследовании факторов.

4.2.1. Прямой метод

Если имеются сведения о распределении сравниваемых групп населения по какому-либо структурному признаку (например, по возрасту), а также о распределении единиц наблюдения (случаев смерти, случаев заболеваний и т.п.) по этому же признаку и, следовательно, можно вычислить повозрастные коэффициенты (смертности, заболеваемости и т.п.), то для расчетов избирают *прямой метод стандартизации*.

Последовательность реализации этого метода можно проследить на примере стандартизации показателей распространенности онкологических болезней в одной из областей РФ. В 1996 году интенсивный показатель онкологической заболеваемости населения этой области составил 332,1 случая на 100 тыс. населения. К 2011 году он возрос до 349,2 случая на 100 тыс. населения. За этот же период изменилась и возрастная структура населения: доля лиц в возрасте 50 лет и старше, например, увеличилась с 30,9% до 41,1%. А эта возрастная группа, как известно, характеризуется наиболее существенным «вкладом» в общий уровень онкологической заболеваемости. В то же время доля лиц в возрасте до 39 лет, среди которых распространённость онкологических болезней, как правило, всегда меньше, снизилась с 54,9% в 1996 году до 44,6% в 2011 году.

С целью проверки рабочей гипотезы о том, что рост общих показателей онкологической заболеваемости может быть объяснен старением населения, проведем сравнительный анализ стандартизованных по возрасту показателей онкологической заболеваемости жителей этой области РФ в 2006 и 2011 гг. Для этого:

1. Вычислим повозрастные коэффициенты онкологической заболеваемости населения области в 1996 и 2011 гг. (табл. 62).

Таблица 62

Расчет интенсивных показателей онкологической заболеваемости населения области

Возрастные группы	1996 год			2011 год		
	Численность населения	Число больных	Заболев. на 100 тыс.	Численность населения	Число больных	Заболев. на 100 тыс.
до 30 лет	481 326	95	19,74	357 396	35	9,79
30-39	205 117	147	71,67	165 538	137	82,76
40-49	177 707	454	255,48	167 920	354	210,81
50 и ст.	386 007	3 456	895,32	481 172	3 567	741,31
Итого	1 250 157	4 152	332,12	1 172 026	4 093	349,22

2. Далее выбираем стандарт возрастного состава. Следует отметить, что не существует единого, постоянного стандарта для всех вариантов различий интенсивных показателей общественного здоровья. В данной ситуации для элиминации возрастного фактора за стандарт целесообразно взять или распределение возрастного состава одного из районов области, или усредненное возрастное распределение населения всей области. Допускается использование в качестве стандарта, что наименее предпочтительно, возрастного состава «посторонней» совокупности (состав населения республики или всей страны). Здесь существует определенная закономерность: чем ближе структура стандарта к структуре сравниваемых совокупностей, тем более точными будут итоговые результаты.

3. За стандарт можно принять усредненное по сравниваемым годам возрастное распределение населения области в процентах или долях от единицы (табл. 63).

Расчет стандарта возрастного распределения населения области

Возрастные группы	Численность населения			Возрастной состав (стандарт)
	1996	2011	В среднем, за 1996 и 2011 годы	
до 30 лет	481 326	357 396	419 361,0	0,346
30-39	205 117	165 538	185 327,5	0,153
40-49	177 707	167 920	172 813,5	0,143
50 и ст.	386 007	481 172	433 589,5	0,358
Итого	1 250 157	1 172 026	1 211 091,5	1,000

4. За Вычисленные интенсивные коэффициенты смертности или заболеваемости для каждой возрастной группы населения перемножаются на соответствующие числа возрастного стандарта. *Например*, для группы в возрасте до 30 лет в 1996 году $19,74 \cdot 0,346 = 6,83$; в 2011 году – $9,79 \cdot 0,346 = 3,39$. Для группы в возрасте 30-39 лет в 1996 году $71,67 \cdot 0,153 = 10,97$; в 2011 году – $82,76 \cdot 0,153 = 12,66$ и т.д. Эти повозрастные стандартизованные показатели суммируются по столбцам 1996 и 2011 гг. (табл. 64).

5. Полученные итоги свидетельствуют о том, что если бы возрастной состав населения области в обеих временных точках был бы одинаков и в 2011, и в 1996 гг., то общие показатели смертности были бы выше не в 2011, а в 1996 году. Соответственно 311,58 и 374,85 случаев на 100 тыс. населения (табл. 64).

Таблица 64

Расчет стандартизованных по возрасту интенсивных показателей смертности населения области (на 100 тыс. чел.)

Возрастные группы, лет	Заболеваемость на 100 тыс. населения		Возрастной состав (стандарт)	Показатели заболеваемости при стандартном возрастном составе	
	1996 год	2011 год		1996 год	2011 год
до 30	19,74	9,79	0,346	6,83	3,39
30-39	71,67	82,76	0,153	10,97	12,66
40-49	255,48	210,81	0,143	36,53	30,15
50 и ст.	895,32	741,31	0,358	320,52	265,39
Итого	332,12	349,22	1,000	374,85	311,58

4.2.2. Косвенный метод

Косвенный метод стандартизации применяется, когда нет сведений о структурном распределении сравниваемых совокупностей (случаев заболеваний, случаев смертей и т.д.) на подгруппы по возрасту, стажу работы и т.п. Или когда числа наблюдений в отдельных подгруппах малы и, следовательно, репрезентативность вычисленных интенсивных коэффициентов сомнительна.

В этом случае целесообразно принять за стандарт заведомо репрезентативные интенсивные коэффициенты, вычисленные для аналогичных подгрупп какой-либо определенной (иногда совершенно посторонней) совокупности населения. Итогом расчетов здесь будут «ожидаемые числа» заболевших, умерших и т.п. в сравниваемых совокупностях населения, на основании которых рассчитываются «ожидаемые» интенсивные коэффициенты.

В качестве практического примера рассмотрим стандартизацию по возрасту показателей заболеваемости с временной утратой трудоспособности (ЗВУТ) работников одинаковых профессий, трудившихся в 2012 году в 2-х цехах предприятия. При этом повозрастные показатели ЗВУТ сравниваемых совокупностей (работники 1-го и 2-го цехов) неизвестны, а известно только повозрастное распределение численности работников этих цехов (табл. 65). За стандартное распределение показателей ЗВУТ принимаем данные о ЗВУТ по всему предприятию за 2012 год.

Итоги расчетов стандартизованных коэффициентов ЗВУТ свидетельствуют о том, что если бы повозрастная ЗВУТ в обоих цехах была бы одинаковой, как в целом на предприятии, то общее число случаев ЗВУТ составило бы 260,3 в 1-м цехе и 279,7 – во 2-м цехе. Соответственно, стандартизованные по возрасту уровни ЗВУТ в пересчете на 100 работающих составили в 1-м цехе $260,3/344*100=82,4$, во 2-м – $279,7/461*100=103,2$ случая

за 2012 год. При этом фактический уровень ЗВУТ в 1-м цехе составлял 78,5, а во 2-м – 78,8 случая на 100 работающих за 2012 год.

Нетрудно заметить, что фактические уровни ЗВУТ работников обоих цехов практически не различаются, в то время как стандартизованные показатели заболеваемости во 2-м цехе заметно выше, чем в 1-м. Отсюда вытекает вывод: если бы возрастной состав в обоих цехах был бы одинаков, то заболеваемость работников 2-го цеха была бы выше.

Таблица 65

Вычисление стандартизованных по возрасту показателей заболеваемости с временной утратой трудоспособности (ЗВУТ) работников 2-х цехов предприятия

Возраст работников (лет)	Стандарт ЗВУТ (в случ. за 2012 год на 100 раб. предприятия)	Повозрастная численность работников в 2011 году (абс. числа)		«Ожидаемые» абс. числа случаев ЗВУТ в 2012 году	
		Цех 1	Цех 2	Цех 1	Цех 2
до 25	18,0	19	84	3,4	15,1
25-29	47,0	28	95	13,2	44,7
30-34	51,0	51	77	26,0	39,3
35-39	86,3	74	69	63,9	59,5
40-44	87,4	83	81	72,5	70,8
45 и ст.	91,4	89	55	81,3	50,3
Всего	79,4	344	461	260,3	279,7

4.2.3. Обратный метод

Обратный метод в отличие от прямого и косвенного методов стандартизации не требует знания структуры сравниваемых совокупностей. Основой расчетов здесь является знание общей численности сравниваемых совокупностей, а также структуры заболевших, умерших и т.п. В качестве стандарта применяются известные структурные показатели повозрастной смертности, заболеваемости и т.п. Для примера рассмотрим стандартизацию показателей первичной патологической пораженности (число заболеваний, впервые выявленных на профилактических медицинских осмотрах, в пересчете на 100 чел.) в двух статистических совокупностях: работников отдела логистики и транспорта и работников основных профессий одного из

предприятий Санкт-Петербурга. Фактические данные, полученные по результатам углубленных профилактических медицинских осмотров, свидетельствуют о фактическом равенстве заболеваемости работников отдела логистики и транспорта и заболеваемости работников основных профессий. Соответственно, 362,09 и 362,14 случая первичных заболеваний на 100 осмотренных. Поскольку условия и характер труда работников отдела логистики и транспорта более благоприятны, чем у работников основных профессий, была выдвинута рабочая гипотеза о том, что на уровне заболеваемости работников этих проф. групп влиял их неодинаковый возрастной состав.

С целью проверки выдвинутой гипотезы было проведено вычисление обратным методом стандартизованных по возрасту показателей заболеваемости (патологической пораженности). Для этих вычислений необходима информация о числах больных в каждой возрастной группе изучаемых совокупностей и стандартное распределение повозрастных показателей заболеваемости. Здесь в качестве стандарта могут использоваться данные повозрастного распределения показателей заболеваемости (первичной патологической пораженности) по РФ, по области, по районам, по предприятию в целом и т.п. В данном случае за стандарт приняты повозрастные показатели первичной пораженности, полученные в целом по предприятию (табл. 66).

Расчет значений в столбце «ожидаемая» численность работников производился следующим образом: для возрастной подгруппы «до 30 лет» отдела логистики и транспорта $45/188,0 \cdot 100 = 24$, для подгруппы «30-39 лет» – $78/208,8 \cdot 100 = 37$ и т.д. Соответственно для возрастной подгруппы «до 30 лет» у работников основных профессий $382/188,0 \cdot 100 = 203$, для подгруппы «30-39 лет» – $532/208,8 \cdot 100 = 252$ и т.д. для остальных возрастных групп в обеих совокупностях.

Вычисление стандартизованных коэффициентов распространенности выявленных на профилактических мед. осмотрах заболеваний у работников отдела логистики и транспорта и работников основных профессиональных групп предприятия

Возраст (лет)	Повозрастные коэффициенты первичной выявляемости заболеваний	Логистика и транспорт		Основные проф. группы	
		Число выявленных болезней	«ожидаемая» численность работников	Число выявленных болезней	«ожидаемая» численность работников
до 30	188,0	45	24	382	203
30-39	208,8	78	37	532	255
40-49	308,1	262	85	1464	475
50-59	419,2	393	94	2641	630
60 и ст.	557,2	150	27	747	134
Итого	338,0	928	267	5 766	1 697

В итоге «ожидаемая» численность работников отдела логистики и транспорта оказалась равной 267 чел., а численность работников основных профессий 1697 чел., в то время как фактическая численность, соответственно, 257 и 1637 чел. Для получения стандартизованных показателей заболеваемости работников отдела логистики и транспорта умножаем принятый за стандарт общий коэффициент первичной заболеваемости (338,0) на соотношение «ожидаемой» численности работников к фактической: $338,0 \cdot (267/257) = 351,15$. Для работников основных профессий – $338,0 \cdot (1697/1637) = 350,39$. Поскольку практическое отсутствие различия уровней общей заболеваемости обследованных подразделений предприятия сохранилось и после стандартизации, схожесть заболеваемости обследованных проф. групп работников не может объясняться возрастным фактором.

5. Показатели описательной статистики

Одна из основных задач математической статистики вытекает из общей методологии статистического анализа – изучение явлений в массе наблюдений. Таким образом, перед исследователем всегда автоматически возникает проблема определения характеристик этой массы наблюдений, предметно называемой *статистической совокупностью*. Такого рода характеристики, называемые показателями описательной статистики, описывают различные свойства статистической совокупности (табл. 67).

Таблица 67

Показатели описательной статистики³²

Показатели	Назначение	Примеры показателей
Средние величины	Описывают положение середины распределения	Степенные средние: - среднее арифметическое; - среднее геометрическое; - среднее гармоническое; - среднее квадратическое; - среднее кубическое; - среднекубическое. Структурные средние: - мода; - медиана
Показатели разброса	Описывают степень разброса (вариабельности, изменчивости) данных	Лимит; амплитуда; дисперсия; среднеквадратическое отклонение; коэффициент вариации; квантили
Показатели формы, концентрации распределения	Отвечают на вопрос об островершинности и симметричности распределения данных около центра	Коэффициент асимметрии; эксцесс; гистограмма; полигон распределения

Знание этих показателей распределения делает возможным последовательное решение следующих вопросов:

³² Практическое вычисление этого комплекса показателей см. в разделе «Показатели описательной статистики в MS Excel» настоящего издания.

- установления закона распределения эмпирических (полученных опытным путем) статистических совокупностей и параметров этого распределения (свойств эмпирических совокупностей);
- оценки причинно-следственных отношений между явлениями на основе оценки взаимосвязи распределений в статистических совокупностях, описывающих эти явления;
- решения проблем, связанных с репрезентативностью (представительностью) выборочных исследований, т.е. соответствия распределений выборочных и генеральных совокупностей, что и определяет в конечном счете точность статистического прогноза.

5.1. Ряды распределения. Вариационные ряды

Отдельные учетные признаки в независимости от того, качественные они или количественные, в любой статистической совокупности могут принимать различные значения. *Например:* в качестве статистической совокупности рассматривается группа студентов. Каждый студент – отдельная единица наблюдения, а учетным признаком принимается, скажем, масса тела. Понятно, что величина массы тела колеблется, варьирует от одного студента к другому. У первого студента масса тела 51 кг, у второго – 67 кг и т.д. Эти варьирующие значения массы тела (учетного признака) называются *вариантами наблюдений*, или просто *варианты*.

Количественные 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 промежуточных значений. *Например*, число ударов пульса, число дней госпитализации и т.п. Прерывный признак всегда является результатом счета.

Непрерывный – это признак, варианты которого 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 многих других вспомогательных операций почти полностью исключено. Однако базовые понятия и термины, используемые в процессе статистического анализа, могут быть поняты только на основе усвоения логической последовательности основных, «ручных» операций статистической обработки данных.

При наблюдении какого-либо варьирующего признака ведется регистрация полученных значений. *Например*, зарегистрированные значения массы тела у обследованных студентов составили: 64, 57, 63, 57, 61, 61, 59, 60, 60, 63, 59, 62, 59, 64, 60, 59, 60, 60, 60, 60, 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. 43).

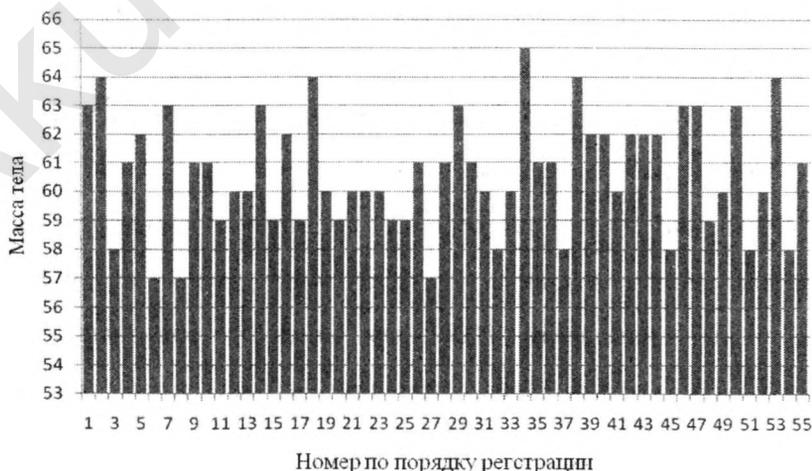
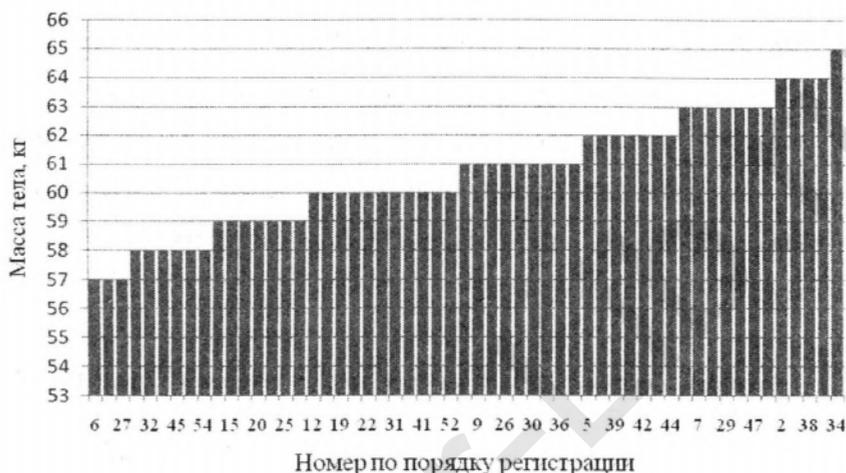


Рис.43. Неупорядоченный ряд данных

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, 61, 62, 62, 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 65 (рис. 44).



определенным интервалом. Если в вариационном ряду все частоты равны 1, т.е. каждая варианта встречается только 1 раз, то такой ряд называют *простым вариационным рядом*. Если в ряду есть частоты, не равные 1, то вариационный ряд называется *взвешенным*.

Распределение частот может быть представлено либо абсолютными, либо относительными величинами. Иногда используют термин «показатели частот», которые показывают долю частот отдельных вариантов (отдельных групп вариант) от общего числа наблюдений. Выражаются они в относительных единицах – процентах от общего числа наблюдений или долях от единицы. Например, лица с массой тела 60 кг составляли 20% от всей совокупности (табл. 68 и рис. 45).

Таблица 68

Распределение обследованных студентов с учетом массы тела (кг)

Вес, кг (<i>V</i>)	Частоты <i>P</i>		Частости		
	Число студентов (частоты)	Накопленные частоты	%	<i>P</i>	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
	$N=\sum P=53$	–	100,0	1,00	–

Для повышения наглядности распределения бывает уместным отобразить ряд накопленных частот или частостей. *Накопленные частоты и частости* позволяют при оценке распределений игнорировать неравную величину интервала в отдельных группах (рис. 45).

Если ряд распределения состоит из вариантов, представленных в виде *дискретных величин*, то он называется *дискретным вариационным рядом*. Графически дискретный вариационный ряд изображается как многоугольник, называемый *полигоном распределения* (рис. 46).

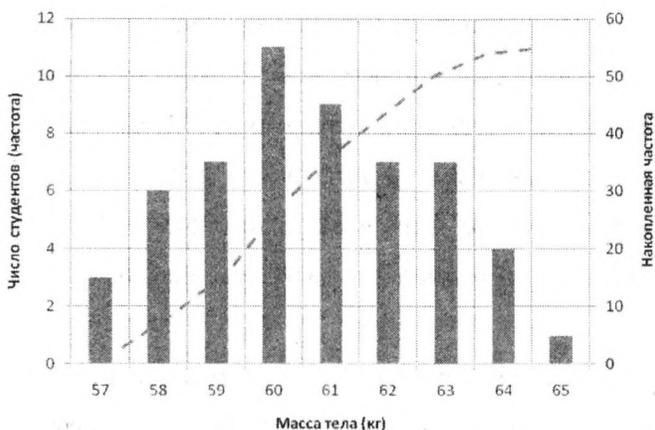


Рис.45. Распределение обследованных студентов с учетом массы тела (кг)

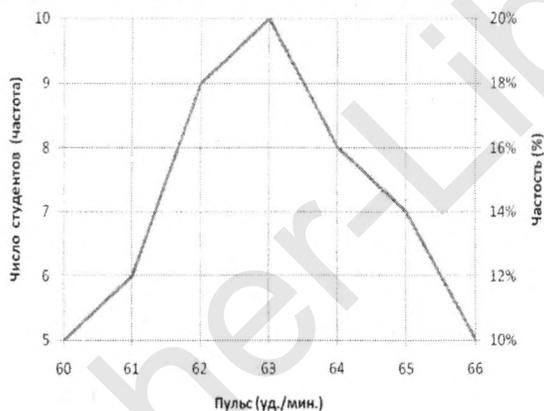


Рис.46. Полигон распределения обследованных студентов по частоте пульса

Приведенные формы ряда распределения применимы лишь для тех случаев, когда дискретный варьирующий признак принимает небольшое количество значений. Если таких вариантов 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)*, то есть групп, имеющих определенные пределы значений. Эти группы образуют *интервальный сгруппированный вариационный ряд*. Графически такой ряд изображается **гистограммой** распределения (рис. 47).

Следует обратить внимание, что гистограмма и столбиковая диаграмма не одно и то же. Здесь следует помнить, что одним из параметров распределения является площадь многоугольника, которую очерчивает гистограмма. Поэтому для нее более оптимальным представляется такое изображение, в котором исключены межгрупповые интервалы (рис. 45), затрудняющие зрительное восприятие размеров суммарной площади, особенно при неравных интервалах группировки.

Всякая сводка или группировка уничтожают особенности отдельных единиц наблюдения, размывая их в группе. Правильная группировка является необходимым условием сохранения основных тенденций распределения признака и, соответственно, залогом успешного решения задач любого статистического анализа.

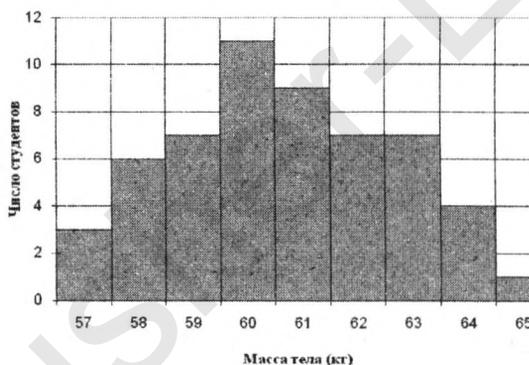


Рис.47. Гистограмма распределения студентов по массе тела

В группах пределы обозначаются или подразумеваются «от» (верхняя граница) и «до» (нижняя граница). Желательно, чтобы интервалы во всех группах конкретного ряда были одинаковы, т.к. ряды распределений, где группировка данных проведена в неодинаковых интервалах, требуют применения специальных методик 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 нехарактерные, случайные колебания были бы сглажены. В практической статистике принято считать, что лучше допустить некоторую потерю в точности, но зато выиграть в наглядности распределений, в аналитических возможностях их обработки.

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 несколько 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. Последовательность операций определения интервала группировки, в этом случае выглядит следующим образом:

- найти разность между максимальным и минимальным значением вариант в ряду и разделить на число групп, которое хотят получить;
- полученную в результате деления величину округлить и таким образом получить интервал.

При незначительном разбросе вариант для определения оптимального интервала группировки можно воспользоваться формулой Стерджесса:

$$i = \frac{V_{\max} - V_{\min}}{1 + 3,322 \text{Lg}(n)}$$

где n – число наблюдений, V_{\max} и V_{\min} – соответственно, минимальное и максимальное значения вариант. Для подобных целей можно использовать и формулу, основанную на рекомендациях К. Брукса и Н. Краузерса:

$$i = \frac{V_{\max} - V_{\min}}{5 \times \text{Lg}(n)}$$

Интервалы могут быть *открытыми*. Такие интервалы имеют одну границу: либо верхнюю, либо нижнюю. *Например*, «от 100 лет и более» или «до 10 лет!» Необходимо помнить, что открытые интервалы существенно влияют на выбор методов статистической обработки данных, сгруппированных таким образом. *Например*, в открытых рядах, т.е. там, где используются открытые интервалы, невозможно вычислить среднее арифметическое.

Закрытые интервалы имеют обе конечные границы, нижнюю и верхнюю. Соответственно, формируются *открытые или закрытые вариационные ряды* (табл. 69). Чтобы не возникало сомнений, в какую группу относится та или иная варианта, границы интервала (границы групп) не должны пересекаться, то есть 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 такого рода неопределенности делаются специальные оговорки («от» и «до»). Оговорки однозначно указывают, в какую группу попадают пограничные значения вариант.

Таблица 69

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 непонятно, куда отнести варианту 80 при групповых границах 75-80 и 80-85, то в обе группы добавляются по 0,5 частоты. Вариант как бы делится между двумя группами (табл. 70).

Распределение частот вариант при совпадении границ группировок

V	P ₁	P ₂
75-80	5	5,5
80-84	7	6,5
Итого	12	12

Когда вариационный ряд сгруппирован, необходимо *определить середину* в каждой группе ряда. Если группа состоит из дискретных величин, то середина группы определяется, как полусумма крайних значений группы. Если группа представлена непрерывными величинами, то середина группы определяется, как полусумма начальных вариант данной и последующей групп (табл. 71).

Таблица 71

Примеры определения середины групп

Дискретный ряд		Непрерывный ряд	
Исходные группы	Середина группы	Исходные группы	Середина группы
V	V _{ср}	V	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* (табл. 72).

Различные варианты группировок
интервального ряда распределения женщин по росту

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

Рост (см) V	Частота P
151	1
152	2
153	3
154	4
155	5
156	7
157	9
158	4
159	3

Рост (см) V	Частота P
151-153	6
154-156	16
157-159	16

Группировка, в которой величины интервалов завышены, приводит к образованию крупных групп, в которых основные тенденции распределения теряются. Мелкая группировка перегружает распределение частными деталями, не отражающими основных тенденций, что неизбежно затрудняет понимание характера вариации. Кроме того, во многих ситуациях высокая точность исходных данных не имеет практического смысла (рис. 48).

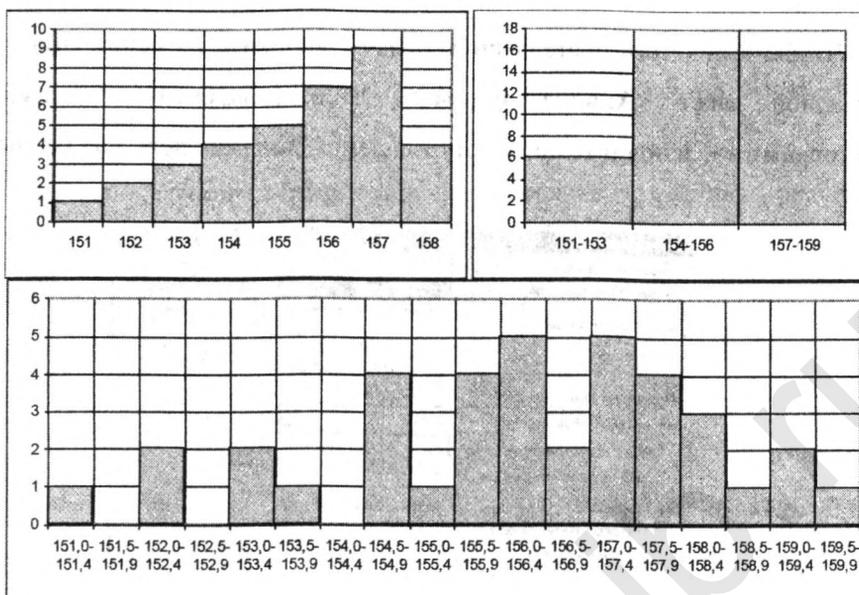


Рис.48. Варианты диаграмм по итогам различных группировок в табл. 72

5.1.1. Построение вариационных рядов в MS Excel

В качестве примера создадим таблицу с результатами обработки измерений массы тела студентов. Для этого введите на лист исходные данные (рис. 49).

	A	B	C	D	E	F	G	H	I	J
1	64	63	60	65	62	60				57
2	57	59	63	61	58	64				58
3	63	62	60	61	63	63				59
4	62	59	59	58	63	58				60
5	57	64	59	64	59	61				61
6	61	60	64	62	60					62
7	61	59	61	62	58					63
8	59	60	58	60	63					64
9	60	60	61	62	58					65
10	60	60	61	62	57					
11										
12										

Рис.49. Расположение исходных данных для построения вариационного ряда в MS Excel

Чтобы выполнить поставленную задачу, вызовите из меню <Данные>³³ прикладной пакет <Анализ данных>. В открывшемся окне выберите <Гистограмма>. Заполните поля в открывшемся окне, как показано на рис. 50.

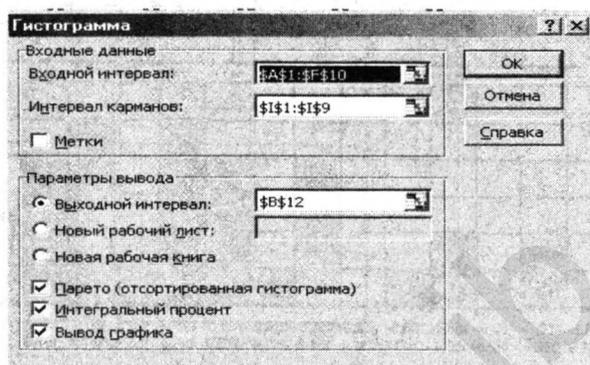


Рис.50. Вариант заполнения окна «Гистограмма»

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 и 65 были введены заранее, последовательно в клетки первого листа (от П до I9). Благодаря этому в процессе группировки все варианты наблюдения с весом 57 кг будут отнесены в группу 57. Все варианты с весом 58 кг – 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.

³³ В более ранних версиях MS Excel, пакет <Анализ данных> располагается в меню <Сервис>.

4. В разделе «**Параметры вывода**» указывается место, куда будет выводиться результат. В этом же разделе указываются интервалы ячеек, в которых будут располагаться результаты.

5. **Парето** – группы сортируются не в порядке возрастания значений вариант, а в порядке роста частот.

6. **Интегральный процент** – накопленные частоты, выраженные в %.

7. **Вывод графика** – диаграмма распределения анализируемого ряда значений. Обратите внимание! На диаграмме варианты (группы вариант) по оси X располагаются в ранжированном по частоте порядке – от большей частоты встречаемости к меньшей. Т.е. первая группа вариант (60) – это варианты с наибольшей частотой (11), затем идут варианты (61) с меньшей частотой (9) и т.д. (рис. 51).

Карман	Частота	Интегральный %	Карман	Частота	Интегральный %
57	3	5,45%	60	11	20,00%
58	6	16,36%	61	8	34,55%
59	7	29,09%	59	7	47,27%
60	11	49,09%	62	7	60,00%
61	8	63,64%	63	7	72,73%
62	7	76,36%	58	6	83,64%
63	7	89,09%	64	5	92,73%
64	5	98,18%	57	3	98,18%
65	1	100,00%	65	1	100,00%
Еще	0	100,00%	Еще	0	100,00%

Рис.51. Результат построения вариационного ряда в Excel

Рядом выводится диаграмма полученного распределения (рис. 52). Внешний вид гистограммы и отдельные элементы экспликации дорабатываются пользователем в соответствии с его предпочтениями с помощью обычных приемов обработки графических изображений в Excel.

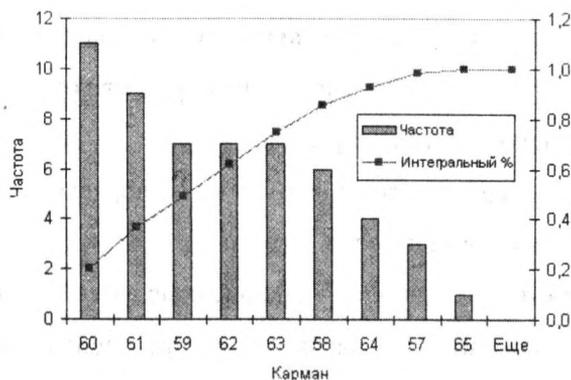


Рис. 52. Диаграмма распределения параметров вариационного ряда, сформированного в Excel (см. рис. 51)

5.2. Показатели центра распределения. Средние величины

Важным свойством статистической совокупности является положение центра распределения. На рис. 53 видно, что, несмотря на незначительную разницу объемов наблюдений в 1 и 2 группах подростков (28 чел. в 1-й группе и 33 – во 2-й), полигоны распределения имеют существенные различия: визуально сдвинуты один от другого, а также имеют другие особенности. Одной из этих особенностей является разное положение центров распределений, которые характеризуются средними величинами. По определению, средняя величина выражает характерную, типичную для данного ряда величину признака. Эта величина образуется в данных условиях места и времени под воздействием совокупности действующих факторов и является усредненной равнодействующей. В средней величине погашаются индивидуальные различия отдельных единиц наблюдения (вариант), обусловленные случайными, приходящими обстоятельствами. Это позволяет выявить закономерности, обусловленные общими для всей совокупности данных причинами. Собственно говоря, обобщающая функция – одна из главных практических функций средних. Отсюда следует, что средняя величина, отражая общие свойства, присущие всей совокупности данных, одним числом характеризует сколь угодно большое количество

индивидуальных значений признака. Т.е. средняя масса тела, вычисленная после обследования 100 человек или 100 тыс. человек, выражается одним числом, одной единицей – средним весом.

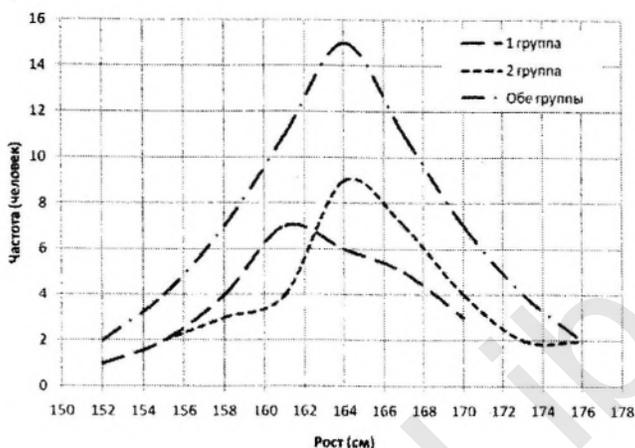


Рис.53. Полигоны распределения 2-х групп подростков по росту

Традиционно подразумевается, что средние величины – это различные виды среднего арифметического. Соответственно, считается, что они достаточно хорошо известны любому человеку, чей образовательный уровень не ниже начального. Однако в статистике выделяют принципиально различающиеся по информативности виды средних величин. В том числе *степенные* и *структурные* средние.

5.3. Степенные средние

К ним, в частности, помимо среднего арифметического, относят **среднее геометрическое, среднее гармоническое и среднее квадратическое** (табл. 73). Называются они степенными потому, что имеют единое алгебраическое выражение, отличающееся только коэффициентом (показателем степени) k , который является коэффициентом статистической размерности признака

$M = \left(\frac{\sum V_i^k}{n} \right)^{\frac{1}{k}}$. Разница между этими средними величинами прямо

пропорциональна вариабельности и обратно пропорциональна числу наблюдений. При низкой вариабельности вариант и достаточно большом числе наблюдений эта разница вообще может отсутствовать.

К сожалению, кроме среднего арифметического, остальные степенные средние крайне редко применяются в практической медицинской статистике, что является источником большого числа ошибок, особенно в интерпретации данных лабораторных исследований, для которых характерны существенная вариабельность данных и малое число наблюдений. В определенной мере способствуют правильному применению степенных средних сопроводительные описания разного рода лабораторного оборудования и методики обработки, получаемых с его помощью результатов лабораторных исследований, если они представляются в виде такого рода средних.

Таблица 73

Виды степенных средних величин

Степенные средние, k	Алгебраическое выражение	
	простое	взвешенное
Гармоническая, $k = -1$	$M = \frac{n}{\sum \frac{1}{V_i}}$	$M = \frac{\sum P_i \cdot V_i}{\sum \frac{P_i \cdot V_i}{V_i}}$
Геометрическая, $k = 0$	$M = \sqrt[n]{V_1 \cdot V_2 \cdot V_3 \dots V_n}$	$M = \sqrt[n]{V_1^{P_1} \cdot V_2^{P_2} \cdot V_3^{P_3} \dots V_n^{P_n}}$
Арифметическая, $k = 1$	$M = \frac{\sum V_i}{n}$	$M = \frac{\sum P_i \cdot V_i}{\sum P_i}$
Квадратическая, $k = 2$	$M = \sqrt{\frac{\sum V_i^2}{n}}$	$M = \sqrt{\frac{\sum P_i \cdot V_i^2}{\sum P_i}}$
Кубическая, $k = 3$	$M = \sqrt[3]{\frac{\sum V_i^3}{n}}$	$M = \sqrt[3]{\frac{\sum P_i \cdot V_i^3}{\sum P_i}}$

В приведенных формулах: k – показатель степени (коэффициент статистической размерности); n – число наблюдений; V – варианты. Если формулы предназначены для использования во взвешенных рядах, т.е. тогда, когда частоты хотя бы в одном наблюдении не равны 1, то в формулы вводится P .

Среднее гармоническое применяется, когда дело имеют с *обратными величинами* (коли-индексы), сложными абсолютными величинами (тонна-километр, килограмм на метр) и т.п. Использование в этих случаях обычного

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 будем считать, что общее число мужчин и женщин одинаково. Среднее арифметическое двух показателей (для мужчин и женщин) $(25+100)/2=62,5$, т.е. каждый из 62,5 чел. был посетителем психиатра, 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, проживавших в районе, 1 был посетителем психиатра. Этот результат мог быть получен и через статистические коэффициенты (интенсивные показатели). Но в данной ситуации вычисление среднего гармонического значительно проще, а результат – точнее.

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 имеет с числовым рядом, отдельные значения в котором распределяются в геометрической прогрессии (резко отличаются друг от друга). Целесообразно вычисление этого показателя и 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 некоего Нощолини и Галилея (XVII век). Лошадь, стоящая 100 крон, одним лицом оценивается в 10 крон, другим – в 1000 крон. Если рассматривать вопрос с арифметической точки зрения, то получим в одном случае ошибку в 90 крон, а во втором – в 900 крон. Если оценивать, во сколько раз ошиблись покупатели, то получим одинаковый ответ для обоих – в 10 раз³⁴.

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 представлены единицами объемов (объемы плазмы, крови, объемы эритроцитарной массы и т.п.).

Таким образом, при вычислении среднего необходимо принимать во внимание фактический состав исходных данных.

5.3.1. Среднее арифметическое. Статистическое взвешивание.

Суммирование вариационных рядов и средних

Среди степенных средних наиболее употребительным является среднее арифметическое (англ. – «mean»). Среднее арифметическое может обозначаться различным символом (M , A , X , \bar{A} и др.). В медицинской статистике чаще всего для его обозначения применяется символ M (от латинского «*media*» – *середина*).

Поскольку в подавляющем большинстве практических исследований решаются задачи исследования выборочных наблюдений, такого рода средние являются, по сути, *выборочными средними*. В качестве среднего

³⁴ Пример взят из книги Ц.Б.Урланиса «Общая теория статистики», 1962.

арифметического генеральной совокупности используется величина, называемая *математическим ожиданием*. Для ее обозначения обычно применяется греческая буква μ . В терминах теории вероятности μ – есть *вероятностное среднее* или *математическое ожидание* случайной величины. Т.е. фактически разница между математическим ожиданием и выборочным средним арифметическим состоит в том, что математическое ожидание является ненаблюдаемым параметром генеральной совокупности, а выборочным средним – результат непосредственного наблюдения. Не вдаваясь в детали методологии статистических оценок параметров генеральной совокупности с позиций теории вероятности, можно отметить, что выборочное среднее используется для оценки неизвестного математического ожидания.

Практически выборочное среднее арифметическое вычисляется следующим образом. Для простого среднего арифметического, т.е. среднего, вычисляемого в простом вариационном ряду, используется формула:

$$M = \frac{1}{n} \sum_{i=1}^{j=n} V_i. \text{ Или в сокращенном виде: } M = \frac{\sum V_i}{n}, \text{ где } n \text{ – число наблюдений,}$$

V_i – исходные числовые данные, называемые вариантами: $(V_1, V_2, V_3, V_4 \dots V_n)$. В основе этого хорошо известного алгоритма вычисления простого среднего арифметического лежат две операции: сложение значений всех вариантов и деление полученной суммы на число наблюдений.

В практике проведения углубленного статистического анализа при изучении причин более высокого или, наоборот, более низкого уровня изучаемого явления применяется *среднее прогрессивное* и *среднее регрессивное*.

Среднее прогрессивное рассчитывается из вариант, чье значение превышает величину среднего арифметического, а *среднее регрессивное* – величина, рассчитанная из вариант, каждая из которых меньше среднего арифметического всего вариационного ряда.

Например, при изучении средней длительности приема амбулаторных больных в городской районной поликлинике были получены следующие результаты. Средняя длительность приема врачом общей практики одного больного составила 15 минут. При этом в группе, где средняя длительность одного приема была ниже среднего уровня (15 мин.), средняя (регрессивная) длительность составляла 12,5 минуты. В группе, где средняя длительность приема была больше среднего уровня (15 мин.), средняя (прогрессивная) длительность приема одного больного составляла 17,5 минуты (табл. 74). Таким образом, разница в длительности приема больных в 1 и 2 группах составляла в среднем $17,5 - 12,5 = 5$ минут.

Таблица 74

**Примеры расчета простой,
регрессивной и прогрессивной средних величин**

Расчеты средних величин		
Простая	Регрессивная	Прогрессивная
V_i (мин.)	$V_i < M$	$V_i > M$
11	11	
12	12	
13	13	
14	14	
15		
16		16
17		17
18		18
19		19
$M=15,0$	$M_{рег} 12,5$	$M_{прог}=17,5$

Вычисления средних прогрессивных и регрессивных величин проводятся как в простых вариационных, так и во взвешенных, сгруппированных рядах. Вычисление на основе взвешенного ряда *взвешенного среднего арифметического* осуществляется тогда, когда наблюдений достаточно много и

они представлены в виде табличных данных. При расчетах здесь учитываются не только значения вариантов, но и частоты (повторяемость) одинаковых значений вариантов. Понятно, что чем больше частота повторений той или иной варианты (группы вариант), тем большую роль, *большой вес* она имеет в статистической совокупности. Среднее арифметическое, рассчитанное в таком 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 ряду определяется, как сумма частот $n = \sum P_i$. Соответственно, формулу для вычисления среднего можно представить в виде алгебраического аналога:

$$M = \frac{\sum V_i P_i}{\sum P_i}.$$

При практическом вычислении среднего взвешенного последовательно выполняются следующие операции.

Каждая варианта в таком ряду умножается на частоту ее встречаемости, как бы «взвешивается» ($V_1 \times P_1, V_2 \times P_2, V_3 \times P_3 \dots V_n \times P_n$). Когда среднее арифметическое определяется в интервальном ряду, где варианты разбиты на интервальные группы (*например*, 15-19, 20-24, 25-29 и т.д.), частоты перемножаются на срединные значения этих групп. Полученные произведения суммируются.

Сумма произведений делится на число наблюдений, в результате чего получается среднее арифметическое (табл. 75).

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 $M = A + \frac{\sum d_i P_i}{\sum P_i} \cdot h$			
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
$\Sigma V=85$	n=5	$\Sigma P=15$		$\Sigma VP=258$	$\Sigma P=15$		$\Sigma PD=3$	
$M=85/5=17$		$M=258/15=17,2$			$M=17+(3/15) \times 1=17,2$			

Упрощенным вариантом ручного вычисления среднего арифметического в сгруппированном ряду является вычисление по способу моментов. Не вдаваясь в математическое обоснование этого способа, здесь можно выделить следующие этапы вычисления (табл.75):

- 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);

- выставляются *условные отклонения* d. Их абсолютные значения последовательно увеличивают на единицу, начиная от 0, который соответствует варианту, принятой за условное среднее. Знак минус обозначает уменьшение вариант от условного среднего. Плюс – соответствующее увеличение вариант;

- произведения условных отклонений на соответствующие им частоты (Pd) суммируются с учетом отрицательных знаков (ΣPd);

- 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); $A + \frac{\sum Pd}{n} h$.

Иногда встречается ситуация, когда необходимо вычислить суммарное среднее или «среднее из нескольких средних», – в ситуации неравных средних и неравных по численности исходных групп данных. В этом случае среднее арифметическое вычисляют, рассматривая каждую группу как самостоятельную совокупность. В каждой из этих групп сначала вычисляется своё среднее. Затем на основе этих данных определяют общее среднее, учитывая число наблюдений в каждой группе $M_{общ} = \frac{\sum M_j n_j}{\sum n_j}$. Для наглядности рассмотрим пример табл. 76:

Таблица 76

Вычисление средних арифметических двух групп

Варианты V_1	Частоты P_1	$V_1 P_1$	Варианты V_2	Частоты P_2	$V_2 P_2$
2	6	12	3	26	78
3	7	21	6	30	180
5	5	25	9	29	261
$n_1 = \sum P_1 = 18$		58	$n_2 = \sum P_2 = 85$		519
$M_1 = 58/18 = 3,2$			$M_2 = 6,1$		

Порядок вычисления общего среднего арифметического может быть одним из представленных способов.

Один способ целесообразно применять, когда известны исходные варианты и их частоты. В этом случае вычисляем сумму произведений вариант и соответствующих им частот $= 12 + 21 + 25 + 78 + 180 + 216 = 577$. Затем находим сумму наблюдений в обеих группах $n_1 + n_2 = 18 + 85 = 103$. После чего вычисляем общее среднее арифметическое $577/103 \approx 5,6$.

Другой способ применяется, когда неизвестны исходные варианты и их частоты. Здесь вычисляем сумму произведений средних первой и второй групп на соответствующее им число наблюдений: $18 \times 3,2 + 85 \times 6,1 = 577$. Затем находим

сумму наблюдений в обеих группах $n_1+n_2=18+85=103$. После чего вычисляем общее среднее арифметическое $576,1/103 \approx 5,6$.

Нетрудно заметить, что попытка вычислить общее среднее прямым способом приведет к ошибочному результату: $\frac{3,2 + 6,1}{2} = 4,7$.

Таким образом, *общее среднее равно среднему арифметическому исходных средних, взвешенных по объемам наблюдений*. Для наглядности рассмотрим нередко возникающую в клинической и лабораторной практике задачу получения смеси 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 подобных задач состоит в применении правила суммирования взвешенных величин. Исходные данные в виде сгруппированного ряда будут выглядеть так (табл. 77):

Таблица 77

Исходные данные для вычисления средней концентрации спирта

Крепость спирта	Вес (кг)
40	10
70	30
96	50

Вычисление взвешенного среднего арифметического и будет ответом на поставленную задачу. $M = (40 \times 10 + 70 \times 30 + 96 \times 50) / (10 + 30 + 50) = 81,1$. Путем несложного алгебраического преобразования можно решать и обратные задачи. *Например*, сколько и какой крепости спирта надо добавить, чтобы получить необходимую концентрацию в заданном количестве?

Иногда требуется определять не общие средние арифметические нескольких массивов наблюдений, а усредненные значения наблюдений в нескольких вариационных рядах. Эта задача обычно возникает, когда ряды наблюдений отличаются друг от друга по симметричности распределения и (или) по граничным значениям вариант, а также когда наблюдения представлены в виде качественных признаков, что делает вычисление общего

среднего арифметического в принципе невозможным, а наиболее типичной суммарной характеристикой центра распределения является структурная средняя – мода. Кроме того, усреднение наблюдений, сгруппированных изначально в несколько вариационных рядов, требуется в ситуациях, когда программой исследования предусмотрено получение целого комплекса статистических характеристик, так называемых показателей описательной статистики.

Например, требуется построить суммарный взвешенный вариационный ряд распределения 2-х групп подростков по результатам измерения роста (табл. 78). Для этого:

- вычисляется среднее значение частоты встречаемости каждой варианты;
- находятся произведения каждой варианты на соответствующую ей среднюю частоту;
- для получения общего среднего арифметического сумма этих произведений делится на сумму усредненных частот: $M_{cp} = 4741/29,0 = 163,5$.

Таблица 78

Суммирование взвешенных вариационных рядов

Рост (см) V	Число наблюдений (частоты)			Средние частоты P_{cp}	V*P _{cp}
	P ₁ 1 группа	P ₂ 2 группа	P ₁ +P ₂		
152	1	*	1	0,5	76,0
155	2	2	4	2,0	310,0
158	4	3	7	3,5	553,0
161	7	4	12	6,0	966,0
164	6	9	15	7,5	1230,0
167	5	7	10	5,0	835,0
170	3	4	6	3,0	510,0
173	–	2	2	1,0	173,0
176	–	2	1	0,5	88,0
Σ	27	31	58	29,0	4741,0

Таким образом, при вычислении любого обобщенного показателя, будь то среднее в отдельном взвешенном вариационном ряду или групповое среднее, должны обязательно учитываться весовые значения (частоты встречаемости)

отдельных вариантов. Собственно говоря, именно поэтому такие средние величины и называются взвешенными. При таких вычислениях абсолютные значения частот могут заменяться их процентным выражением (частостями), то есть 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 жителей города NN больных бронхиальной астмой, лиц, нуждавшихся в госпитализации в течение года, в одном районе (I) было 24%, в другом (II) – 16%, а в третьем (III) – 11%. Нужно найти средний процент (экстенсивный показатель) больных астмой, нуждавшихся в госпитализации.

Самый простой вариант расчета $\frac{24\%+16\%+11\%}{3}=17\%$. Однако при этом будет допущена ошибка, поскольку при таком расчете не учитывается фактическое число больных астмой в каждом из трех районов города. Для того, чтобы избежать неточности, нужно взвесить процент нуждающихся в госпитализации по численности известных больных (табл. 79).

Таблица 79

Вычисление средневзвешенного процента больных бронхиальной астмой (БА), нуждающихся в госпитализации

Район	Больных БА, нуждавшихся в госпитализации		Общая численность известных больных БА
	%	Абс.	Абс.
I	24	77	$77/24 \times 100 = 328$
II	16	166	$166/16 \times 100 = 1038$
III	11	296	$296/11 \times 100 = 2691$
Всего	–	539	4048
Средний взвешенный процент госпитализации $539/4048 \times 100 = 13,3\%$			

Как свидетельствуют полученные результаты, средневзвешенный процент больных бронхиальной астмой, нуждавшихся в госпитализации, (13,3%) был существенно ниже показателя, вычисленного прямым способом (17%). В абсолютных числах эта разница составила $4048 \cdot 0,17 - 4048 \cdot 0,133 \approx 149$ больных бронхиальной астмой.

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 можно толковать, как среднее число обращений, приходящееся на 1000 человек населения. Или показатель соотношения 2,9 штатные единицы среднего медперсонала на 1 врача: 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.), т.е. без параметров распределения исходных данных. Аналогия статистических коэффициентов и средних предъявляет при суммировании коэффициентов требование, аналогичное требованию к суммированию средних: необходимость учета весовых значений суммируемых коэффициентов. Это обеспечивается суммированием относительных величин путем их пересчета на основе суммирования исходных абсолютных чисел (см. раздел «Относительные величины»).

5.3.2. Статистическое взвешивание. Вычисление некоторых индексов

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 индексы как методы статистического анализа детально не рассматриваются, представляется целесообразным разобрать примеры, демонстрирующие роль статистического весового оценивания при вычислении и анализе такого рода величин.

Рассмотрим пример оценки изменения цен на некоторые продукты питания в г. Москве по данным минимального уровня их стоимости в сети магазинов «Пятерочка» на середину 2009 и 2012 гг. на основе использования агрегатного индекса цен³⁵. Простой агрегатный индекс представляет собой отношение суммы цен за отчетный период к сумме цен за те же товары (услуги) в базисный, то есть исходный для расчетов период.

Обычно за основу такого рода расчетов берется стандартный набор самых необходимых товаров. Для продовольственных продуктов этот набор называется потребительской корзиной продуктов. В РФ, например, согласно нормам потребительской корзины, в 2011-2012 гг. трудоспособному человеку полагалось в месяц среди прочего примерно: 9 кг картофеля; 3 кг мясных продуктов; 2 кг свежих фруктов; 1 кг рыбных продуктов³⁶. Применительно к разным социальным, демографическим категориям населения и месту его проживания потребительская корзина может существенно различаться, поэтому состав потребительской корзины является постоянным объектом споров и политических дебатов. Для упрощения расчетов рассмотрим динамику цен только на некоторые продукты, традиционно включаемые в состав потребительской корзины (табл. 80).

Таблица 80

Расчет агрегатного индекса цен

Продукты	Цены (руб. за кг)	
	2009 год	2012 год
Мука пшеничная	22,7	20,9
Картофель	16,9	33,9
Сахар	21,9	27,2
Молоко	26,2	40,4
Итого	2096,7	2134,4

³⁵ По данным сайта <http://yaprotrebitel.ru/product-price>.

³⁶ Утверждена Федеральным законом Российской Федерации от 31 марта 2006 г. № 44-ФЗ.

Опираясь на изменения суммарного уровня цен, их рост можно представить в виде соотношения: $\frac{2096,7 \times 100}{2134,4} = 101,8\%$

Это соотношение можно пояснить так: стоимость указанного набора продуктов возросла на 1,8%. Как видно из представленных расчетов, простой агрегатный индекс представляет собой соотношение двух сумм. Однако такой показатель является весьма грубым и приблизительным. Это связано с тем, что реальное потребление различных товаров и услуг различно³⁷, и поэтому такого рода индексы оказываются лишенными практического смысла.

Для получения более объективной картины динамики цен необходимо произвести расчет взвешенного агрегатного индекса, который учитывал бы весовое значение каждого продукта (табл. 81).

Таблица 81

Расчет взвешенного агрегатного индекса цен

Продукты	Потребительские цены		Объем потребления в год на 1 чел.		
			Нижние границы нормы потребления	Стоимость руб./год	
	2009	2012		2009	2012
Мука пшеничная	22,7	20,9	30	474,43	436,81
Картофель	16,9	33,9	95	572,91	1149,21
Сахар	21,9	27,2	24	595,68	739,84
Молоко	26,2	40,4	320	1058,48	1632,16
Сумма	2096,7	2134,4		2701,5	3958,02

Рост (прирост) взвешенного индекса цен составит: $\frac{2701,5 \times 100}{3958,02} = 146,5\%$.

Отсюда темп прироста цен на указанный набор продуктов в случае использования взвешенного индекса цен будет равен 46,5%.

Аналогичным образом можно просчитать рост реального потребления продуктов всей потребительской корзины за сравниваемые годы в сопоставимых ценах. Так получают взвешенный агрегатный индекс количеств.

³⁷ По данным Министерства здравоохранения и социального развития РФ (Приказ МЗ и СР РФ от 2 августа 2010 г. № 593н «Об утверждении рекомендаций по рациональным нормам потребления пищевых продуктов, отвечающим современным требованиям здорового питания»).

Следует отметить, что индексы, как и любые статистические показатели, не обладают исчерпывающей информативностью. Т.е. их конечные величины во многом зависят от факторов, не учитываемых при вычислениях. Например, представление индексов в рублевом эквиваленте создает определенную вероятность ошибочных заключений, обусловленных изменением реальной стоимости рубля (величины обменного валютного курса). Для учета этого фактора, в свою очередь, используются специальные показатели, в т.ч. межбанковские валютные индексы. Среди них обращает на себя внимание получивший в последнее десятилетие широкую известность оригинальный индекс *Биг-Мака*.

Индекс Биг-Мака – способ определения реальных курсов национальных валют, когда за основу расчетов берется стоимость одного бутерброда, выпускаемого компанией *McDonald's* по строго стандартной технологии. Применение этого индекса в качестве эталона считается оправданным по нескольким причинам: сеть ресторанов *McDonald's* имеется практически во всех странах мира; бутерброды производятся на местах (не экспортируются и не импортируются), т.е. их стоимость не зависит от особенностей национальной таможенной политики, а зависит только от объемов выпуска, цены аренды, сырья (хлеб, сыр, мясо и овощи), рабочей силы, энергоносителей и т.п. «местных» факторов.

Существует большое число других индексов³⁸, и практически во всех из них обязательно учитываются весовые значения анализируемых статистических данных.

5.3.3. Упрощенные способы вычисления среднего арифметического

Расчет среднего арифметического с помощью современных статистических программ, установленных на компьютерах, в принципе сводится к вводу простого ряда исходных числовых данных. Однако приемы упрощенного «ручного» или «полуавтоматического» вычисления средних

³⁸ В этом издании они не рассматриваются.

могут оказаться не только более быстрыми, но и более точными, чем на ПЭВМ, за счет сокращения ошибок, неизбежно возникающих при ручном вводе в компьютер больших объемов информации, особенно в ситуациях:

- когда требуется быстрое, ориентировочное оценивание среднего уровня показателей;
- необходимо вычислить среднее на основе готовых табличных данных, имеющихся в так называемых «твердых копиях» (на бумаге), особенно когда варианты представлены многозначными числами и большими объемами наблюдений.

В независимости от числа наблюдений и размерности исходных данных в вариационном ряду быстро получить среднее арифметическое можно, если:

- распределение вариант (групп вариант) строго симметрично и их число нечетное, тогда вычислений вообще проводить не требуется, т.к. среднее арифметическое равно варианту, занимающей срединное положение в таком ряду;
- распределение вариант (групп вариант) строго симметрично, но их число четное, то среднее арифметическое – есть полусумма крайних вариант:

$$M = \frac{V_{\max} + V_{\min}}{2};$$

- распределение вариант незначительно отличается от симметричного, то полусумма крайних вариант даст приближенное значение среднего арифметического. *Например*, взвешенное среднее арифметическое, вычисленное по данным вариационного ряда в таблице, равно 3618,5 г. Среднее арифметическое, вычисленное упрощенным способом, равно $(3350-3900)/2=3625,0$ г. Разница между этими результатами, отражающими вес ребенка, $3625,0-3618,5=6,5$ г явно несущественна для практики.

Более точные результаты при несимметричном распределении вариант можно получить, используя *способ моментов*. *Например*, требуется определить средний вес новорожденных детей. Исходные данные представлены в виде таблицы (табл. 82), где варианты наблюдений – вес детей в граммах (четырёхзначные числа). Объем наблюдений достаточно велик – 2500 детей. Если

просто вводить в компьютер весь этот числовой массив в развернутом виде или следовать обычному порядку расчетов среднего взвешенного, то придется оперировать большим количеством громоздких величин. Так, для получения среднего взвешенного все частоты P необходимо перемножить на соответствующие им варианты V (вес в граммах): 3350×100 , 3400×150 , 3450×175 и т.д. Затем суммировать эти произведения и разделить на число наблюдений.

При вычислении по способу моментов проводить громоздких операций не требуется, поскольку вместо вариантов и частот используются условные отклонения и частоты вариант. Т.е. исходные данные (табл. 82) заменяются другими, менее громоздкими числами. Последовательность операций при вычислении среднего арифметического упрощенным способом выглядит следующим образом:

1. Определяем частоты (ω) вариант в ряду распределения. Вычислить эти частоты 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 (D) 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 грамм.

3. Выставляем условные отклонения (D), последовательно увеличивая их значения на единицу, начиная от 0, который соответствует варианту, принятой за условное среднее, до самой большой (со знаком плюс) и самой малой (со знаком минус).

4. После этого находим произведения частот на условные отклонения (ωD): $0,04 \times -5 = -0,20$; $0,06 \times -4 = -0,24$ и т.д.

5. Для получения искомого среднего арифметического эти произведения суммируются: $0,37$. После чего эта сумма умножается на величину интервала $50 \times 0,37$ (в нашем примере интервал = 50 грамм), и к этой сумме прибавляется условное среднее (в данном примере 3600 грамм): $M = 50 \times 0,37 + 3600 = 3618,5$.

Вычисление среднего арифметического способом моментов

Вес в граммах V	Число детей P	Частоты ω	Условные отклонения D	ωD
3350	100	0,04	-5	-0,20
3400	150	0,06	-4	-0,24
3450	175	0,07	-3	-0,21
3500	250	0,10	-2	-0,20
3550	275	0,11	-1	-0,11
$A=3600$	300	0,12	0	0
3650	375	0,15	1	0,15
3700	275	0,11	2	0,22
3750	225	0,09	3	0,27
3800	200	0,08	4	0,32
3850	125	0,05	5	0,25
3900	50	0,02	6	0,12
Всего	2500	1,00	-	0,37

В итоге, несмотря на кажущуюся сложность расчетов, для нахождения среднего арифметического по способу моментов громоздких вычислений не понадобилось.

5.3.4. Вычисление степенных средних в MS Excel

Для вычисления степенных средних в *MS Excel* имеется целый набор специальных функций, использование которых не представляет труда, если эти вычисления опираются на простые ряды (блоки) числовых значений исходных данных:

СРЗНАЧ возвращает (синоним – позволяет получить) среднее арифметическое из нескольких массивов (аргументов) чисел. Число 1, число 2 и т.д. – это от 1 до 30 массивов, для которых вычисляется среднее.

СРГАРМ возвращает среднее гармоническое множества данных. Среднее гармоническое – это величина, обратная среднему арифметическому обратных величин.

СРГЕОМ возвращает среднее геометрическое значений массива или интервала положительных чисел.

Если исходные данные представлены в виде взвешенных вариационных рядов, то требуется применение формул, komponуемых «вручную» на основе

отдельных функций. Рассмотрим относительно несложную последовательность решения такого рода задач.

1. В электронной таблице расположите данные, как указано на рис. 54. При этом в столбце А блока ячеек А1:А9 расположите варианты результатов измерения роста (см), а в столбце В блока В1:В9 – значения частот³⁹.

	А	В	С
1	152	1	
2	155	2	
3	158	4	
4	161	7	
5	164	6	
6	167	5	
7	170	3	
8	173	2	
9	176	1	
10			

Рис.54. Исходный взвешенный вариационный ряд

2. Для получения взвешенного среднего арифметического введите прямо с клавиатуры в клетку D1 формулу = СУММПРОИЗВ(А:А;В:В)/СУММ(В:В). После нажатия на клавишу <ENTER> в клетке D1 появится значение вычисленного взвешенного среднего арифметического (рис. 55). Обратите внимание, что в формуле указаны не конкретные адреса ячеек, где расположены исходные числа (А1:А9 и В1:В9), а только наименования столбцов (А:А и В:В).

³⁹ Исходные ряды вариант и их частот при этих вычислениях необязательно ранжировать.

	A	B	C	D
1	152	1		163,61
2	155	2		
3	158	4		
4	161	7		
5	164	6		
6	167	5		
7	170	3		
8	173	2		
9	176	1		
10				

Рис.55. Исходный взвешенный вариационный ряд и значение искомого среднего взвешенного

Это позволяет вводить в эти столбцы практически любое количество исходных значений вариант и соответствующих им частот, получая соответствующее среднее взвешенное. Важно только, чтобы количество вариант и количество их частот совпадало.

4. Для получения взвешенного среднего квадратического введите в клетку D2 формулу

=СТЕПЕНЬ (СУММПРОИЗВ(A:A;A:A;B:B)/СУММ(B:B);1/2).

После нажатия на клавишу <ENTER> в клетке D2 появится значение вычисленного взвешенного среднего квадратического (рис. 56).

4. Для получения взвешенного среднего кубического введите в клетку D3 формулу =СТЕПЕНЬ (СУММПРОИЗВ(A:A;A:A;A:A;B:B)/СУММ(B:B);1/3). После нажатия на клавишу <ENTER> в клетке D3 появится значение вычисленного взвешенного среднего кубического (рис. 56).

5. Для получения взвешенного среднего гармонического необходимо вычислить дополнительные величины. В ячейку C1 введите формулу B1/A1. Затем скопируйте эту формулу до конца рядов данных, занесенных в столбце A и B (в данном примере до ячейки C9). Для получения величины среднего гармонического активизируйте указателем мыши на клетку D4. Введите непосредственно с клавиатуры в эту клетку формулу =СУММ (C:C)/СУММ(B:B). После нажатия на клавишу <ENTER> в клетке D4 появится значение вычисленного среднего гармонического (рис. 56).

	A	B	C	D
1	152	1	0,007	163,61
2	155	2	0,013	163,71
3	158	4	0,025	163,80
4	161	7	0,043	163,42
5	164	6	0,037	
6	167	5	0,030	
7	170	3	0,018	
8	173	2	0,012	
9	176	1	0,006	
10				

Рис.56. Исходный взвешенный вариационный ряд, значения взвешенных: среднего арифметического, среднего квадратического, среднего кубического и среднего гармонического

Для получения значения взвешенного среднего геометрического в электронной таблице расположите данные условного примера вычисления средней длительности пребывания больных в отделении стационара в 1 квартале года, как указано на рис. 57.

	A	B	C
1	15,00	20	2,70805
2	12,00	15	2,48491
3	11,00	14	2,39790
4			

Рис.57. Исходный вариационный ряд

При этом в столбце А (блок А1:А3) расположите варианты результатов регистрации средней длительности пребывания больных в стационаре (дней) за январь, февраль и март месяцы, а в столбце В (блок В1:В3) – значения частот наблюдений (число больных в эти месяцы).

В ячейку С1 введите формулу $Lp(A1)$, затем скопируйте эту формулу до конца блоков данных, занесенных в столбце А или В (в данном примере до строки 3). Введите непосредственно с клавиатуры в клетку D1 формулу $=EXP((1/(СУММ(В:В))\times(СУММПРОИЗВ(В:В;С:С)))$. После нажатия на клавишу <ENTER> в клетке D1 появится значение взвешенного среднего геометрического, равного 10,9 дня.

5.4. Структурные средние. Мода и медиана

Большое значение при выборе характеристики среднего уровня признака имеет и распределение вариантов в вариационных рядах. В ряде ситуаций вместо степенных средних более целесообразно использовать структурные средние. К ним относятся мода и медиана.

Мода (M_o) – наиболее часто встречающаяся варианта. Она дает представление о центре распределения вариационного ряда. Используется:

- для определения центра распределения в открытых вариационных рядах, т.е. когда нет однозначно определенных верхних и (или) нижних границ ряда;
- для определения среднего уровня в рядах исходных данных (вариационных рядах) с резко асимметричным распределением, когда среднее арифметическое теряет одно из своих главных свойств – типичность для данного ряда.

Примером асимметричного распределения может быть ряд наблюдений длительности послеоперационной жизни больных, прооперированных по поводу рака желудка (Белицкая Е.Я., 1974). Основная масса больных погибает на 4-5 году жизни, но незначительная часть больных живет больше 10 лет⁴⁰. Если определять среднюю длительность послеоперационной жизни с помощью простой средней арифметической, то из-за последней группы больных среднее арифметическое может достигать 10 лет. Но такая длительность послеоперационной жизни нетипична для данных больных, поскольку основная масса из них погибает значительно раньше. Более точное значение средней продолжительности послеоперационной жизни здесь даст мода.

Во взвешенном ряду принципиально сохраняется такая же проблема. Например, требуется определить среднюю длительность госпитализации работников промышленного предприятия в связи с производственным травматизмом.

⁴⁰ Данные условные.

При визуальном анализе графического изображения распределения результатов наблюдений (табл. 83 и рис. 58) видно, что вариационный ряд несимметричен: вершина распределения сдвинута в начало ряда. Если определять среднюю величину на основе среднего арифметического (M), то средняя длительность одной госпитализации составит 4,1 дня. Однако чаще всего (M_0) длительность госпитализации составляла 3 дня.

Таблица 83

Распределение работников по длительности госпитализации

Число дней госпитализации V	Число работников P	VP	Частоты (%)	Накопленные частоты (%)
2	7	14	6,6	6,6
3	18	54	25,5	32,1
4	11	44	20,8	52,8
5	6	30	14,2	67,0
6	4	24	11,3	78,3
7	3	21	9,9	88,2
8	2	16	7,5	95,8
9	1	9	4,2	100,0
Итого	52	21212	100,0	—
	$M=212/52 \approx 4,1$		—	—

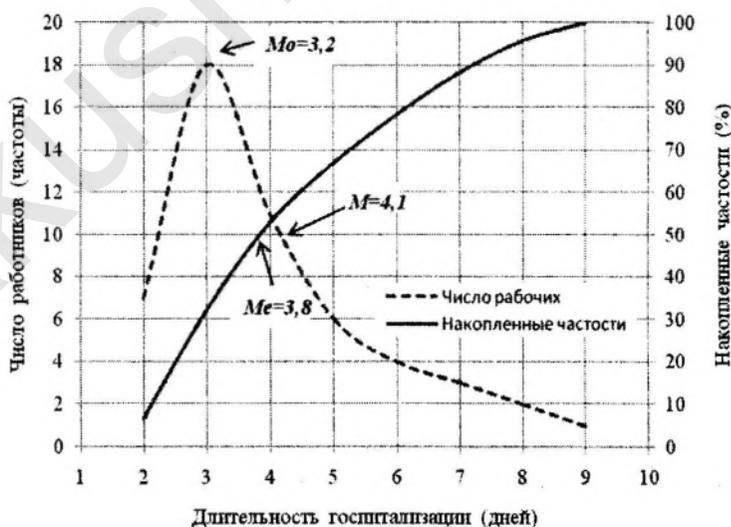


Рис. 58. Распределение обследованных работников по длительности госпитализации

Установить моду в взвешенном 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 (см. второй вариант табл. 72), наибольшая частота соответствует группе

156-157 см. Середина группы соответствует числу работников $\frac{52}{2} \approx 26$,

т.е. варианте, находящейся между 3 и 4 днями. Более точный результат можно

получить путем вычисления по формуле: $M_o = V_o + h \frac{P_{MM} - P_{MM-1}}{2 \times P_{MM} - P_{MM-1} - P_{MM+1}}$, где:

V_o – нижняя граница модального интервала; h – величина интервала;

m''' – частоты модального интервала; $m'''-1$ – частоты предмодального

интервала; $m'''+1$ – частоты после модального интервала. Отсюда точное

значение моды таково $M_o = 3,2$ дня.

Медиана (Median) – это срединная 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, представленной в табл. 84, 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. 85), то медиана равна полусумме двух

средних вариант $(21+22)/2 = 21,5$.

Таблица 84

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

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, медиана – 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 полусумма вариант с рангами $\frac{8}{2}+1 = 5$ и $\frac{8}{2} = 4$, то есть полусумма 5 и 4 по счету вариант.

В сгруппированном интервальном вариационном ряду положение медианы устанавливается по *накопленным частотам 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 (рис. 58).

Для точного определения медианы в интервальном ряду используется формула: $Me = V_o + h \left(\frac{\frac{N}{2} - SM_{e-1}}{PMe} \right)$, где V_o – нижняя граница медианного

интервала, h – величина интервала, SM_{e-1} – накопленные частоты предмедианного интервала, SM_e – накопленные частоты медианного интервала, Pm_e – частота медианного интервала, N – число наблюдений. Ошибка медианы $m_e = (\pi / 2N)^{1/2}$.

Медиана применяется:

– для определения среднего уровня признака в числовых рядах с неравными интервалами в группах;

– для обозначения среднего уровня признака, когда исходные данные представлены в виде качественных признаков и когда единственным способом обозначить 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 учреждений и т.п. Имеется в виду учет оптимальной удаленности учреждений от всех объектов обслуживания (оптимальный радиус обслуживания).

В настоящее время 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 показателей балльных оценок медианы или моды.

В *MS Excel* есть специальные функции, позволяющие определять моду и Медиану в негруппированных рядах.

5.5. Показатели разброса значений рядов наблюдений

В предыдущих разделах рассматривались методы вычисления средних величин. Помимо них как показателей центра распределений, большое значение имеет и характер распределения (разброс, рассеивание) отдельных значений в ряду наблюдений. *Например*, среднее арифметическое в двух представленных статистических совокупностях (рис. 59) одинаково: $M=10$. Однако структура этих совокупностей совершенно различна.

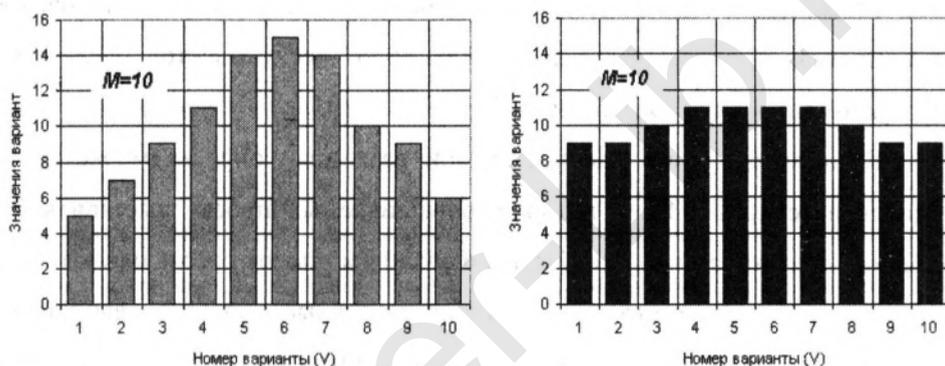


Рис. 59. Распределение исходных данных в двух группах (при одинаковых средних арифметических)

Более того, при решении практических задач возможна ситуация, когда в сравниваемых группах вообще 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.

В практической статистике, учет разброса значений вариант в сравниваемых группах играет весьма существенное значение. В частности, разброс выборочных данных пропорционально связан с их репрезентативностью: чем больше разброс значений, тем больше статистическая ошибка. Кроме того, на практике при статистической оценке

влияния определенных факторов на показатели здоровья нередко возникает ситуация, когда средняя величина оцениваемого параметра здоровья остается практически неизменной, но происходит существенное расслоение или, наоборот, консолидация значений наблюдений в опытных группах по сравнению с контрольными.

Определенное методологическое значение имеет и тот факт, что высокая вариабельность, изменчивость медико-биологических объектов в массе наблюдений – одно из их фундаментальных свойств объективной реальности. Это свойство, *например*, достаточно широко используется при оценке экологической ситуации: снижение видовой, межвидовой и индивидуальной вариабельности представителей флоры и фауны рассматривается как признак деградации, вырождения. Другой пример: с точки зрения формальной статистики, существенное сужение границ вариабельности показателей общественного здоровья может быть итогом «подгонки» результатов в сторону наиболее благоприятных значений либо как симптом сильного внешнего давления на ситуацию.

В практике работы учреждений медико-социальной помощи учет вариабельности тех или иных характеристик здоровья также имеет весьма существенное значение. *Например*, учет крайних значений частоты и длительности заболеваний с временной утратой трудоспособности (ЗВУТ), с одной стороны, позволяет выявлять случаи необоснованной выдачи больничных листков (обычно при 2-3-дневной длительности одного случая ЗВУТ). С другой – позволяет выделять группу часто и длительно болеющих (4-5 случаев ЗВУТ и (или) 30-40 дней нетрудоспособности и более за год), которая играет существенную роль в структуре экономических потерь, связанных с заболеваемостью работающих.

В практической статистике используются различные *показатели разброса, рассеивания* значений в рядах наблюдений, часто называемых *критериями разнообразия или однородности* статистических совокупностей (табл. 86). Сам по себе факт существования определенного множества

критериев разброса значений свидетельствует об определенных ограничениях информативности, иначе говоря – о недостатках каждого из этих критериев.

Простейшими среди критериев разброса значений в рядах ранжированных значений наблюдений являются *лимит* и *амплитуда*.

Лимит (*Lim*) указывает границы вариационного ряда. *Например*, во время осмотра студентов были зарегистрированы следующие значения массы тела: 48, 49, 52, 55, 56, 62, 72, 79 и 92 кг. Самая большая масса тела – 95 кг, самая малая 48 кг. $Lim=(48\div 95 \text{ кг})$.

Амплитуда, или, как еще говорят, вариационный размах (*Ampl; Range*) исчисляется как разность между максимальным и минимальным значениями признака. Например, в приведенном выше ряду значений массы тела: $Ampl=95-48=47 \text{ кг}$.

Таблица 86

Основные статистические характеристики разнообразия признака (разброса вариант)

Критерии	Формулы для расчетов
1. Лимит (<i>Lim</i>)	$Lim = V_{max} \div V_{min}$
2. Амплитуда (<i>Ampl</i>)	$Ampl = V_{max} - V_{min}$
3. Дисперсия (<i>D</i>)	$D = \frac{\sum d_i^2}{n}$ для простого ряда $D = \frac{\sum d_i^2 \cdot p_i}{\sum p_i}$ для взвешенного ряда
4. Среднеквадратическое, стандартное отклонение (σ)	$\sigma = \sqrt{D}$
5. Коэффициент вариации (<i>Cv</i>)	$Cv = \frac{\delta}{M} * 100\%$

Основными недостатками этих критериев является то, что, во-первых, оба этих критерия – есть величины именованные. Таким образом, они могут применяться только для сравнения разброса в однородных и близких по числу

наблюдений рядах. Например, масса тела – в кг. Во-вторых, они полностью зависят от крайних значений признака в вариационном ряду. При этом не учитывается разброс значений внутри ряда.

5.5.1. Дисперсия

Достаточно несложно определить однородность числового ряда с учетом всех значений, составляющих этот ряд, – через усредненное отклонение всех вариантов 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 для этого ряда называется отклонением от среднего $d_i = (V_i - M)$. При усреднении всех отклонений числового ряда, т.е. при делении суммы всех отклонений на число наблюдений, должен получиться средний квадрат отклонений.

Однако для получения обобщающей характеристики числового ряда использовать простую сумму отклонений каждой варианты от среднего нельзя. Это связано с тем, что сумма всех отрицательных и положительных отклонений вариантов от среднего всегда равна нулю, поскольку происходит компенсация всех отклонений. Отклонений со знаком «+» и со знаком «-». Т.е. алгебраически сумма всех отклонений от среднего арифметического всегда равна 0. Можно избежать взаимной компенсации отклонений, используя квадраты отклонений, т.к. при возведении в квадрат отрицательные и положительные числа дают только положительные значения. Таким образом, получают обобщающую характеристику изменчивости, называемую суммой квадратов отклонений. Ее производная величина, частное от деления суммы квадратов отклонений на число наблюдений, и есть *дисперсия*, которая широко применяется и как самостоятельный показатель распределения, и как элемент вычисления других сложных статистических критериев.

Поскольку дисперсия применяется как обязательный параметр нормального (Гауссова) распределения, ее и другие статистические критерии,

при вычислении которых используют дисперсию, относят к группе параметрических статистических критериев (табл. 87).

Это обстоятельство обуславливает необходимость проверки любой исследуемой выборочной совокупности на соответствие нормальному распределению, в противном случае применение тех или иных параметрических методов статистики может оказаться неправомерным.

Таблица 87

**Практическое применение параметрических критериев
разнообразия признака**

Дисперсия D	1) Для оценки изменчивости явлений, варибельности характеристик этих явлений. 2) Для факторного анализа (дисперсионный анализ). 3) Для статистической оценки различий двух совокупностей с одинаковыми или близкими значениями средних (критерий <i>Фишера</i>).
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 (<i>правило трех сигм</i>) в изолированном ряду. 3) Для нормирования параметров различных оценок медико-биологических объектов. В том числе и при диагностической оценке показателей физического развития. 4) Для определения доверительных интервалов статистических величин и репрезентативности выборочных исследований.
Коэффициент вариации S_v	1) Используется для сравнения варибельности значений разноименных признаков. 2) Для получения нормированных оценок варибельности значений (малая, средняя, большая).

Алгебраическое выражение дисперсии: $D = \frac{\sum d_i^2}{n}$ или $D = \frac{\sum V_i^2}{n} - M^2$, где n –

число наблюдений, d_i – отклонения отдельных вариантов (V_i) от среднего арифметического (M) $d_i = (V_i - M)$. Во взвешенном ряду дисперсия вычисляется по

формуле: $D = \frac{\sum d_i^2 P_i}{\sum P_i}$, где P_i – частоты отдельных вариантов (табл. 88).

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 $D = \frac{\sum V_i^2 P_i}{\sum P_i} - M^2$			
Vi	di	di ²	Vi	Vi ²	Vi	Pi	ViPi	Vi ² Pi
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		$\sum d^2=10$	$\sum V^2=1455$		$\sum P=15$	$\sum VP=258$	$\sum V^2P=4456$	
$D=10/5=2$			$D=1455/5 - 17^2=2$		$M=258/15=17,2$ $D=4456/15 - 17,2^2=1,2$			

Упрощенный способ расчета дисперсии позволяет избежать вычислений – отклонений от среднего арифметического. При вычислении этим способом дисперсия равна разности между среднеарифметическим квадратом результатов наблюдений и квадратом среднего арифметического. Понятно, что при вычислении дисперсии с помощью микропроцессорной техники способ расчета дисперсии, как правило, не имеет никакого значения.

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 единая 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 и их последующим простым сложением можно заменить непосредственное вычисление общей дисперсии. Это свойство аддитивности дисперсий имеет большое теоретическое и практическое значение, являясь основой широко применяющегося в медико-биологических исследованиях дисперсионного анализа, с помощью которого оценивается статистическая значимость влияния, т.е. взаимодействия какого-либо фактора или групп факторов и конечного результата. Поскольку группировка наблюдений чаще всего происходит на

основе учета их качественных характеристик, а результат дисперсионного анализа описывается численной характеристикой, то и считается, что этот анализ позволяет оценивать взаимодействие качественных и количественных характеристик каких-либо явлений⁴¹.

5.5.2. Среднеквадратическое отклонение

Существенный недостаток дисперсии, которая является именованной величиной, – несоответствие её размерности и фактической размерности отдельных вариантов наблюдаемого 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 распределения, *Std. Deviation*). Стандартным оно называется потому, что на его величину не влияет размерность исходных единиц наблюдения. Однако применяться среднеквадратическое отклонение может только для сравнения разброса значений в однородных, одноименных рядах наблюдений.

Среднеквадратическое отклонение – есть корень квадратный из дисперсии $\sigma = \pm\sqrt{D}$.

В развернутом виде формула вычисления среднеквадратического отклонения простого ряда представлена следующим выражением: $\sigma = \pm \frac{\sum d_i^2}{n}$.

Для взвешенного вариационного ряда: $\sigma = \pm \frac{\sum d_i^2 p_i}{n}$.

При малом числе наблюдений ($n < 30$) и распределении их значений в соответствии с распределением Гаусса можно использовать упрощенный способ приближенного определения величины среднеквадратического отклонения: $\sigma = (V_{\max} - V_{\min})/6$, где V_{\max} и V_{\min} – минимальная и максимальная варианты ряда.

⁴¹ Более детально расчеты, применяемые при дисперсионном анализе, и оценки его параметров рассмотрены в соответствующем разделе настоящего издания.

В практической статистике среднеквадратическое отклонение гораздо чаще, чем любой другой показатель разброса, используется как мера неопределенности и типичности среднего арифметического или отдельно взятой варианты для вариационного ряда. В основе этих оценок лежит *правило трех сигм*, согласно которому практически все значения нормально распределенных вариант находятся в интервале от $M+3\delta$ до $M-3\delta$. Т.е. в 99,7% случаев все распределенные варианты ряда лежат в этом интервале⁴². Таким образом, если та или иная варианта не укладывается в этот интервал, то с высокой вероятностью можно утверждать, что эта варианта не относится к этому ряду, нетипична для данного ряда, «выскакивающая варианта». Или наоборот: если в этот интервал не укладывается больше 0,3% всех вариант вариационного ряда, то весьма вероятно, что среднее арифметическое нетипично для данного ряда.

Среднеквадратическое отклонение широко используется в практической нормологии⁴³. При этом сигма принимается квантом изменчивости, а нулевой точкой отсчета – среднее арифметическое нормальной выборки. Условные границы принимаются в градуированных шкалах, где градации определяются значениями σ , исходя из выражения $M \pm \sigma$. Основой такого нормирования является кривая нормального распределения, которая, как известно, обладает наибольшей общностью с распределением почти любого типа. При небольшом числе наблюдений диапазон изменчивости целесообразно определять 6 зонами, где 1 и 2 зоны включают в себя малые значения признака или его отрицательную атопию, 3 и 4 зоны – норму, а 5 и 6 – положительную атопию или большие значения признака.

Примером применения такого статистического нормирования в практической медицине может служить оценка физического развития человека по методу *сигмальных шкал*. Например, среднеквадратическое отклонение

⁴² Согласно этому правилу, в интервале от $M+2\delta$ до $M-2\delta$ лежат около 95,5% вариант; в интервале от $M+\delta$ до $M-\delta$ – около 68,3% всех вариант вариационного ряда.

⁴³ *Нормология* – наука о закономерностях нормальной жизнедеятельности человека.

используется⁴⁴ 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 границами – *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 $\pm 2\sigma$, 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 параметры нормального распределения при вычислении сложных параметрических критериев и проведения параметрического статистического анализа.

В то же время дисперсия и среднеквадратическое отклонение как статистические критерии рассеивания имеют следующие недостатки:

- эти критерии – 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 незначителен разброс в изолированном ряду. Они могут использоваться только для сравнения двух или нескольких рядов однородных распределений.

⁴⁴ Последние десятилетия в отечественной педиатрии для оценки физического развития используются центильные таблицы. См. раздел настоящего издания «Центильный метод оценки физического развития».

5.5.3. Коэффициент вариации

Недостатков, свойственных дисперсии и среднеквадратическому отклонению, лишен коэффициент вариации C_v (*Kurtosis*). Этот коэффициент представляет процентное отношение среднеквадратического отклонения к среднему арифметическому $C_v = \frac{\delta}{M} 100\%$. Арифметически отношение M и σ нивелирует влияние абсолютной величины этих характеристик, а процентное соотношение делает коэффициент вариации величиной неименованной. Таким образом, появляется возможность использовать его для сравнения совокупностей, состоящих из разнородных величин.

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 параметрических критериев разброса наглядно демонстрирует табл. 89.

Таблица 89

Состав работников промышленного предприятия *N*

Учетный признак	Среднее арифметическое M	Средне-квадратическое отклонение σ	Коэффициент вариации C_v
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.

5.5.4. Квантили

В некоторых видах распределений среднеквадратическое отклонение и дисперсия (σ и D) отсутствуют или не могут служить характеристиками рассеяния вариант по другим причинам. В частности, указанное обстоятельство может быть связано с тем, что D и σ вычисляются по отклонениям от среднего арифметического, которое не вычисляется в открытых вариационных рядах и в рядах распределений качественных признаков. Поэтому для сжатого описания такого рода распределений используется другой параметр разброса – **квантиль** (синонимы – перцентиль, персентиль⁴⁵), пригодный для статистического описания качественных и количественных признаков при любой форме их распределения. Этот параметр может использоваться и для перевода количественных признаков в качественные. В этом случае качественные оценки присваиваются в зависимости от того, какому по порядку квантилю соответствует та или иная конкретная варианта.

Особым свойством квантилей является их полная независимость от конкретных значений каждой отдельно взятой единицы наблюдения. Значение имеет только порядковый номер, месторасположение этой единицы в ряду распределения. Понять статистическую суть квантилей помогает графический метод, т.е. построение диаграммы распределения частот значений признака в квантилях, которые могут выражаться в процентах или долях единицы.

Например, чтобы разбить исследуемую группу на четыре равные подгруппы, выделяют точки, делящие исходный ряд по 25% наблюдений в каждой подгруппе (рис. 60).

⁴⁵ Указанные различия объясняются национальными особенностями перевода.

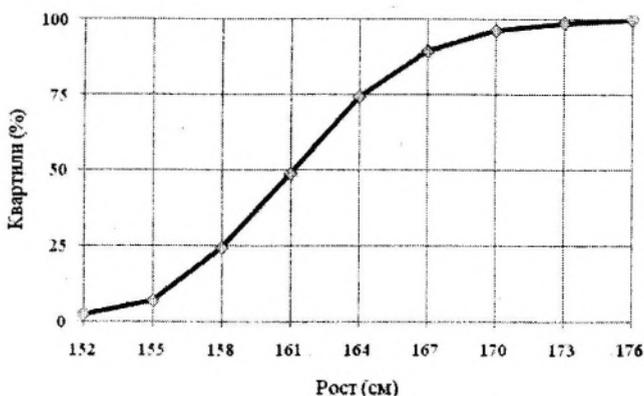


Рис. 60. Распределение квантилей

Для этого необходимо использовать три значения, которые называются *квантилями (quartile)*, поскольку в результате деления получаем 4 группы – *кварты*. Средняя точка деления из них называется *медианой*. Если использовать два *тертиля (tertile)*, можно разбить данные на три группы, четыре *квинтиля (quintile)* – пять групп и так далее. Так получают *децили (decile)*, которые делят данные на 10 частей, и *центили (centile)*, которые делят данные на 100 частей (их также называют *процентилями*). Конкретные значения квантилей могут быть выражены через центили: самый левый квантиль равен 25-му центиллю ($V_{0.25}=158$ см), медиана – 50-му центиллю ($V_{0.50}=161$ см) и т.д.

В медико-биологических исследованиях чаще всего 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 варианту, чье значение не превышает значения у 3% всех вариантов данного ряда. Варианты меньше 10-го ($V_{0.10}$) центиля встречаются у 10% всех вариантов ряда и т.д.

Как уже отмечалось, значения центильных границ могут использоваться для шкалирования (квантификации) данных. Например, показатели физического развития, значения которых у конкретного пациента меньше 3-го центиля, оцениваются как резко пониженные. Находящиеся по своему значению между 3 и 10 центилем – как пониженные, между 10 и 25 центилем –

ниже среднего, между 25 и 75 – средние, между 75 и 90 – выше среднего, между 90 и 97 – повышенные, выше 97 – резко повышенные.

В практике статистического анализа наиболее часто используются следующие квантили;

$V_{0,5}$ – медиана;

$V_{0,25}$, $V_{0,5}$, $V_{0,75}$ – квантили (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).

Вычисление квантилей требует специальной подготовки исходной информации. В частности, исходные данные должны быть ранжированы (упорядочены) от 1-го до конечного в порядке возрастания. Значения конкретного центиля получаются вычислением величины $q=k*(n+1)/100$ и ее последующей интерполяцией между двумя ближайшими (большим и меньшим) к q значениями данных. Рассмотрим пример (J.M. Bland, D.G. Altman, 1994). Для 5-го центиля выборки из 145 наблюдений мы имеем $q=5*146/100=7,3$. Таким образом, 5-й центиль находится на 3/10 расстояния от 7-го к 8-му наблюдению ранжированного ряда. Если значения этих данных равны 11,4 и 14,9, соответственно, искомым центиль равен 12,45. Доверительные интервалы могут быть построены для любого квантиля.

5.5.5. Использование MS Excel для нахождения квантилей

В MS Excel имеются статистические функции, предназначенные для нахождения любых квантилей. Пример, необходимо рассчитать квантили следующих значений: 0,03; 0,10; 0,25; 0,50; 0,75; 0,90; 0,97 в массиве результатов измерения роста в группе подростков. Для этого введите в столбцы А и В (блок клеток А1:В10) значения роста, в столбец С (блок С1:С7) – значения квантилей, которые необходимо получить (рис. 61).

Функция ПЕРСЕНТИЛЬ представляет квантили (синонимы – персентили, перцентили) заданного числового массива. Окно функции имеет

поля ввода исходной информации **Массив** и **К**. В окно «**Массив**» могут вводиться или числовые последовательности, или ссылки на ячейки, содержащие числа. В окно «**К**» – искомые значения квантилей. Эти значения указываются в долях от единицы: 0,1; 0,25; 0,50 и т.д. Для того, чтобы вычислить ряд центилей, после ввода исходных данных в столбцы А, В, С установите функцию **ПЕРСЕНТИЛЬ** (рис. 62) в клетку D1. Затем введите адрес блока ячеек A1 и B10. Обратите внимание! Адрес блока ячеек A1:B10 закрепляется знаком \$, для чего нужно установить указатель мышки на адрес (например, A1) и нажать на клавишу F4. Затем операция повторяется для адреса B:10.

После нажатия клавиши <ОК> в клетке D1 появится результат вычисления – значение центиля 0,03=157,0. Затем нажмите левую клавишу мыши и, не отпуская ее, перетащите указатель до клетки D7. Тем самым будет скопирована формула, содержащаяся в клетке D1, в клетки D2:D7 и получен соответствующий результат.

Из полученных данных видно, что 3% обследованных имели рост не выше 157,0 см, 10% – не более 157,9 см и т.д.

	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		
11				

Рис. 61. Исходные данные и соответствующие квантили

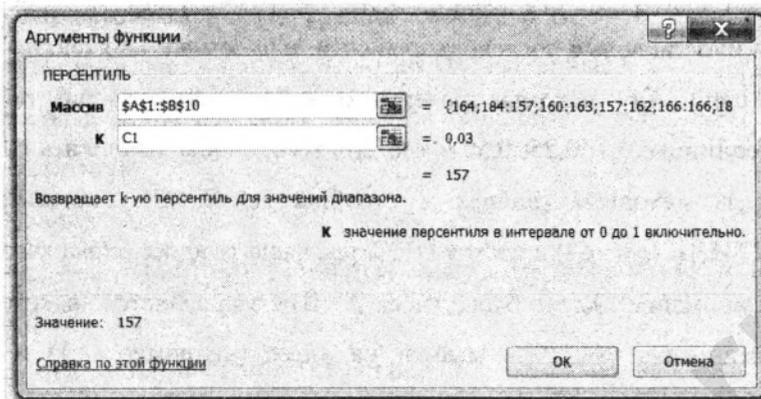


Рис. 62. Пример заполнения рабочего окна функции ПЕРСЕНТИЛЬ

Для получения более детальной характеристики распределения центилей нужно вызвать <Пакет анализа>, в котором находится функция <Ранг и перцентиль> (рис. 63). Заполните поля окна функции в соответствии с расположением исходных данных:

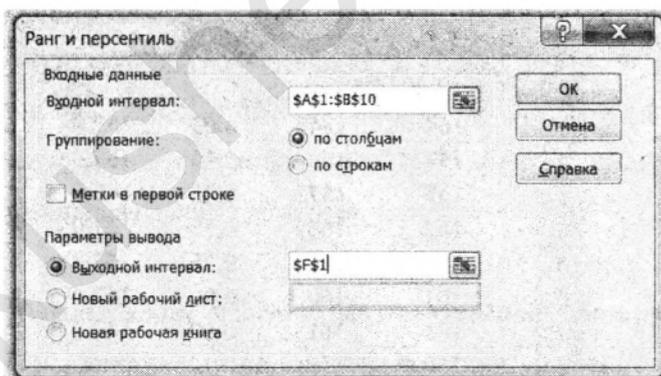


Рис. 63. Заполнение окна исходных данных функции «Ранг и перцентиль»

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. Результат представляется в виде таблицы (рис. 64).

F	G	H	I	J	K	L	M
<i>Точка</i>	<i>Столбец1</i>	<i>Ранг</i>	<i>Процент</i>	<i>Точка</i>	<i>Столбец2</i>	<i>Ранг</i>	<i>Процент</i>
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	166	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	0,00%	3	157	10	0,00%

Рис.64. Вывод результатов функции «Ранг и перцентиль»

Обратите внимание! Несмотря на то, что исходные данные вводились одним массивом, результат представлен в виде двух блоков, 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 «столбец» указываются варианты, расположенные в ранговом порядке (в порядке убывания); в следующем ряду, «ранг», указывается ранговый номер варианты, в ряду «процент» указывается процент (перцентиль) вариант, которые имеют такое же и меньше значение, чем варианта данного ранга.

5.6. Асимметрия и эксцесс. Статистические моменты

Существенное значение для оценки групп статистического наблюдения имеют такие характеристики распределения, как коэффициент **асимметрии** (*Skewness*) и **эксцесса** (*Coefficient_of_excess* – англ.).

На рисунке (рис. 65) полигон частот имеет явно скошенное в правую сторону асимметричное распределение. Количественное описание такой скошенности дается с помощью *коэффициента асимметрии* (As).

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 ($As < 0$). Если распределение симметрично относительно среднего арифметического (математического ожидания), то $As = 0$.

Количественное описание островершинности распределения дается с помощью коэффициента *эксцесса* (Ex). При $Ex > 0$ распределение принято считать островершинным, при $Ex < 0$ – туповершинным (рис. 65).

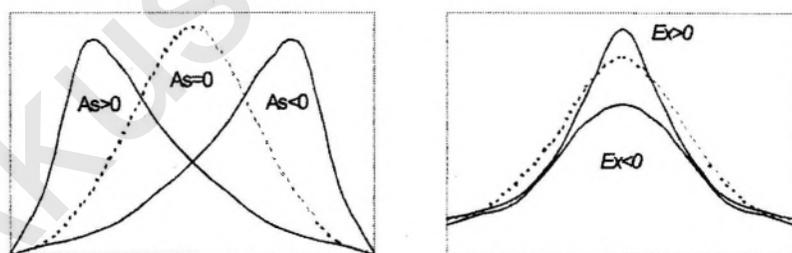


Рис. 65. Примеры статистических распределений и соответствующих им параметров As и Ex

Для практического вычисления многих статистических характеристик, в том числе As и Ex , удобно пользоваться эмпирическими (т. е. вычисляемыми на основе результатов конкретных наблюдений) моментами. Формула для

вычислений обычных эмпирических моментов: $M_k^* = \frac{\sum P_i V_i^k - C^k}{n}$, где V_j – наблюдаемые варианты, P_i – 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: $M_1^* = \frac{\sum P_i V_i - 0^1}{n} = \frac{\sum P_i V_i}{n}$.

Нетрудно заметить, что конечный вариант этой формулы 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. Отсюда *центральный момент второго порядка равен дисперсии.* С помощью центральных моментов третьего и четвертого порядка удобно вычислять A_3 и E_3 (табл. 90).

Таблица 90

Начальные и центральные эмпирические моменты⁴⁶

Эмпирические моменты	Центр распределения	К	Формулы расчета	Область применения (статистический критерий)
Начальный момент (M_1^*)	0	1	$\frac{\sum P_i V_i}{n}$	Среднее арифметическое ($M=M_1^*$)
Центральный момент второго порядка (m_2)	Среднее арифметическое	2	$\frac{\sum P_i V_i^2 - M^2}{n}$	Дисперсия $D = m_2$
Центральный момент третьего порядка (m_3)	Среднее арифметическое	3	$\frac{\sum P_i V_i^3 - M^3}{n}$	Асимметрия $A_3 = \frac{m_3}{\sigma^3}$
Центральный момент четвертого порядка (m_4)	Среднее арифметическое	4	$\frac{\sum P_i V_i^4 - M^4}{n}$	Экспесс $E_3 = \frac{m_4}{\sigma^4} - 3$

⁴⁶ Формально алгоритмы расчетов моментов для генеральной (несмещенные моменты) и выборочной (смещенные моменты) совокупностей различаются. Однако при практических вычислениях разница результатов этих расчетов, как правило, укладывается в границы ошибок репрезентативности.

Оценки показателей эксцесса можно представить следующим образом. Если $E > 0$, то имеется положительный эксцесс, т.е. вершина распределения сильно вытянута. Если $E < 0$, то эксцесс отрицателен и распределение низковоершинное. При $E = 0$ распределение нормальное (Гауссово), при $E = -2$ имеются две самостоятельные кривые.

Пример показателей описательной статистики (среднего арифметического дисперсии, асимметрии, эксцесса), а также графического отображения полигонов распределения результатов измерения роста в 2-х группах подростков представлен на рис. 66.

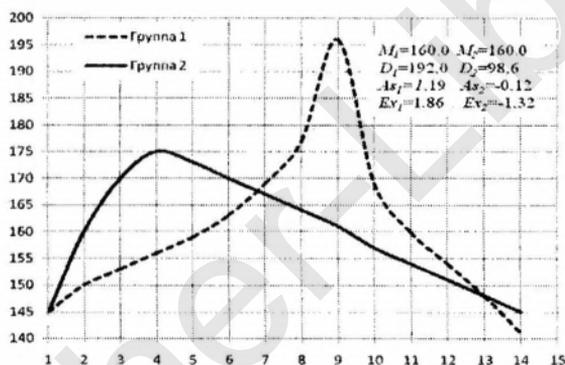


Рис. 66. Примеры статистических распределений и соответствующих им параметров A_s и E_x

5.6.1. Описательная статистика в MS Excel

Набор статистических характеристик, объединяемых названием «Показатели описательной статистики», как правило, применяется для первичного, так называемого «разведочного», анализа итогов обработки массивов данных. В MS Excel имеются отдельные статистические функции, предназначенные для нахождения большинства характеристик описательной статистики. В том числе:

СРЗНАЧ позволяет получать (математический син. – возвращать) среднее арифметическое из набора чисел, а также других источников, в т.ч. ссылки на диапазоны ячеек, содержащих числа. Если ссылки указывают на ячейку, содержащую текст, или на пустую ячейку, то такие значения

игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Если ячейки содержат логические значения и текстовые представления, используется функция **СРЗНАЧА**. Если требуется вычислить среднее значение только для тех значений, которые удовлетворяют определенным критериям, применяются функции **СРЗНАЧЕСЛИ** или **СРЗНАЧЕСЛИМН**.

МЕДИАНА возвращает значение медианы.

МОДА возвращает значение моды.

ДИСП возвращает значение дисперсии из выборочной совокупности. Из генеральной совокупности используется функция **ДИСПР**. Если исходный массив данных находится в базе данных, используются функции **БДДИСП** или **БДДИСПП**.

СТАНДОТКЛОН – возвращает стандартное отклонение (среднеквадратическое отклонение, сигму) из выборочной совокупности. Для генеральной совокупности используется функция **СТАНДОТКЛОНП**. Если ссылки указывают на ячейку, содержащую текст, или на пустую ячейку, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются. Если ячейки содержат логические значения и текстовые представления, используется функция **СТДОТКЛОНА**.

ЭКЦЕСС возвращает функцию эксцесса для массива данных.

СКОС возвращает функцию асимметрии массива данных относительно среднего арифметического.

МАКС возвращает наибольшее значение в указанном массиве чисел.

МИН возвращает наименьшее значение в указанном массиве чисел.

СЧЕТ подсчитывает количество фактических данных, являющихся числами, текстовым представлением чисел (например, число, заключенное в кавычки, такое, как «1»), датами и логическими значениями. Пустые ячейки, логические значения, текст и значения ошибок в массиве или ссылке пропускаются. Для подсчета логических значений, элементов текста или значений ошибок используется функция **СЧЕТЗ**. Если требуется подсчитать

только те числа, которые соответствуют определенным критериям, используется функция **СЧЁТЕСЛИ** или **СЧЁТЕСЛИМН**.

НАИБОЛЬШИЙ и **НАИМЕНЬШИЙ** возвращают заданные по счету и по величине значения из множества данных. Эти функции позволяют выбрать значение по его относительному местоположению.

Практически все перечисленные функции обладают одинаковым интерфейсом ввода данных. Например, окно ввода данных функции **СРЗНАЧ** (рис. 67). Кроме того, общей чертой этих функций является возможность их применения для обработки массивов данных, представленных в несимметричных или нескольких изолированных блоках ячеек.

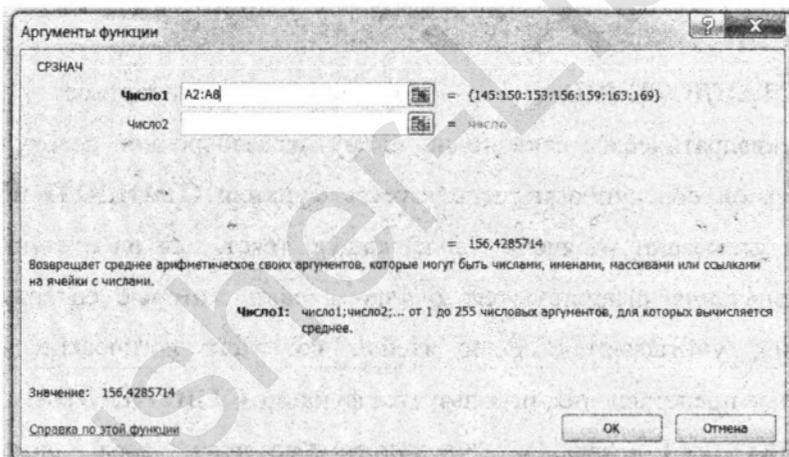


Рис. 67. Пример заполнения рабочего окна функции **СРЗНАЧ**

Для получения комплекса характеристик описательной статистики используется <Пакет анализа>, в котором находится функция <Анализ данных>. В качестве примера рассмотрим технологию получения характеристик описательной статистики на основе исходных данных, расположенных в блоке ячеек A1:B15. Для этого заполните соответствующие поля окна функции:

входной интервал, т.е. диапазон ячеек, содержащий исходные данные (\$A\$1:\$A\$15);

выходной интервал – область вывода итогов. Может отражать только верхнюю правую клетку блока ячеек, куда будут выводиться итоги (\$A\$15).

Полные размеры области вывода данных будут определяться автоматически;

группирование (по столбцам);

метки (входной диапазон содержит или не содержит метки, т.е. названий строк и столбцов);

альфа (уровень значимости, обычно равный 0,05).

Итоговая статистика. Установите флажок, тогда в выходном диапазоне будут размещены следующие характеристики: среднее (арифметическое), стандартная ошибка среднего, медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметричность, амплитуда (максимум – минимум), сумма вариантов и размер выборки (счет).

Установите флажок, чтобы в итоги включить строку для уровня надежности. Уровень надежности – это половина доверительного интервала для генерального среднего арифметического.

После завершения заполнения полей параметров нажмите кнопку [OK]. Результаты анализа принимают вид (рис. 68).

Группа 1	Группа 2		
145	145		
150	160		
153	170		
156	175		
159	173		
163	170		
169	167		
177	164		
196	161		
169	157		
160	154		
154	151		
148	148		
141	145		
Группа 1	Группа 2		
Среднее	160	Среднее	160
Стандартная ошибка	3,84	Стандартная ошибка	2,75
Медиана	157,50	Медиана	160,50
Мода	169,00	Мода	145,00
Стандартное отклонение	14,38	Стандартное отклонение	10,30
Дисперсия выборки	206,77	Дисперсия выборки	106,15
Экссесс	1,86	Экссесс	-1,32
Асимметричность	1,19	Асимметричность	-0,12
Интервал	55	Интервал	30
Минимум	141	Минимум	145
Максимум	196	Максимум	175
Сумма	2240	Сумма	2240
Счет	14	Счет	14
Наибольший(1)	196	Наибольший(1)	175
Наименьший(1)	141	Наименьший(1)	145
Уровень надежности(95,0%)	8,30	Уровень надежности(95,0%)	5,95

Описательная статистика

Входные данные
 Видный интервал: \$A\$1:\$B\$15 [F5]
 Группирование: по столбцам по строкам
 Метки в первой строке

Параметры вывода
 Видный интервал: \$A\$1:7 [F5]
 Новый рабочий лист
 Новая рабочая книга
 Подготовка статистики
 Уровень надежности: 95 %
 К-ый наименьший: 1
 К-ый наибольший: 1

OK Отмена Справка

Рис.68. Заполнение окна исходных данных и вывод результатов функции «Описательная статистика»

После нажатия на <ОК> появится результат в виде комплекса показателей описательной статистики.

Обратите внимание! Во-первых, исходные данные, несмотря на то, что они введены в виде единого блока ячеек A1:B15, воспринимаются функцией «Описательная статистика» как два отдельных столбца данных: A1:A15 и B1:B15.

Во-вторых, полученная с помощью этих функций информация требует обязательной проверки ее репрезентативности. Для этого требуется использование специальных методов проверки статистических гипотез.

6. Статистическая проверка статистических гипотез

Познание окружающего нас мира неразрывно сопряжено с процессом выдвижения тех или иных гипотез и их проверки. Многообразие и сложность окружающей нас объективной реальности определяют разнообразие возможных гипотез о свойствах этой реальности. Методология и методика проверки выдвинутых гипотез зависит от специфики изучаемых процессов и потребностей практики.

В процессе практической деятельности гипотезы всегда должны выдвигаться *перед статистическим анализом*. *Итоги статистического анализа сами по себе могут только подтверждать или не подтверждать выдвинутые ранее гипотезы*. Разумеется, логика научного познания базируется на использовании полученных ранее результатов, которые в свою очередь могут быть основой новых гипотез и последующих их проверок.

Таким образом, основа любого конкретного научного исследования – так называемая *рабочая гипотеза*, которую на основании полученных данных можно либо 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 (математические ожидания) 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.

Обычно при любом статистическом исследовании изначально выдвигают статистическую гипотезу, называемую *основной* или *нулевой* (H_0). Ее

практический смысл сводится к предположению, что различия между параметрами исследуемых выборочных совокупностей случайны⁴⁷. Итогом статистического исследования становится либо подтверждение справедливости нулевой гипотезы, либо ее отрицание. Гипотезу, которая противоречит нулевой и является ее логическим отрицанием, называют *конкурирующей* или *альтернативной* (H_1)⁴⁸. Наличие гипотез H_0 и H_1 предоставляет выбор только одного из двух вариантов, т.е. выбор одного предположения. Например, если нулевая гипотеза предполагает, что среднее арифметическое $M=15$, то логическим отрицанием будет $M \neq 15$. Коротко это записывается так: $H_0: M=15$, $H_1: M \neq 15$.

Обратите внимание! Нулевая гипотеза не может быть сама по себе «плохой» или «хорошей». Все зависит от конкретной цели конкретного исследования. Например, если целью является доказательство эффективности препарата, с точки зрения его способности вызывать торможение неблагоприятного патологического процесса, то отсутствие различий параметров, характеризующих развитие этого процесса, (нулевая гипотеза H_0 принимается, альтернативная H_1 отвергается), будет свидетельством эффективности препарата. И наоборот, если конечная цель исследования – доказательство способности препарата вызывать положительную динамику здоровья, то различие соответствующих показателей здоровья (нулевая гипотеза H_0 отвергается, альтернативная H_1 принимается) будет подтверждением неэффективности препарата.

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.).

⁴⁷ В более простом варианте это звучит так: при оценках различий каких-либо величин в качестве нулевой гипотезы обычно принимают гипотезу об отсутствии различий.

⁴⁸ В ряде случаев альтернативную гипотезу обозначают H_a .

Сложной называют гипотезу, которая состоит из конечного или бесконечного множества простых гипотез. Например, сложная гипотеза $H: D > 15$ ⁴⁹ 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 верная нулевая гипотеза (H_0), а принимается альтернативная гипотеза (H_1). 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. В медицине эти ошибки называют *ложноположительными результатами* (анализов, тестов и т.п.).

Ошибка второго рода – отвергается верная альтернативная гипотеза (H_1), а принимается неверная нулевая гипотеза (H_0). Соответственно, *ошибка второго рода* – ложноотрицательный результат. Встречается, когда по итогам диагностического обследования считается, что заболевание отсутствует, а на самом деле оно есть. Значимость ошибки второго рода обозначают символом β (табл. 91).

⁴⁹ D — дисперсия.

Обозначения значимости статистических ошибок

Источники ошибок		Истинные гипотезы	
		H_0	H_1
Принятые гипотезы	H_0	Нет ошибки	β
	H_1	α	Нет ошибки

Понятно, что последствия этих ошибок могут быть различны. Если, например, при отсутствии заболевания человек признается больным, т.е. допускается ошибка первого рода, в результате которой здорового человека начинают лечить, то последствия одни. А если при наличии заболевания человека не признают больным и, соответственно, его не лечат, то последствия этой ошибки (второго рода) другие, чаще всего более тяжкие.

Другие варианты ошибок: а) человек опоздал на нужный ему поезд; б) сел на поезд, следующий в другом направлении. Последствия этих ошибок явно неоднозначны. 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 (K)*. Вероятность ошибки первого рода при проверке статистических гипотез с помощью того или иного критерия называют *уровнем значимости критерия*. Вероятность ошибки второго рода определяет другую характеристику критерия – *мощность критерия*.

Существует жесткая функциональная связь между значениями ошибок первого и второго рода. Например, повышение чувствительности теста обычно неизбежно ведёт к увеличению вероятности *ошибки первого рода*, т.е. росту числа ложноположительных результатов, а понижение чувствительности – к увеличению риска *ошибок второго рода* – риску ложноотрицательных результатов. Поэтому в статистике обычно приходится идти на компромисс между приемлемым уровнем *ошибок первого и второго рода*. Зачастую для принятия решения здесь используется пороговое значение, которое может варьироваться, делая тест более строгим или, наоборот, более мягким. Этим

пороговым значением является уровень значимости, которым задаются при проверке статистических гипотез.

Еще одной важной характеристикой статистических критериев является область их применения. Например, если проверяют гипотезу о равенстве дисперсий в двух совокупностях *Но* $D_1=D_2$, то в качестве критерия (K)⁵⁰ истинности нулевой гипотезы (*Но*) принимают определенную величину отношения 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 ($K_{набл}$), сравнивают с известными (фактическими) значениями таких критериев ($K_{факт}$). Последние часто именуются как *теоретические критерии* ($K_{теоретич}$).

Главной проблемой здесь является правильный выбор статистического критерия с учетом допустимой области его применения. Область же применения того или иного критерия, как уже отмечалось, задается законом его распределения. В частности, весьма существенное значение в этом аспекте имеет факт, относится ли избранный критерий к семейству параметрических или к семейству непараметрических критериев. К *параметрическим* относят критерии, значения которых могут определяться только в пределах нормального распределения или, иначе говоря, при нормальном распределении значений избранной статистической совокупности. *Непараметрическими* – критерии, для которых это условие необязательно.

⁵⁰ В практике описания методик или результатов статистического анализа нередко допускаются ошибки использования термина «критерий». В частности, его путают с термином «коэффициент». Однако *критерий* – это признак, правило, на основании которого формируются оценки объекта, процесса. А коэффициент – число или алгебраическое выражение, являющееся составной частью математического выражения, уравнения, алгоритма решения задачи. Справедливости ради нужно отметить, что в целом ряде случаев один и тот же объект может выступать и как критерий, и как коэффициент. Например: критерий Стьюдента (t) используется для оценки существенности различий средних арифметических, а коэффициент (t) Стьюдента используется для определения размеров доверительных границ средних величин.

Например, в большинстве медико-биологических исследований для статистической проверки гипотез существенности различий средних арифметических используется параметрический критерий Стьюдента (t). Методология его применения основана на предположении о соответствии сравниваемых выборочных совокупностей нормальному распределению и равенстве дисперсий этих совокупностей. На практике зачастую никакой проверки соответствия исходных 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.

Наиболее оптимальный вариант – использование компьютерных статистических программ. С их помощью все необходимые данные (теоретические распределения, критические значения статистических критериев и т.п.) представляются пользователю автоматически. Результаты их сравнений с данными эмпирических наблюдений выводятся тоже автоматически. На практике это осуществляется использованием встроенных функции математической статистики компьютерных программ (в том числе и *Microsoft Excel*). С помощью этих функций при необходимости генерируются и числовые ряды различных теоретических распределений.

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, при которых нулевая гипотеза отвергается, называется *критической областью*. Другая часть возможных значений критерия, при

которых нулевая гипотеза принимается, называется *областью принятия гипотезы*. Если полученные (наблюдаемые) значения критерия $K_{набл.}$ принадлежат критической области, то нулевую гипотезу отвергают. Если они принадлежат области принятия нулевой гипотезы, то ее принимают. Критическая область и область принятия гипотезы являются интервалами, разделяемыми *критическими точками*.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области (рис. 69).

Правосторонняя критическая область определяется неравенством $K_{кр} > k_{кр}$.

Левосторонняя критическая область определяется неравенством $K_{кр} < k_{кр}$.

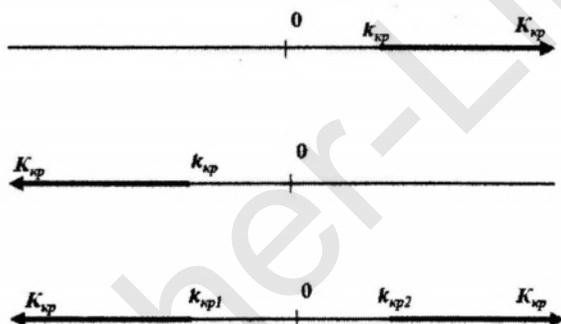


Рис. 69. Схема распределения критических областей принятия гипотез.

Двусторонней называют область, определяемую неравенством $K_{кр} > k_{кр2}$ и $K_{кр} < k_{кр1}$. Нулевую гипотезу отвергают, если вычисленное значение критерия $K_{набл}$ окажется в критической области. Если же $K_{набл}$ оказывается за пределами критической области, то оснований отвергнуть нулевую гипотезу нет.

В медико-биологических исследованиях используют как односторонние, так и двусторонние критерии. Существенным достоинством односторонних критериев является их большая мощность⁵¹ по сравнению с двухсторонними.

⁵¹ Мощность критерия – вероятность совершить ошибку второго рода, т.е. вероятность пропуска (вероятность недиагностированного состояния).

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 есть, причем весьма существенные⁵².

Например, при исследовании различий веса печени двух групп 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 $F_{набл}=25/7=3,6$. 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 большая 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 $k_{кр}=2,7^{53}$. 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 $\alpha=0,05$. При $P=0,01$ $F_{набл} < Fk_{кр}$. 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 (статистически

⁵² В наиболее тривиальном виде различия между односторонним и двухсторонним критериями учитываются так: если интересует только абсолютное значение различий, то используются односторонние критерии. Если интересует и их знак, то двухсторонние.

⁵³ Для нахождения критических точек с помощью *Microsoft Excel* можно использовать встроенную функцию *FRASПОБР*. Эта функция воспроизводит $k_{кр}$ с заданной вероятностью и степенями свободы.

достоверным) признать нельзя. Окончательное заключение в данном случае зависит от исследователя, проводившего опытное наблюдение.

Применение одностороннего критерия может быть, например, следующим: 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 показателей, составило 1,8. Число степеней свободы в этом наблюдении было равно 25. При этих условиях критическая точка распределения t критерия, если $P=0,05$, 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 незначимы. В случае использования одностороннего критерия критическая точка того же критерия $t = 1,8$ будет равна 1,71 ($1,8 > 1,71$). При таком уровне значимости *нулевая гипотеза H_0 может быть отвергнута*. Соответственно, различия можно признать статистически достоверными. Поскольку в данной ситуации ожидать улучшения здоровья в опытной группе под воздействием неблагоприятного фактора внешней среды не представляется реальным, т.е. в опытной группе может быть только хуже, чем в контрольной, можно ориентироваться на результаты анализа с помощью одностороннего критерия. 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 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. Таким образом, здесь следует придерживаться общего правила: на основании однократной или косвенной проверки статистическую гипотезу **доказать 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 неверная и, следовательно, справедливо 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 правильной (ошибка второго рода) равна β , то мощность критерия $\alpha=1-\beta$. Таким образом, 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 – увеличить объем выборок.

6.1. Оценка статистических параметров выборочных данных

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 характеристикам генеральной совокупности?

Обычно значения параметров разных выборок из одной генеральной совокупности не совпадают. Числовые значения параметров выборок являются лишь результатом приближенного статистического оценивания значений параметров в генеральной совокупности. Таким образом, статистическое оценивание выборочных данных в силу изменчивости наблюдаемых явлений позволяет получать только приближенные значения параметров генеральной совокупности⁵⁴.

Различают **точечные** и **интервальные** оценки.

6.1.1. Точечные оценки. Состоятельность, несмещенность и эффективность точечных оценок

Точечная оценка генеральной совокупности – характеристика параметра генеральной совокупности, вычисляемая на основе выборки и отражаемая на числовой оси одной точкой. Проиллюстрируем точечные оценки простым

⁵⁴ Строго говоря, в статистике *оценка* – это правило вычисления оцениваемого параметра, а термин «оценить», т.е. провести оценивание, означает указать приближенное значение.

условным примером. Пусть мы имеем генеральную совокупность N , состоящую всего из 10 вариантов. Среднее значение генеральной совокупности составляет:

$$M=(16+18+20+22+24+26+28+30+32+34)/10=25,0^{55}$$

Затем получим среднее арифметическое выборочным путем. Для этого сформируем случайным способом три выборки с числом наблюдений, равным 3, 4 и 5 (табл. 92). Полученные выборочные средние (22,7; 24,0; 25,2) являются *точечными оценками* генерального среднего арифметического (25,0).

Любая выборочная характеристика, используемая в качестве приближенного значения генеральной характеристики и получаемая вычислением одного числа (точки), называется *точечной статистической оценкой*. Таким образом, выборочные средние – есть *точечные оценки*. При избрании способа получения точечных оценок учитывается, что они должны обладать свойствами *состоятельности, несмещенности и эффективности*.

Таблица 92

Пример условных выборочных совокупностей

Генеральная совокупность		1-я выборка	2-я выборка	3-я выборка
N	V	V ₁	V ₂	V ₃
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	30
9	32			
10	34			34
Число наблюдений	10	3	4	5
Среднее арифметическое	25,0	22,7	24,0	25,2
Отклонение выборочных средних от генерального среднего		2,3	1,0	-0,2

Состоятельная оценка – точечная оценка, которая при неограниченном увеличении объема выборки приближается к истинному значению оцениваемой генеральной характеристики. *Например*, по данным табл. 90, среднее

⁵⁵ Поскольку ряд явно симметричен, то вычислить среднее арифметическое можно проще: $(16+34)/2=25,0$.

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 большей численности выборок.

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 генерального. 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 получаем *смещенную точечную оценку генеральной дисперсии*. Для получения несмещенной точечной оценки генеральной дисперсии из выборочных данных используют формулу расчета исправленной (несмещенной) дисперсии $D = \frac{\sum x_i^2}{n-1}$. При сравнении формул видно, что они отличаются лишь знаменателями. Очевидно, что при больших объемах выборки смещенная и несмещенная дисперсии отличаются мало. На практике, пользуются формулой вычисления несмещенной дисперсии, если число наблюдений в выборке не превышает 30 вариантов ($n < 30$). При большем числе наблюдений влияние $- 1$ становится несущественным и его можно игнорировать.

Эффективная оценка – точечная оценка, которая гарантирует наименьшее отклонение выборочной оценки от такой же оценки генеральной совокупности. Или в терминах математической статистики: эффективная оценка параметра – его точечная оценка, дисперсия которой меньше, чем дисперсия другой какой угодно его оценки.

6.1.2. Интервальные оценки. 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. 92) 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 (характеристики непосредственно 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 **стандартной ошибкой среднего**, или **ошибкой среднего**.

Вычисляется этот параметр в случае повторного отбора⁵⁶ по формуле:

$m = \pm \frac{\delta}{\sqrt{n}}$ или $m = \pm \sqrt{\frac{\delta^2}{n}}$, где: δ – среднеквадратическое отклонение выборки, n – число наблюдений в выборке (объем выборки), $\delta = \pm\sqrt{D}$ – среднеквадратическое (стандартное) отклонение, D – дисперсия.

Если выборка, объем которой известен (n), сформирована из генеральной совокупности бесповторным отбором, то в формулу вводится поправочный

множитель, и она приобретает вид: $m = \pm \sqrt{\left(\frac{\delta^2}{n}\right)\left(1 - \frac{n}{N}\right)}$.

Очевидно, что при большой генеральной совокупности, в т.ч. гипотетической генеральной совокупности, когда $N \rightarrow \infty$, этот множитель стремится к единице.

При определении ошибки выборочной доли, выраженной в долях от единицы (0,25 или 0,45), используют формулу $m = \pm \sqrt{\frac{P(1-P)}{n-1}}$, где P – величина

доли. Если доля выражена в % (25% или 45%), используют формулу

$$m = \pm \sqrt{\frac{P(100-P)}{n-1}}$$

Указанным способом ошибки доли определяются, если число наблюдений достаточно велико. Соответствие величины выборки такому способу вычисления ошибки можно определить, найдя произведение доли (в %) на число наблюдений, которое 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 обеспечивающий точность такого прогноза интервал, называемый **доверительным интервалом**.

⁵⁶ При *повторном* отборе попавшая в выборку единица после регистрации возвращается в генеральную совокупность и снова участвует в дальнейшем отборе. Таким образом, объем генеральной совокупности не меняется. При *бесповторном* отборе единица, попавшая в выборочную совокупность, в генеральную не возвращается, т.е. объем генеральной совокупности в процессе исследования сокращается.

Статистическая оценка, которая определяется двумя числами – концами интервала, называется *интервальной оценкой*.

Величина доверительного интервала задается вероятностью безошибочного прогноза, эту вероятность принято называть «*доверительной вероятностью*», или *вероятностью безошибочного прогноза*⁵⁷. 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 , доверительной вероятности и уровня значимости нулевой гипотезы (P) соотносятся следующим образом (табл. 93).

Таблица 93

Соотношение критериев статистической достоверности выборочных характеристик

Доверительный коэффициент t	Доверительная вероятность (%)	Уровень значимости (P)
1	68,3	0,32
2	95,5	0,05
3	99,7	0,01

При малых числах наблюдений значения коэффициента Стьюдента с учетом уровня доверительной вероятности или уровня значимости можно установить по специальным таблицам. Для этой же цели служит функция СТЬЮДРАСПОБР, которая воспроизводит (син. – возвращает) значение доверительного коэффициента Стьюдента, как функцию уровня значимости и числа степеней свободы.

Выбор того или иного уровня значимости или, соответственно, доверительной вероятности в общем является произвольным, так же как и доверительный интервал для средних, равный произведению tm . В медико-биологических исследованиях допускается доверительная вероятность не менее 95,5%. Поэтому в медицинской статистике доверительный интервал для средних, при достаточно большом числе наблюдений ($n > 30$) берется

⁵⁷ Иногда здесь используют термин «*надежность*».

не меньше $\pm 2m$. Отсюда предельная ошибка выборки $\Delta = M \pm 2m$. При доверительной вероятности 99,7% доверительный интервал составит $\pm 3m$, или $\Delta = M \pm 3m$. В целом чем больше доверительная вероятность, тем больше доверительный интервал и предельная ошибка.

Граничные точки доверительного интервала называются *доверительными пределами*.

Как уже отмечалось (табл. 93), каждому значению доверительной вероятности соответствует свой уровень значимости (P). *Уровень значимости* выражает значимость, вероятность нулевой гипотезы. Т.е. в случае оценки вероятности различий выборочной и генеральной средних существует вероятность того, что выборочная и генеральная 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 94

Интервальные оценки среднего арифметического

	$M=25,2$	$m=3,1$	$n=50$
Критерий Стьюдента t	1	2	3
Доверительная вероятность	63,8%	95,5%	99,7%
Уровень значимости P	0,32	0,05	0,01
Доверительный интервал $t m$	$\pm 3,1$	$\pm 6,2$	$\pm 9,3$
Предельная ошибка выборки Δ	$25,2 \pm 3,1$	$25,2 \pm 6,2$	$25,2 \pm 9,3$
Доверительные пределы	$28,3 \div 22,1$	$31,4 \div 19,0$	$34,5 \div 15,9$

Если выборки небольшие по объему, то 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, т.е. соответствия генеральной совокупности. Уровень значимости находится в убывающей зависимости 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 статистической значимости $P \leq 0,05$ содержит довольно большую вероятность ошибки (5%). Поэтому в тех случаях, когда требуется особая уверенность в достоверности полученных результатов, в медицинской статистике 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. 95.

Таблица 95

Интерпретация уровней значимости (P)

Показатели значимости (P)	Интерпретация
$\geq 0,1$	Данные согласуются с нулевой гипотезой (H_0).
$> 0,05$	Есть сомнения как в нулевой (H_0), так и в альтернативной гипотезе (H_1).
$< 0,05$	Нулевая гипотеза (H_0) может быть отвергнута.
$\leq 0,01$	Нулевая гипотеза (H_0) может быть отвергнута. Сильный довод.
$\leq 0,001$	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 обратно пропорционально зависит от числа наблюдений в выборке (n). Интервальные оценки коэффициентов асимметрии, эксцесса и коэффициента вариации, вычисление которых проводится на основе стандартных ошибок этих коэффициентов, сохраняют эту закономерность.

Вычисление ошибки показателя асимметрии может производиться по

формуле:
$$m_{As} = \pm \sqrt{\frac{6}{n} - (1 + Cv^2)}$$

Ошибка показателя эксцесса:
$$m_{Ex} = \pm \sqrt{\frac{24}{n}} = 2m_{As}$$

Ошибка коэффициента вариации:
$$m_{Cv} = \pm \frac{Cv}{n} \sqrt{\frac{1}{2} + \left(\frac{Cv}{100}\right)^2}$$

В медико-биологических исследованиях объекты наблюдения, как правило, весьма переменчивы. 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, что, в свою очередь, делает бессмысленным точное измерение исходных данных. Это обстоятельство необходимо учитывать при планировании исследований для предупреждения 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 арифметических $M_1 + M_2 + \dots + M_n$ можно использовать формулу: $m = \pm \sqrt{m_1^2 + m_2^2 + \dots + m_n^2}$.

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 $M_1 - M_2 - \dots - M_n$ вычисляется аналогично: $m = \pm \sqrt{m_1^2 + m_2^2 + \dots + m_n^2}$.

Ошибка деления средних арифметических: $m = \pm \frac{M_2}{M_1} \sqrt{\left(\frac{m_1}{M_1}\right)^2 + \left(\frac{m_2}{M_2}\right)^2}$.

Ошибка произведения средних арифметических: $m = \pm M_1 M_2 \sqrt{\left(\frac{m_1}{M_1}\right)^2 + \left(\frac{m_2}{M_2}\right)^2}$.

7. Статистические 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 распределением называют распределение, которое выбирается как образец (стандарт) для описания закона фактического, эмпирического распределения, полученного в опыте. Описание закона распределения может производиться с помощью математических формул или графических отображений.

К задачам, решаемым на основе анализа соответствия эмпирического и теоретического распределений, следует в первую очередь отнести:

- обоснование выбора конкретных методов статистического анализа генеральной и выборочной совокупностей;
- изучение причин, обуславливающих тот или иной закон распределения эмпирических данных (данных, полученных опытным путем);
- вскрытие на этой основе 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, в основе которых лежат, соответственно, дискретные и непрерывные величины.

Наиболее распространенные среди дискретных распределений – биномиальное и пуассоновское распределения. Среди непрерывных особенно широко используются при построении доверительных интервалов и статистической проверке гипотез: нормальное (Гауссово), а также связанные с ним распределения Стьюдента, χ^2 -квадрат и F – распределение Фишера.

7.1. Биноминальное распределение и распределение Пуассона

Биноминальное распределение (*распределение Бернулли*) – дискретное распределение вероятностей случайной величины, принимающей целочисленные значения K с вероятностями $P_K = \binom{n}{K} P^K (1-P)^{n-K}$, где P – параметр биномиального распределения, иногда называемый вероятностью положительного исхода; одно из основных распределений вероятностей, порождаемых конечным множеством независимых случайных экспериментов (испытаний). Это распределение возникает тогда, 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 невозможна, поскольку изделие, прошедшее испытание, не подлежит дальнейшему использованию. Поэтому для контроля из партии на удачу выбирают определенное количество образцов изделий (n). Эти образцы всесторонне проверяют, 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 о качестве всей партии изделий лежит сравнение распределения результатов контроля, т.е. данных, полученных опытным путем, и теоретического распределения, при котором гарантируется 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, чем 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 биномиальное распределение связаны между собой и с некоторыми другими типами распределений. Среди них наиболее существенным представляется схожесть с нормальным распределением.

7.2. 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 (Gauss, 1777-1855), исследовавшего его в начале XIX века.

Возьмем некоторое биномиальное распределение с параметром $K=9$ (см. раздел «Биномиальное распределение»). Затем увеличим этот параметр ($K=17$). При этом пропорционально уменьшим ширину каждого интервала (разряда), чтобы второе распределение имело бы ту же ширину, что и первое. Кроме того, если пропорционально уменьшить и частоту каждого интервала, то оба распределения будут иметь одинаковый объем. После этих операций над вторым распределением (сужение каждого разряда и уменьшение их частот) оба распределения будут лишь несколько различаться своей формой (рис. 70).

Если продолжить эту процедуру, беря все большие значения ($K=33$), то в предельном значении $K \rightarrow \infty$ верхние стороны столбиков сольются в гладкую кривую, которую называют *кривой нормального распределения непрерывной величины*.

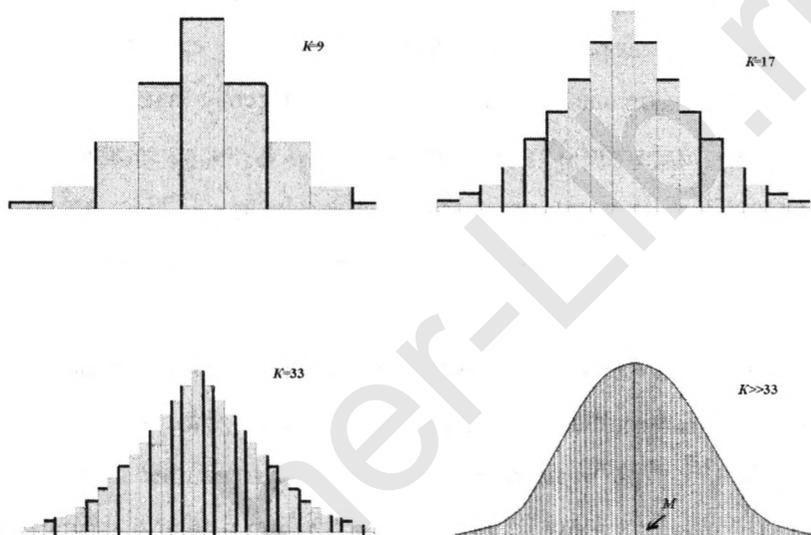


Рис. 70. Трансформация различных вариантов распределения

Как и всякую кривую, кривую нормального распределения можно описать некоторой функцией:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(M-x)^2}{2\sigma^2}}$$

где M – среднее теоретическое значение, e – основание натуральных логарифмов ($e=2,718..$), δ – среднеквадратическое отклонение.

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.

Такое распределение называется нормальным или гауссовым (распределением Гаусса). Следовательно, оно покоится на тех же предпосылках, что и биномиальное, т.е. на гипотезе о малых независимых, случайно сочетающихся факторах. Условие $K \rightarrow \infty$ 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.

Диаграмма нормального распределения симметрична относительно точки M , т.е. положительные и отрицательные отклонения равной величины от центра распределения (среднего арифметического) 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 δ – среднеквадратическое отклонение, характеризует степень сжатия или растяжения (плотности) диаграммы. Чем больше δ , тем шире заштрихованная область, обозначающая на рисунке часть вариант, покрываемых границами значений δ , а ее максимальная высота ниже. Кривая как бы расплывается в стороны. При малых δ заштрихованная область, наоборот, стягивается к середине. При этом подразумевается, что совокупности имеют одинаковый объем наблюдений (рис. 71).

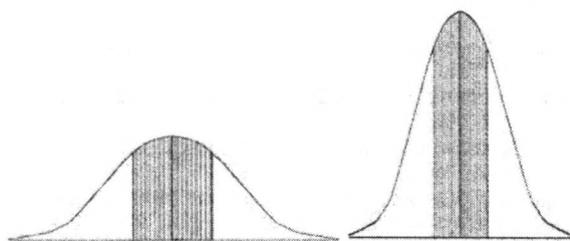


Рис.71. Различные варианты нормального распределения

На представленных рисунках заштрихованная площадь отражает часть вариант, отклоняющихся от среднего значения не более чем на δ , т.е. от $-\delta$ до $+\delta$. 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 δ . В пределах от -2δ до $+2\delta$ лежат 95,5% всех вариант. В пределах от -3δ до $+3\delta$ – 99,7% и т.д. Эта закономерность трактуется как *правило трех сигм*.

M и δ являются параметрами, описывающими различные модификации нормального распределения. Т.е. статистические критерии, при вычислении которых используют эти параметры, принято называть параметрическими критериями. Отсюда вытекает важное методическое условие применения параметрических критериев: исследуемая совокупность должна подчиняться закону нормального распределения.

Среди причин отклонения эмпирических распределений от нормального, достаточно часто встречающихся в медико-биологических исследованиях, является то, что в одну общую совокупность сводятся две или несколько нормальных совокупностей, каждая из которых имеет свои параметры распределения. Например, если 2 совокупности показателей физического развития мужчин и женщин объединить в одну совокупность, то в итоге нормального, Гауссова, распределения заведомо не получится, если даже каждая из 2-х исходных совокупностей данных подчинялась закону нормального распределения.

При этом если разница между средними арифметическими двух распределений небольшая, то при равных или близких среднеквадратических отклонениях вершина суммарного распределения будет тупой (рис. 72А). Если разница средних арифметических больше, чем среднеквадратическое отклонение каждой из них, то кривая распределения будет двухвершинной, «двугорбой» (рис. 72В). Такая форма распределения в статистике называется *бимодальной* и служит абсолютным доказательством того, что результаты обработки этого ряда будут некорректны из-за неоднородности исходных данных.

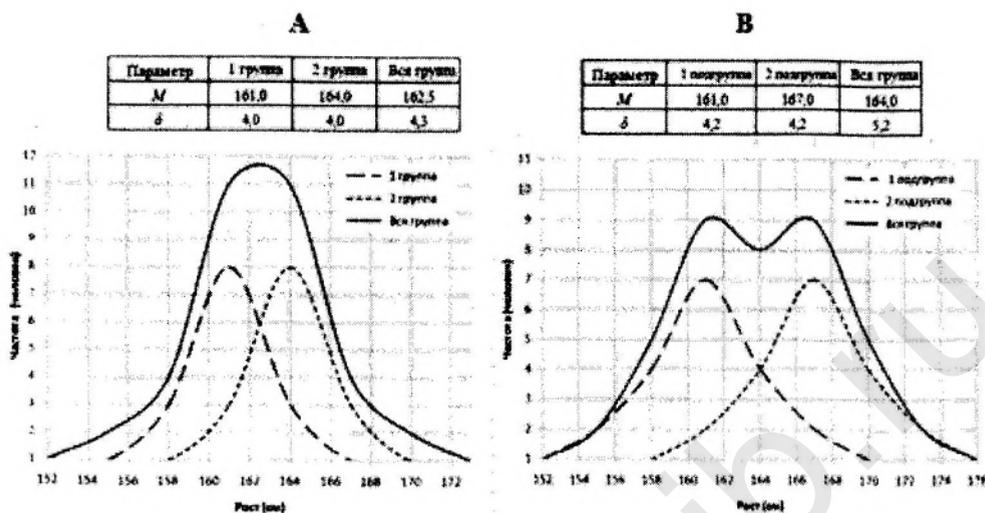


Рис. 72. Некоторые причины несоответствия эмпирических наблюдений и нормальных (Гауссовых) распределений

Понятно, что грамотный исследователь не будет объединять в одну группу характеристики физического развития мужчин и женщин. Понятно и то, что перечень возможных вариантов отклонения исследуемой совокупности от нормального распределения и возможные причины этих отклонений достаточно многообразны и не всегда лежат на поверхности. Вместе с тем, отклонение исследуемой совокупности от нормального распределения должно быть поводом для обязательного тщательного анализа причин этих отклонений и в случае объективной несовместимости исходных данных и закона нормального распределения поводом использования методов непараметрической статистики.

7.2.1. Нахождение нормального распределения с помощью MS Excel

Известно, что нормальное распределение задается двумя основными параметрами: центром распределения (средним арифметическим) и разбросом вариант (дисперсией или среднеквадратическим отклонением). Поэтому при формировании того или иного нормально распределенного ряда теоретических частот обязательно пользуются эти параметры.

Для примера построим ряд теоретических частот нормального распределения с помощью функции плотности нормального распределения. Для этого сформируйте таблицу исходных данных и итогов расчетов (рис. 73). В том числе:

1. Введите в ячейки А3:А9 и В3:В9 исходные данные.
2. Чтобы рассчитать среднее взвешенное $M = \frac{\sum VF}{n}$, установите курсор в ячейку В12 и наберите формулу =СУММПРОИЗВ(А3:А9; В3:В9)/В10.

	А	В	С	Д	Е
1	Варианты	Фактические		Теоретические	
2	<i>V</i>	Частоты	Частоты	Частоты	Частоты
3	21	2	0,032	0,022459	1,4
4	22	9	0,145	0,079033	4,9
5	23	6	0,097	0,17854	11,1
6	24	11	0,177	0,258926	16,1
7	25	18	0,290	0,241062	14,9
8	26	15	0,242	0,144078	8,9
9	27	1	0,016	0,055282	3,4
10	Итого	62	1,000		60,7
11					
12	<i>M</i> =	24,34			
13	<i>D</i> =	2,3			
14	СКО=	1,50			
15					

Рис. 73. Нахождение параметров эмпирического и теоретического нормального распределения

3. Формулу вычисления дисперсии введите в ячейку В13: =СУММПРОИЗВ(А3:А9; А3:А9; В3:В9)/В10)-В12*В12.
4. Среднеквадратическое отклонение (СКО) находим через дисперсию. Для этого в ячейку В14 введите формулу =КОРЕНЬ(Д13).
5. Теоретические частоты находим с помощью функции НОРМРАСП. Для этого установите курсор в ячейку Д3, вызовите указанную функцию и заполните ею окна, как показано на рис. 74. Затем скопируйте содержимое ячейки F3 в блок ячеек D4:D9.

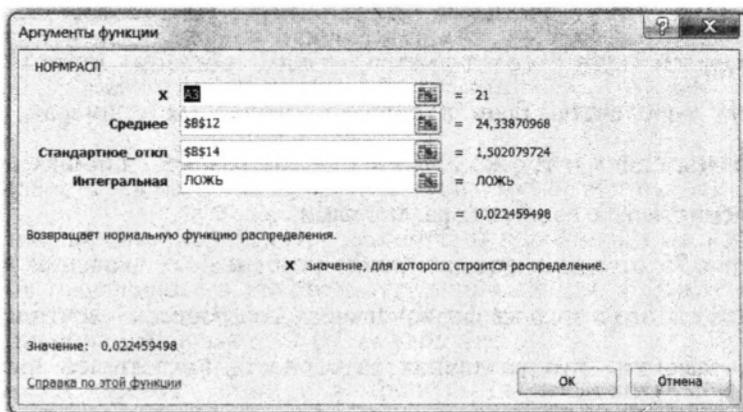


Рис. 74. Пример заполнения окна функции НОРМРАСП

6. Для того чтобы заменить теоретические частоты теоретическими частотами, в ячейку E3 введите формулу: =D3*\$B\$10. Затем скопируйте ее в остальные ячейки столбца E4:E9.

Обратите внимание, что сумма теоретических частот (60,7), расположенных в ячейках E3:E9, не равна сумме фактических частот (62,0). Эта незначительная погрешность объясняется точностью метода и находится в допустимых пределах ($\approx 2,1\%$).

7.3. Преобразования эмпирических распределений.

Логит- и Пробит-анализ

Соответствие распределения эмпирических, т.е. полученных опытным путем, наблюдений известным законам распределения, и в первую очередь закону нормального (Гауссова) распределения, существенно ограничивает аналитические возможности статистической обработки информации. Это несоответствие можно устранить, если использовать замену численных характеристик оцениваемого статистического признака. Например, с точки зрения формальной логики, в качестве единицы счета частоты сердечных сокращений можно принять не число ударов пульса, а обратную единицу – время между ударами пульса. Или для статистики распределения каких-либо объектов шаровидной формы единицей счета можно выбрать диаметры шаров

или их веса (при плотности шара= $const$). Как известно, вес шара пропорционален кубу его диаметра, таким образом можно перейти к кубической мере счета. Если в обоих приведенных примерах представить распределения одних и тех же объектов в этих разных единицах счета, то эти распределения неизбежно будут различными.

На рис.75 отражено распределение усредненных значений двух разных показателей одного и того же физиологического процесса – вентиляции легких. Нетрудно заметить, что различная размерность показателей интенсивности вентиляции легких ведет к прямо противоположной тенденции их распределения.

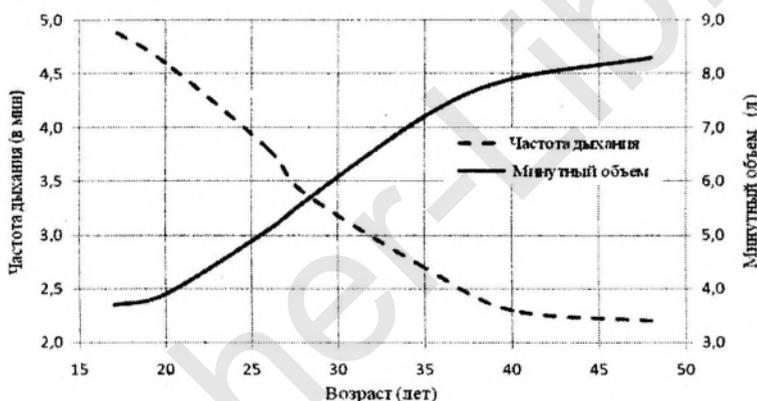


Рис.75. Распределение показателей вентиляции легких в зависимости от возраста мужчин

Существенные изменения характера статистических распределений происходят и тогда, когда единицам счета придают другую размерность с помощью математических преобразований: $x' = \frac{1}{x}$, $x' = \sqrt{x}$, $x' = \ln(x)$ и т.д. В ряде случаев основой таких преобразований служат более сложные функции⁵⁸.

Такой подход используется в статистике для приведения «неудобных» форм рядов наблюдений в формы, удобные для аналитической обработки. Эта операция, называемая *линеаризацией* кривых, осуществляет преобразование

⁵⁸ Обосновать правомерность такого рода преобразований современная наука может не всегда, и их допустимость зачастую доказывается эмпирически, т.е. опытным путем.

сложных аналитических кривых в прямую линию, которая доступна обработке относительно простыми методами регрессионного анализа линейной зависимости.

Формально сущность линеаризации заключается в том, что исходные результаты наблюдений переводятся из системы координат x y в координаты, определяемые производными некоторых функций от x и y , между которыми существует линейная зависимость $F(y) = a + bf(x)$.

Указанный метод лежит в основе применения логит-анализа (логистического анализа), использующего результаты логарифмических преобразований исходных данных. При этом чрезвычайно важным является то, что логистические кривые, т.е. кривые логарифмического распределения, отражают законы роста и умирания – фундаментальных явлений, присущих всему живому. По логистическому закону происходит увеличение массы особи, ее отдельных органов, рост численности различных популяций и т.п. Обратный процесс – уменьшение численности популяции, понижение жизнеспособности особей, их деградация – происходит по обратной логистической кривой. Таким образом, отражается динамика процессов зарождения, развития и умирания всех известных форм жизни. Это важное свойство всего живого дает возможность, используя логистический закон, находить аналитическим путем критические точки смертельных (LD) и (или) эффективных доз (E) различных препаратов, лучевых облучений и т.п.⁵⁹

Одним из способов практической реализации такого подхода, когда в одном и том же наблюдении одни единицы измерения заменяются другими единицами, обладающими другими свойствами, является пробит-анализ. В отличие от логит-анализа, иногда называемого логистическим анализом, в основе которого лежит использование логистического распределения, пробит-анализ может использовать и параметры нормального распределения качественных бинарных признаков (да – нет; есть эффект – нет эффекта; есть

⁵⁹ Методика логистического анализа в данном издании не рассматривается.

гибель – нет гибели и т.п.). Понятно, что в статистике, оперирующей только групповыми оценками, рассматриваются вероятности, вычисляемые для некоторой совокупности подопытных объектов⁶⁰.

Например, при изучении способности ионизирующего излучения вызывать летальный исход наиболее часто критерием оценки действия на биологические объекты является гибель биологического объекта – интегральный показатель, финал многочисленных нарушений. На практике для того, чтобы дать оценку летальному действию этого фактора, несколько групп животных подвергают воздействию разных доз и для каждой группы (т.е. для каждой дозы) находят процент погибших животных. Ту дозу, при которой погибает половина (50%) подопытных животных, принимают за усредненную характеристику летального действия ионизирующего излучения для данного вида животных. Эту характеристику обычно обозначают LD_{50} (LD_{50}).

Проблема заключается в том, что при практическом проведении такого эксперимента обычно ни одна из применяемых доз точно не соответствует LD_{50} , поэтому эта доза получается каким-либо расчетным путем, в том числе и с помощью уравнения линейной регрессии, для чего необходимо привести данные к форме, позволяющей получать распределение в линейном виде.

В ходе пробит-анализа графическим способом, при построении диаграммы распределения результатов воздействия ионизирующего излучения, по оси абсцисс откладывают величины доз облучения, а по оси ординат – частоты гибели биологических объектов, облученных в различных дозах. Итоговая диаграмма в этом случае называется *дозовой кривой гибели (смертности)*. Если по оси ординат откладывать не частоту погибших, а выживших особей, то диаграмма будет носить название *дозовая кривая выживаемости*. Обычно такого рода диаграммы отражают достаточно сложные зависимости, статистическая обработка которых весьма затруднена.

⁶⁰ Срок наблюдения, в течение которого наблюдают за гибелью (или выживаемостью), зависит, в частности, от вида биологического объекта. Так, при исследовании летального действия ионизирующего излучения на большинство видов позвоночных животных срок наблюдения составляет обычно 30 или 60 суток (для мышей и крыс – 30 суток).

С целью решения этой проблемы и выполняются определенные преобразования исходной информации, в частности, логарифмическое преобразование.

Дозовая кривая гибели⁶¹ после логарифмического преобразования исходных данных обычно имеет типичную S-образную форму (рис. 76).

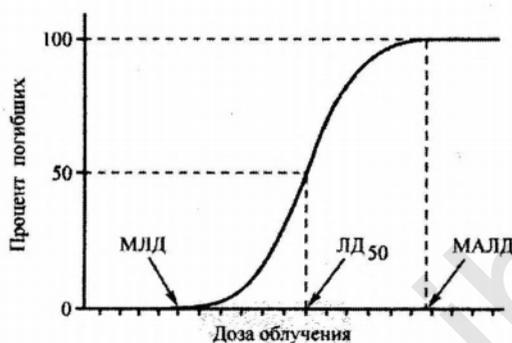


Рис. 76. Дозовая кривая гибели облученных животных

На ней можно выделить несколько характеристических точек. Одна из них — это точка, значение ординаты которой составляет 50% гибели облученных организмов. Величина дозы облучения, соответствующая абсциссе этой точки, называется *полуметальной дозой*. Таким образом, *полуметальная доза* — это доза облучения, при действии которой погибает 50% облученных организмов. Полуметальную дозу (при использовании кириллицы) обозначают: $ЛД_{50}$ или $ЛД_{50/30}$, $ЛД_{50/60}$, где в знаменателе подстрочного индекса указывают срок (в сутках), в течение которого наблюдают за гибелью облученных организмов. Эта дозовая величина широко используется в радиобиологии для характеристики *радиочувствительности* (т.е. радиопоражаемости) биологических объектов.

Вторая характеристическая точка дозовой кривой гибели имеет в качестве абсциссы т.н. *минимальную абсолютно летальную дозу* (сокращенно обозначают, как $МАЛД$, $ЛД_{100}$, $ЛД_{100/30}$), т.е. минимальную дозу, вызывающую гибель всех облученных организмов (ордината — 100% гибели).

⁶¹ Из работы Платонова А.Г. и Ахалая М.Я. (2006).

Третья характеристическая точка дозовой кривой гибели имеет в качестве абсциссы т.н. *минимальную летальную дозу* (сокращенно обозначают, как *МЛД*), т.е. практически ту максимальную дозу, которая не вызывает гибели ни одного из облученных организмов (ордината – 0% гибели).

Строго говоря, термины «минимальная абсолютно летальная доза» и «минимальная летальная доза» не совсем правомерны, т.к. кривая гибели приближается к отметкам 100% и 0% асимптотически (т.е. бесконечно). Поэтому обычно за *МАЛД* принимают дозу, вызывающую гибель 90 или 99% объектов, а за *МЛД* – дозу, вызывающую гибель 1% или 10% объектов.

Дозы облучения, лежащие в диапазоне от *МЛД* до *МАЛД*, называют *летальными дозами*. Дозы, превышающие *МАЛД*, определяют, как *сверхлетальные*, а дозы, лежащие ниже *МЛД*, – как *сублетальные*.

В фармакологических и токсикологических исследованиях препаратов рассчитывают эффективные дозы, действующие в 50% случаев (ED_{50}), а также ED_{16} и ED_{84} и токсические дозы: LD_{16} , LD_{50} , LD_{84} и LD_{100} .

Зная LD_{16} и LD_{84} , можно рассчитать стандартную ошибку LD_{50} по формуле:

$$m_{LD_{50}} = \frac{LD_{84} - LD_{16}}{\sqrt{2N}}$$

где N – количество животных в пределах групп, использованных для испытания этих доз.

Решение задачи определения значений перечисленных доз затрудняется тем, что измерение летальной или эффективной дозы воздействия изучаемого фактора для отдельной особи практически невозможно, т.к. итог воздействия этих доз проявляется только через несколько дней или даже недель. Понятно, что если бы изучаемый фактор обладал свойством постепенного накопления, то можно было бы, прибавляя через определенные промежутки времени дополнительную порцию, зафиксировать дозу, при которой проявляется желаемый эффект. Однако полного накопления обычно не происходит, т.к. в живом организме, как правило, протекают защитные процессы,

дезактивирующие чужеродный фактор. В таких условиях методом пробит-анализа можно рассчитать дозы, вызывающие любой процент встречаемости такого эффекта.

При использовании метода пробит-анализа логарифмическое преобразование данных, располагаемых по оси X координатной сетки диаграммы, дополняется преобразованием и оси ординат (Y). На ней откладываются не частоты (P) встречаемости наблюдаемого эффекта (например, гибели), а значения статистической функции $\Psi(P)$, соответствующие этим частотам (P). Таким образом, координатная сетка диаграммы определяется осью абсцисс с десятичными или натуральными логарифмами доз и осью ординат, где отражается функция $\Psi(P)$ -- величина, обратная интегралу вероятностей⁶² частот бинарных событий.

Существует несколько вариантов пробит-анализа, которые, в конечном счете, различаются адекватностью параметрам исходной информации и набором результирующих показателей⁶³. Все различия существующих ныне вариантов пробит-анализа обычно сводятся к следующему:

1. Логарифмические преобразования шкалы доз воздействия (концентраций) проводятся на основе разных логарифмов или вообще не проводятся.
2. Расчеты проводятся без учета или с учетом «веса» экспериментальных точек.
3. Дозы, вызывающие 0% и 100% эффекта, либо вовсе исключаются из расчетов, либо для их обработки применяются разные подходы.
4. Для выравнивания аналитической кривой применяются различные методы.
5. Определение ошибок репрезентативности и, соответственно, доверительных интервалов проводится не только для ЛД₅₀, но и для других доз.

⁶² Отсюда и термин «*пробит*» (от англ. «*probability*» – вероятность).

⁶³ Учет этого обстоятельства особенно важен, если существует необходимость сопоставлять конечные результаты с результатами других, аналогичных исследований, выполненных в рамках определенного набора результирующих показателей.

На практике реализация пробит-анализа осуществляется с помощью специальных статистических программ. Однако эти программы зачастую довольно сложны и неудобны для обработки данных медико-биологических экспериментов. Использование разного рода «самоделок» в виде макросов и тому подобных приложений для *Excel*, во-первых, не позволяет реализовать многие аналитические возможности пробит-анализа. Во-вторых, чреваты возможностью получения ошибочных результатов.

В то же время внимание начинающих пользователей пробит-анализа, особенно на этапе первичного освоения этого метода или при необходимости быстрого получения ориентировочных данных, могут привлечь относительно простые приемы работы в *Ms Excel* – широко используемого программного пакета, способного осуществлять расчеты с применением широкого комплекса методов статистической обработки, хранения и визуализации данных.

7.3.1. Выполнение пробит-анализа в *Ms Excel*

В качестве *примера* реализации пробит-анализа в *Ms Excel* проведем обработку экспериментальных данных рентгеновского облучения 7 групп подопытных мышей⁶⁴. Для этого создадим специальную таблицу исходных данных (рис. 77). После заполнения заголовков столбцов на листе *Ms Excel* в ячейках A1:D1 выполним ввод числовой информации:

- в ячейках A2:A8 расположим дозы облучения (в рентгенах);
- в B2:B8 – число мышей в каждой из подопытных подгрупп;
- в C2:C8 – число погибших мышей в каждой из подгрупп;
- в D2:D8 введем расчетные данные – частоту гибели мышей в каждой

из подопытных групп. Чтобы получить эти частоты, в ячейку D2 введите формулу $=C2/B2$ и нажмите клавишу ENTER. После этого скопируйте эту формулу в ячейки D3:D8. Для этого щелчком мыши выделите ячейку D2, затем, установив указатель мыши на нижний правый угол ячейки, не отпуская клавишу мыши, перетащите содержимое ячейки D2 в ячейки D3:D8.

⁶⁴ Исходные данные условные.

Итогом этих операций должна стать таблица исходных данных, изображенная на рис. 77.

Для проверки характера распределения исходных данных построим, используя возможности *Ms Excel*, точечную диаграмму, в варианте «с гладкими кривыми», распределения частот гибели опытных групп (рис. 78)⁶⁵.

На этой диаграмме величины доз (столбец А) отражены по горизонтальной оси, частоты гибели животных (столбец D) – по вертикальной.

	А	В	С	Д
1	Доза	Животных	Погибли	P
2	350	32	2	0,0625
3	400	35	6	0,1714
4	450	34	16	0,4706
5	500	37	32	0,8649
6	550	36	34	0,9444
7	600	31	30	0,9677
8	650	32	31	0,9688

Рис.77. Расположение данных для пробит-анализа в *MS Excel*

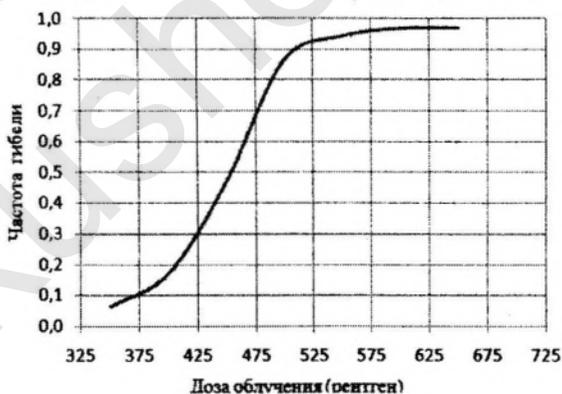


Рис. 78. Кривая гибели животных опытных групп

⁶⁵ См. раздел «Построение диаграмм в *MS Excel*».

Чтобы выполнить пробит-анализ, необходимо провести аналитические преобразования рядов исходных данных. В том числе, частоты преобразовать с помощью функции $\Psi(P)$, а дозы перевести в логарифмические меры. В качестве логарифмической меры возьмем натуральные логарифмы⁶⁶.

Для выполнения этих операций нужно обозначить названия блоков ячеек, где будут расположены итоги преобразований данных. Соответственно, в ячейке A11 – «Доза (Ln)», а в ячейке B11 – « $\Psi(P)+5$ ». Затем вводим в ячейку A12 формулу =LN(A2), а в ячейку B12 – формулу =НОРМСТОБР(D2/1)+5. После заполнения этих ячеек копируем их содержимое в ячейки A13:A18 и B13:B18. Результат должен соответствовать рис. 79.

	A	B	C	D
1	Доза	Животных	Погибли	P
2	350	32	2	0,0625
3	400	35	6	0,1714
4	450	34	16	0,4706
5	500	37	32	0,8649
6	550	36	34	0,9444
7	600	31	30	0,9677
8	650	32	31	0,9688
9				
10				
11	Доза (Ln)	$\Psi(P)+5$		
12	2,54	3,47		
13	2,60	4,05		
14	2,65	4,93		
15	2,70	6,10		
16	2,74	6,59		
17	2,78	6,85		
18	2,81	6,86		
19				

Рис. 79. Расположение результатов преобразований

Используя информацию, расположенную в ячейках A12:B18, получаем итоговую диаграмму, отражающую выполненные преобразования (рис. 80).

⁶⁶ Выбор именно этой меры в данном случае продиктован удобством отображения хода операции.

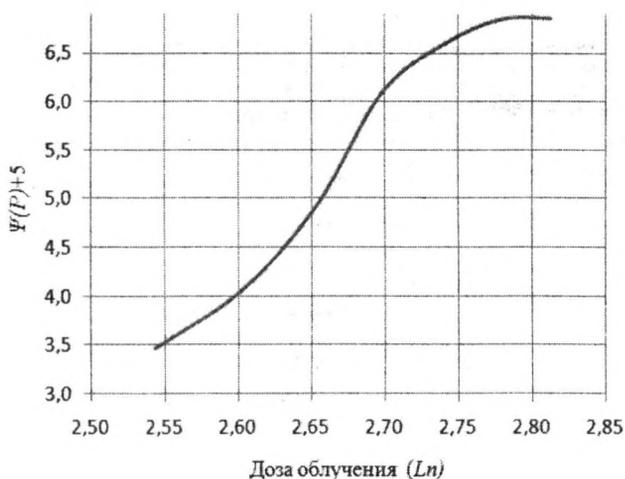


Рис.80. Кривая распределения преобразованных данных

Для получения значения LD_{50} через линейную регрессию проведем обработку полученной аналитической кривой с помощью функции *MS Excel* «Добавить линию тренда». Для этого выделите мышкой сформированную кривую, нажатием на левую клавишу мышки вызовите меню мастера диаграмм (Рис. 81) и выделите закладку «Добавить линию тренда». В открывшемся окне установите параметры линии тренда «линейная», «показывать уравнение на диаграмме» и «поместить на диаграмму величину аппроксимации R^2 » (рис. 82).

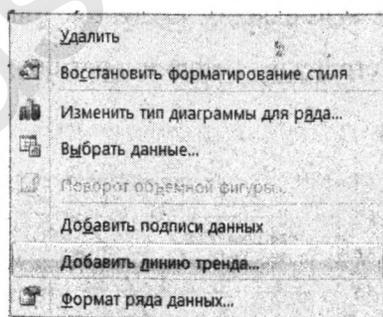


Рис.81. Меню мастера диаграмм

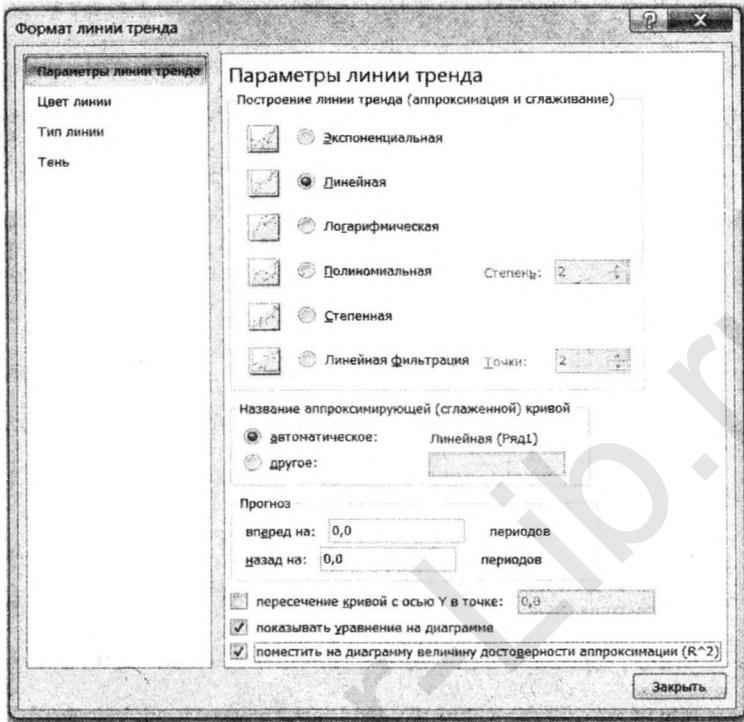


Рис.82. Параметры тренда линейной регрессии

После выполнения команды «Закреть», появится диаграмма линейной регрессии и уравнение регрессии, полученные путем аппроксимации распределения методом наименьших квадратов (рис. 83). Показатель R^2 в данном случае демонстрирует детерминированность (точность линейной аппроксимации).

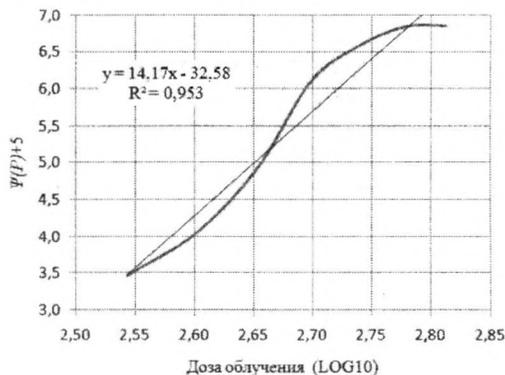


Рис.83. Диаграмма и уравнение линейной регрессии

На пересечении абсциссы $\Psi(P)+5''=5,0$, что соответствует $P=0,5$ или 50%, ордината равна 2,65. Поскольку это число – есть десятичный логарифм дозы облучения, через функцию, обратную натуральному логарифму (в *Excel* эта функция обозначена, как **EXP**)⁶⁷, находим значение дозы ЛД₅₀, используя формулу $=\text{EXP}(2,65*2,303591)\approx 448,8$ рентгена.

Понятно, что в данном случае прямое использование осей диаграммы, может дать неточные значения ЛД₅₀. Его точное значение получают расчетным путем, опираясь на параметры полученной линейной регрессии. Из уравнения линейной регрессии $y=14,17x-32,58$ путем несложных алгебраических преобразований получаем формулу для определения десятичных логарифмов доз облучения (x) при разных значениях пробитов (y). Для пробита, равного 5 (ЛД₅₀), $x=(5+32,58)/14,17=2,65$, что соответствует определенной ранее дозе облучения 448,8 рентгена.

Таким же расчетным путем, через параметры линейной регрессии, определяют необходимые для вычисления ошибки ЛД₅₀ значения ЛД₁₆ и ЛД₈₄, которым, исходя из функции преобразований доз облучения, соответствуют пробиты $\Psi(0,16)+5=4,1$ и $\Psi(0,84)+5=6,0$.

$$\text{ЛД}_{16}=\text{EXP}((4,1+32,58)/14,17)=387,8 \text{ рентгена};$$

$$\text{ЛД}_{84}=\text{EXP}((6,0+32,58)/14,17)=528,0 \text{ рентгена}.$$

Отсюда ошибка ЛД₅₀, $m_{\text{ЛД}50}=(528,0-387,8)/\sqrt{2*106}=9,6$ рентгена, где 106 – число животных, облученных дозами, находящимися в интервале 387,8-528,0 рентгена.

⁶⁷ При необходимости потенцировать десятичный логарифм можно воспользоваться соотношением $\text{LN}(x)=\text{LOG}10*2,302591$.

8. Сравнение статистических распределений, нахождение различий

Среди задач статистических оценок совпадений рядов распределений выделяются две группы задач:

- 1) задачи оценок совпадения (согласия) или, что практически одно и то же, расхождения эмпирических и теоретических распределений;
- 2) задачи оценок различий эмпирических распределений.

Для того, чтобы более четко обозначить разницу между критериями, с помощью которых решаются две эти группы задач, критерии оценок совпадений эмпирических и теоретических распределений принято называть критериями согласия.

8.1. Критерии согласия эмпирических данных и теоретических распределений

Оценка меры совпадения эмпирических данных, полученных опытным путем, и теоретических распределений, т.е. оценка соответствия результатов опыта известным законам распределений – одна из ключевых задач многих статистических исследований. Связано это, как минимум, с двумя обстоятельствами.

Во-первых, широким использованием при медико-биологических исследованиях методов параметрической статистики, приложимых только к нормально распределенным рядам наблюдений.

Во-вторых, во многих случаях соответствие одному закону распределения двух или более разных, на первый взгляд, выборочных совокупностей может служить доказательством равенства этих выборок между собой. В общем здесь действует логика: если $A=B$ и $B=C$, значит $A=C$. Иначе говоря, отличия этих выборок случайны, и на самом деле они принадлежат одной генеральной совокупности.

Самый распространенный дефект статистического анализа в научных работах медико-биологического профиля – оценка различий производных величин без учета соответствия исходных рядов наблюдений известным

законам статистических распределений. *Например*, прямое применение средних и дисперсий, а также математических формул и сложных алгоритмов, предусматривающих их использование, без проверки согласия исходных данных с законом нормального распределения.

Единственный способ избежать необходимости такого рода проверки – использование методов статистики, не требовательных к параметрам распределения данных, так называемых методов *непараметрической статистики*. Однако использование методов *параметрической статистики*, т.е. методов, основанных на параметрах нормального распределения, должно предваряться оценкой соответствия эмпирических совокупностей (опытных выборок) закону нормального, Гауссова, распределения.

Для оценивания меры соответствия (расхождения) эмпирических данных и каких-либо теоретических распределений применяются различные статистические критерии согласия, *например*, критерий Колмогорова-Смирнова, *хи-квадрат* Пирсона и др. Наглядную картину дает и визуальная проверка схожести распределений с помощью гистограммы, которая показывает частоту попаданий вариант ряда распределения в отдельные интервалы.

Существует большое количество самых разнообразных критериев согласия. Это разнообразие определяется большим разнообразием конкретных задач, решаемых такого рода критериями. Это могут быть задачи оценки расхождений теоретической и эмпирической плотности распределений, расхождения функций распределения вероятностей, расхождения параметров сравниваемых распределений и т.п. Различаются критерии согласия и по видам теоретических распределений, берущихся за основу анализа. Понятно, что основой расчета здесь является наличие теоретического распределения с заданными параметрами. В то же время практическое построение такого теоретического распределения является трудоемкой задачей. Например, для нахождения нормального (Гауссова) распределения, наиболее часто употребляемого в статистических расчетах, требуется использовать формулу

плотности нормального распределения с применением некоторых исходных параметров эмпирического распределения:

$$\varphi(x) = \frac{N}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Использование встроенных функций *MS Excel*, как и возможностей специальных программ статистической обработки данных, значительно облегчает задачу.

В ряде случаев ориентировочную оценку расхождения или совпадений распределений позволяет давать графический метод.

8.1.1. Критерий Пирсона χ^2 (хи-квадрат)

Среди критериев согласия большое распространение получил критерий К. Пирсона – **критерий χ^2** (хи-квадрат). Одной из причин его популярности является возможность 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 $n \rightarrow \infty$ 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). Формула для расчета величины критерия хи-квадрат:

$$\chi^2 = \sum \frac{(P - P'')^2}{P''}$$

где P – фактические, эмпирические частоты или численности; P'' – теоретические частоты или численности.

Практическое вычисление критерия χ^2 (хи-квадрат) на конкретных данных, когда в качестве теоретического рассматривается нормальное (Гауссово) распределение, представлено в табл. 96⁶⁸.

Таблица 96

Пример вычисления критерия χ^2 (хи-квадрат)

Масса тела детей (кг)	Частоты (число детей)		Объединенные частоты		$\frac{(P-P'')^2}{P''}$
	Эмпирич. P	Теоретич. P''	Эмпирич. P	Теоретич. P''	
21	2	1,4			
22	9	4,9	11	6,3	3,5
23	6	11,1	6	11,1	2,3
24	11	16,1	11	16,1	1,6
25	18	14,9	18	14,9	0,6
26	15	8,9	16	12,4	1,1
27	1	3,4			
Итого	62	58	62	60,7	9,1

Условием применения этого метода является наличие не менее 5 наблюдений (частот) в каждой клетке таблицы. Если в исходных данных это условие не выполняется, необходимо объединить, сгруппировать малые частоты. Нужно помнить, что любая группировка данных всегда таит угрозу снижения информативности статистики. Поэтому необходимо очень тщательно обосновывать выбор порядка укрупнения групп.

После укрупнения интервалов значение критерия вычисляется в следующей последовательности:

1. Вычисляется разность между частотами эмпирических P и теоретических P'' распределений.
2. Чтобы погасить отрицательные значения, эта разность возводится в квадрат.
3. Затем для уменьшения числового размера результата делится на величину эмпирических частот. Итог: $\chi^2=9,1$.

В целом расчетная величина критерия χ^2 тем меньше, чем меньше различаются парные частоты сравниваемых распределений.

⁶⁸ Вычисление MS Excel теоретических частот нормального распределения см. на рис. 73 и рис. 74.

Число степеней свободы k , которое необходимо знать для оценки справедливости нулевой гипотезы, устанавливается в зависимости от вида распределения. При биномиальном (Пуассоновом) распределении $k=S-2$, при нормальном $k=S-3$. Где k – число степеней свободы, S – число групп вариант с учетом объединения групп.

Полученное значение критерия χ^2 сравнивают с табличным или найденным с помощью компьютера⁶⁹ теоретическим значением при заданном уровне значимости. В случае, когда вычисленный критерий χ^2 больше табличного (теоретического) значения критерия, нулевая гипотеза, которая предполагает соответствие эмпирического и теоретического распределений, отвергается.

Обратите внимание, что встроенная в MS Excel функция ХИ2РАСП дает значения односторонней вероятности. Получение значений двухсторонней вероятности путем удваивания односторонней P возможно только при симметричном распределении исходных данных.

Способом определения достоверности различий распределений, с помощью которого можно выполнять оценку вручную и без использования таблиц, является оценка χ^2 по правилу Романовского. Согласно этому правилу, нулевая гипотеза отвергается, если неравенство $(\chi^2 - k) / \sqrt{2k} \geq 3$, т.е. если неравенство будет равно или больше 3. В данном примере $(9,1-2)/2 \approx 3,6$. Следовательно, опытное эмпирическое распределение не отличается от нормального.

Следует отметить, что критерий χ^2 , как и другие критерии согласия, применяемые к группированным данным, обладает существенным недостатком. Он не позволяет иногда обнаружить реально существующие различия, поскольку некоторые из них «скрадывает» группировка, которая осуществляется исследователем в ходе подготовки рядов распределений к анализу. Пример такого последствия перегруппировки данных приведен

⁶⁹ В Ms Excel функция ХИ2РАСП.

в табл. 97, когда укрупнение интервала в первой подгруппе наблюдения привело к полному исчезновению очевидных различий в структуре сравниваемых групп наблюдения.

Таблица 97

Сравнительное распределение показателей структуры групп наблюдения (в % к итогу)

Исходные группы наблюдения			После дополнительной группировки		
Возраст (лет)	Мужчин	Женщин	Возраст (лет)	Мужчин	Женщин
до 20	16	5			
20-29	17	28	до 29	33	33
30-39	34	34	30-39	34	34
40 и ст.	33	33	40 и ст.	33	33
Итого	100	100	Итого	100	100

Методики вычисления критерия χ^2 , которые могут использоваться для решения задач оценки согласия эмпирических распределений, представлены в разделах «Оценка различий эмпирических распределений с помощью *Excel*», «Коэффициенты сопряженности Пирсона (С) и Чупрова (К)», а также «Вычисление критерия сопряженности в *MS Excel*».

8.2. Критерии согласия эмпирических распределений

8.2.1. Критерий Мостеллера

Этот непараметрический критерий, один из самых простых по технологии применения, позволяет сравнивать распределение однородных наблюдений в двух выборках одинакового, в том числе и малого объема ($n \approx 5$). Достоинство критерия: не требует каких-либо вычислений вообще (!). Недостаток: имеет низкую мощность и может быть использован только для быстрой, грубой проверки статистической достоверности сдвига сравниваемых распределений.

Гипотеза о равенстве двух выборок отклоняется с доверительной вероятностью 0,95 ($P=0,05$), если 5 (при $n < 25$) или 6 (при $n > 25$) наибольших или наименьших значений содержатся в одной и той же выборке.

Например, очевидно, что из двух выборок X_1 и X_2 при $n=12$ (табл. 98) 7 наибольших значений принадлежат выборке X_2 (21, 22, 23, 25, 26, 28, 29). Это

число превышает критическое значение критерия Мостеллера, для такой выборки равное 5. Следовательно, можно признать наличие статистически значимой разницы двух выборок при $P=0,05$.

Таблица 98

Ряды распределения двух выборок случайных величин

X_1	11	15	20	19	18	16	20	18	14	12	13	17
X_2	11	21	22	19	26	15	25	29	23	19	12	28

8.2.2. Критерий Розенбаума

Является аналогом критерия Мостеллера, поэтому им обоим присущи одинаковые достоинства и недостатки. Может использоваться для быстрой, приближенной сравнительной оценки распределения однородных наблюдений в двух выборках одинакового объема.

Если число наблюдений не менее 16 ($n \geq 16$) и не менее 5 наблюдений значений одной выборки находятся вне размаха второй выборки, то нулевая гипотеза об отсутствии сдвига на уровне достоверности $P=0,05$ отклоняется. Если число наблюдений не менее 20 ($n \geq 20$) и не менее 7 наблюдений значений одной выборки находятся вне размаха другой выборки, то нулевая гипотеза об отсутствии сдвига отклоняется при $P=0,01$.

8.2.3. Критерий знаков (Z)

Критерий знаков (Z) – простой непараметрический критерий, основанный на подсчете числа пар однонаправленных эффектов в парных выборках. При его определении учитываются не числовые значения, а только лишь знак различия попарно связанных вариантов. Например, при пульмонологическом обследовании в группе пациентов, состоящей из 20 человек, у 5 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 («плюс», если есть бронхоспазм; и «минус», если нет), которые меньше всего встречались. В приведенном примере таких случаев 5. Критическое значение

критерия знаков (см. Приложение 6) 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 целом нет), если бы число эпизодов было не менее 6. Поскольку это условие не выполнено (фактическое число различий 5), можно считать появление 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), притом 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.

8.2.4. Критерий Вилкоксона U (Вилкоксона – Манна – Уитни)

Существует множество вариантов этого критерия, отличающихся методиками вычислений. Соответственно имеется множество различных вариантов наименований по авторам.

Односторонний критерий Вилкоксона – Манна – Уитни позволяет сравнивать 2 независимые выборки с разным числом наблюдений на основе их ранговых распределений. Таким образом, существенными являются не числовые значения наблюдений, а их ранговое расположение в ряду. *Например:* при оценке частоты пульса у мужчин и женщин в конце рабочего дня были получены следующие показатели: у мужчин – 72, 70, 74, 73, 80, 79, 76; у женщин – 75, 78, 71, 77 уд./мин. В табл. 99 представлен ранжированный ряд этих значений. Различия между двумя группами можно считать наиболее существенными, если после объединения обеих исходных групп в один ряд все значения одной исходной группы будут располагаться после значений второй группы. Однако в реальном распределении существуют некоторые отклонения – *инверсии*. Число инвер-

сий подсчитывают через U . В исходных группах распределения значений пульса у мужчин и женщин (табл. 99) 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 (см. Приложение 7) 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 99

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

8.2.5. Критерий гомогенности Брандта-Снедекора

Это один из несложных критериев оценки принадлежности двух взвешенных выборок однородных единиц наблюдения к одной совокупности, т.е. определения схожести параметров этих выборок (средних, дисперсий и т.п.). Иначе говоря, этот критерий применяется для проверки гипотезы о том, что две группы наблюдений, представленных в виде двух взвешенных вариационных рядов, относятся к одной общей совокупности. Все расчеты выполняются только на частотах распределения, выраженных в абсолютных числах. Формально расчет ведется через критерий χ^2 (хи-квадрат), величина которого определяется по формуле:

$$\chi^2 = \frac{N_{xy}^2}{N_x N_y} \left(\sum \frac{f_x^2}{f_{xy}} - \frac{N_x^2}{N_{xy}} \right),$$

где N_x и N_y – суммы частот сравниваемых рядов; $N_{xy}=N_x+N_y$ – сумма частот обоих рядов; f – частоты первого ряда; f_{xy} – частоты объединенного ряда.

Пример определения гомогенности результатов измерений массы тела в двух группах детей (табл. 100).

Таблица 100

Сравнительное распределение результатов измерения массы тела в двух группах детей

Масса детей (кг)	Число детей		f_x^2	f_{xy}
	1 группа f_x	2 группа f_y		
21	2	11	4	13
22	9	15	81	24
23	6	21	36	27
24	11	32	121	43
25	18	17	324	35
26	15	14	225	29
27	1	9	1	10
Итого	62	119		181

Величина первой группы $N_x=62$ чел.; второй – $N_y=119$ чел., $N_{xy}=119+62=181$ чел. Подставляя исходные данные, получаем:

$$\chi^2 = \frac{181^2}{62 \cdot 119} \left(\frac{4}{13} + \frac{81}{24} + \frac{36}{27} + \frac{121}{43} + \frac{324}{35} + \frac{225}{29} + \frac{1}{10} - \frac{62^2}{181} \right) = 16,5$$

При числе степеней свободы $k=S-1=7-1=6$ значимость результата равна $P=0,01$. Т.о., нулевая гипотеза, свидетельствующая о том, что обе группы относятся к одной общей совокупности, т.е. об отсутствии статистически значимых различий между параметрами распределения двух групп, не подтвердилась.

8.2.6. Модифицированный критерий гомогенности хи-квадрат

Может применяться для оценки однородности распределения совокупности, состоящей из двух и более взвешенных выборок. Нулевой гипотезой считается гипотеза о том, что сравниваемые совокупности одинаково распределены.

Такая задача обычно ставится, поскольку однородность состава сравниваемых совокупностей является обязательным условием правомерности утверждения, что разные результаты не определяются неоднородностью сравниваемых совокупностей наблюдений. При этом для этого критерия не имеет никакого значения, в каких единицах оценивались сравниваемые ряды. Например, единицы наблюдения в одном ряду – в кг, в другом – в см и т.п. Однако схожесть полученных по этим выборкам конкретных величин, которые учитывают значения отдельных наблюдений (средние, дисперсии и т.п.), может быть установлена только путем обычных сравнений однородных данных.

При вычислении модифицированного критерия гомогенности хи-квадрат все расчеты выполняются только на относительных частотах распределения, рассчитываемых относительно итоговой суммы наблюдений. Понятно, что итоговая сумма частот не может быть больше или меньше 1.

Например, определение гомогенности (однородности) группы наблюдения по возрастно-половому составу (табл. 101). Величина всей группы $n=159$ чел. Пересчитываем абсолютные численности подгрупп в частоты относительно общего числа наблюдений: $29/159=0,182$; $36/159=0,226$; $14/159=0,088$ и т.д.

Таблица 101

Распределение выборочной группы наблюдения по возрасту и полу ($n=159$)

Абс. числа			
Возраст (лет)	Пол		Оба пола
	М	Ж	
<20	29	36	65
20-39	14	24	38
≥40	51	5	56
Все	91	65	159

Относительные частоты			
Возраст (лет)	Пол		Оба пола
	М	Ж	
<20	0,182	0,226	0,409
20-39	0,088	0,151	0,239
≥40	0,321	0,031	0,352
Все	0,519	0,409	1,00

Величина модифицированного критерия гомогенности χ^2 (хи-квадрат) определяется по формуле:

$$\chi^2 = n \left[\left(\sum_{i=1}^L \sum_{j=1}^L \frac{f_{ij}^2}{f_i f_j} \right) \cdot 0,01(n-1) \right]$$

где n — общее число наблюдений; f_{ij} — частоты по i строке и j столбцу.

По данным Таблицы :

$$\chi^2 = 159 * ((0,182^2/(0,409*0,519) + 0,226^2/(0,409*0,409) + 0,088^2/(0,239*0,519) + 0,151^2/(0,239*0,409) + 0,321^2/(0,352*0,519) + 0,031^2/(0,352*0,409)) * 0,01 * 158) = 37,1$$

При числе степеней свободы $k=S-1=7-1=6$ значимость результата равна $P < 0,01$. Т.о. нулевая гипотеза, свидетельствующая о том, что эти группы относятся к одной общей совокупности, т.е. однородно распределены, не подтвердилась.

8.2.7. Критерий Колмогорова

Несложен для практического применения и критерий Колмогорова, обозначаемый $K(\lambda)$:

$$K(\lambda) = D_{\max} \cdot \sqrt{N},$$

где N — число наблюдений в статистическом ряду, D_{\max} — максимальная разница в накопленных эмпирических и теоретических частотах. Ограничением при его применении является возможность сравнивать с его помощью лишь взвешенные ряды. Число наблюдений при этом должно быть достаточно большим ($N > 50$). 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 $P = 0,95$, если $K(\lambda) > 1,63$ при $P = 0,98$, 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, который играет наиболее существенную роль в расхождении распределений. В табл. 102 представлены частоты распределения обследованных лиц по показателям форсированной жизненной емкости легких (ФЖЕЛ). Теоретические частоты, распределение которых соответствует закону нормального распределения, получены, как и в предыдущем примере вычисления критерия согласия χ^2 (см. табл. 96). Значения вариант (показатели ФЖЕЛ) в таблице не приводятся, так как они не играют роли при непосредственном вычислении критериев согласия χ^2 (хи-квадрат) и $K(\lambda)$.

Таблица 102

Пример вычисления критерия Колмогорова $K(\lambda)$

Частоты		Накопленные частоты		Накопленные частоты		D
Эмпирические	Теоретические	Эмпирические	Теоретические	Эмпирические	Теоретические	
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	$K(\lambda)=0,055 * \sqrt{328}=0,99$				

Последовательность вычисления критерия $K(\lambda)$ 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.

3. 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.

4. В столбец 5 заносятся накопленные эмпирические частоты. Получаются они делением накопленной частоты на общее число наблюдений: $1/328*0,003$; $43/328*0,131$; $108/328*0,329$ и т.д.

5. Таким же образом рассчитываются накопленные теоретические частоты в столбце 6: $3,41/328*0,01$. Затем $25,04/328*0,076$ и т.д.

6. В столбец 7 заносим разность накопленных эмпирических и теоретических частот: $0,003-0,01=0,007$, $0,131-0,076=0,055$ и т.д.

Для дальнейших расчетов нужна только одна разность из столбца 7, максимальная. При выборе числа его знак (плюс или минус) не учитывается. В данном случае это 0,055.

По формуле $K(\lambda)=D_{max} \cdot \sqrt{N}$ получаем критерий $K(\lambda)=0,99$. Поскольку это значение меньше, чем 1,36, делаем вывод о том, что нулевая гипотеза не отвергается. Таким образом, предположение о соответствии распределения эмпирических частот и нормального распределения остается в силе.

Поскольку в ходе расчета критерия Колмогорова выделяется наиболее существенная разность распределений в отдельном диапазоне признака, по данным, приведенным в табл. 102, можно утверждать, что это диапазон признака (величина ФЖЕЛ не приведена в таблице), который встретился 42 раза.

9. Сравнение параметров распределений, нахождение различий

Статистический анализ различий между группами наблюдений, образованных по принципу «опыт-контроль», является одной из коренных задач медико-биологических исследований. Например, характеристика эффективности фармакологического препарата может складываться из статистических оценок различий каких-либо физиологических параметров опытных и контрольных групп лабораторных животных, групп больных с различными вариантами течения заболевания и его исходов. К аналогичным задачам относятся задачи оценки влияния разных факторов окружающей среды, социально-экономических условий жизни на здоровье тех или иных групп населения и т.п.

Для того чтобы здесь получить исчерпывающий ответ, необходимо решить следующие задачи, которые, на первый взгляд, не всегда связаны напрямую с конечной целью – оценкой существенности различий:

- 1) Относится та или иная варианта (отдельная единица) к данной статистической совокупности (к данной группе наблюдений)?
- 2) Отличается ли эмпирическое (полученное опытным путем) распределение совокупности наблюдений от того или иного теоретического распределения?
- 3) Соответствует ли распределение эмпирических данных одной группы распределению данных о другой группе?
- 4) Являются ли конкретные выборочные группы наблюдения выборками из одной и той же, другой совокупности?

Необходимость решения первого вопроса диктуется сильной вариабельностью данных медико-биологических исследований. Существенные, экстремальные разбросы значений отдельных вариант нередко вызывают сомнения, действительно ли та или иная варианта относится к конкретной группе или она включена в группу ошибочно? Значимость этой задачи определяется тем, что зачастую именно из-за этих так называемых «выскакивающих вариант» создается впечатление существенности различий сравниваемых результатов.

Ответ на второй вопрос позволяет обосновать выбор того или иного метода статистического анализа, в том числе и выбор метода оценки достоверности (статистической репрезентативности) различий.

Решение третьего вопроса дает ответ о существенности, значимости различий отдельных статистических параметров исследуемых групп. Следует отметить, что такие параметры распределений, как средние величины, ошибки средних, дисперсии и т.п., чаще всего не являются конечной целью статистического исследования, а являются лишь производными величинами, с помощью которых изучается вопрос о различиях совокупностей наблюдений в сравниваемых группах.

Четвертый вопрос так же как, и третий, ставит ту же проблему оценки значимости различий, но с других позиций: действительно ли различаются выборочные группы или различия могут быть объяснены обычной вариабельностью разных выборок, принадлежащих, по большому счету, единому целому или, как говорят в статистике, к одной генеральной совокупности?

Нетрудно заметить, что, по существу, все четыре вопроса сводятся к одному – является ли наблюдаемая разница статистических данных, по своей сути являющихся выборочными, случайной характеристикой, отражением реально существующих различий?

9.1. Принадлежность варианты к совокупности.

«Выскакивающие» варианты

В ряду наблюдений одна или несколько варианты могут резко отличаться по величине от других, существенно влияя на конечный результат статистического анализа. В этой ситуации можно предположить, что наличие таких «выскакивающих» вариант может быть результатом ошибочных измерений или записей.

В практике, если объем наблюдения достаточно велик, проблема принадлежности очень большой или очень маленькой варианты к конкретной совокупности решается довольно просто. Крайние варианты, так называемые «экстремальные варианты» или «выбросы», просто исключают. Это вполне

допустимо, ибо в большой совокупности роль нескольких крайних вариантов при вычислении средних величин, как правило, весьма незначительна. Выборка, оставшаяся после этой операции, называется цензурированной выборкой. Такая операция, проведенная путем отбрасывания некоторого числа экстремальных наблюдений по признаку их выхода за пределы заданного интервала возможных значений, называется цензурированием первого типа. Здесь общий объем выборки остается случайной величиной. Если отбрасывается заранее заданная, фиксированная часть крайних значений, то это называется цензурированием второго типа⁷⁰. Не вдаваясь в детальное рассмотрение последствий применения каждого из этих способов, следует отметить, что любое цензурирование должно применяться весьма осторожно. Необходимо помнить, что после его многократного применения обычно требуется использовать специальные методы статистической обработки выборки. Кроме того, потеря даже одной варианты может исказить результат всех последующих вычислений.

Наиболее простым приемом цензурирования является использование правила трех сигм. Согласно этому правилу, в пределах средняя арифметическая $\pm 3\sigma$ находится 99,7% всех вариантов. Поэтому если вариант попадает в этот интервал, то она может считаться принадлежащей к данной совокупности. Если не попадает, то она является «выскакивающей» и 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 значений $40+3 \times 2,5$ кг или от 47,5 до 32,5 кг должно находиться 99,7% всех вариантов. Вариант 58 кг явно не попадает в этот интервал. Таким образом, с высокой степенью вероятности можно утверждать, что данная вариант не относится к указанному ряду.

Рассмотренная выше методика определения «выскакивающих» вариантов неудобна для практического использования. Во-первых, она требует

⁷⁰ Иногда такие операции по имени автора, обосновавшего правомерность такого рода действий, называют простым оцениванием Диксона (Кобзарь А.И., 2006).

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 следующего алгоритма. Во-первых, вычисляются величины:

$$\tau' = \frac{V_n - V_{n-1}}{V_n - V_1} \quad \text{и} \quad \tau'' = \frac{V_1 - V_2}{V_{n-1} - V_1}, \text{ где}$$

τ' – показатель наибольшей варианты; а τ'' – наименьшей варианты. Критические значения этих показателей с учетом уровней значимости приводятся в специальных таблицах (приложение 1).

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 наименьшей в этом ряду 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 $P=0,01$ $m=0,92$. Поскольку вычисленное значение показателя меньше табличного (теоретического), варианту $V_1 = 45$ 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.

Достаточно широкое применение для оценки допустимости цензурирования первого типа получил критерий Романовского, который применяется следующим образом:

- определяются среднее арифметическое (\bar{x}) и среднеквадратическое, стандартное отклонение (σ) вариационного ряда без включения в него «выскакивающих» вариант;
- по специальной таблице, рассчитанной В.И. Романовским (Приложение 2), определяется значение вспомогательного коэффициента (f) для данного числа наблюдений (N);
- если «выскакивающая» варианта $V_{\text{выск}} - x > f\sigma$, то она исключается из выборки. В противном случае, эта варианта не исключается.

Исключать «выскакивающие» значения можно при $P=0,05$ (5%) и $P=0,01$ (1%). Если принять $P=0,05$, то риск ошибки признать варианту «выскакивающей», в то время как она не является таковой, составляет 5%. Значения $P=0,01$ дают возможность уменьшить вероятность аналогичной ошибки до 1 случая из 100.

*Например*⁷¹: При пяти определениях живых микробных клеток в 24-часовой культуре золотистого стафилококка получены следующие результаты (в млн. на 1 мл): 0,7; 0,4; 2,1; 0,9 и 0,6. Третья величина резко отличается от остальных. Можно ли это «выскакивающее» значение считать результатом случайной ошибки и исключить из статистической обработки?

Среднее арифметическое без «выскакивающей» варианты равно:

$$M=0,7+0,4+0,9+0,6/4=0,65.$$

Среднеквадратическое отклонение равно:

$$\sigma = \sqrt{\frac{(0,7-0,65)^2 + (0,4-0,65)^2 + (0,9-0,65)^2 + (0,6-0,65)^2}{4}} = 0,21$$

$V_{\text{выск}} - x = 2,1 - 0,65 = 1,45$. При $n=4$ и $f_{0,05} = 3,558$. Следовательно, при $P=0,05$ $\sigma f = 0,21 \times 3,558 = 0,7472$. Т.е. $V_{\text{выск}} - x > \sigma f$. Следовательно, есть все основания считать результат третьего измерения ошибочным и исключить его из статистической обработки.

9.1.1. Определение «выскакивающей» варианты с помощью MS Excel

В MS Excel имеется функция ZТЕСТ, с помощью которой можно определить принадлежность варианты к совокупности. *Например*, имеется некая статистическая совокупность, образованная рядом числовых значений: 14, 15, 17, 18, 22, 24, 19. Необходимо определить, соответствует ли варианта 24 этому ряду?

Порядок решения задачи в MS Excel:

- введите в ячейки A1:A7 исходный ряд чисел;
- установите курсор на ячейку (в позицию A9);

⁷¹ Пример взят из книги Меркова А.М., Полякова Л.Е. Санитарная статистика (пособие для врачей). – Л. – 1974. – 384 с.

- выполните команду <Функции> из меню <Формулы> или выберите на панели инструментов пиктограмму;

- выберите в появившемся диалоговом окне категорию «Статистические», а затем функцию «СТАНДОТКЛОНП» и нажмите [ОК];

- введите в окошко «Число» координаты числового ряда A1:A7 и нажмите [ОК];

Таким образом, в клетке A9 появится значение среднеквадратического отклонения (σ). Это даст возможность продолжить вычисления, для чего:

- установите курсор в позицию A10;

- выберите в диалоговом окне «Мастер функций» категорию «Статистические», а затем функцию «ZТЕСТ» и нажмите [ОК];

- в открывшемся окне установите параметры: Массив – (A1:A7); X – число, имеющее сомнительную принадлежность к совокупности, в данном случае 24, или координаты клетки, в которой это число расположено (A6);

- сигма – значение стандартного (среднеквадратического) отклонения, вычисленного вами ранее (A9).

После ввода исходных данных нажмите [ОК].

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 сомнительную варианту числа 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 полностью остается за исследователем.

С помощью встроенной функции *MS Excel* «УРЕЗСРЕДНЕЕ» можно оценивать возможные последствия исключения из анализируемого ряда «выскакивающей» 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 среднее арифметическое вычисленное из урезанного, цензурированного числового ряда. Если урезанное и исходное среднее не отличаются друг от друга или отличаются несущественно, то варианта (группа вариант), исключенная из урезанного ряда, не является «выскакивающей» и ее можно оставить в составе исходного ряда. Если средние будут различаться, то присутствие «выскакивающих» вариант будет, безусловно, сказываться на результатах дальнейшего анализа исходного числового ряда.

9.2. Сравнение средних арифметических и относительных показателей

Оценки различий между группами наблюдений с помощью таких статистических параметров, как средние величины или доли (%), являются одними из самых частых операций в практике обработки данных медико-биологических исследований. Уже более сотни лет многие из них, в большей части совершенно необоснованно, заканчиваются именно этой операцией.

Такая существующая и поныне «традиция» ограниченного применения возможностей статистики сформировалась во времена малой доступности «машинных» средств при обработке данных. С точки же зрения современных возможностей сплошной компьютеризации, ограничение доказательной базы оценкой одного-двух параметров выборки оправдано лишь для предварительного, разведочного рассмотрения итогов наблюдений или при отсутствии исходных выборок. Анализ только отдельных параметров выборок безусловно, обедняет доказательную базу любого исследования.

Порядок получения параметров выборок по схеме «опыт-контроль» на практике обычно строится следующим образом. *Например*, при оценке токсичности какого-либо вещества обычно берутся две группы биологических объектов – лабораторных животных. Подбираются животные одинакового

возраста, пола, одинакового содержания и т.п. I.e. делается все, чтобы эти объекты наблюдения представляли собой единую, как можно более однородную статистическую совокупность, что позволяет максимально снизить исходную вариабельность анализируемых данных. Оптимальной считается ситуация, когда сравниваемые группы могут отличаться только тем, что одна из групп (опытная) подвергается воздействию токсического вещества, а другая (контрольная) – нет. Однако на практике в любом случае, произошли ли после воздействия токсического вещества изменения в опытной группе или нет, в силу присутствия элемента случайности различия и внутри групп наблюдения, и между группами обязательно будут⁷².

Соответственно, при оценке результатов исследования возникает проблема: являются ли различия двух групп только следствием обычной случайности или разница возникла из-за того, что произошли определенные сдвиги физиологических функций животных опытной группы по сравнению с контролем? Иначе принадлежат ли животные опытной и контрольной групп к той же самой исходной совокупности или опытная группа принадлежит к другой совокупности, совокупности с измененными физиологическими параметрами?

Наиболее широко для решения этих задач в практике медико-биологических исследований используется параметрический критерий Стьюдента (t). Применение этого критерия недопустимо в случаях:

- несоответствия распределения вариант в сравниваемых группах закону нормального распределения;
- дискретного (балльного) характера сравниваемых числовых признаков;
- неподтвержденного равенства несмещенных дисперсий (разброса вариант в выборке) сравниваемых групп.

⁷² Не останавливаясь на перечислении всех обстоятельств, которые, несмотря на тщательный предварительный отбор объектов наблюдения, обуславливают случайные различия (случайность наблюдений), можно напомнить, что всему живому свойственна вариабельность, изменчивость.

Следует помнить, что задача сравнения средних двух нормально распределенных совокупностей при неравных дисперсиях известна, как проблема Беренса-Фишера, по имени авторов, впервые ее сформулировавших. Точного решения этой задачи до настоящего времени нет. На практике обычно используются различные критерии приближенного сравнения совокупностей, например: критерий Кохрана-Кокса, критерий Уэлча, критерий Сатервайта и т.п. (Кобзарь А.И., 2006).

При выборе метода вычисления значения критерия Стьюдента учитывается и объем выборок: малые выборки ($n < 30$) или большие выборки ($n > 30$).

Рассмотрим несколько примеров, в которых распределение исходных данных будем считать соответствующим нормальному распределению, а число наблюдений – менее 30.

Если значение среднего арифметического из выборки нужно сравнить с генеральным средним или с каким-нибудь теоретическим стандартом, то для вычисления значения критерия Стьюдента (t) используют формулу:

$$t = \frac{M - M_c}{m} \quad , \quad \text{где } m = \pm \sqrt{\frac{D}{n}} \quad , \quad D - \text{дисперсия выборки,}$$

n – число наблюдений в выборке, M – среднее арифметическое выборки, M_c – среднее стандартной выборки.

Например, требуется оценить различия между опытными (выборочными) данными и теоретическим стандартом. При определении значений жизненной емкости легких (ЖЕЛ) у 6 пациентов опытной группы были получены следующие результаты в процентах отклонений от должных величин ЖЕЛ: 91, 76, 80, 85, 84, 90. Можно ли считать с доверительной вероятностью $P=0,05$, что средний уровень этих отклонений ЖЕЛ выходит за границы условной нормы отклонений ЖЕЛ, равной $85-90$ ⁷³, если в опытной группе среднее отклонение составило $M=84,3$ при $m=\pm 2,6$.

⁷³ По Л.Л. Шяку, Н.Н. Канаеву, 1980.

В этом случае $t = \frac{64,3 - 55,0}{2,6} = -0,26$ или 0,26 (поскольку знак здесь не учитывается).

При числе степеней свободы $k=n-1=6-1=5$ и уровне значимости $P=0,05$ критическое значение критерия Стьюдента $t=2,57$, что значительно больше вычисленного значения t . Таким образом, нулевая гипотеза, свидетельствующая об отсутствии статистически значимых различий оцениваемых значений, не отвергается. Иначе говоря, выход результатов, полученных в опытной группе, за границы условной нормы ЖЕЛ статистически не подтвержден.

Особенностью, учитываемой при оценке различий, является попарная связанность данных, так называемые *парные* или *попарносвязанные выборки*, когда, например, рассматриваются различия физиологических показателей одной и той же группы животных до и после опыта.

Пример: результаты обследования одних и тех же лабораторных животных до опыта и после опыта (табл. 103). Дисперсии в этом случае могут быть неизвестны и во внимание не приниматься. Более того, в этом случае формально нет необходимости вычислять и средние. Несмотря на то, что при вычислениях t -критерия таким способом не используются эти параметры нормального распределения, этот метод считается параметрическим, поскольку распределение значений парной разности d должно подчиняться закону нормального распределения.

Парное сравнение зависимых результатов обследования

Номер животного	Результаты обследования		Разность d	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

Для вычисления абсолютного значения⁷⁴ критерия t в случае парных сравнений используется формула:

$$t = \frac{\sum d}{\sqrt{\frac{n \sum d^2 - (\sum d)^2}{n-1}}}$$

$$t = \frac{11}{\sqrt{\frac{(8 \cdot 25 - 11^2)}{(8-1)}}} = 3,27$$

Число степеней свободы $k=n-1=8-1=7$ дает возможность найти при доверительной вероятности $P=0,05$ критическое значение $t=2,4$. Поскольку вычисленное (фактическое) значение критерия Стьюдента превосходит критическое (см. Приложение 5), нулевая гипотеза отвергается. Таким образом, различия двух рядов наблюдений, «до» и «после», следует признать статистически значимыми.

Если вычислить значение критерия t , используя методiku, применяемую для оценки различий в несвязанных выборках, то получим значение $t=2,08$. Т.е. различия должны быть признаны несущественными.

Из этих результатов видно, что использование вместо одного другого метода определения достоверности различий может кардинально менять

⁷⁴ Т.е. без учета знака + или -.

результаты. Объясняется это тем, что с помощью *парного метода* рассматриваются просто разности между двумя измерениями («до опыта» и «после опыта») для каждого объекта наблюдения. Вычитая первые значения из вторых 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 парных, зависимых выборок в сравнении с *t*-критерием для несвязанных (независимых) выборок – такой подход дает всегда более объективный результат.

В ситуациях, когда задачей является оценка различий средних арифметических из двух независимых выборок (табл. 104), можно использовать следующий способ вычисления *t*-критерия.

Таблица 104

**Сравнительная длительность пребывания в стационарах
больных хроническим бронхитом**

Номер больного <i>j</i>	Число дней госпитализации			
	Первая больница		Вторая больница	
	y_1	y_1^2	y_2	y_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

1. Среднее первой выборки $M_1=124/8=15,5$ дня. Второй выборки – $M_2=135/8=16,9$ дня.
2. Несмещенные дисперсии, если число наблюдений небольшое и одинаково в обеих выборках, вычисляем по формуле:

$$D = \frac{\sum y_i^2 - M_j^2 n}{n - 1}$$

$$D = \frac{1936 - 15,5^2 \cdot 8}{8-1} = 2,0 \text{ дня}; \quad D = \frac{2293 - 16,9^2 \cdot 8}{8-1} = 2,1 \text{ дня}$$

3. С помощью критерия Фишера проверяем равенство дисперсий при уровне значимости $P=0,05$ и для степеней свободы, соответственно равных 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:

$$m = \sqrt{\frac{D}{n-1}}; \quad m_1 = \sqrt{\frac{2,0}{8-1}} = 0,53 \text{ дня}; \quad m_2 = \sqrt{\frac{2,1}{8-1}} = 0,55 \text{ дня}.$$

5. Находим абсолютное значение критерия Стьюдента:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{15,5 - 16,9}{\sqrt{0,55^2 + 0,53^2}} = 1,79$$

Строго говоря, значение критерия Стьюдента в приведенном примере равно 1,79. Однако при вычислении этого критерия принимается только его абсолютное значение.

6. Находим для него число степеней свободы: $k=(n_1-1)+(n_2-1)-2=$, 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 $k=14$ (см. Приложение 5).

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 несущественными.

По аналогичной методике производятся вычисления и оценки критерия Стьюдента в случаях, когда исследуются различия средних арифметических из совокупностей неодинаковых объемов.

Если сравниваемые средние арифметические получены при большом объеме нормально распределенных наблюдений, то определение достоверности различий этих средних можно провести упрощенным способом:

$$\frac{(\bar{M}_1 - \bar{M}_2)^2}{m_1^2 + m_2^2}$$

Если полученное значение будет больше или равно 9 (≥ 9), то различия средних арифметических можно считать статистически достоверными. Если меньше, то нет.

Следует помнить, что правомерность использования того или иного метода вычисления и оценки значения критерия Стьюдента может определяться обстоятельствами, оказавшимися вне поля зрения исследователя. Поэтому в сомнительных случаях, особенно при получении значений критерия, близких к критическим, следует вычислить коэффициент корреляции или иной показатель статистической связи между сравниваемыми рядами наблюдений. Если корреляция между рядами окажется существенной, то необходимо проверить результаты, полученные путем вычисления значения критерия Стьюдента одним из рассмотренных выше методов, другим методом оценки различий.

Особой задачей является задача нахождения различий частот. Одним из способов ее решения является угловое преобразование Фишера. Это преобразование служит основой для вычисления одноименного, по автору, критерия F , который применяется для сравнения двух частот из 2-х выборочных совокупностей⁷⁵.

Общий вид таких выборок, где частоты представлены в виде процентов, соответствует четырехпольной таблице 2x2 (табл. 105).

⁷⁵ Иногда для этого применяют критерий Стьюдента t .

Пример четырехпольной таблицы 2x2 (распределение по столбцам)

	Группа А (%)	Группа В(%)
Есть признак С	10	45
Нет признака С	90	55
Сумма	100	100

Суть углового преобразования Фишера состоит в переводе процентных величин в радианы центрального угла ϕ . Правомерность применения этого метода в статистике вызывает много споров среди специалистов, но его применение вполне допустимо при использовании в качестве ориентировочных оценок **различий частот**.

Достоинством данного критерия является то, что он может применяться к качественным данным или вообще разнородным данным, поскольку проценты чаще всего – это показатели распределения (частот) тех или иных вариантов. Понятно, что при этом размерность и качественные особенности самих вариантов не учитываются. Например, можно сравнивать процентные оценки изменения массы тела (кг) и роста (см). Другое достоинство метода – минимальный объём выборки может быть небольшим, что достаточно характерно для медико-биологических исследований. При этом верхней границы объема наблюдений не существует – выборки могут быть сколь угодно большими.

Недостатками метода является то, что ни одна из сопоставляемых долей не должна быть равной нулю при минимальном пределе – 2 наблюдения, т.к. итоговое значение критерия будет существенно завышенным. При этом, если в одной выборке – 2 наблюдения, то в другой необходимо иметь не менее 30. Если в одной выборке – 3 наблюдения, во второй должно быть не менее 7, если в одной выборке – 4 наблюдения, во второй должно быть не менее 5. Несоблюдение этих условий не позволяет получить оценку репрезентативности различий.

Если число наблюдений в одной выборке более 5, то формально для другой не существует ограничений.

Не вдаваясь в детали обоснования правомерности применения этого алгебраического преобразования, формулу для такой операции можно представить в виде: F , где p – частота в долях от единицы, т.е. 40%; $p = 0,4$. На основании этих данных производится вычисление значения критерия Фишера:

$$F = \frac{(\varphi_1 - \varphi_2)^2 N_1 N_2}{N_1 + N_2}$$

где N_1 и N_2 – числа наблюдений в первой и второй группах.

Пример: сравниваемые показатели равны $P_1 = 66,0\% = 0,66$ и $P_2 = 82,6\% = 0,826$. Объемы выборок $N_1 = 325$ и $N_2 = 23$. По таблице приложения находим, что P_1 соответствует $\varphi_1 = 1,90$ и P_2 соответствует $\varphi_2 = 2,28$. Тогда:

$$F = (1,9 - 2,28)^2 \cdot \frac{325 \cdot 23}{325 + 23} = 885$$

Проверка значимости полученного результата на уровне 95% ($P = 0,05$) осуществляется путем сравнения значения вычисленного F и критического значения $F_{кр}$ (см. Приложение 3), при этом первое число степеней свободы всегда принимается равным 1, а второе – $325 + 23 - 2 = 336$. Такой степени свободы соответствует $F_{кр} = 3,84$, что значительно меньше вычисленного значения F . Таким образом, сравниваемые показатели статистически различаются.

Для сравнения показателей в случае, когда выборки, из которых они вычислены, достаточно велики, часто используется формула:

$$t = \frac{P_1 - P_2}{\sqrt{m_{P_1}^2 + m_{P_2}^2}}$$

где P_1 и P_2 – сравниваемые показатели (%). Ошибки долей m могут быть вычислены по формуле:

$$m_p = \sqrt{\frac{Pq}{n-1}}$$

где: P – показатель в %, а $q = 100 - P$. Одной из весьма распространенных ошибок является использование для определения различий показателей этой формулы не для альтернативного варьирования, т.е. не для тех ситуаций, когда

возможен один из двух вариантов: «болел» или «не болел» (так называемое дихотомическое распределение). Однако, *например*, для оценки интенсивных показателей общей заболеваемости по обращаемости, заболеваемости с временной утратой трудоспособности, а также в ситуации, когда случаи болезни у одного человека могут встречаться по несколько раз в исследуемый период (год, квартал и т.п.), эта формула непригодна.

В.А. Мозглякова (1964) предложила использовать в этих случаях формулу:

$$m = \sqrt{\frac{M}{n}},$$

где M – среднее число заболеваний (за год, квартал и т.п.); n – число наблюдений. Эта формула используется, поскольку, по мнению автора, показатели заболеваемости распределяются в соответствии с законом распределения Пуассона или законом асимметричного биномиального распределения (при относительно невысоких показателях и небольшом числе наблюдений (<100)).

При больших числах наблюдений или высоких показателях заболеваемости (например, частота ЗВУТ >150 случая на 100 рабочих и т.п.) определять статистические ошибки следует только через построение вариационных рядов распределения. *Например*, при расчете показателей заболеваемости с временной утратой трудоспособности на 100 работающих строится вариационный ряд (табл. 106).

Таблица 106

Распределение рабочих по кратности заболеваний

Кратность заболеваний в году (V)	Число рабочих (P)
0	150
1	103
2	85
3 и более	17

И далее см. в разделах «Ряды распределений. Вариационные ряды».

9.2.1. Статистическая проверка различий средних арифметических в MS Excel

В MS Excel имеется несколько вариантов прямой оценки статистической достоверности различий средних арифметических.

Функция ТТЕСТ представляет результаты t -теста (Стьюдента) 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 (рис. 84). В эти окна могут вводиться числовые последовательности или ссылки на ячейки, содержащие числа.

В окне «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).

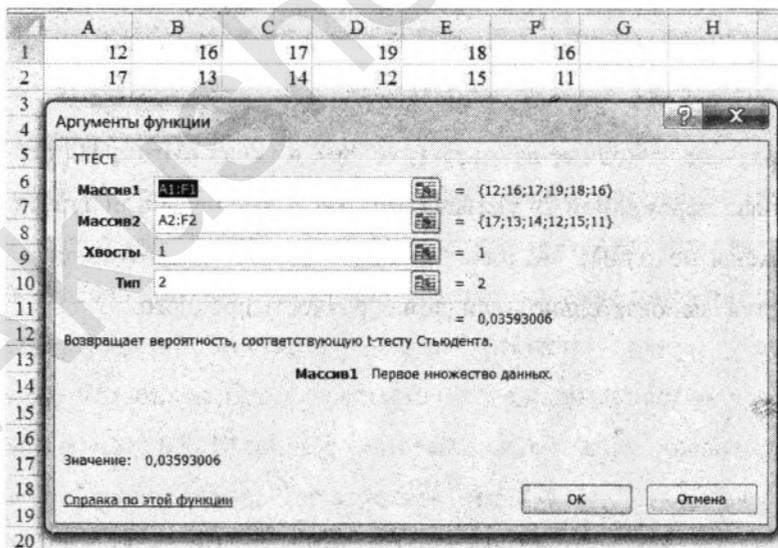


Рис.84. Окно функции ТТЕСТ

При анализе данных, указанных для примера на рис. 84, получается ответ: односторонняя вероятность сходства средних парной выборки $P=0,17$. Таким образом, нулевая гипотеза H_0 не отвергается, поскольку $P>0,05$. Следовательно, имеющиеся различия можно признать несущественными.

Более глубокий статистический анализ позволяет проводить пакет анализа. Здесь предусмотрены три инструмента анализа различий средних арифметических (рис. 85), однако результаты тестирования различий более детальны.

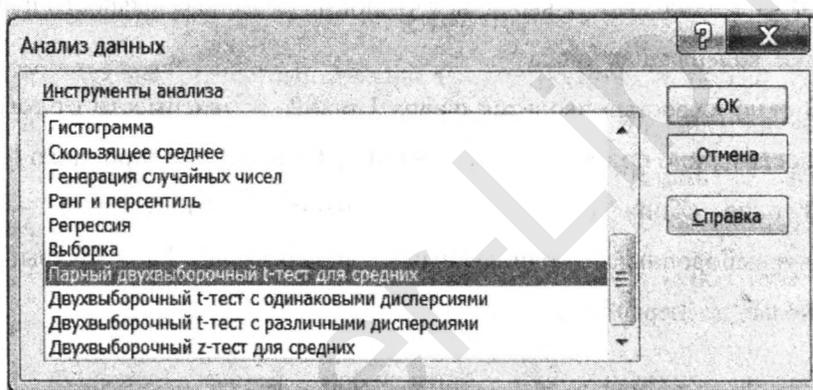


Рис.85. Выбор методов статистического анализа различий средних арифметических

В разделе «Входные данные» (рис. 86), в окнах «Интервал переменной 1» и «Интервал переменной 2» указываются числовые ряды или адреса клеток, где расположены исходные числовые данные. В окне «Гипотетическая разность» указывается (необязательно) ожидаемая разность среднего.

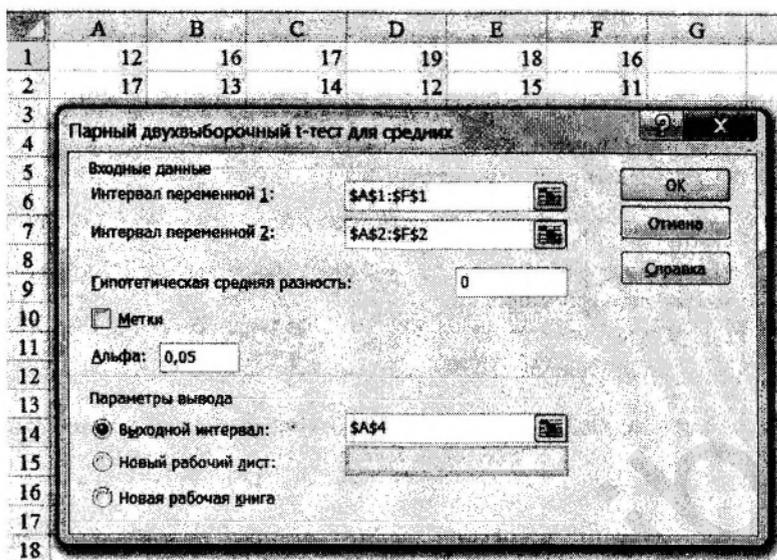


Рис.86. Параметры ввода данных для парного двухвыборочного t-теста средних арифметических по двум выборкам

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. В разделе «Параметры вывода» указывается место вывода информации.

Выходной интервал – место вывода информации на том же листе MS 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) MS Excel. Результаты анализа двух числовых рядов представлены следующими характеристиками (рис. 87).

	Переменная 1	Переменная 2
Среднее	16,333	13,667
Дисперсия	5,867	4,667
Наблюдения	6	6
Корреляция Пирсона	-0,5861	
Гипотетическая разность средних	0	
df	5	
t-статистика	1,6	
P(T<=t) одностороннее	0,08525	
t критическое одностороннее	2,01505	
P(T<=t) двухстороннее	0,17050	
t критическое двухстороннее	2,57058	

Рис.87. Результаты парного двухвыборочного *t*-теста средних арифметических по двум выборкам

Среднее – средние арифметические обоих рядов распределений признаков. Соответственно, 16,3 и 13,7.

Дисперсия – дисперсии двух выборок (5,9 и 4,7).

Наблюдения – числа наблюдений в выборках (6 и 6).

Корреляция Пирсона (0,59) – коэффициент корреляции Пирсона, с помощью которого оценивается 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 несвязанных выборок. Если будет установлена сильная связь анализируемых выборок, то полученные с помощью этого метода результаты будут представляться весьма сомнительными.

$df=5$ – число степеней свободы.

Далее выводятся фактические и критические значения *t*-критерия Стьюдента, а также значимости (*P*) различий выборок. Поскольку значимыми считаются различия при $P < 0,05$, в данном примере различия можно признать статистически неподтвержденными.

Аналогично осуществляется ввод и вывод информации при анализе существенности различий с помощью других *t*-тестов, имеющихся в пакете анализа *MS Excel*.

9.3. Сравнение нескольких средних.

Модифицированный критерий Стьюдента (Q)

Критерий достаточно прост при вычислении, но может применяться только в равных по числу наблюдений рядах при условии нормального распределения всей совокупности данных. Этот критерий не рекомендуется применять в рядах с числом наблюдений (в каждом) более 10.

Для понимания алгоритма его вычисления рассмотрим *пример* определения статистически значимых различий результатов измерения массы тела (кг) подростков в 5 разных группах, имевших среднюю массу тела 53,7 кг; 56,0 кг; 52,7 кг; 49,4 кг; 74,6 кг (табл. 107).

Порядок вычисления:

1) Вычисляются суммы в каждом ряду наблюдений: 376, 392, 369, 346, 522. Из их числа определяются максимальные и минимальные суммы (522 и 346).

2) Отсюда модифицированный критерий Стьюдента равен:

$$Q = \frac{522 - 346}{174} = 1,01$$

3) Критическое значение критерия находим по таблице (см. Приложение 4). Для $k=5$ и $n=7$ при $P=0,05$ $Q_{\text{критич}}=0,61$.

Поскольку вычисленное значение критерия больше критического, то нулевая гипотеза, говорящая о равенстве средних, отвергается, т.е. следует признать статистически значимой разницу средних величин в этих рядах данных.

Таблица 107

Промежуточные результаты вычисления модифицированного критерия Стьюдента (Q)

Группы (i)	$n_1=n_2=n_3=n_4=n_5=7$							Суммы	Среднее арифметич. (M_i)	max _i	min _i
1	41	44	48	51	60	62	70	376	53,7	70	41
2	45	49	51	52	56	68	71	392	56,0	71	45
3	38	41	44	49	61	62	74	369	52,7	74	38
4	34	35	39	48	54	61	75	346	49,4	75	34
5	54	56	62	78	84	92	96	522	74,6	96	54
$k=5$									$\sum \max_i - \sum \min_i = 174$		

10. Статистика динамики изменений явлений. Динамические ряды

Анализ динамики изменений различных характеристик объектов наблюдения, разнесенных во времени и пространстве, – один из самых сложных разделов математической статистики. В зависимости от предметности этого анализа используются существенно различающиеся методики. Не останавливаясь детально на их особенностях, следует отметить, что результаты оценки динамических рядов обладают более высокой информационной значимостью, чем отдельно взятые, изолированные оценки. Это обстоятельство объясняется тем, что доказательной базой таких исследований становится численная оценка тренда процессов развития (или регресса) явлений. Поэтому, опираясь на статистическое оценивание тренда изучаемого явления, можно пренебречь некоторой неточностью ошибок наблюдений (в т.ч. и групповых) или даже их отсутствием в отдельных дискретных точках динамического ряда.

С помощью статистического анализа динамики явлений решаются следующие задачи:

- выявление, фильтрация шума (помех) и описание характерных тенденций изменения (тренда) наблюдаемого явления;
- подбор статистической модели, описывающей эти изменения;
- отыскание на основе этой модели отсутствующих промежуточных значений показателей наблюдаемого явления (**интерполяция**);
- предсказание дальнейших изменений (**прогноз, экстраполяция**) показателей за пределами изученного ряда.

В медицинской статистике на первом этапе оценивания результатов большинства динамических наблюдений обычно используются относительно простые, но вместе с тем достаточно информативные приемы анализа.

Во-первых, применяются графические изображения, которые при удачной экспликации исходных данных могут дать самую исчерпывающую информацию. Во-вторых, вычисляется комплекс относительно несложных в

расчетах специальных относительных величин, называемых показателями динамического ряда.

Динамический ряд – ряд производных статистических величин, показывающий изменение какого-либо явления во времени и пространстве.

Различают следующие виды динамических рядов:

Простой динамический ряд – исходные данные представляют собой абсолютные величины.

Производный динамический ряд – исходный ряд составлен из средних или относительных величин.

Сложный – динамический ряд, отражающий параллельное изменение нескольких явлений.

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.

10.1. Показатели динамического ряда

К числу наиболее распространенных производных величин (показателей), характеризующих динамический ряд, относятся: абсолютный прирост или убыль, темп роста или снижения, темп прироста, абсолютное значение 1% прироста или убыли, показатели наглядности. В зависимости от точек счета могут использоваться два вида перечисленных величин: *базисные и цепные*. Если точкой отсчета всей цепочки исходных значений изучаемого явления служит первая точка, т.е. все показатели динамического ряда рассчитываются относительно этой точки, то такие показатели называются *базисными*. Например: базисные показатели наглядности смертности жителей РФ за период 1999, 2000, 2001, 2002 гг. последовательно вычисляются, как процентное отношение показателя смертности в 2000 году к показателю смертности

в 1999 году. Затем показателя смертности в 2001 году опять к показателю 1999 года и показателя смертности в 2002 году снова к показателю 1999 года.

Если расчет значений каждого последующего показателя ведется относительно значений предыдущей точки счета, то такие показатели называются *цепными*. Например: цепной показатель прироста уровня смертности жителей РФ за период 1999, 2000, 2001, 2002 гг. вычисляется последовательно, как разность уровня смертности в 2000 и 1999 гг., затем как разность смертности в 2001 и 2000 гг., затем – в 2002 и 2001 гг.

Объективный анализ динамики любого явления требует одновременного (параллельного) применения нескольких показателей динамического ряда. Использование одного, изолированного показателя может приводить к существенным ошибкам оценки общей динамики (тренда) явлений в ряду наблюдений, поскольку каждый из показателей обладает своими достоинствами и недостатками.

Абсолютный прирост или убыль характеризует динамику изменения явления за один интервал наблюдений. Этот цепной показатель получается путем вычитания из данных **последующей точки счета** данных **предыдущей точки**. По мере последовательного «продвижения» по ряду исходных данных эти точки последовательно сдвигаются. Если ряд возрастает, то прирост положителен. Если убывает – прирост отрицателен (убыль). Этот показатель не может использоваться при сравнении динамики разнородных данных (вес – в кг, рост – в см и т.п.). Кроме того, на его значение оказывает влияние и абсолютный размер анализируемой характеристики. *Например*, рост в сантиметрах – трехзначное число, окружность бедра, тоже в сантиметрах, – двухзначное. Поэтому абсолютный прирост, равный 5 сантиметрам в первом и втором случаях, имеет различное смысловое значение.

Темп роста или снижения – также цепной показатель, показывающий соотношение в процентах **последующего** уровня и **предыдущего**. Обезличенная форма показателя темпа роста (в %) позволяет использовать его при сравнительном анализе динамики разнородных величин. *Например*,

можно параллельно сравнивать динамику роста (в см) и веса (в кг) детей по мере увеличения продолжительности жизни. Вычисляется цепной показатель темпа роста путем деления последующего уровня на предыдущий и умножения на 100.

Темп прироста – цепной показатель, который показывает, на сколько процентов увеличился или уменьшился уровень явления. В отличие от предыдущего показателя (темпа роста) показывает изменение не всего уровня, а только той части, на которую увеличился или уменьшился предыдущий уровень. По существу отражает относительную скорость изменения явления от одной точки отсчета к другой. Вычислять цепной вариант темпа прироста можно двумя способами. Первый способ: абсолютный прирост за последующий момент времени делится на уровень предыдущего момента времени наблюдения и умножается на 100. Второй способ более простой: из темпа роста вычитается 100.

По сути дела, темп роста в 150% и темп прироста 50% отражают одну и ту же динамику явления. Близкое звучание терминов «темп роста» и «темп прироста» позволяет широко их использовать в недобросовестных целях. В частности, темп роста используется для того, чтобы отразить успехи, а темп прироста, для того чтобы уменьшить видимость растущих недостатков.

Абсолютное значение 1% роста характеризует значение (стоимость) 1% прироста изучаемого явления. Этот показатель (в цепном варианте) может вычисляться делением показателя предыдущего уровня исходных данных на 100 или абсолютного прироста на темп прироста. Является одним из самых существенных, поскольку «стоимость» одного процента роста может быть разной. Например, число районов города N с высоким уровнем загрязнения атмосферного воздуха в 1995 году было равно 4, в 1996 таких районов стало 8. Темп роста – 200%. В городе NN таких районов в 1995 году было 10, стало – 20. Темп роста в обоих случаях – 200%. Однако в первом случае число неблагополучных районов увеличилось на 4, а во втором – на 10. Более

того, даже в одном динамическом ряду значение цепного показателя одного процента роста может существенно различаться на разных отрезках времени.

Базовые показатели наглядности характеризуют динамику явления в процентах относительно исходного уровня. Исходный уровень принимается за 100%. Все последующие уровни через пропорцию отсчитываются от первого. В отличие от цепных показателей на всем протяжении временного ряда «стоимость» одного процента этого показателя остается неизменной.

Существуют и другие показатели динамики, которые применяются в условиях определенных целей анализа и разных вариантах исходных данных. Большинство из них отличаются большой специфичностью и трудоемкостью вычислений. В наиболее общем виде расчет перечисленных выше показателей динамического ряда выглядит следующим образом (табл. 108).

Таблица 108

**Динамика случаев заболеваний
с временной утратой трудоспособности (ЗВУТ)**

Год	Уровень ЗВУТ	Абсолютный прирост	Темп роста (%)	Темп прироста (%)	1% прироста	Показатели наглядности (%)
	У	А	Тр	Тп	П	Н
2001	65,8					100,0
2002	90,2	24,4	137,1	37,1	0,7	137,1
2003	67,4	-22,8	74,7	-25,3	0,9	102,4
2004	94,3	26,9	139,9	39,9	0,7	143,3
2005	55,4	-38,9	58,7	-41,3	0,9	84,2
2006	45,1	-10,3	81,4	-18,6	0,6	68,5
2007	48,2	3,1	106,9	6,9	0,5	73,3

Абсолютный прирост или снижение заболеваемости в 2002 году по сравнению с 2001 годом и в 2003 году по сравнению с 2002 годом:

$$A_{2002} = Y_{2002} - Y_{2001} = 90,2 - 65,8 = 24,4 \text{ случая на 100 работающих за год;}$$

$$A_{2003} = Y_{2003} - Y_{2002} = 67,4 - 90,2 = -22,8 \text{ случая на 100 работающих за год.}$$

Темп роста или снижения заболеваемости в 2002 году по сравнению с 2001 годом и в 2003 году по сравнению с 2002 годом:

$$T_{p2002} = (Y_{2002}/Y_{2001}) \times 100 = 90,2/65,8 \times 100 = 137,1\%;$$

$$T_{p2003} = (Y_{2003}/Y_{2002}) \times 100 = 67,4/90,2 \times 100 = 74,7\%.$$

Темп прироста заболеваемости в 2002 году по сравнению с 2001 годом и в 2003 году по сравнению с 2002 годом:

1 способ расчета:

$$T_{п2002} = (A_{2002}/Y_{2001}) \times 100 = 24,4/65,8 \times 100 = 37,1\%;$$

$$T_{п2003} = (A_{2003}/Y_{2002}) \times 100 = -22,8 / 90,2 \times 100 = -25,3\%;$$

2 способ расчета:

$$T_{п2002} = T_{2002} - 100\% = 137,1\% - 100\% = 37,1\%;$$

$$T_{п2003} = T_{2003} - 100\% = 74,7\% - 100\% = -25,3\%.$$

Абсолютное значение 1% прироста заболеваемости в 2002 году по сравнению с 2001 годом и в 2003 году по сравнению с 2002 годом:

1 способ расчета:

$$П_{2002} = Y_{2001}/100 = 65,8/100 \approx 0,66 \approx 0,7;$$

$$П_{2003} = Y_{2002}/100 = 90,2/100 \approx 0,90 \approx 0,9;$$

2 способ расчета:

$$П_{2002} = A_{2002}/T_{п2002} = 24,4/37,1 \approx 0,7;$$

$$П_{2003} = A_{2003}/T_{п2003} = -22,8/-25,3 \approx 0,9.$$

Показатели наглядности прироста ЗВУТ в 2002 и 2003 годах по сравнению с 2001 годом:

$$H_{2002} = (Y_{2002}/Y_{2001}) \times 100 = 90,2/65,8 \times 100 = 137,1\%;$$

$$H_{2003} = (Y_{2003}/Y_{2001}) \times 100 = 67,4/65,8 \times 100 = 102,4\%.$$

В ряде случаев целесообразно вычислять обобщенные показатели динамики явления за все анализируемые периоды наблюдений. Этот вариант является наиболее приемлемым, когда интервалы между отдельными промежутками наблюдений не равны между собой.

В качестве обобщающего показателя вычисляются:

1. *Средний уровень динамического ряда за 2001-2007 гг.:*

– в моментном ряду с равными промежутками между датами

$$Y_{\text{средн.}} = (1/2 Y_{2001} + Y_{2002} + Y_{2003} + \dots + 1/2 Y_{2007})/n, \text{ где } n - \text{ число точек счета};$$

– в моментном ряду с неравными промежутками между датами

$$Y_{\text{средн.}} = (1/2 Y_{2001} \times t_{2001} + Y_{2002} \times t_{2002} + \dots + 1/2 Y_{2007} \times t_{2007})$$

$$(t_{2001} + t_{2002} + \dots + t_{2007}), \text{ где } t_n - \text{ число дней в году};$$

- в интервальном ряду = $(Y_{2001} + Y_{2002} + Y_{2003} + \dots + Y_{2007})/n$

$$Y_{\text{средн.}} = (65,8 + 90,2 + 67,4 + 94,3 + 55,4 + 45,1 + 48,2)/7 = 66,6.$$

2. *Средний абсолютный прирост:*

$$A_{\text{средн.}} = (A_{2002} + A_{2003} + \dots + A_{2007})/n$$

$$A_{\text{средн.}} = (24,4 - 22,8 + 26,9 - 38,9 - 10,3 + 3,1)/7 = -2,5.$$

3. *Средний темп роста (среднее хронологическое) вычисляется в виде среднего геометрического:*

$$\sqrt[n]{Tr_{2002} \times Tr_{2003} \times Tr_{2004} \times Tr_{2005} \times Tr_{2006} \times Tr_{2007}}$$

$$Tr_{\text{средн.}} = \sqrt[6]{137,1 \times 74,7 \times 139,9 \times 58,7 \times 81,4 \times 106,9} = 94,9\%.$$

4. *Средний темп прироста:*

$$Tp_{\text{средн.}} = Tr_{\text{средн.}} - 100\%$$

$$Tp_{\text{средн.}} = 94,9 - 100\% = -5,1\%.$$

Изменение перечисленных показателей динамического ряда, несмотря на их кажущуюся однонаправленность, может принимать самые разнообразные формы. Например, абсолютные приросты могут быть стабильными, а темпы роста (прироста) при этом увеличиваться или уменьшаться.

10.2. Обработка динамических рядов в MS Excel

В программе *MS Excel* имеется целый ряд возможностей статистической обработки динамических рядов. В качестве примера создадим в *Excel* таблицу «Динамика распространенности ВИЧ в РФ» (рис. 88).

	А	В	С	Д	Е	Ф	Г	Н
1	Заболываемость ВИЧ в РФ (Российский статистический ежегодник, 2010 г.)							
	Год	Зарегистрировано больных ВИЧ инфекцией (на 100 тыс.чел. населения РФ)	Абс.прирост или убыль (на 100 тыс.чел)	Темп роста (%)	Темп прироста (%)	Значение 1%	Показатель наглядности (%)	
2								
3	2000	53,4					100,0	
4	2001	106,7						
5	2002	130,4						
6	2003	147,6						
7	2004	159,6						
8	2005	164,5						
9	2006	166,8						
10	2007	188,4						
11	2008	212,2						
12	2009	234,6						
13								
14								
15								

Рис.88. Динамика распространенности ВИЧ в РФ (исходные данные)

После запуска *Excel* введите в ячейку A2 слово «год» и нажмите клавишу [Enter]. В ячейки A3-A9 введите года: 2000...2009. Запишите в ячейки B1-G2 названия столбцов, а в ячейки B3-B9 – числовые данные.

Заполните пустые столбцы таблицы. В графу «Абс. прирост или убыль» занесите разность между последующим и предыдущим уровнями. Для этого введите в ячейку C4 формулу: =B4-B3 (для записи адресов используется только латинский шрифт).

В графу «Темп роста» заносится отношение (в %) каждого последующего уровня к предыдущему. Для этого введите в ячейку D4 формулу: =B4/B3*100.

В ячейку графы «Темп прироста» E5 занесите =D4-100.

Заполните графу «Значение 1%». Для этого введите в ячейку F4 формулу: =B3/100.

В графу «Показатель наглядности» заносится отношение (в %) каждого уровня к исходному уровню за 2000 г. Для этого введите в ячейку G3 формулу:

= B4/\$B\$3*100. Знак \$ включается в формулу, чтобы адрес ячейки B3 не изменялся, как обычно при копировании.

Скопируйте формулы из ячеек C3:G3 в ячейки C4:G9, для этого:

– выделите мышкой блок ячеек C3:G3;

– установите указатель мыши в левый нижний угол выделенного блока.

Этот угол обычно отмечен небольшим квадратиком. Когда вы к нему подведете курсор мыши, стрелка курсора примет вид знака +. После этого, нажмите правую клавишу мыши и, не отпуская её, переместитесь в ячейку G9. В результате этой операции область ячеек C3:G9 окажется заполненной скопированными формулами, т.е. подготовленный ранее макет таблицы окажется наполненным результатами вычислений (рис. 89).

	A	B	C	D	E	F	G	H
1	Зболеваемость ВИЧ в РФ (Российский статистический ежегодник, 2010 г.)							
	Год	Зарегистрировано больных ВИЧ инфекцией (на 100 тыс.чел. населения РФ)	Абс.прирост или убыль (на 100 тыс.чел)	Темп роста (%)	Темп прироста (%)	Значение 1%	Показатель являемости (%)	
2								
3	2000	53,4					100,0	
4	2001	106,7	53,3	199,8	99,8	0,53	199,8	
5	2002	130,4	23,7	122,2	22,2	1,07	244,2	
6	2003	147,6	17,2	113,2	13,2	1,30	276,4	
7	2004	159,6	12,0	108,1	8,1	1,48	298,9	
8	2005	164,5	4,9	103,1	3,1	1,60	308,1	
9	2006	166,8	2,3	101,4	1,4	1,65	312,4	
10	2007	188,4	21,6	112,9	12,9	1,67	352,8	
11	2008	212,2	23,8	112,6	12,6	1,88	397,4	
12	2009	234,6	22,4	110,6	10,6	2,12	439,3	
13								
14								

Рис.89. Динамика распространенности ВИЧ в РФ (результаты расчетов)

10.3. Углубленный анализ динамических рядов

Как уже отмечалось, основной задачей статистического анализа динамических рядов является:

- выявление систематической, регулярной компоненты динамики ряда;
- выделение (фильтрация) шума, что позволяет увидеть регулярную компоненту более отчетливо.

Когда речь идет о систематической, регулярной компоненте, в медицинской статистике обычно подразумевают либо сезонность, либо тренд.

Сезонность – это периодически повторяющаяся компонента. *Например*, постоянный «скачок» инфекционной заболеваемости в определенные месяцы года. Особенностью показателей сезонности являются их цикличность и неравные промежутки времени между точками счета.

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).

10.3.1. 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 сравнение значений показателей в определенных интервалах времени (декабрь одного года с декабрем другого, январь – с январем и т.д.). Кроме того, показатели сезонности позволяют исключить влияние на уровни показателей разной длительности временных интервалов: январь – 31, февраль – 28 дней и т.п. (табл. 109).

Таблица 109

Распределение случаев 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

Таблица 110

Порядок вычисления показателей сезонности

1. Определяется среднеедневное 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 среднеедневное число заболеваний за год:	
	$326/365=0,89$
3. Вычисляются показатели сезонности в абсолютных числах, как отношение среднеедневных месячных уровней к среднеедневному годовому:	
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	$0,939 \times 100=93,9\%$
в феврале	$0,960 \times 100=96,0\%$
в марте	$0,903 \times 100=90,3\%$ и т.д.

10.3.2. Вычисление показателей сезонности в MS Excel

В качестве примера создадим в Excel таблицу «Распределение острых кишечных инфекций (ОКИ) за год» (рис. 90).

	A	B	C	D	E	F
1	Месяц года	Дней в месяце	Всего ОКИ	В среднем за 1 день	Сезонность (абс.)	Сезонность (%)
2	январь	31	26	0,84	0,94	93,9
3	февраль	28	24	0,86	0,96	96,0
4	март	31	25	0,81	0,90	90,3
5	апрель	30	21	0,70	0,78	78,4
6	май	31	25	0,81	0,90	90,3
7	июнь	30	31	1,03	1,16	115,7
8	июль	31	29	0,94	1,05	104,7
9	август	31	34	1,10	1,23	122,8
10	сентябрь	30	33	1,10	1,23	123,2
11	октябрь	31	28	0,90	1,01	101,1
12	ноябрь	30	25	0,83	0,93	93,3
13	декабрь	31	25	0,81	0,90	90,3
14	Итого	365	326	0,89	1,00	100,0
15						
16						

Рис. 90. Пример формирования в Excel исходных данных сезонности

После запуска Excel сформируйте таблицу, содержащую названия колонок и исходные данные (рис.91).

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.

Для определения общего числа заболеваний за год введите в ячейку C15 формулу: =СУММ(C3:C14). В результате получите число 326.

Для определения среднедневного месячного числа заболеваний введите в ячейку D3 формулу: =C3/B3.

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].

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, выполните копирование формулы по аналогии с пп.4.

Для заполнения графы «Показатель сезонности в процентах»:

- в ячейку F3 введите формулу $=E3*100$;
- скопируйте формулу в ячейки F4:F15.

Для задания в столбцах таблицы нового числового формата:

- выделите соответствующий блок ячеек и 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].

	A	B	C	D	E	F
	Месяц года	Дней в месяце	Всего ОКИ	В среднем за 1 день	Сезонность (абс.)	Сезонность (абс.)
1						
2	январь	31	26	0,84	0,94	93,9
3	февраль	28	24	0,86	0,96	96,0
4	март	31	25	0,81	0,90	90,3
5	апрель	30	21	0,70	0,78	78,4
6	май	31	25	0,81	0,90	90,3
7	июнь	30	31	1,03	1,16	115,7
8	июль	31	29	0,94	1,05	104,7
9	август	31	34	1,10	1,23	122,8
10	сентябрь	30	33	1,10	1,23	123,2
11	октябрь	31	28	0,90	1,01	101,1
12	ноябрь	30	25	0,83	0,93	93,3
13	декабрь	31	25	0,81	0,90	90,3
14	Итого	365	326	0,89	1,00	100,0
15						

Рис. 91. Итоговые показатели сезонности

10.4. Повышение наглядности 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 временных 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 1980, 1981 и 1982 годы ($84+94+92=270$), затем за 1985, 1986 и 1987 годы и т.д. (табл. 111);

– *вычисление групповых средних*, которые определяются на основе данных по укрупненным интервалам ($270/3=90$, $263/3=88$ и т.д.).

Таблица 111

Сглаживание динамического ряда укрупнением интервалов
и скользящим средним

Учетный годы	Число рождений	Суммы по 3 годам	Средние по 3 годам	Скользящие средние
1995	84			–
1996	94		90,0	90,0
1997	92	270		89,7
1998	83			88,7
1999	91		87,3	87,3
2000	88	262		87,0
2001	82			86,7
2002	90		83,0	83,0
2003	77	249		82,3
2004	80			82,3
2005	90		82,7	82,7
2006	78	248		–

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 временного ряда.

Он основан на переходе от начальных значений ряда к значениям, усредненным в определенном интервале времени, например, за 3 года (табл. 112). В этом случае интервал времени при вычислении каждого последующего показателя как бы скользит по временному ряду.

Применение скользящего среднего особенно полезно для детального выявления закономерностей распределения (тенденций) динамического ряда в ситуациях, когда на показатели многократно воздействуют резко выделяющиеся данные, так называемые интервенции.

В приведенном примере временной интервал для вычисления скользящего среднего принят равным 3 годам. В результате проведенного сглаживания основная тенденция динамического ряда стала более наглядной. В частности, при оценке динамики рождаемости можно утверждать, что наиболее интенсивно рождаемость снижалась в период с 1984 по 1986 гг. и с 1988 по 1990 гг. Периоды с 1986 по 1988 гг. и с 1990 по 1993 гг. отличались относительной стабильностью, несмотря на размахи колебаний годовых показателей.

Вычисление скользящего среднего

Годы	Число рождений	Арифметические операции
1995	84	—
1996	94	$(95+96+97)/3=90,0$
1997	92	$(96+97+98)/3=89,7$
1998	83	$(97+98+99)/3=88,7$
1999	91	$(98+99+00)/3=87,3$
2000	88	$(99+00+01)/3=87,0$
2001	82	$(00+01+02)/3=86,7$
2002	90	$(01+02+03)/3=83,0$
2003	77	$(02+03+04)/3=82,3$
2004	80	$(03+04+05)/3=82,3$
2005	90	$(04+05+06)/3=82,7$
2006	78	—

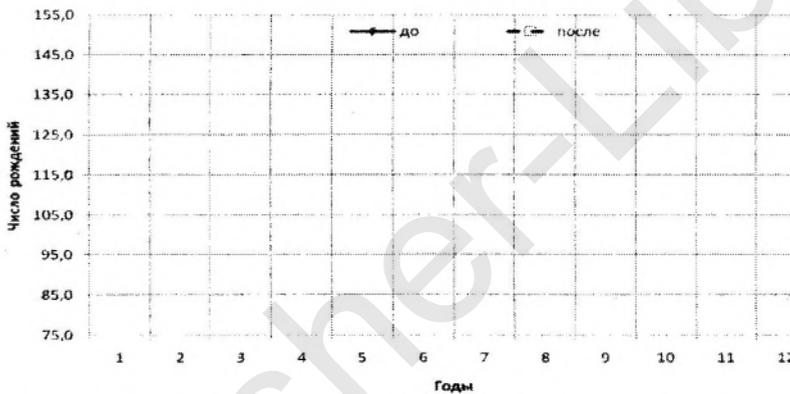


Рис. 92. Сравнительное распределение итоговых показателей исходного и выровненного (скользящей средней) рядов данных из табл. 112

В целом чем больше интервал сглаживания, тем более плавный вид имеет диаграмма скользящих средних. При выборе величины интервала сглаживания необходимо исходить из размеров динамического ряда (числа отдельных временных точек) и содержательного смысла отражаемой динамики. Большая величина динамического ряда с большим числом точек наблюдения позволяет использовать более крупные временные интервалы сглаживания (5, 7, 10 и т.д.). Если процедура скользящего среднего используется для сглаживания сезонного ряда, то чаще всего величину интервала сглаживания принимают равной 3.

Весьма результативным методом, хотя в своей основе более сложным, является сглаживание (выравнивание) рядов динамики с помощью различных математических функций аппроксимации.

При помощи этих функций формируются 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 экспоненциальное сглаживание. Не вдаваясь в детальное математическое описание этого метода, следует отметить, что в отличие от скользящего среднего или группового среднего методикой экспоненциального сглаживания учитываются сразу все наблюдения ряда, а не те несколько, 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 α равно 1, то предыдущие наблюдения полностью игнорируются. Если α равно 0, то игнорируются текущие наблюдения. Значения α между 0, 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 «10.4.1. 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.

Показателем правильности выбора того или иного уравнения аппроксимации служит коэффициент R^2 . Чем больше его значение приближается к единице, тем большее соответствие фактического и выровненного распределений. Максимальное значение, которое R^2 может принимать в предельном случае, равно 1.

Следует отметить, что при решении проблемы выбора для выравнивания (аппроксимации) исходных данных прямой или какой-либо кривой нельзя исходить из формальных соображений типа: 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 существенных усилий со стороны исследователя. Имея эти результаты, можно проводить математическую **экстраполяцию**, то есть давать прогноз показателей в продолжение ряда исходных данных, или проводить **интерполяцию**, то есть определять утраченные или отсутствующие показатели в любой промежуточной точке интервала анализируемого временного ряда.

Говорить о достоверности прогнозов динамики каких-либо явлений можно лишь при сохранении общих тенденций, то есть при наличии статистической

устойчивости, инерционности изучаемых явлений. Здесь имеется в виду инерционность, которая обеспечивает сохранение в общих чертах механизма формирования явления и инерционность характера динамики процесса (темп, направление) 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 $R^2 < 0,6$.

10.4.1. Обработка динамических рядов и прогноз динамики в MS Excel

В программе *MS Excel* имеется целый ряд возможностей углубленной статистической обработки динамических рядов.

Как уже отмечалось, математическое выражение закономерностей динамики данных можно получить с помощью **функции экспоненциального сглаживания**. Например, необходимо провести сглаживание числового ряда, отражающего динамику рождаемости за 1995-2006 гг.

Предварительно в *MS Excel* необходимо разместить исходные данные (табл. 112). Затем строим линейную диаграмму с помощью функции мастера диаграмм «Добавить диаграмму», выбираем линейную диаграмму. В *Excel* она названа «График». На графике по оси X вместо фактических значений каждого года выставляем точки отсчета (1, 2, 3 и т.д.). Связано это с тем, что в формуле выравнивания (аппроксимации) тренда будут использоваться не конкретные значения данных по оси X, а номера точек отсчета. (Чтобы в эти формулы попадали конкретные значения данных по оси X, в *Excel* необходимо использовать диаграмму «Точечная»). Для повышения изобразительности диаграммы по вертикальной оси в качестве точки отсчета выбираем не 0, а 75 (рис. 93).



Рис. 93. Распределение числа родившихся в 1995-2006 гг.

Щелкнув правой кнопкой мыши по линии графика, вызываем всплывающее меню, из которого выбираем функцию «Добавить линию тренда».

Затем последовательно в открывшейся закладке выбираем «Тип» линии тренда «Линейная». Затем, переключив ярлык закладки на **Параметры**, устанавливаем параметры «Прогноз вперед на 1 период» и флажки: «Показывать уравнение на диаграмме и поместить на диаграмму величину достоверности аппроксимации R^2 » (рис. 94).

После этого нажмите на клавишу [OK]. На экране появится изображение графика исходного ряда данных, аппроксимирующей линии (линии сглаживания) и уравнения аппроксимирующей функции. По своему виду эта функция представляет собой уравнение регрессии. Параметр R^2 показывает, насколько точно соответствует вычисленное уравнение регрессии истинной тенденции динамического ряда (См. раздел «Коэффициент линейной корреляции», «коэффициент детерминации»). Максимально возможное значение $R^2=1,0$ или 100%. В данном примере в случае использования аппроксимации (сглаживания) с помощью линейной функции $R^2=0,239$ или около 24%. Если из значения $R^2=0,239$ извлечь квадратный корень, то получим

$|R|=0,49$, т.е. абсолютное значение коэффициента корреляции. Поскольку отраженная на диаграмме линия регрессии очевидно демонстрирует обратную связь, то коэффициент корреляции $R = -0,49$.

С помощью коэффициента детерминации в *Excel* можно подбирать функцию наиболее полно аппроксимирующую ту или иную тенденцию (тренд). Соответственно, это дает возможность проводить прогнозирование значений Y . Если период прогноза равен единице, на экране изображение диаграммы приводится с отображением точки прогноза во времени ($12+1=13$ по оси X).

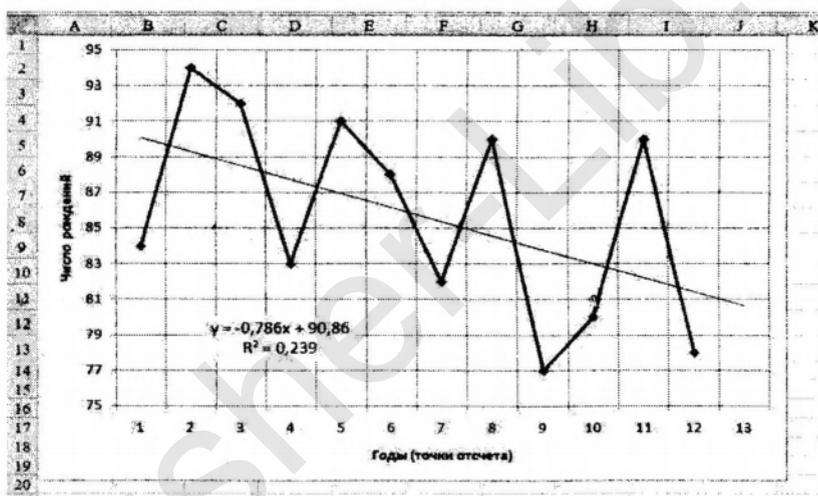


Рис.94. Диаграмма выравнивания динамического ряда с помощью линейной функции $Y = -0,7867x + 90,86$

11. Статистические оценки связей

Большая часть статистической доказательной базы любого медико-биологического исследования строится на выявлении и численной оценке связи (взаимодействий, влияний и т.п.) различных объектов и явлений. Например, эффективность какого-либо препарата, метода лечения или профилактики оценивается в конце концов через оценки связи показателей здоровья пациентов и факта применения того или иного препарата, метода лечения, профилактики и т.п. Формальная постановка задачи здесь, как правило, выглядит следующим образом: «Определить наличие и силу связи какого-либо фактора с другим фактором» или «Определить наличие и силу зависимости какого-либо результивного признака от одного или нескольких факторных признаков» и т.п.

Следует обратить внимание на то, что термин «зависимость» при статистической обработке данных должен использоваться весьма осторожно. Это обусловлено природой статистического анализа, который, во-первых, оперирует только с явлениями, проявляющимися в массе наблюдений, иначе говоря, с вероятностями явлений. То есть в отдельных случаях между объектами наблюдения связи может не быть вовсе или она может иметь свои, уникальные черты, отличаясь от связи, проявляющейся в группе наблюдений (групповой тенденции). Во-вторых, выявленная статистическая связь сама по себе не вскрывает истинных причинно-следственных отношений между объектами, взаимодействие между которыми может быть опосредовано третьими факторами. Причем эти третьи факторы часто лежат вообще вне поля зрения исследователя. **Попытки механически перенести данные статистических расчетов в объективную реальность могут привести к ошибочным выводам.** Например, утверждение: «Чем громче утром кричат воробьи, тем выше встанет солнце», несмотря на свою явную несуразность, с точки зрения формальной статистики, вполне правомерно.

Визуализация любой статистической связи может быть обеспечена следующими приемами:

- графическим методом с помощью диаграмм;
- с помощью статистических таблиц;
- с помощью специальных численных показателей (коэффициентов корреляции, параметров регрессии и т.п.).

Использование графического метода служит обычно свидетельством добросовестности статистического анализа. В частности, графический метод позволяет не только сделать наглядными основные закономерности связей, но и обеспечить избрание исследователем адекватного метода оценивания взаимосвязей рассматриваемых явлений.

Статистические таблицы решают аналогичные задачи повышения наглядности статистических зависимостей и выбора метода их численной оценки. Более того, в ряде случаев, например, при изучении явлений, представленных частотами качественных или количественных признаков, таблицы – неременный атрибут технологии оценивания взаимосвязей.

Понятно, что цели и задачи любого исследования определяются его уникальностью. Вместе с тем, существует общая, мало зависящая от этой уникальности методология статистической оценки связей. Она включает следующие «узловые точки»:

- установление характера связи (функциональная или статистическая), ее направления (прямая или обратная) и формы (линейная или нелинейная);
- оценка тесноты (силы, плотности) связи, в том числе определение величины детерминации взаимосвязанных явлений (факторов);
- оценка репрезентативности оценок взаимосвязей, полученных по выборочным данным (величина ошибки, доверительный интервал, уровень значимости);
- построение математической модели, описывающей взаимосвязь и обеспечивающей статистическое прогнозирование конечных результатов при возможных изменениях исходных данных.

Одной из основных, базовых характеристик связи явлений⁷⁶ является ее направленность, которая 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 увеличении температуры газа, возрастает его давление (при неизменном объеме газа). При уменьшении температуры снижается и давление газа.

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 основным элементом всех управляющих систем, в том числе и систем регулирования процессов жизнедеятельности любого организма.

Всякая из перечисленных зависимостей по **характеру** связи может быть функциональной или статистической.

Функциональная связь – такой вид связи, когда каждому значению одного признака соответствует точное, единственное значение другого. *Например*, взаимосвязь площади круга (S) и радиуса его окружности (R). Известно, что площадь и радиус круга (окружности) связаны вполне определенным отношением: $S = \pi R^2$, где число $\pi = 3,141592\dots$ – константа, а R^2 – квадрат радиуса круга. Т.е. изменение одного признака точно связано с изменением другого.

В медико-биологических исследованиях сталкиваться с функциональной связью приходится крайне редко, поскольку объекты этих исследований имеют большую вариабельность (изменчивость). Кроме того, существующие взаимосвязи различных явлений могут быть сведены к отношению двух-трех факторов лишь условно. Как правило, они всегда зависят от большого числа различных малоустойчивых по своей природе факторов.

⁷⁶ В статистике, для обозначения характеристик изучаемых явлений используется множество разнообразных терминов, являющихся, по сути, синонимами: факторы, результативные и факторные признаки, зависимые и независимые переменные, предикторы, регрессоры и т.д.

Для медико-биологических объектов характерна *статистическая связь*. То есть такая связь, когда изменение величины одного признака связано с тенденцией изменения значений или с вероятностью появления другого признака. Иначе говоря, 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 изменение величины сопровождается, как правило, определенным изменением величины Y . И это правило существует только в общих чертах как *тенденция*, наблюдаемая в массе наблюдений. Например: при изменении роста человека меняется и масса его тела. Однако эта зависимость не является полной, т.е. функциональной. У людей с одинаковым ростом может быть разная масса тела, поскольку ее величина обычно обусловлена и многими другими факторами, в том числе и неявляющимися, по своей сути, характеристиками физического развития (*например*, образ жизни и т.п.).

Следует отметить, что часто, особенно в случае исследования биологических объектов, направление связи, ее плотность (сила) или вообще существование сохраняются только в определенном интервале значений. За пределами этого интервала связь может ослабнуть, стать прямо противоположной по направлению или совсем исчезнуть. *Например*, при увеличении возраста человека, начиная от момента его рождения, сила скелетной мускулатуры увеличивается. Но после достижения пика развития организма, по мере дальнейшего увеличения возраста происходит снижение этой силы.

Исторически сложилось так, что по мере развития статистических методов оценок взаимосвязей возникла и закрепилась определенная путаница при применении термина «корреляция». Его стали применять и для обозначения явления – связи статистического характера, и для обозначения методологии численной оценки таких связей.

Считается, что термин «корреляция» как словесное описание объективно существующей в природе статистической по характеру взаимосвязи различных явлений впервые употребил в науке французский палеонтолог Жорж Кювьё в XVIII веке. Этот термин использовался им без

численных оценок для отображения, *например*, связи размера костей скелета и массы скелетной мускулатуры. В дальнейшем применение термина «корреляция» получило широкое развитие в самых разнообразных по тематике исследованиях: и при характеристике вида связей, и для описания принципа их численной оценки⁷⁷.

Развитие методологии статистического анализа привело к тому, что в настоящее время существуют два разных принципа численной оценки статистических связей: принцип *ковариаций* и принцип *взаимной сопряженности*. Причем разница между ними, нередко оставляемая без внимания, весьма существенна.

Если основанием для заключения о существовании связи служит наличие тенденции параллельного изменения, взаимной вариации, ковариации отдельных пар, триад и т.п., логически связанных количественных характеристик наблюдаемых явлений, то подобное заключение основано на принципе учета *ковариации*. В математическом отношении задача сводится к определению численной меры ковариации, и в качестве мер таких изменений могут использоваться коэффициенты ковариации, коэффициенты линейной корреляции Пирсона, коэффициенты ранговой корреляции Спирмена, Кендала, корреляционное отношение и др.

Принцип взаимной сопряженности предполагает установление связи в тех случаях, когда появление одного события или изменения его характеристики с определенной вероятностью связано с возникновением другого события. Например: появление кашля с мокротой может с определенной вероятностью свидетельствовать о возникновении у пациента хронического бронхита. Если в серии повторяющихся наблюдений одно из событий появляется одновременно с другим чаще, чем можно объяснить случайным стечением обстоятельств, то это служит основанием говорить о взаимосвязи, сопряженности появления этих событий. При этом могут учитываться либо дихотомические варианты (есть или нет то или иное

⁷⁷ Говоря о развитии научных исследований взаимоотношений костей скелета и мягких тканей, следует упомянуть уникальный метод восстановления утраченных человеческих образов, предложенный М.М. Герасимовым (1907-1970). В основе этого метода, как известно, лежит понимание связи (корреляции) толщины, вида и рельефа кожных покровов со скелетом, структурой и рельефом костей черепа и т.п.

событие), либо изменение частоты отдельных классов (градаций), вероятности сопряженных событий.

Статистические критерии оценки такого рода связей, основанные на учете частот, составляют довольно многочисленную группу критериев взаимной сопряженности или просто *сопряженности*: $K\phi$ (критерий знаков Фехнера), Q (коэффициент ассоциации Юла), Φ (коэффициент сопряженности, контингенции), C (коэффициент взаимной сопряженности Пирсона), K (коэффициент Чупрова) и др.

Здесь важно понимать, что критерии, основанные на учете ковариации и сопряженности, дают, каждый по-своему, оценку одного явления, называемого *статистической связью*. Следует помнить, что эта связь – объективно существующее явление, выявление которого и оценка в зависимости от конкретной ситуации может быть осуществлена различным способом, в том числе и через вычисление специальных коэффициентов.

11.1. Оценки статистических связей на принципе ковариаций

11.1.1. Коэффициент ковариации

В прямом переводе с английского языка *ковариация* – соизменимость, то есть взаимная изменчивость, взаимная вариация двух или более наблюдаемых величин (рис. 95).

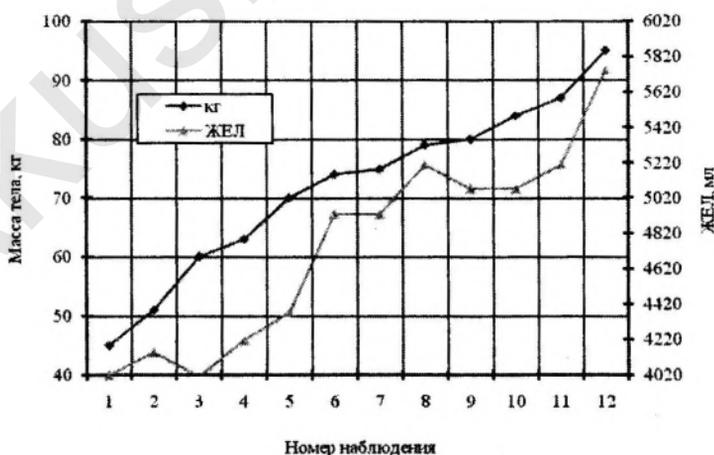


Рис.95. Диаграмма связи variability массы тела (кг) и жизненной емкости легких (ЖЕЛ, мл)

Если в случае двух рядов X и Y объективно существующие колебания (вариации) значений величин совпадают, то признается наличие между X и Y связи. Расчет коэффициента ковариации производится вычислением усредненной суммы произведений отклонений отдельных значений каждого из наблюдаемых признаков от своих групповых средних (табл. 113):

$$COV = \frac{\sum(X - M_x)(Y - M_y)}{N} = \frac{48683,0}{10} = 4868,3.$$

Знак коэффициента линейной ковариации (COV) указывает на направление связи: если $COV > 0$ – это означает прямую связь (при росте значений одного признака растут и значения другого), $COV < 0$ – указывает на обратную связь. При $COV = 0$ связь отсутствует. Достоинством этого коэффициента является относительная простота вычисления. Существенный недостаток – зависимость значения коэффициента ковариации от масштаба единиц измерений наблюдаемых величин. Этого недостатка лишен коэффициент корреляции Пирсона.

Таблица 113

Расчет ковариации (COV) массы тела и жизненной емкости легких (**ЖЕЛ**)

Номер наблюдения	X (масса, кг)	Y (ЖЕЛ, мл)	X - M_x	Y - M_y	(X - M_x)(Y - M_y)
1	50	4020	-18,4	-597,5	10994,0
2	50	4150	-18,4	-467,5	8602,0
3	60	4010	-8,4	-607,5	5103,0
4	65	4215	-3,4	-402,5	1368,5
5	70	4374	1,6	-243,5	-389,6
6	75	4927	6,6	309,5	2042,7
7	75	4927	6,6	309,5	2042,7
8	79	5412	10,6	794,5	8421,7
9	80	5070	11,6	452,5	5249,0
10	80	5070	11,6	452,5	5249,0
Σ	684	14675	-	-	48683,0

11.1.2. Коэффициент линейной корреляции Пирсона

В отличие от коэффициента ковариации для вычисления коэффициента корреляции Пирсона⁷⁸ используются не сами отклонения $(X - M_X)$ и $(Y - M_Y)$, а безразмерное отношение этих отклонений к дисперсиям. Поэтому значение коэффициента корреляции не зависит от единиц измерения. Например, показатель корреляции между ростом и весом будет одним и тем же, независимо от того, проводились измерения в дюймах и футах или в сантиметрах и килограммах. Коэффициент линейной корреляции Пирсона после соответствующего алгебраического преобразования коэффициента ковариации выглядит следующим образом:

$$R_{XY} = \frac{\text{cov}}{\delta_X \delta_Y} = \frac{\sum (x - M_X)(y - M_Y)}{\sqrt{\sum (x - M_X)^2 \sum (y - M_Y)^2}},$$

где M_X и M_Y – средние значения в коррелируемых рядах значений признаков X и Y, а δ_X и δ_Y – среднеквадратические отклонения в этих рядах. Для удобства «ручных» вычислений может применяться алгебраический аналог приведенной формулы, в котором отсутствуют средние величины:

$$R_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{N} \right]}}, \text{ где } N - \text{число коррелируемых пар.}$$

Коэффициент корреляции VJ; TN принимает значения от -1 до +1 в следующих границах качественных оценок силы связи по шкале Чеддока (табл. 114):

Таблица 114

Распределение оценок коэффициента корреляции

Сила связи	Прямая	Обратная
Полная (функциональная)	+1,0	-1,0
Сильная, плотная	От +0,1 до +0,7	От -0,1 до -0,7
Средней силы, средней плотности	От +0,7 до +0,3	От -0,7 до -0,3
Связь слабая, низкая плотность	От +0,3 до 0,0	От -0,3 до 0,0
Связь отсутствует	0,0	0,0

⁷⁸ Для обозначения коэффициента корреляции Пирсона часто применяется символ r . Поскольку в настоящем издании рассматривается пакет анализа Excel, то авторами применяется символ R , который используется в Excel.

Например, по данным табл. 113, коэффициент корреляции $R=0,94$. Следовательно, связь по направлению прямая, по силе – сильная, плотная.

Графическим, визуальным отображением силы и направленности корреляционной связи служат так называемые корреляционное поле и линия регрессии, которые представляют собой график, отображающий распределение отдельных значений X и Y в виде точек с соответствующими абсциссами и ординатами, а также аппроксимацию распределения этих точек с помощью уравнения регрессии. Например, подросток Петров имел рост (X) 130 см, а массу тела (Y) – 55 кг. Соответственно, точка его параметров будет обозначена на графике координатами: $X=130$ см, $Y=55$ кг. Сидоров имел рост 170 см и массу тела 75 кг. Его точка будет иметь координаты: $X=170$ см, $Y=75$ кг и т.д. (рис. 96). На этом же рисунке представлены линия и соответствующее ей уравнение линейной регрессии. Коэффициент детерминации позволяет определить значение коэффициента линейной корреляции Пирсона: $R=\sqrt{R^2}=\sqrt{0,605}=0,78$.

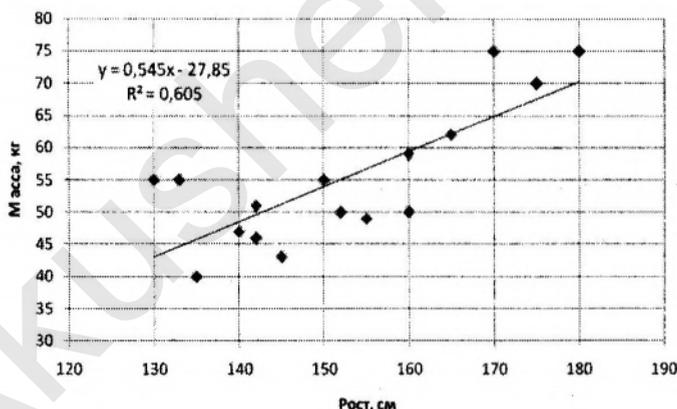


Рис.96. Распределение группы подростков по росту и массе тела

Как и при вычислении любого другого статистического коэффициента или показателя, полученного на основе выборочных данных, необходима оценка статистической репрезентативности (значимости) коэффициента корреляции. Приближенная оценка этой репрезентативности может приближенно осуществляться исходя из того, что абсолютное значение

коэффициента R должно превышать ошибку не менее чем в два раза. Более точно оценка репрезентативности производится через ошибку. Ошибка коэффициента линейной корреляции определяется по формуле:

$$m_{R_{XY}} = \sqrt{\frac{1 - R_{XY}^2}{N - 2}}$$

Оценка значимости коэффициента корреляции производится через

критерий Стьюдента: $t = \frac{|R_{XY}|}{m_{R_{XY}}}$, где $|R_{XY}|$ – абсолютная величина коэффициента корреляции, а $m_{R_{XY}}$ – ошибка коэффициента корреляции. В свою очередь, критерий t оценивается с учетом числа степеней свободы $(N-2)$, где N – число парных вариантов. Если для оценивания используется таблица критических значений, то значение критерия t должно быть равно или больше табличного, соответствующего вероятности $P \leq 0,05$. При этом учитывается характер статистической связи – односторонняя или двухсторонняя. В приведенном

примере (рис. 96): $m_{R_{XY}} = \sqrt{\frac{1 - 0,605}{16 - 2}} = 0,17$; $t = \frac{|R_{XY}|}{m_{R_{XY}}} = \frac{0,17}{0,14} = 1,2$, $P > 0,05$. Таким образом, коэффициент корреляции следует признать статистически незначимым.

Оценку статистической значимости (репрезентативности) полученных результатов можно произвести и с помощью специальных таблиц (см. Приложение 9). Обычно репрезентативность коэффициента корреляции определяется с учетом односторонней или двухсторонней направленности связи. В приведенном примере связь явно односторонняя, т.к. рост может определять массу тела, но не наоборот. При оценке приведенными выше способами статистической значимости коэффициента корреляции не учитывается его знак, оценка репрезентативности которого является отдельной задачей. Такого рода оценку несложно произвести с помощью специальной таблицы (табл. 115).

Число пар наблюдений (N), необходимое для подтверждения знака коэффициента корреляции. (М.Б. Славин, 1989)

R	P		R	P	
	0,05	0,01		0,05	0,01
0,01	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

Доверительный интервал коэффициента корреляции можно рассчитать путем несложных операций в *MS Excel*, воспользовавшись преобразованием Фишера (Виноградов А., 2010). Эти расчеты будут приведены в разделе настоящего издания, посвященного применению *MS Excel* для вычисления показателей статистической связи.

Вне зависимости от способа оценки различий коэффициентов корреляции, необходимо помнить, что репрезентативность этих различий обусловлена не только величиной разности сравниваемых коэффициентов и объемом наблюдений, но и величиной коэффициентов. Например, разница между коэффициентами корреляции, равными, соответственно, 0,10 и 0,20, может быть незначимой. А разница между коэффициентами, равными 0,80 и 0,90, полученными на тех же самых массивах наблюдений, может оказаться значимой.

Для вычисления коэффициента корреляции целесообразно использовать негруппированные ряды исходных данных, поскольку их непродуманная группировка может существенно исказить результаты любого статистического анализа. Кроме того, при использовании коэффициентов корреляции и ковариации необходимо помнить, что:

- их применение возможно только при наличии достаточного количества наблюдений, которое должно не менее чем в 5-6 раз превышать

число коррелируемых факторов. При меньшем объеме наблюдений статистическая ошибка коэффициентов возрастает до размеров, делающих бессмысленным их вычисление;

- в коррелируемых рядах недопустимы пропуски;
- любой показатель статистической связи не служит безусловным доказательством объективного существования непосредственных причинно-следственных отношений между рассматриваемыми факторами, которые могут быть взаимосвязаны опосредованно, через не учтенные в исследовании другие факторы;
- коэффициенты ковариации и корреляции Пирсона применимы только для оценки линейных взаимосвязей. В случае нелинейных зависимостей они дают заниженные значения.

Для оценки статистической достоверности корреляции необходимо, чтобы распределение совокупности значений факторных и результативных признаков подчинялось Гауссову (нормальному) распределению. Проверка этого условия может производиться либо путем использования специальных статистических тестов, либо приближенно, путем визуальной оценки графического отображения распределения коррелируемых признаков.

Еще одной особенностью как коэффициентов ковариации, так и корреляции является отсутствие у них свойства аддитивности. Т.е. их нельзя суммировать и, соответственно, вычислять усредненные значения результатов, полученных в нескольких аналогичных группах наблюдений. Если все же необходимо получить, например, средний коэффициент корреляции, то следует преобразовать отдельные коэффициенты корреляции в такую меру зависимости, которая будет аддитивной. Например, получить *коэффициенты детерминации*, которые при одинаковом или близком числе взаимодействующих факторов в группах наблюдений аддитивны, и путем извлечения из их суммы квадратного корня можно получить суммарный (усредненный) коэффициент корреляции.

11.1.3. Коэффициент детерминации

Рассматривая приведенную в примере связь массы тела и веса подростков (рис. 96), с помощью *коэффициента детерминации* R^2 можно определить формальную долю влияния факторного признака (в данном случае роста) на результирующий признак (массу тела). Численно коэффициент детерминации представляет собой квадрат коэффициента линейной корреляции. Если принять во внимание, что на величину массы тела влияют и другие факторы, то R^2 показывает долю тех изменений численных значений массы тела, которые обусловлены ростом обследованных подростков. В данном случае $R^2=0,605$, или $60,5\%$ ⁷⁹.

Коэффициент детерминации широко применяется в статистике как универсальная мера при оценке парных и множественных, линейных и нелинейных взаимосвязей (корреляционный и регрессионный анализ, дисперсионный анализ и т.п.). Многими исследователями считается, что коэффициент детерминации является более адекватным показателем связи, чем коэффициент корреляции.

Недостаток R^2 заключается в том, что его значение увеличивается при увеличении числа взаимодействующих факторов. Поэтому сравнение значений R^2 в исследованиях с разным числом взаимодействующих факторов некорректно. Дабы было возможным сравнивать результаты исследования с разным числом факторов так, чтобы число регрессоров (факторов) не влияло на статистику R^2 , используется *скорректированный коэффициент детерминации*:

$$R_{\text{ск}}^2 = \left(1 - (1 - R^2) \frac{(n-1)}{(n-k)} \right),$$

Где n — количество наблюдений, k — количество регрессоров.

Практическим аналогом коэффициента детерминации в случае криволинейной связи является квадрат корреляционного отношения η^2 (см. раздел «Нелинейная связь. Корреляционное отношение»).

⁷⁹ Понятно, что статистическая значимость этих показателей определяется значимостью предварительно вычисленных коэффициентов корреляции.

11.1.4. Коэффициент Фехнера

В некоторых исследованиях приходится сталкиваться с ситуацией, когда учетные признаки представлены не точными, а приближенными числовыми значениями. Наиболее простой коэффициент корреляции, который может использоваться в этой ситуации, – коэффициент парной корреляции Фехнера. Для того чтобы понять алгоритм его вычисления, рассмотрим пример его применения для доказательства вины конкретного источника загрязнения в ухудшении экологического состояния открытого водоема. Для решения этой задачи требуется оценить зависимость интенсивности запаха воды (в баллах) водоема от удаленности (в км) места забора проб от возможного источника загрязнения (табл. 116).

Таблица 116

Вычисление коэффициента парной корреляции Фехнера

Значения		Отклонения от средних		С, Н	
Х, км	У, баллы	d_x	d_y	С	Н
1,0	7	-1,9	2,9		1
1,5	6	-1,4	1,9		1
2,0	5	-0,9	0,9		1
2,5	4	-0,4	-0,1	1	
4,0	3	1,1	-1,1		1
4,5	2	1,6	-2,1		1
5,0	2	2,1	-2,1		1
$M_x=2,9$	$M_y=4,1$	–	–	1	6

Коэффициент парной корреляции Фехнера вычисляется по формуле: $f = \frac{(c-n)}{(c+n)}$, где С – число признаков, знаки отклонений которых от своих средних совпадают. Н – число признаков, знаки отклонений которых не совпадают. $f = \frac{(1-6)}{(1+6)} = \frac{-5}{7} = -0,71$.

Коэффициент Фехнера f оценивается по тем же правилам, что и рассмотренный выше коэффициент корреляции Пирсона. Помимо простоты

вычисления, коэффициент корреляции Фехнера обладает еще одним, не всегда приемлемым, свойством – большой приближенностью оценок связи. Более точные значения дают коэффициенты ранговой корреляции. Кроме того, они пригодны и для оценки качественных признаков.

11.1.5. Коэффициент ранговой корреляции Спирмена

Если исходные данные, вне зависимости от их качественного или числового выражения, расположить в порядке смыслового возрастания или убывания, иначе говоря, ранжировать, то порядковый номер каждой единицы наблюдения будет номером ее ранга. К коэффициентам, способным оценивать статистическую связь на основании таких рядов, относится **коэффициент ранговой корреляции Спирмена**.

Когда анализируемые ряды признаков X и Y взаимосвязаны, а изменения числовых значений их рангов в большую или меньшую сторону будут совпадать, то разность рангов будет минимальна, и наоборот. При этом, если между рядами X и Y имеется «полная прямая зависимость» – в том смысле, что ранги объектов полностью совпадают, то выборочный коэффициент ранговой корреляции будет равен единице. Если между признаками X и Y имеется полная противоположная зависимость, то коэффициент ранговой корреляции будет равен минус единице.

Достоинством коэффициентов корреляции рангов считается простота вычислений. Повсеместное применение компьютерных программ снизило существенность этого достоинства, что в совокупности с присущим коэффициентам ранговой корреляции «генетическим» недостатком – приближенностью результатов, сделало их применение малооправданным. В настоящее время применение ранговых коэффициентов корреляции может быть оправдано в случаях:

- когда необходимо быстро, ориентировочно определить связь между какими-то явлениями;
- если необходимо оценить связь между качественными признаками;
- когда распределение значений учетных признаков не соответствует нормальному распределению или распределение неизвестно.

Существует несколько различных коэффициентов ранговой корреляции. Наиболее распространенный из них – коэффициент корреляции рангов Спирмена, вычисляется он по формуле:

$$\rho = 1 - \frac{6 \sum d^2}{n^3 - n}$$

где d – разность между рангами, n – число сопоставляемых пар Y и X , 6 – константа уравнения.

Для примера рассмотрим похожие данные уже приводившейся выше задачи о зависимости интенсивности запаха воды (в баллах) водоема от удаленности (в км) места забора проб от возможного источника загрязнения. Вычисления выполняются в следующей последовательности (табл. 117):

Таблица 117

Вычисление коэффициента корреляции рангов Спирмена

Значения		Ранги		Разности	
X, км	Y, баллы	X	Y	d	d ²
1,0	7	1,5	8	-6,5	42,25
1,0	6	1,5	7	-5,5	30,25
2,0	5	3	6	-3,0	9,00
2,5	4	4	5	-1,0	1,00
4,0	3	5	4	1,0	1,00
4,5	2	6	2,5	3,5	12,25
5,0	2	7	2,5	4,5	20,25
5,5	1	8	1	7,0	49,00
$\sum d^2 = 165$					

- располагаем значения признака X , ранжируя их в первом столбце таблицы по мере возрастания;

- в соседнем, втором, столбце располагаем соответствующие признаку X значения Y . Специального ранжирования в этом столбце уже не проводится. Все значения во втором столбце располагаются только соответственно парным Y значениям X ;
- в третьем столбце выставляются ранги значений по ряду X . Самому малому значению $X=1,0$ (вверху столбца) нужно присвоить ранг 1. Но следующее значение X опять равно 1,0, поэтому каждому из них присваивается средний ранг $(1+2)/2=1,5$. Т.е. каждому из одинаковых значений приписывается ранг, равный среднему арифметическому рангу одинаковых объектов;
- в четвертом столбце таблицы размещаются ранги значений Y . Номера рангов выставляются согласно логике ранжирования данных первого столбца: начиная от самых малых значений Y . Самое малое значение $Y=1$ (внизу столбца) получает ранг 1. Затем идет значение $Y=2$, которое получает ранг 2. Но на третьем месте стоит значение запаха в баллах, равное 2. Т.е. эти баллы занимают 2-е и 3-е места. Соответственно, им обоим присваивается усредненное значение ранга $(2+3)/2=2,5$;
- после этого нумерация рангов продолжается в той же последовательности. В пятом столбце записывается разность рангов ряда $X-Y$, в шестом – квадрат этой разности;
- полученные значения квадратов суммируются и подставляются в формулу:

$$\rho = 1 - \frac{6 \sum 107}{8 \cdot (8^2 - 1)} = -0,96$$

Таким образом, связь между интенсивностью запаха воды и удаленностью от возможного источника загрязнения сильная, обратная, т.е. чем больше удален источник загрязнения, тем меньше интенсивность запаха.

Статистическая значимость коэффициента Спирмена оценивается аналогично коэффициенту линейной корреляции Пирсона.

11.2. Частная (парциальная) корреляция

При оценке взаимосвязи факторных и результативных признаков наиболее существенной проблемой является анализ взаимодействия не двух, а гораздо чаще встречающихся взаимодействий 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, *например*, на группе больных; в коллективе рабочих; среди определенных, ограниченных по численности контингентов населения и т.п., этот способ малопригоден. Главная причина – практическая невозможность подобрать в натуральных условиях лиц с набором одинаковых исходных характеристик (возраст, пол, стаж и т.п.) и подвергающихся воздействию «в чистом виде» только одного или максимум двух факторов. Другой подход – применение специальных статистических методов, с помощью которых производится последовательная элиминация влияния одних факторов и выделение результатов влияния других факторов. К таким методам относится частная (парциальная) корреляция.

Например, в ходе вычисления коэффициентов частной корреляции для трех признаков последовательно элиминируется влияние одного из трех 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_1 и X_2 . Реализуются эти расчеты следующим образом.

Элиминирование влияния третьего признака и выявление связи между первым и вторым производится по формуле:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Аналогичным образом производится элиминирование влияния второго признака и выявление связи между первым и третьим:

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Затем производится выявление связи между третьим и вторым признаком при элиминировании третьего:

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Например, требуется провести сравнительную оценку взаимосвязи факторов: длительности рабочего времени, усталости (число ошибок набора) и интенсивности труда (количество набранных страниц текста) на персональном компьютере (табл. 118).

Таблица 118

Исходные данные для расчета частной корреляции

Время работы (час)	Число ошибок	Число страниц текста (за час работы)
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 связи частоты ошибок (проявление усталости) и длительности рабочего времени ($R_{12} = 0,41$ – связь прямая, средней силы) и о снижении производительности труда по мере увеличения времени работы ($R_{13} = -0,36$ – связь обратная, средней силы). Показатели увеличения усталости и снижения производительности труда также статистически связаны ($R_{23} = -0,74$ – связь обратная, сильная).

На основе рассчитанных коэффициентов парных корреляций можно вычислить соответствующие коэффициенты частной корреляции:

$$r_{12.3} = \frac{0,4 - (-0,7)(-0,4)}{\sqrt{(1 - (-0,7^2))(1 - (-0,4^2))}} = 0,2$$

$$r_{13.2} = \frac{0,4 - 0,4(-0,7)}{\sqrt{(1 - (-0,7^2))(1 - 0,4^2)}} = -0,1$$

$$r_{23.1} = \frac{-0,7 - 0,4(-0,4)}{\sqrt{(1 - (-0,4^2))(1 - 0,4^2)}} = -0,7$$

Анализ полученных коэффициентов частной корреляции (табл. 119) показывает, что при устранении фактора усталости ($R_{12.3}$) существенно уменьшается связь между качеством труда и длительностью рабочего времени.

Таблица 119

Итоговые значения коэффициентов корреляции

R_{12}	0,41	$R_{12.3}$	0,24
R_{13}	-0,36	$R_{13.2}$	-0,08
R_{23}	-0,74	$R_{23.1}$	-0,70

11.3. Множественная корреляция. Корреляционная матрица

Метод множественной корреляции – метод многомерного анализа, обычно применяемого для характеристики совместного, совокупного влияния комплекса факторов на резульативный признак (фактор). Изменяется величина коэффициента множественной корреляции, оставаясь всегда положительной, по абсолютному значению может изменяться в пределах от 0 до 1, как и коэффициенты парной или частной корреляции. Например, для данных, приведенных в табл. 118, коэффициент корреляции между первым признаком и двумя другими (вторым и третьим) будет выглядеть так:

$$r_{1.23} = \sqrt{1 - (1 - r_{12}^2)(1 - r_{13}^2)} = \sqrt{1 - (1 - 0,4^2)(1 - (-0,7^2))} = 0,75$$

На основании коэффициента множественной корреляции определяется соответствующий коэффициент детерминации: $(0,75^2=0,6)$, который показывает долю совместного влияния 2-го и 3-го признака на 1-й (60%).

Более точный способ вычисления коэффициента множественной корреляции – с помощью метода, родственного для множественной корреляции, – множественной регрессии (см. раздел «Параметрические показатели статистической связи в MS Excel»).

Корреляционная матрица – объект, элементами которого являются показатели различных вариантов парной взаимосвязи нескольких признаков (факторов), при этом число рассматриваемых признаков практически не ограничено. Представляет собой симметричную квадратную матрицу размером $M \times M$, где M – число исследуемых факторов, главная диагональ которой заполнена единицами (или нулями – для удобства дальнейшего анализа), а недиагональные элементы представляют собой меру тесноты связи между парой факторов (коэффициент корреляции, корреляционное отношение, индекс Фехнера и т.д.).

Другими словами, для заполнения корреляционной матрицы необходимо провести анализ исходных данных по принципу «каждый с каждым любым известным способом».

Для данных, представленных в табл. 118, корреляционная матрица, где мерой связи выступают коэффициенты линейной корреляции Пирсона, будет выглядеть следующим образом (табл. 120):

Таблица 120

**Пример корреляционной матрицы
(в показателях линейной корреляции Пирсона)**

Учетные признаки	Время работы (час)	Число ошибок	Число страниц текста
Время работы (час)	1,00	0,41	-0,36
Число ошибок	0,41	1,00	-0,74
Число страниц текста	-0,36	-0,74	1,00

коэффициенты линейной корреляции, то матрица будет абсолютно симметричной относительно главной диагонали. Поэтому при представлении такого рода данных обычно заполняют только одну из частей матрицы, выше или ниже диагонали.

Если же в качестве меры связи используются показатели, описывающие нелинейные корреляции, *например*, показатели корреляционного отношения, то матрица заполняется несимметричными данными (табл. 121). В этом случае в матрице должны быть отражены результаты всех вариантов взаимосвязей.

Таблица 121

**Пример корреляционной матрицы
(в показателях корреляционного отношения η)**

Учетные признаки	Время работы (час)	Число ошибок	Число страниц текста
Время работы (час)	1,00	0,71	0,52
Число ошибок	0,85	1,00	0,82
Число страниц текста	0,77	0,60	1,00

Корреляционная матрица является исходным объектом при использовании практически всех методов многомерного статистического анализа. С помощью нее в ходе предварительного анализа (так наз. «разведочного» анализа) определяются признаки (факторы), которые:

- предельно тесно (функционально) взаимосвязаны между собой, что дает основание исключать их из массива наблюдений;
- нелинейно связаны ($R < 0,8$). Такого рода признаки в случае анализа линейных взаимосвязей либо элиминируются из совокупности, либо при анализе этой совокупности используются специальные методы, позволяющие анализировать распределение нелинейно взаимосвязанных признаков.

11.4. Нелинейная связь. Корреляционное отношение

Если значение коэффициента линейной корреляции оказалось невысоким, это не означает, что связь действительно слабая. Возможно, что связь факторного и результирующего признаков носит криволинейный характер, не улавливается коэффициентами линейной корреляции. Универсальной характеристикой, применение которой не ограничено формой статистической связи, является *корреляционное отношение* η , принимающее в зависимости от силы взаимосвязи всегда положительные значения в пределах от 0 до 1. Этот показатель нередко именуется как *универсальное корреляционное отношение*.

В ситуации, когда анализируемая связь носит линейный характер, значение коэффициента корреляционного отношения равно коэффициенту линейной корреляции. Исключением здесь является то, что коэффициент линейной корреляции указывает направление связи (положительная или отрицательная), а корреляционное отношение – нет. Такая особенность имеет определенное смысловое значение: нелинейная связь, характеристикой которой является корреляционное отношение, может менять свою направленность даже в пределах одного ряда наблюдений. Например, парабола на своем восходящем отрезке отражает прямую связь, в нисходящем – обратную.

Еще одной особенностью корреляционного отношения является раздельное оценивание влияния X на Y и Y на X . Т.е. при оценке нелинейной зависимости обычно вычисляются два варианта: η_{xy} и η_{yx} .

В общем виде формула корреляционного отношения выглядит следующим образом: $\eta_{xy} = \frac{\sigma_{xy}}{\sigma_y}$ и $\eta_{yx} = \frac{\sigma_{yx}}{\sigma_x}$, где σ_{yx} – среднее квадратическое отклонение, представляющее изменчивость Y под влиянием только X , σ_{xy} – изменчивость X под влиянием только Y , σ_y и σ_x – среднее квадратические отклонения, характеризующие изменчивость признаков X и Y в целом.

Во многих программных пакетах статистической обработки квадрат корреляционного отношения и квадрат коэффициента корреляции обозначаются одинаково – R^2 . Однако, строго говоря, это разные показатели, характеризующие одно явление – статистическую связь. Схожи они только в случае линейной связи. Если связь линейная, то коэффициент детерминации линейной связи и корреляционные отношения равны: $R^2_{xy} = \eta^2_{xy} = \eta^2_{yx}$. Если связь не линейна, то равенство не выполняется: $R^2_{xy} \neq \eta^2_{xy} \neq \eta^2_{yx}$.

Например, при использовании коэффициента линейной корреляции Пирсона для оценки зависимости второго фактора – «развития усталости (число допускаемых ошибок)» – от первого – «продолжительности рабочего времени» (табл. 118) – получено невысокое значение коэффициента линейной корреляции: $R_{12} = 0,24$. Одной из причин такой оценки связи могла быть нелинейная зависимость этих факторов: как правило, в течение рабочего дня работоспособность сначала повышается (происходит «вработывание»), затем начинает прогрессивно снижаться. Эта гипотеза подтверждается и диаграммой, отражающей сравнительные итоги линейной и нелинейной аппроксимации исходного распределения (рис. 97).

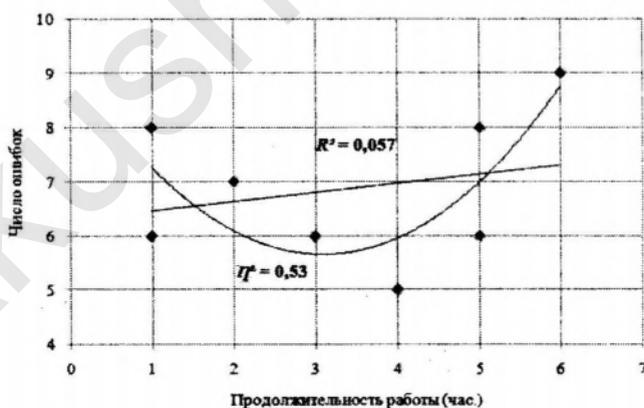


Рис. 97. Сравнительные результаты линейной и нелинейной аппроксимации распределения точек парных значений продолжительности рабочего времени и числа допускаемых ошибок

Как видно из представленных данных, в случае линейной аппроксимации коэффициент линейной детерминации $R^2_{12} = 0,057$ и, соответственно,

коэффициент линейной корреляции $R_{12} = \sqrt{0,057} = 0,24$. Нелинейная аппроксимация этого же массива данных полиномом третьей степени позволила получить $R_{12}^* = 0,53$, соответствующее корреляционное отношение $\eta_{12} = \sqrt{0,53} = 0,74$. То есть во втором случае была выявлена сильная статистическая связь между длительностью работы и развитием усталости (числом ошибок)⁸⁰.

Как уже отмечалось, при нелинейной взаимосвязи параметры регрессии первого учетного признака (продолжительности рабочего времени) и второго (числа ошибок) существенно отличаются. На рис. 98 представлены распределения, обратные распределению, отображенному на рис. 97⁸¹.

Нетрудно заметить, что коэффициент детерминации в первом и втором случаях не отличается друг от друга ($R^2_{21} = 0,057$; $R^2_{12} = 0,057$). При нелинейной связи различия η_{12} и η_{21} существенны ($\eta_{12} = 0,53$; $\eta_{21} = 0,38$).

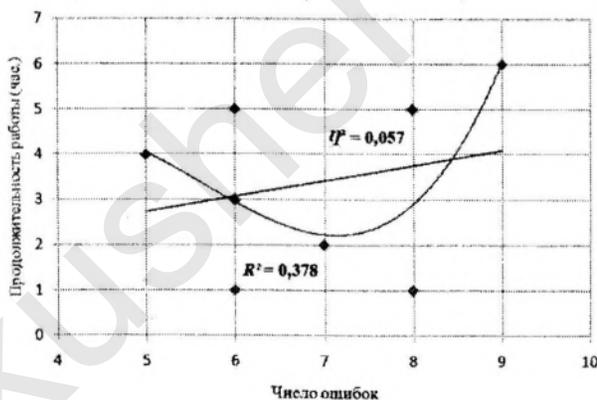


Рис.98. Сравнительные результаты линейной и нелинейной аппроксимации распределения парных значений числа ошибок и продолжительности рабочего времени

⁸⁰ Подробнее о практическом получении этими способами коэффициентов детерминации и корреляции см. в разделе «Показатели статистической связи в MS Excel».

⁸¹ Понятно, что взаимоотношение факторов, которое представлено на Рис. 97, не имеет логического смысла и представлено лишь с познавательной целью.

Оценка существенности различий корреляционного отношения и коэффициентов линейной корреляции может быть проведена несколькими способами. Наиболее простой, но наименее точный: связь считается нелинейной, если разность квадрата корреляционного отношения и квадрата коэффициента линейной корреляции превышает 0,1; то есть $(\eta_{xy}^2 - R_{xy}^2) > 0,1$. Более точные результаты оценки статистической достоверности (существенности) различий этих коэффициентов даются с помощью F критерия Фишера на основе показателя:

$$L = \frac{\eta^2(n-k)}{(k-2)(1-\eta^2)}$$

Если значение $L \geq F_{0,05}$, то гипотеза о линейности корреляции отвергается при $P=0,05$. Число степеней свободы определяется по формулам: $f_1=k-2$; $f_2=n-k$, где k – число коррелируемых пар; n – общее число вариант в обоих коррелируемых рядах.

Например, $R=0,51$; $R^2=0,51*0,51=0,26$; $\eta^2=0,72$ при $n=20$ и $k=10$. Тогда:

$$L = \frac{0,72(20-10)}{(10-2)(1-0,72)} = 3,21$$

При степенях свободы $f_1=10-2=8$ и $f_2=20-10=10$, $F_{0,05}=3,07$, т.е. $L > F_{0,05}$, что свидетельствует о статистически значимом отсутствии линейной корреляции.

При практических вычислениях в *MS Excel* удобней пользоваться вычислением и оценкой значимости корреляционных отношений с помощью регрессионного анализа тренда распределений в графическом пакете или данными двухфакторного (при парной связи) дисперсионного анализа. Практические приемы этих вычислений будут рассмотрены в разделе «Вычисление показателей статистической связи в *MS Excel*» настоящего издания.

11.5. Оценка взаимосвязи динамических рядов. Автокорреляция

В практической медицинской статистике оценка степени взаимосвязи изменений уровней 2-х или более рядов динамики чаще всего решается методами линейной корреляции. При этом в качестве единиц счета берутся численные значения исходных уровней взаимодействующих рядов динамики или отклонения фактических уровней динамических рядов от их трендов.

Особенностью оценивания взаимосвязи динамических рядов является необходимость предварительного анализа зависимых рядов на присутствие в них явления, называемого *автокорреляция*.

Автокорреляция – корреляция ряда с самим собой с задержкой на k интервалов (лаг, порядок автокорреляции). Следует помнить, что зависимость последующих уровней от предыдущих характерна для большинства динамических рядов, поэтому выявление и определение существенности автокорреляции является весьма актуальной задачей при изучении динамики явлений.

Если в динамических рядах имеется автокорреляция, то попытка оценки статистической связи между этими рядами неизбежно ведет к серьезным ошибкам. Поэтому прежде, чем анализировать взаимосвязь отдельных динамических рядов, необходимо проверить каждый из них на наличие автокорреляции.

Анализ набора результатов вычисления (коррелограммы) показателей автокорреляции позволяет выявить присутствие в динамических рядах той или иной статистически устойчивой компоненты, взятой при разных лагах. Как правило, при последовательных значениях $k=1, 2, 3$ и 4 . Обычно ограничением здесь является протяженность динамических рядов, т.е. количество временных точек счета.

Если существование автокорреляции нашло свое статистическое подтверждение и наиболее высоким оказался коэффициент автокорреляции первого порядка, то исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции, рассчитанный с более высоким лагом ($k=m$), то ряд содержит циклические колебания с

периодичностью в m интервалов наблюдаемого ряда. Если ни один из коэффициентов автокорреляции не значим, то:

- либо значения уровней ряда определяются только случайной компонентой;
- либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ, т.к. линейные коэффициенты автокорреляции отражают только линейную связь⁸².

Принципиально методики расчета коэффициента парной линейной корреляции Пирсона и автокорреляции почти не отличаются. Не отличаются и правила толкования их результатов. Вместе с тем, необходимо помнить, что по знаку коэффициента автокорреляции нельзя делать вывод о направленности тенденции ряда.

При вычислении коэффициента автокорреляции фактические уровни одного ряда рассматриваются как значения *факторного признака* X , а уровни этого же ряда со сдвигом на один или несколько периодов (лаг автокорреляции) принимаются в качестве *результативного признака* Y (табл. 122 и табл. 123).

Коэффициент автокорреляции с лагом, равным единице ($k=1$), будет равен (табл. 122):

$$R_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{N} \right]}} = \frac{1597212 - \frac{3015,4 \cdot 3066,3}{9}}{\sqrt{\left[1020274 - \frac{3015,4^2}{9} \right] \cdot \left[1030099 - \frac{3066,3^2}{9} \right]}} = 0,82$$

Его ошибка $m_{R_{XY}} = \sqrt{\frac{1 - R_{XY}^2}{N - 2}} = \sqrt{\frac{1 - 0,82^2}{9 - 2}} = 0,2$, тогда $t = \frac{|R_{XY}|}{m_{R_{XY}}} = \frac{0,82}{0,23} \approx 3,6$ при

значимости $P > 0,05$. Таким образом, установлена статистически подтвержденная, сильная прямая связь – автокорреляция ряда с лагом $k=1$.

⁸² Для динамических рядов, распределенных нелинейно, коэффициент автокорреляции может приближаться к нулю.

Анализ динамического ряда показателей первичной заболеваемости жителей Санкт-Петербурга злокачественными новообразованиями (случ. на 100 тыс. населения), присутствие автокорреляции при $k=1$

Годы	Злокач. новообразования (на 100 тыс.нас.)	k=1				
		X	Y	X ²	Y ²	XY
1987	323,6	323,6	319,6	104 717,0	102 144,2	103 422,6
1988	319,6	319,6	323,9	102 144,2	104 911,2	103 518,4
1989	323,9	323,9	326,1	104 911,2	106 341,2	105 623,8
1990	326,1	326,1	324,2	106 341,2	105 105,6	105 721,6
1991	324,2	324,2	343,6	105 105,6	118 061,0	111 395,1
1992	343,6	343,6	355,6	118 061,0	126 451,4	122 184,2
1993	355,6	355,6	348,2	126 451,4	121 243,2	123 819,9
1994	348,2	348,2	350,6	121 243,2	122 920,4	122 078,9
1995	350,6	350,6	374,5	131 299,7	122 920,4	699 447,0
1996	374,5	—	—	—	—	—
Сумма		3 015,4	3 066,3	1 020 274,4	1 030 098,5	1 597 212,5

Коэффициент автокорреляции с лагом, равным двум ($k=2$), будет равен (табл. 123):

$$R_{XY} = \frac{919093,2 - \frac{2664,8 \cdot 2746,7}{8}}{\sqrt{\left[88974,7 - \frac{2664,8^2}{8}\right] \cdot \left[945284,2 - \frac{2746,7^2}{8}\right]}} = 0,68$$

Его ошибка $m_{R_{XY}} = \sqrt{\frac{1 - 0,68^2}{8 - 2}} \approx 0,26$, тогда $t = \frac{|R_{XY}|}{m_{R_{XY}}} = \frac{0,68}{0,26} \approx 2,3$ при значимости $P < 0,05$.

Таким образом, статистически подтвержденной связи не установлено, т.е. автокорреляции с лагом $k=2$ не выявлено.

Поскольку статистически значимого коэффициента автокорреляции не выявлено, для выявления возможной нелинейной тенденции необходимо вычислить коэффициент автокорреляции из логарифмов исходных уровней. Отличное от нуля значение этого коэффициента будет свидетельствовать о возможной нелинейной тенденции ряда.

Анализ динамического ряда показателей первичной заболеваемости жителей Санкт-Петербурга злокачественными новообразованиями (случ. на 100 тыс. населения), отсутствие автокорреляции при $k=2$

Годы	Злокач. новообразования (на 100 тыс. населения)	$k=2$				
		X	Y	X ²	Y ²	XY
1987	323,6	323,6	323,9	104 717,0	104 911,2	104 814,0
1988	319,6	319,6	326,1	102 144,2	106 341,2	104 221,6
1989	323,9	323,9	324,2	104 911,2	105 105,6	105 008,4
1990	326,1	326,1	343,6	106 341,2	118 061,0	112 048,0
1991	324,2	324,2	355,6	105 105,6	126 451,4	115 285,5
1992	343,6	343,6	348,2	118 061,0	121 243,2	119 641,5
1993	355,6	355,6	350,6	126 451,4	122 920,4	124 673,4
1994	348,2	348,2	374,5	121 243,2	140 250,3	130 400,9
1995	350,6					
1996	374,5					
Сумма		2 664,8	2 746,7	888 974,7	945 284,2	916 093,2

Поскольку статистически значимого коэффициента автокорреляции не выявлено, для выявления возможно присутствующего нелинейного тренда необходимо попробовать провести выравнивание ряда с помощью нелинейной регрессии.

Поскольку по результатам проведенного анализа нашло свое статистическое подтверждение только автокорреляция первого порядка, то исследуемый ряд содержит статистически устойчивую тенденцию. Направленность и численное описание этой тенденции можно дать с помощью традиционных показателей динамического ряда.

11.6. Регрессия

Особым методом статистического описания взаимосвязей количественных данных является регрессионный анализ (регрессия), описывающий зависимость усредненного значения какой-либо численной характеристики, часто называемой зависимой переменной, одного явления (фактора) от характеристик независимых переменных (других факторов). Этот метод включает целый комплекс отдельных методик вычисления параметров

регрессии и статистического моделирования взаимосвязей. Конечной целью здесь является определение характера и степени детерминированности зависимой переменной (Y) одной (X) или несколькими ($X_1, X_2 \dots X_i$) независимыми переменными, часто называемыми *регрессорами*.

Уравнение линейной парной регрессии (одна переменная зависимая, одна независимая) выглядит следующим образом: $Y = ax + b$, где Y – зависимая переменная, x – независимая переменная, a и b – коэффициенты регрессии.

В ходе анализа любого вида регрессий (прямой, линейной или нелинейной; парной, двухфакторной или многофакторной регрессии) решаются следующие задачи:

- определение формы связи;
- определение параметров уравнения регрессии (модели регрессии), описывающей выявленную связь;
- статистическая проверка регрессионной модели, в т.ч. ее адекватности изучаемой совокупности данных.

Регрессионный анализ, несмотря на относительно большую громоздкость и методическую сложность получения его результатов по сравнению с получением результатов корреляционного анализа, отличается существенно большей информативностью. Например, полученные с помощью коэффициентов линейной корреляции Пирсона сравнительные оценки применения 2-х различных препаратов говорят об одинаковой силе и направленности влияния этих препаратов на один из биохимических параметров, характеризующих угнетение патологического процесса (его значения отражены по вертикальной оси). Основа этого сравнения, коэффициент линейной корреляции Пирсона, в обоих случаях был почти предельно высоким и одинаковым: $R_1=0,97$ и $R_2=0,97$. Однако регрессионный анализ распределения тех же исходных данных свидетельствует о существенном различии итогов применения этих препаратов (рис. 99).

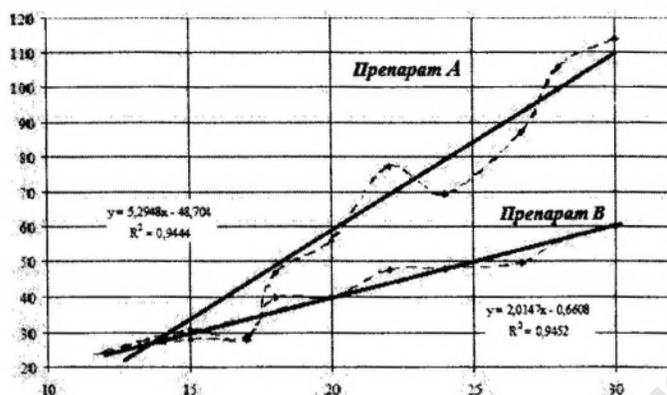


Рис.99. Итоги регрессионного анализа результатов применения двух препаратов

Собственно статистический термин «регрессия» родился при изучении Ф.Гальтоном устойчивости наследования в поколениях такой физической характеристики человека, как рост. Иначе говоря, насколько рост детей связан (коррелирует) с ростом родителей. В частности, Ф. Гальтон установил, что у высоких родителей рождаются высокие дети. Однако высокий рост детей чаще всего оказывался несколько ниже роста их родителей. Таким образом, происходит постепенная *регрессия* этого признака физического развития. Статистический метод описания выявленной закономерности, примененный исследователем, и получил название *регрессионный анализ*, который повсеместно стал применяться в математической статистике уже вне зависимости от особенностей изучаемых взаимосвязей: обратных или прямых, линейных или нелинейных и т.п.

Линии регрессии служат показателями формы статистической связи. Алгебраические описания, т.е. уравнения этих линий, называют *уравнениями регрессии*. Иногда эти уравнения называют *уравнениями корреляционной связи*, или *уравнениями корреляции*. Параметры уравнений зависят от формы связи, которая определяется по форме *корреляционного поля*. Распределение точек на графике визуально представляет тесноту (плотность, силу) связи. Если связь слабая или отсутствует, то точки распределяются бессистемно по всей площади

графика в пределах значений, которые могут принимать x и y . Если связь сильная (плотная), то точки располагаются плотно, вдоль некоторой результирующей линии, с помощью которой можно *аппроксимировать* (заменить) массу точек близким по характеру распределения объектом – одной линией, называемой *линией регрессии*. Чем более тесна корреляционная связь, тем более плотно около линии регрессии располагаются точки. На графике видно и направление связи: прямая или обратная. Как уже говорилось, в случае линейной зависимости y от x уравнением регрессии является уравнение прямой в системе двух координат: $y=ax+b$, где y – значение результирующего признака (зависимая переменная), x – значение факторного признака (независимая переменная, регрессор), a и b – коэффициенты линейной регрессии.

Форма связей (линейная, нелинейная) проявляется в виде разнообразных форм линии регрессии, описываемых различными уравнениями регрессии с разными параметрами (рис. 100).

Вычисление параметров линий регрессии из-за громоздкости расчетов рекомендуется осуществлять с помощью специальных программ статистической обработки данных. Большими возможностями для такого рода операций обладает и *MS Excel*.

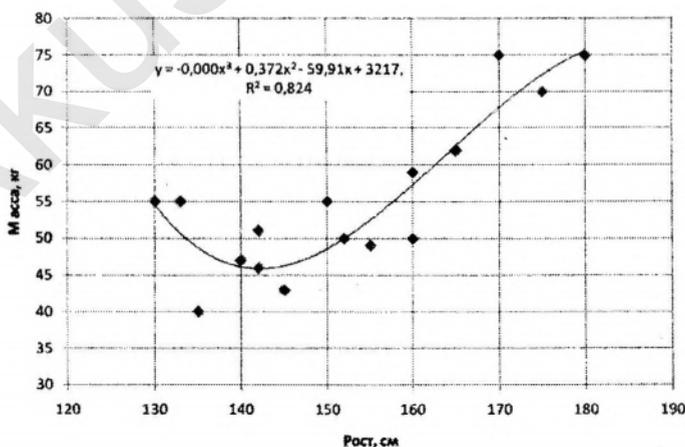


Рис.100. Аппроксимация распределения группы подростков по росту и весу (уравнение регрессии, полином третьей степени)

Регрессия весьма широко используется при статистических оценках результатов медико-биологических исследований. Большое распространение она получила и при стандартизации (нормировании) разного рода физиологических показателей. Простейшим примером такого применения линейной регрессии может служить индекс Брока, который используется как росто-весовой индекс для исчисления стандартного, нормального веса: из роста (в см.) вычитают 100 и получают нормальный вес (кг), который соответствует этому росту. Алгоритм расчета этого индекса записывается в виде уравнения линейной регрессии $y=ax+b$, где y – вес, x – рост, $b=-100$, a – поправочный коэффициент, который изменяется в зависимости от возрастной группы. Другие антропометрические индексы, построенные на основе регрессии и давно применяющиеся на практике, представлены в разделе «Физическое развитие» настоящего издания.

Иногда при измерении расстояний на местности прибегают к счету пройденного расстояния шагами. Соотношение длины шага и роста человека описывается уравнением линейной регрессии: $L=37+h/4$, h – рост в см., L – длина его шага. Отсюда, после несложного алгебраического преобразования, можно получить уравнение для определения роста человека по следам его шагов: $h=(L-37)*4$.

Полная статистическая оценка взаимосвязи требует нахождения уравнения регрессии не только для зависимости y от x , но и для зависимостей x от y . В силу многих причин результаты этих вычислений не будут зеркально похожи. В любом случае, использование современных средств статистической обработки данных не требует в обоих случаях проведения сложных операций. Однако, для понимания методических основ статистики взаимосвязей полезно будет остановиться на их деталях. Поскольку порядок вычисления в обоих случаях будет похожий, ограничимся рассмотрением основ обработки уравнения $y=ax+b$ (зависимость y от x).

В уравнении $y=ax+b$ коэффициент b равен высоте точки пересечения линией регрессии оси ординат (от 0 точки отсчета). Коэффициент a равен тангенсу угла наклона линии регрессии. Этот коэффициент, называемый коэффициентом регрессии, имеет большой практический 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°C частота пульса увеличивается в среднем на 10 ударов в минуту.

Статистический анализ подразумевает решение уравнения регрессии, т. е. отыскание параметров a и b этого уравнения. Математически решение уравнения регрессии сводится к вычислению указанных параметров таким образом, чтобы точки исходных данных корреляционного поля как можно ближе лежали к прямой регрессии, обеспечивая наименьший квадрат отклонений этих точек от линии регрессии (метод наименьших квадратов):

$$a = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$

Пример: найти выборочное уравнение регрессии по данным 8 наблюдений (табл. 124) зависимой (y – уровень систолического давления в мм рт. ст.) и независимой переменной (x – возраст в годах).

Таблица 124

Расчетная таблица параметров регрессии

Пары наблюдений	x возраст (лет)	y систолическое давление (мм рт.ст.)	x^2	xy
1	19	115	361	2185
2	20	117	400	2340
3	21	120	441	2520
4	21	120	441	2520
5	22	120	484	2640
6	23	125	529	2875
7	25	130	625	3250
8	24	125	576	3000
Итого	175	972	3857	21330

Согласно приведенным выше уравнениям, параметры a и b равны:

$$a = \frac{8 * 21330 - 175 * 972}{8 * 3857 - 175^2} = 2,33$$

$$b = \frac{972}{8} - 2,33 * \frac{175}{8}$$

Отсюда искомое уравнение регрессии $y=2,33*x+70,36$. График, уравнение линейной регрессии и коэффициент детерминации R^2 и представлены на (рис. 101).

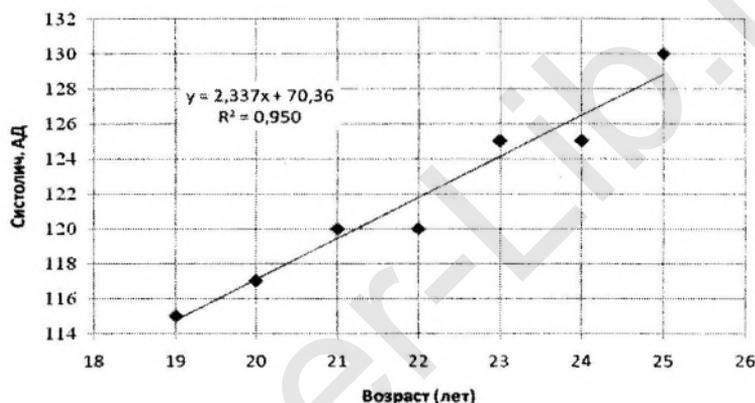


Рис.101. Корреляционное поле, график и уравнение линейной регрессии (по данным табл. 124)

Одним из найденных параметров регрессии был коэффициент регрессии, равный 2,33. Этот коэффициент показывает, что при изменении в рассмотренном интервале значений независимой переменной (возраста) на 1 год зависимая переменная (систолическое давление) будет увеличиваться в среднем на 2,33 мм рт. ст.

Как уже отмечалось выше, полная оценка взаимосвязи с помощью регрессии требует нахождения параметров регрессии не только для зависимости X от Y , но и Y от X . Но, разумеется, необходимо помнить о целесообразности такого рода вычислений. Так, задача определения зависимости возраста от уровня систолического давления в клинической практике представляется абсурдной.

Коэффициент детерминации показывает, какая доля вариации зависимого признака Y учтена в построенной регрессионной модели и обусловлена влиянием на него факторного, независимого признака (признаков). Чем ближе R^2 к 1, тем выше качество модели. Этот коэффициент является универсальным, так как отражает и тесноту связи, и точность модели, и может использоваться при оценке любой формы статистической связи. Поскольку в данном примере рассмотрена линейная зависимость, по значению R^2 легко определить значение коэффициента линейной корреляции Пирсона $R = \sqrt{R^2} = \sqrt{0,959} = 0,979$.

Другой способ определить качество регрессионной модели – проверить, насколько близки расчетные и фактические (исходные) данные. Это делается путем подстановки исходных значений (из табл. 125) в уравнение регрессии $y = 2,337 * x + 70,36$ (рис. 101). Например, для первого наблюдения, когда возраст равен 19, расчетное значение $y_1 = 2,33 * 19 + 70,36 \approx 115$ соответствует исходному.

Нетрудно заметить, что между исходными (фактическими) и расчетными значениями в ряде случаев существует определенная разница, так наз. *остатки* регрессии (табл. 125).

Таблица 125

Различия фактических и вычисленных показателей систолического давления (остатки регрессии)

Возраст	Систолическое давление		Остатки
	Фактическое	Расчетное	
19	115	115	0
20	117	117	0
21	120	119	-1
21	120	119	-1
22	120	122	+2
23	125	124	-1
25	130	129	-1
24	125	126	+1

Анализ этих остатков дает представление, насколько адекватна регрессионная модель. Согласно общим положениям, если регрессия

удовлетворительно описывает истинную зависимость между y и x , то остатки должны быть независимыми, нормально распределенными случайными величинами с нулевым средним, а в значениях остатков должен отсутствовать тренд. Нормальность распределения остатков может быть установлена одним из критериев согласия. Гипотеза об отсутствии тренда проверяется одним из критериев тренда и случайности. На практике для такого рода проверок можно использовать гистограммы или критерий Фишера (F).

В данном примере (табл. 125) эти остатки вызваны, скорее всего, явно недостаточным числом наблюдений, которое в 5-6 раз должно превышать число анализируемых факторов, и примененным методом регистрации артериального давления.

Параметры выборочных уравнений регрессии, как и любые выборочные статистические характеристики, оцениваются в определенных интервалах. В том случае, если уравнение регрессии имеет вид $y=ax+b$, выборочные значения коэффициентов a и b являются оценкой соответствующих генеральных коэффициентов и отличаются от них в среднем на величину соответствующих им ошибок. «Ручные» вычисления этих ошибок достаточно трудоемки, поэтому методика их получения в полуавтоматическом режиме будет рассмотрена в разделе «Показатели статистической связи в MS Excel».

Рассмотренные выше параметры регрессии касаются так называемой двухмерной зависимости. Т.е. рассматривались варианты, при которых линейно (рис. 101) или нелинейно (рис. 100) взаимодействуют два признака – зависимый (результативный) и независимый (факторный). В объективной реальности чаще приходится сталкиваться с многофакторными системами взаимодействий, а двухмерные зависимости рассматриваются либо как фрагменты многофакторных систем, либо как некие условные допущения, когда другие независимые факторы этих систем преднамеренно игнорируются или просто неизвестны.

Соответственно, если рассматривается число независимых факторов больше 2, то расчеты проводятся с учетом трехмерного, четырехмерного и т. п.

пространства распределения. С математической точки зрения, число пространственных распределений, в принципе, не ограничено.

Уравнение множественной регрессии, т.е. уравнение, описывающее влияние на зависимую переменную нескольких независимых переменных (регрессоров) таково:

$$Y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$$
, где x_1, x_2, \dots, x_n — независимые переменные a или b , a_1, a_2, \dots, a_n и b — коэффициенты регрессии.

Например: должные величины показателей жизненной емкости легких (ДЖЕЛ) вычисляются для мужчин в возрасте 18-25 лет по уравнению 3 мерной регрессии (Р.Ф. Клемент и др., 1988): ДЖЕЛ= $b_0 + b_1x_1 + b_2x_2$, где b_0 — константа, равная 6,908; b_1 — коэффициент по росту (x_1), равный 5,8; b_2 — коэффициент по возрасту (x_2) = 0,085. С помощью этого уравнения, опираясь на фактические данные о конкретном человеке, путем несложных вычислений можно определить должную величину ЖЕЛ этого человека. Так, для мужчины в возрасте 19 лет, имеющему рост 1,8 метра, ДЖЕЛ= $-6,908 + 5,8 \cdot 1,8 + 0,085 \cdot 19 = 5,2$. С точки зрения клинической практики, снижение фактической ЖЕЛ по сравнению с ДЖЕЛ может говорить о рестриктивных нарушениях вентиляционной способности легких, являющихся следствием нарушения процесса расправления легких при вдохе.

Относительная простота практического 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 параметров уравнений регрессии требует большой и кропотливой работы.

Одной из причин, снижающих точность регрессионного анализа, по аналогии с коэффициентами корреляции, является несоответствие теоретического распределения, взятого за основу расчетов, и фактического распределения исходных данных. Например, линия регрессии может представлять собой не прямую, а какую-либо кривую. Соответственно, форма

уравнения регрессии должна соответствовать реально существующей криволинейной зависимости. Как видно из примера (рис. 102), аппроксимация распределения одного и того же ряда исходных данных с помощью различных функций приводит к разным результатам.

При аппроксимации исходного распределения с помощью уравнения линейной зависимости (1) $Y = -0,316x + 16,996$ коэффициент R^2 , описывающий, по существу, качество регрессии, самый низкий (0,0456).

При аппроксимации того же распределения с помощью полинома второй степени (2) $Y = -0,2892x^2 + 4,3113x + 3,8838$ коэффициент качества регрессии выше более чем на порядок ($R^2 = 0,6073$).

При аппроксимации с помощью полинома третьей степени (3) $Y = -0,0521x^3 - 0,15387x^2 + 12,568x - 8,8611$ коэффициент качества регрессии самый высокий среди перечисленных ($R^2 = 0,860$).

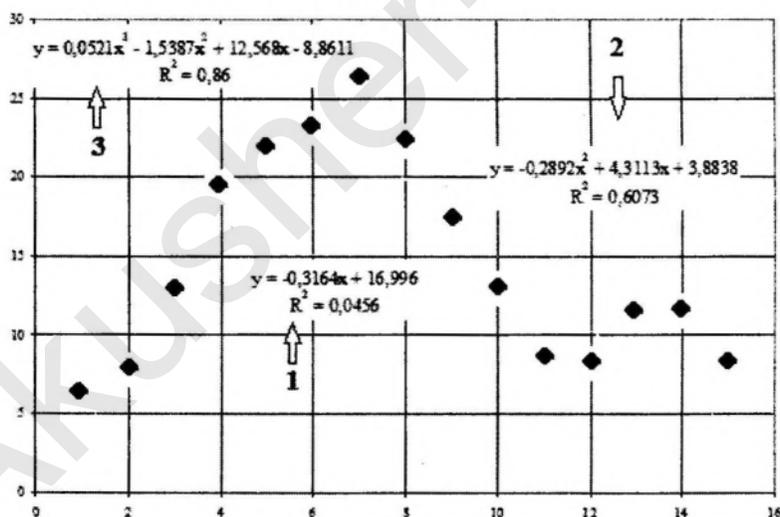


Рис.102. Сравнительная аппроксимация ряда данных с помощью уравнений регрессии: линейной (1), полиномов второй (2) и третьей (3) степеней

Таким образом, становится понятным, почему для получения точной интерпретации регрессии необходим предварительный анализ двумерных диаграмм парных распределений исходных данных.

Однако на практике это условие часто игнорируется, поскольку при малых вариациях исходных значений в исследуемых рядах отдельные параметры регрессии зачастую мало зависят от характера распределения.

В математической статистике в ситуации, когда существует необходимость точного анализа нелинейных связей, применяется так называемая *линеаризация*. Это весьма сложный и спорный метод, позволяющий заменять нелинейные системы эквивалентными линейными системами. Однако указанная эквивалентность сохраняется лишь для ограниченных в пространстве или времени систем и обычно позволяет выяснить лишь некоторые параметры взаимодействия. Считается, что методы статистической линеаризации могут дать вполне удовлетворительную точность, если:

- анализируемая нелинейная система близка к линейной, что и дает возможность с достаточной степенью точности заменять нелинейные преобразования приближенными линейными;
- известны параметры закона распределения вероятностей изучаемых нелинейных процессов, что позволяет определить по статистическим характеристикам линейное преобразование, в достаточной степени эквивалентное нелинейному.

Среди многочисленных условий проведения регрессионного анализа выделяются еще два, без которых вообще начинать этот анализ не следует.

Во-первых, обязательно наличие корреляционной связи между факторными (независимыми) и зависимым (результативным) признаками. Если $R^2 > 0,5$, то обычно считается, что данная регрессионная модель достаточна.

Во-вторых, должна отсутствовать сильная корреляционная ($R^2 \geq 0,7$) связь между отдельными независимыми признаками (регрессорами).

Понятно, что невыполнение первого условия делает бессмысленным весь регрессионный анализ. Невыполнение второго создает специфичную для множественной регрессии проблему «непредвиденного» взаимодействия отдельных анализируемых факторов или их некоторых групп, так называемую

«коллинеарность или мультиколлинеарность», существенно искажающую конечный результат анализа.

Практически все перечисленные проблемы адекватности исходных данных условиям применения регрессионного анализа могут решаться путем исключения «неподходящих» независимых переменных из регрессионной модели (см. раздел «Показатели статистической связи в MS Excel»). В ситуации, когда простое механическое исключение таких переменных невозможно или спорно, используются специальные сложные методы, описание которых не входит в задачи данного издания.

Говоря об интерпретации результатов регрессионного анализа, нельзя еще раз не упомянуть о том, что он позволяет обнаружить только формализованные, числовые зависимости, а не объективные связи.

11.7. Оценки взаимосвязи качественных признаков на принципе взаимной сопряженности

В основу статистической оценки взаимосвязи качественных признаков положен принцип взаимной сопряженности, который реализуется путем анализа взаимного распределения частот качественных характеристик (переменных значений характеристик, *переменных*) наблюдаемых вариантов.

В наиболее общем виде такой подход означает отказ от использования при вычислении различных критериев взаимосвязи числовых значений отдельных единиц наблюдения (вариант), а использование только частот встречаемости (частот) той или иной переменной. Это позволяет анализировать взаимосвязь любых единиц наблюдения вне зависимости от их однородности, качественного или количественного единства. Существенным недостатком, органически присущим такому подходу, является рост величины ошибок второго рода, т.е. вероятности ложноотрицательных результатов.

При вычислении показателей взаимосвязи на основе принципа сопряженности исходные данные обычно представляются в виде групповых таблиц сопряженности.

Таблица сопряженности – универсальное средство отображения распределения двух зависимых переменных с целью исследования связи между ними. Строки таблицы сопряженности представляют распределение одной качественной характеристики (переменной), столбцы – распределение другой переменной. На пересечении строки и столбца указывается частота совместного появления f_{ij} – единиц наблюдения с соответствующими значениями двух характеристик (переменных). Сумма частот по строке f_i называется маргинальной частотой строки; сумма частот по столбцу f_j – маргинальной частотой столбца. Сумма маргинальных частот равна общему объему наблюдений n .

Например: в обследованной совокупности работников одного из предприятий число наблюдений (n) составило 238 человек (табл.127). Из них, имели профессию станочник – 161 человек (частота одной из значений переменной «профессия»), прочие профессии – 122 человека (частота другого значения той же переменной). Общее число людей по строкам – есть маргинальные частоты встречаемости той или иной профессиональной группы f_i . Среди всех обследованных болели – 108 человек (частота одного из значений переменной – «случаи заболеваний в году»), не болели – 175 человек (частота другого значения переменной – «случаи заболеваний в году»). Общее число людей по столбцам – есть маргинальные частоты встречаемости болевших и неболевших работников f_j .

В последние годы в литературе встречаются различные расширительные понимания таблицы сопряженности. Предполагается, что в качестве ее элементов могут фигурировать не только частоты, но и многие другие числовые данные: скажем, в клетках половозрастной таблицы могут стоять средние значения зарплаты тех людей, которые характеризуются отвечающим клетке значениям пола и возраста. О том же расширительном понимании таблицы сопряженности говорится в описании известного пакета статистической обработки данных *SPSS*.

Соответственно этому утверждению, в клетки таблицы сопряженности могут быть помещены мода, медиана, дисперсии, величины квадратов

отклонений от средних, разница между эмпирической и теоретической частотами и т.д. (Ростовцев и др., 1997). Однако при таком подходе теряется всякий смысл существования категории *таблица сопряженности*, поскольку практически полностью исчезают все ее отличия от тривиальной групповой статистической таблицы. А это, в свою очередь, создает предпосылки возникновения серьезной методологической ошибки – недоучета принципиальных различий методик статистической обработки частот, рядов производных величин и сгруппированных данных измерений в разнородных числовых шкалах.

Визуально с помощью таблицы сопряженности можно видеть взаимосвязь распределения частот, однако количественную характеристику связи могут дать только коэффициенты сопряженности. Выбор того или иного коэффициента диктуется характером исходных данных и необходимой точностью анализа. Общими достоинствами коэффициентов сопряженности являются:

- относительная простота вычисления;
- возможность с их помощью обрабатывать не только качественные, но и количественные признаки. Для этого количественным признакам с помощью группировки придают характер качественных.

При использовании таблиц сопряженности 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 оценки, формально свидетельствующей о направлении связи, зависит в этой ситуации только от произвольного расположения данных в таблице. Но сам факт наличия или отсутствия связи в целом, бесспорно, устанавливается на основе абсолютного значения соответствующих коэффициентов взаимной сопряженности.

Другой особенностью коэффициентов сопряженности является их заниженные значения в случае резко асимметричного распределения данных в клетках таблицы, а также несимметричных таблиц. Поэтому рекомендуется

использовать параллельно несколько разных (но одномерных!) коэффициентов сопряженности, соответствующих исходным данным. Сообразно уверенности в точности исходных данных из них выбирается либо коэффициент, имеющий наибольшее числовое выражение, либо наименьшее.

11.7.1. Коэффициенты ассоциации и контингенции

На основе распределения исходных данных в четырехпольных групповых таблицах сопряженности, состоящих всего из двух строк и двух столбцов (без итогов, 2×2), с помощью относительно простых при вычислении коэффициентов *ассоциации* или *контингенции* можно дать некоторую количественную оценку взаимосвязи двух признаков, распределение которых представлено в виде абсолютных или относительных частот (табл. 126).

Таблица 126

Макет четырехпольной таблицы взаимной сопряженности (2×2)

Признаки	<i>A</i>	не <i>A</i>	—
<i>B</i>	<i>a</i>	<i>b</i>	<i>a+b</i>
не <i>B</i>	<i>c</i>	<i>d</i>	<i>c+d</i>
—	<i>a+c</i>	<i>b+d</i>	$N=a+b+c+d$

Например: требуется доказать, существует ли связь между профессией и заболеваемостью (табл. 127).

Таблица 127

Распределение рабочих по случаям заболеваний в году

Профессия	Болели	Не болели	Итого
Станочник	78	83	161
Прочие	30	92	122
Всего	108	175	283

Коэффициент ассоциации Q (коэффициент Юла) равен:

$$Q = \frac{ad - cb}{ad + cb} = \frac{78 * 92 - 30 * 83}{78 * 92 + 30 * 83} \approx 0,48$$

Связь считается установленной, если коэффициент ассоциации по абсолютному значению выше 0,5. В приведенном примере статистическая достоверность связи вызывает сомнение.

Для более точной оценки статистической достоверности коэффициента ассоциации используется значение его среднеквадратического отклонения, равному корню квадратному из его дисперсии, который несложно вычислить по формуле:

$$\sqrt{D_q} = \sqrt{\frac{1}{4}(1 - Q^2) \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)} = \sqrt{\frac{1}{4}(1 - 0,48) \left(\frac{1}{78} + \frac{1}{92} + \frac{1}{30} + \frac{1}{92} \right)} = 0,115$$

Поскольку значение коэффициента ассоциации превышает свое среднеквадратическое отклонение более чем в 4 раза, то выявленную связь следует признать статистически достоверной.

Особенностью коэффициента ассоциации является односторонняя оценка взаимосвязи, поэтому числовые значения этого коэффициента всегда ниже, чем значения других коэффициентов сопряженности, дающих двухсторонние оценки. К таким двухсторонним коэффициентам для таблиц сопряженности 2x2 относится коэффициент контингенции (сходства) Φ :

$$\Phi = \frac{ad - cb}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$

$$\Phi = \frac{78 * 92 - 30 * 83}{\sqrt{(78 + 83)(30 + 92)(78 + 30)(83 + 92)}} = 0,24$$

Для приближенной оценки достоверности результата можно воспользоваться и правилом, согласно которому коэффициент Φ считается статистически достоверным, если его численные значения $\geq 0,3$.

Точную оценку достоверности коэффициента Φ можно проводить с помощью критерия согласия Пирсона χ^2 (хи-квадрат), учитывая степень свободы, равной 1, а n , равным $a+b+c+d$. Поскольку исходные данные могут быть представлены неабсолютными числами, то для расчета значения χ^2 (хи-квадрат) допускается использование формулы:

$$\chi^2 = \frac{n(ad - cb - \frac{n}{2})^2}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{263 \cdot (76 \cdot 92 - 30 \cdot 83 - \frac{263}{2})^2}{\sqrt{(76+83) \cdot (30+92) \cdot (83+92) \cdot (76+30)}} = 3,97.$$

Так как критическое значение $\chi^2=3,84$ при $P=0,05$ и степени свободы =1, то связь считается статистически достоверной.

11.7.2. Коэффициенты сопряженности

Коэффициенты Фишера-Филча (ϕ), Крамера (V), Пирсона (C) и Чупрова (K), величины которых всегда находятся в пределах от 0 до 1, вычисляются на основе различных вариантов соотношения величин критерия согласия Пирсона χ^2 (хи-квадрат) и числа наблюдения. Математическая логика применения перечисленных коэффициентов базируется на том, что если теоретическое распределение говорит о независимости (несвязанности) признаков, то отклонение фактического распределения от теоретического будет тем больше, чем больше признаки взаимосвязаны. Это отклонение оценивается с помощью критерия χ^2 хи-квадрат, который включен в формулы в качестве основного элемента. Сам по себе критерий хи-квадрат может свидетельствовать о наличии или отсутствии взаимосвязи, однако его значение во многом зависит от числа наблюдений. Таким образом, значения χ^2 напрямую не могут использоваться для сравнения статистических связей в совокупности с разным числом наблюдений. Этот недостаток и устраняется при использовании помимо численных значений χ^2 и числа наблюдений. Такие общие для всех перечисленных коэффициентов особенности определяют их общие достоинства и недостатки.

Существенные достоинства:

- относительная простота получения оценок связи в таблицах сопряженности, где число строк и столбцов выходит за пределы 2×2 (за исключением критерия Фишера-Филча), в том числе и в несимметричных таблицах;
- упрощение процедуры оценок статистической достоверности значений этих критериев, так как величина χ^2 , на основе которой эта

достоверность и оценивается, используется по ходу, как промежуточный результат вычислений. Т.е. не требуется специальных дополнительных расчетов для процедуры оценки репрезентативности полученных значений этих критериев.

Недостатки перечисленных коэффициентов сопряженности:

- ограничение области их применения только таблицами, где данные представлены абсолютными числами (абсолютными частотами), и недопустимость для таблиц с производными величинами (относительные частоты, средние и т.п.);
- недопустимость применения в случаях, когда число наблюдений в отдельных клетках таблицы сопряженности менее 5. В этих случаях обычно производят объединение соседних столбцов или клеток, однако этим самым снижается общая точность критерия χ^2 ;
- общее число наблюдений в таблице сопряженности не должно быть меньше 50, а оптимальным считается 100 и более. При этом рекомендуется применять как можно более дробные группировки.

Эти условия, диктуемые особенностями критерия χ^2 , нередко забываются исследователями.

Расчет каждого из коэффициентов: Фишера-Филча (φ), Крамера (V), Пирсона (C) и Чупрова (K) строится на основе значения χ^2 , которое

определяется по классической формуле: $\chi^2 = \frac{(P_i - P_i^*)^2}{P_i^*}$, где P_i – фактические частоты, а P_i^* – ожидаемые частоты. Чтобы избежать вычисления ожидаемых частот, можно использовать более практичный вариант:

$$\chi^2 = n \left(\sum_{i=1}^r \sum_{j=1}^c \frac{f_{ij}^2}{f_j f_i} - 1 \right)$$

где f – частота в клетке таблицы, i – номер строки, j – номер столбца,
 n – число наблюдений.

Для данных, приведенных в табл. 127:

$$\chi^2 = 283 \cdot \left(\frac{78^2}{108 \cdot 161} + \frac{63^2}{175 \cdot 161} + \frac{30^2}{108 \cdot 122} + \frac{92^2}{175 \cdot 122} - 1 \right) = 16,74$$

Поскольку число степеней свободы принимается равным 1 и величина вычисленного значения χ^2 выше $\chi_{\text{кр}}^2 = 10,83$ при $P < 0,01$ (см. Приложение 8), то принимается заключение о высокой статистической достоверности результата. Соответственно, все приведенные ниже коэффициенты сопряженности тоже принимаются как статистически значимые.

Критерий Фишера-Филча (φ) может применяться только для таблиц 2x2:

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{16,74}{283}} = 0,244$$

Критерий Крамера (V), представляющий модификацию критерия Фишера для любых таблиц сопряженности:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{16,74}{283(2-1)}} = 0,244$$

где k – наименьшее из количеств строк и столбцов (табл. 127).

Критерий сопряженности Пирсона (C), пригодный для любых таблиц сопряженности:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{16,74}{16,74 + 283}} = 0,236$$

Критерий сопряженности Чупрова (K) дает более осторожные результаты, когда таблица сопряженности имеет квадратную форму, а число строк и столбцов не превышает 5.

$$K = \sqrt{\frac{\chi^2}{n\sqrt{(s-1)(t-1)}}} = \sqrt{\frac{16,74}{283\sqrt{(2-1)(2-1)}}} = 0,243$$

где s – число строк, t – число столбцов.

Несмотря на функциональное и алгебраическое сходство перечисленных коэффициентов сопряженности, особенно в случае оценки таблицы 2×2 , существуют определенные особенности в интерпретации результатов, полученных с помощью коэффициента C . Формально предел значений коэффициента сопряженности C так же, как и в других приведенных выше коэффициентах, равен 1,0. К этому значению они стремятся при увеличении силы связи между признаками. Но коэффициент C 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 (C_{max}) позволяет получить его нормированное значение, которое является более точной характеристикой для каждого конкретного случая.

Для симметричной таблицы, где число строк и число столбцов равно между собой:

$$C_{norm} = \frac{C}{C_{max}}, \text{ а } C_{max} = \sqrt{\frac{t-1}{t}},$$

где t – число строк. В случае таблицы 2×2 , в приведенном примере:

$$C_{max} = \sqrt{\frac{2-1}{2}} \approx 0,71, \text{ а нормированное значение } C_{norm} = \frac{0,236}{0,71} = 0,34.$$

Несмотря на то, что разница между нормированным и ненормированным значениями коэффициента сопряженности в данном случае оказалась незначительной, численное значение C_{norm} заметно повышает убедительность наличия связи.

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. Кендалл (1960) считали, что в таблицах 5×5 необходимость в корректировке вообще отпадает. Однако, учитывая, что на

практике разбиение качественных признаков на такое количество диапазонов встречается достаточно редко из-за традиционно малого числа наблюдений, в клинических исследованиях применение нормированных значений оценок сопряженности представляется весьма целесообразным.

Эта же закономерность сохраняется и в случае использования таблицы сопряженности 3x3 (табл. 128).

Таблица 128

Заболееваемость работников различных профессиональных групп

Проф. группы	Кратность заболеваний за год			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

Для данных, приведенных в табл. 128:

$$\chi^2 = 212 \cdot \left(\frac{11^2}{100 \cdot 46} + \frac{12^2}{39 \cdot 46} + \frac{23^2}{73 \cdot 46} + \frac{17^2}{100 \cdot 57} + \frac{16^2}{39 \cdot 57} + \frac{24^2}{73 \cdot 57} + \frac{72^2}{100 \cdot 109} + \frac{11^2}{39 \cdot 109} + \frac{26^2}{73 \cdot 109} - 1 \right) = 33,4$$

Критерий Крамера:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{33,4}{212(3-1)}} = 0,33$$

Критерий сопряженности Пирсона:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{33,4}{33,4 + 212}} = 0,37$$

для которого $C_{max} = \sqrt{\frac{3-1}{3}} \approx 0,81$ и, соответственно, $C_{норос} = \frac{0,87}{0,82} = 0,45$.

Критерий сопряженности Чупрова

$$K = \sqrt{\frac{33,4}{212\sqrt{(3-1)(3-1)}}} = 0,28$$

11.8. Вычисление показателей статистической связи в MS Excel

11.8.1. Показатели парной корреляции и параметры регрессии

Расчет параметрических показателей линейной корреляции и регрессии – тесно взаимосвязанные задачи, их решение с помощью MS Excel может проводиться несколькими способами. Главным условием выбора того или иного способа является форма связи (линейная или нелинейная), поэтому первым шагом, обязательно предвещающим всю последовательность дальнейших операций корреляционного анализа, является проверка линейности взаимосвязи. Эта задача относительно несложно может быть решена с помощью пакета MS Excel «Гистограмма» (рис. 103), путем визуального анализа распределения исходных данных на диаграмме «Точечная».

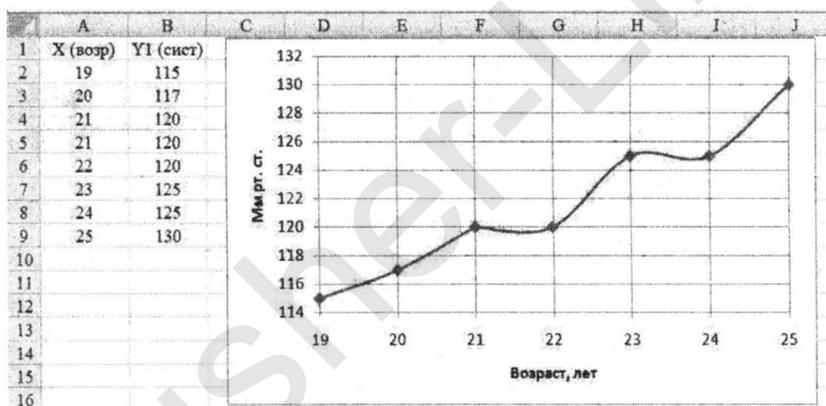


Рис.103. Линейность распределения коррелирующих переменных

Если линейность взаимосвязи не вызывает сомнения, то наиболее простым способом получения коэффициента парной линейной корреляции Пирсона является использование из меню «Формулы» пакета «Мастер функций»⁸³. В открывающейся закладке «Статистические» (мастер функций шаг 1 из 2) можно выбрать функцию «КОРЕЛ» (рис. 104).

⁸³ Мастер функций можно вызвать и через указатель командной строки 

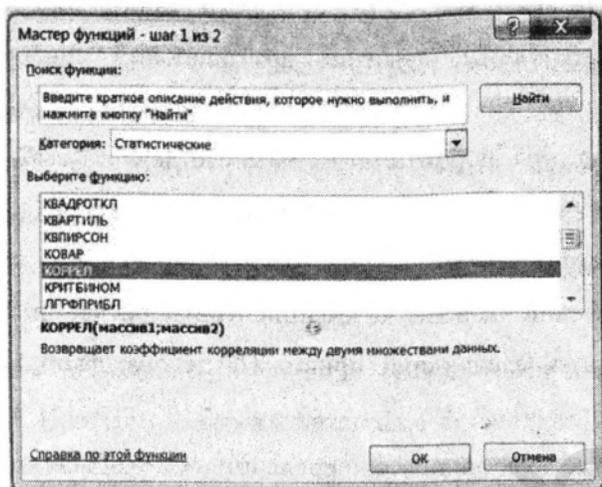


Рис.104. Окно мастера функции

После активизации щелчком правой клавиши мыши функции «КОРЕЛ» появится закладка «Аргументы функции». В окна **Массив1** и **Массив2** нужно ввести адреса массивов исходных данных без заголовков по ряду X и ряду Y (рис. 105).

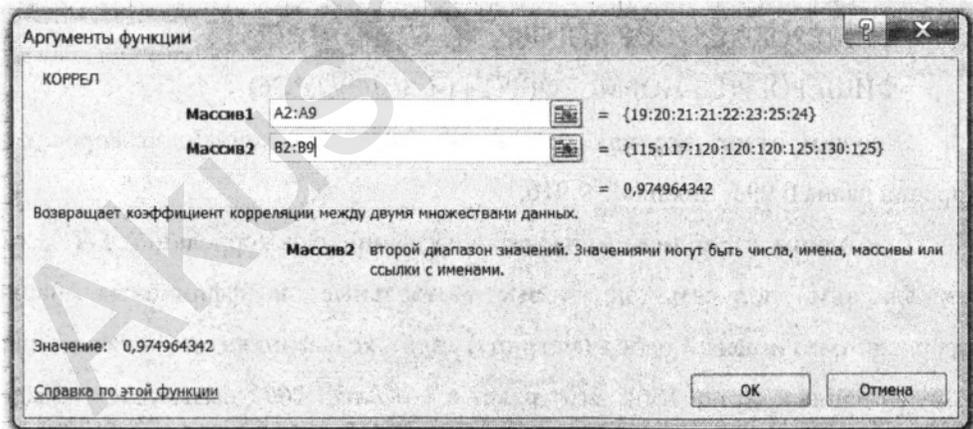


Рис.105. Аргументы функции «КОРЕЛ»

После ввода исходных данных и активизации виртуальной клавиши <ОК> появится результат: значение коэффициента линейной корреляции Пирсона $\approx 0,975$.

Как и результаты любого статистического анализа выборочных данных, коэффициент корреляции должен оцениваться с учетом значений ошибок репрезентативности. Несмотря на то, что готовой функции прямого определения значений ошибок коэффициентов корреляции в *MS Excel* нет, можно использовать следующий прием. Последовательно введем в ячейки исходные данные:

C1 – значение коэффициента корреляции и C2 – объем выборки;

C3 – доверительный коэффициент (например, 95).

Затем формулы промежуточных расчетов в ячейки;

C5 – формулу, вычисляющую преобразование Фишера для коэффициента корреляции =ФИШЕР(C1);

C6 – формулу для подсчета стандартной ошибки: =1/КОРЕНЬ(C2-3).

В ячейки C7 и C8 введем формулы определения верхней и нижней границ доверительного интервала для коэффициента корреляции:

=ФИШЕРОБР(C5+НОРМСТОБР(1-(1-C3/100)/2)*C6);

=ФИШЕРОБР(C5-НОРМСТОБР(1-(1-C3/100)/2)*C6).

Получим ответ: верхняя граница коэффициента линейной корреляции Пирсона равна 0,993, нижняя – 0,916.

Еще одним инструментом расчета коэффициентов корреляции в *MS Excel*, позволяющим получать не только отдельные коэффициенты парной корреляции, но и целый набор (матрицу) таких коэффициентов, является пакет «Анализ данных» (рис. 106). Этот пакет в *MS Excel 2007* вызывается из меню «Данные». В более ранних версиях *MS Excel* – из меню «Сервис».

Указав инструмент анализа <Корреляция> (Рис.) и нажав <ОК>, вводим исходную информацию для расчета коэффициентов корреляции (рис. 107):

- блок ячеек с входными данными (входной интервал). При этом в блок могут включаться несколько числовых данных нескольких факторных признаков (независимых переменных);
- указываем ориентацию исходных данных (группирование) по строкам или столбцам;
- если есть заголовки столбцов, то активизируем «Метки в первой строке»;
- указываем «выходной интервал» в разделе «Параметры вывода» расчетных данных. При этом выходной интервал ячеек может быть обозначен только одной левой верхней клеткой блока клеток, где будет размещен конечный результат.

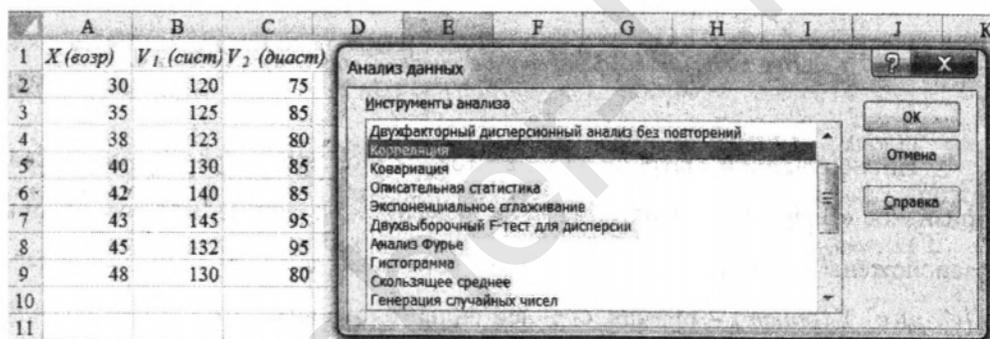


Рис.106. Пакет «Анализ данных». Корреляция

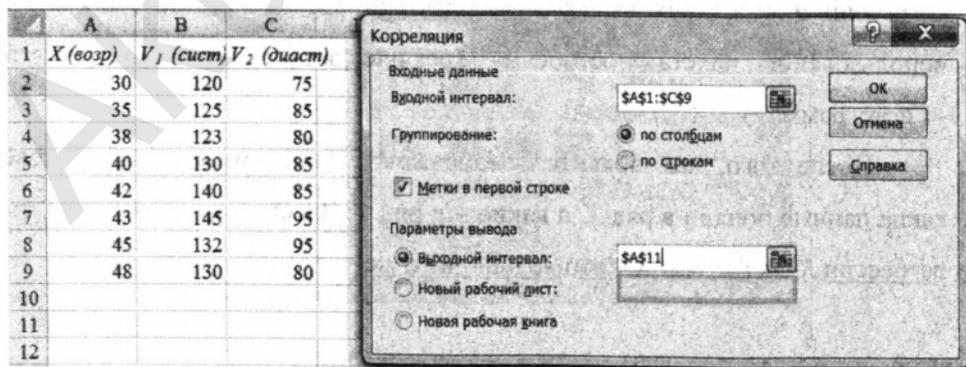


Рис.107. Ввод параметров линейной корреляции Спирмена

После активизации <ОК> появляется результат вычислений в виде матрицы коэффициентов линейной корреляции (рис. 108).

	A	B	C	D
1	<i>X (возр)</i>	<i>V₁ (сист)</i>	<i>V₂ (диаст)</i>	
2	30	120	75	
3	35	125	85	
4	38	123	80	
5	40	130	85	
6	42	140	85	
7	43	145	95	
8	45	132	95	
9	48	130	80	
10				
11		<i>X (возр)</i>	<i>V₁ (сист)</i>	<i>V₂ (диаст)</i>
12	<i>X (возр)</i>	1		
13	<i>V₁ (сист)</i>	0,621	1	
14	<i>V₂ (диаст)</i>	0,527	0,718	1

Рис.108. Матрица коэффициентов линейной корреляции Спирмена

На полученной матрице результатов видно, что корреляционная связь признаков «Возраст» и «Систолическое давление», числовые значения которых расположены в рядах $X(\text{возр})$ и $Y_1(\text{сист})$ прямая, средней силы. Признаков $X(\text{возр})$ и $Y_2(\text{диаст})$ – прямая, средней силы (плотности). Признаков $X(\text{возр})$ и $Y_2(\text{сист})$ – прямая, сильная (плотная).

Одним из самых практичных способов определения параметров регрессии и через них вычисление коэффициента корреляции в *MS Excel* – использование пакета построения графических изображений (<Вставка>, <Гистограмма>)⁸⁴.

После того, как появится изображение диаграммы, нужно проверить, какие данные попали в ряд X , а какие – в ряд Y . Здесь важно внимание, так как регрессии X от Y и Y от X – это не одно и то же!

⁸⁴ Если построение диаграммы производится в *MS Excel 2007*, то предварительно рекомендуется выделить на листе с помощью мышки область исходных данных. В более ранних версиях *MS Excel* область данных задается при дальнейшем проходе мастера диаграмм.

В окне мастера диаграмм необходимо выбрать точечный тип диаграммы и ее вид без соединений точек (рис. 109). Затем нажать [OK].

С помощью редактора диаграмм можно откорректировать внешний вид диаграммы (см. раздел «Графические изображения»). В частности, чтобы упростить для этого примера зрительное восприятие информации, нужно:

- убрать лишние, в данном случае оформительские элементы – название диаграммы и легенды. Для этого каждый из этих элементов поочередно: выделить и нажать клавишу компьютера **DELETE**;

- выставить параметры формата горизонтальной оси (абсцисс): <фиксированное>, минимальное значение – 28,0; <фиксированное>, максимальное значение – 52,0; цена основных делений – 4,0;

- выставить параметры формата вертикальной оси (ординат): <фиксированное>, минимальное значение – 118,0; <фиксированное>, максимальное значение – 145,0; цена основных делений – 3,0;

- на диаграмме выделить щелчком левой клавиши мыши ряд полученных точек, а затем нажатием на правую клавишу мыши вызвать список команд. Выбрать из них: <Параметры линии тренда>, «Линейная» (рис. 110).

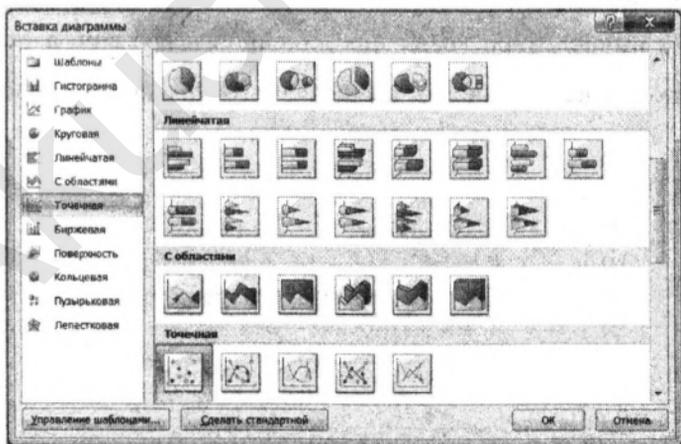


Рис.109. Выбор диаграммы в MS Excel 2007

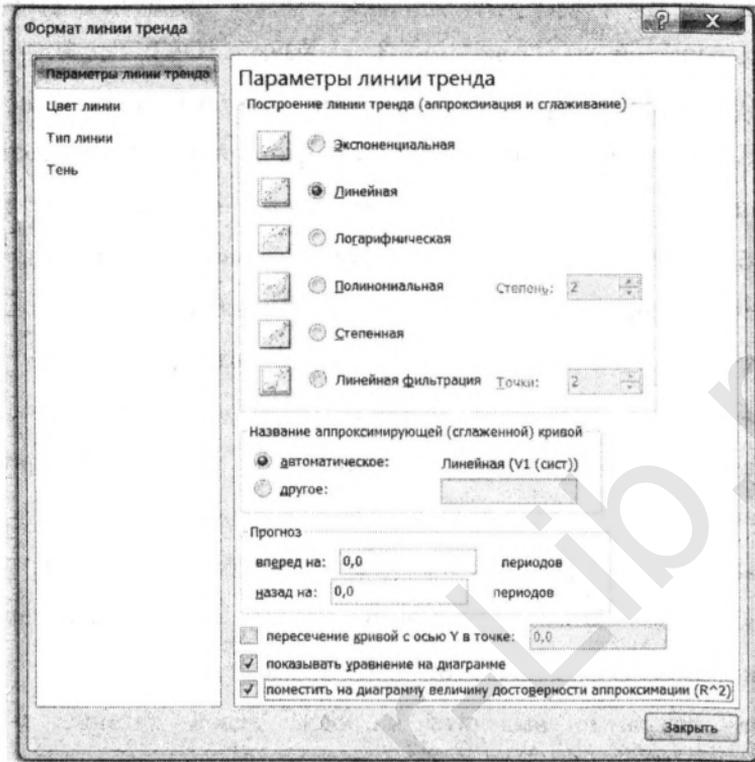


Рис.110. Установки параметров линии тренда в MS Excel 2007

На диаграмме появится аппроксимирующая линия регрессии, автоматически вычисленное уравнение линейной регрессии, а также значение R^2 , которое показывает точность вычисленного уравнения регрессии Y по X (рис.111).

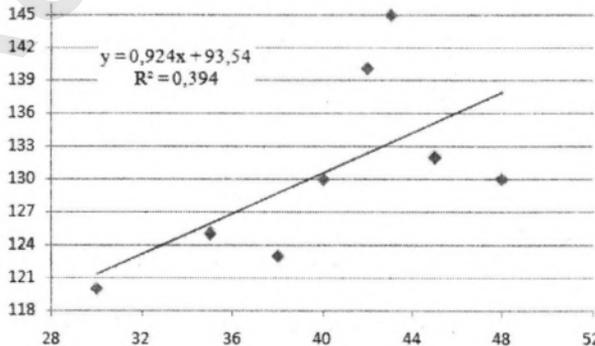


Рис. 111. Диаграмма с параметрами линейной регрессии (линия и уравнение регрессии, величина линейной аппроксимации R^2)

Максимально возможное значение $R^2=1,0$ или 100%. В данном примере в случае использования аппроксимации (сглаживания) с помощью линейной регрессии $R^2=0,394$ или 39,4% (см. разделы «Коэффициент линейной корреляции», «Коэффициент детерминации»). С помощью этого коэффициента, поскольку в случае линейной аппроксимации R^2 – есть квадрат коэффициента линейной корреляции, можно легко найти коэффициент корреляции: $R=\sqrt{0,394}=0,63$. Так как угол наклона линии регрессии соответствует прямой связи, то и коэффициент корреляции положительный.

Ошибка уравнения регрессии, т.е. величины Y , вычисляемой уравнением регрессии, можно определить с помощью функции **СТОШХУ**.

С помощью мастера диаграмм можно производить прогнозирование значений Y . Для этого при выборе параметров тренда нужно:

- в окне <Формат линии тренда> активизировать опцию «Прогноз»;
- установить величину X (единицу интервала ряда X вперед или назад).

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. В *Excel* можно путем подбора выбрать параметры аппроксимации, более точно характеризующие ту или иную взаимосвязь. Например, при анализе использованных выше исходных данных полином 3-й степени дает $R^2 = 0,711$ или 71,1%. Тогда нелинейная корреляция, точнее – корреляционное отношение, такова: $r=\sqrt{0,711} = 0,843$ (рис. 112).

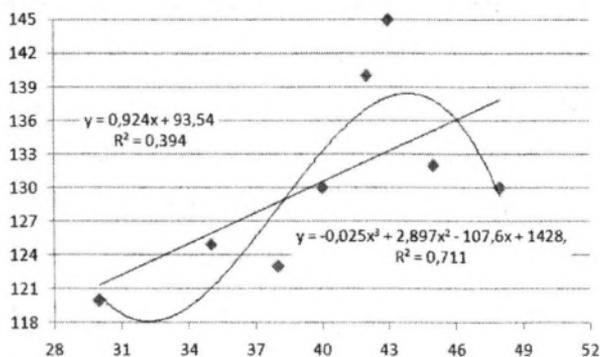


Рис.112. Диаграмма сравнительных параметров линейной и нелинейной регрессий (линии и уравнения регрессии, величины аппроксимации R^2)

Более глубокий регрессионный анализ позволяет проводить пакет *MS Excel* <Анализ данных>, обеспечивающий возможность всесторонней оценки достоверности полученных результатов. Для выполнения расчетов с помощью пакета <Анализ данных> в *MS Excel 2007*:

Вызовите команду <Анализ данных> из меню <Данные>. Должно появиться диалоговое окно (рис. 113).

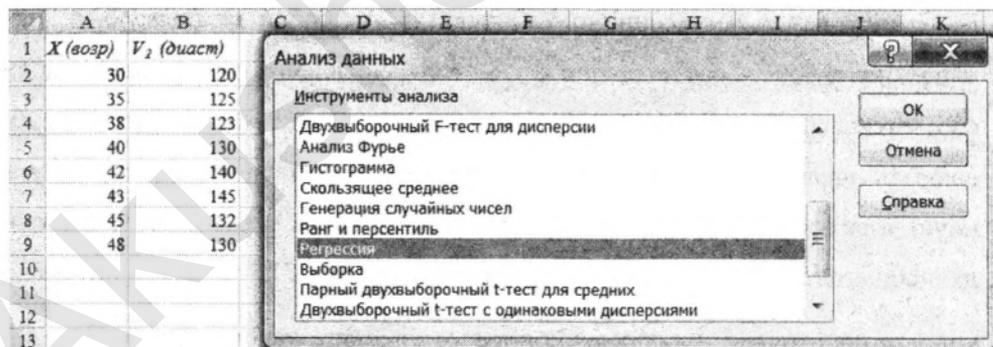


Рис.113. Пакет «Анализ данных». Регрессия

Выберите метод «Регрессия» и нажмите [ОК]. В появившемся окне «Регрессия» введите входной интервал Y, т.е. ссылку на диапазон зависимых данных: **\$B\$2:\$B\$10**.

Примечание: Этот диапазон должен состоять из одного столбца.

Затем в разделе <Входные данные> введите входной интервал X (регрессора), т.е. диапазон независимых данных, подлежащих анализу: $\$A\$2:\$A\10 .

Примечание: Независимые переменные могут располагаться в одном или нескольких столбцах, расположенных слева направо.

Флажок <Метки> устанавливается, если первая строка входного интервала содержит заголовки.

Флажок <Уровень надежности> 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.

В разделе <Параметры вывода> установите переключатель <Выходной интервал>, чтобы указать место на рабочем листе, куда будут выводиться результаты анализа. Для получения результатов на другом листе установите переключатель параметров вывода в положение <Новый рабочий лист> (рис. 114).

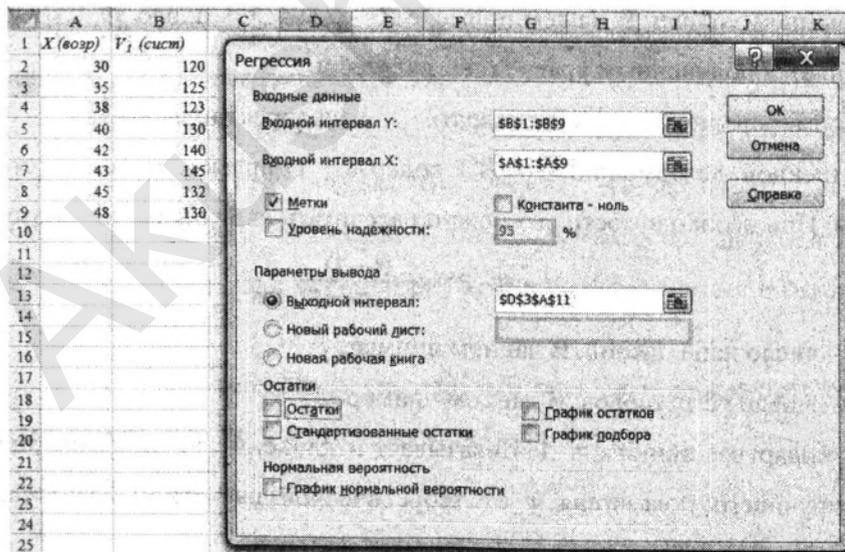


Рис.114. Ввод данных для вычисления основных параметров регрессии

Кроме того, в окне <Регрессия> можно установить следующие параметры:

<График подбора>, чтобы построить диаграммы наблюдаемых и предсказанных (вычисленных по уравнению регрессии) значений Y для каждой независимой переменной X .

<Остатки>, чтобы включить остатки в выходной диапазон.

<Стандартизированные остатки>, чтобы включить стандартизированные остатки в выходной диапазон.

<График остатков>, чтобы построить диаграмму остатков.

<График нормальной вероятности>, чтобы построить диаграмму нормальной вероятности.

После завершения настройки параметров нажмите [ОК].

Результаты регрессионного анализа при вводе перечисленных параметров (табл. 129) состоят из 3-х таблиц. Первая таблица содержит значение корреляции и коэффициента детерминации (квадрата коэффициента корреляции – R^2), являющегося показателем качества подобранной регрессионной модели. В нашем примере $R^2 \approx 0,394$, т.е. около 39,4% вариации Y объясняется полученным уравнением регрессии.

Нормированный «R-квадрат» (скорректированный индекс множественной детерминации – R^2) содержит поправку на число степеней свободы. При необходимости его можно рассчитать «вручную» по формуле:

$$1 - (1 - R^2) \cdot \frac{(n - 1)}{(n - m - 1)}, \text{ где:}$$

n – число наблюдений. В данном примере 8.

m – число регрессоров. В данном примере 1.

Стандартная ошибка = 7,1. Показывает отклонение фактических значений результирующего показателя Y от теоретической расчетной величины, или (грубо говоря) какой разброс данных присущ выборке.

Вторая таблица содержит результаты дисперсионного анализа, с помощью которого проверяется статистическая значимость вычисленной регрессии. Эта таблица включает следующие параметры:

df – число степеней свободы,

SS – сумма квадратов,

MS – средний квадрат (дисперсия),

F -статистика Фишера – фактическое значение критерия Фишера при данной дисперсии и ее степени свободы.

Значимость F – значимость критерия Фишера.

В данном случае основной результат дисперсионного анализа состоит в том, что уравнение линейной регрессии является статистически незначимым, т.к. полученная значимость $P=0,095$ больше $0,05$. Таким образом, принимается гипотеза о незначимости (статистической недостоверности) уравнения регрессии.

В третьей таблице представлены результаты вычисления коэффициентов уравнения линейной регрессии $Y=aX+b$: коэффициент $b=93,54$ и коэффициент $a=0,924$. Таким образом, уравнение регрессии имеет вид $Y=0,924X+93,54$.

Остальные результаты позволяют проверить значимость полученных коэффициентов a и b уравнения регрессии.

Коэффициент a незначим, т.к. абсолютное значение критерия t (t -статистика) $\approx 2,0$ при значимости $P=0,095$. Коэффициент b значим, т.к. абсолютное значение критерия t (t -статистика) $\approx 4,9$ при значимости $P=0,003$. Т.е. нулевая гипотеза принимается для коэффициента a и отвергается для b .

Соответственно величинам стандартных ошибок принимаются размеры доверительных границ этих коэффициентов (табл. 129).

11.8.2. Параметры многофакторной линейной регрессии

Рассмотренные выше примеры касались так называемой двухмерной зависимости, когда рассматривается взаимодействие двух факторов (признаков) – зависимого (результативного) и независимого (факторного, регрессора). В реальной ситуации чаще приходится сталкиваться с многомерными зависимостями.

В общем виде формула для расчета коэффициента множественной линейной регрессии выглядит так: $y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$, где a, b_1, b_2, \dots, b_n – коэффициенты регрессии (регрессоров x_1, x_2, \dots, x_n). В качестве примера рассмотрим расчет параметров многомерной регрессии комплекса санитарно-гигиенических показателей загрязнения воды и пошаговое построение информативной регрессионной модели.

Справочка. Среди критериев загрязнения водоемов наиболее часто в санитарно-гигиенической практике используют специальные тесты, свидетельствующие о присутствии разных по происхождению и составу загрязнителей. В том числе показатели биологического потребления кислорода (БПК5 и БПК20), наличие взвешенных веществ и т.п. Интегральным показателем для этого ряда тестов служит химическая потребность в кислороде (ХПК), равная числу миллиграммов кислорода, поглощаемого 1 литром пробы воды из горячего подкисленного раствора бихромата калия. ХПК считается одним из наиболее информативных показателей антропогенного загрязнения природных и сточных вод.

Оценка взаимосвязи результатов отдельных тестов загрязнения и интегрального показателя ХПК позволяет:

- установить роль того или иного загрязнителя в общем загрязнении водоема;
- выбрать наиболее эффективную, с точки зрения технических и экономических возможностей, схему улучшения экологического состояния источника водоснабжения, водосброса и т.п.

В табл. 130 приведены условные результаты регрессионной оценки перечисленных тестов.

Таблица 130

Характеристики загрязнения сточных вод

Номер пробы	Окисляемость	БПК5	БПК20	Взвешенные в-ва	ХПК
	x_1	x_2	x_3	x_4	y
1	20,3	1,72	5,10	33,40	85,0
2	21,8	1,76	5,10	33,50	84,9
3	21,4	1,73	5,30	33,40	85,2
4	20,3	1,75	5,20	33,40	85,2
5	20,3	1,75	5,20	33,50	85,2
6	20,5	1,78	5,40	34,50	85,6
7	20,5	1,80	5,20	33,60	85,9
8	20,1	1,80	5,10	34,10	85,6
9	20,2	1,80	5,00	33,40	85,3
10	20,6	1,95	5,10	31,60	85,9
11	21,2	2,00	5,40	40,15	100,4
12	21,5	2,10	4,42	40,10	106,2
13	22,2	2,60	4,50	33,00	106,3
14	22,0	2,90	5,10	33,30	107,0
15	22,0	2,80	5,00	33,10	106,6

Первым шагом, предваряющим непосредственную процедуру многофакторного регрессионного анализа, является расчет корреляционной матрицы.

Для получения этой матрицы с помощью команды <Анализ данных> из меню <Данные> нужно вызвать диалоговое окно «Анализ данных» (рис. 115 и рис. 113). Выбирается метод «Корреляция» и после нажатия кнопки [OK] в появившемся окне «Корреляция» вводится входной интервал данных: \$A\$1; \$F\$16 и начальные координаты выходного интервала данных (\$A\$18).

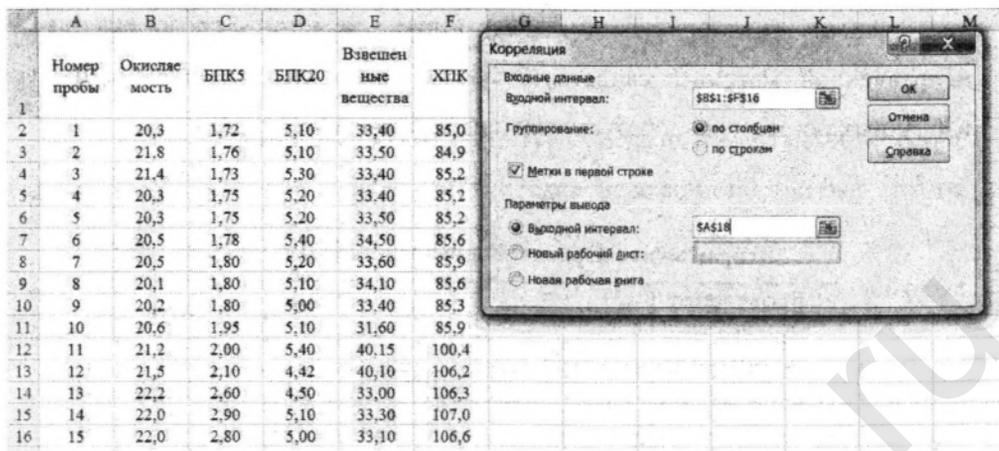


Рис.115. Параметры ввода данных для корреляционной матрицы

После нажатия на кнопку [ОК] на листе Excel в отмеченном месте появится готовая корреляционная матрица (табл.131).

Таблица 131

Корреляционная матрица (к табл.130)

Характеристики загрязнения	Окисляемость	БПК5	БПК20	Взвешенные вещества	ХПК
Окисляемость	1,00				
БПК5	0,758	1,00			
БПК20	-0,433	-0,424	1,00		
Взвешенные вещества	0,128	-0,035	-0,164	1,00	
ХПК	0,777	0,885	-0,568	0,405	1,00

Анализ распределения коэффициентов линейной корреляции, представленных в этой матрице, свидетельствует, что связь фактора «Взвешенные вещества» с результирующим признаком «ХПК» крайне низкая (практически отсутствует). Это обстоятельство делает не только бессмысленным расчет параметров регрессии, описывающих взаимосвязь регрессора «Взвешенные вещества» и зависимой переменной «ХПК», но и способно исказить итоги взаимодействия всего комплекса реально взаимодействующих факторов. Поэтому из массива исходных данных целесообразно исключить столбец «Взвешенные вещества» и повторить без

него заново расчет матрицы. Результаты расчетов, представленные в исправленной матрице, свидетельствуют о сильной линейной связи всех избранных регрессоров и результативного признака (табл. 132).

Таблица 132

Исправленная корреляционная матрица (к табл. 130)

Характеристики загрязнения	Окисляемость	БПК5	БПК20	ХПК
Окисляемость	1,00			
БПК5	0,758	1,00		
БПК20	-0,433	-0,424	1,00	
ХПК	0,777	0,885	-0,568	1,00

Итоги регрессионного анализа при вводе значений регрессоров, ограниченных перечнем, указанным в матрице табл. 132, свидетельствуют:

- показатель качества подобранной регрессионной модели $R^2 \approx 0,972$, т. е. около 97,2% вариации Y объясняется полученным уравнением регрессии;
- стандартная ошибка=2,61.

Основной результат дисперсионного анализа – уравнение линейной регрессии статистически значимо ($F < 0,001$).

Коэффициенты отдельных факторных признаков (регрессоров) позволили определить общий вид уравнения регрессии:

$$Y = 153,3 + 1,39X_1 - 0,9X_2 - 18,2X_3$$

Соответственно величинам стандартных ошибок принимаются размеры доверительных границ этих коэффициентов.

Однако оценка статистической репрезентативности параметров регрессии с помощью дисперсионного анализа (табл. 133) свидетельствует:

- коэффициенты a (Y -пересечение) и коэффициент b_3 (регрессор «БПК20») статистически значимы, т.к. для t -статистики этого критерия $P < 0,01$;
- коэффициенты b_1 и b_2 для регрессоров «окисляемость» и «БПК5» не значимы, т. к. для соответствующих им t $P > 0,05$.

Таким образом, нулевая гипотеза отвергается для коэффициентов a и b_1 , но принимается для b_2 и b_3 .

В этой ситуации, из-за отсутствия статистической достоверности отдельных коэффициентов, возникает высокая вероятность ошибки итогов всего регрессионного анализа. Для того чтобы устранить эту ошибку, необходимо исключить из массива исходных данных регрессоры, роль которых статистически незначима.

После исключения регрессоров x_1 и x_2 и пересчета параметров регрессии получено уравнение, в котором единственным «действующим» регрессором остается БПК₂₀ (табл. 134). Уравнение регрессии при высокой статистической достоверности всех параметров регрессии приобретает следующий вид:

$$Y=185,3-19,2x_3$$

Понятно, что исключение или замена того или иного регрессора должны осуществляться с учетом всех факторов объективной реальности. Например, низкая статистическая репрезентативность того или иного регрессора может быть обусловлена не его объективно низкой ролью, а недостаточным числом наблюдений, низким качеством проведенных измерений и т.п.

11.8.3. Критерии сопряженности

MS Excel не имеет встроенных функций прямого вычисления коэффициентов сопряженности. Вместе с тем, существуют возможности полуавтоматического решения такого рода задач, в том числе и на основе критерия χ^2 . Первым этапом здесь является вычисление χ^2 и его параметров. В качестве примера возьмем исходные данные из табл. 128. Для удобства работы расположим их в *MS Excel* в диапазоне клеток A1:D4, как представлено на рис. 116 (полужирным шрифтом выделены итоги по строкам и столбцам таблицы):

	A	B	C	D
1	11	12	23	46
2	17	16	24	57
3	72	11	26	109
4	100	39	73	212
5				
6	21,698	8,462	15,840	
7	26,887	10,486	19,627	
8	51,415	20,052	37,533	212,0
9				
10				
11	Значимость критерия χ^2			1,0025E-06
12	Величина критерия χ^2			33,3714981
13	Критерий сопряженности C			0,36878716
14	Критерий Чупрова K			0,28054654

Рис. 116. Размещение данных при вычислении коэффициентов сопряженности

Первым шагом вычисления критерия χ^2 является вычисление «ожидаемых» значений частот. Для этого в ячейке с координатами A6 наберите формулу **A\$4*\$D1/\$D\$4**. В этой формуле **A\$4** – итог по столбцу таблицы исходных данных, **\$D1** – итог по строке, а **\$D\$4** – общий итог исходной таблицы. Знак \$ используется для полной или частичной фиксации адреса при копировании⁸⁵.

⁸⁵ Все адреса клеток в таблицах *MS Excel* указываются латинским шрифтом!

Скопируйте набранную формулу в область A6:C8. После копирования в этой области разместятся «ожидаемые» значения частот. В клетке D8 рекомендуется для контроля разместить функцию СУММ(A6:C8), т.к. при правильном вычислении суммы ожидаемых (клетка D8) и фактических частот (клетка D4) должны совпадать.

Разместите последовательно в ячейках A11, A12, A13, A14 пояснительные надписи в соответствии с рис. 116.

После этого установите курсор в ячейку D11 и занесите в нее функцию ХИ2ТЕСТ (<Формулы>, <Вставить функцию>, <Мастер функций>). В открывшемся окне функции укажите аргументы функции «Фактический интервал» – A1:D4, «Ожидаемый интервал» – A6:C8. Нажмите [OK]. В клетке появится итог вычисления значимости χ^2 (в зависимости от формата, установленного для клетки): 0,000001003 или 1,0025E-6.

Для того, чтобы вызвать абсолютное значение χ^2 , в клетку D12 вставьте функцию ХИ2ОБР и укажите в строке «Вероятность» адрес вычисленной вероятности χ^2 – D11, а в строке «Степени свободы» – число степеней свободы в исходной таблице = (3-1)*(3-1)=4. Нажмите <OK>. В клетке появится значение χ^2 .

Значение критерия сопряженности C вычисляется в соответствии с формулой: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$. Для этого в клетку D13 введите функцию КОРЕНЬ(D14/(D14+D4)), где D14 — адрес клетки, содержащей значение χ^2 , D4 – число наблюдений.

Значение критерия Чупрова K вычисляется по формуле: $K = \sqrt{\frac{\chi^2}{n \cdot (r-1) \cdot (c-1)}}$

Для этих вычислений в клетку D14 введите функцию КОРЕНЬ(D14/(D4*корень(4))), где D14 – адрес клетки, содержащей значение χ^2 , D4 – число наблюдений = (3-1)*(3-1).

Результаты после введения всей поясняющей информации будут представлены в следующем виде (рис. 116).

Общим итогом выполнения этих операций является не только получение набора результатов, но и алгоритм расчета критериев сопряженности, который создан на листе *MS Excel* и пригоден для многократного использования при практических расчетах в таблицах размерности 3x3.

При необходимости использовать таблицы другой размерности нужно менять адреса ссылок на ячейки в соответствии с логикой анализа, представленного в разделе. При этом необходимо «вручную» менять число степеней свободы, на которое есть ссылки в формулах.

11.9. Дисперсионный анализ

В наиболее общем виде *дисперсионный анализ* – метод математической статистики, позволяющий оценивать различия разных групп наблюдений и на этой основе находить взаимосвязь качественных и количественных характеристик изучаемых статистических совокупностей. С помощью дисперсионного анализа устанавливаются изменения дисперсии результатов наблюдений при различных уровнях изучаемого фактора. Эти уровни составляют основу качественной группировки результатов наблюдения, а различие (изменчивость) средних этих групп и является главным компонентом, на котором основывается вывод о значимости влияния фактора на наблюдаемые величины.

Дисперсионный анализ может применяться и для определения различий. И, наоборот, схожести нескольких групп наблюдения. Уверенность в схожести отдельных групп позволяет в ходе обработки результатов укрупнять группы путем объединения и тем самым получать более детальную и статистически более достоверную информацию.

11.9.1. Общие принципы

Дисперсионный анализ позволяет давать численные оценки существенности влияния некоторого качественного фактора на изучаемую величину, точнее – на ее результирующие показатели. В зависимости от

глубины исследования может оцениваться влияние одного или нескольких факторов (многофакторный анализ).

Условия применения дисперсионного анализа

В классическом варианте дисперсионный анализ основывается на следующих предпосылках: исходные величины распределены нормально; дисперсии исходных данных одинаковы для всех групп. Отсюда вытекает одно из требований, предъявляемых к исходным величинам, – перед проведением дисперсионного анализа необходимо проверить нормальность распределения и равенство дисперсий величин, образующих группы наблюдений.

Другой базовой основой дисперсионного анализа является свойство аддитивности дисперсий, т.е. возможности прямого сложения дисперсий отдельных фрагментов (групп, подгрупп) статистической совокупности для получения общей дисперсии всей совокупности.

Практической основой дисперсионного анализа является оценка 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 этих отклонений. Затем их суммируют и делят на число наблюдений, т.е. усредняют отклонения, получая значение дисперсии.

Таким образом, суть дисперсионного анализа заключается в расчленении общей дисперсии изучаемого признака на отдельные компоненты – межгрупповые дисперсии, обусловленные влиянием конкретных факторов, и остаточную

дисперсию. Сравнивая эти компоненты общей дисперсии, посредством F -критерия Фишера можно определить, какая часть общей изменчивости результивного признака обусловлена действием факторного признака.

В наиболее общем виде формулы вычисления параметров дисперсионного анализа выглядят следующим образом. Дисперсия в простом вариационном ряду:

$$D = \frac{\sum d^2}{n}$$

В варианте однофакторного анализа формула для расчета общей дисперсии значений признака всей совокупности относительно общего среднего представлена выражением:

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}}$$

Внутригрупповая дисперсия, остаточная дисперсия или дисперсия ошибки – есть среднее арифметическое групповых дисперсий, взвешенных по объемам отдельных групп:

$$D_{\text{внгр}} = \frac{\sum N_j D_j}{n}$$

где n – объем всей совокупности; N_j – объем группы j ; D_j – дисперсия группы j .

Межгрупповая или факторная дисперсия обусловлена влиянием какого-либо фактора:

$$D_{\text{межгр}} = \frac{\sum N_j (M_j - M)^2}{n}$$

где M_j – групповое среднее группы j , M – общее среднее; n – объем всей совокупности, N_j – объем группы j .

Для удобства расчетов в ходе дисперсионного анализа при большей части арифметических операций применяется основной элемент дисперсии – *сумма квадратов отклонений от среднего*. В литературе эта величина нередко именуется «*Варiances*». В *MS Excel* она обозначается как *SS*.

Группировка отдельных вариант производится на основе качественных дискретных признаков (пол, профессия, диагноз заболеваний и т.п.). *Например,*

распределение обследованных пациентов по полу: мужчины и женщины (дискретные группы). При числовых, непрерывно меняющихся признаках группировка производится по интервалам значений признаков. Например, распределение обследованных по возрасту: 20-29 лет, 30-39 лет и т.д.

Не вдаваясь в детали математического обоснования алгоритма вычислений различных параметров дисперсионного анализа, можно выделить следующие последовательные этапы:

1. Вычисление общей суммы квадратов отклонений $SS_{общ}$.
2. Вычисление факториальной суммы квадратов отклонений $SS_{факт}$ и суммы квадратов отклонений повторностей $SS_{повт}$.
3. Вычисление остаточной суммы квадратов отклонений $SS_{ост}$.
4. Нахождение степеней свободы для всех вычисленных квадратов отклонений df .
5. Вычисление факториальной дисперсии, дисперсии повторностей и остаточной дисперсии: $D_{факт}$, $D_{повт}$, $D_{ост}$.
6. Проведение парного сравнения факториальной и остаточной дисперсий. Дисперсии повторностей и остаточной дисперсии.
7. Оценка значимости различий сравниваемых групп с помощью теоретических значений распределения Фишера-Снедекора.
8. Оценка линейной взаимосвязи факториального и результативного признаков с помощью коэффициента детерминации R^2 .

Для подробного рассмотрения метода ограничимся относительно простым случаем однофакторного анализа, когда на количественный результирующий признак воздействует один качественный фактор (факториальный признак).

Главным условием использования однофакторного дисперсионного анализа является повторность опытов (наблюдений). В этой связи следует отметить, что во многих справочниках и пособиях не указывается на необходимость учета роли повторностей. В такой ситуации дисперсия повторностей автоматически включается в состав остаточной дисперсии.

Однако здесь, при малом числе повторностей или при недостаточно однородном составе повторностей, возникает угроза ошибочного оценивания результатов анализа. Включение в общий алгоритм дисперсионного анализа процедур определения дисперсии повторностей и оценки ее значимости позволяет предотвратить такого рода ошибки.

Например, требуется методом дисперсионного анализа проверить нулевую гипотезу о равенстве средней продолжительности заболеваний хроническим бронхитом, повлекшим временную нетрудоспособность, среди рабочих 3-х основных профессиональных групп: станочники, слесари, прочие. Четыре повторные выборки одинакового объема сделаны на предприятии из цехов с близкими условиями и характером труда работников (табл. 135).

Таблица 135

Длительность заболевания хроническим бронхитом (исходные данные)

Номер цеха (повторности)	Факториальный признак (профессия)		
	Уровни фактора (число дней болезни)		
	Станочники	Слесари	Прочие
1	10	5	4
2	14	11	6
3	9	7	11
4	12	8	4
В среднем	11,25	7,25	6,25
Общее среднее $M=8,42$ дня			

Для упрощения расчетов вычтем из каждой варианты исходных данных число, равное общему среднему (8,42). На конечный результат такое преобразование не окажет влияния, но значительно облегчит «ручные» вычисления и понимание принципа дисперсионного анализа (табл. 136):

Длительность заболевания хроническим бронхитом
(исходные данные уменьшены на величину среднего арифметического =8,42)

Номер цеха <i>i</i> (повторности)	Факториальный признак (профессия)			Сумма (<i>R</i> повт)	Сумма (<i>R</i> 2повт)
	Уровни фактора (<i>R</i> =число дней болезни=8,42)				
	Станочники	Слесари	Прочие		
1	1,6	-3,4	-4,7	-6,3	39,1
2	5,6	2,6	-2,4	5,8	33,1
3	0,6	-1,4	2,6	1,8	3,1
4	3,6	-0,4	-4,4	-1,3	1,1
Сумма (<i>R_i</i>)	11,4	-2,7	-8,7	-	-
Сумма (<i>R</i> 2 _{<i>i</i>})	128,4	7,1	75,1	-	-
Сумма $R_i^2=210,7$				Сумма $R_{2повт}=76,8$	

Все значения, находящиеся в клетках таблицы, возведем в квадрат (табл. 137).

Таблица 137

Длительность заболевания хроническим бронхитом
(данные табл. 136 возведены в квадрат)

Номер цеха <i>i</i> (повторности)	Факториальный признак (профессия)		
	Уровни фактора ($P=R^2$)		
	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
Сумма (P_i^2)	46,9	20,5	51,5
Сумма $P_i^2=118,9$			

Используя полученные промежуточные данные и учитывая, что число уровней фактора 3, число повторов 4, получаем:

- Общую сумму квадратов отклонений $SS_{общ} = 118,9$;
- Факториальную сумму квадратов отклонений и сумму квадратов отклонений повторностей $SS_{факт} = 210,7/4 = 52,6$; $SS_{повт} = 76,8/3 = 25,6$;
- Остаточную сумму квадратов отклонений $SS_{ост} = SS_{общ} - SS_{факт} - SS_{повт} = 118,9 - 52,6 - 25,6 = 40,7$;
- Число степеней свободы:
 - $f_{факт} = \text{число уровней факториального признака} - 1 = 3 - 1 = 2$;

- $f_{повт} = \text{число повторностей} - 1 = 4 - 1 = 3$;
- $f_{общ} = \text{число групп наблюдения} - 1 = 12 - 1 = 11$;
- $f_{ост} = f_{общ} - f_{факт} = 11 - 3 - 2 = 6$.

5. Находим факториальную дисперсию, дисперсию повторностей и остаточную дисперсии: $D_{факт} = 52,7/2 = 26,3$; $D_{повт} = 25,6/3 = 8,5$ и $D_{ост} = 40,7/6 = 6,7$;

6. Сравниваем факториальную и остаточную дисперсии (критерий Фишера):

$$F = \frac{D_{факт}}{D_{ост}} = \frac{26,3}{6,7} = 3,9$$

7. Сравниваем дисперсии повторностей и остаточную дисперсию (критерий Фишера):

$$F = \frac{D_{повт}}{D_{ост}} = \frac{8,5}{6,7} = 1,3$$

8. Проводим оценку значимости различий сравниваемых групп с помощью теоретических значений распределения Фишера-Снедекора. Учитывая, что число степеней свободы числителя равно 2, а знаменателя – 6 (число всех наблюдений – 1; число групп – 1) и уровень значимости – 0,05, с помощью встроенной функции *Ms Excel* или по таблице критических значений критерия Фишера (Приложение 3 3) находим критическую точку $F = 5,1$.

Аналогичную операцию проводим и для оценки статистической значимости повторностей. Так как полученные значения критерия Фишера F не превышают значения этой точки, нулевую гипотезу отвергать нет основания. Другими словами, источники вариации (группы наблюдения) в целом статистически не различаются (табл. 138).

Основные итоги однофакторного дисперсионного анализа

Источники вариации	Суммы квадратов отклонений SS	Число степеней свободы f	Дисперсии	$F_{выч}$	$F_{критич}$
Профессии (фактор.)	52,7	2	26,3	3,9	5,1
Повторности	25,5	3	8,5	1,3	4,7
Остаточная	40,7	6	6,7		
Итого	118,9	11			

9. Проводим оценку линейной взаимосвязи факториального и резульативного признаков с помощью коэффициента детерминации:

$$R^2 = \frac{52,7}{118,9} = 0,44, \text{ откуда коэффициент корреляции } R = \sqrt{0,44} = 0,7.$$

Статистическая значимость коэффициента корреляции определяется величиной критерия Фишера. Знак коэффициента корреляции, т.е. направленность связи, определяется сообразно логике распределения исходных данных. Понятно, что в случае качественных характеристик определение однозначной направленности связи, как правило, представляется сомнительным. Например, в данном случае нельзя утверждать, что какая-то профессия по своей сути лучше (первая), а какая-то хуже (вторая) и т.п.

Поскольку значимых различий распределения показателей длительности заболевания в разных профессиональных группах не выявлено, говорить о статистически существенном влиянии профессионального фактора на длительность болезни и связи заболеваемости хроническим бронхитом с профессиональным фактором нет никаких оснований.

Если такой итог не устраивает исследователя, то можно попробовать сравнить отдельные пары наиболее ярко различающихся групповых средних с помощью критерия Стьюдента. При достаточно большом объеме наблюдений и адекватности группировок разделения факториальных признаков этот прием может дать положительный результат.

В практике медико-биологических исследований нередко встречается необходимость анализировать большее число взаимодействующих факторов. Решение таких задач вызывает резкое увеличение громоздкости дисперсионного анализа. Поэтому обычно используются только однофакторные, максимум двухфакторные дисперсионные комплексы. Считается, что многофакторные комплексы можно обрабатывать путем последовательного анализа выделяемых из всей анализируемой совокупности одно- или двухфакторных комплексов.

Существенный недостаток дисперсионного анализа – трудность формирования однородных групп наблюдения. Среди причин, обуславливающих существование этого рода трудностей, является и объективная закономерность: все живое изменчиво, поэтому отдельные единицы наблюдения всегда обладают своими особыми, неповторимыми, индивидуальными чертами. Эти индивидуальные черты служат источником статистических погрешностей. Отсюда и возникает трудность выполнения одного из главных требований дисперсионного анализа — соблюдение однородности, т.е. одинаковой дисперсии численных характеристик сравниваемых совокупностей объектов. Но на практике при воздействии внешнего фактора зачастую проявляется еще одна черта биологических объектов – отдельные единицы наблюдения в силу своей индивидуальности, могут реагировать на разного рода воздействия неодинаково. Даже если будет существовать некая общая тенденция реагирования на влияние фактора, группа неизбежно будет расслаиваться в силу индивидуальных особенностей реагирования. И, как следствие, неизбежно будет возникать изменение разброса значений регистрируемых данных (внутригрупповая вариация). Понятно, что если факторное воздействие будет отсутствовать или будет очень слабым (контрольные или исходные уровни наблюдений), то и изменение внутригрупповой вариации будет отсутствовать или будет относительно слабым. При существенном воздействии фактора (опытная группа) изменения внутригрупповой вариации будут более существенны. Отсюда очевидно: чем

больше давление того или иного фактора на наблюдаемую группу, тем больше изменение вариабельности регистрируемых величин в этой группе⁸⁶.

Устраняется неоднородность сравниваемых групп специальными методами отбора. Но эти методы отбора почти неизбежно нарушают другое обязательное правило статистического анализа – случайность формирования выборок.

11.9.2. Дисперсионный анализ в MS Excel

В программе *MS Excel* имеются функциональные возможности реализации задач дисперсионного анализа. В их числе:

- «**Пакет анализа**», позволяющий проводить дисперсионный анализ следующих видов: однофакторный, двухфакторный без повторений, двухфакторный с повторениями;
- **Сводные таблицы**, которые обеспечивают решение задачи выборки из больших объемов наблюдений совокупностей для дисперсионного анализа.

11.9.3. Методики дисперсионного анализа, основанные на применении функции MS Excel «Пакет анализа»

Однофакторный дисперсионный анализ.

Обязательным условием применения такого анализа является повторность данных. Следует отметить, что прямое использование дисперсионного анализа в *MS Excel* не предусматривает учета повторностей при однофакторном анализе. В ряде случаев это может приводить к существенным ошибкам, поэтому рекомендуется обращать внимание на соблюдение определенных правил, не упомянутых в многочисленных пособиях по применению *MS Excel*.

⁸⁶ Закон больших чисел, согласно которому в массе наблюдений последствия воздействия на итоги случайных обстоятельств взаимно погашаются, здесь не может работать, поскольку давление исследуемого фактора нарушает элемент случайности.

Например, необходимо оценить влияние условий и характера труда (в данном случае профиля цеха) на заболеваемость рабочих гастритом. Исходные данные расположены в таблице *MS Excel* (рис. 117).

Согласно таблице, исследуемая результативная характеристика представлена числами заболеваний (на 100 раб. за год), которые рассматриваются в 3-х качественно разных подгруппах одного факториального признака: условий и характера труда (доменный, мартеновский и прокатный цеха). В каждой подгруппе имеется по 4 повторности. Повторность данных обеспечена результатами наблюдений за работниками 4-х разных заводов, схожих по профилю и технологии производства.

	A	B	C	D
1	Заболеваемость гастритом на 100 раб. за год			
2	Заводы		Цеха (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	36,6	21,9
7	4	21,5	29,6	38,1
8				

Рис.117. Исходные данные однофакторного комплекса

Для решения задачи в *MS Excel*:

- Выберите пакет <Анализ данных> из меню <Данные>⁸⁷.
- В соответствии с условиями задачи выберите в появившемся диалоговом окне метод «**Однофакторный дисперсионный анализ**» и нажмите [OK].

В появившемся окне «**Однофакторный дисперсионный анализ**» установите:

⁸⁷ В более ранних версиях *MS Excel* этот пакет расположен в меню «Сервис».

- в разделе <Входные данные> входной интервал (\$B\$4:\$D\$7);
- в разделе <Группирование> активизируйте переключатель <по столбцам>;
- метки (выбранный нами входной диапазон не содержит метки, т.е. названий строк и столбцов);
- альфа (уровень значимости =0,05).
- В разделе <Параметры вывода> установите переключатель в положение «**Выходной интервал**» и укажите начальную координату блока клеток, куда будут выводиться результаты анализа (\$A\$9).
- После завершения настройки параметров нажмите [OK].

Диалоговое окно с заполненными исходными параметрами должно выглядеть следующим образом (рис. 118):

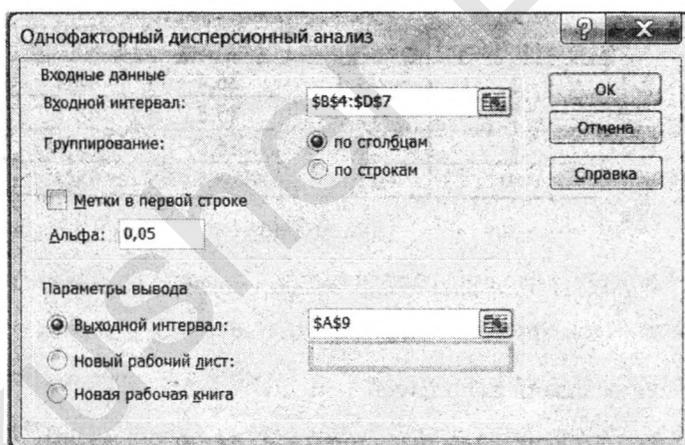


Рис.118. Окно данных однофакторного дисперсионного анализа

Результаты однофакторного дисперсионного анализа представлены в двух блоках (рис. 119). В первом – располагаются показатели описательной статистики групп наблюдения, распределенных по столбцам: числа наблюдений, суммы вариантов, средние арифметические, дисперсии. Во втором блоке приводятся следующие параметры:

SS – суммы квадратов по отдельным источникам вариации,

df – степени свободы,

MS – средние квадраты (дисперсии) отдельных источников вариации,

F – фактическое значение F -критерия Фишера,

P значение – значимость вычисленного F -критерия,

F критическое – критическое значение F -критерия при задаваемых пользователем значениях Альфа (0,05; 0,01 и т.п.).

В приведенном примере сумма квадратов, обусловленная влиянием исследуемого факториального признака (межгрупповая сумма), $SS_{факт}=607,87$. Остаточная сумма квадратов (внутригрупповая) $SS_{ост}$ равна 702,49.

Дисперсии: межгрупповая (факторная) $MS_{факт}=607,87/2=303,9$; остаточная $MS_{ост}=702,49/9=78,1$.

Соотношение этих дисперсий $MS_{факт}/MS_{ост}=303,9/78,1=3,89$, что при соответствующих числах степеней свободы df (2 и 9) определяет его значимость $P=0,06$, т.е. $>0,05$.

Основной вывод: поскольку значение вычисленного F -критерия Фишера меньше значения критического F -критерия, различия дисперсий распределения показателей заболеваемости гастритом работников считаются статистически неподтвержденными.

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	Счет	Сумма	Среднее	Дисперсия		
Столбец 1	4	85,6	21,4	30,673333		
Столбец 2	4	138,1	34,525	71,675833		
Столбец 3	4	151,6	37,9	131,81333		
Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	607,875	2	303,9375	3,8939305	0,06048005	4,256494729
Внутри групп	702,4875	9	78,0541667			
Итого	1310,363	11				

Рис.119. Результаты однофакторного дисперсионного анализа

Дополнительным аргументом в пользу этого вывода может быть оценка различий средних величин, полученных в ходе проведенного дисперсионного анализа, с помощью «обычного» – критерия Стьюдента.

Например, числа наблюдений, групповые средние по столбцам 1, 2 и 3, а также соответствующие им дисперсии представлены в первой таблице итогов дисперсионного анализа (рис. 119). Используя эти данные, значение t -критерия Стьюдента можно определить по формуле:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}, \text{ где } m_i = \sqrt{\frac{D_i}{n-1}}$$

Так, чтобы определить достоверность различий средних группы 1 и группы 2, последовательно находим значения:

$$m_1 = \sqrt{\frac{D_1}{n-1}} = \sqrt{\frac{30,67}{4-1}} = 3,2 \quad \text{и} \quad m_2 = \sqrt{\frac{D_2}{n-1}} = \sqrt{\frac{71,8}{4-1}} = 4,9$$

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{34,5 - 21,4}{\sqrt{3,2^2 + 4,9^2}} = 1,73$$

С помощью функции СТЬЮДРАСП находим значение доверительной вероятности (нулевой гипотезы) $P \approx 0,3$. Так как $P > 0,05$, нулевая гипотеза, говорящая об отсутствии различий средних, не отвергается, что соответствует выводу об отсутствии различий дисперсий, полученному ранее.

Вместе с тем, как уже отмечалось выше, при таком алгоритме расчетов различий дисперсий и средних не учитывается влияние на конечный результат повторностей наблюдений, которые при определенных обстоятельствах (малое число наблюдений, неточности при формировании выборочных повторностей и т.п.) способны вызывать существенные погрешности анализа. Предотвратить их можно с помощью оценок значимости повторностей. Для этого весьма подходит имеющийся в пакете <Анализ данных> двухфакторный анализ с неповторяющимися данными. Несмотря на то, что это название свидетельствует о том, что этот вариант дисперсионного анализа в *MS Excel* не предназначен для однофакторного анализа, он вполне подходит для

решения задач однофакторного анализа, позволяя при этом учитывать роль повторностей наблюдений.

Для удобства работы возьмем уже использованные в предыдущем примере исходные данные. Однако рассмотрим их несколько иначе. Будем считать, что мы оцениваем влияние на заболеваемость рабочих гастритом уже двух факториальных признаков: «Условия труда» (цех) и «Повторности» (рис. 120).

1. Выполним команду <Анализ данных> из меню <Данные>;
2. В соответствии с поставленной задачей в появившемся диалоговом окне выбираем метод «**Двухфакторный дисперсионный анализ без повторений**» и нажимаем [ОК]. Обратить внимание! Несмотря на то, что анализ, по существу, однофакторный используется один из имеющихся в пакете <Анализ данных> вариантов двухфакторного дисперсионного анализа;
3. В окне «Двухфакторный дисперсионный анализ без повторений» установим для входных данных следующие параметры:
4. входной интервал (\$B\$4:\$D\$7);
5. метки (выбранный нами входной диапазон не содержит метки, т.е. названий строк и столбцов);
6. альфа (уровень значимости =0,05).

	А	В	С	Д
1	Заболеваемость гастритом на 100 раб.за год			
2	Заводы		Цеха (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	36,6	21,9
7	4	21,5	29,6	38,1
8				

Рис.120. Исходные данные однофакторного комплекса с учетом повторностей

7. В разделе <параметры вывода> установим переключатель в положение «**Выходной интервал**» и укажем начальную координату блока клеток, куда будут выводиться результаты анализа (\$A\$9);

8. После завершения настройки параметров нажимаем [OK].

Диалоговое окно с заполненными исходными параметрами должно выглядеть следующим образом (рис. 121):

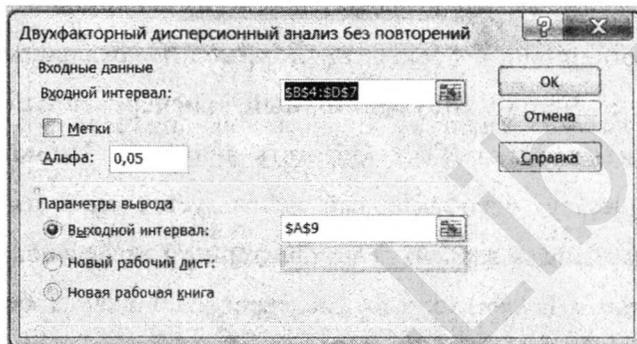


Рис.121. Окно данных однофакторного комплекса с повторностями

Результаты двухфакторного дисперсионного анализа, как и в предыдущем случае, представлены в виде двух блоков (рис. 122). В первом – располагается описательная статистика повторностей (заводы), распределенных по строкам: числа наблюдений, суммы вариант, средние арифметические, дисперсии и аналогичная статистика факториального признака, группировка которого проведена по столбцам (цеха).

В другом блоке результатов дисперсионного анализа представлены:

SS – суммы квадратов по отдельным источникам вариации. По повторности (строки) равна 96,1; по факториальному признаку (столбцы) равна 607,9. Остаточная, внутригрупповая сумма квадратов (погрешность) равна 606,3. Итоговая сумма – 1310,363.

Исходя из этих данных, путем несложных вычислений можно получить значения соответствующих коэффициентов детерминации, позволяющие

оценить вклад отдельных источников вариации в общую вариацию исследованной совокупности:

$$R^2_{повтор} = 96,2/1310,4 = 0,073 \text{ или } 7,3\%; R^2_{факт} = 607,9/1310,4 = 0,464 \text{ или } 45,4\%;$$

$$R^2_{погр} = 606,3/1310,4 = 0,463 \text{ или } 46,3\%.$$

Определение статистической достоверности (значимости) дисперсионного анализа проводится на основе следующих данных второй таблицы итогов:

- df – степени свободы, соответственно, 3, 2, 6 и 11.
- MS – средние квадраты (дисперсии) отдельных источников вариации 32,1; 303,9; 101,1.
- F – фактическое значение F - критерия Фишера.
- P -значение – значимость вычисленного F критерия для факторов, расположенных по строкам, – 0,317; по столбцам – 3,008.
- F критическое – критическое значение F при $P=0,05$.

F -критерии Фишера для повторностей: $F_1=32,1/101,1=0,317$. Для факториального признака «Условия труда»: $F_2=303,9/101,1=3,008$. Соответственно значимость для повторностей $P=0,81$; для факториального признака «Условия труда» $P=0,124$.

Двухфакторный дисперсионный анализ без повторений

ИТОГИ	Счет	Сумма	Среднее	Дисперсия
Строка 1	3	92,9	30,9666667	112,66333
Строка 2	3	107,8	35,9333333	369,66333
Строка 3	3	85,4	28,4666667	55,863333
Строка 4	3	89,2	29,7333333	68,903333
Столбец 1	4	85,6	21,4	30,673333
Столбец 2	4	138,1	34,525	71,675833
Столбец 3	4	151,6	37,9	131,81333

Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F критическое
Строки	96,17583	3	32,0586111	0,3172488	0,8130051	4,757062664
Столбцы	607,875	2	303,9375	3,0077353	0,12451779	5,14325285
Погрешность	606,3117	6	101,051944			
Итого	1310,363	11				

Рис.122. Результаты однофакторного дисперсионного анализа с повторностями

Из полученных результатов можно сделать следующий **вывод**: есть основания не отвергать нулевую гипотезу об отсутствии влияния двух рассмотренных факторов («условий труда» и «повторностей наблюдений») на заболеваемость рабочих гастритом, так как величина значимости критерия Фишера превышает в обоих случаях 0,05.

Следует обратить внимание на то, что факториальный признак «Условия труда», в случае когда проводился однофакторный анализ без учета повторностей и когда проводился с их учетом, оставаясь постоянно незначимым для отбрасывания нулевой гипотезы, в первом случае все же был существенно ближе к критической точке значимости. Причиной такого явления является то, что во втором случае проводился более «чистый опыт», поскольку из совокупности результатов выделялся фактор «повторности». Тем самым снижался информационный шум, т.е. уменьшалось влияние на конечный результат анализа неточностей формирования повторностей, которые играли роль помехи. Отсюда следует вывод: использование однофакторного дисперсионного анализа без учета повторностей чревато возможными ошибками.

Двухфакторный анализ с неповторяющимися данными.

Поскольку в предыдущем примере в качестве повторностей использовались выборки вполне определенного происхождения, которое само по себе могло играть заметную роль в формировании изучаемой заболеваемости, можно рассмотреть предыдущие исходные данные под другим углом зрения, т.е. принять повторности как самостоятельный факториальный признак: «Место труда (завод)».

Для решения задачи в *MS Excel* оформим таблицу с исходными данными несколько иначе (рис. 123).

	A	B	C	D
1	Заболеваемость гастритом на 100 раб.за год			
2	Заводы	Цеха (фактор)		
3		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	36,6	21,9
7	Завод 4	21,5	29,6	38,1
8				

Рис.123. Исходные данные двухфакторного (без повторностей) комплекса

Весь остальной алгоритм действий и конечные итоги остаются такими же, как и в предыдущем примере. Единственным отличием является замена термина «Повторности» на название второго факториального признака «Завод».

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) – факториальный признак 1 и влияние заводских условий (№ завода) – факториальный признак 2. В качестве повторностей наблюдения будем рассматривать повторения наблюдений на каждом заводе в течение трех лет (рис. 124).

	A	B	C	D	E
1	Год	№	Цеха		
2	повторности	Заводов	Доменный	Мартен	Прокатный
3	2010	Завод 1	23,1	25,4	43,1
4	2011		22,0	26,5	47,0
5	2012		25,0	27,2	39,2
6	2010	Завод 2	11,1	45,5	52,2
7	2011		18,3	48,7	54,3
8	2012		12,0	42,2	38,7
9	2010	Завод 3	24,5	38,3	18,5
10	2011		28,9	25,4	22,1
11	2012		27,4	33,1	25,2
12	2010	Завод 4	18,9	33,5	38,8
13	2011		22,2	25,5	41,1
14	2012		23,3	29,9	34,5

Рис.124. Исходная таблица для двухфакторного дисперсионного комплекса с повторениями

Для решения задачи в *MS Excel*:

1. Вызовите <Пакет анализа> из меню <Данные>.
2. В появившемся диалоговом окне нужно выбрать метод «Двухфакторный дисперсионный анализ с повторениями» и нажать [OK].
3. В окне «Двухфакторный дисперсионный анализ с повторениями» установите для входных данных следующие параметры (рис. 125):
4. в разделе <Входные данные> – входной интервал (\$B\$2:\$E\$14). Обязательно в этот интервал должен быть включен один ряд клеток с названиями столбцов (градациями одного факториального признака) и названием ряда другого факторного признака исходной таблицы (сверху);
5. число строк для выборки – 3;
6. Альфа (уровень значимости =0,05).
7. В разделе <Параметры вывода> установите переключатель в положение <Выходной интервал>. Выходной интервал можно задать в виде одной координаты \$G\$1 начального блока клеток, куда будет выведен результат.
8. После завершения ввода этих параметров нажмите [OK].

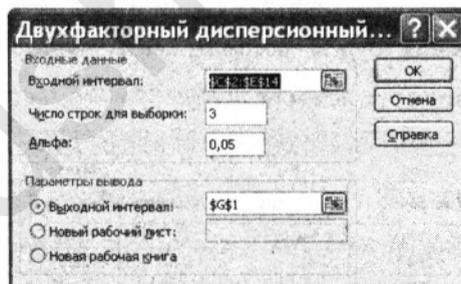


Рис.125. Окно исходных параметров двухфакторного дисперсионного анализа с повторениями

Результаты дисперсионного анализа будут представлены в двух группах. В первой – для каждой строки и каждого столбца исходной таблицы приведена описательная статистика: количество чисел, сумма, среднее и дисперсия по первому, второму фактору и итогам (рис. 126).

Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	Доменные Мартен		Прокатны Итого	
	Завод 1			
Счет	3	3	3	9
Сумма	70,1	79,1	129,3	278,5
Среднее	23,3667	26,3667	43,1	30,9444
Дисперсия	2,30333	0,82333	15,21	89,3853
Завод 2				
Счет	3	3	3	9
Сумма	41,4	136,4	145,2	323
Среднее	13,8	45,4667	48,4	35,8889
Дисперсия	15,39	10,5633	71,67	300,474
Завод 3				
Счет	3	3	3	9
Сумма	80,8	96,8	65,8	243,4
Среднее	26,9333	32,2667	21,9333	27,0444
Дисперсия	5,00333	42,1233	11,2433	34,6203
Завод 4				
Счет	3	3	3	9
Сумма	64,4	88,9	114,4	267,7
Среднее	21,4667	29,6333	38,1333	29,7444
Дисперсия	5,24333	16,0533	11,2233	60,2203
Итого				
Счет	12	12	12	
Сумма	256,7	401,2	454,7	
Среднее	21,3917	33,4333	37,8917	
Дисперсия	30,239	70,0679	126,866	

Рис.126. Результаты двухфакторного (с повторностями) дисперсионного анализа в MS Excel. Первая группа

Вторая группа результатов представлена на рис. 127. Содержательная часть этой группы ссылок уже рассматривалась ранее (однофакторный без повторений, с повторениями и двухфакторный дисперсионный комплекс). В данном варианте дисперсионного анализа эта группа итогов в отличие от рассмотренных ранее содержит и данные о результатах оценок взаимодействия факторов (взаимодействие). Кроме того, особенностью пакета *Microsoft Excel* является то, что в итогах двухфакторного с повторениями дисперсионного анализа данные по строкам, отражающие значения одного из факторных признаков, именованы как «Выборка», а значения другого фактора именованы как «Столбец».

Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F критическое
Выборка	369,823	3	123,274	7,15153	0,00135355	3,00878657
Столбцы	1748,51	2	874,257	50,7183	2,4068E-09	3,402826105
Взаимодействие	1715,38	6	285,897	16,5858	1,7758E-07	2,508188823
Внутри	413,7	24	17,2375			
Итого	4247,42	35				

Рис.127. Результаты двухфакторного (с повторностями) дисперсионного анализа в MS Excel

Согласно представленным итогам, можно заключить, что конкурирующая гипотеза может быть принята для факториального признака 1 и для факториального признака 2, а также их взаимного влияния, т.к. во всех случаях результаты неравенства $F > F_{кр}$ соответствуют величине значимости нулевой гипотезы $P < 0,05$.

Примечание. Уровни значимости (P) в строках итоговой таблицы «Столбцы» и «Взаимодействие» приведены в компактной, экспоненциальной форме. Например, число 24068E-09 выглядит в обычной записи как 0,0000000024068. E-9 обозначает, что первая значащая цифра стоит на девятом месте после запятой.

Таким образом, можно считать доказанным, с высокой степенью статистической достоверности, влияние условий и характера труда на показатели заболеваемости хроническим гастритом.

11.9.4. Методики дисперсионного анализа с использованием сводных таблиц и «Пакета анализа» MS Excel

Несмотря на кажущуюся громоздкость описание этих методик, их практическое использование чрезвычайно просто и доступно любому пользователю, знакомому с основами работы в MS Excel.

Достоинством совмещенного применения сводных таблиц и «Пакета анализа» MS Excel является простота группировок исходных данных и расчетов параметров дисперсионного анализа. При этом для однофакторного дисперсионного анализа обязательны повторности наблюдений.

Следует отметить, что прямой дисперсионный анализ содержимого сводных таблиц с помощью <Пакета анализа> неизбежно приводит к существенным ошибкам из-за функциональных особенностей *MS Excel*. Получать здесь надежные результаты можно, только используя особые технологические приемы.

Например, предположим, что необходимо оценить связь места жительства (по факту регистрации: город или пригород) и такого показателя физического развития мужчин трудоспособного возраста, как рост. Распределение 48 обследованных мужчин по признаку «Регистрация» представлены в сводной таблице *MS Excel* (рис 128). Имеются 4 повторности наблюдений.

	A	B	C	D	
1	Количество по полю Рост	Регистрация			
2	Повторности	Город	Пригород	Общий итог	
3	Повторность 1		9	3	12
4	Повторность 2		6	6	12
5	Повторность 3		4	8	12
6	Повторность 4		5	7	12
7	Общий итог		24	24	48

Рис.128. Сводная таблица распределения чисел наблюдений

Весь ход анализа для краткости изложения будем рассматривать, не вдаваясь в алгебраическое обоснование алгоритма вычислений. Для реализации дисперсионного анализа данные сводной таблицы должны быть только в показателях несмещенной дисперсии. Использование здесь полей сводной таблицы с любыми другими параметрами данных недопустимо, так как неизбежно ведет к ошибочным итогам.

Однофакторный дисперсионный анализ без повторностей.

Разместим на листе *MS Excel* два макета таблиц результатов: сводную таблицу, содержащую только итоги по столбцам⁸⁸, что достаточно для решения задачи однофакторного дисперсионного анализа без повторностей в блоке ячеек A1:D3 (рис. 129) и макет стандартной таблицы итогов дисперсионного анализа в блоке ячеек A6:F10.

	A	B	C	D	E	F
1	Несмещенная дисперсия по полю Рост	Регистрация г.				
2		Город	Пригород	Общий итог		
3	Итого	20,25	19,33	32,04		
4						
5						
6	Источники вариации	SS	df	MS	F	P
7	Фактор "Проф. группа"					
8	Внутригрупповая					
9						
10	Итого					
11						

Рис. 129. Сводная таблица несмещенных дисперсий результативного признака и макет таблицы итогов дисперсионного анализа без повторностей

1. Вычислим значение общей суммы квадратов отклонений $SS_{общ} = D_{общ} * N = 32,04 * 48 = 1537,92$, где N – общее число наблюдений, а $D_{общ}$ – общая дисперсия (**Общий итог**). Практическая формула этой операции в ячейке **B10** выглядит следующим образом: **=D3*48**.

В ячейку с адресом **C10** введем число степеней свободы $df_{общ} = N - 1 = 48 - 1 = 47$, где $N = 48$ – общее число наблюдений (рис. 128).

2. Вычисляем сумму квадратов отклонений по группам факториального признака. При равенстве чисел наблюдения по градациям факторного признака $SS_{факт} = SS_{общ} - (D_1 + D_2) * k = 1537,92 - (20,25 + 19,33) * 2 = 588,0$, где:

- k – число наблюдений по любому столбцу, кроме столбца **Общий итог**.

В примере (рис. 128) $k = 24$;

- D_1 – дисперсия первой градации факториального признака (**Город**);

⁸⁸ Предполагается, что повторности наблюдений существуют, только не выводятся в итоговую таблицу, чтобы не загромождать результаты.

- D_2 – дисперсия второй градации факториального признака (Пригород).

Практическая формула этой операции, вводимая в ячейку **B7**, имеет вид:
=B10-(B3+C3)*24.

3. Для $SS_{факт}$ число степеней свободы $df_{факт}$ определяется, как число градаций факториального признака (место регистрации) – 1. $df_{факт}=2-1=1$. Это значение вводим в ячейку **C7**.

4. Остаточная (внутригрупповая) сумма квадратов отклонений вычисляется по формуле: $SS_{ост}=SS_{общ}-SS_{факт}=1537,92-588,0=949,92$, для чего в ячейку **B8** рабочего листа *MS Excel* введем формулу этой операции: **=B10-B7**.

5. Для $SS_{ост}$ число степеней свободы $df_{ост}=df_{общ}-df_{факт}=47-1=46$. Это значение вводим в ячейку **C8**.

6. Знание вариации (суммы квадратов отклонений) факториального признака, внутригрупповой вариации и соответствующих им степеней свободы позволяет вычислить соответствующие дисперсии:

- $D_{факт}=SS_{факт}/df_{факт}=588,0/1=588,0$ и $D_{ост} = SS_{ост}/df_{ост}=949,92/46=20,65$;
- их соотношение (критерий Фишера) $F=588/20,65=28,47$.

7. Из меню командной строки <Формулы> с помощью команды <Вставить функцию> вызываем <Мастер функций>. В открывшемся окне из раздела «Категория» выбираем функцию **FPACП** (рис. 130).

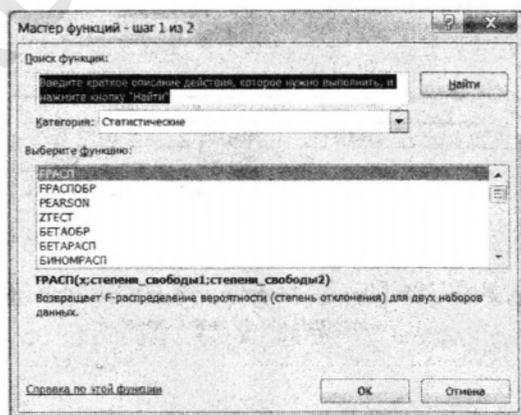


Рис. 130. Окно мастера функций. Выбор функции **FPACП**

8. Затем в открывшемся окне <Аргументы функции> указываем адреса ячеек итоговой таблицы, где расположены значения аргументов (Рис.): вычисленного критерия F – в ячейке **E7**; степени свободы 1 – в ячейке **C7**; степени свободы 2 – в ячейке **C8** (рис. 131).

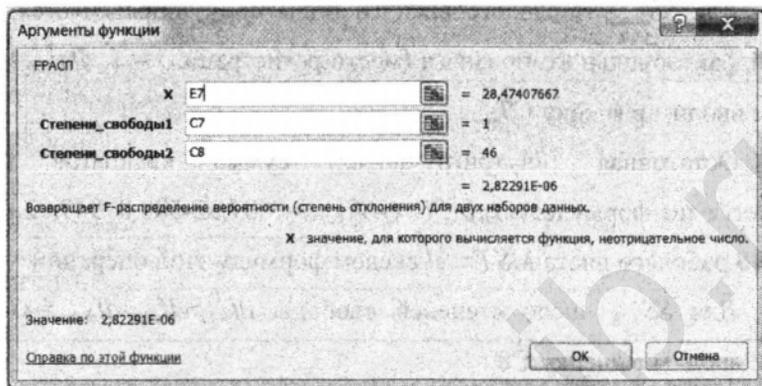


Рис.131. Окно аргументов функции FRASPI

Согласно итогам, представленным на рис. 132, фактор «Проф. группа» статистически значимо влияет на распределение показателей роста, т.к. есть основания для отбрасывания нулевой гипотезы о незначимости влияния места проживания на распределение показателей значений роста ($F > F_{кр}$ при $P < 0,05$).

	A	B	C	D	E	F
1	Несмещенная дисперсия по полю Рост	Регистрация				
2		Город	Пригород	Общий итог		
3	Итого	20,25	19,33	32,04		
4						
5						
6	Источники вариации	SS	df	MS	F	P
7	Фактор "Проф.группа"	588,00	1	588,00	28,47	2,8229E-06
8	Внутригрупповая	949,92	46	20,65		
9						
10	Итого	1537,92	47			

Рис.132. Итоговая таблица однофакторного дисперсионного анализа в MS Excel без учета повторностей наблюдений

Однофакторный дисперсионный анализ с повторностями.

Разместим на листе *MS Excel* две таблицы:

1. Сводную таблицу распределения факторного признака «Регистрация» и признака «Повторности».

2. Макет таблицы итогов такого дисперсионного анализа (рис. 133).

Поскольку исходные данные представлены той же совокупностью данных, что и в предыдущем примере, первые четыре операции практически зеркально повторяют операции, выполняемые при однофакторном дисперсионном анализе без повторностей. Изменяться будут только адреса ячеек (из-за того, что во вторую таблицу итогов добавилась информация по повторениям).

	A	B	C	D	E	F
1	Несмещенная дисперсия по полю Рост	Регистрация				
2	Повторности	Город	Пригород	Общий итог		
3	Повторность 1	18,17	18,67	23,41		
4	Повторность 2	18,14	18,47	34,31		
5	Повторность 3	18,19	17,48	35,22		
6	Повторность 4	18,80	18,78	31,69		
7	Общий итог	20,25	19,33	32,04		
8						
9						
10	Источники вариации	SS	df	MS	F	P
11	Фактор "Проф. группа"					
12	Повторности					
13	Внутригрупповая					
14						
15	Итого					
16						

Рис.133. Сводная таблица несмещенных дисперсий результативного признака и макет таблицы итогов однофакторного дисперсионного анализа с повторностями

1. Вычислим значение общей суммы квадратов отклонений $SS_{общ} = D_{общ} * N = 32,04 * 48 = 1537,92$, где N – общее число наблюдений, а $D_{общ}$ – общая дисперсия (**Общий итог**). Практическая формула этой операции на листе *MS Excel* в ячейке с адресом **B15** выглядит следующим образом: **=D7*48**.

2. В ячейку с адресом **C15** введем число степеней свободы $df_{общ} = N - 1 = 48 - 1 = 47$.

3. Вычисляем сумму квадратов отклонений по группам факториального признака «Регистрация». При равенстве чисел наблюдения по градациям факторного признака $SS_{факт} = SS_{общ} - (D_1 + D_2) * k = 1537,92 - (20,25 + 19,33) * 2 = 588,0$, где:

k – число наблюдений по любому из столбцов, кроме столбца «Общий итог». В примере (рис. 128) $k=24$;

- D_1 – дисперсия первой градации факториального признака (Город);
- D_2 – дисперсия второй градации факториального признака (Пригород).

Практическая формула этой операции, вводимая в ячейку В11, имеет вид: $=B15-(B7+C7)*24$.

4. Для $SS_{факт}$ число степеней свободы $df_{факт}$ определяется, как число градаций факториального признака – 1. $df_{факт} = 2 - 1 = 1$. Это значение вводим в ячейку рабочего листа С11.

5. Вычисляем сумму квадратов отклонений по признаку «Повторности». При равенстве чисел наблюдения по повторностям:

$SS_{повт} = SS_{общ} - (D_{повт1} + D_{повт2} + D_{повт3} + D_{повт4}) * k_{повт} = 1537,92 - (23,41 + 34,31 + 35,22 + 31,69) * 12 = 42,42$, где:

k – число наблюдений по любой из строк «Повторности», кроме строки «Общий итог». В примере (рис. 128) $k=12$;

- $D_{повт1}$ – дисперсия строки «Повторность 1»;
- $D_{повт2}$ – дисперсия строки «Повторность 2»;
- $D_{повт3}$ – дисперсия строки «Повторность 3»;
- $D_{повт4}$ – дисперсия строки «Повторность 4».

Практическая формула этой операции, вводимая в ячейку В12, имеет вид: $=B15-(D3+D4+D5+D6)*12$.

6. Число степеней свободы для повторностей $df_{повт}$ определяется, как число повторностей – 1. $df_{повт} = 4 - 1 = 3$. Это значение вводим в ячейку рабочего листа С12.

7. Остаточная (внутригрупповая) сумма квадратов отклонений вычисляется по формуле: $SS_{ост} = SS_{общ} - SS_{факт} - SS_{ном} = 1537,92 - 588,0 - 42,42 = 907,50$, для чего в ячейку **B13** рабочего листа *MS Excel* введем практическую формулу этой операции: **=B15-B11-B12**.

8. Для $SS_{ост}$ число степеней свободы $df_{ост} = df_{общ} - df_{факт} - df_{ном} = 47 - 1 - 3 = 43$. Это значение вводим в ячейку рабочего листа **C13**.

9. Знание вариации (суммы квадратов отклонений) факториального признака, внутригрупповой вариации и соответствующих им степеней свободы позволяет вычислить соответствующие дисперсии:

- $D_{факт} = SS_{факт} / df_{факт} = 588,0 / 1 = 588,0$ и $D_{ост} = SS_{ост} / df_{ост} = 949,92 / 46 = 20,65$;
- их соотношение (критерий Фишера) $F = 588 / 20,65 = 28,47$.

10. Из меню командной строки <Формулы> с помощью команды <Вставить функцию> вызываем <Мастер функций>. В открывшемся окне из раздела «Категория» выбираем функцию **FPАСП** (рис. 134).

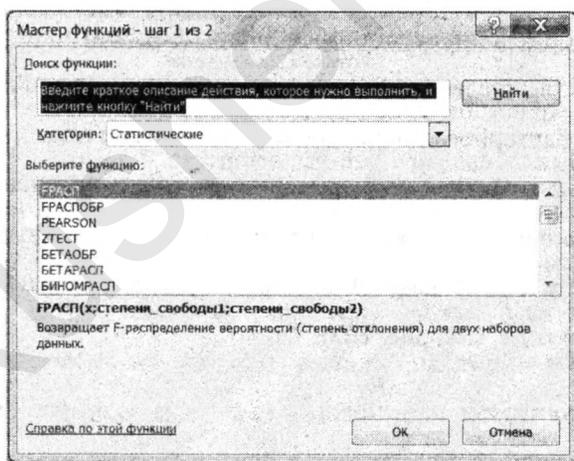


Рис. 134. Окно мастера функций. Выбор функции **FPАСП**

11. Затем в открывшемся окне <Аргументы функции> указываем адреса ячеек итоговой таблицы, где расположены значения аргументов для фактора «Регистрация»: вычисленного критерия F – в ячейке **E11**; степени свободы 1 – в ячейке **C11**; степени свободы 2 – в ячейке **C13**.

12. Аналогичная операция повторяется и для фактора «Повторности». При этом соответственно меняются адреса вычисленного критерия F в ячейке E12; степени свободы 1 – в ячейке C12; степень свободы 2 (остается прежней) – в ячейке C13. Согласно итогам, представленным на рис. 135, и фактор «Проф. группа», и фактор «Повторности» статистически значимо влияют на распределение показателей роста, т.к. основания для отбрасывания нулевой гипотезы об отсутствии влияния каждого из них – нет (в обоих случаях $F > F_{кр}$ при $P < 0,05$).

Источники вариации	SS	df	MS	F	P
Фактор "Проф. группа"	588,00	1	588,00	27,86	4,0490E-06
Повторности	42,42	3	14,14	0,67	5,7515E-01
Внутригрупповая	907,50	43	21,10		
Итого	1537,92	47			

Рис.135. Итоговая таблица однофакторного дисперсионного анализа (с учетом значимости повторностей наблюдений)

Двухфакторный дисперсионный анализ без повторностей.

Алгоритм вычисления всех его параметров полностью аналогичен алгоритму однофакторного дисперсионного анализа с повторениями. Различия заключаются только в том, что название учетного признака «Повторности» заменяется названием второго учетного признака. Соответственно, обозначения градаций этого признака заменяют названия повторностей «Повторность 1», «Повторность 2» и т.д. (см. рис. 133).

12. Оценка различий показателей заболеваемости

Одной из самых распространенных практических задач медицинской статистики является анализ различий показателей заболеваемости. Этот анализ обычно проводится с целью выяснения роли конкретных факторов, определяющих частоту и структуру заболеваемости, вероятности возникновения тех или иных форм и исходов различной патологии. Организационно-методическим проблемам изучения заболеваемости, в том числе и оценкам ее статистических показателей, посвящено большое количество специальных изданий по санитарной и медицинской статистике. Не останавливаясь на вопросах теории и практики решения общих проблем изучения заболеваемости, познакомимся со способами практической оценки достоверности показателей заболеваемости. Этот выбор объясняется большим числом такого рода задач и большим количеством ошибок, допускаемых в этой области применения прикладной медицинской статистики.

12.1. Типичные ошибки, допускаемые при анализе показателей заболеваемости

Предпосылки к возникновению такого рода ошибок нередко появляются уже на первых этапах статистического исследования показателей заболеваемости: при планировании и сборе исходных данных. В частности, когда из-за большой 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 часто приводят исследователей к существенным просчетам. Наиболее грубые просчеты, как правило, возникают при использовании данных о заболеваемости по обращаемости (общей заболеваемости, инфекционной заболеваемости, травматизме и т.п.). Ошибки здесь чаще всего возникают по трем причинам:

1. Непонимание разницы между абсолютным числом заболеваний и показателями заболеваемости, как явлением. *Например*, разность в абсолютном числе заболеваний «до того» и «после» может составлять 3 случая. Если «до того» этих заболеваний было всего 5, то такую разность можно считать существенной, когда обследовалось всего 20 человек. Но когда это произошло среди 10 тысяч человек – ситуация совершенно другая. Соответственно они (эти различия) стоят в этих ситуациях совершенно неодинаково, и оцениваться должны иначе. Поэтому применение методов статистики, допускающих обработку данных о заболеваемости, представленных только в абсолютных числах (*например*, хи-квадрат,) ограничивается относительно небольшим общим числом наблюдений (50-100 единиц). Для других ситуаций требуются методы, допускающие обработку производных величин, полученных в исходных совокупностях наблюдений существенно больших размеров.

2. Из-за искаженного представления или неполных (по содержательной части) данных. В статистике такую информацию называют информацией, содержащей систематические ошибки. Такого рода огрехи обычно возникают в результате неполной регистрации наблюдений, плохо налаженной работы управленческих и организационных структур, из-за желания приукрасить истинное положение вещей и т.п. По этому поводу существует известное высказывание: «Оценка труда врачей на основе обращаемости невозможна, поскольку эта обращаемость регулируется и учитывается самими врачами».

3. Из-за того, что полученные сплошным наблюдением, но на ограниченной территории (в пределах небольшого населенного пункта, городского или сельского района и т.п.) механически переносятся на более обширные территории. Таким образом, подразумевается некое выборочное исследование, когда по части судят о целом. Но такой подход тоже порождает *систематические ошибки* (см. раздел настоящего издания «Генеральная и выборочная совокупности»). Формально выборочным методом, как известно, повсеместно получают характеристики генеральной совокупности путем перенесения результатов изучения ограниченной по численности группы

наблюдений (выборки) на всю генеральную совокупность. Однако этот прием требует соблюдения определенных условий и правил, которые позволяют учесть вероятности ошибок выборок и, соответственно, дать им числовую характеристику (см раздел «Методы сплошного исследования»). Такие ошибки, называемые *ошибкам репрезентативности*, иначе говоря «представительностью выборочных данных», по своей природе не тождественны *систематическим* (механическим) ошибкам. *Систематические ошибки* наблюдения в отличие от *ошибок репрезентативности*, непредсказуемо искажают результат наблюдения. Поэтому распространять принцип суждения «по части о целом», игнорируя жесткие правила формирования выборок, недопустимо.

Известно, что в ряде ситуаций значения ошибок могут быть уменьшены увеличением объема выборок. Но систематические ошибки не могут быть устранены таким образом (поэтому они и называются «систематическими»), и попытки оправдать использование данных, содержащих систематические ошибки, большой численностью наблюдений несостоятельны.

В любом случае при использовании данных официальной статистики необходимо иметь в виду реальное положение дел на местах с регистрацией заболеваний. *Например*, при оценке распространенности заболеваний по обращаемости большое значение имеет учет таких местных особенностей, как доступность медицинской помощи (радиус обслуживания медучреждений, укомплектованность штатов медработников, наличие квалифицированных врачей-специалистов), близость крупных мегаполисов с развитой сетью специализированных медучреждений и т.п.

Другая группа ошибок связана с неправомерным применением методик, которые требуют обязательного учета того, как распределены отдельные наблюдения в группах, послуживших основой вычисления: альтернативно или нет, подчиняются закону Гауссова распределения или нет и т.п.

Например, известно, что для статистической оценки достоверности результатов выборочного исследования, в том числе и оценки различий

производных величин, для относительных величин и средних вычисляются значения ошибок репрезентативности (m). Здесь между статистическими коэффициентами (интенсивные, наглядности, соотношения и т.п.) и средними существует много общего. Недаром 100 лет назад относительные величины иногда называли средними. Принципиальное отличие средних от относительных величин (стат. коэффициентов) состоит в том, что последние, как правило, не сопровождаются параметрами распределения совокупностей, которые они характеризуют. А средние величины, например стандарты физического развития, биохимические показатели и т.д., представляются всегда с доверительными границами, значениями ошибок и т.п.

Наиболее просто вычислять ошибки относительных величин в случае, если они отражают так называемое альтернативное распределение. При этом подразумевается, что исходные совокупности соответствовали закону нормального распределения или были близки к нему. При этом, в любом случае рекомендуется соблюдать определенную осторожность: при значении итоговых оценок различий, близких к критическим, обязательно проводить повторную проверку результатов, используя другой, более адекватный метод.

12.2. Параметрические оценки различий показателей заболеваемости

12.2.1. Альтернативное распределение

Исходные данные о заболеваемости могут быть представлены в форме так называемых альтернативных распределений, которые широко используют в отечественной статистике при анализе практически любой заболеваемости. *Например*, интенсивные показатели частоты болевших и неболевших лиц, число летальных исходов болезней. Уровень онкологической заболеваемости, частота альтернативного события (заболел, не заболел раком). Поскольку методики определения различий такого рода показателей менее проблематичны, целесообразно проводить сводку первичных данных о любой заболеваемости таким образом, чтобы потом можно было использовать именно альтернативные показатели.

Рассмотрим пример оценки различий показателей заболеваемости с временной утратой трудоспособности (ЗВУТ). Основной проблемой здесь является то, что на практике наиболее широко используются небинарные показатели ЗВУТ: частота временной утраты трудоспособности в случаях и днях заболеваемости. Вычисляются они предельно просто: число случаев или дней заболеваний за год (или за другой период времени) делится на число работников и умножается на 100. Но за эту простоту приходится расплачиваться тем, что оценивать статистическую достоверность различий таких показателей традиционными методами параметрической статистики, в том числе и методами обработки бинарных данных, нельзя. Дело спасает то обстоятельство, что и оценка-то статистических различий таких показателей во многих случаях невозможна в принципе! Связано это с тем (а на практике это часто забывается), что данные о заболеваемости, например работников какого-либо предприятия или цеха, по своей сути – данные сплошного наблюдения о всей ЗВУТ всех работников этого предприятия или цеха. Так как здесь рассматриваются генеральные совокупности, т.е. совокупности, в которые включены все возможные единицы наблюдения (все работники предприятия, цеха и т.п., соответственно и все утраты их трудоспособности), статистическая оценка каких-либо различий неправомерна вообще, т.к. она проводится только в условиях выборочных исследований.

Другое дело, если данные о заболеваемости работников одного цеха рассматриваются как выборочные характеристики заболеваемости аналогичного, более крупного контингента. *Например*, по данным о заболеваемости работников одного цеха проводится оценивание заболеваемости работников группы других, аналогичных цехов предприятия и т.п. Или данные о ЗВУТ какого-либо контингента получены выборочным путем, при проведении углубленного научно-практического исследования. Здесь, разумеется, приходится использовать специальные статистические методики.

Рассмотренные ниже способы оценки различий можно использовать и при анализе различий большинства других показателей заболеваемости. При

углубленном изучении многих видов заболеваемости, в том числе и ЗВУТ, как уже отмечалось, зачастую можно уйти от решения проблемы оценки различий неальтернативных показателей заболеваемости, заменив их альтернативными. Например, на показатели болевших или не болевших в течение года лиц, которые обычно рассчитываются на основе численности «круглогодовых» контингентов (т.е. отработавших на момент исследования не менее 1 года). При таком подходе предусматривается возможность учета только одной альтернативы: работник болел (например, 2 раза) или не болел, 2 раза, в течение года. По сути дела предусматривается учет только одной из двух альтернатив: «да» или «нет».

Средняя ошибка для альтернативных показателей неболевших лиц может быть рассчитана по формуле:

$$m = \pm \sqrt{\frac{Pq}{n}}$$

где P – % не болевших лиц, исчисленный на постоянно (круглый год) работавших на предприятии лиц; $q = 100 - P$; n – число наблюдений (численность «круглогодовых» работников, заболеваемость которых изучается). Аналогичным образом вычисляется ошибка числа болевших лиц или числа болевших за год 1, 2, 3, 4 раза и более.

Например, в течение 1999 года на предприятии круглый год работали 1447 человек, из которых 827 человек имели случаи утраты трудоспособности в связи с ЗВУТ. Показатель болевших лиц:

$$P_0 = \frac{827 \times 100}{1447} = 57,1\%$$

Показатель неболевших лиц:

$$P_n = \frac{(1447 - 827) \times 100}{1447} = 42,9\% \text{ . Или проще } P_n = 100 - 51,7 = 42,9\%.$$

Величины соответствующих ошибок:

$$m_0 = \pm \sqrt{\frac{57,1 \times (100 - 57,1)}{1447}} = 1,3;$$

$$m_n = \pm \sqrt{\frac{42,9 \times (100 - 42,9)}{1447}} = 1,3.$$

Обратите внимание: ошибки равны, т.к. множители в числителе формулы при таком распределении одинаковы.

Величина средней ошибки, как известно, позволяет определить 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 и не быть учтены в данном исследовании⁸⁹.

В нашем примере показатель болевших лиц, т.е. лиц, имевших в течение года случаи ЗВУТ, равен 57,1%. На основании этого результата с вероятностью 68% (0,68) можно утверждать, что во всех аналогичных производственно-профессиональных группах рабочих (при одинаковом возрастно-половом составе), не болевших среди «круглогодных» рабочих, будет в пределах 57,1%±1,3%, т.е. от 55,8% до 58,4%.

Если же необходимо получить выводы, степень вероятности которых должна составлять 95,5% или 0,95 (обычная вероятность при медико-биологических исследованиях), то тогда показатели неболевших будут в интервале 57,1%±1,3%*2, то есть от 54,5% до 59,7%.

Для того, чтобы оценить статистическую достоверность разности между этими показателями (при условии соответствия распределения исходных данных нормальному, Гауссову, распределению), можно использовать параметрический критерий Стьюдента (t)⁹⁰:

$$t = \frac{P_1 - P_2}{\sqrt{m_1^2 + m_2^2}}$$

где P_1 – % болевших лиц одной группы наблюдений, P_2 – другой группы наблюдений. Соответственно, m_1 и m_2 – средние ошибки наблюдений одной и другой группы.

Другим способом оценки достоверности разности показателей, полученных из совокупностей с альтернативным, бинарным варьированием,

⁸⁹ Подробнее смотри в разделе настоящего издания «Интервальные оценки. Доверительная значимость, доверительная вероятность, доверительный интервал, доверительный предел».

⁹⁰ Подробнее – в разделе «Статистические критерии различия».

является оценка с помощью критерия Фишера по методу «фи», иногда называемому **точным методом Фишера**, или **методом углового преобразования Фишера** (Гублер Е.В., 1990). В основе этого метода лежит следующая особенность показателей, выраженных в процентах: если оценивается разность только двух возможных значений показателей: «есть» или «нет», то показатели (P в %), преобразованные по формуле $\varphi = 2 \arcsin \sqrt{P}$, распределяются близко к нормальному, Гауссову, распределению. В тех случаях, когда показатель в % будет менее 20% или более 80%, такой 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 ($F_{\text{преобр}}$) необходимо иметь две таблицы. Одна (Приложение 10) – таблица преобразований доли ($P\%$) в $\varphi = 2 \arcsin \sqrt{P}$, другая (Приложение 11) – таблица преобразованных значений критерия Фишера ($F_{\text{преобр}}$).

Вычисления производятся по формуле:

$$F_{\text{преобр}} = (\varphi_1 - \varphi_2)^2 \frac{n_1 n_2}{n_1 + n_2}$$

где $F_{\text{преобр}}$ – преобразованный критерий Фишера; φ_1 и φ_2 , – углы «фи» в радианах, которые находят по специальной таблице (Приложение 10); n – число наблюдений в 1 и 2-й группах.

Например, в обследованной профессиональной группе было 150 «круглогодичных» 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:

$$P_1 = \frac{4 \times 100}{150} = 2,7\% \quad \text{и} \quad P_2 = \frac{7 \times 100}{600} = 1,2\%$$

По таблице (Приложение 10) находим для 2,7% значение $\varphi_1 = 0,33$. Для 1,2% значение $\varphi_2 = 0,22$. Отсюда:

$$F_{\text{пробр}} = (0,33 - 0,22)^2 \frac{150 \cdot 600}{150 + 600} = 1,45$$

Этому значению согласно таблице критических значений (Приложение 11) соответствуют: для односторонней вероятности – $P=0,074$; для двухсторонней – $P=0,148$.

Указанные результаты свидетельствуют, что статистически значимых различий между двумя показателями нет. Для того, чтобы нулевую гипотезу о равенстве показателей отвергнуть, как известно, P не должно превышать 0,05, а в данном примере оно составило 0,148. Таким образом, в любом случае, как при односторонней оценке, так и при двухсторонней, исследуемые различия показателей частоты заболеваний хроническим бронхитом можно считать статистически несущественными.

С помощью метода углового преобразования можно определять и **доверительные границы показателей**. Для этого используются следующие формулы:

– для первого показателя нижняя доверительная граница:

$$\varphi_{\text{ни}} = \varphi_1 - F_{\text{пробр}} / \sqrt{n}$$

– для первого показателя верхняя доверительная граница:

$$\varphi_{\text{вн}} = \varphi_1 + F_{\text{пробр}} / \sqrt{n}$$

Аналогичным образом вычисляются верхняя и нижняя границы для второго показателя. В примере, условия которого указаны выше, решение задачи определения доверительных границ будет выглядеть следующим образом:

• Для доверительного уровня, который в медико-биологических исследованиях обычно принимается равным 95%, $F_{\text{пробр}}$ соответствует 1,96. Если принимается более высокий доверительный уровень – 99%, то $F_{\text{пробр}} = 2,56$. У показателя $P_1=2,7\%$, $\varphi_1 = 0,33$. Соответственно нижняя доверительная граница при $P=95\%$ $\varphi_{\text{ни}} = 0,33 - 1,96 / \sqrt{150} = 0,17$. По Приложению 10 находим соответствующее доверительное значение нижней границы

показателя $P_1=0,7\%$. Верхняя доверительная граница $\varphi_{n1} = 0,33 + 1,96 / \sqrt{150} = 0,49$. Соответствующее значение верхней границы показателя (по Приложению 10) – 5,9%.

• У показателя $P_2=1,2\%$, $\varphi_1 = 0,22$. Нижняя доверительная граница при доверительной вероятности 95% соответствует $\varphi_{n2} = 0,22 - 1,96 / \sqrt{600} = 0,06$. По Приложению 10 находим соответствующее доверительное значение нижней границы показателя – 0,09%. Верхняя доверительная граница $\varphi_{n2} = 0,22 + 1,96 / \sqrt{600} = 0,30$. Соответствующее доверительное значение верхней границы показателя = 2,3%.

12.2.2. Неальтернативное распределение

Особой проблемой в практической медицинской статистике является оценка значимости различий показателей заболеваемости, когда эти показатели получены из данных, не имеющих альтернативного распределения. К таким показателям 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 обострением 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 – среднее квадратическое (стандартное отклонение), а n – число наблюдений.

Расчеты по этой формуле, обоснованные с точки зрения теории статистики, невозможно осуществить, опираясь на большинство официальных статистических учетных форм заболеваемости, т.к. для получения необходимого аргумента функции σ необходимо строить вариационные ряды. Эта операция возможна только при полицейском учете всего наблюдаемого контингента. На практике осуществить такой учет можно только при проведении специальных исследований. Так, при изучении заболеваемости с

временной утратой трудоспособности необходимо получать параметры распределения, строя, например, взвешенные вариационные ряды. Вариантами (V) в таких рядах может быть кратность случаев ЗВУТ в связи с определенным заболеванием или по всем болезням вместе взятым, которые возникли в коллективе в течение года. Частотами (P) для вариантов будут числа лиц, утративших трудоспособность по этим причинам в течение года определенное число раз (0, 1, 2, 3 и т.д.). Например (табл. 139):

Таблица 139

Распределение случаев ЗВУТ работников предприятия в 2012 году

Кратность ЗВУТ в году (V)	Число рабочих (P)	VP	V^2P
0	29	0	0
1	22	22	22
2	14	28	56
3	13	39	117
4	2	8	32
Итого	80	97	227

Откуда заболеваемость в среднем на одного рабочего за год: $97/80=1,212$ случая. На 100 рабочих – 121,2 случая за год. Дисперсия:

$$D = \frac{\sum V^2 P_i}{\sum P_i} - M^2 = \frac{227}{80} - 1,212^2 = 1,367 \text{ случая.}$$

Среднеквадратическое отклонение: $\sigma = \sqrt{1,367} = 1,169$ случая.

Ошибка показателя: $m_1 = \frac{1,169}{\sqrt{80}} \approx 0,131$ на одного рабочего, или 13,1 случая 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 1 человека в год. В.Ю. Урбах (1967) считает, что эти формулы пригодны для оценки различий, когда вычисленное значение доверительного критерия (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 для других видов заболеваемости по обращаемости. Однако здесь нужно помнить, что использовать доверительный критерий Стьюдента t следует весьма осторожно. Это связано с тем, что t -критерий применим только в случае нормального распределения сравниваемых совокупностей. Поэтому проводить оценку различий показателей можно только после проверки полученных (эмпирических) распределений на их соответствие нормальному распределению, что особенно важно, когда значение доверительного критерия (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 временного интервала выборки, иначе говоря, 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 $3650/2500 \times 1000 = 1460$ случаев за год на тысячу

взрослого населения поселка АА. При пересчете за 1 календарный день (в абсолютных числах) получаем $3650/365=10$ случаев заболеваний; $10/2500=0,004$ 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:

$$-L = \frac{1}{2n} x^2 \left(2mt - \frac{1 + \alpha}{2} \right).$$

Верхняя граница:

$$+L = \frac{1}{2n} x^2 \left(2(m - 1) + \frac{1 + \alpha}{2} \right),$$

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 (рис. 136).



Рис. 136. Интервалы двух выборочных оценок заболеваемости

В случае а) различие выборочных оценок уровней заболеваемости достоверно; в случае б) – недостоверно.

Практические расчеты удобно проводить в MS Excel.

12.2.3. Анализ различий показателей заболеваемости в MS Excel с помощью метода доверительных интервалов

После ввода в блок ячеек A1:C2 исходных данных, как указано на рис. 137, необходимо:

1. рассчитать число случаев заболеваний на 1 день. Для этого в ячейку MS Excel B4 записываем функцию =B3/B2;
2. в ячейке B5 указываем необходимую доверительную вероятность (0,95 или 0,99);
3. в ячейку B6 – формулу: =1/(2*B2)*ХИ2ОБР((1-B5)/2;2*(B3+1));
4. в ячейку B7 вводим формулу: =1/(2*B2)*ХИ2ОБР((1+B5)/2;2*B3);
5. копируем содержимое блока ячеек B4:B7 в ячейки C4:C7.

	A	B	C
1	Показатели	Выборка 1	Выборка 2
2	Объем выборок	2500	3100
3	Число случаев заболеваний на 1 день	28	13
4	Уровень заболеваемости на 1 день	0,0112	0,00419
5	Доверительная вероятность	0,95	0,95
6	Верхняя граница заболеваемости	0,01619	0,00717
7	Нижняя граница заболеваемости	0,00744	0,00223
8			

Рис.137. Вывод исходных данных и результатов анализа различий показателей заболеваемости

Поскольку нижняя граница большего показателя (Выборка 1) меньше верхней границы меньшего показателя (Выборка 2), т.е. $0,00744 > 0,00717$, различия показателей следует признать статистически достоверными.

12.3. Непараметрические критерии оценки различий показателей заболеваемости

Непараметрические критерии позволяют обнаружить различие между сравниваемыми совокупностями, в чем бы это различие ни заключалось: в средних значениях или в 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, как правило, нельзя.

Существует большое число непараметрических критериев, каждый из которых имеет свои условия применения. Так, с помощью парного критерия Вилкоксона оценивают различия парносопряженных совокупностей. При его расчете принимают во внимание величины вариант, что позволяет расположить их в определенной последовательности или даже ранжировать с тем, чтобы получить соответствующие производные величины, которые приобретают свойства формальных критериев, т.е. инструментов оценивания существенности различий сравниваемых совокупностей. С помощью непараметрических коэффициентов корреляции рангов Спирмена и Кендалла 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.

12.3.1. Простейшие критерии быстрой оценки различий (критерий Мостеллера, Розенбаума, критерий знаков)

Эти критерии⁹¹ позволяют получать быстрые ориентировочные оценки статистической достоверности различий распределений нескольких показателей заболеваемости или других показателей, применяющихся при изучении здоровья. Данные могут быть представлены в виде абсолютных чисел или производных величин, вычисленных на относительно небольших, одинаковых по численности выборках ($n < 30$), выраженных в любых единицах измерения: случаи или длительность заболеваний, случаи рождений или смерти, параметры физического развития и т.п. Непременным условием здесь остается соблюдение фундаментальных требований любого статистического анализа: единство методов регистрации исходной информации, единство времени, например, данные о заболеваемости могут быть только за определенный по длительности период времени, сопоставимые единицы

⁹¹ Подробнее смотри в разделе «Сравнение распределений» настоящего издания.

сравнения и т.п. Кроме того, если сравниваются ряды производных величин (например, средних), то разброс значений отдельных наблюдений, на основании которых вычислена каждая из этих производных, должен быть незначителен.

Из достоинств указанных критериев вытекает их существенный недостаток: незначительная точность. Отсюда возникает важная особенность их практического применения: если итоги (любой, положительный или отрицательный) анализа различий сравниваемых рядов определены какой-нибудь одной единицей наблюдения, то целесообразно перепроверить эти результаты другим, более точным критерием.

Критерий Мостеллера. В качестве *примера* его применения определим достоверность различий показателей частоты случаев заболеваний с временной утратой трудоспособности хроническим бронхитом в течение года до оздоровления в профилактории и года после. Общие показатели частоты заболеваний составили: в первой выборке – 10,8, во второй — 9,3 случая на 100 обследованных за год (0,05 табл. 140). Результаты анализа этих различий с помощью критерия Мостеллера можно сформулировать следующим образом: рабочая гипотеза равенства частот заболеваний в двух выборок «До оздоровления» и «После оздоровления» отклоняется с вероятностью ($\alpha \geq 0,95$ или $P \leq 0,05$), поскольку 6 из 14 показателей выборки «После профилактория» (6,0; 4,9; 5,6; 6,8; 7,1; 7,1) превышают наименьший показатель выборки «До профилактория» (7,8).

Критерий Розенбаума. Вне размаха выборки (0,05 табл. 140) «До профилактория» (7,8-14,5) находится 6 значений выборки «После профилактория», что больше критического числа 5. Это позволяет отвергнуть гипотезу об одинаковых параметрах обеих выборок при $P \leq 0,05$.

Заболееваемость с временной утратой трудоспособности хроническим бронхитом лиц, проходивших оздоровление в профилактории (случаев заб. на 100 чел. за год)

Номера групп наблюдения	Случаев заболеваний в течение года		Направленность изменения
	До профилактория	После профилактория	
1	9,2	7,1	+
2	12,2	11,5	-
3	8,4	5,6	-
4	8,0	4,9	-
5	10,2	7,1	+
6	14,5	14,1	-
7	10,4	9,3	-
8	13,4	12,2	-
9	12,0	11,4	-
10	13,1	14,0	-
11	10,8	10,2	-
12	11,7	10,6	-
13	7,8	6,0	+
14	8,4	6,8	-
В среднем	10,8	9,3	

Критерий знаков (Z). С помощью этого критерия оценивают направленность изменений показателей заболеваемости в сравнимых взаимосвязанных парах вариантов. Достоверность определяют при подсчете числа однонаправленных эффектов. В экспериментальных и клинических исследованиях этот критерий обычно применяют в случаях, когда анализу подвергают сопряженные между собой варианты, полученные на одной совокупности.

В приведенном примере (0,05 табл. 140) из 14 групп в 11 отмечено снижение частоты случаев временной утраты трудоспособности из-за хронического бронхита. Меньшее число раз встречался рост заболеваемости – 3 раза. Число степеней свободы – $14-1=13$. По таблице критических значений (Приложение 6) находим, что для числа степеней свободы 13 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 подтверждается.

Обратите внимание! Критерий знаков устанавливает статистическую значимость различий, но не величину этих различий.

Критерий Уайта (K) – 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 табл. 141.

Таблица 141

Сравнительная заболеваемость жителей районов А и В
(случаев заб. на 1000 чел. за год)

Год	Район А		Район В	
	Заболеваемость	Ранг R_1	Заболеваемость	Ранг R_2
1992	1061,3	6	1071,3	7
1993	1075,2	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
-	$n_1=7$	$\sum R_1=49$	$n_2=7$	$\sum R_2=56$

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 (Приложение 12). Для района А сумма рангов $\sum R_1=49$, района 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. Меньшая из найденных сумм равна 49, что превышает критическое значение критерия Уайта. По условиям применения критерия Уайта, если меньшая сумма рангов превышает его критическое значение, то имеющиеся различия в распределении показателей заболеваемости нельзя признать статистически достоверными.

12.3.2. Критерий соответствия (согласия) Пирсона χ^2

Является одним из самых распространенных в практике статистических исследований непараметрических критериев. Основываясь на оценку распределений, этот критерий может применяться для сравнительной оценки показателей заболеваемости, выраженных в абсолютных числах и распределенных по группам наблюдений (по возрасту, по времени и т.п.).

Методика его применения для решения такого рода задач детально описана десятки лет назад (Мерков А.М., Поляков Л.Е., 1974), но из-за необходимости учета численности больных в абсолютных числах в практической статистике используется относительно редко. В основе применения критерия χ^2 лежит анализ соответствия фактического и ожидаемого распределения показателей заболеваемости. При этом общий для сравниваемых групп показатель становится основой для нахождения ожидаемого распределения показателя заболеваемости в каждой из этих групп (возрастной, половой, и т.п.).

1. Как уже отмечалось, оценки различий показателей заболеваемости строятся на основе исходных данных, представленных абсолютными числами. Однако по ходу вычисления, опираясь на эти абсолютные числа, можно вычислять практически любые производные величины. Рассмотрим порядок оценки показателей заболеваемости. В данном примере (табл. 142) порядок вычисления следующий:

2. Вычисляются показатели частоты случаев заболеваемости на 100 рабочих.

3. Определяются «ожидаемые» числа P . Для возрастной группы «до 30 лет» число обследованных работников – 55 человек. Например, если бы заболеваемость в группе «до 30 лет» была такая же, как и в целом по всем обследованным (см. группу «Все»), то численность болевших $P_{до\ 30}$ составила бы $55 \times 23,8/100 = 13$ человек.

4. Затем получают разность фактических и «ожидаемых» чисел для возрастной группы «до 30 лет»: $11-13,1=-2,1$. Каждую такую разность возводят в квадрат и делят на P_j : $(-2,1^2/13,1)=0,33$.

Таблица 142

Оценка различий повозрастных показателей заболеваемости с помощью критерия χ^2

Вычисляемые данные	Возрастные группы (лет)				Все
	До 30	30-39	40-49	50 и ст.	
Число работников, абс.	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_i)	13,1	14,3	28,6	11,2	—
$(P-P_i)$	-2,1	2,7	4,4	-5,2	—
$(P-P_i)^2$	4,4	7,4	19,7	26,9	—
$(P-P_i)^2/P_i$	0,33	0,52	0,69	2,40	3,94

Полученные таким образом частные в других возрастных группах суммируют. Эта сумма и будет вычисленным значением критерия $\chi^2=3,94$.

С помощью функций *Microsoft Excel* или по таблице критических значений χ^2 определяют значимость результата. В данном 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 7,81 (Приложение 8). При уровне значимости 0,01 – 11,34. Поскольку вычисленное значение критерия меньше указанных, следует принять нулевую гипотезу, говорящую об отсутствии различий в распределении анализируемых показателей. С помощью функции *MS Excel* **ХИ2РАСП** можно определить конкретную одностороннюю вероятность $P=0,6997$.

Аналогичные вычисления проводятся, если заболеваемость анализируется на основе других показателей заболеваемости. Например, можно использовать процент неболевших 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 и другие варианты методики оценок различий показателей заболеваемости с помощью критерия 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. 143).

Таблица 143

**Оценка различий обращаемости по поводу ИБС
в населенном пункте N за первое полугодие**

Месяц (j)	Число обращений, P	$P - P_i$	$(P - P_i)^2$	$(P - P_i)^2 / P_i$
1	17	1	1	0,063
2	15	-1	1	0,063
3	18	2	4	0,250
4	14	-2	4	0,250
5	16	0	0	0,000
Всего	80	—	—	0,626
P	$80/5=16$			

Поскольку вычисленное значение критерия $\chi = 0,625$ при числе степеней свободы $5-1=4$ значительно меньше критического (определяется с помощью таблицы критических значений χ^2 ⁹² 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.

⁹² Смотри Приложение 8.

Медианный критерий с использованием статистики χ^2 – простой в практическом применении. Позволяет получать быстрые приближенные оценки нескольких распределений любых величин, в том числе и показателей заболеваемости, в нескольких группах, которые могут быть неравными по численности показателей. Если оценки этого критерия находятся на уровне, близком к критическому, то рекомендуется использовать более точные методы.

Формула его расчета выглядит следующим образом:

$$\chi^2 = 4 \sum_{i=1}^k \frac{1}{n_i} m_i^2 - N,$$

где N – объединенное число всех наблюдений, по всем группам обследуемой совокупности. Единицей наблюдения считается отдельно взятый показатель, будь то производная величина или другой какой-либо численный результат наблюдений; n_j – число наблюдений в j группе; m_j – число наблюдений в каждой j группе, превосходящих значение медианы, определенной для всей совокупности наблюдений; k – число групп наблюдений. Число степеней свободы для статистики χ^2 определяется, как $k-1$.

Например: требуется сравнить заболеваемость населения 3-х районов области за период 1992-1998 гг. (табл.144)⁹³.

Таблица 144

Сравнительная заболеваемость населения районов А, Б и В области (случаев на 1000 чел. за год)

Годы наблюдения	Районы области		
	А	Б	В
1992	1061,3	1071,3	1008,3
1993	1076,2	1082,2	1015,2
1994	1022,9	1102,9	1024,7
1995	1038,2	1073,4	1043,6
1996	1077,4	1001,9	1076,7
1997	1086,3	1016,1	1038,9
1998	1049,8	1099,3	1086,3

⁹³ Обратить внимание! В примере оцениваются не распределения чисел заболеваний, а распределения показателей явления – заболеваемости.

Последовательность вычислений следующая:

1. Определяется медиана для всех наблюдений. В настоящем примере это 1061,3 случая заболеваний на 10000 чел. за год. Найти этот показатель можно с помощью функции *MS Excel* **МЕДИАНА**.

В каждом столбце таблицы определяется число показателей. В данном случае оно одинаково (табл. 144) – 7,7 и 7.

2. В каждом столбце определяем число показателей, превосходящих медиану. В первом (район А) таких значений было 3, во втором (район Б) – 5 и в третьем (район В) – 2.

3. Для каждого столбца таблицы находим квадраты этих чисел и умножаем на число наблюдений в каждом столбце. Для первого столбца $(3*3)/7=1,29$; для второго $(5*5)/7=3,57$; для третьего $(2*2)/7=0,57$.

4. Суммируем эти значения: $1,29+3,57+0,57=5,43$. Умножаем эту сумму на 4 (константа функции) и вычитаем объединенное число всех наблюдений: $5,43*4-21=0,74$.

5. Оцениваем результат по таблице критических значений критерия χ^2 (Приложение 8). Для степени свободы $3-1=2$ критическое значение, удовлетворяющее условию $P \leq 0,05$, должно быть не менее 5,99. С помощью функции *MS Excel* **ХИ2РАСП** можно определить конкретную одностороннюю вероятность $P=0,6997$.

Таким образом, нулевая гипотеза, свидетельствующая об отсутствии различий, не отвергается, т.е. статистически значимых различий распределений показателей заболеваемости в сравниваемых группах (районы А, Б и В) не выявлено.

ПРИЛОЖЕНИЯ

Приложение 1

Граничные значения τ_{α} при $P=0,001$

n	τ_{α}	n	τ_{α}	n	τ_{α}
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

Приложение 2

Значения коэффициента f для исключения
выскакивающих вариант (по В.И. Романовскому)

N	0,05	0,02	0,01	0,001
3	4,969	8,042	11,460	36,486
4	3,558	5,077	6,530	14,468
5	3,041	4,105	5,043	9,432
6	2,777	3,635	4,355	7,409
7	2,616	3,360	3,963	6,370
8	2,508	3,180	3,711	5,733
9	2,431	3,053	3,536	5,314
10	2,372	2,959	3,409	5,014
11	3,327	2,887	3,310	4,791
12	2,291	2,829	3,233	4,618
13	2,261	2,782	3,170	4,481
14	2,236	2,743	3,118	4,369
15	2,215	2,710	3,075	4,276
16	2,197	2,683	3,038	4,198
17	2,181	2,658	3,006	4,131
18	2,168	2,637	2,997	4,074

N	0,05	0,02	0,01	0,001
19	2,156	2,618	2,953	4,024
20	2,145	2,602	2,932	3,979
20	2,145	2,602	2,932	3,979
21	2,135	2,587	2,912	3,941
22	2,127	2,575	2,895	3,905
23	2,119	2,562	2,880	3,874
24	2,112	2,552	2,865	3,845
25	2,105	2,541	2,852	3,819
26	2,099	2,532	2,840	3,796
27	2,094	2,524	2,830	3,775
28	2,088	2,517	2,820	3,755
29	2,083	2,509	2,810	3,737
30	2,079	2,503	2,802	3,719
40	2,048	2,456	2,742	3,602
60	2,018	2,411	2,683	3,402
120	1,988	2,368	2,628	3,388

Значения $F_{критич}$ при $P=0,05$

Число степеней свободы	1	2	3	4	5	8	12	24
1	161,45	199,50	215,71	224,58	230,16	238,88	243,91	249,05
2	18,51	19,00	19,16	19,25	19,30	19,37	19,41	19,45
3	10,13	9,55	9,28	9,12	9,01	8,85	8,74	8,64
4	7,71	6,94	6,59	6,39	6,26	6,04	5,91	5,77
5	6,61	5,79	5,41	5,19	5,05	4,82	4,68	4,53
6	5,99	5,14	4,76	4,53	4,39	4,15	4,00	3,84
7	5,59	4,74	4,35	4,12	3,97	3,73	3,57	3,41
8	5,32	4,46	4,07	3,84	3,69	3,44	3,28	3,12
9	5,12	4,26	3,86	3,63	3,48	3,23	3,07	2,90
10	4,96	4,10	3,71	3,48	3,33	3,07	2,91	2,74
11	4,84	3,98	3,59	3,36	3,20	2,95	2,79	2,61
12	4,75	3,89	3,49	3,26	3,11	2,85	2,69	2,51
13	4,67	3,81	3,41	3,18	3,03	2,77	2,60	2,42
14	4,60	3,74	3,34	3,11	2,96	2,70	2,53	2,35
15	4,54	3,68	3,29	3,06	2,90	2,64	2,48	2,29
16	4,49	3,63	3,24	3,01	2,85	2,59	2,42	2,24
17	4,45	3,59	3,20	2,96	2,81	2,55	2,38	2,19
18	4,41	3,55	3,16	2,93	2,77	2,51	2,34	2,15
19	4,38	3,52	3,13	2,90	2,74	2,48	2,31	2,11
20	4,35	3,49	3,10	2,87	2,71	2,45	2,28	2,08

Значения $F_{критич}$ при $P=0,01$

Число степеней свободы	1	2	3	4	5	8	12	24
1	4052,18	4999,50	5403,35	5624,58	5763,65	5981,07	6106,32	6234,63
2	98,50	99,00	99,17	99,25	99,30	99,37	99,42	99,46
3	34,12	30,82	29,46	28,71	28,24	27,49	27,05	26,60
4	21,20	18,00	16,69	15,98	15,52	14,80	14,37	13,93
5	16,26	13,27	12,06	11,39	10,97	10,29	9,89	9,47
6	13,75	10,92	9,78	9,15	8,75	8,10	7,72	7,31
7	12,25	9,55	8,45	7,85	7,46	6,84	6,47	6,07
8	11,26	8,65	7,59	7,01	6,63	6,03	5,67	5,28
9	10,56	8,02	6,99	6,42	6,06	5,47	5,11	4,73
10	10,04	7,56	6,55	5,99	5,64	5,06	4,71	4,33
11	9,65	7,21	6,22	5,67	5,32	4,74	4,40	4,02
12	9,33	6,93	5,95	5,41	5,06	4,50	4,16	3,78
13	9,07	6,70	5,74	5,21	4,86	4,30	3,96	3,59
14	8,86	6,51	5,56	5,04	4,69	4,14	3,80	3,43
15	8,68	6,36	5,42	4,89	4,56	4,00	3,67	3,29
16	8,53	6,23	5,29	4,77	4,44	3,89	3,55	3,18
17	8,40	6,11	5,18	4,67	4,34	3,79	3,46	3,08
18	8,29	6,01	5,09	4,58	4,25	3,71	3,37	3,00
19	8,18	5,93	5,01	4,50	4,17	3,63	3,30	2,92
20	8,10	5,85	4,94	4,43	4,10	3,56	3,23	2,86

Критические значения модифицированного критерия Стьюдента (Q)

k	n								
	2	3	4	5	6	7	8	9	10
P=0,05									
1	18,10	7,60	6,20	5,80	5,70	5,70	5,80	5,90	6,00
2	3,50	2,50	2,30	2,30	2,40	2,40	2,50	2,60	2,60
3	1,79	1,44	1,41	1,43	1,47	1,52	1,57	1,62	1,68
4	1,18	1,01	1,01	1,03	1,07	1,11	1,14	1,19	1,23
5	0,89	0,78	0,78	0,81	0,84	0,87	0,91	0,94	0,97
6	0,71	0,64	0,64	0,66	0,69	0,71	0,74	0,77	0,80
7	0,59	0,53	0,54	0,56	0,59	0,61	0,64	0,66	0,68
8	0,50	0,46	0,47	0,49	0,51	0,53	0,56	0,58	0,60
9	0,44	0,41	0,42	0,43	0,45	0,47	0,49	0,51	0,53
10	0,39	0,36	0,37	0,39	0,40	0,42	0,44	0,46	0,47
P=0,01									
1	90,00	17,50	11,20	9,30	8,60	8,20	8,00	8,00	8,00
2	8,30	4,00	3,40	3,20	3,10	3,10	3,10	3,20	3,20
3	3,40	2,10	1,91	1,87	1,87	1,90	1,93	1,98	2,00
4	2,00	1,42	1,33	1,32	1,33	1,36	1,39	1,43	1,47
5	1,43	1,07	1,01	1,01	1,03	1,06	1,09	1,12	1,15
6	1,09	0,86	0,82	0,82	0,84	0,86	0,89	0,92	0,95
7	0,89	0,71	0,69	0,70	0,72	0,74	0,76	0,78	0,80
8	0,74	0,61	0,60	0,60	0,62	0,64	0,66	0,68	0,70
9	0,64	0,53	0,52	0,53	0,55	0,56	0,58	0,60	0,62
10	0,56	0,47	0,47	0,47	0,49	0,50	0,52	0,54	0,55

Критические значения двустороннего критерия Стьюдента t

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
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	2,00	2,66	91	1,66	1,99	2,63
72	1,67	2,00	2,66	92	1,66	1,99	2,63

Критические значения Z-числа (n) реже встречающихся знаков * (по В.Ю. Урбаху, 1963)

n \ P	0,05	0,01	n \ P	0,05	0,01	n \ P	0,05	0,01
7	1		23	7	5	34	11	10
8	1	1	24	7	6	35-36	12	10
9-11	2	1	25	8	6	37-38	13	11
12-14	3	2	26-27	8	7	39	13	12
15-16	4	3	28	9	7	40-41	14	12
17	5	3	29	9	8	42-43	15	13
18-19	5	4	30-31	10	8	44-46	16	14
20	6	4	32	10	9	47-48	17	15
21-22	6	5	33	11	9	49-50	18	16

**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							
	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							
	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

Критические значения одностороннего критерия χ^2 (хи-квадрат)

Число степеней свободы	Уровни значимости		
	$P=0,05$	$P=0,01$	$P=0,001$
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,12
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,72	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,31

Число степеней свободы	Уровни значимости		
	$P=0,05$	$P=0,01$	$P=0,001$
21	32,67	38,93	46,80
22	33,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70
31	44,99	52,19	61,10
32	46,19	53,49	62,49
33	47,40	54,78	63,87
34	48,60	56,06	65,25
35	49,80	57,34	66,62
36	51,00	58,62	67,99
37	52,19	59,89	69,35
38	53,38	61,16	70,70
39	54,57	62,43	72,05
40	55,76	63,69	73,40

ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

Минимальные значения коэффициентов двухсторонней корреляции,
при которых связь между двумя рядами наблюдений
можно считать значимой с уровнем надежности P ,

N – число сравниваемых пар

(по Л.Н. Большову и Н.В. Смирнову, 1965)

$N \setminus P$	0,05	0,025	0,01	0,005	0,003	0,001
4	0,729	0,811	0,822	0,917	0,941	0,974
5	0,669	0,754	0,833	0,875	0,905	0,950
6	0,621	0,707	0,789	0,834	0,870	0,924
7	0,582	0,666	0,750	0,798	0,836	0,898
8	0,549	0,632	0,713	0,765	0,805	0,872
9	0,521	0,602	0,685	0,735	0,776	0,847
10	0,497	0,576	0,658	0,708	0,730	0,823
11	0,476	0,553	0,634	0,684	0,726	0,801
12	0,457	0,532	0,612	0,661	0,703	0,780
13	0,441	0,514	0,592	0,641	0,683	0,760
14	0,426	0,497	0,574	0,623	0,664	0,742
15	0,412	0,482	0,558	0,606	0,647	0,725
16	0,400	0,468	0,543	0,590	0,631	0,708
17	0,389	0,456	0,529	0,573	0,616	0,693
18	0,378	0,444	0,516	0,561	0,602	0,679
19	0,369	0,433	0,503	0,549	0,589	0,665
20	0,360	0,423	0,492	0,537	0,576	0,652
25	0,323	0,381	0,445	0,487	0,524	0,597
30	0,269	0,349	0,409	0,449	0,484	0,554
35	0,275	0,325	0,331	0,418	0,452	0,519
40	0,257	0,304	0,358	0,393	0,425	0,490
45	0,243	0,288	0,338	0,372	0,403	0,463
50	0,231	0,273	0,322	0,354	0,384	0,443
60	0,211	0,250	0,295	0,325	0,352	0,408
70	0,195	0,232	0,274	0,302	0,327	0,380
80	0,183	0,217	0,257	0,283	0,307	0,357
90	0,173	0,205	0,242	0,267	0,290	0,338
100	0,164	0,195	0,230	0,254	0,276	0,321

Преобразованные доли выборки (P – долям от 1,0, например, 20%=0,20):

$$\varphi = \arcsin \sqrt{P}$$

Доли выбор- ки	Последний знак в дробном числе доли									
	0	1	2	3	4	5	6	7	8	9
0,000	0,000	0,020	0,028	0,035	0,040	0,045	0,049	0,053	0,057	0,060
0,001	0,063	0,066	0,069	0,072	0,075	0,077	0,080	0,082	0,085	0,087
0,002	0,089	0,092	0,094	0,096	0,098	0,100	0,102	0,104	0,106	0,108
0,003	0,110	0,111	0,113	0,115	0,117	0,118	0,120	0,122	0,123	0,125
0,004	0,127	0,128	0,130	0,131	0,133	0,134	0,136	0,137	0,139	0,140
0,005	0,142	0,143	0,144	0,146	0,147	0,148	0,150	0,151	0,152	0,154
0,006	0,155	0,156	0,158	0,159	0,160	0,161	0,163	0,164	0,165	0,166
0,007	0,168	0,169	0,170	0,171	0,172	0,173	0,175	0,176	0,177	0,178
0,008	0,179	0,180	0,181	0,182	0,184	0,185	0,186	0,187	0,188	0,189
0,009	0,190	0,191	0,192	0,193	0,194	0,195	0,196	0,197	0,198	0,199
0,01	0,200	0,210	0,220	0,229	0,237	0,246	0,254	0,262	0,269	0,277
0,02	0,284	0,291	0,298	0,304	0,311	0,318	0,324	0,330	0,336	0,342
0,03	0,348	0,354	0,360	0,365	0,371	0,376	0,382	0,387	0,392	0,398
0,04	0,403	0,408	0,413	0,418	0,423	0,428	0,432	0,437	0,442	0,446
0,05	0,451	0,456	0,460	0,465	0,469	0,473	0,478	0,482	0,486	0,491
0,06	0,495	0,499	0,503	0,507	0,512	0,516	0,520	0,524	0,528	0,532
0,07	0,536	0,539	0,543	0,547	0,551	0,555	0,559	0,562	0,566	0,570
0,08	0,574	0,577	0,581	0,584	0,588	0,592	0,595	0,599	0,602	0,606
0,09	0,609	0,613	0,616	0,620	0,623	0,627	0,630	0,633	0,637	0,640
0,10	0,644	0,647	0,650	0,653	0,657	0,660	0,663	0,666	0,670	0,673
0,11	0,676	0,679	0,682	0,686	0,689	0,692	0,695	0,698	0,701	0,704
0,12	0,707	0,711	0,714	0,717	0,720	0,723	0,726	0,729	0,732	0,735
0,13	0,738	0,741	0,744	0,747	0,750	0,752	0,755	0,758	0,761	0,764
0,14	0,767	0,770	0,773	0,776	0,778	0,781	0,784	0,787	0,790	0,793
0,15	0,795	0,798	0,801	0,804	0,807	0,809	0,812	0,815	0,818	0,820
0,16	0,823	0,826	0,828	0,831	0,834	0,837	0,839	0,842	0,845	0,847
0,17	0,850	0,853	0,855	0,858	0,861	0,863	0,866	0,868	0,871	0,874
0,18	0,876	0,879	0,881	0,884	0,887	0,889	0,892	0,894	0,897	0,900
0,19	0,902	0,905	0,907	0,910	0,912	0,915	0,917	0,920	0,922	0,925
0,20	0,927	0,930	0,932	0,935	0,937	0,940	0,942	0,945	0,947	0,950
0,21	0,952	0,955	0,957	0,959	0,962	0,964	0,967	0,969	0,972	0,974
0,22	0,976	0,979	0,981	0,984	0,986	0,988	0,991	0,993	0,996	0,998
0,23	1,000	1,003	1,005	1,007	1,010	1,012	1,015	1,017	1,019	1,022
0,24	1,024	1,026	1,029	1,031	1,033	1,036	1,038	1,040	1,043	1,045
0,25	1,047	1,050	1,052	1,054	1,056	1,059	1,061	1,063	1,066	1,068
0,26	1,070	1,072	1,075	1,077	1,079	1,082	1,084	1,086	1,088	1,091
0,27	1,093	1,095	1,097	1,100	1,102	1,104	1,106	1,109	1,111	1,113
0,28	1,115	1,117	1,120	1,122	1,124	1,126	1,129	1,131	1,133	1,135
0,29	1,137	1,140	1,142	1,144	1,146	1,148	1,151	1,153	1,155	1,157
0,30	1,159	1,161	1,164	1,166	1,168	1,170	1,172	1,175	1,177	1,179

Приложение 10 (продолжение)

Преобразованные доли выборки ($P\%$ = долям от 1,0, например, 20%=0,20):

$$\varphi = 2 \arcsin \sqrt{P}$$

Доли выбор- ки	Последний знак в дробном числе доли									
	0	1	2	3	4	5	6	7	8	9
0,31	1,181	1,183	1,185	1,187	1,190	1,192	1,194	1,196	1,198	1,200
0,32	1,203	1,205	1,207	1,209	1,211	1,213	1,215	1,217	1,220	1,222
0,33	1,224	1,226	1,228	1,230	1,232	1,234	1,237	1,239	1,241	1,243
0,34	1,245	1,247	1,249	1,251	1,253	1,256	1,258	1,260	1,262	1,264
0,35	1,266	1,268	1,270	1,272	1,274	1,277	1,279	1,281	1,283	1,285
0,36	1,287	1,289	1,291	1,293	1,295	1,297	1,299	1,302	1,304	1,306
0,37	1,308	1,310	1,312	1,314	1,316	1,318	1,320	1,322	1,324	1,326
0,38	1,328	1,330	1,333	1,335	1,337	1,339	1,341	1,343	1,345	1,347
0,39	1,349	1,351	1,353	1,355	1,357	1,359	1,361	1,363	1,365	1,367
0,40	1,369	1,371	1,374	1,376	1,378	1,380	1,382	1,384	1,386	1,388
0,41	1,390	1,392	1,394	1,396	1,398	1,400	1,402	1,404	1,406	1,408
0,42	1,410	1,412	1,414	1,416	1,418	1,420	1,422	1,424	1,426	1,428
0,43	1,430	1,432	1,434	1,436	1,438	1,440	1,442	1,444	1,446	1,448
0,44	1,451	1,453	1,455	1,457	1,459	1,461	1,463	1,465	1,467	1,469
0,45	1,471	1,473	1,475	1,477	1,479	1,481	1,483	1,485	1,487	1,489
0,46	1,491	1,493	1,495	1,497	1,499	1,501	1,503	1,505	1,507	1,509
0,47	1,511	1,513	1,515	1,517	1,519	1,521	1,523	1,525	1,527	1,529
0,48	1,531	1,533	1,535	1,537	1,539	1,541	1,543	1,545	1,547	1,549
0,49	1,551	1,553	1,555	1,557	1,559	1,561	1,563	1,565	1,567	1,569
0,50	1,571	1,573	1,575	1,577	1,579	1,581	1,583	1,585	1,587	1,589
0,51	1,591	1,593	1,595	1,597	1,599	1,601	1,603	1,605	1,607	1,609
0,52	1,611	1,613	1,615	1,617	1,619	1,621	1,623	1,625	1,627	1,629
0,53	1,631	1,633	1,635	1,637	1,639	1,641	1,643	1,645	1,647	1,649
0,54	1,651	1,653	1,655	1,657	1,659	1,661	1,663	1,665	1,667	1,669
0,55	1,671	1,673	1,675	1,677	1,679	1,681	1,683	1,685	1,687	1,689
0,56	1,691	1,693	1,695	1,697	1,699	1,701	1,703	1,705	1,707	1,709
0,57	1,711	1,713	1,715	1,717	1,719	1,721	1,723	1,725	1,727	1,729
0,58	1,731	1,734	1,736	1,738	1,740	1,742	1,744	1,746	1,748	1,750
0,59	1,752	1,754	1,756	1,758	1,760	1,762	1,764	1,766	1,768	1,770
0,60	1,772	1,774	1,776	1,778	1,780	1,782	1,784	1,786	1,789	1,791
0,61	1,793	1,795	1,797	1,799	1,801	1,803	1,805	1,807	1,809	1,811
0,62	1,813	1,815	1,817	1,819	1,821	1,823	1,826	1,828	1,830	1,832
0,63	1,834	1,836	1,838	1,840	1,842	1,844	1,846	1,848	1,850	1,853
0,64	1,855	1,857	1,859	1,861	1,863	1,865	1,867	1,869	1,871	1,873
0,65	1,875	1,878	1,880	1,882	1,884	1,886	1,888	1,890	1,892	1,894
0,66	1,897	1,899	1,901	1,903	1,905	1,907	1,909	1,911	1,913	1,916
0,67	1,918	1,920	1,922	1,924	1,926	1,928	1,931	1,933	1,935	1,937
0,68	1,939	1,941	1,943	1,946	1,948	1,950	1,952	1,954	1,956	1,958
0,69	1,961	1,963	1,965	1,967	1,969	1,971	1,974	1,976	1,978	1,980
0,70	1,982	1,984	1,987	1,989	1,991	1,993	1,995	1,998	2,000	2,002

Преобразованные доли выборки ($P\%$ = долям от 1,0, например, 20%=0,20):

$$\Phi = \text{zarc sin } \sqrt{P}$$

Доли выборки	Последний знак в дробном числе выборки									
	0	1	2	3	4	5	6	7	8	9
0,71	2,004	2,006	2,009	2,011	2,013	2,015	2,018	2,020	2,022	2,024
0,72	2,026	2,029	2,031	2,033	2,035	2,038	2,040	2,042	2,044	2,047
0,73	2,049	2,051	2,053	2,056	2,058	2,060	2,062	2,065	2,067	2,069
0,74	2,071	2,074	2,076	2,078	2,081	2,083	2,085	2,087	2,090	2,092
0,75	2,094	2,097	2,099	2,101	2,104	2,106	2,108	2,111	2,113	2,115
0,76	2,118	2,120	2,122	2,125	2,127	2,129	2,132	2,134	2,136	2,139
0,77	2,141	2,144	2,146	2,148	2,151	2,153	2,156	2,158	2,160	2,163
0,78	2,165	2,168	2,170	2,172	2,175	2,177	2,180	2,182	2,185	2,187
0,79	2,190	2,192	2,194	2,197	2,199	2,202	2,204	2,207	2,209	2,212
0,80	2,214	2,217	2,219	2,222	2,224	2,227	2,229	2,232	2,234	2,237
0,81	2,240	2,242	2,245	2,247	2,250	2,252	2,255	2,258	2,260	2,263
0,82	2,265	2,268	2,271	2,273	2,276	2,278	2,281	2,284	2,286	2,289
0,83	2,292	2,294	2,297	2,300	2,302	2,305	2,308	2,310	2,313	2,316
0,84	2,319	2,321	2,324	2,327	2,330	2,332	2,335	2,338	2,341	2,343
0,85	2,346	2,349	2,352	2,355	2,357	2,360	2,363	2,366	2,369	2,372
0,86	2,375	2,377	2,380	2,383	2,386	2,389	2,392	2,395	2,398	2,401
0,87	2,404	2,407	2,410	2,413	2,416	2,419	2,422	2,425	2,428	2,431
0,88	2,434	2,437	2,440	2,443	2,447	2,450	2,453	2,456	2,459	2,462
0,89	2,465	2,469	2,472	2,475	2,478	2,482	2,485	2,488	2,491	2,495
0,90	2,498	2,501	2,505	2,508	2,512	2,515	2,518	2,522	2,525	2,529
0,91	2,532	2,536	2,539	2,543	2,546	2,550	2,553	2,557	2,561	2,564
0,92	2,568	2,572	2,575	2,579	2,583	2,587	2,591	2,594	2,598	2,602
0,93	2,606	2,610	2,614	2,618	2,622	2,626	2,630	2,634	2,638	2,642
0,94	2,647	2,651	2,655	2,659	2,664	2,668	2,673	2,677	2,681	2,686
0,95	2,691	2,695	2,700	2,705	2,709	2,714	2,719	2,724	2,729	2,734
0,96	2,739	2,744	2,749	2,754	2,760	2,765	2,771	2,776	2,782	2,788
0,97	2,793	2,799	2,805	2,811	2,818	2,824	2,831	2,837	2,844	2,851
0,98	2,858	2,865	2,872	2,880	2,888	2,896	2,904	2,913	2,922	2,931
0,99	2,941	2,952	2,962	2,974	2,987	3,000	3,015	3,032	3,052	3,078

Определение значимости P по известному $F_{гробр}$
(по Е.В. Гублеру, 1990)

P равно или меньше	P равно или больше (последний знак)									
	0	1	2	3	4	5	6	7	8	9
0,00		2,91	2,81	2,70	2,62	2,55	2,49	2,44	2,39	2,35
0,01	2,31	2,28	2,25	2,22	2,19	2,16	2,14	2,11	2,09	2,07
0,02	20,5	2,03	2,01	1,90	1,97	1,96	1,94	1,92	1,91	1,89
0,03	1,88	1,86	1,85	1,84	1,82	1,81	1,80	1,79	1,77	1,76
0,04	1,75	1,74	1,73	1,72	1,71	1,70	1,68	1,67	1,66	1,65
0,05	1,64	1,64	1,63	1,62	1,61	1,60	1,59	1,58	1,57	1,56
0,06	1,56	1,55	1,54	1,53	1,52	1,52	1,51	1,50	1,49	1,48
0,07	1,48	1,47	1,46	1,46	1,45	1,44	1,43	1,43	1,42	1,41
0,08	1,41	1,40	1,39	1,39	1,38	1,37	1,37	1,36	1,36	1,35
0,09	1,34	1,34	1,33	1,32	1,32	1,31	1,31	1,30	1,30	1,29
0,10	1,29									

Критические значения критерия Уайта (K)

п/п	4		5		6		7		8		9		10	
	0,05	0,01	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

СПИСОК ЛИТЕРАТУРЫ

1. Алехнович М.В. К анализу контингентов длительно и часто болеющих / М.В. Алехнович (и др.) // Сов. здравоохран. – 1979. – № 1. – С. 28–32.
2. Анисимов Р.П. Изучение заболеваемости с временной утратой трудоспособности рабочих машиностроительной промышленности / Р.П. Анисимов // Сов. здравоохранение. – 1966. – № 4. – С. 12–18.
3. Бабкин С.И., Белькович В.И., Вишняков И.И. Основные проблемы массового медицинского обследования населения // Медицинская техника. – 1972. – № 6. – С. 12–19.
4. Белицкая Е.Я. Учебное пособие по медицинской статистике. – Ленинградское отделение: Медицина, 1972. – С. 178.
5. Биометрические методы / Ред. Н.А. Плохинский. – М.: Издательство Московского университета, 1975. – С. 168.
6. Вентцель Е.С. Теория вероятности. – М.: Наука, Физматгиз, 1969. – С. 576.
7. Вигдорчик Н.А. Естественный профессиональный отбор с точки зрения профессиональной гигиены. – Л., 1928. – С. 53.
8. Вилянский М.П. Скрининг в массовых профилактических осмотрах (автоматизированная система) / М.П. Вилянский, Б.С. Кибрин, А.А. Чумороков. – М.: Медицина, 1987. – С. 160.
9. Вишняков Н.И. Актуальные проблемы здравоохранения. Сборник научных трудов. – СПб., 2003. – С. 230.
10. Власов В.В. Эффективность диагностических исследований. – М.: Медицина, 1988. – С. 256.
11. Войцехович Б.А. Новые тенденции распространенности социально значимых заболеваний / Б.А. Войцехович, Е.И. Клещенко // Медико-социальные проблемы социально-обусловленных заболеваний: научные труды Российской научно-практической конференции. 26–27 мая 2004 г. – М.: РИО ЦНИИОИЗ, 2004. – С. 11–15.
12. Вопросы статистики // Ежемесячный научно-информационный журнал (Государственный комитет по статистике Российской Федерации) – издается с 1994 года.
13. Государственный доклад о состоянии здоровья населения Российской Федерации в 1994 г.; 1995 г.; 1996 г. // Здравоохран. Рос. Федерации. – 1996. – № 4. – С. 9–21. – 1996. – № 6. – С. 3–9. – 1997. – № 4. – С. 11–23. – 1998. – № 4. – С. 3–16.
14. Гриненко А.Я. Состояние здоровья населения Северо-Западного федерального округа России и проблемы реформирования здравоохранения. – СПб., 2003. – С. 160.
15. Гублер Е.В., Генкин А.А. Применение непараметрических критериев статистики в медико-биологических исследованиях. – Л.: Медицина, 1973. – С. 144.

16. Данишевский К.Д. Прогноз развития демографической ситуации и эпидемиологии инфекционных и неинфекционных заболеваний в Российской Федерации в 2002–2010 гг. / К.Д. Данишевский, А.В. Бобрик // Проблемы здоровья населения. Сборник научно-практических статей. – М.: Открытый институт здоровья, 2005. – С. 72–81.
17. Догле Н.В., Юркевич А.Я. Заболеваемость с временной утратой трудоспособности. – М.: Медицина, 1984. – С. 183.
18. Доклад о состоянии здравоохранения в мире – 2002. Снижение риска, укрепление здорового образа жизни. – ВОЗ. – Женева, 2002. – С. 248.
19. Дьяченко В.Г. Экспертиза качества медицинской помощи. // Вопросы теории и практики. – М.: ЗдравРеформ, 1996. – С. 203.
20. Ежова Т.В. Информационные технологии на службе здравоохранения / Т.В. Ежова, Н.Е. Ставская // Здравоохранение Рос. Федерации. – 2009. – № 3. – С. 30–34.
21. Жданов А.М. Зарубежные автоматизированные системы многопрофильной оценки здоровья // Обзорная информация ЦБНТИ Медпром. Серия: Промышленность медицинской техники. – М., 1980. – № 10.
22. Зайцев В.М. Медицинская статистика в амбулаторно-поликлинических учреждениях промышленных предприятий / В.М. Зайцев, Л.А. Аликбаева, А.А. Сидоров и др. – СПб., 2009. – С. 416.
23. Зайцев Г.Н. Математический анализ биологических данных. – М.: Наука, 1991. – С. 184.
24. Здоровье населения и приоритеты здравоохранения: Сборник / Ред. Медик. – М., 2005. – С. 192.
25. Здравоохранение в России. 2007: Стат. сб. / Росстат. – М., 2007. – С. 355.
26. Иванова А.Е. Состояние здоровья и инвалидность / А.Е. Иванова, Э.В. Кондракова // Общественное здоровье и профилактика заболеваний. – М., 2006. – № 2. – С. 15–28.
27. Ильясов Ф.Н. Репрезентативность результатов опроса в маркетинговом исследовании // Социологические исследования. – 2011. – № 3. – С. 112–116.
28. Казначеев В.П. Донозологическая диагностика в практике массовых обследований населения / В.П. Казначеев, Р.М. Баевский, А.П. Берсенев. – Л.: Медицина, 1980. – С. 208.
29. Кайгародова Т.В. Информационные ресурсы Всемирной организации здравоохранения в анализе состояния здоровья населения при установлении приоритетов в здравоохранении / Т.В. Кайгародова. Информационно-аналитический вестник. Социальные аспекты здоровья населения. – 2007. – № 1. – С. 11–24.
30. Капра Ф. Скрытые связи. / пер. с англ. – М.: ООО Издат. дом «София», 2004. – С. 336.
31. Катлер Д.Л., Мешалкин Л.Д. Некоторые вопросы, связанные с широким использованием программ массовых осмотров населения // Здравоохранение Рос. Федерации. – 1973. – № 5. – С. 39–42.

32. Клюковкин К.С., Рывкин А.Ю., Солдатенкова Ж.М., Стожарова С.И., Борисевич Е.М., Лебедева Д.Н. Динамика общей и первичной заболеваемости населения Санкт-Петербурга в 1998–2007 гг. // Проблемы городского здравоохранения: Сб. научных трудов. – СПб., 2008. – вып. 13. – С. 11–14.
33. Кобец А.В. Оценка рисков смерти в различных группах российского населения / А.В. Кобец // Обществ. здоровье и профилактика заболеваний. – 2006. – № 3. – С. 14–17.
34. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006. – С. 816.
35. Колодяжный Н.Г. Подводные камни информатизации медицинских учреждений / Колодяжный Н.Г. // ГлавВрач. – 2006. – № 7. – С. 82–86.
36. Корневский Н.А. Построение автоматизированных компьютерных медицинских систем. – Курск: Изд-во КГТУ, 2003. – С. 185.
37. Коровина О.В., Маринкин В.И., Зайцев В.М. Программный комплекс автоматизированного скрининга и анализа статистики по хроническим неспецифическим заболеваниям легких // Материалы 1-й научно-технической конференции «Информатика в медицине». – Дагомыс, 1992. – С. 15–16.
38. Лебедев А.А. Рыночная экономика и медицина. – Самара: Дом печати, 2006. – С. 201.
39. Лисицын Ю.П. Социальная гигиена и организация здравоохранения. Проблемные лекции. – М.: Медицина, 1992. – С. 148.
40. Максимова Т.М. Современное состояние, тенденции и перспективы оценки здоровья населения. – М.: ПЕРСЭ, 2002. – С. 192.
41. Мартыненко В.Ф., Вялкова Г.М., Полесский В.А., Беляев Е.Н., Гройсман В.А., Серегина И.Ф. / Под ред. А.И. Вялкова. Информационные технологии в системе управления здравоохранением Российской Федерации. – М.: «Гэотар-Медиа», 2006. – С. 128.
42. Медик В.А., Токмачев М.С. Руководство по статистике здоровья и здравоохранения. – М.: Медицина, 2006. – С. 527.
43. Медков В.М. Демография. Учебник. – М.: ИНФРВ-М, 2005. – С. 576.
44. Мельникова З.М. О соотношении общей заболеваемости (по обращаемости) и заболеваемости с временной утратой трудоспособности // Здравоохранение Рос. Фед. – 1961. – № 8. – С. 13–17.
45. Мерков А.М. Здоровье населения и методы его изучения: избранные произведения / Под ред. М.С. Бедного. – М.: Статистика, 1979. – С. 232.
46. Мерков А.М., Поляков Л.Е. Санитарная статистика (пособие для врачей). – Л., 1974. – С. 384.

47. Минцер О.П., Молотков В.Н. Кибернетическое прогнозирование в пульмонологии. – М.: Медицина, 1983. – С. 176.
48. Новгородцев Г.А., Полонский М.А., Демченкова Г.З. Диспансеризация населения в СССР. – М.: «Медицина», 1984. – С. 336.
49. Ноткин Е.Л. Об углубленном анализе данных заболеваемости с временной утратой трудоспособности // Гигиена и санитария. – 1979. – № 5. – С. 40–46.
50. Овчаров В.К., Медик В.А. Заболеваемость населения: история, современное состояние и методология изучения. – М.: Медицина, 2003. – С. 509.
51. Орел В.И. Современные подходы к использованию информационных технологий в здравоохранении / В.И. Орел, Т.Ш. Кутушев // Вестн. СПб. гос. мед. Академии им. И.И. Мечникова. – 2008. – № 4. – С. 7–10.
52. Отдельнова К.А. Методика сравнительной оценки состояния здоровья рабочих промышленных предприятий: в кн. «Социально-гигиенические исследования». Труды 2-го МОЛМИ им. Н.И. Пирогова. – М.: МЗ РСФСР, 1980. – С. 174–180.
53. Перекопская Л.Г. Социально-гигиенические аспекты медицинской демографии / Научный обзор МЗ СССР, ВНИИМИ. Обзорная информация // Под ред. Ю.П. Лисицина. – М., 1978. – С. 74.
54. Петленко В.П. Интегральная медицина XXI века. – Медицинский сервер MedCom, – РИА Медицина Северо-Запада, выпуск № 7(10), сентябрь 1998.
55. Платонов А.Г., Ахалая М.Я. Дозовая зависимость постлучевой гибели. Расчет полудетальной дозы ЛД50 методом пробит-анализа. – М.: МГУ им. Ломоносова, 2006. – С. 33.
56. Покровский А.В., Пискарев Ю.Г., Генрих К.Р., Фандесв В.А. Состояние здоровья допризывной и призывной молодежи // Гигиеническая наука и практика на рубеже XXI века: Мат. IX Всеросс. съезда гигиенистов и санитарных врачей. – М., 2001. – Т. 2. – С. 440–442.
57. Поляков И.В. Актуальные проблемы организации здравоохранения современной России / И.В. Поляков // Макроэкономика: качество жизни и проблемы охраны здоровья населения: труды Всероссийской заочной научно-практической конференции с международным участием. – СПб, 2008. – С. 210–214.
58. Руководство по сбору данных о рождаемости и смертности. Методологические исследования / ООН, Нью-Йорк, 2005. – Серия F. – № 92. – С. 132.
59. Савельев С.И., Зайцев В.М., Котова Г.Н., Вишняков Н.И. Оценка гигиенической активности современной молодежи по результатам углубленного статистического анализа образа жизни / Под редакцией академика РАМН, профессора Г.Г. Онищенко. – Л., 2004. – С. 416.
60. Семашко Н.А. Избранные произведения. – М.: Медицина, 1967. – С. 379.

61. Толмачев В.А. Основы характеристики медицинской услуги // Бюллетень НИИ СГЭ и УЗ им. Н.А. Семашко. – 1995. – № 3. – С. 57–61.
62. Урбах В.Ю. Математическая статистика для биологов и медиков. – М.: Изд-во АН СССР, 1963. – С. 325.
63. Утка В.Г. Современное состояние и пути развития отечественной медицинской статистики // Михайлова Ю.В. и др. / Специальное приложение к сборнику научных трудов «Новые технологии в современном здравоохранении»: ФГУ ЦНИИОИЗ. – М., 2007. – С. 36.
64. Фролова Е.В. Скрининг с позиций доказательной медицины: необходимость, стоимость, эффективность / Е.В. Фролова // Российский семейный врач. – 2000. – № 4. – С. 33–36.
65. Шахгельдянц А.Е. Лечебно-профилактическая помощь рабочим промышленных предприятий. – М.: Медицина, 1968. – С. 168.
66. Шведова Н.А. Здравоохранение: американская модель. – Наука, 1993. – С. 144.
67. Юл Дж. Э., Кендалл М. Дж. Теория статистики / Пер. с англ. – М.: ГосИздат, 1960. – С. 780.

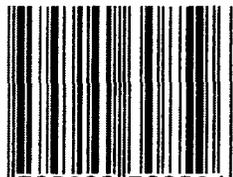
В.М. ЗАЙЦЕВ, С.И. САВЕЛЬЕВ

ПРАКТИЧЕСКАЯ МЕДИЦИНСКАЯ СТАТИСТИКА

УЧЕБНОЕ ПОСОБИЕ

Под редакцией
академика РАН, профессора, д.м.н.,
заслуженного деятеля науки Российской Федерации
А.И. Потапова
и профессора, д.м.н. О.Г. Хурцилава

ISBN 978-5-9903085-9-6



9 785990 308596

Корректор – О.Ж. Белкина

Сдано в набор 01.10.2013 г. Подписано к печати 18.11.2013 г.
Формат 70x100/16. Бумага офсетная.
Гарнитура HeliosCondBlack, HeliosCond, Times New Roman. Усл. печ. л. 46,76
Тираж 1000 экз. Заказ № 478

Подготовлено в печать ИП Луговина Н.Н.

Отпечатано с готового оригинал-макета в ООО «Цифра»
392000, г. Тамбов, ул. Чичканова, д. 17