ФГБУ ВЦЭРМ им. А.М. Никифорова МЧС России

Н.В. Макарова

Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NCSS, SYSTAT

Методическое пособие

Методика и условия применения статистических пакетов
Примеры
Словарь терминов
Справочник формул

Санкт-Петербург
2012

УДК 61:311(075.8)
ББК 51.1(2)я73

М-15

Макарова Н.В.

Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NCSS, SYSTAT : методическое пособие / Н.В. Макарова ; Всерос. центр экстрен. и радиац. медицины им. А.М. Никифорова МЧС России – СПб.: Политехника-сервис, 2012. – 178 с.

Это пособие предназначено для научных работников, аспирантов, врачей с разным уровнем статистической подготовки, которые самостоятельно осуществляют анализ данных на компьютере. Большое число примеров позволит исследователям освоить технологии использования компьютерных программ, выбрать адекватные варианты статистической обработки и грамотно трактовать полученные результаты. В книге проведено сравнение возможностей статистического анализа в нескольких универсальных пакетах программ.

Справочная информация о понятиях и методах статистики, описанных в пособии, а также формулы для вычисления приведены в «Приложении».

Рецензент:

В.И. Кувакин – доцент кафедры автоматизации управлением медицинской службой (с военно-медицинской статистикой) Военно-медицинской академии им. С.М. Кирова заслуженный работник высшей школы РФ доктор медицинских наук профессор

ISBN 5-9231-0269-2

Оглавление

Предисловие научного редактора

Данное пособие подготовлено в ФГБУ Всероссийский центр экстренной и радиационной медицины им. А.М. Никифорова МЧС России (ВЦЭРМ), в научно-исследовательском отделе «Медицинский регистр МЧС России» (начальник – Астафьев О.М.). Автор, являясь начальником лаборатории статистического анализа данного сектора, проводит статистическую обработку и анализ данных разнообразных исследований, осуществляемых во ВЦЭРМ, в течение многих лет.

Пособие предназначено для научных работников, аспирантов и врачей с разным уровнем статистической подготовки, которые самостоятельно осуществляют анализ данных биологических, эпидемиологических, клинических, психологических, лабораторных и других исследований на компьютере. Большое число примеров позволит исследователям освоить технологии использования компьютерных программ, выбрать адекватные варианты статистической обработки и грамотно трактовать полученные результаты.

В книге проведено сравнение возможностей статистического анализа в нескольких универсальных пакетах программ, наиболее распространенных среди исследователей медико-биологического профиля. Это позволит выбрать наиболее подходящий для целей исследования инструмент. Например, для оценки риска неблагоприятного исхода в связи с влиянием нескольких факторов наиболее широкие возможности предоставляет программа NCSS.

Большинство использованных в книге примеров основано на данных из реальных медицинских исследований, проведенных в НИО «Медицинский регистр МЧС России» и других подразделениях ВЦЭРМ. В частности, использовались данные исследований сотрудников НИО Санникова М.В. и Шевченко Т.И. Ряд примеров связан с оценкой медико-биологических последствий для здоровья ликвидаторов, которые проживают на территории Северо-Запада России и включены в систему Национального радиационно-эпидемиологического регистра (НРЭР). Эта территориальная база данных НРЭР ведется и обрабатывается в НИО «Медицинский регистр МЧС России» с 1997 года и включает сведения о здоровье более 11000 ликвидаторов за весь период наблюдений – с 1986 по 2012 гг.

Справочная информация о понятиях и методах статистики, описанных в пособии, а также формулы для вычисления приведены в «Приложении».

Доктор медицинских наук профессор С.С. Алексанин

В результате применения статистического метода мы получаем не истину в последней инстанции, а всего лишь оценку вероятности того или иного предположения. Кроме того, каждый статистический метод основан на собственной математической модели и результаты его правильны настолько, насколько эта модель соответствует действительности.

С. Гланц. Медико-биологическая статистика.

ВВЕДЕНИЕ.

При экспериментальных исследованиях медико-биологических систем их характерной особенностью является отсутствие полной воспроизводимости и стабильности. Это связано с очень большим числом факторов, влияющих на исход опыта, в том числе и не поддающихся измерению. Поэтому статистические методы являются основным способом количественного описания медико-биологических объектов и явлений.

Существует множество учебников, в том числе и хороших, посвященных статистическим методам обработки данных. Однако для практического использования они не совсем удобны, поскольку описание методов начинается и заканчивается на уровне математических формул. Для понимания метода формулы незаменимы, но для реальных расчетов сейчас естественно использовать профессионально созданные пакеты прикладных статистических программ. Использование пакетов – это тоже технология, которую надо знать. С одной стороны, скорость получения статистических выводов стала неизмеримо выше, чем была при ручных подсчетах (не говоря о том, что многие методы были просто недоступны без качественного программного обеспечения и современных компьютеров), и это позволяет применять несколько альтернативных методов для проверки и подтверждения полученных выводов. С другой стороны, вопрос правильного использования статистических методов не снимается, и это относится как к выбору соответствующих модулей в пакете программ, так и конкретных опций, задающих параметры исследуемой модели.

Кроме того, исторически сложилась традиционная тематика в учебниках по статистике на русском языке, и в этой традиции не нашлось места описанию ряда очень важных методов обработки данных. В частности, автору не удалось найти удовлетворительных описаний метода Мантеля-Ханзела (Mantel-Haenszel) вычисления

объединенного риска, логистической регрессии, логлинейного анализа. Эти методы изложены в настоящем руководстве, и их применение проиллюстрировано рядом примеров. Традиционные учебники, как правило, уделяют основное внимание статистическим выводам, основанным на предположении о нормальном распределении переменных, в том числе многомерным статистическим методам, таким как регрессионный, факторный, дисперсионный анализ. В практических же задачах часто требуется выявить отличия, связи, структуры для переменных, имеющих порядковую или дискретную структуру, или же для набора переменных разного типа. Для них также требуется выбор наилучшего метода, представление о границах применимости каждого из возможных способов обработки данных. Кроме того, даже для переменных непрерывного типа далеко не всегда возможно применение параметрических методов анализа, особенно при малых объемах выборок.

В самом общем виде основные задачи, для решения которых применяются методы статистического анализа, можно сформулировать следующим образом:

1) выявление отличий анализируемых показателей и их связей в двух и более выборках;

2) определение наличия и величины связи (или зависимости) одного или нескольких факторов с другими показателями или процессом;

3) проведение анализа структуры данных.

Для проведения статистического анализа, как правило, достаточно владения одной из компьютерных систем анализа данных. Наиболее распространенные и универсальные системы, такие как STATISTICA, SPSS, SAS, NCSS, SYSTAT, предлагают примерно совпадающий арсенал основных методов анализа данных. В предлагаемом пособии мы будем давать ссылки на все вышеуказанные системы, но в качестве основной выбрана STATISTICA. Такой выбор связан прежде всего с удобством экспорта-импорта данных и результатов в этой системе, а также наибольшей доступностью для русскоязычных пользователей.

Для грамотного применения любого статистического пакета анализа данных нужно, во-первых, сформулировать задачу таким образом, чтобы для ее решения можно было использовать статистические методы, то есть создать математико-статистическую модель исследования. Для начинающих исследователей это наиболее трудный пункт плана работ. Как правило, статистическая модель практического исследования многовариантна и допускает

использование нескольких методов. Поэтому надо представлять себе возможности, которые есть в арсенале прикладной статистики, и понимать, какие из них могут быть использованы для решения конкретной задачи. Наконец, важно учитывать, какие требования к данным сопровождают использование этих методов, и проверять выполнение этих требований.

Поскольку существует несколько возможных вариантов решения задачи, следует выбрать наиболее подходящие из них. Кроме того, использование нескольких методов позволяет проверить полученные выводы. Приведенные далее схемы показывают спектр методов, применяемых для решения наиболее распространенных задач. На схемах указаны самые известные методы, содержащиеся практически во всех статистических программах анализа данных.

Схема 1. Выявление отличий анализируемого показателя в двух и более выборках

Схема 2. Определение наличия и величины связи (или зависимости) двух показателей

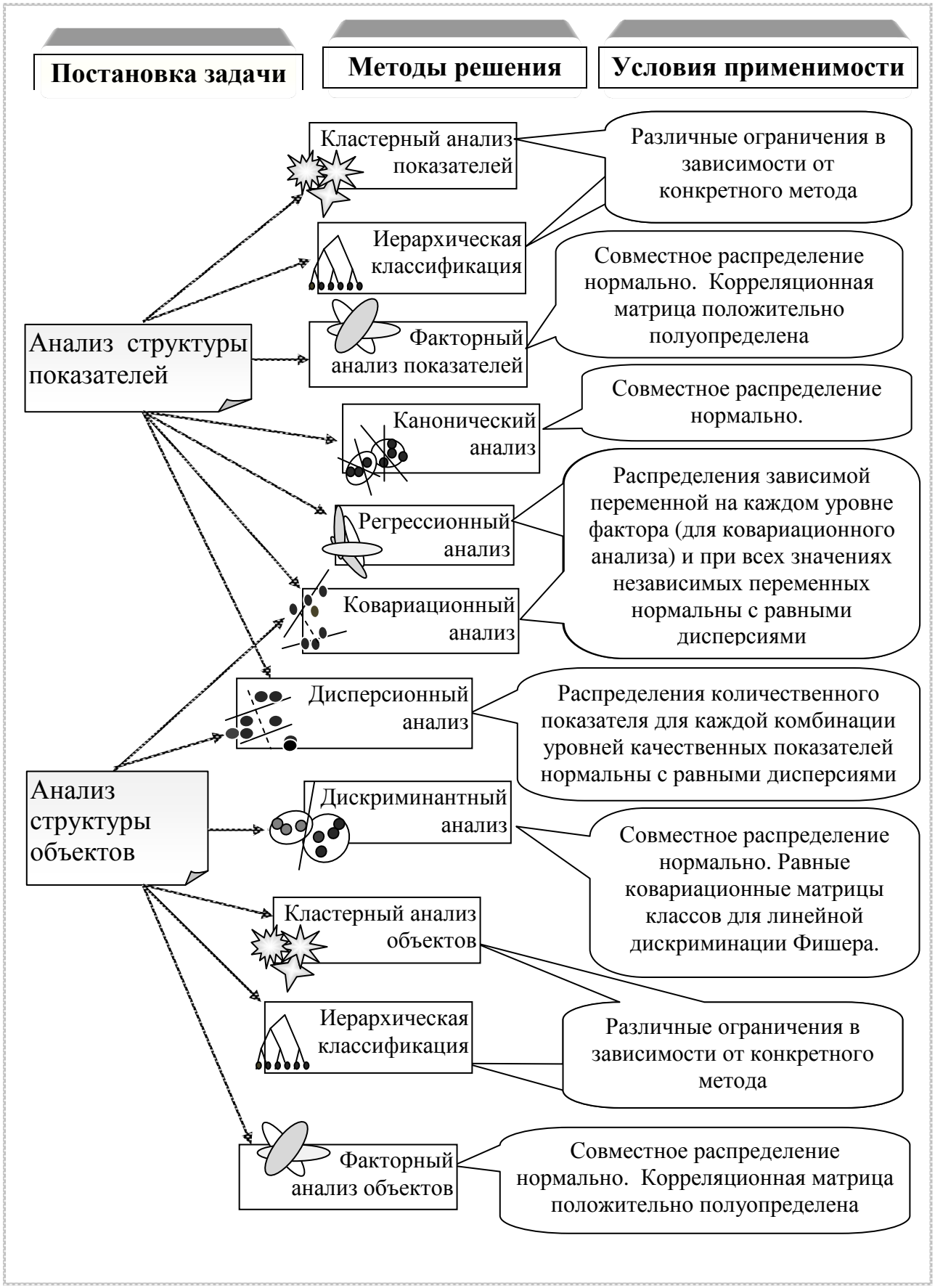


Схема 3. Исследование структуры данных.

Комментарии к схемам

При изучении схем 1 (1') и 2 (2') нетрудно заметить, что одни и те же методы могут быть использованы для решения различных задач. С другой стороны, одну и ту же задачу можно сформулировать как в рамках схемы 1, так и в рамках схемы 2. Приведем примеры.

1. Был вычислен индекс функциональных изменений (ИФИ) для пожарных и спасателей, проходивших ежегодную диспансеризацию. Исследователь хочет выяснить, есть ли отличия по этому индексу у обследованных двух профессиональных групп. Вопрос можно сформулировать следующим образом: а) «отличается ли ИФИ у спасателей и пожарных?» или б) «зависит ли ИФИ от характера работы?».

Для ответа на вопрос (а) следует искать методы на схеме 1, причем конкретизация вопроса может быть различной: (1) - отличаются ли средние значения ИФИ в группах спасателей и пожарных; (2) – отличаются ли распределения индекса в этих группах, включая и отличие средних; (3) – отличаются ли доли лиц с ИФИ более 3.1 (неудовлетворительная адаптация или срыв адаптации) в исследуемых группах.

Для ответа на вопрос (б) методы следует искать на схеме 2. В рамках этой схемы вопрос может быть уточнен следующим образом: (1) - отличаются ли средние значения ИФИ при разных уровнях показателя «характер работы»; (3) – отличаются ли доли лиц с неудовлетворительной адаптацией для групп с различным характером работы.

(а) и (б) отличаются только формулировкой вопроса, а не существом задачи, поэтому на обеих схемах присутствуют методы для ее решения. В данном случае это может быть двухвыборочный критерий Стьюдента, если выполнены необходимые условия нормальности и конкретная форма вопроса – (1) – проверка равенства средних значений. Может быть также использован критерий Вилкоксона, Манна-Уитни, если проверяется (2) – совпадение распределений.

2. Инструментом широкого применения является критерий χ^2. В частности, он может использоваться для сравнения двух и более выборочных распределений. В этом случае говорят, что проверяется гипотеза об однородности двух или нескольких выборок. Особенно часто критерий χ^2 используется в тех случаях, когда интересующий нас показатель является качественным, т.е. его значения не связаны

отношением порядка. Характерная задача, в которой целесообразно использовать данный критерий, – сравнение структуры заболеваемости или смертности в нескольких группах наблюдения (например, на разных территориях). Значения исследуемого показателя – отдельные нозологические классы: инфекционные болезни, болезни органов дыхания и т.д. Количество выявленных болезней по этим классам (количество умерших по отдельным причинам) на каждой территории определяет выборочное распределение.

С другой стороны, критерий χ^2 может быть использован для проверки гипотезы о независимости двух показателей. Показатели могут быть как количественными, так и качественными, но следует иметь в виду, что при применении этого критерия не учитывается информация о порядке и величине значений показателей. Поэтому разумно применять критерий χ^2 для проверки независимости в том случае, когда по крайней мере один показатель является качественным. Например, для выяснения связи профессиональной принадлежности со статусом курения.

Пособие написано для практических исследователей, поэтому применение методов анализа разбирается на конкретных примерах, с использованием наиболее распространенных пакетов статистических программ. Особое внимание уделено вопросам применимости известных методов и выбора наиболее подходящего способа обработки в зависимости от параметров задачи: объемов, типов данных, постановки вопроса. В основном, предметом анализа в данном пособии являются задачи, в которых требуется проанализировать частоты различных событий. Такой выбор обусловлен, во-первых, распространенностью этих задач, а во-вторых, недостаточно полным освещением способов решения их в специальной литературе. Максимально возможное количество формул вынесено в Приложение, для того чтобы облегчить процесс чтения руководства исследователями, не обладающими математической подготовкой.

Первая глава пособия посвящена определению основных понятий, которые используются на этапе планирования статистического анализа. Для того, чтобы получить достоверные результаты, необходимо корректно подготовить материал исследования для статистической обработки.

Вторая глава содержит описание задач, для решения которых используются частотные таблицы и таблицы сопряженности. Основная часть главы посвящена способам и условиям применения критерия χ^2 .

В третьей главе подробно обсуждаются способы получения статистических выводов о частотах появления события в одной или нескольких выборках как частном случае обработки таблиц сопряженности и частотных таблиц. Описаны основные типы задач и способы решения.

Глава 4 посвящена способам вычисления рисков при наличии мешающих параметров. В ней приведены способы вычисления объединенных рисков с проверкой однородности и значимости и способы стандартизации.

В главе 5 изложена методика и практика применения логистической регрессии.

Глава 6 содержит описание логлинейного анализа.

В Приложении приведены точные формулы и определения статистических понятий и методов.

Нумерация формул, рисунков и таблиц двойная: первая цифра соответствует номеру главы, вторая является порядковым номером формулы, рисунка, таблицы в данной главе. Схемы пронумерованы последовательно, независимо от глав. Таблицы и рисунки в примерах нумеруются в соответствии с номером примера: Таблица П1-1, Рис. П1-1. Текст примеров и обсуждение полученных результатов приведено более мелким шрифтом.

ГЛАВА 1. ОПРЕДЕЛЕНИЕ ОСНОВНЫХ ПОНЯТИЙ

Выборка (sample) –

часть популяции (генеральной совокупности), полученная путем отбора. Исследования выполняются обычно на выборках.

1.1. Виды данных

При различных исследованиях в статистическом анализе могут участвовать данные разных типов. Для корректного использования статистических методов важно представлять, какого типа данные будут обрабатываться. Упрощенно можно разделить их на два основных типа: качественные и количественные.

Качественные данные (nominal data)

Также называются классификационными, неупорядоченными. Это признаки, которые нельзя выразить количественно: диагноз, место проживания, пол. Говорят, что такие показатели измерены в номинальной шкале. При использовании статистических пакетов для анализа данных признаки могут (или должны) быть оцифрованы: например, 1 – «Калининградская область», 2 – «Ленинградская область», 3 – «СПб». Смысла эти числа не имеют, это только удобная форма записи.

Частным случаем номинальных являются дихотомические данные (признаки, имеющие только два значения, типа «да – нет», называются также бинарными). Для их оцифровки принято использовать числа 0 и 1: 0 – «нет», 1 – «да». Для некоторых статистических программ такая кодировка обязательна.

Порядковые данные (ordinal data)

Встречаются также названия: признаки с упорядоченными состояниями, ординальные. Показатели, измеряемые в шкале порядка – промежуточные между качественными и количественными (стадии болезни, оценки – «плохо», «удовлетворительно», «хорошо»). Такие признаки могут быть осмысленно оцифрованы, поскольку порядок состояний имеет смысл. Часто к таким показателям следует относить балльные оценки, полученные при проведении тестов или экспертиз.

Количественные данные (numerical data)

Признаки, выражаемые в числовой форме: возраст, вес, количество детей в семье. В свою очередь, они делятся на непрерывные и дискретные.

Непрерывные данные (continuous data)

Количественные данные, которые могут принимать любое значение на непрерывной шкале. Другое название – признаки, измеряемые в интервальной шкале (температура, АДС, рост).

Дискретные данные (discrete data)

Количественные данные, измеряемые в шкале отношений. Они принимают, как правило, конечное число значений, хотя иногда и очень большое: количество смертей в течение года в исследуемой когорте, количество пропущенных по болезни рабочих дней.

1.2. Подготовка данных

Для использования статистических программ обработки исходные данные должны быть представлены в виде таблицы. Каждая строка таблицы соответствует одному объекту выборки (например, одному человеку), а каждый столбец – одному показателю. При этом вся необходимая для вычислений информация должна содержаться в таблице. Например, если вся выборка состоит из двух частей – «опыт» и «контроль», то одним из столбцов таблицы будет показатель «группа». Этот показатель для объектов, относящихся к опыту, заполняется словом «опыт» или каким-либо числовым кодом, например, 1. Для объектов, относящихся к контролю, он заполняется словом «контроль» или другим числовым кодом, например, 0. Если какой-либо показатель у объекта не известен (не измерялся или не имеет смысла), соответствующая ячейка таблицы должна остаться незаполненной.

Эти требования одинаковы для всех существующих программ статистической обработки.

Образец подготовки данных.

Номер пациента	Пол	Возраст	Вес	Рост	ИМТ (вычисляемый показатель)	Группа наблюдения
1	2	41	68	168	1.76	1
2	2	44		165		1
3	1	52	73	174	1.81	1
4	1	50	65	168	1.74	2
5	2	42	59	168	1.65	2

В тех случаях, когда требуется обработать динамические наблюдения, появляются разные варианты подготовки данных, связанные с разными видами анализа. Например, если показатель (ЧСС) был измерен у испытуемых до, во время и после нагрузки, при обработке можно использовать (схема 1) как критерий Стьюдента для связанных выборок или критерий Вилкоксона, так и дисперсионный анализ с повторными измерениями.

Для применения программ дисперсионного анализа все измерения должны быть внесены в один столбец, а номер измерения записан как отдельная переменная. В этом случае таблица на схеме символически обозначена значком 

Номер пациента	Номер измерения	Измерение ЧСС
1	1	85
2	1	65
3	1	71
4	1	73
5	1	69
1	2	91
2	2	76
3	2	81
4	2	69
5	2	78
1	3	91
2	3	81
3	3	81
4	3	75
5	3	89

При статистической обработке связанных выборок (критерии Стьюдента или Вилкоксона) таблица должна иметь столбцы, соответствующие всем измерениям показателя, а критерии применяются к парам столбцов, например: сравнение «Измерения 1» и «Измерения 2». Такой вид таблицы на схеме отражен значком 

Номер пациента	Измерение 1	Измерение 2	Измерение 3
1	85	91	91
2	65	76	81
3	71	81	81
4	73	69	75
5	69	78	89

1.3. Использование данных разных видов в анализе

Разные виды данных используются в статистическом анализе неодинаково. Качественные данные обычно задают группировки в исследованиях. В этом случае они называются группирующими. Группирующие переменные могут определять как группы сравнения, так и разделение исходной выборки на более однородные части для проведения анализа по каждой части выборки отдельно. Иногда качественный показатель, в частности, бинарный, рассматривается в качестве фактора: например, наличие заболевания как целевой фактор, курение как фактор риска.

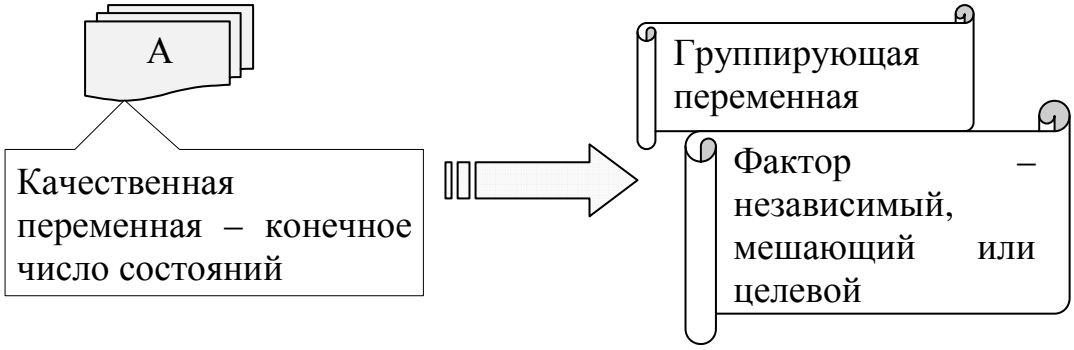


Рис.1.1. Использование качественных переменных в статистическом анализе

Количественные показатели могут являться 1) целевыми для исследования – например, зависимые переменные (dependent variable) в дисперсионном, ковариационном, регрессионном анализе; 2) объясняющими или независимыми переменными в ковариационном, регрессионном, дискриминантном анализе; 3) также могут быть группирующими, если количественный показатель является дискретным. Если же показатель непрерывен, то для использования его как группирующего вводятся градации состояний, и он преобразуется в порядковый.

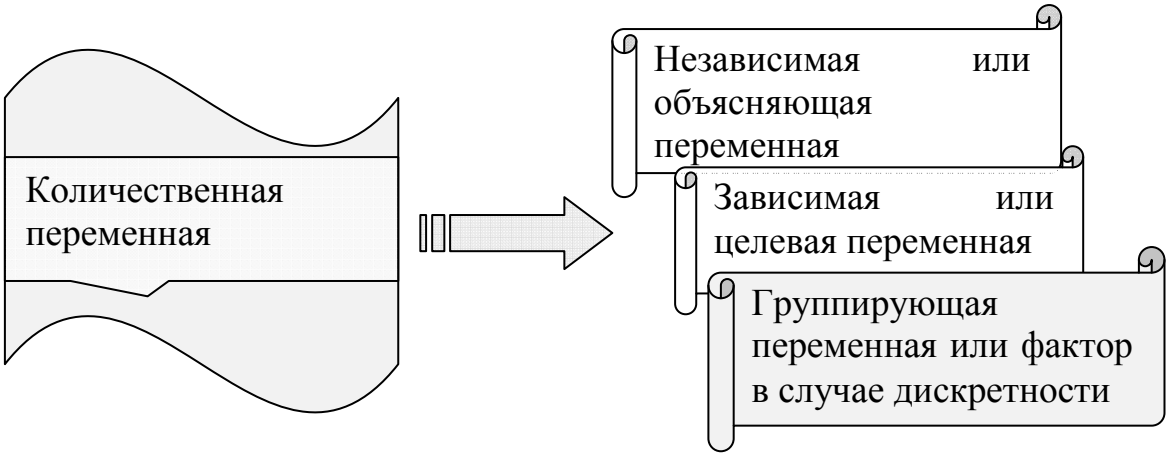


Рис.1.2. Использование количественных переменных в статистическом анализе

1.4. Предположения

Для того чтобы получить статистические выводы о значениях параметров или связях между ними, сначала требуется сделать определенные предположения. Основные предположения касаются того, какого рода случайность присуща интересующим нас показателям в популяции. Типы случайностей описываются разными законами распределения. В зависимости от того, какой закон распределения мы выбрали для описания показателей, мы располагаем и соответствующим набором статистических средств (видов статистического анализа) для вычисления характеристик изучаемых показателей и получения статистических выводов.

Как правило, с определенным типом данных связан определенный закон распределения или несколько основных законов.

Самый простой случай – дихотомический показатель, то есть показатель, имеющий ровно два возможных значения. Его моделируют биномиальным законом распределения (точные определения в «Приложении»).

Для непрерывных показателей существует большой выбор различных законов распределения, но в первую очередь проверяется согласие выборочного распределения с нормальным (гауссовским) законом распределения. Для этого используются критерии согласия (см. «Приложение»). Основные статистические методы анализа непрерывных показателей разработаны для случая, когда показатели подчинены нормальному распределению. Это касается регрессионного, дисперсионного, ковариационного, дискриминантного, факторного анализа. Когда говорят о параметрических методах и критериях, как правило, подразумевается, что анализируемые показатели имеют именно нормальное распределение.

Для дискретных данных чаще всего используется моделирование полиномиальным распределением (конечное число исходов) или распределением Пуассона.

В случае, когда мы имеем дело с порядковыми данными, при их анализе используют методы, свободные от предположений о конкретном виде распределения – непараметрические или ранговые методы. Эти методы применяются и в тех случаях, когда непрерывные показатели имеют распределения, существенно отличающиеся от нормального (проверка производится с помощью критериев согласия).

Предельные теоремы теории вероятностей позволяют использовать более мощные параметрические методы для данных, распределение которых отличается от нормального, при достаточно большом объеме выборки. Величина «достаточного объема» может отличаться для различных методов и различных характеристик показателей.

Например, t-статистика Стьюдента менее чувствительна к отклонениям от нормальности, нежели F-статистика Фишера. Поэтому выводы, полученные при применении методов, основанных на F-статистике, (регрессионный, дисперсионный, дискриминантный анализ), для малых выборок могут быть недостоверными.

Далее, при сравнении частот событий в двух выборках (например, уровней заболеваемости) достаточными для применения нормальной аппроксимации будут объемы выборок n_1, $n_2 > 100$ или даже 50, но при условии, что события происходят не очень редко и не очень часто: если частота в пределах от 0.1 до 0.9. Подробнее условия применимости нормальной аппроксимации при сравнении частот обсуждаются в главе 3 «Сравнение частот событий».

1.5. Анализ мощности и оценка объема выборки в планировании эксперимента

Оценка необходимого объема выборки возможна и необходима только в том случае, когда исследователь заранее сформулировал проверяемую гипотезу, причем весьма точно: нужно не только зафиксировать интересующий исследователя параметр, но и величину его изменения, которую требуется обнаружить. При этом расчеты объема выборки будут зависеть от того, какой критерий планируется применить.

Назовем исходную гипотезу "нулевая гипотеза" - H_0. Как правило, она состоит в предположении, которое мы заинтересованы опровергнуть – в том, что интересующий нас параметр не изменился

(или не отличается в двух группах, или равен конкретной величине). Соберем данные. Используя статистическую теорию, проверим, верна ли гипотеза H_0 или ее следует отвергнуть. Отвергая H_0, мы обосновываем то, во что действительно верим. Эта ситуация, типичная во многих областях приложения, называется критерий отвержения-принятия - "Reject-Support testing," (RS testing): отвергая нулевую гипотезу, мы подтверждаем теорию.

Нулевая гипотеза либо справедлива, либо ошибочна, и статистическая процедура указывает на это. Нулевая гипотеза либо отвергается, либо не отвергается. Следовательно, до проведения эксперимента мы постулируем, что имеют место только 4 возможности, показанные ниже:

		Реальная ситуация	
		H_O	H_1
Решение	H_0	Правильное принятие	Ошибка II рода β
	H_1	Ошибка I рода α	Правильное отвержение

Соответственно, возможны ошибки двух типов, и они показаны в этой таблице. Обычно придерживаются такой точки зрения, что ошибка I рода α должна принимать значение 0.05 или ниже, тогда как ошибка II рода β должна быть столь малой, насколько это возможно при фиксированном уровне ошибки I рода. "Статистическая мощность", которая равна 1 - β, соответственно, должна быть максимально высокой. Идеальный вариант, когда мощность равна, по крайней мере, 0.80, чтобы обнаружить разумные уклонения от нулевой гипотезы.

Для определения объема выборки требуется заранее задать следующие параметры:

1. Мощность (1-β) – вероятность обнаружения эффекта заданной величины как статистически значимого, если он существует. β – это вероятность ошибки II рода, состоящей в неправильном принятии нулевой гипотезы (**не** обнаружение реально существующих отличий). Мощность обычно выбирается равной 0.7-0.8 (70-80%).
2. Уровень значимости α принятия нулевой гипотезы. Обычно выбирается равным 0.05. α – это вероятность ошибки I рода, состоящей в неправильном отвержении нулевой гипотезы (обнаружение отличий там, где их в действительности нет).

3. Характеристика вариабельности наблюдений – как правило, стандартное отклонение.
4. Наименьший значимый эффект – это та величина эффекта, которую считают клинически важной и которую желательно обнаружить. Чаще всего это разность средних значений или пропорций.

Задав эти параметры, можно воспользоваться несколькими способами вычисления необходимого объема выборки. Для наиболее часто используемых критериев – парного и непарного критериев Стьюдента и критерия χ^2 Пирсона можно применить оценки с помощью номограммы Альтмана или быстрой формулы Лера. Для их применения вычисляется «стандартизованная разность» - для двухвыборочного критерия Стьюдента это δ/σ, где δ – наименьшая клинически значимая разность средних значений, σ – стандартное отклонение, одинаковое в обеих группах. Соотношение δ/σ иногда обозначается символом φ и называется «параметром нецентральности». При сравнении частот стандартизованная разность равна

$$(p_1 - p_2) / [(\tilde{p}(1-\tilde{p})]^{1/2},$$

где $(p_1 - p_2)$ – наименьшая клинически важная разность долей (пропорций) явления в двух группах,

$$\tilde{p} = (p_1 + p_2) / 2$$

Тогда, согласно быстрой формуле Лера, для получения мощности 80% и уровня значимости 0.05 требуется взять в каждой из групп 16/(стандартизованную разность) наблюдений. Для достижений 90% мощности в числителе вместо 16 нужно взять 21.

ГЛАВА 2. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ТАБЛИЦ

Одной из самых распространенных задач прикладной статистики является проверка гипотез, касающихся распределений одного или нескольких дискретных показателей с конечным числом возможных значений (исходов), причем количество таких значений невелико. Такие показатели могут быть по существу дискретными (профессия, пол) или дискретизированными в результате обработки (возраст → возрастная группа); на них может быть определен порядок значений (оценка ответа в баллах) или они могут быть номинальными, качественными (цвет волос). Все эти обстоятельства следует учитывать при выборе метода обработки.

Наиболее известным и универсальным методом решения статистических задач, связанных с дискретными показателями, является критерий χ2. Далее будут показаны основные способы его применения и возможные ограничения.

2.1. Использование критерия χ^2.

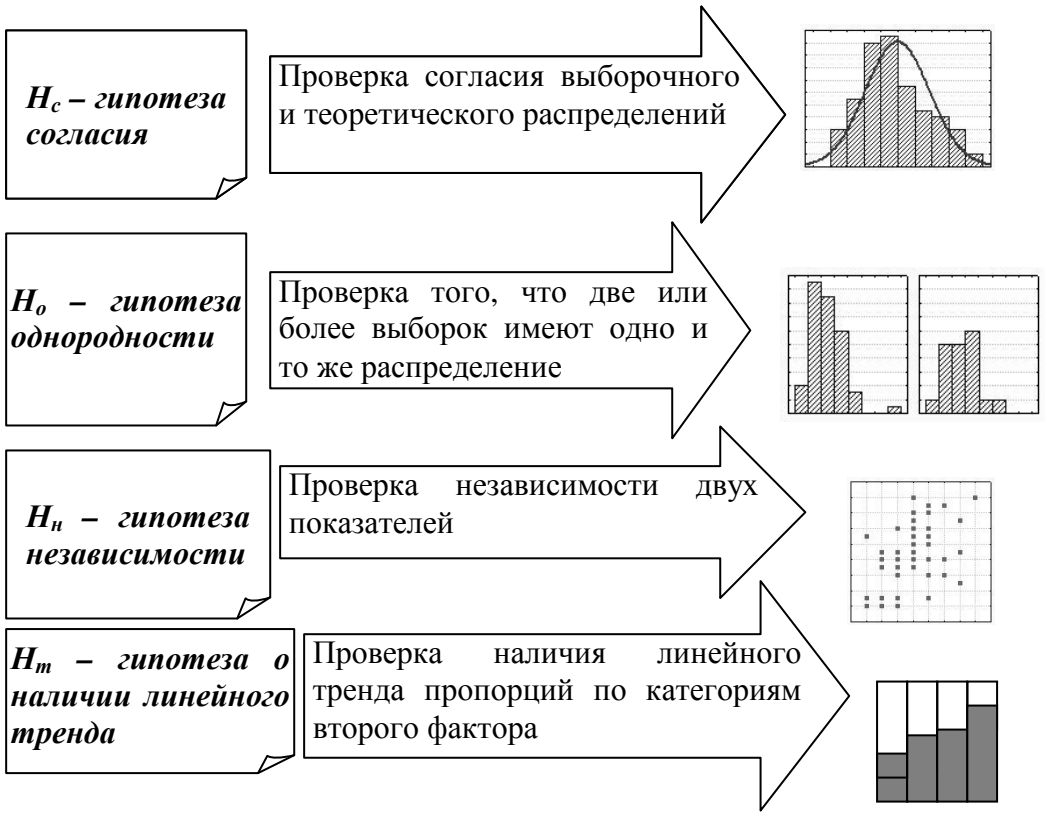


Схема 4. Виды статистических гипотез, в которых используется критерий χ^2.

На приведенной выше схеме приведены основные статистические гипотезы, для проверки которых используется критерий χ^2.

Статистика Пирсона $x^2 = \Sigma_{k=1}^{r}(n_k - np_k)^2/np_k$ объединяет индивидуальные расхождения между наблюдаемыми и ожидаемыми частотами в общую меру расстояния. При больших отклонениях отдельных наблюдаемых частот от ожидаемых значения статистики будут большими, при малых отклонениях всех наблюдений статистика будет мала по величине. Вопрос о границе малых значений, которые еще можно трактовать как случайные отклонения, решается в терминах выборочного распределения статистики x^2, приближенно совпадающего с распределением χ^2 (хи-квадрат), и поэтому статистику Пирсона x^2 часто называют статистикой Пирсона χ^2 или просто статистикой (критерием) χ^2.

2.2. Проверка гипотезы согласия H_c.

Статистика Пирсона применяется в качестве критерия согласия для проверки гипотезы о виде распределения. В этом случае осуществляется сравнение теоретических и выборочных частот (как для дискретных, так и для непрерывных переменных).

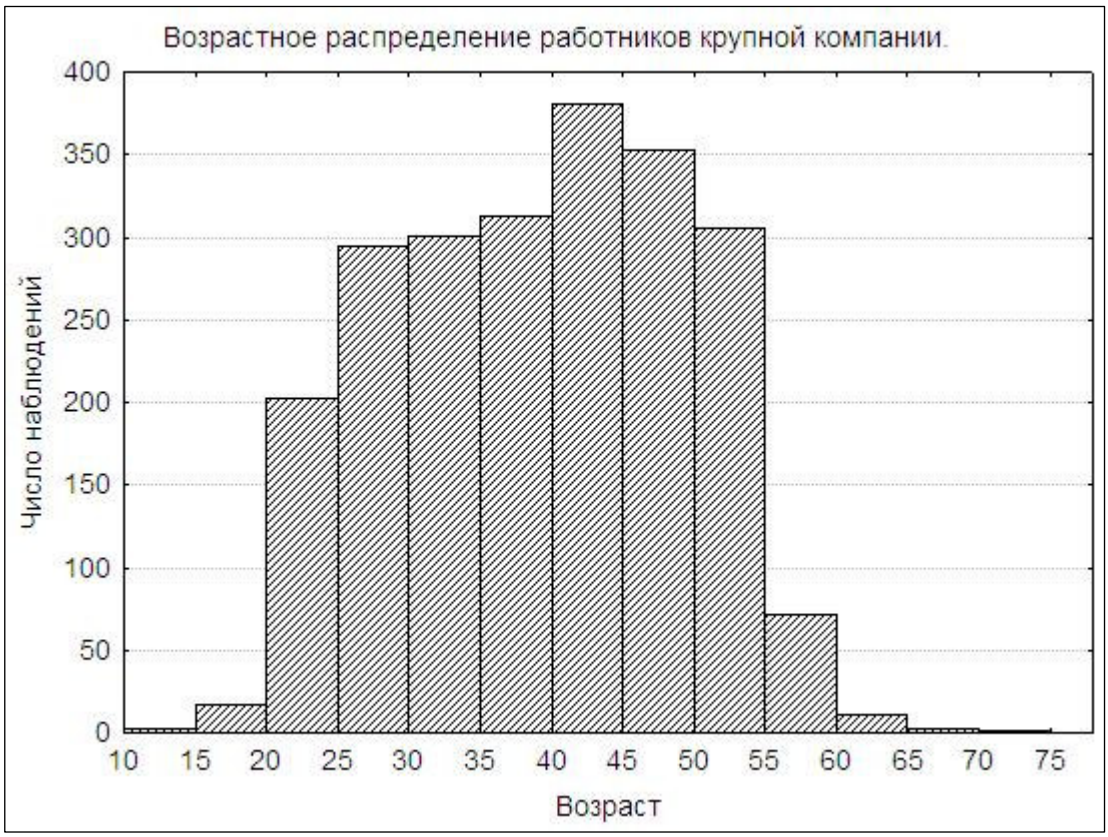


Рис.2.1. Гистограмма возрастного распределения

Если проверяется согласованность выборки с генеральной совокупностью, то p_i – относительные частоты событий A_i в генеральной совокупности. На рис.2.1 приведено возрастное распределение работников крупной компании, которое может рассматриваться как генеральная совокупность в случае, когда в исследованиях принимает участие небольшая часть работников этой компании (выборка), для которой требуется проверка согласованности возрастного распределения, чтобы можно было говорить о ее репрезентативности. При проверке согласованности выборочного распределения с теоретическим p_i – вероятности появления событий A_i для теоретического закона. Исходные данные – выборочное распределение $\{n_i\}_{i=1,\ldots,r}$ и ожидаемое распределение $\{np_i\}_{i=1,\ldots,r}$ – заносятся в таблицу.

Таблица 2.1. Подготовка данных для проверки гипотезы согласия

Ряд значений	Выборочное распределение	Ожидаемое распределение выборки
A_1	n_1	np_1
A_2	n_2	np_2
…	…	…
A_r	n_r	np_r
Сумма	n	n

На приведенных ниже рисунках выборочные частоты - это число наблюдений для каждой из возможных градаций дискретной переменной (общее число заболеваний) или число наблюдений в каждом из выбранных интервалов непрерывной переменной (возраст). Теоретические частоты $\{p_i\}_{i=1,\ldots,r}$ можно получить с помощью таблиц для соответствующих законов распределения (в данном случае - Пуассона и нормального).

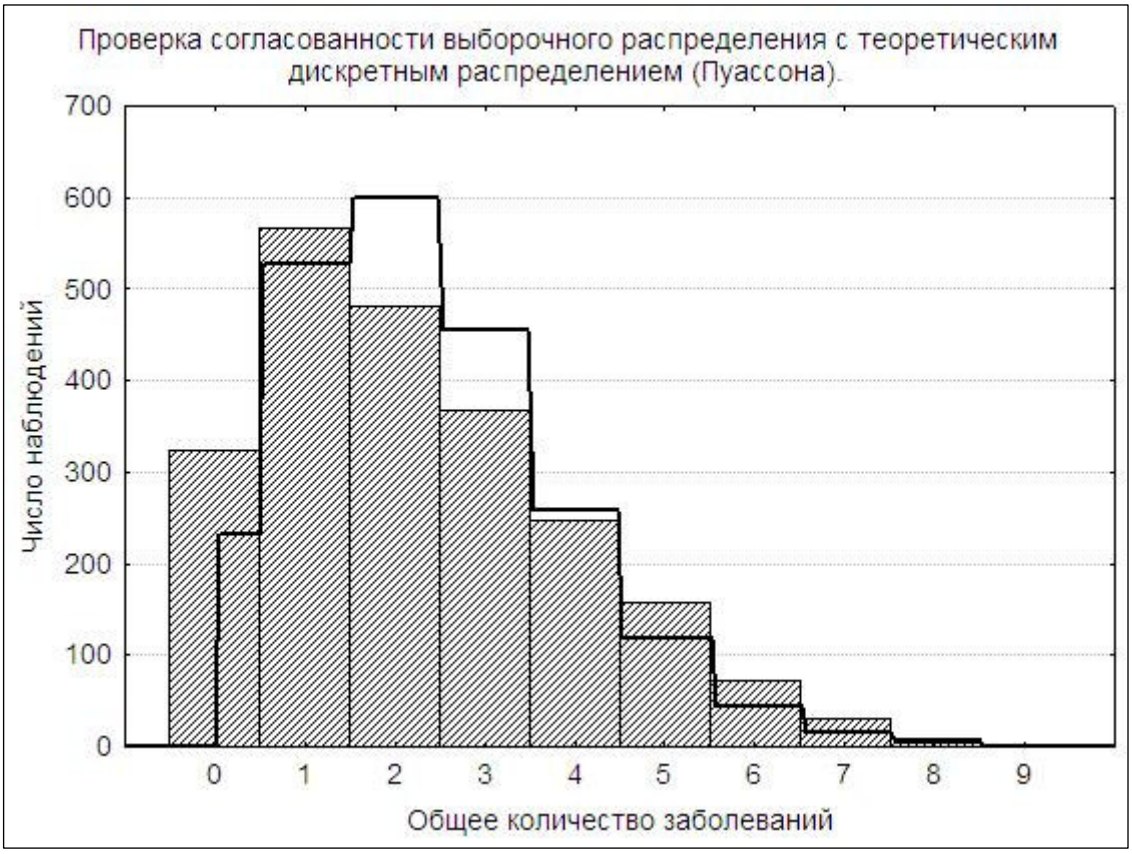


Рис.2.2. Гистограмма распределения общего количества заболеваний

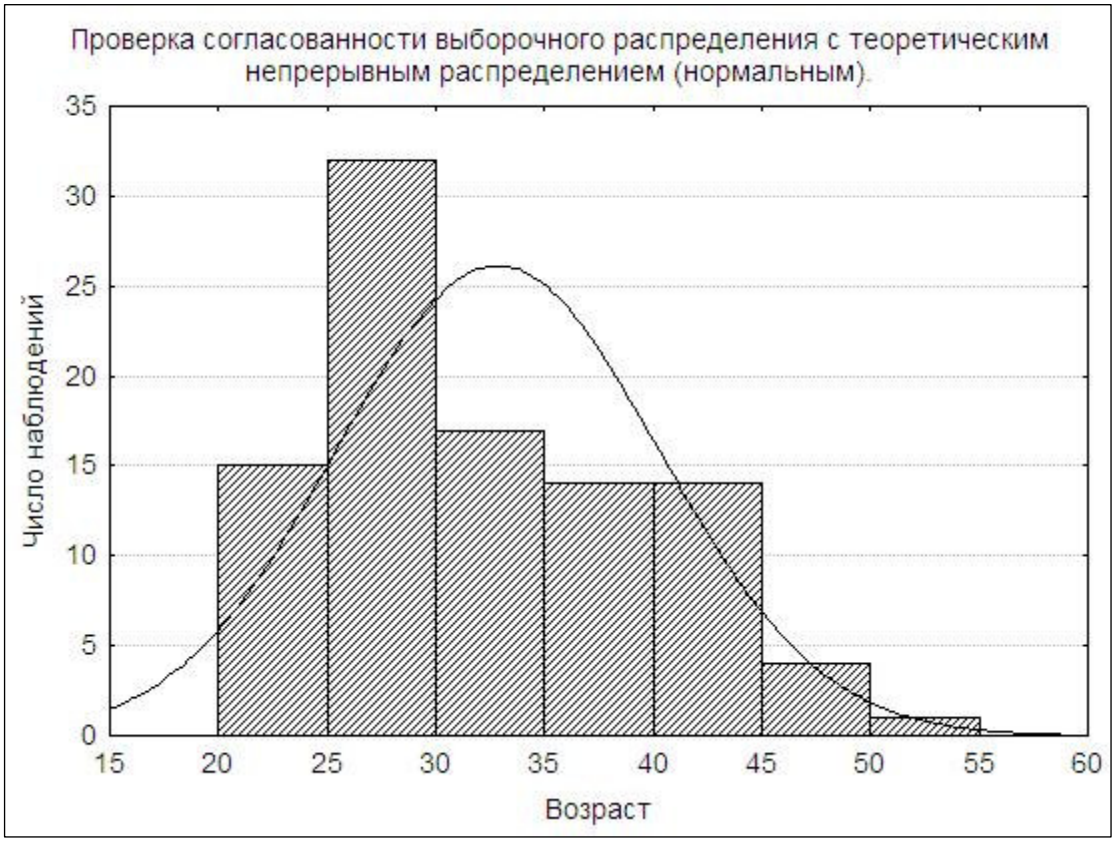


Рис.2.3. Гистограмма возрастного распределения выборки

Вычисляется выборочная статистика критерия

$$\chi^2_{в} = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} \quad (2.1)$$

где r – количество ячеек (разных значений для дискретного распределения или интервалов для непрерывного).

Число степеней свободы d вычисляется по формуле

$$d = r - l - 1 \quad (2.2),$$

где l – количество параметров теоретического распределения, которые оценены по выборке (например, среднее и среднеквадратическое отклонение для нормального закона, $l = 2$). При сравнении данных выборки с распределением генеральной совокупности $l = 0$. Если проверяется согласованность с теоретическим распределением, то $l \geq 0$. Число степеней свободы d должно быть не менее 1.

Для принятия решения о виде распределения выборки формулируется нулевая и альтернативная гипотезы.

Нулевая гипотеза $H_с$: выборка согласуется с теоретическим распределением.

Альтернативная гипотеза $H_{нс}$: выборка не согласуется с теоретическим распределением.

Гипотеза $H_с$ принимается на уровне α, если

$$\chi^2_{в} < \chi^2_{1-\alpha}(d) \quad (2.3)$$

Здесь $\chi^2_{1-\alpha}(d)$ – (1-α) квантиль распределения χ^2 с d степенями свободы.

Если неравенство (2.3) не выполнено, гипотеза отклоняется – принимается решение, что распределение выборки отличается от теоретического (альтернативная гипотеза).

Выбор уровня значимости определяет приемлемую для исследования вероятность ошибочно отклонить нулевую гипотезу: уровень значимости α – это вероятность того, что нулевая гипотеза $H_с$ верна, но при этом выборочное значение статистики $\chi^2_{в}$ больше квантили $\chi^2_{1-\alpha}(d)$, то есть, в соответствии с правилом (2.3), нулевая гипотеза будет отклонена.

Далее на примерах покажем, как применяется данный критерий.

(а) Проверка гипотезы о согласии распределения дискретного показателя с конечным числом возможных значений (исходов) A_1, A_2, …, A_r с распределением генеральной совокупности (популяции).

Пример 1. В распоряжении исследователей имеются данные о наблюдениях за группой взрослых испытуемых с различным образовательным статусом: 10 человек с высшим образованием, 15 человек со средним специальным образованием, 10 человек с общим средним образованием, 20 человек с неполным средним образованием и 5 человек с начальным образованием. В целях дальнейшего исследования требуется проверить согласованность имеющейся выборки по образовательному статусу со всем населением города.

В данном примере изучаемым признаком является образовательный статус человека. Он имеет 5 возможных значений (r=5). В качестве генеральной совокупности рассматривается население города.

Из статистических справочников получены данные по городу (в процентах): среди взрослого населения лиц с высшим образованием 11%, со средним специальным образованием 20%, с общим средним образованием 19%, с неполным средним образованием 38% и с начальным образованием 12%. Таблица для дальнейших вычислений имеет вид:

Таблица П1-1.

Образование	Выборочное распределение	Теоретическое распределение	Ожидаемое распределение
Высшее	10	0.11	6.6
Средн.специальное	15	0.20	12
Общее среднее	10	0.19	11.4
Неполное среднее	20	0.38	22.8
Начальное	5	0.12	7.2
Сумма	60	1	60

По формуле (2.1) получим: $\chi^2_в = 3.69$

Число степеней свободы $d = 5 - 1 = 4$.

На уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = \chi^2_{0.95}(4) = 9.49$
Неравенство (2.3) выполнено, существенных отличий выборочного распределения от генерального не обнаружено.

► Ограничения

При применении критерия χ^2 используется тот факт, что каждая из случайных величин $(n_k - np_k)/(np_k)^{1/2}$ имеет распределение, близкое к нормальному N(0,1). Эта аппроксимация следует из теоремы Лапласа. Поскольку данная теорема является предельной, то для достаточной корректности при применении критерия должны

выполняться некоторые условия, а именно: это утверждение достаточно точно, если

(С2.1) все ожидаемые частоты $\mathbf{np_k \geq 5}$.

Это наиболее строгое условие. Исследования Кокрейна позволили ему сформулировать более мягкие ограничения для некоторых типов задач:

(С2.1a) если проверяется согласие с одномодальным распределением, где ожидаемые частоты малы только на «хвостах» распределений, следует добиться минимальной ожидаемой частоты на каждом из «хвостов» не менее 1.

(С2.1b) при проверке согласия с непрерывным распределением (нормальным, логнормальным, экспоненциальным и т.д.) минимальные ожидаемые частоты на «хвостах» распределений должны быть не меньше 1, а размер ячеек должен быть выбран таким, чтобы частоты были не слишком велики. Это существенно для повышения чувствительности критерия. Например, для n=200 наблюдений максимальная частота должна быть не более 12, для n=400 максимальная частота $\leq$ 20, для n=1000 максимальная частота ≤ 30. ►

Пример 1 – продолжение. В таблице П1-1 все ожидаемые частоты (столбец «ожидаемое распределение») больше 5, поэтому в данной задаче применение критерия χ^2 правомерно.

В случае, если для какой-нибудь ячейки условия (С2.1 или С2.1a, С2.1b) не выполнены, для применения критерия χ^2 следует модифицировать исходную таблицу: объединить или разделить соседние ячейки, чтобы ожидаемая частота для событий удовлетворяла условиям. Данная операция особенно важна в том случае, когда выборочное значение статистики мало отличается от критического.

(б) Проверка согласованности распределения выборки с непрерывным (например, нормальным) законом распределения.

В этом случае весь диапазон значений переменной должен быть разбит на несколько непересекающихся интервалов, и для каждого интервала вычисляется количество элементов выборки, попавших в этот интервал – наблюдаемые частоты. Параметры непрерывного закона (для нормального закона это среднее и среднеквадратическое отклонение) могут быть известны заранее или оцениваться по выборке. Ожидаемое

распределение вычисляется в соответствии с непрерывным законом для каждого интервала значений.

Пример 2. Заданы следующие выборочные значения возраста испытуемых (вариационный ряд).

Таблица П2-1. Исходные данные

Возраст	20	21	22	24	26	27	29	30	33	34	36	37	39	42	44	47	Сумма
Количество	3	1	2	4	6	3	1	5	4	8	6	6	3	1	3	4	60

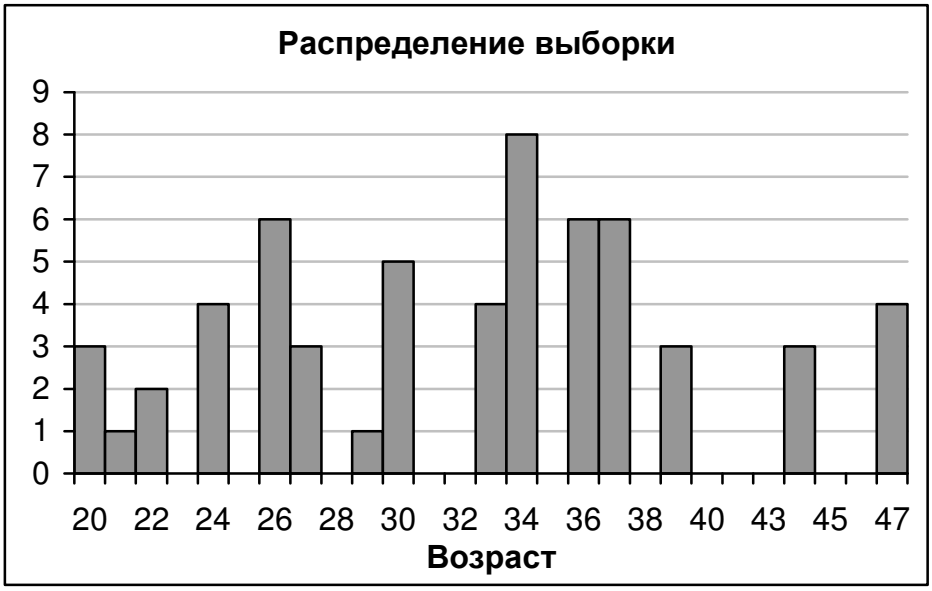


Рис. П2-1. Гистограмма возрастного распределения

Требуется проверить, согласуется ли распределение выборки с нормальным распределением на уровне $\alpha = 0.05$. Параметры нормального закона заранее не известны.

1. Оценим по выборке значения среднего и среднеквадратического отклонения.

Среднее значение $M_x = 32.63$

Дисперсия $D_x = 53.66$

Среднеквадратическое отклонение $s_x = (D_x)^{1/2} = 7.325$

По выборке оценены два параметра распределения, $l = 2$.

2. Разобьем весь диапазон значений на 7 стандартных интервалов с граничными точками 20, 25, 30, 35, 40 и 45 лет и вычислим выборочные и теоретические частоты.

Для вычисления теоретических частот можно воспользоваться таблицами для функции распределения стандартизованной нормальной величины – функции Лапласа $\Phi(x)$. Тогда $P\{x_1 < x \le x_2\} = \Phi[(x_2 - M_x)/s_x] - \Phi[(x_1 - M_x)/s_x]$. Определим соответствующие границам выделенных интервалов точки стандартизованного распределения и значения

функции Лапласа для них. Для отрицательных значений используется соотношение $\Phi(-x) = 1-\Phi(x)$.
Таблица П2-2.

Возраст	Стандартизованные значения	Ф(х)
20	-1.72	1-0.9573
25	-1.04	1-0.8508
30	-0.36	1-0.6406
35	0.32	0.6255
40	1.01	0.8438
45	1.69	0.9545

Таблица П2-3.

Возраст	Выборочное распределение	Теоретическое распределение	Ожидаемые значения
≤ 20	3	0.0427	2.56
21-25	7	0.1065	6.39
26-30	15	0.2102	12.61
31-35	12	0.2661	15.97
36-40	15	0.2183	13.10
41-45	4	0.1107	6.64
> 45	4	0.0455	2.73
Сумма	60	1	60

3. Вычислим статистику критерия: $\chi^2_в = 3.48$.
Степеней свободы $d = 7-2-1 = 4$.
На уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = \chi^2_{0.95}(4) = 9.49$
Выборка согласуется с нормальным распределением в соответствии с неравенством (2.3), причем выборочное значение статистики существенно меньше критического.

Проверим выполнение условий применимости критерия: C2.1 или C2.1a, C2.1b.
В таблице П2-3 ожидаемые частоты на хвостах распределений более 1. Однако в интервале 26-40 лет ожидаемые частоты слишком велики для 60 наблюдений. Это означает, что мы производим очень грубое сравнение выборочного распределения с нормальным вокруг среднего.

Для увеличения чувствительности критерия разобьем ячейки на более мелкие части: диапазон с 25 до 40 лет разделим на интервалы по 3 года. В таблице станет r = 9 ячеек.

Таблица П2-4.

Возраст	Выборочное распределение	Теоретическое распределение	Ожидаемые значения
≤ 20	3	0.0427	2.56
21-25	7	0.1065	6.39
26-28	9	0.1151	6.91
29-31	6	0.1486	8.92
32-34	12	0.1624	9.74
35-37	12	0.1504	9.02
38-40	3	0.1156	6.94
41-45	4	0.1132	6.79
> 45	4	0.0455	2.73
Сумма	60	1	60

$\chi^2_в$= 7.198; d = 9-2-1 = 6; $\chi^2_{0.95}(6)$ = 12.6. Выборка согласуется с нормальным распределением.

Если операцию разделения или объединения ячеек произвести не удается, применение критерия χ^2 для такой задачи будет некорректным и требуется использование других методов.

Расчеты с использованием статистических пакетов.

Для проверки гипотезы согласия с помощью критерия χ^2 можно использовать стандартные программы.

Statistica v.6.0 → **Distribution Fitting.** Проверяется согласие с одним из известных распределений: нормальное, лог-нормальное, экспоненциальное, хи-квадрат, - одной из переменных исходной таблицы данных. Пользователь может задать число интервалов, минимальное и максимальное значения. Вычисляются наблюдаемые и ожидаемые значения.

В остальных программах для проверки согласия используются другие статистики (см.Словарь – статистика).

2.3. Проверка гипотезы однородности H_o

Если имеется два или более выборочных распределений (серий) одного и того же показателя с множеством возможных значений A_1, A_2, …, A_k (B_1, B_2, …, B_l – номера или названия отдельных серий), то можно ли утверждать, на некотором уровне значимости, что частоты появления событий A_i в этих сериях совпадают, т.е. что серии выбраны из одной генеральной совокупности? В данном случае проверяется гипотеза H_o: серии B_1, B_2, …, B_l одинаково распределены (гипотеза об однородности выборок). Эта гипотеза может проверяться как относительно всех серий в совокупности, так и относительно каждой пары серий в отдельности. Альтернативная гипотеза $H_{но}$: распределения в сериях отличаются. На Рис.2.4 в качестве серий выступают различные возрастные группы, а событие А – оценка достоверности ответов на вопросы MMPI. Эта оценка имеет 3 градации: A_1 – недостоверно; A_2 – сомнительно; A_3 – достоверно.

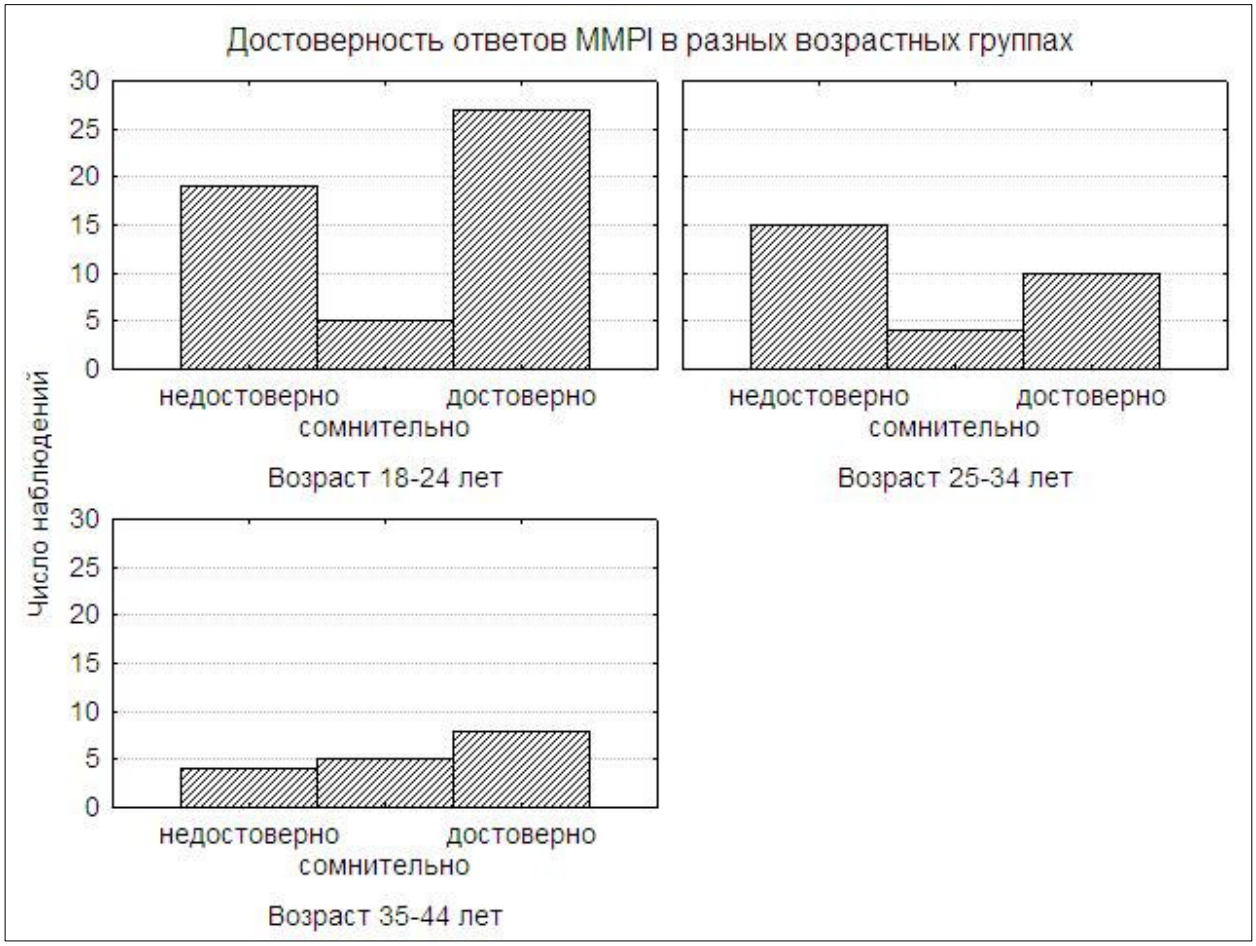


Рис.2.4. Гистограммы распределений достоверности ответов в разных возрастных группах

Выборочные распределения обычно представлены в виде таблиц сопряженности. Их элементами являются частоты {nij}i=1÷k, j=1÷l (Таблица 2.2).

Таблица 2.2. Таблица сопряженности факторов A и B

Ряды значений	B_1	B_2	…	B_l	Сумма по строке
A_1	n_{11}	n_{12}	…	n_{1l}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	…	n_{2l}	$n_{2\bullet}$
…	…	…	…	…	…
A_k	n_{k1}	n_{k2}	…	n_{kl}	$n_{k\bullet}$
Сумма по столбцу	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet l}$	N

Оценкой вероятности появления события A_i является его относительная частота $\tilde{p}_i = n_{i\bullet} / N$; для серии B_j это относительная частота $\tilde{q}_j = n_{\bullet j} / N$. Нулевая гипотеза – гипотеза об однородности – утверждает, что вероятность осуществления события A_i в серии B_j есть произведение вероятностей их появления, $\tilde{p}_{ij} = \tilde{p}_i \bullet \tilde{q}_j$, то есть ожидаемые частоты в ячейке

$$\tilde{n}_{ij} = n_{i\bullet} \times n_{\bullet j} / N \quad (2.4)$$

Статистика критерия – мера отклонения наблюдаемых частот от ожидаемых

$$\chi^2_{в} = \Sigma_{i=1}^{k} \Sigma_{j=1}^{l} (n_{ij} - \tilde{n}_{ij})^2 / \tilde{n}_{ij} \quad (2.5)$$

В предположении нулевой гипотезы критерий распределен как $\chi^2(d)$, где

$$d = (k-1) \times (l-1) \quad (2.6)$$

Гипотеза об однородности выборок (одинаковом распределении) принимается на уровне α, если

$$\chi^2_{в} < \chi^2_{1-\alpha}(d)$$

В противном случае гипотеза отклоняется (принимается альтернативная гипотеза).

► Ограничения

(С2.2) Критерий применим, если все ожидаемые частоты $\tilde{n}_{ij} \geq 4$.
Или, если объем выборки и количество ячеек в таблице сопряженности достаточно большие, то минимальная ожидаемая частота может быть равна 1, то есть
(С2.3) Критерий применим, если $d \geq 8$ и $N \geq 40 \Rightarrow \tilde{n}_{ij} \geq 1$. ►

В статистических пакетах в качестве условия применимости критерия используются ограничения

►(С2.2а) Критерий применим, если не более чем 20% ожидаемых частот в таблице меньше 5.

Пример 3. Для испытуемых из примера 1 кроме образовательного статуса известен пол. Вопрос: отличаются ли распределения по образовательному статусу для представителей разного пола в выборке? Таблица перекрестного табулирования по двум признакам для дальнейших вычислений:

Таблица П3-1.Исходные данные

Пол / Образование	Мужчины	Женщины	Сумма по строке
Высшее	4	6	10
Среднее специальное	10	5	15
Общее среднее	3	7	10
Неполное среднее	12	8	20
Начальное	4	1	5
Сумма по столбцу	33	27	60

Можно сказать, что в данной задаче требуется сравнить структуру образования у женщин и мужчин.

Задачи сравнения структур часто встречаются при анализе заболеваемости и смертности у различных групп населения.

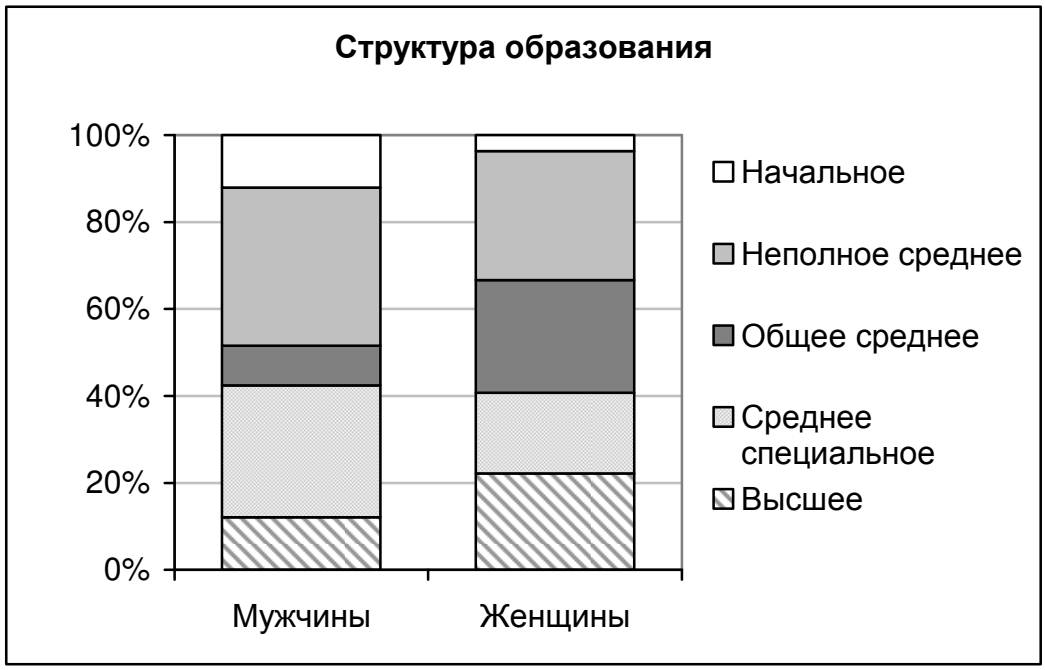


Рис. П3-1. Структура образования у мужчин и женщин в выборке

По таблице П3-1 вычислим ожидаемые частоты $\tilde{n}_{ij}$ (формула 2.4).

Таблица П3-2. Ожидаемые частоты.

Пол / Образование	Мужчины	Женщины	Сумма
Высшее	5.5	4.5	10
Среднее специальное	8.25	6.75	15
Общее среднее	5.5	4.5	10
Неполное среднее	11	9	20
Начальное	2.75	2.25	5
Сумма	33	27	60

Статистика критерия $\chi^2_в = 5.724$

Степеней свободы d=4

На уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = \chi^2_{0.95}(4) = 9.49$

Неравенство (2.3) выполнено, распределение показателя «образование» для мужчин и женщин не отличается, но минимальное ожидаемое значение 2.25. Ограничение (С2.3) не выполнено, поэтому для данной задачи требуется или использовать другой критерий, или объединить ячейки.

При объединении последних двух ячеек в признаке «образование» получим следующие результаты.

Таблица П3-3.

Пол / Образование	Мужчины	Женщины	Сумма
Высшее	4	6	10
Среднее специальное	10	5	15
Общее среднее	3	7	10
Неполное среднее или начальное	16	9	25
Сумма	33	27	60

Статистика критерия $\chi^2_в = 5.077$

Степеней свободы d=3

На уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = \chi^2_{0.95}(3) = 7.815$.

Неравенство (2.3) выполнено, распределение показателя «образование» однородно по «полу» на уровне значимости 0.05, минимальное ожидаемое значение 4.5, то есть ограничения выполнены, критерий применим.

Частным, но очень важным случаем является проверка равенства **пропорций**. Такая задача возникает, если интересующий нас показатель имеет ровно два возможных значения (жив – умер; нет заболевания – есть заболевание; курит – не курит). В этом случае показатель распределен по биномиальному закону. Биномиальное распределение определяется одним параметром – вероятностью появления события в одном испытании. Пропорции являются оценками параметра. Если требуется проверить равенство параметров двух биномиальных распределений, критерий χ^2 применяется к таблице сопряженности 2×2 (таблица 2.3).

Таблица 2.3. Таблица сопряженности для биномиальных факторов

Серия / Значение	1	2	Сумма
A	n_{11}	n_{12}	$n_{1\bullet}$
не A	n_{21}	n_{22}	$n_{2\bullet}$
Сумма	$n_{\bullet 1}$	$n_{\bullet 2}$	N

Гипотеза p1= p2 (частота события A в первой серии равна его частоте во второй серии) эквивалентна гипотезе о том, что выборки извлечены из одной генеральной совокупности, т.е. однородны. При

этом формула вычисления выборочного значения критерия принимает вид:

$$\chi^2_{в}= (N - 1)\cdot(n_{11}\cdot n_{22} - n_{12}\cdot n_{21})^2 / n_{\cdot 1}\cdot n_{\cdot 2}\cdot n_{1\cdot}\cdot n_{2\cdot}; \; d=1. \quad (2.7)$$

► Ограничения

(С2.4) Для таблицы 2×2 критерий χ^2 можно использовать, если объем выборки и ожидаемые частоты удовлетворяют следующим условиям: **при N > 20, все ожидаемые частоты $\bar{n}_{ij}$ должны быть > 3 ;**

(С2.5) **если N ≤ 20, наблюдений в первой серии $n_{\cdot 1}$ должно быть > 5, а во второй $n_{\cdot 2} > n_{\cdot 1} / 3$.**

Кокрейн рекомендует для выборок с числом наблюдений **N ≤ 20** использовать точный критерий Фишера. ►

Проверка гипотез об однородности для таблиц k×l с проверкой условий применимости критерия χ^2 осуществляется в программе chi_sq_ru.stb.

Расчеты с использованием статистических пакетов.

SPSS → Descriptive Statistics → Crosstab,
NCSS → Descriptive Statistics → Crosstab,
SYSTAT → Tables → Crosstab → Two-Way,
Statistica v.6.0 **→ Basic Statistics/Tables → Tables and Banners**

В этих программах по переменным исходной таблицы данных вычисляется статистика критерия, проверяется гипотеза однородности показателей.

2.4. Проверка гипотезы независимости $H_н$

В том случае, когда имеется одна выборка, у которой зафиксированы значения двух показателей, A и B, с множествами возможных значений A_1, A_2, …, A_k и B_1, B_2, …, B_l соответственно, Таблица 2 является таблицей перекрестного табулирования. Тогда обычно требуется ответить на вопрос: зависят ли показатели A и B, т.е. зависят ли частоты появления событий A_i от того, какому уровню j показателя B (B_j) они соответствуют? При такой постановке задачи будет проверяться нулевая гипотеза $H_н$: предположение, что показатели A и B независимы при заданном уровне значимости. Проверка этой гипотезы осуществляется так же, как и проверка гипотезы об однородности распределений (формулы (2.4) – (2.7)). Альтернативная гипотеза $H_{зав.}$: показатели зависимы, распределения частот A_i в

различных столбцах таблицы отличаются. Ограничения, приведенные в предыдущем пункте (С2.2 – С2.5), также должны соблюдаться при проверке гипотезы о независимости с помощью критерия χ^2.

В случае, если гипотеза о независимости отвергается, характеристикой величины связи между показателями может быть один из следующих коэффициентов связи, вычисляемых на основе статистики χ^2 : φ, С (коэффициент контингации) или V Крамера.

$$\varphi = \sqrt{\frac{\chi^2}{n}} \ , \ C = \sqrt{\frac{\chi^2}{n+\chi^2}} \ , \ V = \sqrt{\frac{\chi^2}{n*(q-1)}} \ , \text{ где q = min (k, l)}$$

Все коэффициенты, как правило, имеют значения от 0 до 1, хотя коэффициент φ может и превышать 1 в некоторых случаях, а С и V не достигают значения 1 (приближаются асимптотически). Значение 0 показывает отсутствие связи показателей (независимость), статистическая значимость всех коэффициентов определяется статистикой χ2в - если на выбранном уровне отвергается гипотеза о независимости, то на том же уровне коэффициенты φ, С, V отличаются от 0.

Пример 4. При обследовании работников нефтедобывающей компании в целях медицинского страхования был вычислен индекс функциональных изменений (ИФИ). Все обследованные были разбиты на 3 возрастные группы в соответствии с возрастной структурой контингента: (1) до 40 лет; (2) 40 – 49 лет; (3) 50 и более лет. Известна также профессиональная принадлежность каждого обследованного.

Таблица П4-1. Таблица сопряженности для ИТР и администрации

Градации ИФИ	Возрастные группы			Всего
	(1) до 40 лет	(2) 40 – 49 лет	(3) 50 и более лет	
Удовлетворительная адаптация.	57	12	3	72
Напряжение механизмов адаптации	26	36	16	78
Неудовлетворительная адаптация	9	13	17	39
Срыв адаптации	2	12	22	36
Всего	94	73	58	225

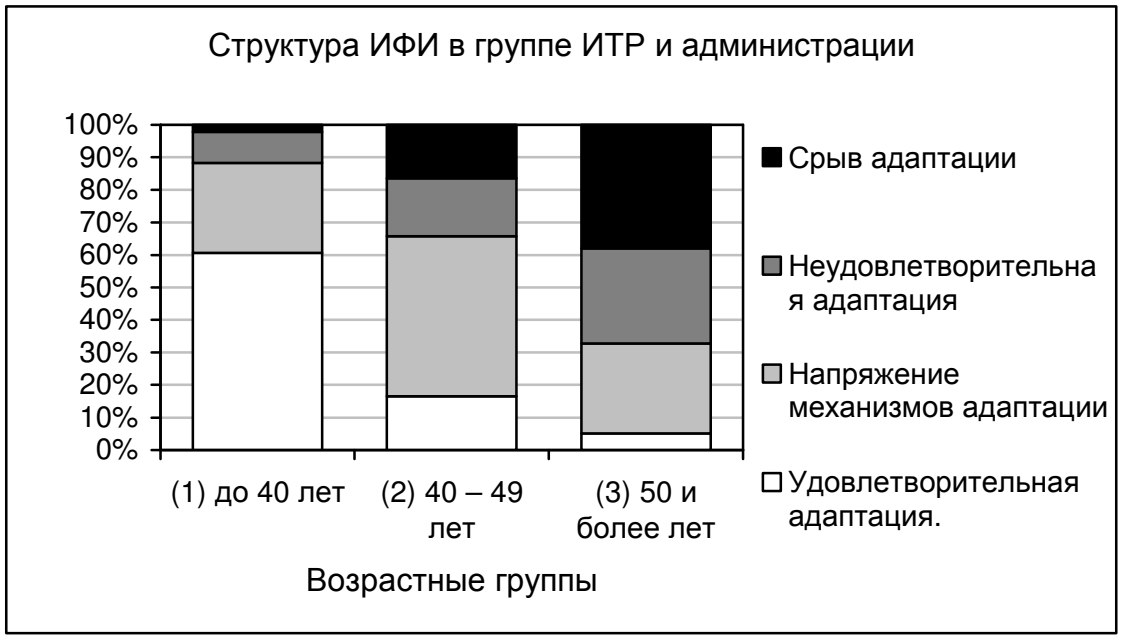


Рис. П4-1. Структура ИФИ в разных возрастных группах для ИТР и администрации

Требуется выяснить, есть ли связь между возрастом и градациями ИФИ в каждой профессиональной группе.

Результаты применения программы chi_sq_ru.stb

Критерий χ^2 применим (ограничения выполнены)

Наблюдений n = 225, степеней свободы d = 6, минимальное ожидаемое значение 9.28,

$\chi^2_в = 86.17$, на уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = 12.59$.

Гипотеза независимости отвергается, показатели зависимы.

Коэффициент связи $\varphi = 0.62$

Вывод: в профессиональной группе «ИТР и администрация» связь между показателями «градации ИФИ» и «возрастные группы» существует, причем довольно сильная.

Те же результаты можно получить, применяя стандартную процедуру статистического пакета Statistica, однако при этом не проводится проверка применимости критерия к данной таблице, пользователь должен осуществить ее самостоятельно.

Таблица П4-2. Таблица сопряженности для профессиональной группы «рабочие»

Градации ИФИ	Возрастные группы			Всего
	(1) до 40 лет	(2) 40 – 49 лет	(3) 50 и более лет	
Удовлетворительная адаптация	22	13	3	38
Напряжение механизмов адаптации	11	11	5	27
Неудовлетворительная адаптация	7	3	2	12
Срыв адаптации	1	4	4	9
Всего	41	31	14	86

Наблюдений n = 86, степеней свободы d = 6, минимальное ожидаемое значение 1.46,

$\chi^2_в = 10.81$, на уровне $\alpha = 0.05$ квантиль $\chi^2_{1-\alpha}(d) = 12.59$.

Применение критерия χ^2 неадекватно задаче (ожидаемые значения малы).

Исходя из этого, делать дальнейшие выводы о наличии или отсутствии связи показателей с помощью критерия χ^2 не следует.

Для Таблицы П4-2 программы Statistica и SYSTAT не дают предупреждений о неприменимости критерия χ^2, поэтому на основании сравнения выборочного значения статистики и ее критического значения можно сделать вывод о независимости показателей.

Программы SPSS и NCSS предупреждают о наличии слишком малых ожидаемых значений: At least one cell had an expected value less than 5.

Все стандартные программы могут по запросу пользователя вывести таблицу ожидаемых значений, но делать дальнейшие выводы и осуществлять проверку других ограничений должен сам исследователь.

Для того, чтобы получить решение поставленной задачи, можно объединить 3-ю и 4-ую строки Таблицы П4-2. Получится Таблица П4-3.

Таблица П4-3. Таблица сопряженности для профессиональной группы «рабочие»

Градации ИФИ	Возрастные группы			Всего
	(1) до 40 лет	(2) 40 – 49 лет	(3) 50 и более лет	
Удовлетворительная адаптация	22	13	3	38
Напряжение механизмов адаптации	11	11	5	27
Неудовлетворительная адаптация или срыв	8	7	6	21
Всего	41	31	14	86

Однако и в этом случае получим:

Наблюдений n = 86, степеней свободы d = 4, минимальное ожидаемое значение 3.42.

Применение критерия χ^2 неадекватно задаче (ожидаемые значения малы).

Значит, нужно применить другой метод для выявления связи между показателями.
В данном случае оба показателя не являются номинальными в точном понимании, их значения естественным образом упорядочены. Поэтому, в соответствии со Схемой 2, можно использовать коэффициенты ранговой корреляции. Поскольку таблица небольшая (4×3), повторяющихся значений много (до 22 в ячейке), лучше всего здесь применить коэффициент γ (см. Приложение).

С помощью блока **Nonparametrics** в программе **Statistica** получим:

Gamma = 0.33, p-level < 0.003.

Таким образом, можно сказать, что существует ранговая связь между исследуемыми показателями в профессиональной группе «рабочие».
Для сравнения силы ранговой связи показателей у «рабочих» и «ИТР-администрации» вычислим коэффициент ранговой корреляции для таблицы П4-1.
Gamma = 0.64, p-level < 0.00001.

Таким образом, связь в группе «рабочие» значительно слабее, нежели в группе «ИТР-администрация».
Различные коэффициенты связи, в данном случае φ и γ, по абсолютной величине находятся в одном и том же диапазоне – от 0 до 1, но

вычисляются они разным образом. Поэтому для сравнения силы связи в нескольких группах лучше вычислять коэффициенты одного вида для всех групп.

Расчеты с использованием статистических пакетов.

Расчеты производятся теми же средствами, как в случае проверки гипотезы однородности.

2.5. Проверка гипотезы наличия линейного тренда $H_т$.

При анализе изменения параметров ряда биномиальных распределений (обработка таблиц перекрестного табулирования двух показателей, из которых первый имеет ровно два возможных значения, а значений второго показателя более двух и они упорядочены) иногда требуется оценить, имеется ли тренд возрастания или убывания параметров. В качестве примера можно привести анализ тренда уровней смертности в зависимости от возраста или дозовой нагрузки.

В этой задаче критерий χ^2 применяется к таблице сопряженности $2\times k$ (таблица 2.4), причем каждой серии поставлена в соответствие дозовая нагрузка $x_1, x_2, \ldots, x_k$. В зависимости от смысла задачи числа $\{x_i\}_{i=1,\ldots,k}$ могут представлять собой или последовательные натуральные числа, или середину возрастного интервала, или соответствующую серии дозу препарата и т.д. Линейный тренд – это регрессия пропорций $\{n_{1i} / n_{\bullet i}\}$ на дозы $\{x_i\}$.

Грубый анализ данных таблицы состоит в проверке гипотезы о том, что пропорции, соответствующие k столбцам таблицы, не отличаются (формула (2.5), вычисление $\chi^2_в(k-1)$). Однако при таком анализе не используется информация об упорядоченности столбцов по дозовой нагрузке.

Таблица 2.4. Подготовка данных для проверки линейного тренда

Серия / Значение	1	2	…	k	Сумма
A	n_{11}	n_{12}	…	n_{1k}	$n_{1\bullet}$
не A	n_{21}	n_{22}	…	n_{2k}	$n_{2\bullet}$
Сумма	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet k}$	N
Дозовая нагрузка	x_1	x_2	…	x_k	

Для проверки нулевой гипотезы H_0: в пропорциях отсутствует линейный тренд против альтернативной гипотезы $H_т$: есть линейный тренд - вычисляется статистика критерия

$$\chi^2_в = \frac{(\sum_{i=1}^{k} x_i * n_{1i} - n_{1\bullet} * \sum_{i=1}^{k} \frac{x_i * n_{\bullet i}}{N})^2}{\frac{n_{1\bullet}}{N} * (1 - \frac{n_{1\bullet}}{N}) * (\sum_{i=1}^{k} n_{\bullet i} * x_i^2 - N * (\sum_{i=1}^{k} \frac{x_i * n_{\bullet i}}{N})^2)} \qquad (2.8)$$

В предположении нулевой гипотезы критерий распределен как $\chi^2(1)$.

Гипотеза об отсутствии тренда принимается на уровне α, если

$$\chi^2_в < \chi^2_{1-\alpha}(1)$$

В противном случае гипотеза отклоняется (принимается альтернативная гипотеза о наличии линейного тренда).

Разность $\chi^2_в(d-1)$ и $\chi^2_в(1)$, распределенная как $\chi^2(d-2)$, используется для проверки значимости отклонения пропорций от линейного тренда.

Пример 5. (Данные Т.И.Шевченко, НИС Медицинский регистр). Проведено психологическое тестирование 97 пожарных по методике ТОР Залевского. Было обнаружено увеличение доли лиц с высоким уровнем актуальной ригидности (АР) при возрастании возраста и стажа обследованных. С учетом возраста и стажа сформировано 5 групп. Получена следующая таблица.

Таблица П5-1

Номер / Значение	1	2	3	4	5	Сумма
Высокий уровень АР	1	3	2	5	7	18
Низкий и умеренный уровень АР	23	23	9	14	10	79
Сумма	24	26	11	19	17	97
% лиц с высоким уровнем АР	4.2	11.5	18.2	26.3	41.2	
Код группы по стажу и возрасту	1	2	3	4	5	

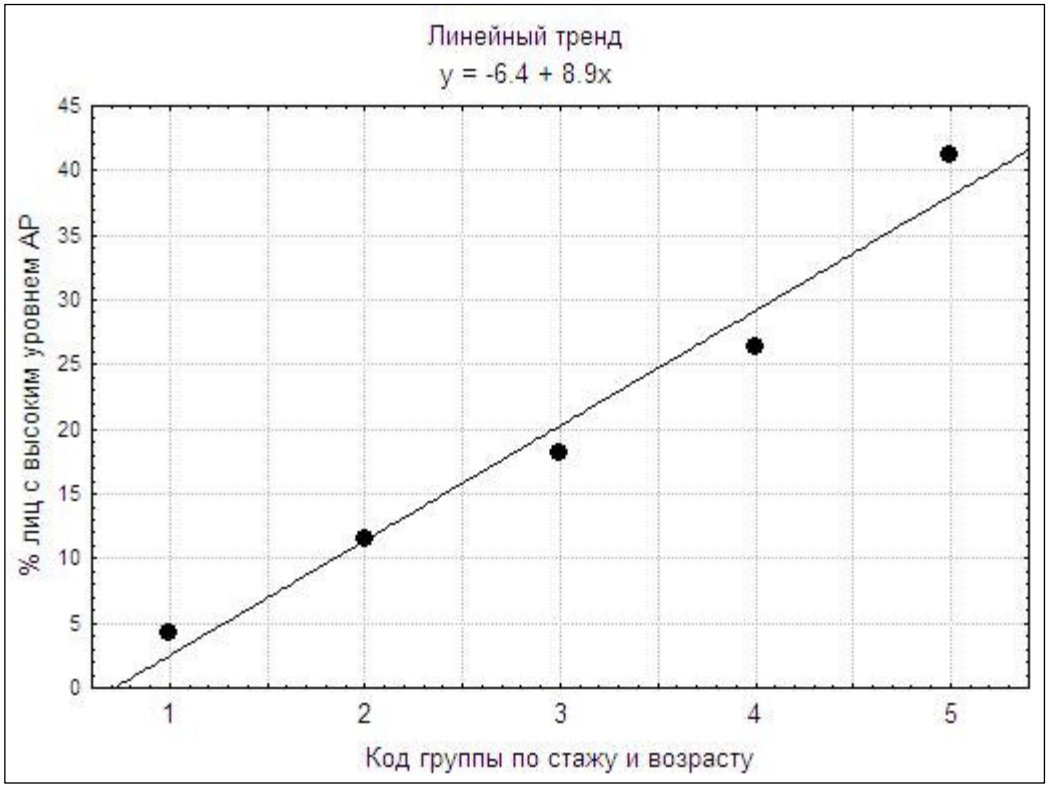


Рис. П5-1. Линейный тренд по коду группы

Результаты проверки наличия линейного тренда с помощью статистических программ.

SYSTAT →Tables → Crosstabs → Two-way… (задать **Cochran's test of linear trend** в выборе **Statistics…**)

Test statistic	Value	df	Prob
Pearson Chi-square	10.649	4	0.031
Cochran's Linear Trend	10.366	1	0.001

WARNING: More than one-fifth of fitted cells are sparse (frequency < 5).
Significance tests computed on this table are suspect.

В первой строке таблицы – статистика Пирсона хи-квадрат для проверки однородности столбцов таблицы (с 4 степенями свободы).
Во второй строке, с названием «линейный тренд Кокрейна», статистика Пирсона хи-квадрат для проверки наличия линейного тренда (с 1 степенью свободы).
Выведено предупреждение о том, что в таблице слишком много малых частот.
Тесты показывают значимое отклонение от равенства пропорций по столбцам и наличие линейного тренда.

Более четко данные выводы приведены в программе NCSS, но в этой программе проверка наличия линейного тренда проводится с помощью другой статистики (описание в Приложении).

NCSS → Descriptive Statistics → Cross Tabulation

Armitage Test for Trend in Proportions		
Ho: p1 = p2 = p3 = ... = pk		
Armitage S	-666	
Standard Error of S	210.2	
Z-Value (Standardized S)	-3.168	
Prob (Ha: Increasing Trend – *альтернативная гипотеза – возрастающий тренд*)	0.001	Reject Ho - *отвергается нулевая гипотеза*
Prob (Ha: Decreasing Trend)	0.999	Accept Ho
Prob (Ha: Any Trend)	0.002	Reject Ho
WARNING: At least one cell had an expected value less than 5.		

В данном тесте, в отличие от статистики Пирсона хи-квадрат, проверяются также и односторонние альтернативные гипотезы: о наличии увеличивающегося и уменьшающегося линейного тренда. В данном примере подтверждено наличие возрастающего тренда.

ГЛАВА 3. СРАВНЕНИЕ ЧАСТОТ СОБЫТИЙ

Частным, но наиболее распространенным случаем задачи о сравнении распределений дискретных показателей является получение статистических выводов о частоте появления определенного события. В качестве события может рассматриваться наличие конкретного заболевания, например ИБС, у отдельного испытуемого из выборки или когорты; наличие определенных особенностей заболевания у каждого наблюдаемого из группы больных, - например, наличие очаговых изменений в легких у больных ХОБЛ. Как правило, исследователя интересует не только частота события, но и сравнение ее с частотой в другой выборке или в популяции.

На Схеме 5 приведены основные типы задач. При внешней схожести задач, соответствующих типам I (выборка-популяция) и II (выборка-выборка), способы их решения существенно отличаются.

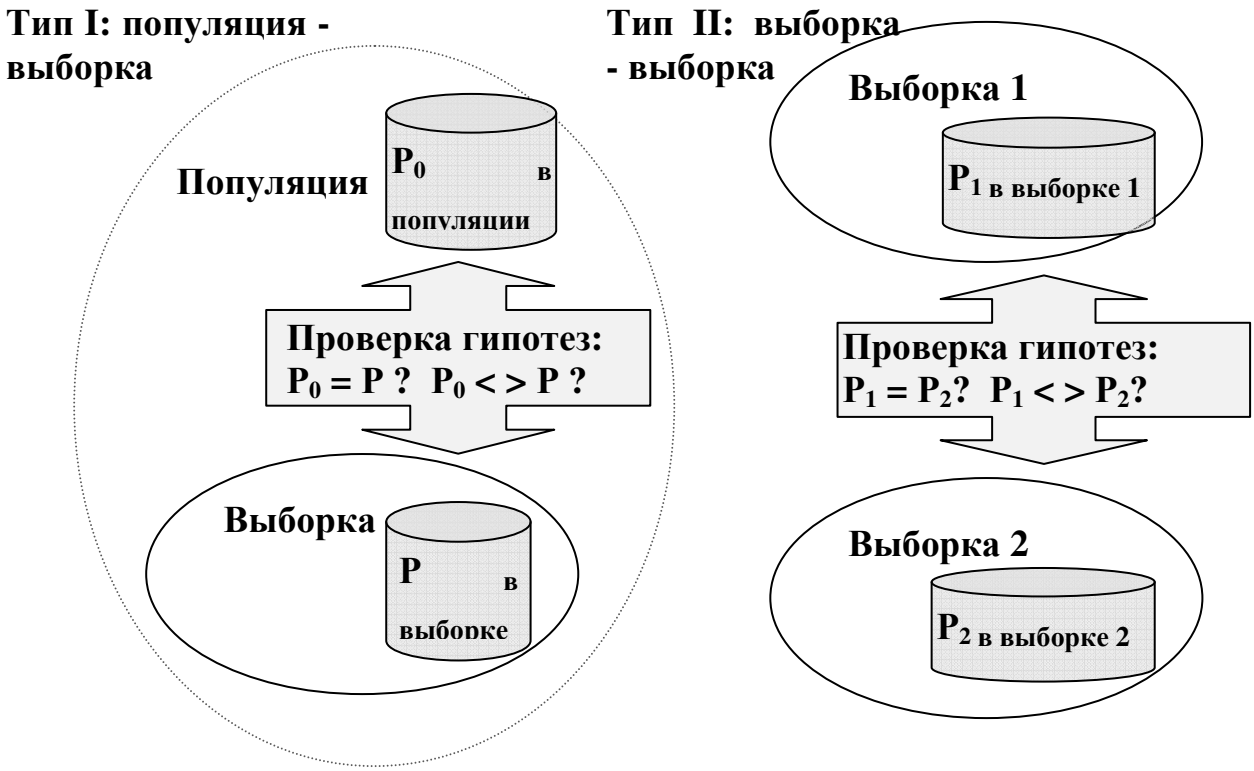


Схема 5. Основные типы задач сравнения частот событий

Статистически решение этих задач сводится к оценке параметров биномиальных распределений, то есть вероятности осуществления некоторого события (А) по данным одной или нескольких серий испытаний (выборок), и сравнении этих параметров между собой или с

параметром популяции. Для этого нужно осуществить проверку гипотез о величине и соотношении параметров.

3.1. Оценка параметров биномиальных распределений и проверка гипотез

Формально задачи I типа сводятся к следующему описанию:

►Проведена серия из n испытаний, в которой событие A появилось x раз. Согласуется ли частота появления события в данной серии с заранее известной частотой p? Как правило, слова «серия из n испытаний» означают, что у нас есть выборка объема n.

Для оформления задач используется таблица 3.1, являющаяся частным случаем таблицы 2.1 (глава 2).

Таблица 3.1. Подготовка данных для проверки гипотезы согласия

Ряд значений	Выборочное распределение	Ожидаемое распределение – распределение в выборке, если бы частота события A равнялась p.
A	x	np
не A	n-x	n(1-p)
Сумма	n	n

Пример 6. В поселке N в течение года родились 20 младенцев, из них 3 мальчика. Известно, что в среднем вероятность появления на свет мальчика в генеральной совокупности (по стране) равна 0.51. Можно ли утверждать, что выборочное распределение согласуется с генеральным (распределением в популяции)?

Пример 7. (Данные НРЭР по Северо-Западному региону РФ и Комитета по здравоохранению администрации СПб).

Среди ликвидаторов СПб в возрастной группе 40-44 лет в 2000 г. умер 1 человек из 501 наблюдаемых. В базовом распределении (по СПб, аналог генеральной совокупности) повозрастной уровень смертности для мужчин 40-44 лет составил в 2000 г. составил 13 человек на 1000 населения. Можно ли утверждать, что выборочное распределение согласуется с базовым (генеральным) на уровне 0.05?

Задачи II типа могут быть сформулированы так:

►Проведены две серии испытаний. Можно ли утверждать, что частоты появления события A в этих сериях совпадают, т.е. что эти серии выбраны из одной генеральной совокупности, распределенной по биномиальному закону?

►Проведено несколько серий испытаний в разных условиях. Можно ли утверждать, что условия проведения испытаний повлияли на частоту появления события А?

Таблица 3.2 для этих задач является частным случаем таблицы 2.2 (глава 2).

Таблица 3.2. Подготовка данных для проверки гипотезы однородности

Ряд значений	Выборочное распределение 1	Выборочное распределение 2	Выборочное распределение 3
А	x_1	x_2	x_3
не А	n_1-x_1	n_2-x_2	n_3-x_3
Сумма	n_1	n_2	n_3

Пример 8. (данные НРЭР по Северо-Западному региону РФ)
Среди ликвидаторов СПб в возрастной группе 40-44 лет в 1990 г. умерло 8 человек из 875 наблюдаемых, в 1995 г. умерло 6 человек из 672 наблюдаемых и в 2000 г. умер 1 человек из 501 наблюдаемого. Можно ли утверждать, на уровне значимости 0.05, что за этот период повозрастной уровень смертности не менялся?

В примерах 6, 7 и 8 подлежит оценке параметр частоты встречаемости некоторого события в одной или нескольких выборках и сравнение полученных оценок между собой или с другими, известными заранее, величинами. В качестве модели используется выборка из генеральной совокупности, описываемой биномиальным законом распределения, т.е. при решении указанных и подобных им задач мы будем предполагать, что для каждого элемента выборки интересующее нас событие может осуществиться с одной и той же определенной вероятностью. Можно сказать, что моделирование биномиальным распределением применимо, если изучаемое явление представляет собой характеристику или качество каждого элемента выборки: или оно есть, или его нет (в примерах – для человека это пол или факт смерти). Поэтому при сравнении уровней заболеваемости в нескольких выборках методы, основанные на оценках параметра биномиального распределения, применимы только в случаях бесповторных заболеваний (редких или хронических).

Для решения задач I и II типов применяются различные методы расчетов. Они могут быть осуществлены вручную, с помощью формул, приведенных в «Приложении», или же с помощью стандартных статистических программ. Здесь описаны расчеты с использованием

пакетов SPSS, NCSS, SYSTAT, STATISTICA v.5.x, 6.1 и приложений к ней, написанных на Statistica Basic.

3.2. Расчеты для задач I типа с использованием статистических пакетов

Оценка выборочного значения параметра биномиального распределения (вероятности появления события в каждом испытании) и сравнение его с параметром генеральной совокупности (популяции). Сравнение можно проводить или с помощью вычисления доверительного интервала для оцениваемого параметра, или с помощью проверки гипотез о совпадении и отличии параметров.

В зависимости от объема выборки при решении этой задачи могут использоваться как приближенные (нормальная аппроксимация), так и точные формулы (Приложение. Доверительный интервал для параметра биномиального распределения p).

Для вычислений можно использовать известные программы:

SPSS → Analyze → Nonparametric Tests → Binomial… - для переменной из файла данных вычисляется оценка параметра и сравнивается с введенным значением параметра генеральной совокупности. Приводится p-значение статистики, полученное на основе нормальной аппроксимации, поэтому для малых выборок или редких событий программа неприменима.

NCSS → Proportions… → Proportion - 1. – вводится число успехов (событий), число наблюдений, параметр генеральной совокупности и уровень значимости. Может использоваться для любых выборок, поскольку приводятся и точные, и приближенные оценки. Выбор нужной оценки и нужного статистического вывода из полученного списка – задача пользователя.

Программа NCSS. Результаты вычислений для Примера 6**:**

Число успехов (Number of Successes) – количество родившихся мальчиков. Их доля в выборке (Sample Proportion) – P, доля в популяции (Hypothesized Proportion) – P0.

One Proportion Report

Data Section

(описание входных данных и вычисление частоты)

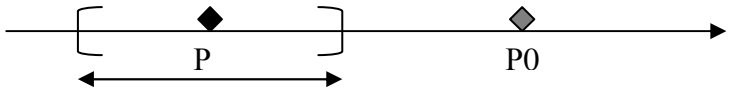
Sample Size (n)	Number of Successes (X)	Sample Proportion (P)	Hypothesized Proportion (P0)	Confidence Alpha	Hypothesis Alpha
20	3	0.150	0.51	0.05	0.05

Confidence Limits Section

(Доверительный интервал для частоты. Используется и точный метод, и аппроксимация нормальным распределением)

	Lower 95%	Upper 95%
Calculation Method	**Confidence Limit**	**Sample Proportion (P)**	**Confidence Limit**
Exact (Binomial)	0.032	0.150	0.379
Approximation (Uncorrected)	0.000	0.150	0.306
Approximation (Corrected)	0.000	0.150	0.331
Wilson Score	0.052	0.150	0.360

На основании результатов, приведенных в этом разделе, можно сделать вывод об отличии P и P0: границы доверительных интервалов параметра P, вычисленные любым из методов, не накрывают значение P0.



Доверительный интервал

Hypothesis Test Section

(Проверка гипотез. Проверяется нулевая гипотеза H0: P=P0 против двусторонней альтернативной гипотезы H1 и односторонних альтернативных гипотез H2 и H3)

Alternative Hypothesis	Exact (Binomial)		Normal Approximation using (P0)			Normal Approximation using (P)		
	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)
H1: P<>P0	0.001	Reject H0	-3.	0.003	Reject H0	-4.2	0.000	Reject H0
H2: P<P0	0.001	Reject H0	-3.	0.001	Reject H0	-4.2	0.000	Reject H0
H3: P>P0	0.999	Accept H0	-3.	0.999	Accept H0	-4.2	0.999	Accept H0

По результатам, приведенным в разделе «Проверка гипотез», также можно утверждать, что P и P0 отличаются:

- отвергается гипотеза H0 о равенстве параметров P = P0

при двусторонней альтернативной гипотезе H1: параметры P < > P0 неравны и

при односторонней альтернативной гипотезе H2: параметр P < P0,

т.е. на уровне значимости α < 0.002 принимается решение о том, что P < P0, и эти выводы совпадают для точного и приближенных вычислений, однако применимо только точное из-за малого объема выборки.

Программа **bin_coef_ru.stb** (приложение к **Statistica v 5.**x). Для этой программы

Входная информация – таблица

Количество событий **X**	Параметр генеральной совокупности p
Количество наблюдений n	

Выходная информация:

Доверительный 95% интервал для оценки параметра биномиального распределения выборки (частоты) и его сравнение с параметром генеральной совокупности **p**.

►В данной программе вычисления производятся, в зависимости от величины x и n, или по приближенным (нормальная аппроксимация), или по точным формулам, без участия пользователя. Поэтому существенных ограничений здесь нет.

Входная таблица (Пример 6)

	Исходное распределение	p сравнения
Событий	3	0.51
Наблюдений	20	

Выходная таблица

95% доверительный интервал

	p	лев.граница p	прав.граница p	p сравнения
Оценки	0.15	0.032	0.379	**0.51**

Вывод: на уровне значимости 0.05 можно утверждать, что вероятность появления мальчика в этом поселке отличается от 0.51 (значимо меньше), т.е. выборка имеет распределение, отличное от распределения генеральной совокупности.

При решении примера 7 после вычисления параметра базового распределения можно проводить вычисления так же, как и в примере 6. Для вычислений могут быть использованы те же программы.

Также можно использовать программу **binomial_ru.stb** (приложение к **Statistica v 5.**x). В этом случае предварительные вычисления не нужны.

Входная информация – таблица

Тестовая выборка	Базовая выборка	Параметр
Количество событий $\mathbf{X_1}$	Количество событий $\mathbf{X_2}$	α –уровень значимости
Количество наблюдений n_1	Количество наблюдений n_2	

Выходная информация:
Доверительный 95% интервал для оценки параметра биномиального распределения (частоты) $\mathbf{p_1}$, соответствующего первому столбцу таблицы, и его сравнение с параметром $\mathbf{p_2}$, соответствующим второму столбцу.

►Как и в предыдущей программе, вычисления производятся, в зависимости от величины x_1 и n_1, по приближенным или по точным формулам. Существенных ограничений нет.

Пример 7 (данные НРЭР по Северо-Западному региону РФ и Комитета по здравоохранению администрации СПб).

Среди ликвидаторов СПб в возрастной группе 40-44 лет в 2000 г. умер 1 человек из 501 наблюдаемого. В базовом распределении (по СПб, аналог генеральной совокупности) повозрастной уровень смертности для мужчин 40-44 лет составил в 2000 г. составил 13 человек на 1000 населения. Можно ли утверждать, что выборочное распределение не отличается от базового (генерального) при выбранном уровне значимости 0.05?

Входная таблица (Пример 7)

	Исходное распределение	Распределение сравнения	Альфа - уровень значимости
Событий	1	13	0.05
Наблюдений	501	1000	

Выходная таблица

(1-альфа)% доверительный интервал (*Пример 7*)

	p исходное	лев.граница p	прав.граница p	p сравнения
Оценки	0.002	0.00005	0.011	**0.013**

Вывод: на уровне значимости 0.05 можно утверждать, что в 2000 г. вероятность смерти среди ликвидаторов 40-44 лет отличалась от 0.013, т.е. тестовая и базовая выборки имели разные параметры.

Графической иллюстрацией полученного вывода может быть следующий рисунок. Для выборочного параметра указан 95% доверительный интервал.

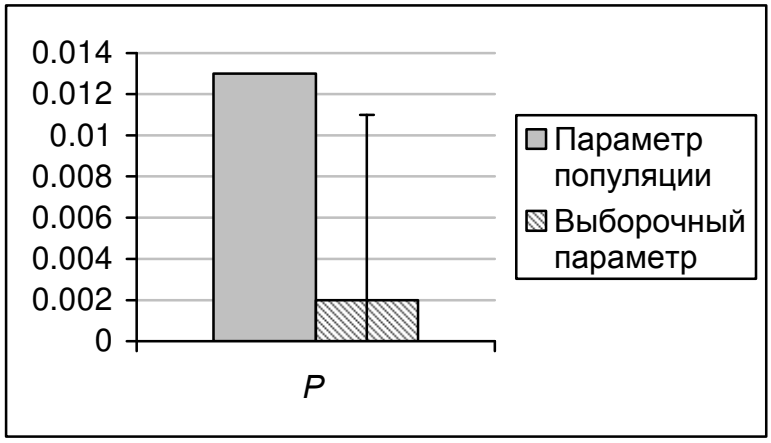


Рис. П7-1. Сравнение параметра популяции P_0 с выборочным параметром P.

Результаты, полученные для примера 7 при использовании программы NCSS.

NCSS → … One Proportion Report

Confidence Limits Section

Calculation Method	**Lower 95% Confidence Limit**	**Sample Proportion (P)**	**Upper 95% Confidence Limit**
Exact (Binomial)	0.000	0.002	0.011
Approximation (Uncorrected)	0.000	0.002	0.006
Approximation (Corrected)	0.000	0.002	0.007
Wilson Score	0.000	0.002	0.011

Hypothesis Test Section

Alternative Hypothesis	Exact (Binomial)		Normal Approximation using (P0)			Normal Approximation using (P)		
	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)
H1: P<>P0	0.026	Reject H0	-1.98	0.048	Reject H0	-5.02	0.000	Reject H0
H2: P<P0	0.011	Reject H0	-1.98	0.024	Reject H0	-5.02	0.000	Reject H0
H3: P>P0	0.999	Accept H0	-1.98	0.979	Accept H0	-5.02	1.000	Accept H0

Вывод. Принимается гипотеза H2: P<P0.

3.3. Расчеты для задач II типа с использованием статистических пакетов

При сравнении оценок параметров двух и более биномиальных распределений можно использовать несколько возможных подходов. Поскольку в основе их всех лежат парные сравнения, далее будем рассматривать различные методы проверки гипотезы о равенстве параметров именно **двух** распределений. Если перед нами встает задача сравнения параметров двух и более выборок, то определение наилучшего метода решения существенно связано как с содержанием, так и с численными параметрами задачи.

(а) Сравнение оценок параметров выборочных распределений. Использование нормальной аппроксимации и точного критерия Фишера.

Суть метода состоит в проверке гипотезы H_0: $p_1 = p_2$, где p_1 и p_2 - два параметра биномиальных распределений. При проверке используется аппроксимация нормальным распределением и фактически проверяется гипотеза о том, что разность оценок параметров равна 0.

Недостаток этого метода – в ограничениях, которые связаны с аппроксимацией:

► Ограничения

При использовании аппроксимация нормальным распределением все ожидаемые частоты в ячейках таблицы должны быть более 5 (С2.1).

В случаях, когда можно использовать нормальную аппроксимацию для сравнения параметров двух и более выборок, спектр применяемых методов анализа расширяется: например, для сравнения частот можно использовать критерий Стьюдента и однофакторный дисперсионный анализ.

Метод (а) реализован в следующих программах:

Statistica v.5.x, 6.0 → **Basic Statistics → Difference tests: r, %, means (or Other significance tests, v.5.x) → Difference between two proportions.**

NCSS → NCSS Navigator → NCSS – Data Analysis, … → Test of Frequencies and Proportions → Two Proportions Test.

SYSTAT → Statistics → Tables → Crosstabs …,→ One-way Tables. Для переменных из файла данных вычисляются частоты их значений и доверительные интервалы для них (в %).

В программе **bin_k_ru.stb** (приложение к **Statistica v 5.**x).

Если мы имеем дело с выборками достаточного объема, а исследуемое событие случается достаточно часто, то можно использовать любую из перечисленных выше программ. Достаточный объем выборки означает, что наблюдений в ней 50 или больше. Достаточная частота события означает, что в каждой выборке оно появлялось и не появлялось более 5 раз.

Если эти ограничения не выполнены, то могут применяться только те программы, в которых вычисляется точный критерий Фишера. Далее приведены примеры вычислений с помощью перечисленных выше программ.

Программа bin_k_ru.stb. Эта программа удобна в том случае, когда нужно получить результаты сравнения параметров более 2-х выборок.

► Ограничения

Программа может использоваться, если количество наблюдений в каждом столбце больше 50, количество случаев и не-случаев в каждом столбце больше 5 ($x_i > 5$, $n_i - x_i > 5$, $n_i > 50$ для всех $i = 1, 2, \ldots, k$).

Входная информация – таблица

1-я выборка	2-я выборка	…	k-я выборка
Количество событий $\mathbf{X_1}$	Количество событий $\mathbf{X_2}$	…	Количество событий $\mathbf{X_k}$
Количество наблюдений n_1	Количество наблюдений n_2	…	Количество наблюдений n_k

Выходная информация:

Оценки параметров распределений $\mathbf{p_1}$, $\mathbf{p_2}$.,…, $\mathbf{p_k}$;

Значения статистик парных сравнений и соответствующие им p-значения.

Пример 8. (данные НРЭР по Северо-Западному региону РФ)

Среди ликвидаторов СПб в возрастной группе 40-44 лет в 1990 г. умерло 8 человек из 875 наблюдаемых, в 1995 г. умерло 6 человек из 672 наблюдаемых, и в 2000 г. умер 1 человек из 501 наблюдаемых. Можно ли утверждать, на уровне значимости 0.05, что за этот период уровень смертности не менялся?

Входная таблица (Пример 8)

	1-е распределен.	2-е распределен.	3-е распределен.
Событий	8	6	1
Наблюдений	875	672	501

Выходная таблица

Параметры биномиальных распределений и p-значения для статистик

	p	COL2	COL3
1	0.009	0.78	0.11
2	0.008		0.17
3	0.002		

Вывод: значимых отличий в уровне смертности ликвидаторов СПб 40-44 лет в 1990, 1995 и 2000 гг. не обнаружено. При этом требуемые ограничения (наблюдений более 50, случаев в каждом столбце более 5) не выполнены для 3-го распределения, поэтому для корректного сравнения уровней смертности требуются дополнительные вычисления.

Вариант уточнения вычислений:

Пример 9. (данные НРЭР по Северо-Западному региону РФ)
Укрупним возрастную группу и рассмотрим уровни смертности ликвидаторов 40-49 лет за те же годы.
Среди ликвидаторов СПб в возрастной группе 40-49 лет в 1990 г. умерло 9 человек из 1172 наблюдаемых, в 1995 г. умерло 18 человек из 1768 наблюдаемых, в 2000 г. умерло 7 человек из 1321 наблюдаемых. Можно ли утверждать, на уровне значимости 0.05, что за этот период уровень смертности не менялся?

Входная таблица (Пример 9)

	1-е распределен.	2-е распределен.	3-е распределен.
Событий	9	18	7
Наблюдений	1172	1768	1321

Выходная таблица

Параметры биномиальных распределений и p-значения для статистик

	p	COL2	COL3
1	0.008	0.49	0.46
2	0.010		0.13
3	0.005		

Вывод: значимых отличий в уровне смертности ликвидаторов СПб 40-49 лет в 1990, 1995 и 2000 гг. не обнаружено. Все ограничения выполнены.
Иллюстрация примеров 8, 9 представлена на следующем рисунке. Снижение повозрастной смертности ликвидаторов в 2000 г. не является статистически значимым.

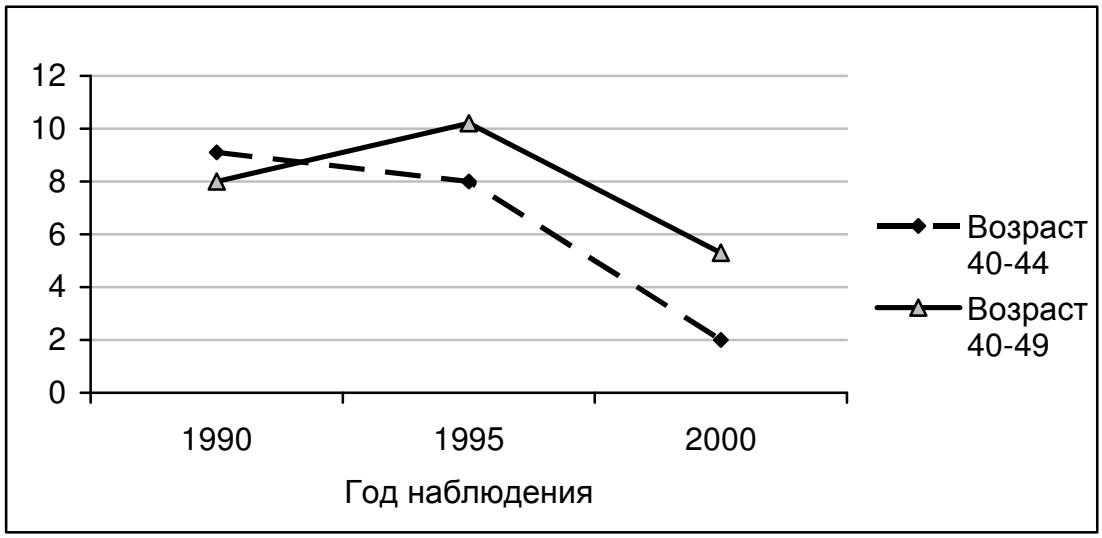


Рис. П9-1. Повозрастные уровни смертности ликвидаторов на 1000 чел.

Программа **NCSS.** Эта программа дает возможность получить корректный результат и для малой выборки тоже. Сравнивать между собой можно только две выборки. Стандартные обозначения для таблиц 2x2, которые использованы в части Table Section программы NCSS, следующие:

Ряд значений	Выборка 1	Выборка 2	Сумма
A	a	b	m_1
не A	c	d	m_2
Сумма	n_1	n_2	n

NCSS → … → Two Proportions Test. (*Пример 8.)*

Производится последовательно сравнение пар распределений: 1-е – 2-е; 1-е – 3-е; 2-е – 3-е.

Two Proportions Report (сравнение 1-го и 3-го распределений)

Table Section

(описание входных данных в виде таблицы)

A	**B**	**C**	**D**	**N1 (A+C)**	**N2 (B+D)**	**M1 (A+B)**	**M2 (C+D)**	**N (N1+N2)**
1	8	500	867	501	875	9	1367	1376

Data Section (описание входных данных и вычисление частот)

Sample	Sample Size	Number in Group One	Number in Group Two	Proportion In Group One	Proportion In Group Two
One	501	1	500	0.0020	0.9980
Two	875	8	867	0.0091	0.9909
Total	1376	9	1367	0.0065	0.9935

Confidence Limits of Difference Section

(доверительный интервал для разности частот. Используется аппроксимация нормальным распределением)

Difference	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit
-0.0071	0.0038	-0.0146	0.0003

Hypothesis Test Section

(проверка гипотез. Проверяется нулевая гипотеза H0: P1=P2 против двусторонней альтернативной гипотезы H1 и односторонних альтернативных гипотез H2 и H3)

Alternative Hypothesis	Fisher's Exact Test		Normal Approximation			Yates Chi-Square Test	
	Prob. Level	Decision (5%)	Z-Value	Prob. Level	Decision (5%)	Chi-Square Value	Prob. Level
H1: P1-P2<>0	0.168	Can't Reject	-1.58	0.114	Can't Reject	1.525	0.217
H2: P1-P2<0	0.104	Can't Reject	-1.58	0.057	Can't Reject		
H3: P1-P2>0	0.983	Can't Reject	-1.58	0.943	Can't Reject		

Для получения достоверных статистических выводов в этой задаче можно использовать только **Prob.Level** для **Fisher's Exact Test.** Поэтому принимается предположение о равенстве параметров (нулевая гипотеза). Она не отвергается при сравнении со всеми альтернативными гипотезами.

Для Примера 9 получим с помощью той же программы следующие результаты:

Two Proportions Report (сравнение 2-го и 3-го распределений)
Table Section

A	B	C	D	N1 (A+C)	N2 (B+D)	M1 (A+B)	M2 (C+D)	N (N1+N2)
18	7	1750	1314	1768	1321	25	3064	3089

Data Section

Sample	Sample Size	Number in Group One	Number in Group Two	Proportion In Group One	Proportion In Group Two
One	1768	18	1750	0.0102	0.9898
Two	1321	7	1314	0.0053	0.9947
Total	3089	25	3064	0.0081	0.9919

Confidence Limits of Difference Section

Difference	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit
0.0049	0.0031	-0.0012	0.0110

Hypothesis Test Section

Alternative Hypothesis	Fisher's Exact Test		Normal Approximation			Yates Chi-Square T	
	Prob. Level	Decision (5%)	Z-Value	Prob. Level	Decision (5%)	Chi-Square Value	Prob. Level
P1-P2<>0	0.158	Can't Reject	1.498	0.134	Can't Reject	1.678	0.195
P1-P2<0	0.958	Can't Reject	1.498	0.933	Can't Reject		
P1-P2>0	0.096	Can't Reject	1.498	0.067	Can't Reject		

В этом случае можно использовать любой столбец (и **Fisher's Exact Test –** точный тест Фишера, и **Normal Approximation –** нормальную аппроксимацию), выводы не противоречат друг другу: существенных отличий параметров не обнаружено.

Statistica v.5.x, 6.0 → **… → Difference between two proportions.**

Сравнение пар распределений: 1-е – 2-е; 1-е – 3-е; 2-е – 3-е. Осуществляются только вычисления на основе нормальной аппроксимации. Для каждого распределения требуется ввести оценку параметра и объем выборки: При сравнении 1-го и 3-го распределений

Примера 8 входная таблица имеет вид:

Pr.1	0.0091	**N1**	875
Pr.2	0.002	**N2**	501

Вычисляется **p**-значение для проверки одно- или двусторонней гипотезы.

Two-sided p=0.1154

One-sided p=0.0577

Эти значения соответствуют столбцу **Prob.Level** в разделе **Normal Approximation** предыдущего раздела (вычисления с помощью программы NCSS) и дают некорректный результат для данной задачи (в одной из ячеек число наблюдений менее 5), однако такие выводы должен делать пользователь, программа не проверяет выполнение ограничений.

В следующем примере можно использовать приближение нормальными с.в.

Пример 10. (данные НРЭР по Северо-Западному региону РФ)

Среди ликвидаторов Северо-Западного региона, по данным РГМДР, за период наблюдений 1987-2004 гг. зафиксировано следующее число наблюдаемых и умерших ликвидаторов на отдельных территориях (Калининградская область, Ленинградская область, Санкт-Петербург, Новгородская область и Псковская область):

Таблица П10-1. Исходные данные

	Калинингр. обл.	Ленинград. обл.	Санкт-Петербург	Новгород. обл.	Псковск. обл.
Умерли	280	359	533	260	165
Наблюдались	1564	2321	4765	1397	1040
Человеко-лет наблюдения	19281	27597	56129	18880	12080
Среднее число лет наблюдения	13.79	14.30	14.43	12.58	14.64

На основании этих данных вычислены показатели смертности на пяти территориях, представленные в следующей таблице.

Таблица П10-2. Показатели смертности и доверительные интервалы для них.

	Калинингр. область	Ленингр. область	Санкт-Петербург	Новгород. область	Псковск. область
% умерших	17.90	15.47	11.19	18.61	15.87
95% доверительный интервал для % умерших	16.08 – 19.88	14.05 – 17.0	10.32 – 12.11	16.66 – 20.74	13.77 – 18.21
Уровни смертности на 1000 чел.-лет наблюдения	14.52	13.01	9.50	13.77	13.66

Можно ли утверждать, на уровне значимости 0.05, что показатели смертности ликвидаторов на различных территориях за этот период не отличаются?

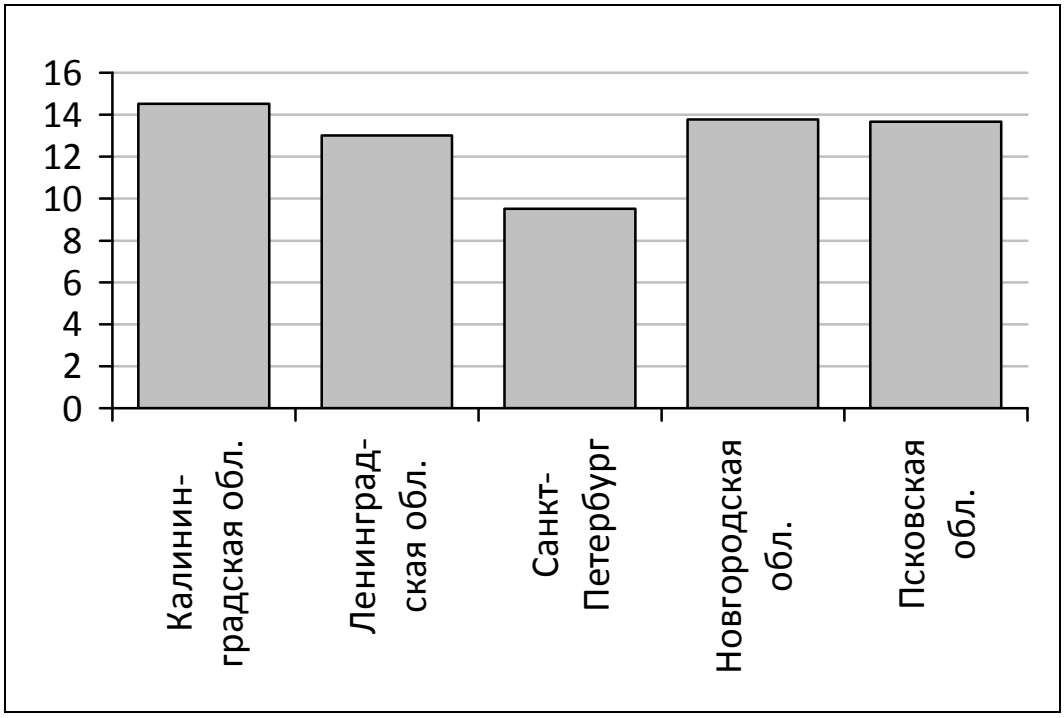


Рис.П10-1. Уровни смертности на 1000 человеко-лет наблюдения

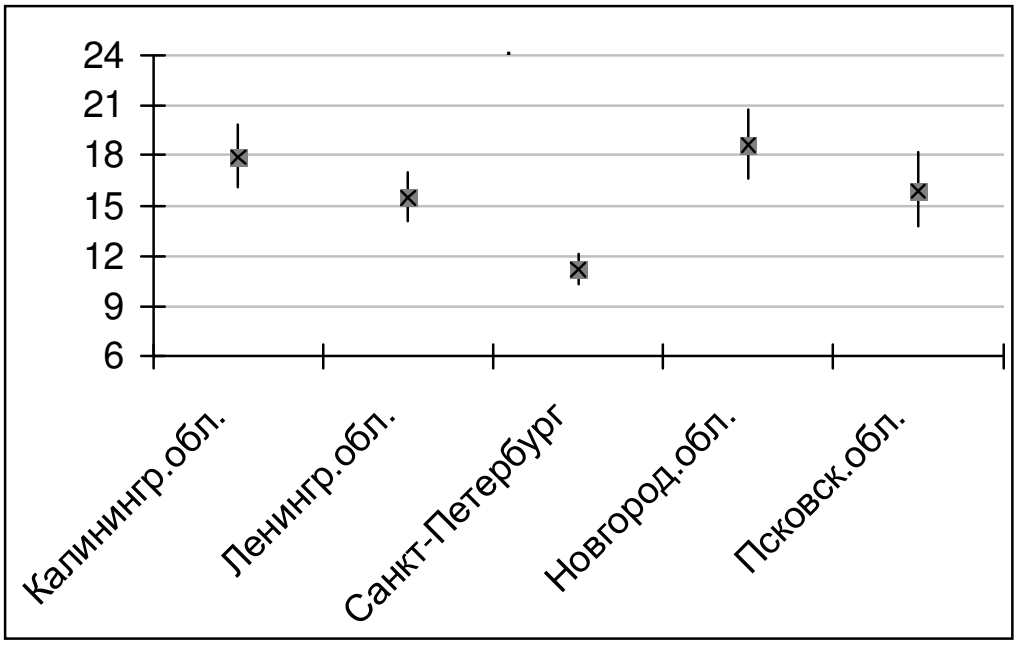


Рис.П10-2. Процент умерших и 95% доверительный интервал

В таблице приведены два показателя смертности – процент умерших среди всех, кто был под наблюдением (частота смерти на территории) и уровни смертности на 1000 человеко-лет наблюдения. Второй показатель более точно отражает численное основание для учета смертности, т.к. ликвидаторы в разное время вставали на учет, кто-то умер, кто-то переехал, поэтому число наблюдавшихся за весь период не совпадает с числом наблюдавшихся в каждый отдельный год. Например, сравнивая два показателя смертности по Новгородской области, мы обнаружим, что в процентном выражении он существенно превышал соответствующий показатель Ленинградской области, Псковской области и даже Калининградской области. Однако в среднем ликвидаторы Новгородской области наблюдались меньшее число лет, чем на остальных территориях. Поэтому уровень смертности в Новгородской области сравним с уровнями Ленинградской и Псковской областей и ниже уровня Калининградской области.

Частоту смерти можно сравнить по территориям с помощью статистики Стьюдента. Объемы выборок вполне позволяют это сделать.

Таблица П10-3. Попарное сравнение частоты смерти на отдельных территориях. P-значения статистики Стьюдента с поправкой Бонферрони на множественность сравнений (k = 10)

	Калинингр. обл.	Ленингр. обл.	Санкт-Петербург	Новгород. обл.	Псковская обл.
Калинингр. обл.		0.450	**0.000**	1.000	1.000
Ленингр. обл.	0.450		**0.000**	0.130	1.000
Санкт-Петербург	**0.000**	**0.000**		**0.000**	**0.000**
Новгород. обл.	1.000	0.130	**0.000**		0.770
Псковская обл.	1.000	1.000	**0.000**	0.770	

Согласно этой таблице, частота смерти в СПб отличается от всех остальных территорий (на уровне $\alpha < 0.0001$), остальные частоты значимо не отличаются.

Таблица П10-4. Попарное сравнение уровней смертности на отдельных территориях. P-значения проверки двусторонней гипотезы (P1=P2 против P1-P2<>0), также с поправкой Бонферрони

	Калинингр. обл.	Ленингр. обл.	Санкт-Петербург	Новгород. обл.	Псковская обл.
Калинингр. обл.		1.000	**0.000**	1.000	1.000
Ленингр. обл.	1.000		**0.000**	1.000	1.000
Санкт-Петербург	**0.000**	**0.000**		**0.000**	**0.010**
Новгород. обл.	1.000	1.000	**0.000**		1.000
Псковская обл.	1.000	1.000	**0.010**	1.000	

Уровни смертности в СПб отличается от всех остальных территорий (на уровне $\alpha < 0.01$), уровни смертности на остальных территориях не отличаются – вывод аналогичен предыдущему.

(б). Использование критерия χ^2

Метод предусматривает использование непараметрического подхода – проверка совпадения (однородности) двух и более распределений осуществляется с помощью критерия χ^2. При применении этого критерия предположения о биномиальности распределений не используются. Описание метода дано в главе 2.

В статистических пакетах **NCSS, SPSS, Statistica** v.5.x, 6.0, **SYSTAT** критерий χ^2 вычисляется в рамках процедуры **Crosstabs,** что требует табуляции данных из файла. Это не всегда удобно – например, при сравнении более чем двух параметров. Часть программ при этом сообщает об ограничениях, которым должны удовлетворять частоты, получаемые в ячейках.

В отличие от программ сравнения биномиальных коэффициентов (метод IIа), при применения критерия χ^2 не используются альтернативные односторонние гипотезы $p_1 < p_2$ или $p_1 > p_2$. Проверяется нулевая гипотеза о совпадении (однородности) распределений против альтернативной о том, что они различны. Поэтому в результатах вычислений по Примеру 8 и Примеру 9 с помощью программы **NCSS,** приведенных выше, последние столбцы в таблице раздела проверки гипотез (**Hypothesis Test Section**), озаглавленные **Yates Chi-Square Test** , содержат результаты вычислений только в первой строке.

В программе **chi_sq_ru.stb** (приложение к **Statistica v 5.**x).

Входная информация – таблица

1-я выборка	2-я выборка	…	k-я выборка
Количество случаев $\mathbf{x_1}$	Количество случаев $\mathbf{x_2}$	…	Количество случаев $\mathbf{x_k}$
Количество не случаев $\mathbf{n_1-x_1}$	Количество не случаев $\mathbf{n_2-x_2}$	…	Количество не случаев $\mathbf{n_k-x_k}$

Выходная информация:

Оценки параметров распределений $\mathbf{p_1}$, $\mathbf{p_2}$.,…, $\mathbf{p_k}$;

Значения статистик парных сравнений и соответствующие им p-значения.

Значение статистики сравнения всей совокупности столбцов и соответствующее ему p-значение.

Сообщения о возможности применения метода для статистических выводов в предлагаемой задаче:

1) Критерий хи-квадрат неадекватен (ожидаемые значения малы)
2) Критерий хи-квадрат применим (ограничения выполнены)

Пример 8. (данные РГМДР по Северо-Западному региону РФ)

Входная таблица (Пример 8)

	1-е распределение	2-е распределение	3-е распределение
Событий	8	6	1
Наблюдений	875	672	501

Выходная таблица

Хи-квадрат статистики попарного сравнения столбцов.
Критическое значение **3.84** на уровне 0.05
Критерий хи-квадрат неадекватен (ожидаемые значения малы)

Выборочные значения статистики

	1-е распределение	2-е распределение	3-е распределение
1		0.002	2.504
2			2.325
3			

Частоты для столбцов входной таблицы

	1-е распределение	2-е распределение	3-е распределение
p на 1 наблюд.	0.0091	0.0089	0.002

Вывод: значимых отличий в уровне смертности ликвидаторов СПб 40-44 лет в 1990, 1995 и 2000 гг. не обнаружено. Требуемые ограничения (наблюдений более 50, случаев в каждом столбце более 5) выполнены не везде, поэтому для корректного сравнения уровней смертности требуются дополнительные вычисления.

Выходная таблица (Пример 9)

Хи-квадрат статистики попарного сравнения столбцов.
Критическое значение **3.84** на уровне 0.05
Критерий хи-квадрат применим (ограничения выполнены)
Выборочные значения статистики

	1-е распределение	2-е распределение	3-е распределение
1		0.485	0.552
2			2.245
3			

Частоты для столбцов входной таблицы

	1-е распределение	2-е распределение	3-е распределение
p на 1 наблюд.	0.0077	0.010	0.0053

Вывод: не обнаружено значимых на уровне 0.05 отличий в уровне смертности ликвидаторов СПб 40-49 лет в 1990, 1995 и 2000 гг.

Если каждая из выборок, параметры которых мы должны сравнить, соответствует уровню какого-либо фактора (территория, пол, возраст, прием лекарств и т.д.), и вычисленные параметры p_i для разных уровней фактора (выборок) отличаются, то можно говорить о связи фактора и события А. Например, связь ИБС с курением, возрастом, избыточной массой тела. Одной из наиболее известных характеристик связи является **риск**.

3.4. Риски

Риском называется вероятность возникновения неблагоприятного исхода, и, как всякая вероятность, она принимает значения в интервале от 0 (риск отсутствует) до 1 (неблагоприятный исход наступит наверняка). В качестве неблагоприятного исхода может рассматриваться болезнь, смерть, определенное осложнение и т.д.

В исследованиях, как правило, встает вопрос оценки риска неблагоприятного исхода в связи с каким-либо фактором. В качестве меры воздействия фактора на частоту (риск) возникновения события используют относительный риск, атрибутивный (абсолютный) риск или отношение шансов.

Относительный риск (relative risk, RR) – это отношение частоты события в той части выборки, где фактор действует, к частоте в части выборки, где фактор не действует.

Часть выборки, на которой действует фактор, называется «экспонированной» данным фактором.

Относительный риск оценивают, чтобы проверить, существует ли мультипликативное взаимодействие между фактором и событием. Если оценка относительного риска статистически не отличается от **1 (**на выбранном **уровне значимости α)**, то гипотеза о наличии мультипликативного взаимодействия отвергается.

Атрибутивный риск (attributable risk, AR) – это разность частот события в экспонированной и не экспонированной фактором риска частях выборки. Атрибутивный риск можно вычислить по отношению к группе риска или всей популяции и выразить как в абсолютных числах, так и в процентах.

Атрибутивный риск предназначен для измерения аддитивного взаимодействия между фактором и событием. Если оценка AR статистически не отличается от **0**, то гипотеза о наличии аддитивного взаимодействия отвергается.

Отношение шансов. (odds ratio, OR) Шансы события – это отношение числа случаев появления события в выборке к числу случаев его непоявления (к числу «не-случаев»). Например, если исследуемое событие – наличие ИБС, то шансы этого события в выборке – это отношение количества наблюдаемых, у которых есть это заболевание, к количеству наблюдаемых, у которых его нет. Отношение шансов – это шансы события в экспонированной фактором части выборки, деленные на шансы события в неэкспонированной части.

Если для сравнения частот событий используется вычисление рисков, то приведенные на схеме 5 типы задач I и II изменяются в блоках, где определены проверяемые гипотезы (Схема 6).

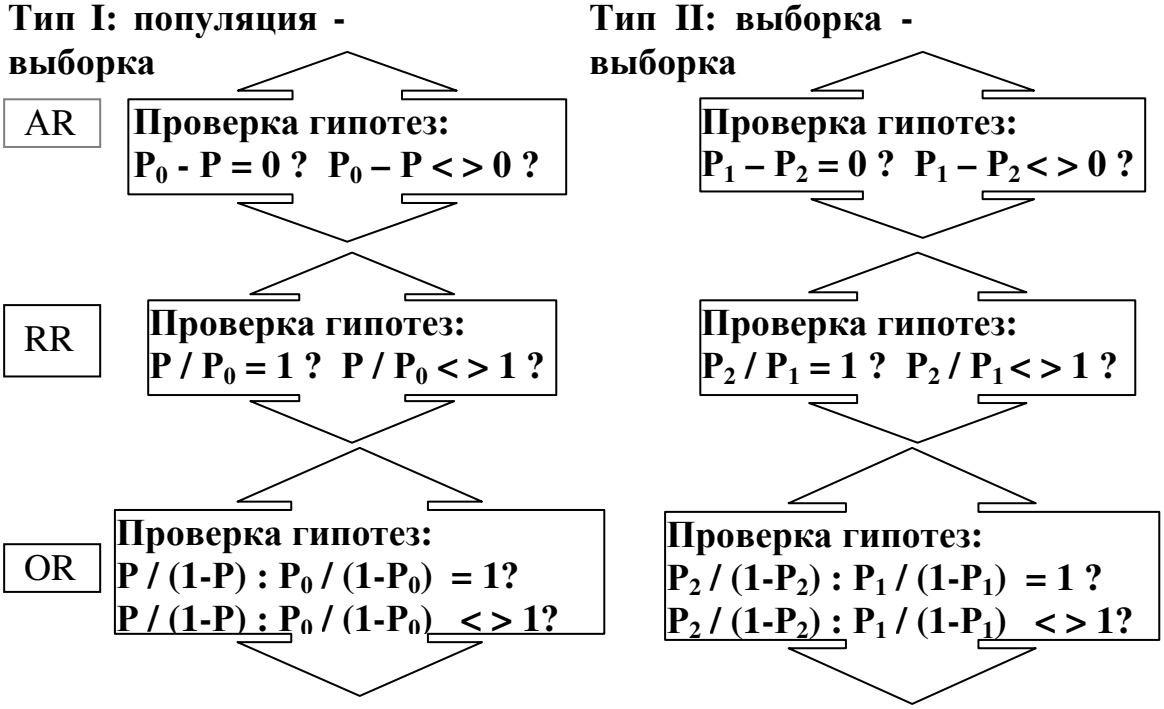


Схема 6. Основные типы задач проверки рисков

Относительный риск RR (relative risk) – это отношение $\frac{p_2}{p_1}$. Если подходить чисто алгебраически, то равенство $\frac{p_2}{p_1}=1$ (RR=1) эквивалентно равенствам $p_1= p_2$ или $p_2 – p_1=0$ (AR=0). Однако, с точки зрения статистики, при проверке каждого из этих равенств, являющихся **символическими** записями статистических гипотез, используется различная техника вычислений. Поэтому и статистические выводы можно получить разные. Например, RR статистически отличен от 1, а гипотеза H_0: AR = 0 не отвергается.

(в). Вычисление относительных рисков

При анализе таблиц 2x2 можно вычислить относительный риск появления события при уровне изучаемого фактора F_2 по сравнению с уровнем F_1 (в выборке F_2 по сравнению с выборкой F_1). Термином «относительный риск» могут быть обозначены три различных (но близких по смыслу) понятия. Все они вычисляются как отношения, используя следующие формулы.

	Уровни изучаемого фактора F или названия выборок	
	F_1	**F_2**
Событий	a_1	a_2
Наблюдений	c_1	c_2

Относительный риск R_{21} для таблицы 2x2:

$$R_{21} = \frac{p_2}{p_1} ,$$

где p_i могут быть (1) пропорциями (частотой), (2) уровнями или (3) шансами осуществления события.

Пропорции и **уровни** p_i вычисляются по одной формуле:

$$p_i = \frac{a_i}{c_i} \quad (3.1),$$

где a_i - число «случаев», c_i - число «наблюдений» (для пропорций) или «человеко-годы наблюдения» (для уровней). Чаще всего мы имеем дело с пропорциями, поэтому в заголовке второй строки таблицы стоит слово «наблюдения». В дальнейшем, если не оговорено противное, символом p_i будет обозначаться пропорция.

Для пропорций и уровней относительный риск обычно обозначается символом **RR** (relative risk). В табличных обозначениях относительный риск события записывается как

$$RR = R_{21} = \frac{a_2 \times c_1}{c_2 \times a_1} \quad (3.2)$$

Шансы события (заболевания, смерти и т.д.) определяются как отношение числа «случаев» к числу «не случаев».

$$\tilde{p}_i = a_i / (c_i - a_i) = p_i / (1 - p_i) \quad (3.3)$$

При сравнении шансов осуществления события в двух выборках

относительный риск – это отношение шансов (odds ratio). Он обычно обозначается символом **OR**. В табличных обозначениях **отношение шансов** записывается как

$$OR = R_{21} = \frac{a_2 \times (c_1 - a_1)}{(c_2 - a_2) \times a_1} \quad (3.4)$$

Риск имеет асимптотически логнормальное распределение, поэтому уровни значимости определяются для проверки гипотезы

H_0: lnR = 0 (R = 1).

Способ вычисления стандартной ошибки $SE(p_i)$ зависит от содержания таблицы.

(в.1) Если строка «наблюдений» означает количество объектов наблюдения, для которых некоторое «событие» обязательно должно или осуществиться, или не осуществиться, причем осуществиться оно может только один раз (например, a_i – количество умерших, c_i – общая численность наблюдаемых), то p_i – *пропорция (частота)* по содержанию, распределение числа событий моделируется биномиальным распределением, и

$$p_i = \frac{a_i}{c_i}, \quad \mathrm{SE}(p_i) = \sqrt{\frac{p_i \times (1 - p_i)}{c_i}}$$

Доверительный интервал для относительного риска приведен в Приложении (риски, формула (I)).

(в.2) Если в строке «наблюдений» - человеко-годы наблюдения за период (общее время под риском), тогда p_i – *уровень* по содержанию. В этом случае распределение числа событий моделируется распределением Пуассона,

$$p_i = \frac{a_i}{c_i}, \qquad SE(p_i) = S(p_i) = \frac{p_i}{\sqrt{a_i}}$$

Доверительный интервал для относительного риска как отношения уровней (Приложение, риски, формула (II)) несколько шире, чем в предыдущем случае.

(в.3) При вычислении OR – отношения шансов - в строке «наблюдений» количество объектов наблюдения. Шансы используются при исследованиях «случай – контроль» или при изучении редких событий. Вместо частоты p_i в этом случае вычисляются *шансы* (осуществления события в группе):

$$\tilde{p}_i = a_i / (c_i - a_i), \; \tilde{p}_i = p_i / (1 - p_i)$$

Это выражение называется *логитом* p_i.

Доверительный интервал для отношения шансов (Приложение, риски, формула (III)) шире, чем для относительных рисков в обоих случаях.

Стандартные программные средства позволяют вычислить отношение шансов (OR). Эти вычисления реализованы в следующих программах.

NCSS → NCSS Navigator →
NCSS – Data Analysis, … → Test of Frequencies and Proportions → Two Proportions Test.

Для Примера 10: сравнение доли умерших среди ликвидаторов Калининградской области и СПб.

	Калининградская обл.	Ленинградская обл.	Санкт-Петербург
Умерли	280	359	533
Наблюдались	1564	2321	4765

Two Proportions Power Analysis
Numeric Results
Null Hypothesis: P1=P2 Alternative Hypothesis: P1<>P2. Continuity Correction Used.

Power	**N1**	**N2**	**Allocation Ratio**	**P1**	**P2**	**Odds Ratio**	**Alpha**	**Beta**
0.80000	4765	1564	0.328	0.112	0.179	1.729	0.000	0.200

Summary Statements
Group sample sizes of 4765 and 1564 achieve 80% power to detect a difference of 0.067 between the null hypothesis that both group proportions are 0.112 and the alternative hypothesis that the proportion in group 2 is 0.179 using a two-sided Chi-square test with continuity correction and with a significance level of 0.00000.
На уровне значимости $\alpha < 0.00001$ отвергается нулевая гипотеза, т.е. отношение шансов OR=1.729 на этом уровне отличается от 1.

SPSS → Analyze → Descriptive Statistics → Crosstabs …, далее определить флажок **Risk** в опции **Statistics ...** Риски определяются для категорий первой переменной (переменная по строкам), в зависимости от категорий второй переменной (переменная по столбцам). Для того, чтобы вычисления рисков в этой программе было возможно, и переменная строки, и переменная столбца должны иметь 2 возможных значения, т.е. результат табуляции – таблица 2×2.

Пример 11. Зависимость актуальной ригидности и стажа работы у пожарных. (Данные Т.И.Шевченко, НИС Медицинский регистр).
Проведено тестирование в пожарных частях СПб по опроснику ТОРЗ Залевского. Исследуется связь возраста, стажа работы и ригидности по основным шкалам.
В соответствии с методикой для каждого испытуемого получены уровни актуальной ригидности (АР): низкий, умеренный и высокий. Далее приведена таблица распределения по уровням актуальной ригидности пожарных двух групп по стажу работы: «2.5 – 5 лет» и «более 5 лет».

AP_ * Гр_стаж_ Crosstabulation		Группа стажа		Всего
		более 5 лет	2.5 - 5 лет	
AP	высокая	14	3	17
	низкая и умеренная	25	29	54
Total		39	32	71

Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for AP_ (высокая/ низкая и умеренная)	5.413	1.394	21.025
N of Valid Cases	71		

OR – отношение шансов иметь АР высокого уровня при стаже свыше 5 лет увеличивается более чем в 5 раз по сравнению с группой по стажу «2.5 – 5 лет», и этот показатель статистически значим на уровне 0.05 (95% доверительный интервал не включает 1).

SYSTAT → Statistics → Tables → Crosstabs …,→ Two-Way Tables далее определить флажок **Odds Ratio** в опции **Statistics ...** Риски определяются для категорий первой переменной (переменная по строкам), в зависимости от категорий второй переменной (переменная по столбцам). Как и в программе **SPSS**, и переменная строки, и переменная столбца должны иметь 2 возможных значения, результат табуляции – таблица 2×2.

Для Примера 11.

Coefficient	Value	Asymptotic Std Error
Odds Ratio	5.413	
Ln(Odds)	1.689	0.692

Для нахождения доверительного интервала требуется произвести вычисления самостоятельно по формуле III Приложения.

В программе **Statistica** v.5.x, 6.0 непосредственно риски не вычисляются. Более сложным образом их можно оценить с помощью модуля «Обобщенные линейные модели – логистическая регрессия». Для примера 11 результирующая таблица с оценкой OR имеет следующий вид.

Model: Logistic regression (logit) N of 0's:54 1's:17 Dep. var: AP
Loss: Max likelihood
Final loss: 35.416 $Chi^2(1)=7.328$ p=.0068

	Группа по стажу
Estimate	1.689
Odds ratio (unit ch)	5.413
Odds ratio (range)	5.413

В программе **risks2x2_ru.STB -** вычисление рисков для таблиц 2×2 - сравнение оценок параметров двух биномиальных распределений – реализованы вычисления всех относительных рисков.

Для Примера 11.
Выходная таблица

	Относительный риск RR	RR-b	RR+b	p1	p2
Для уровней (человеко-годы в строке наблюдений)	3.829	1.100	13.324	0.094	0.359
Для пропорций (количество наблюдений)	3.829	1.205	12.166	0.094	0.359
Для шансов – OR (количество наблюдений, редкие события)	5.413	1.502	19.508	0.103	0.560

В программе вычисляется как относительный риск (RR) для пропорций и уровней, так и отношение шансов (OR), с доверительными 95% интервалами (RR-b, RR+b), а также частота и шансы события в каждой группе (p1, p2). В данном примере можно использовать 2-ю и 3-ю строки таблицы для оценки влияния стажа на повышение уровня актуальной ригидности. Иллюстрация полученных результатов на рис.П11-1.

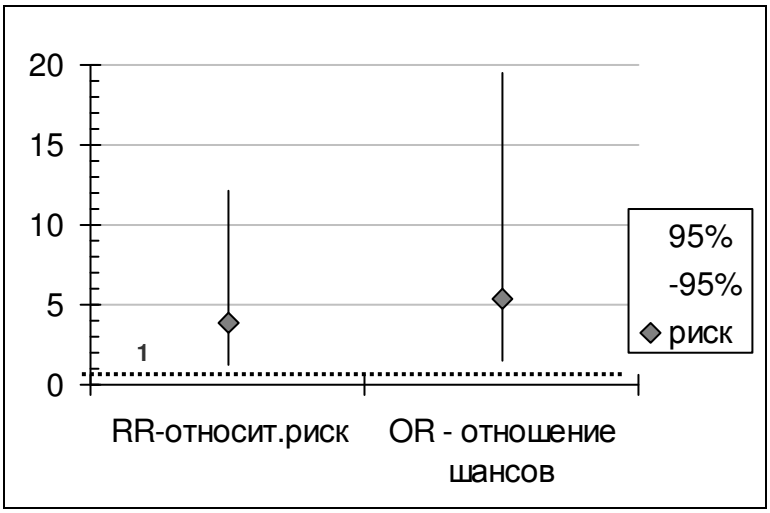


Рис.П11-1. Относительный риск и отношение шансов с 95% доверительными интервалами.

Исследуемый фактор влияния может иметь более двух уровней, то есть он не обязательно дихотомический. В этом случае говорят о нескольких «уровнях экспонированности фактором» и о «дозовом отклике» или «дозовой зависимости». Например, если исследуемый фактор риска - масса тела, обычно вводятся градации этого фактора: масса тела недостаточная, в норме, избыточная. Для трех уровней индекса массы тела может вычисляться риск смерти, связанный с недостаточной массой тела (по сравнению с нормальной), и риск, связанный с избыточной массой тела (также по сравнению с нормальной).

При исследованиях смерти от ИБС в зависимости от величины АДС выделяют, как правило, несколько дозовых групп, например: (1) менее 140 мм рт.ст., (2) 140-159 мм рт.ст., (3) 160 и выше. Количество таких групп может быть значительно больше, если позволяет объем выборки.

Исследование линейного тренда пропорций в зависимости от дозы обсуждалось в Главе 2, как один из способов применения критерия хи-квадрат.

ГЛАВА 4. ОЦЕНКА РИСКА ПРИ НАЛИЧИИ НЕСКОЛЬКИХ ФАКТОРОВ

4.1. Влияние сопутствующих факторов

Посторонние факторы, которые не являются предметом исследования, но также воздействуют на частоту события, называются «мешающими» (confounding). Их воздействие может исказить получаемые выводы о влиянии фактора на исследуемый показатель. Часто в качестве мешающего фактора выступает возраст. Если мы изучаем как фактор риска для здоровья стаж работы на опасном производстве, то в этом случае возраст, несомненно, является мешающим параметром, и его влияние следует исключить.

Для исключения влияния мешающих факторов используют несколько приемов. Одним из них является стратификация выборки по мешающим параметрам. Стратификация - это разбиение выборки на части - страты, - в которых значения мешающих параметров не меняются (например, стратификация по полу) или меняются незначительно (возрастные группы с интервалом в 5 или 10 лет). Риски, относительные или абсолютные, вычисляются для каждой страты в отдельности. Таким образом можно получить свободную от влияния мешающего параметра оценку риска в каждой страте.

Далее, если нужно получить одну совокупную оценку риска, можно осуществить стандартизацию набора рисков по эталонному распределению мешающего параметра или вычислить объединенный риск с проверкой однородности по стратам. Эта методика отражена на Схеме 7.

Все эти приемы, с одной стороны, позволяют делать более обоснованные статистические выводы, исключая влияние посторонних, но связанных с оцениваемым эффектом факторов. С другой стороны, они уменьшают объем выборки, для которой производится отдельная оценка, что снижает достоверность результата. Предположим, что исходная выборка состоит из 1000 человек, из них половина экспонирована фактором риска, а половина – нет, и мы хотим исключить влияние пола и возраста на оценки риска. При стратификации по возрасту и полу, если введено 5 возрастных групп и 2 по полу, объем отдельной страты в среднем в 10 раз меньше, чем исходный. То есть в среднем каждая экспонированная и неэкспонированная фактором риска группа будет содержать 50 человек. Даже для такой большой исходной выборки из-за естественной неравномерности распределений по разным параметрам отдельные

страты могут оказаться слишком мелкими.

В каждой практической задаче объем выборки определен и конечен. Поэтому при решении приходится выбирать между учетом всех возможных влияний и надежностью выводов.

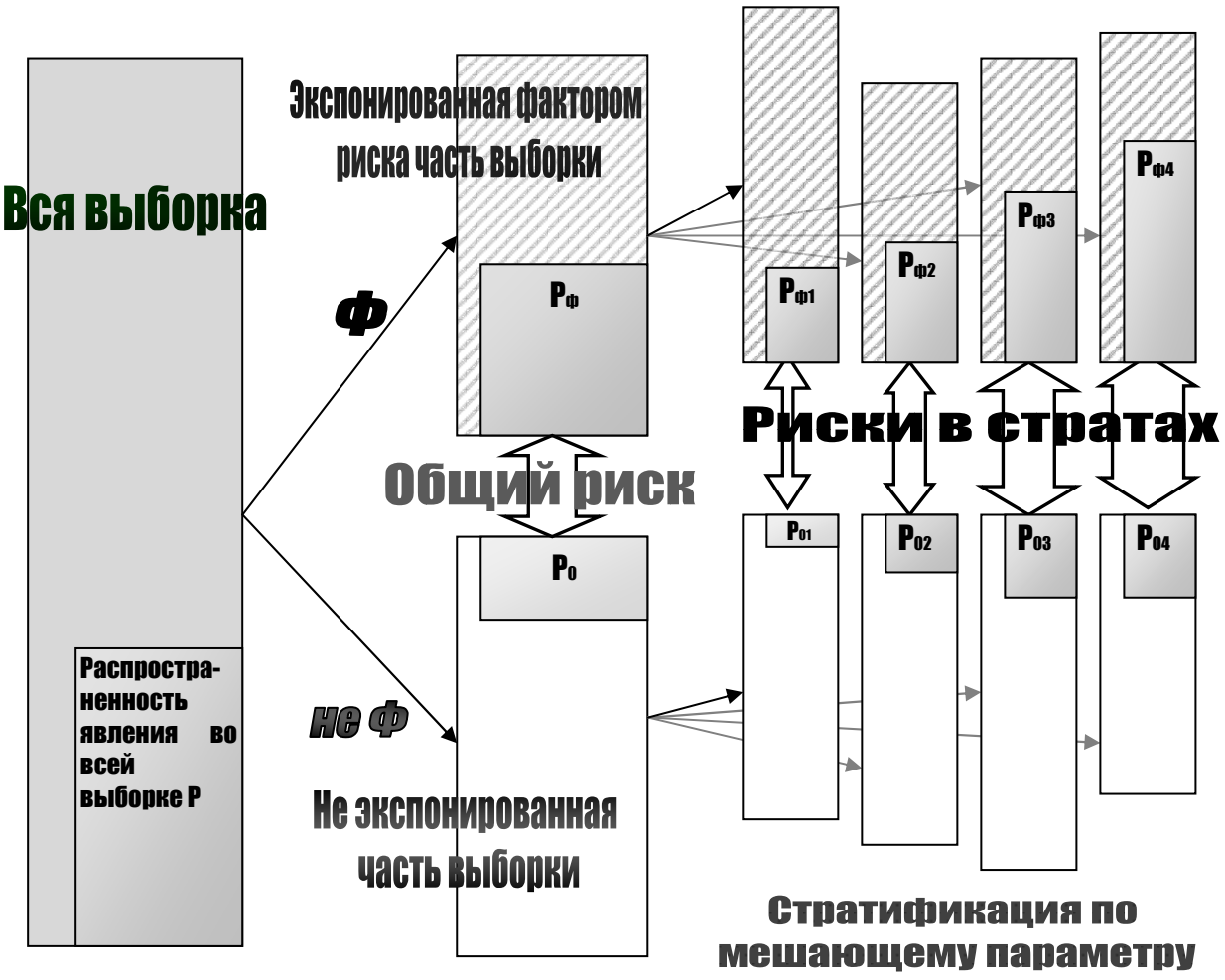


Схема 7. Вычисление рисков при наличии мешающего фактора

4.2. Вычисление объединенных относительных рисков при наличии мешающих факторов

Для вычисления рисков требуется сформировать имеющиеся данные в виде таблиц сопряженности по всем уровням изучаемого и мешающего факторов. Такую таблицу можно представить следующим образом.

Таблица 4.1. Исходные данные для вычисления объединенного риска

Уровни мешающего фактора U	**Содержание таблиц**	**Уровни изучаемого фактора F**			
		F_1	**F_2**	**…**	**F_L**
U_1	Событий	a_{11}	a_{12}	…	a_{1L}
	Наблюдений	c_{11}	c_{12}	…	c_{1L}
U_2	Событий	a_{21}	a_{22}	…	a_{2L}
	Наблюдений	c_{21}	c_{22}	…	c_{2L}
…					
U_K	Событий	a_{K1}	a_{K2}	…	a_{KL}
	Наблюдений	c_{K1}	c_{K2}	…	c_{KL}

Если ограничиться одной из таблиц 2x2, например, соответствующей первым уровням обоих факторов таблицы 4.1, то можно вычислить относительный риск появления события при уровне изучаемого фактора F_2 по сравнению с уровнем F_1 (или на уровне F_1 по сравнению с уровнем F_2) при значении мешающего фактора U_1. Для этого используются формулы, методы и программы, описанные в предыдущей части (Глава 3, п.3.4).

Таблица 4.2. Частичная таблица для вычисления риска

Уровни мешающего фактора U		**Уровни изучаемого фактора F**	
		F_1	**F_2**
U_1	Событий	a_{11}	a_{12}
	Наблюдений	c_{11}	c_{12}

Относительный риск R_{21} для таблицы 4.2:

$$R_{21} = \frac{p_2}{p_1} ,$$

где p_i могут быть уровнями, пропорциями (частотой) или шансами осуществления события, а способ вычисления стандартной ошибки $SE(p_i)$ зависит от содержания таблицы, как указано в главе 3.

При наличии нескольких уровней мешающего фактора U (таблица 4.3) для оценки объединенного мультипликативного воздействия

уровней исследуемого фактора F на частоту осуществления событий используются тест Мантеля-Ханзела (Mantel-Haenszel) и тест Вульфа (Woolf). С помощью теста Вульфа проверяется однородность (равенство) отношений шансов по стратам. Это позволяет оценить связь исследуемого и мешающего факторов. Также вычисляется объединенный риск Вульфа, но он считается менее полезным и используется реже, чем объединенный риск Мантеля-Ханзела. Тест Мантеля-Ханзела позволяет решить две взаимосвязанные задачи: во-первых, проверить равенство единице отношений шансов (OR) в стратах, и во-вторых, вычислить совокупную оценку отношения шансов по уровням изучаемого фактора F с исключением влияния мешающего фактора U.

При использовании этих тестов предполагается, что выполнены следующие условия:

► Наблюдения независимы.

Практически это означает, как минимум, что выборка случайная, без повторений.

► Наблюдения одинаково распределены.

Это значит, что все наблюдения получены одинаковым образом – нельзя смешивать данные исследований разного типа, например, телефонного опроса и личного анкетирования.

Таблица 4.3. Исходные данные для вычисления объединенного риска при двух уровнях изучаемого фактора

Уровни мешающего фактора U	**Содержание таблиц**	**Уровни изучаемого фактора F**	
		F_1	**F_2**
U_1	Событий	a_{11}	a_{12}
	Наблюдений	c_{11}	c_{12}
U_2	Событий	a_{21}	a_{22}
	Наблюдений	c_{21}	c_{22}
…			
U_K	Событий	a_{K1}	a_{K2}
	Наблюдений	c_{K1}	c_{K2}

Весь анализ проводится для двух уровней изучаемого фактора и К (K>=2) уровней мешающего фактора. Если изучаемый фактор имеет более двух уровней, весь анализ нужно повторять для каждой пары уровней.

Каждому i-му уровню фактора U: U_1, U_2, …, U_K, соответствует таблица T_i.

Таблица T_i :

a_{i1}	a_{i2}
c_{i1}	c_{i2}

Статистика (риск) Мантеля-Ханзела для сравнения частот события на двух уровнях изучаемого фактора F_1 и F_2 вычисляется по формуле (в случае, когда уровней всего два, второй уровень – F_2 – может означать отсутствие действия фактора):

$$R_{MH} = \frac{\sum_{i=1}^{K} \frac{a_{i1} \times (c_{2i} - a_{2i})}{n(i)}}{\sum_{i=1}^{K} \frac{a_{2i} \times (c_{1i} - a_{1i})}{n(i)}} \qquad (4.1)$$

Проверяется нулевая гипотеза H_0: $R_{MH}=1$
(точная формулировка проверяемой гипотезы H_0:

все составляющие риски OR(i)=1

против альтернативной гипотезы H_1:

хотя бы один из этих рисков отличен от 1).

Для проверки гипотезы используется статистика χ^2. Формула для вычисления этой статистики приведена в Приложении (тест Мантеля-Ханзела). Вычисление доверительных интервалов для R_{MH} было усовершенствовано Робинсом (Robins). Эти вычисления довольно трудоемки, поэтому для них разумно использовать статистические программы, в первую очередь NCSS, в которой наиболее полно проводится анализ объединенных рисков.

Еще один способ вычисления объединенного риска был предложен Вульфом. С помощью объединенного относительного риска Вульфа вначале проверяется взаимодействие факторов U и F (анализ однородности таблиц). Для этого используется критерий χ^2 для статистики, связывающей риски в стратах и объединенный риск Вульфа.

Предполагается, что в строке «Наблюдений» у нас количество наблюдений, а не число человеко-лет наблюдения, т.е. риск

возникновения события оценивается с помощью пропорций или шансов, но не уровней. Для каждой из таблиц вычислим следующие величины:

1. Отношение шансов OR(i), или перекрестное произведение.
2. Весовые коэффициенты таблиц для вычисления взвешенного риска W(i). Эти коэффициенты обратно пропорциональны дисперсиям ошибок для каждой таблицы.
3. Объем выборки n(i). Объем вычисляется как сумма числа наблюдений по столбцам: $n(i) = c_{i1} + c_{i2}$

   Если выполнено следующее условие:

► $a_{i1}>0$, $a_{i2}>0$, $c_{i1}-a_{i1}>0$, $c_{i2}-a_{i2}>0$ (количество всех «событий» и «не событий» в таблице T_i больше нуля),

$$\text{то} \quad OR(i) = \frac{a_{i2} \times (c_{i1} - a_{i1})}{a_{i1} \times (c_{i2} - a_{i2})}, \tag{4.2}$$

$$W(i) : \frac{1}{W(i)} = \frac{1}{a_{i1}} + \frac{1}{c_{i1} - a_{i1}} + \frac{1}{a_{i2}} + \frac{1}{c_{i2} - a_{i2}} \tag{4.3}$$

Если же хотя бы одно из чисел a_{i1}, a_{i2}, $c_{i1} - a_{i1}$, $c_{i2} - a_{i2}$ равно 0, то все значения в ячейках увеличиваются для вычислений на 0.5 (в некоторых программах допускается увеличение на другое малое число, например, 0.25).

Для вычисления объединенного взвешенного риска R_w (Вульфа) сначала вычисляется логарифм R_w как взвешенная комбинация логарифмов рисков OR(i):

$$\ln(R_w) = \frac{\sum_{i=1}^{K} W(i) \times \ln OR(i)}{\sum_{i=1}^{K} W(i)} \tag{4.4}$$

Логарифм объединенного риска распределен асимптотически нормально. Стандартная ошибка логарифма объединенного взвешенного риска R_w определяется весовыми коэффициентами W(i).

$$\text{SE}\ln(R_w) = \sqrt{\frac{1}{\sum_{i=1}^{K} W(i)}} \tag{4.5}$$

В некоторых работах (в частности, J.F.Osborn. Basic Statistical Methods for Epidemiological Studies) использование взвешенного риска предполагается в случае, когда для измерения взаимодействия факторов

используется относительный риск RR, т.е. используются пропорции или уровни для оценки частоты наблюдаемого явления.

Объединенный взвешенный риск R_w (Вульфа) используется не только как самостоятельная характеристика, но, что более существенно, является составной частью статистики для проверки однородности таблиц (однородности рисков по уровням мешающего фактора), т.е. для проверки наличия взаимодействия мешающего и изучаемого факторов. Для такой проверки используется статистика χ^2. Ее выборочное значение вычисляется как

$$\chi^2_в = \sum_{i=1}^{K} W(i) \times (\ln OR(i))^2 - (\sum_{i=1}^{K} W(i)) \times (\ln R_w)^2 , \qquad (4.6)$$

и она распределена асимптотически как χ^2(K-1).

4.3. Вычисление объединенных рисков с использованием статистических пакетов

В программе **SPSS** вычисляется риск Мантеля-Ханзела и приводится 95% доверительный интервал для R_{MH}.

SPSS → Analyze → Descriptive Statistics → Crosstabs …, далее определить флажки **Risk и Cochran's and Mantel-Haenszel statistics** в опции **Statistics ...** Риски определяются для категорий первой переменной (переменная по строкам), в зависимости от категорий второй переменной (переменная по столбцам) по всем значениям переменной, задающей уровни (**Layers**) и объединенный риск Мантеля-Ханзела. И переменная строки, и переменная столбца должны иметь ровно 2 возможных значения, уровней может быть несколько.

В программе **SYSTAT** вычисляется только сам риск Мантеля-Ханзела, без доверительных интервалов и проверки однородности.

SYSTAT → Statistics → Tables → Crosstabs …,→ Multiway Tables далее определить флажок **Mantel-Haenszel test for 2x2 sub-tables** в опции **Display ...** Риски определяются для категорий первой переменной (переменная по строкам), в зависимости от категорий второй переменной (переменная по столбцам). Как и в программе **SPSS**, и переменная строки, и переменная столбца должны иметь 2 возможных значения, результат табуляции – таблица 2×2.

Наиболее полные вычисления осуществляются в программе **NCSS**. В ней вычисляются OR(i) по стратам, риск Мантеля-Ханзела R_{MH}, R_{MH} с поправкой на непрерывность, приводятся 95% доверительный интервалы для R_{MH}, в том числе с уточнением Робинса, а также объединенный риск Вульфа R_w с доверительным интервалом, тест

проверки однородности рисков и тесты проверки отклонения рисков в стратах от 1. В качестве входной информации используется таблица перекрестного табулирования, заполненная специальным образом.

NCSS → Analysis → Desriptive Statistics → CrossTab → Analysis → Proportions → Mantel-Haenszel Test

Пример 12. (Данные НРЭР Северо-Запада). По результатам наблюдения за ликвидаторами за период 1986-2005 гг. получены следующие данные о смертности в связи с полученной дозой и возрастом участия в работах по ликвидации аварии.

Таблица П12-1. Исходные данные

Номер дозовой группы (dose_gr)	Доза, сЗв	Статус (ind_death)	Age work group (группы по возрасту участия)		
			3	2	1
			18 - 29	30 - 39	40+
3	0-5	жив	167	1207	388
		умер	14	268	108
Всего			181	1475	496
2	5.1-19.9	жив	595	1424	551
		умер	59	340	182
Всего			654	1764	733
1	20+	жив	504	1017	372
		умер	60	247	130
Всего			564	1264	502
Суммы			1399	4503	1731

На приведенном далее Рис.П12-1 видно, что доля умерших увеличивается как с увеличением возраста участия, так и с ростом полученной дозы. Требуется оценить риски смерти, связанные отдельно с возрастом и дозой.

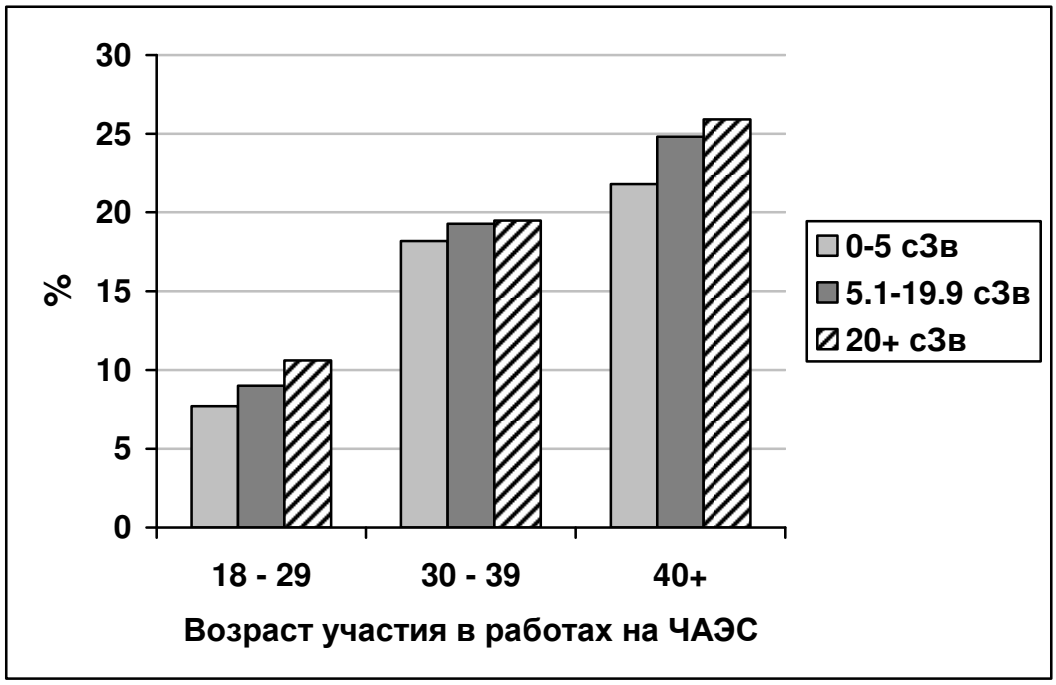


Рис. П12-1. Процент ликвидаторов, умерших к 2006 году, в зависимости от дозы и возраста участия в аварийно-восстановительных работах

Данные таблицы П12-1 для работы программы NCSS - вычисление рисков Мантеля-Ханзела - должны быть представлены следующим образом в виде таблицы NCSS Data.

ind_death (1 – умер, 2 – жив)	dose_gr (по табл.П12-1)	age_work_gr (по табл.П12-1)	Count (количество)
1	1	1	130
2	1	1	372
1	2	1	182
2	2	1	551
1	3	1	108
2	3	1	388
1	1	2	247
2	1	2	1017
1	2	2	340
2	2	2	1424
1	3	2	268
2	3	2	1207
1	1	3	60
2	1	3	504
1	2	3	59
2	2	3	595
1	3	3	14
2	3	3	167

В программе уровни изучаемого фактора сравниваются попарно, поэтому для того, чтобы сравнить между собой три уровня, следует использовать фильтр и выбирать последовательно пары: $F_1 – F_2$, $F_1 – F_3$, $F_2 – F_3$.

12.А. Вначале выберем в качестве мешающего фактора дозу и вычислим относительные риски смерти, связанные с возрастом ликвидаторов.

Для сравнения 2 и 3 возрастной групп задается фильтр: age_work_gr>1.

В качестве Count Variable задается переменная Count,

Disease Variable – ind_death (меньшее значение этой переменной задает интересующее нас событие, риск которого требуется оценить, в данном случае – смерти: ind_death =1 означает «умер», ind_death = 2 – «жив»).

Exposure Variable – age_work_gr (меньшее значение этой переменной соответствует наличию экспозиции, т.е. определяется риск в группе age_work_gr = 2 по сравнению с группой age_work_gr = 3).

Stratum Variable – dose_gr (могут быть заданы от 1 до 4 переменных, определяющих стратификацию). Порядок значений стратифицирующей переменной не имеет значения.

Delta Value – величина, которая добавляется к каждой ячейке таблицы 2×2 (T_i), если хотя в одной из них содержится 0 (формулы для вычисления OR(i)). Традиционно это 0.5, однако недавние исследования показали, что иногда более целесообразно использовать $\delta = 0.25$.

Alpha – уровень значимости, используемый при вычислении доверительных интервалов. Обычно выбирается $\alpha = 0.05$.

Результаты вычислений.

Mantel-Haenszel Test Report

Filter	age_work_gr>1
Counts Variable	count

Strata Count Section (описание данных)

Strata	dose_gr	A	B	C	D	Sample Odds Ratio
1	1	247	1017	60	504	2.0401
2	2	340	1424	59	595	2.4079
3	3	268	1207	14	167	2.6486

A: age_work_gr = 2, ind_death = 1 («умер»)
B: age_work_gr = 2, ind_death = 2 («жив»)
C: age_work_gr = 3, ind_death = 1
D: age_work_gr = 3, ind_death = 2

Strata Detail Section (относительные риски OR и доверительные 95% интервалы для них в отдельных стратах по дозе)

Strata	Lower 95.0% C.L.	1/2-Corrected Odds Ratio	Upper 95.0% C.L.	Exact Test	Proportion Exposed	Proportion Diseased
1	1.4937	2.0342	2.7910	0.0000	0.6915	0.1679
2	1.7805	2.4001	3.2623	0.0000	0.7295	0.1650
3	1.4729	2.6079	4.8477	0.0002	0.8907	0.1703

Результаты вычислений для каждой из страт (таблиц 2×2).

1/2-Corrected Odds Ratio – это отношение шансов, вычисленное с использованием Delta Value, а затем скорректированное с помощью специальной итерационной процедуры.

Lower и **Upper 95.0% C.L. –** доверительный интервал для отношения шансов в каждой страте. Вычисляется по формулам, приведенным в Приложении.

Exact Test - точный тест Фишера для таблицы 2×2. Проверяется нулевая гипотеза H_0: OR=1. Она отвергается, если значение теста меньше выбранного уровня значимости.

Proportion Exposed – доля экспонированных фактором риска лиц во всей таблице, в данном случае – доля лиц из 2 возрастной группы в каждой страте.

Proportion Diseased – доля «случаев» во всей таблице, в данном случае – пропорция умерших.

В нашем примере все риски OR(i) значимо превышают 1.

Mantel-Haenszel Statistics Section

Method	Lower 95.0% C.L.	Estimated Odds Ratio	Upper 95.0% C.L.	Chi-Square Value	DF	Prob Level
MH C.C.	1.8808	2.2825	2.7700	69.83	1	0.000000
MH	1.8827	2.2825	2.7673	70.53	1	0.000000
Robins	1.8754	2.2825	2.7780			
Woolf	1.8651	2.2705	2.7639	66.78	1	0.000000
Heterogeneity Test				0.93	2	0.628853

Рекомендуется следующий порядок рассуждений на основе полученных результатов:

1 – прежде всего проверяется гипотеза об однородности (равенстве) отношений шансов во всех стратах (Heterogeneity Test). В данном случае эта гипотеза принимается (p = 0.629).

2 - используется MH C.C. проверка гипотезы о том, что все отношения шансов равны 1. В нашем примере эта гипотеза отвергается (p < 0.0000005).

3 – для оценки доверительного интервала и R_{MH} предлагается использовать оценку Робинса.

Таким образом, **R_{MH} (2-3) = 2.2825**, доверительный интервал (1.8754, 2.7780).

Отдельно укажем содержание каждой из строк.

MH C.C. – метод Мантеля-Ханзела с поправкой на непрерывность, используется для вычисления R_{MH} и доверительного интервала для риска, а также проверки гипотезы о том, что все риски в отдельных стратах равны 1 против альтернативной гипотезы, что хотя бы один из рисков отличен от 1.

MH - метод Мантеля-Ханзела вычисления R_{MH}, доверительного интервала для риска и проверки гипотезы о равенстве всех рисков в стратах 1, но без поправки на непрерывность.

Robins – модификация Робинса метода Мантеля-Ханзела, Она касается только вычисления доверительного интервала, поэтому в данной строке отсутствуют значения выборочной статистики и p-значения. Оценка объединенного риска в этой строке совпадает с риском MH C.C.

Woolf – объединенный риск (R_w) по методу Вульфа.

Heterogeneity Test – разработанный Вульфом тест для проверки общей гипотезы об однородности рисков в стратах: нулевая гипотеза предполагает, что все риски равны между собой, но не обязательно равны 1.

Соответствующие формулы даны в Приложении.

При сравнении 1 и 2 уровней фактора «age_work_group» получим:

Риски однородны, все отличны от 1, **R_{MH} (1-2) = 1.3614**, доверительный интервал (1.1919, 1.5549).

При сравнении 1 и 3 уровней фактора «age_work_group» получим:

Риски однородны, все отличны от 1, **R_{MH} (1-3) = 3.1714**, доверительный интервал (2.5608, 3.9276).

Таким образом, все полученные риски значимо превышают 1.

12.Б. В качестве фактора риска выбирается полученная доза, а в качестве мешающего параметра – возраст участия в работах.

Сравнивая 1 и 2 уровни изучаемого фактора, получим: риски однородны, но каждый статистически не отличается от 1. Значимого влияния фактора «доза» нет.

Strata Detail Section

Strata	Lower 95.0% C.L.	1/2-Corrected Odds Ratio	Upper 95.0% C.L.	Exact Test	Proportion Exposed	Proportion Diseased
1	0.8082	1.0583	1.3847	0.6894	0.4065	0.2526
2	0.8439	1.0174	1.2260	0.8523	0.4174	0.1939
3	0.8085	1.2004	1.7829	0.3839	0.4631	0.0977

Mantel-Haenszel Statistics Section

Method	Lower 95.0% C.L.	Estimated Odds Ratio	Upper 95.0% C.L.	Chi-Square Value	DF	Prob Level
MH C.C.	0.9087	1.0519	1.2177	0.46	1	0.497938
MH	0.9154	1.0519	1.2088	0.51	1	0.475685
Robins	0.9154	1.0519	1.2088			
Woolf	0.9154	1.0519	1.2089	0.51	1	0.475536
Heterogeneity Test				0.60	2	0.740698

При сравнении 2 и 3 уровней фактора «dose_group» получим те же результаты.

Mantel-Haenszel Statistics Section

Method	Lower 95.0% C.L.	Estimated Odds Ratio	Upper 95.0% C.L.	Chi-Square Value	DF	Prob Level
MH C.C.	0.9590	1.1120	1.2894	1.98	1	0.159733
MH	0.9626	1.1120	1.2846	2.08	1	0.149158
Robins	0.9627	1.1120	1.2845			
Woolf	0.9623	1.1117	1.2842	2.07	1	0.150312
Heterogeneity Test				0.40	2	0.819819

При сравнении 1 и 3 уровней фактора «dose_group» также получим, что влияние дозы статистически незначимо.

Mantel-Haenszel Statistics Section

Method	Lower 95.0% C.L.	Estimated Odds Ratio	Upper 95.0% C.L.	Chi-Square Value	DF	Prob Level
MH C.C.	0.9885	1.1581	1.3568	3.30	1	0.069188
MH	0.9919	1.1581	1.3523	3.45	1	0.063379
Robins	0.9921	1.1581	1.3519			
Woolf	0.9905	1.1567	1.3507	3.38	1	0.065881
Heterogeneity Test				1.07	2	0.586868

После проведения анализа влияния двух факторов на уровень смертности ликвидаторов за период наблюдения 1986-2005 гг. можно сделать вывод, что полученная доза не оказывает заметного влияния на этот уровень, в отличие от возраста ликвидаторов.

4.4. Стандартизация

Еще одним методом, позволяющим исключить влияние мешающих параметров на исследуемый показатель, является стандартизация. Наиболее часто на практике производится стандартизация по возрасту. Необходимость применения стандартизации покажем на примере.

По данным из сборника Комитета по здравоохранению Администрации Санкт-Петербурга «Анализ медицинских данных государственного статистического наблюдения» (В.М.Дорофеев и др., СПб, 2003), в 2001 году показатель общей смертности на 1000 человек

населения СПб занимал 26 место среди 89 субъектов РФ и был на 5.1% выше среднего по России. Однако стандартизованный показатель (стандартизация по европейскому стандарту) занимал 67 место и был ниже среднего по стране. Эти данные приведены в следующей таблице и на рисунках.

Таблица 4.4. Обычные и стандартизованные показатели общей смертности

Территория	Обычные показатели			Стандартизованные показатели		
	М	Ж	Все	М	Ж	Все
РФ	17.9	13.7	15.7	20.9	10.3	14.8
СПб	18.2	15.2	16.5	19.4	9.9	13.8

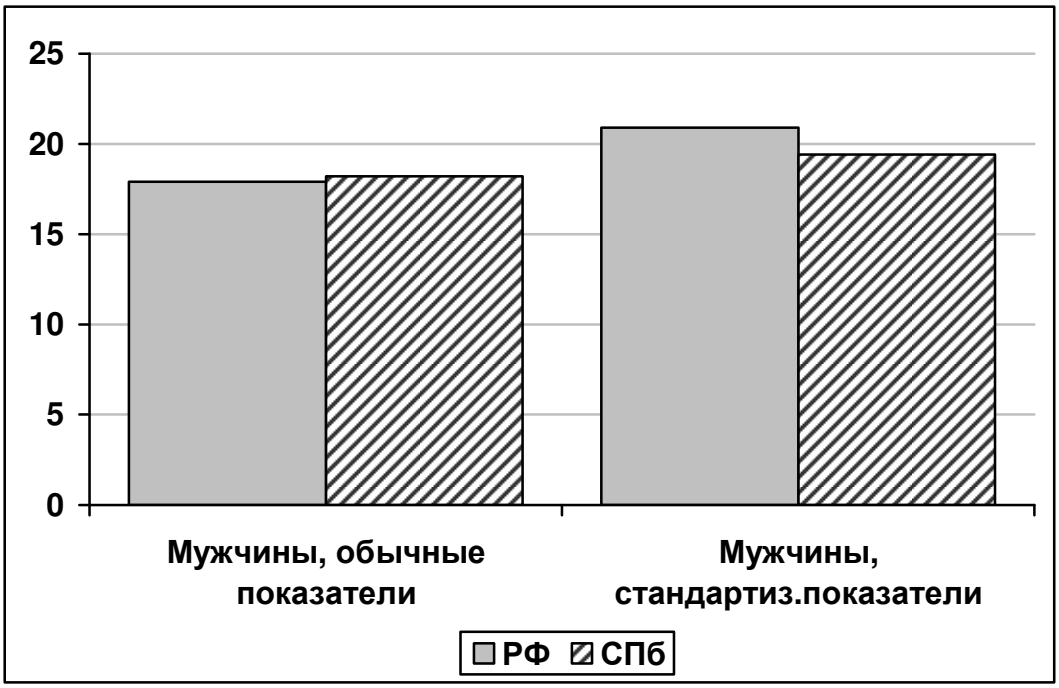


Рис. 4.1. Показатели смертности мужчин РФ и СПб в 2001

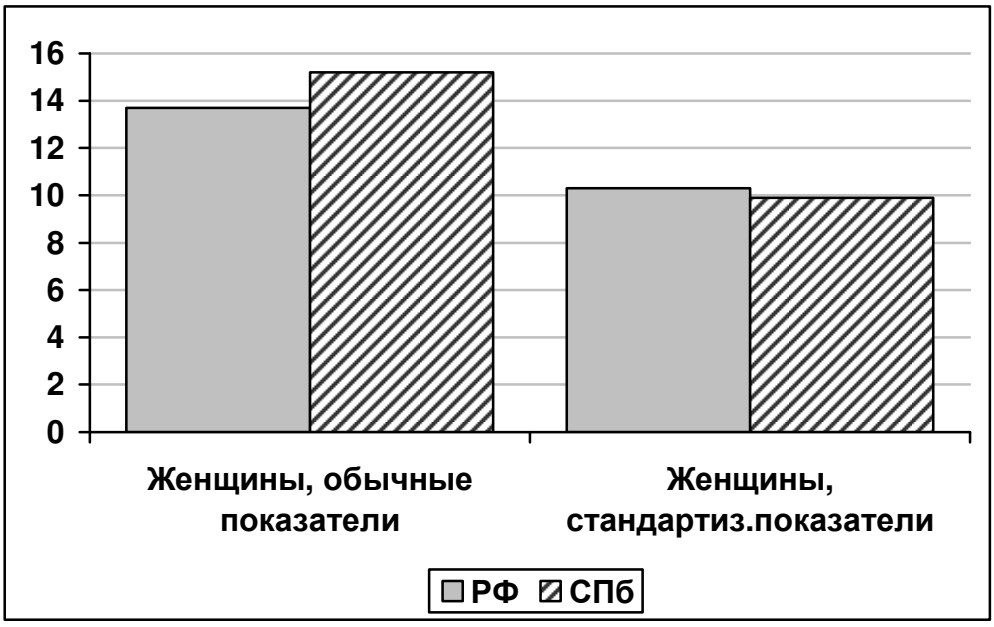


Рис. 4.2. Показатели смертности женщин РФ и СПб в 2001 г.

Таким образом, исключение влияния на результат возрастного распределения населения на территории приводит к изменению выводов о соотношении показателей смертности.

Для формализации методов стандартизации используется следующая таблица.

Таблица 4.5. Исходные данные для проведения стандартизации

Группы мешающего параметра	**Изучаемая популяция**			**Стандартная популяция**		
	Кол-во объектов под риском	Кол-во случаев	Уровень	Кол-во объектов под риском	Кол-во случаев	Уровень
1	n_1	r_1	p_1	N_1	R_1	P_1
2	n_2	r_1	p_2	N_2	R_2	P_2
…						
K	n_k	r_k	p_k	N_k	R_k	P_k
Всего	n	r	P	N	R	P

Если в качестве мешающего параметра выступает возраст, то группы 1, 2. …, k – могут быть стандартными возрастными группами с диапазоном 5 или 10 лет (например, 20-29, 30-39,…) или сконструированными в целях исследования группами с другим диапазоном (например, 7 лет).

Различают несколько методов стандартизации. Чаще всего

используется прямая стандартизация, но также существуют методы непрямой и обратной стандартизации, оценка максимального правдоподобия, метод Юла и некоторые другие. Выбор варианта стандартизации в основном определяется целью ее проведения и наличием необходимой информации. В основном задачи, для которых применяются методы стандартизации, описываются схемой 8. Это означает, что изучаемый нами фактор действует только на выборку, и мы сравниваем распространенность явления в выборке с распространенностью в популяции, с помощью стандартизации исключая влияние мешающих параметров. Для решения задач, соответствующих схеме 7 (сравнение распространенности явления в двух выборках), из всех методов стандартизации может использоваться только один – прямая стандартизация.

Схема 8. Методика применения стандартизации

Прямая стандартизация

Обозначим символом $t_i = N_i/N$ долю i-ой группы мешающего параметра в общем объеме стандартной популяции. Тогда стандартизованный уровень

$$p_{n.cm.} = \sum_{i=1}^{k} t_i \times p_i \qquad (4.7)$$

Стандартная ошибка прямого стандартизованного уровня

$$\text{Ст.ош.}(p_{\text{п.ст.}}) = \sqrt{\sum_{i=1}^{k} t_i^{\,2} \times \frac{p_i(1-p_i)}{n_i}} \qquad (4.8)$$

Сравнительный индекс (CMI для смертности, CII для первичной заболеваемости) при сравнении со стандартной популяцией

$$\text{CMI} = \frac{\sum_{i=1}^{k} p_i t_i}{P} = p_{\text{п.ст.}}/P \qquad (4.9)$$

Стандартная ошибка сравнительного индекса (относительного риска) в этом случае

$$\text{Ст.ош.}(\text{CMI}) = \sqrt{\sum_{i=1}^{k} \frac{N_i^{\,2} p_i(1-p_i)}{P^2 n_i}} \qquad (4.10)$$

Пример 13. (Данные Комитета по здравоохранению Администрации Санкт-Петербурга «Анализ медицинских данных государственного статистического наблюдения» (В.М.Дорофеев и др., СПб, 2003) и учебника «Демография», Медков В.М., Москва, Инфра-М, 2004). Известны повозрастные показатели смертности мужского населения в СПб, общие показатели смертности в 2001 г. в России и СПб (17.9 и 18.18 на 1000 человек населения, соответственно) и структура мужского населения России.

Таблица П13-1. Повозрастные уровни смертности мужчин в СПб на 1000 человек и возрастная структура мужского населения России в 2001 году

Возраст	Повозрастные показатели смертности в СПб (p_i)	Возрастная структура в России (t_i)	Показатели смертности в СПб с учетом доли в возрастной структуре (t_i*p_i)
0-1	10.79	0.0096	0.104
1-4	0.86	0.0478	0.041
5-9	0.47	0.0578	0.027
10-14	0.41	0.0868	0.036
15-19	1.66	0.0894	0.148
20-24	4.09	0.0807	0.330
25-29	5.53	0.0773	0.427
30-34	4.86	0.0713	0.347
35-39	8.21	0.0824	0.677
40-44	13.28	0.0907	1.204
45-49	18.47	0.0810	1.496
50-54	28.70	0.0643	1.845
55-59	31.43	0.0319	1.003
60-64	42.41	0.0536	2.273
65-69	53.47	0.0332	1.775
70-74	73.54	0.0307	2.258
75-79	92.36	0.0123	1.136
80-84	125.77	0.0049	0.616
85+	170.75	0.0039	0.666
Всего	**Общий показатель смертности 18.18**	**1.**	**Стандартизованный показатель смертности 16.409**

Эти данные позволяют вычислить прямой стандартизованный уровень смертности для СПб (16.4) и сравнительный индекс CMI:
CMI = 16.409/17.9 = 0.92
Однако вычислить ошибки уровня и сравнительного индекса по этим данным невозможно – неизвестны структура и количество мужского населения СПб.
Пример 14. (Данные НРЭР Северо-Запада). По данным НРЭР о смертности ликвидаторов, проживающих в СПб и Ленинградской области (сравнимые по численности когорты) можно построить следующий график.

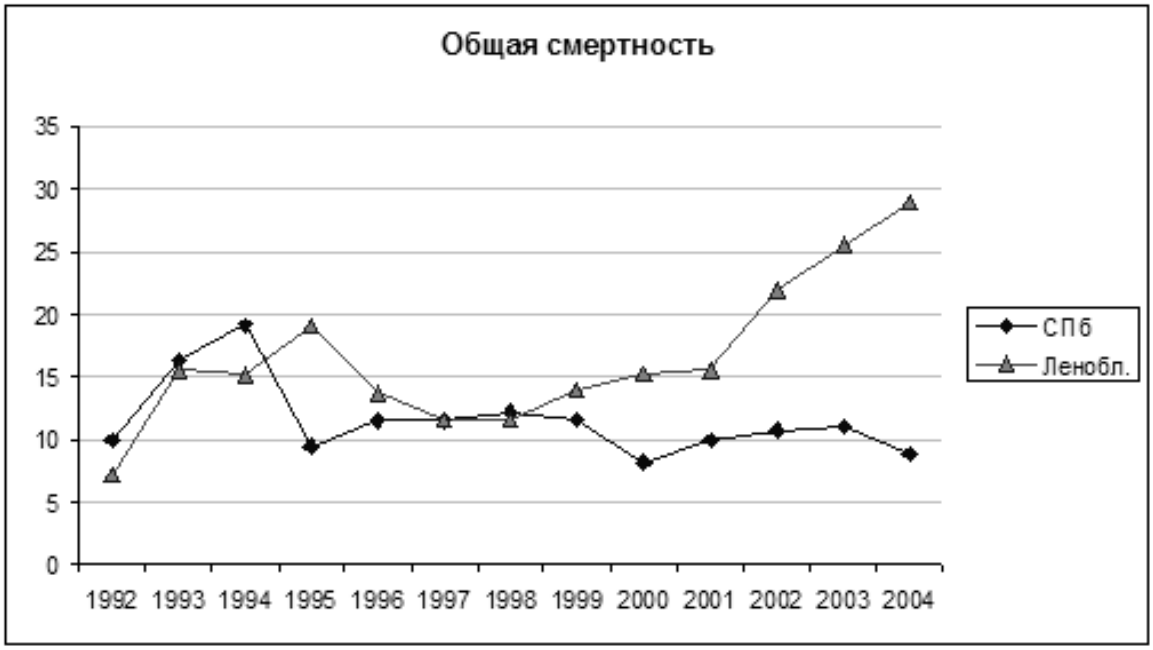


Рис. П14-1. Динамика общей смертности ликвидаторов Ленинградской области и Санкт Петербурга

Начиная с 2000 г. наблюдаются резкие отличия в общих показателях смертности ликвидаторов этих двух субъектов РФ. Для того, чтобы выяснить, объясняются ли эти отличия различной возрастной структурой в регионах, вычислим повозрастные уровни смертности. Повозрастные уровни смертности в Ленинградской области также превышают соответствующие показатели Санкт-Петербурга, однако соотношения повозрастных уровней отличаются. Для получения единой характеристики этого превышения возможно как вычисление объединенных рисков, так и стандартизация.

Таблица П14-1. Повозрастные уровни смертности на 1000 человек и возрастная структура ликвидаторов в 2000 - 2004 гг.

	СПб					**Ленобласть**				
	2000	2001	2002	2003	2004	2000	2001	2002	2003	2004
Возраст	**Количество умерших**									
30-34	0	0	0	0	0	1	0	0	0	0
35-39	1	0	0	0	0	1	1	1	1	0
40-44	1	1	0	0	1	4	2	4	3	3
45-49	6	4	4	2	0	10	7	9	10	6
50-54	7	11	13	10	13	5	12	16	19	12
55-59	6	8	9	14	4	3	2	6	8	16
60-64	5	10	5	4	4	4	4	5	4	11
65-69	3	2	6	6	7	0	0	0	2	4
70-90	2	2	4	6	5	1	1	0	0	0

Возраст	СПб					Ленобласть				
	2000	2001	2002	2003	2004	2000	2001	2002	2003	2004
	Количество наблюдаемых									
30-34	139	82	37	8	3	61	37	14	3	2
35-39	327	333	308	268	207	170	147	138	121	94
40-44	495	435	398	374	371	307	263	238	195	183
45-49	793	710	669	627	580	557	504	457	416	358
50-54	1000	1050	1011	974	894	527	592	629	634	558
55-59	391	485	628	761	871	169	209	239	303	405
60-64	440	467	407	355	357	81	91	114	113	128
65-69	112	131	220	296	367	22	21	30	44	54
70-90	86	117	137	152	172	7	9	10	12	18
	Повозрастная смертность									
30-34	0.00	0.00	0.00	0.00	0.00	16.39	0.00	0.00	0.00	0.00
35-39	3.06	0.00	0.00	0.00	0.00	5.88	6.80	7.25	8.26	0.00
40-44	2.02	2.30	0.00	0.00	2.70	13.03	7.61	16.81	15.39	16.39
45-49	7.57	5.63	5.98	3.19	0.00	17.95	13.89	19.69	24.04	16.76
50-54	7.00	10.48	12.86	10.27	14.54	9.49	20.27	25.44	29.97	21.51
55-59	15.35	16.50	14.33	18.40	4.59	17.75	9.57	25.11	26.40	39.51
60-64	11.36	21.41	12.29	11.27	11.20	49.38	43.96	43.86	35.40	85.94
65-69	26.79	15.27	27.27	20.27	19.07	0.00	0.00	0.00	45.46	74.07
70-90	23.26	17.09	29.20	39.47	29.07	142.86	111.11	0.00	0.00	0.00

В данном примере возможно провести прямую стандартизацию. В качестве стандартного возрастного распределения разумно взять среднюю за рассматриваемый период возрастную структуру в двух регионах. Однако, исходя из того, что младшая и старшая из рассматриваемых возрастных групп в отдельные годы являются малочисленными (что влечет за собой большую величину стандартной ошибки стандартизованного показателя), объединим их с соседними возрастными группами. Получим следующие коэффициенты **(t_i):**

Возраст	**Средняя возрастная структура за период (t_i)**
30-39	0.089
40-44	0.115
45-49	0.200
50-54	0.278
55-59	0.157
60-64	0.090
65-90	0.071

При пересчете показателей смертности с учетом стандартной возрастной структуры получим следующие результаты:

	2000	2001	2002	2003	2004
	Общий показатель смертности				
СПб	8.20	9.97	10.75	11.01	8.90
Ленобласть	15.25	15.48	21.94	25.53	28.89
	Стандартизованный показатель смертности				
СПб	9.18	9.96	10.12	9.34	7.69
Ленобласть	19.27	18.09	21.39	24.95	28.57
	Стандартная ошибка стандартизованного показателя смертности				
СПб	5.17	4.34	4.11	3.82	3.43
Ленобласть	13.13	12.11	8.98	11.06	12.32
	Сравнительный индекс стандартизованных показателей смертности				
CMI	2.1	1.8	2.1	2.7	3.7

Стандартная ошибка стандартизованного показателя смертности весьма велика, особенно для Ленобласти. Это не удивительно, учитывая малые количество наблюдений для младших и старших возрастных групп. Единственный способ улучшения этой ситуации – укрупнение возрастных групп и исключение крайних малочисленных групп. Поэтому рассмотрим 10-летние возрастные интервалы, начиная с 35 лет. Получим следующую структуру стандарта (средняя в двух регионах за период 2000-2004 гг.).

Возраст	**Средняя возрастная структура за период (t_i)**
35-44	0.287
45-54	0.448
55-64	0.209
65+	0.056

Для таких возрастных групп стандартные ошибки стандартизованных показателей смертности становятся уже приемлемыми.

Территория	**2000**	**2001**	**2002**	**2003**	**2004**
	Стандартизованный показатель смертности				
СПб	8.37	9.27	9.16	8.63	6.78
Ленобласть	17.80	21.46	27.13	31.53	25.96
	Стандартная ошибка стандартизованного показателя смертности				
СПб	2.27	1.81	1.84	1.66	1.52
Ленобласть	6.61	6.47	3.31	5.55	5.69

	Сравнительный индекс стандартизованных показателей смертности				
	2000	**2001**	**2002**	**2003**	**2004**
CMI	2.1	2.3	3.0	3.7	3.8

Рис. П14-2. Стандартизованные показатели смертности ликвидаторов

Заметим, что прямая стандартизация не позволяет вычислить ошибки сравнительных индексов стандартизованных показателей, поскольку приведенные выше формулы применимы только в задаче сравнения со стандартной популяцией. Поэтому для оценки кратности превышения объединенных показателей смертности следует использовать методику вычисления объединенных рисков Мантеля-Ханзела. Эта методика вполне адекватна для данного примера. Объединенные риски были рассчитаны с учетом всех исходных возрастных групп.

Вычисление объединенных рисков (NCSS)

Методы	**Оценки**		**2000 г.**	**2001 г.**	**2002 г.**	**2003 г.**	**2004 г.**
Мантеля-Ханзела	**Оценка отношения шансов**		2.21	1.87	2.42	2.87	4.06
Робинса	**95.0% доверит. интервал**	Нижняя граница	1.30	1.13	1.52	1.83	2.56
		Верхняя граница	3.75	3.10	3.84	4.50	6.43
Тест однородности Вульфа	**p-значение**		0.690	0.795	0.795	0.626	0.143

Рис. П14-3. Сравнение показателей смертности с помощью объединенного отношения шансов

Таким образом, оба метода показывают существенное превышение уровней смертности ликвидаторов Ленобласти, по сравнению с ликвидаторами СПб, в 2000 – 2004 гг.

Непрямая стандартизация

Используется, если

1) неизвестны групповые уровни в исследуемой популяции, или

2) величина исследуемой популяции мала и, вследствие этого, количество случаев в каждой группе мешающего параметра мало, что приводит к большой выборочной ошибке.

$$p_{н.ст.} = \frac{rP}{\sum_{i=1}^{k} P_i n_i} \qquad (4.11)$$

В знаменателе – ожидаемое в соответствии со стандартными групповыми уровнями количество случаев в исследуемой популяции. В числителе – реальное число случаев, умноженное на стандартный уровень.

Стандартная ошибка непрямого стандартизованного уровня

$$\text{Ст.ош.}(p_{н.ст.}) = \sqrt{\frac{P^2 \sum_{i=1}^{k} n_i p_i q_i}{(\sum_{i=1}^{k} P_i n_i)^2}} \qquad (4.12)$$

При малых p_i можно использовать приближенное выражение

$$\text{Ст.ош.}(p_{н.ст.}) \approx \frac{стандартиз.уровень}{\sqrt{r}} = \frac{p_{н.ст.}}{\sqrt{r}} \qquad (4.13)$$

Стандартизованное отношение первичной заболеваемости – SIR. Вычисляется аналогично стандартизованному отношению смертности SMR:

$$\text{SMR} = \frac{r}{\sum_{i=1}^{k} P_i n_i} \qquad (4.14)$$

Стандартная ошибка стандартизованного отношения

$$\text{Ст.ош.(SMR)} = \sqrt{\frac{\sum_{i=1}^{k} n_i p_i q_i}{(\sum_{i=1}^{k} P_i n_i)^2}} \approx \frac{SMR}{\sqrt{r}} \qquad (4.15)$$

- приближенная формула для малых значений p_i .

Другие методы стандартизации существуют в большом количестве, но применяются достаточно редко.

ГЛАВА 5. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ: ОЦЕНКА ВЛИЯНИЯ НЕСКОЛЬКИХ ФАКТОРОВ НА РЕЗУЛЬТИРУЮЩИЙ ДИСКРЕТНЫЙ ПОКАЗАТЕЛЬ

5.1. Логистическая регрессия для бинарного отклика.

Часто в исследованиях встает вопрос оценки частот благоприятного и неблагоприятного исходов (бинарного отклика) в связи с несколькими факторами, часть из которых является дискретными переменными, остальные – непрерывными переменными. Для такой оценки применяется логистическая регрессия. В качестве меры воздействия фактора на частоту возникновения события логистическая регрессия использует отношение шансов.

Логистическая регрессия аналогична обычной множественной регрессии, за исключением того, что зависимая переменная (Y) является бинарной (т.е. имеет два значения, 0 и 1), а не непрерывной. Этот метод конкурирует с дискриминантным анализом как способом анализа переменной, задающей бинарный отклик. Вообще для анализа такого отклика, если независимые переменные непрерывны, могут использоваться и дискриминантный анализ, и логистическая регрессия.

Дискриминантный анализ является оптимальным методом анализа бинарного отклика в случае, когда выполнены основные условия его применения: данные получены из двух многомерных нормальных распределений с равными ковариационными матрицами. Если же какие-то из условий не выполнены, а это очевидным образом случится, если некоторые независимые переменные являются дискретными, а не непрерывными, данный метод уже не будет оптимальным, более того, он станет неадекватным исследуемым данным.

В этом случае как раз применима логистическая регрессия. Для этого метода несущественно, являются ли независимые переменные дискретными или непрерывными, выполнены ли условия нормальности, каковы дисперсии переменных. Конечно, вычисления происходят намного медленнее, чем в дисперсионном анализе и регрессионном анализе. Замедление ощутимо при больших объемах выборок и большом количестве переменных. Именно это обстоятельство препятствовало использованию метода в прошлом, однако при нынешнем развитии компьютерных технологий уже не является существенным ограничением.

Линейная логистическая модель для зависимой переменной Y, имеющей два значения (y_1=0 и y_2=1), и независимых переменных X_1, …

, X_p произвольной природы имеет вид: вероятность (Prob) того, что Y принимает значение y_2,

$$\text{Prob}\,(Y = 1) = \frac{1}{1 + \text{Exp}(-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p))} \qquad (5.1)$$

Здесь $\{\beta_i\}$ – логистические регрессионные коэффициенты. Их оценки обозначаются $\{b_i\}$.

Вероятность второго исхода Prob (Y = 0) = 1 – Prob (Y = 1).

Эта модель может быть записана линейным образом с использованием логарифмирования отношения вероятностей двух исходов:

$$\text{Ln}\left(\frac{\text{Prob}\,(Y = 1)}{\text{Prob}\,(Y = 0)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (5.2)$$

Левая часть этого выражения называется логит-преобразованием вероятности или еще логарифмом отношения шансов.

5.2. Логит и логистическое преобразование

В множественной регрессии набор объясняющих переменных используется для того, чтобы предсказать среднее значение зависимой переменной. В логистической регрессии объясняющие переменные используются для предсказания логита зависимой переменной.

Пусть бинарная зависимая переменная имеет значения 0 (отрицательный отклик) и 1 (положительный отклик). Тогда среднее значение этой переменной – это доля положительных откликов, и оно совпадает с вероятностью положительных откликов (1). Обозначим ее p. Тогда 1–p – вероятность отрицательного отклика (0). Отношение p/(1-p), называемое *шансами события*, - это отношение вероятности осуществления события к вероятности его неосуществления. Шансы события являются одной из возможных характеристик распространенности явления, наряду с его вероятностью, но в отличие от нее могут иметь значения от 0 до бесконечности (границы максимальных значений нет). По шансам события можно вычислить его вероятность, а по вероятности – шансы. Для редких событий эти характеристики близки по величине. *Логитом* называется логарифм шансов. Это преобразование удобно для моделирования линейной комбинации объясняющих переменных, поскольку может иметь любое числовое значение.

$$l = \text{logit}(p) = \ln\left(\frac{p}{1 - p}\right) \qquad (5.3)$$

В таблице приведены значения шансов и логитов для различных значений p.

P	P/(1-P)	Logit(P)	P	P/(1-P)	Logit(P)
0.001	0.001	-6.907	0.999	999.0	6.907
0.01	0.010	-4.595	0.99	99.0	4.595
0.05	0.053	-2.944	0.95	19.0	2.944
0.10	0.111	-2.197	0.90	9.0	2.197
0.20	0.250	-1.386	0.80	4.0	1.386
0.30	0.429	-0.847	0.70	2.3	0.847
0.40	0.667	-0.405	0.60	1.5	0.405
0.50	1.000	0.000			

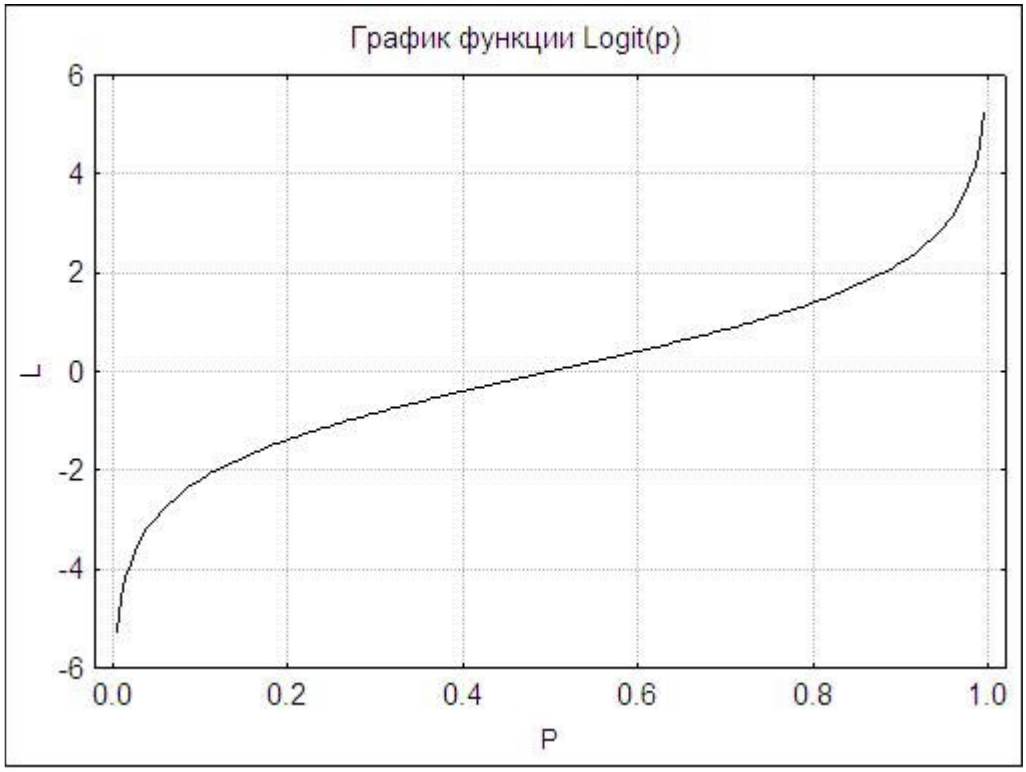


Рис. 5.1. График функции l=logit(p)

Логистическое преобразование является обратным к логит-преобразованию и позволяет определить значение p по значению l.

$$p = \text{logistic}(l) = \frac{e^l}{1 + e^l} \qquad (5.4)$$

Логарифм отношения шансов

Для сравнения двух пропорций (например, частоты положительного отклика для разных полов – мужского и женского) используется разница между двумя логарифмами шансов (глава 3, п.3.4).

$$l_1 - l_2 = \text{logit}(p_1) - \text{logit}(p_2) = (\text{после преобразований}) \ln(OR_{12})$$

Эта разность обычно называется логарифмом отношения шансов. Отношение шансов используют для сравнения пропорций в разных группах. Заметим, что логистическое преобразование тесно связано с отношением шансов. Обратное преобразование

$$OR_{12} = e^{(l_1 - l_2)} \qquad (5.5)$$

5.3. Логистическая регрессия – общие уравнения

В общем виде рассматривают *множественную логистическую регрессию* для описания дискретной зависимой переменной с конечным числом (2 и более) значений. Множественная логистическая регрессия представляет дискретную переменную Y, имеющую G ($G \geq 2$) значений $\{Y_1, Y_2, \ldots, Y_G\}$ через набор из p независимых переменных $X_1, X_2, \ldots, X_p$. Заметим, что при применении логистической регрессии *не предполагается* какое-либо упорядочение значений зависимой переменной, Y используется как номинальная переменная. Одно из ее значений используется для определения базовой или референтной группы, а все остальные выступают равноправно как метки опытных или исследуемых групп. Разница между множественной логистической регрессией и логистической регрессией для бинарного отклика – чисто техническая, определяемая числом групп G. Однако в тех случаях, когда исследователь может выбирать между применением нескольких бинарных и полиномиальной зависимой переменной, следует остановить свой выбор именно на бинарных переменных, поскольку интерпретация полученных результатов будет проще. В частности, независимые переменные, необходимые для описания одной группы, могут оказаться излишними при описании другой. А при использовании множественной логистической регрессии они все должны быть включены в уравнения.

Обозначим набор независимых переменных $X = (X_1, X_2, \ldots, X_p)$, а набор соответствующих всем значениям зависимой переменной параметров β обозначим

$$B_g = \begin{pmatrix} \beta_{g1} \\ \ldots \\ \beta_{gp} \end{pmatrix}$$

Если для бинарной зависимой переменной логистическая модель задается одним уравнением, то в общем случае для этого требуется G-1 уравнение - по количеству значений зависимой переменной минус 1 –

из-за использования одной из групп, обычно первой, в качестве референтной. Необходимость референтной группы связана с тем, что логистическая модель описывает не вероятности, а отношения вероятностей принадлежности к группам:

$$\ln\left(\frac{p_g}{p_1}\right) = \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \qquad (5.6)$$

p_g – это вероятность того, что наблюдение, для которого независимые переменные имеют значения X_1, X_2, …, X_p, относится к группе g, т.е. зависимая переменная Y принимает значение Y_g

$p_g = \text{Prob}(Y = Y_g \mid X)$

Обычно в модель включено пересечение, или свободный член, но это не обязательно. Величины P_1, P_2, … , P_G - это априорные вероятности групп.

Референтной (reference) называется первая по порядку группа в уравнениях. Выбор референтной группы произвольный, но осмысленный. Обычно это наибольшая группа или контрольная группа, с которой сравниваются все остальные группы.

$\{\beta_{ij}\}$ – это множество регрессионных коэффициентов (неизвестных), которые требуется оценить по имеющимся данным. Эти оценки обозначаются $\{b_{ij}\}$.

Оценки максимального правдоподобия параметров $\{\beta_{ij}\}$ получаются с помощью нахождения точки экстремума логарифма отношения правдоподобия. Формулы приведены в Приложении. Там же описаны основные статистики, применяемые для оценки результатов применения логистической регрессии.

5.4. Интерпретация регрессионных коэффициентов

Интерпретация полученных оценок не так проста, как для случая множественной регрессии.

Рассмотрим простой случай бинарной зависимой переменной Y и единственной независимой переменной X. Пусть Y имеет значения 0 и 1. Тогда уравнение логистической регрессии (5.6) имеет вид

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \qquad (5.7)$$

Соответственно, мы получим, что изменение логарифма шансов события (левая часть уравнения) при увеличении независимой переменной X на 1 как раз и составит β_1.

$$\beta_1 = \ln\left(\frac{\text{odds}(X+1)}{\text{odds}(X)}\right)$$

Если независимых переменных более одной, то такая интерпретация каждого коэффициента сохраняется, - имея в виду, что увеличивается на 1 только соответствующая переменная, а остальные переменные не меняются.

Бинарная переменная X

Если X имеет только два значения, 0 и 1, то последняя формула дает простую интерпретацию коэффициента β_1 как логарифма отношения шансов:

$$\beta_1 = \ln\left(\frac{\text{odds}(X=1)}{\text{odds}(X=0)}\right) \qquad (5.8)$$

Если зависимая переменная имеет более двух значений, то модель определяется большим, чем одно, числом уравнений. Для каждого из них в отдельности интерпретация коэффициентов такая же, как для случая бинарного отклика.

5.5. Применение метода логистической регрессии для анализа данных в статистических программах

Пример 15(1) – данные НРЭР Северо-Запада.

Данные НРЭР о ликвидаторах Северо-Запада включают информацию о их возрасте, времени и продолжительности участия в аварийно-восстановительных работах, месте постоянного проживания, полученной дозе внешнего облучения. Будет изучаться влияние этих факторов на статус ИНД_ИБС: нет ИБС(0) – есть ИБС(1) у ликвидаторов на момент исследования (2007 год).
На первом этапе выбраны непрерывные числовые переменные: доза - DOSE, год рождения - Y_BIRTH, день начала работ (в днях после аварии) - moment, продолжительность работ (дней) - srok_days. В качестве модели рассматривается сумма факторов без взаимодействий, в качестве наблюдаемых – все находившиеся под наблюдением к началу 2007 года ликвидаторы Северо-Запада (не снятые с наблюдения по причине смерти, переезда и т.д.), у которых указаны значения всех независимых переменных.

Программа NCSS

В результате проведения логистического анализа получены следующие результаты.

Model R-Squared 0.112

Раздел анализа отклика (Response Analysis Section)

ИНД_ИБС Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
Нет ИБС	2588	2575	0.500	0.143	64.915
есть ИБС	2082	2075	0.500	0.143	65.658
Total	4670	4650			65.246

В данном разделе описано исходное распределение зависимой переменной (Count), количество различных комбинаций независимых переменных в каждой категории зависимой переменной (Unique Rows), выбранные априорные вероятности для каждой категории (Prior), R^2 для регрессии зависимой переменной с предсказанной вероятностью принадлежности к категории (Act vs Pred R-Squared), % наблюдений, правильно классифицированных с помощью построенной логистической модели, в каждой категории и в целом (% Correctly Classified)

Раздел проверки значимости параметров (Parameter Significance Tests Section).

(По отношению к референтной группе - Reference Group: ИНД_ИБС = 0)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	247.528	10.687	23.163	**0.000**	10000+
B1: DOSE	0.006	0.004	1.552	0.121	1.007
B2: moment	-0.001	0.000	-3.914	**0.000**	0.999
B4: Y_BIRTH	-0.127	0.005	-23.150	**0.000**	0.881
B3: srok_days	0.000	0.000	-0.123	0.902	1.000

Список параметров (Parameter) включает все независимые непрерывные переменные в исходном виде. Заметим, что для дискретных переменных, если они участвуют в анализе, генерируются бинарные переменные по числу возможных значений, и каждая из созданных бинарных переменных включается в список параметров. В текущем анализе дискретных независимых переменных нет.

Регрессионные коэффициенты (Regression Coefficient) представляют собой оценки соответствующих коэффициентов регрессии. Стандартные ошибки (Standard Error) регрессионных коэффициентов адекватны для больших выборок. Z-значение теста Вальда (Wald Z-Value) для проверки гипотезы о равенстве 0 регрессионного коэффициента (Beta=0). Используется для больших выборок. Для малых выборок следует использовать критерий отклонений. Уровень значимости критерия Вальда (Wald Prob Level). Переменная статистически значима, если этот уровень меньше заданной величины (обычно 0.05).

В данном примере значимыми для описания переменной «ИНД_ИБС» являются переменные «год рождения» - Y_BIRTH и «день начала работ

(в днях после аварии)» - moment, а влияние переменных «доза» и «продолжительность работ» статистически незначимо.
Отношение шансов (Odds Ratio) – это оценка отношения шансов, связанная с данным регрессионным коэффициентом. Она полезна только для бинарных независимых переменных со значениями 0 и 1. Для других дискретных переменных (у которых количество значений более двух) генерируются соответствующее количество бинарных переменных. Формула для вычисления отношения шансов в этом случае

$$OR = e^b$$

Для непрерывных независимых переменных OR не имеет особого смысла.

Раздел «доверительные интервалы для параметров» (Parameter Confidence Limits Section) (Reference Group: ИНД_ИБС = 0 «нет ИБС»)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio Exp(B)
B0: Intercept	247.528	10.687	226.583	268.473	10000+
B1: DOSE	0.006	0.004	-0.002	0.015	1.007
B2: moment	-0.001	0.000	-0.001	0.000	0.999
B4: Y_BIRTH	-0.127	0.005	-0.138	-0.116	0.881
B3: srok_days	0.000	0.000	-0.001	0.000	1.000

Данный раздел дублирует информацию предыдущего раздела, кроме доверительного 95% интервала для параметра b или β (4-й и 5-й столбцы, Lower 95% Confidence Limit и Upper 95% Confidence Limit). Он построен с помощью статистики Вальда.

Раздел «оценка отношения шансов» (Odds Ratios Section) (Reference Group: ИНД_ИБС = 0 «нет ИБС»)

Parameter	Regression Coefficient (B or Beta)	Odds Ratio Exp(B)	Lower 95% Confidence Limit	Upper 95% Confidence Limit
B0: Intercept	247.528	10000+	10000+	10000+
B1: DOSE	0.006	1.007	0.998	1.015
B2: moment	-0.001	0.999	0.999	1.000
B4: Y_BIRTH	-0.127	0.881	0.872	0.890
B3: srok_days	0.000	1.000	0.999	1.000

В этом разделе оценки регрессионных коэффициентов и отношений шансов те же, что в предыдущих таблицах. Доверительные интервалы для отношения шансов получены с помощью статистики Вальда, из доверительных интервалов для регрессионных коэффициентов путем потенцирования.

Оценка регрессионной логистической модели (Estimated Logistic Regression Model) (Model For ИНД_ИБС = 1)

247.528 + 0.0065*DOSE – 0.0005*moment - 0.127*Y_BIRTH – 0.00003*srok_days

В этом разделе приведено уравнение логистической регрессии в виде регулярного текста, что позволяет ее использовать для преобразований. Все коэффициенты этого уравнения записаны в 1-м столбце предыдущих таблиц (Regression Coefficient (B or Beta)).

Таблица классификации (Classification Table)

	Estimated (оценка)		
Actual	**0**	**1**	**Total**
Нет ИБС	1680	908	2588
есть ИБС	715	1367	2082
Total	2395	2275	4670

Percent Correctly classified (правильно классифицировано) = 65.2%

В таблице представлены результаты классификации наблюдений на основании логистического регрессионного уравнения.

В целом по данному примеру можно сделать вывод, что четыре непрерывные независимые переменные, заданные для описания бинарного отклика – наличия ИБС – недостаточно хорошо его описывают. В качестве суммарных характеристик выступает, во-первых, качество классификации (65.2% правильной классификации) и R^2 модели (0.112).

Пример 15(2).

На втором этапе создания адекватной модели и ее анализа непрерывные независимые переменные преобразованы в дискретные: вместо дозы – дозовая группа (ДОЗ_ГРУП) с тремя градациями: (1) – от 0 до 5 сЗв, (2) – от 5.1 до 19.9 сЗв, (3) 20+ сЗв. Вместо года рождения – возрастная группа участия в работах (Age_group_input) с тремя градациями: (1) – от 18 до 29 лет, (2) – от 30 до 39 лет, (3) – 40+ лет. Вместо дня начала работ (в днях после аварии) – период начала работ (period) с четырьмя градациями: (1) 0-15 дни приезда, (2) – 16-350 дни приезда, (3) – 351-700 дни приезда, (4) – 700+ дни приезда. Вместо продолжительности работ в днях – группы по продолжительности работ (srok_group) с четырьмя градациями: (1) до 1 месяца, (2) 31-60 дней, (3) 61-365 дней, (4) более 1 года. В качестве модели рассматривается сумма факторов без взаимодействий, в качестве наблюдаемых – все находившиеся под наблюдением к началу 2007 года ликвидаторы Северо-Запада.

Для этой модели логистический анализ (**программа NCSS**) позволил получить следующие результаты.

Model R-Squared 0.75

Анализ отклика

ИНД_ИБС Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
Нет ИБС	2588	106	0.5000	0.1216	62.519
есть ИБС	2082	105	0.5000	0.1216	67.003
Total	4670	211			64.518

Таблица анализа отклика несущественно отличается от предыдущего примера: доля правильной классификации снизилась на 0.7%, R^2 модели увеличилась с 0.112 до 0.75. Но в следующем разделе отражена более детальная информация о влиянии отдельных уровней независимых переменных на бинарный отклик.

Раздел проверки значимости параметров (Parameter Significance Tests Section) - по отношению к референтной группе (Reference Group: ИНД_ИБС = 0)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-0.9385	0.1620	-5.793	**0.0000**	0.3912
B1: Age_group_input=2 («30-39 лет»)	1.1909	0.0863	13.800	**0.0000**	3.2901
B2: Age_group_input=3 («40+ лет»)	2.1821	0.1072	20.357	**0.0000**	8.8646
B3: period=2 («16-350 дни приезда»)	0.1944	0.1324	1.469	0.1418	1.2146
B4: period=3 («351-700 дни приезда»)	0.0853	0.1512	0.564	0.5726	1.0890
B5: period=4 («700+ дни приезда»)	-0.3720	0.1743	-2.135	**0.0328**	0.6893
B6:srok_group=2 ("31-60 дней")	-0.4736	0.1051	-4.506	**0.0000**	0.6227
B7:srok_group=3 ("61-365 дней")	-0.4521	0.0900	-5.024	**0.0000**	0.6363
B8:srok_group=4 ("БОЛЕЕ ГОДА")	-0.3496	0.2958	-1.182	0.2373	0.7050
B9: ДОЗ_ГРУП=2 («5.1- 19.9 сЗв»)	0.0510	0.0981	0.520	0.6027	1.0524
B10: ДОЗ_ГРУП=3 («20+ сЗв»)	0.2693	0.1110	2.426	**0.0153**	1.3090

Из этой таблицы уже можно сделать вывод о значимости отдельных параметров и их градаций. Красным цветом выделены статистически значимые на уровне 0.05 градации параметров. По отношению к референтным градациям независимых переменных (с наименьшим номером) значимыми оказались все градации возрастных групп (B1 и

B2), с положительными коэффициентами; последний уровень переменной «период» (B5) – с отрицательным коэффициентом; 2-й и 3-й уровни переменной «срок- группа» (B6 и B7) – также с отрицательными коэффициентами. Для дозовых групп значимым является последний уровень (B10), с положительным коэффициентом. При этом связанные со значимыми положительными влияниями риски (отношение шансов) превышают 1, т.е. увеличивают вероятность заболевания (для возрастных групп и наибольшей дозы внешнего облучения), а для 4-го периода начала работ и средних уровней продолжительности работ отношение шансов менее 1, эти факторы снижают вероятность наличия ИБС.

Обратив внимание на численное значение полученных рисков, на следующем этапе для дискретных независимых переменных «дозовая группа», «период начала работ» и «продолжительность работ» осуществим уменьшение количества уровней - статистически незначимые уровни будут объединены с учетом полученных результатов. Вместо ДОЗ_ГРУП с тремя градациями сформируем переменную dose_20 с 2-мя градациями, объединив первые две группы. Для dose_20 получим уровни: (1) – от 0 до 19.9 сЗв, (2) 20+ сЗв. Это сделано с учетом того, что отношение шансов для дозовой группы «5.1-19.9 сЗв» близко к 1 по сравнению с референтной группой «0 - 5 сЗв». Вместо (period) с четырьмя градациями сформируем period_4 с 2-мя градациями, выделив 4-й период приезда: 1 уровень 0-700 дни приезда (объединены 1, 2 и 3 периоды), 2 уровень – 700+ дни приезда (4 период). Вместо srok_group с четырьмя градациями сформирована переменная srok_1_month с 2-мя градациями: (1) до 1 месяца, (2) 32 дня и более (объединены 2, 3 и 4 уровни переменной srok_group, поскольку отношения шансов для них близки по величине). Возрастная группа участия в работах (Age_group_input) остается без изменений: все ее уровни значимы, отношения шансов отличаются друг от друга значительно.

В качестве модели, как и ранее, рассматривается сумма факторов без взаимодействий, множество наблюдаемых не меняется. Получим следующие результаты.

Model R-Squared 0.916

Анализ отклика

ИНД_ИБС Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
Нет ИБС	3886	21	0.5000	0.1601	74.112
есть ИБС	3244	20	0.5000	0.1601	58.600
Total	7130	41			67.055

Раздел проверки значимости параметров Parameter Significance Tests Section (Reference Group: ИНД_ИБС = 0 (Нет ИБС))

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-0.9443	0.0691	-13.657	**0.0000**	0.3889
B1: Age_group_input=2 («30-39 лет»)	1.2297	0.0674	18.250	**0.0000**	3.4203
B2: Age_group_input=3 («40+ лет»)	2.4667	0.0827	29.844	**0.0000**	11.7838
B3: (dose_20_=2) «20+ сЗв»	0.2250	0.0649	3.470	**0.0005**	1.2524
B4: (period_4=2) «700+ дни приезда»	-0.5583	0.0796	-7.016	**0.0000**	0.5722
B5: (srok_1_month=2) « >1 мес.»	-0.3265	0.0610	-5.354	**0.0000**	0.7215

Эти результаты лучше полученных ранее: все параметры значимы, процент правильной классификации повысился до 67%, R2 модели увеличилась до 0.916. При этом качество деления на группы с ИБС и без ИБС остается неудовлетворительным, особенно это касается группы с заболеванием. Далее возможны следующие шаги: (а) – усложнять модель, добавляя взаимосвязи независимых переменных, и (б) добавлять еще какие-либо независимые переменные. Влияние на переменную-отклик «наличие ИБС» может оказать место проживания наблюдаемых, поэтому добавим в анализ независимую переменную «регион» (REGION) с 5 градациями: (1) Калининградская область, (2) Ленинградская область, (3) Санкт-Петербург, (4) Новгородская область и (5) Псковская область. Это те субъекты РФ, которые относятся к Северо-Западному подразделению РГМДР.

Model R-Squared 0.883

Анализ отклика

ИНД_ИБС Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
0	3883	90	0.5000	0.2081	67.963
1	3244	81	0.5000	0.2081	73.089
Total	7127	171			**70.296**

Раздел проверки значимости параметров
(Reference Group: ИНД_ИБС = 0)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-2.0083	0.1326	-15.146	**0.0000**	0.1342
B1: Age_group_input=2 («30-39 лет»)	1.2847	0.0690	18.618	**0.0000**	3.6137
B2: Age_group_input=3 («40+ лет»)	2.4718	0.0851	29.050	**0.0000**	11.8436
B3: (dose_20_=2) «20+ сЗв»	0.2110	0.0670	3.148	**0.0016**	1.2350
B4: (period_4=2) «700+ дни приезда»	-0.6119	0.0819	-7.474	**0.0000**	0.5423
B5: (REGION=2) Ленобласть	1.2324	0.1233	9.999	**0.0000**	3.4295
B6: (REGION=3) СПб	1.1657	0.1161	10.042	**0.0000**	3.2082
B7: (REGION=4) Новгородская обл.	0.1691	0.1330	1.272	0.2035	1.1843
B8: (REGION=5) Псковская обл.	-0.3864	0.1653	-2.338	**0.0194**	0.6795
B9: (srok_1_month=2) « >1 мес.»	-0.1121	0.0636	-1.762	0.0780	0.8940

В результате использования дополнительной переменной качество классификации существенно улучшилось (до 70.3%). Для того, чтобы оценить вклад каждой из независимых переменных в описании отклика (ИНД_ИБС) следует обратиться к таблицам «Анализ отклонений» и «Логарифм правдоподобия и R-квадрат».

Таблицы в разделах «Анализ отклонений» (Analysis of Deviance Section) и «Логарифм правдоподобия и R-квадрат» (Log Likelihood & R-Squared Section) вычисляются в том случае, когда модель уже определена, и в качестве способа выбора наилучшего подмножества задана опция “None”.

Анализ отклонений (Analysis of Deviance Section)

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi Square)	Prob Level
All	9	9822.751	1583.772	**0.0000**
Age_group_input	2	9243.121	1004.143	**0.0000**
dose_20_	1	8248.897	9.919	**0.0016**
period_4	1	8296.439	57.461	**0.0000**
REGION	4	8617.537	378.559	**0.0000**
srok_1_month	1	8242.080	3.102	0.0782
None(Model)	9	8238.978		

Исключенный член (Term Omitted) – тот член, который проверяется в данной строке. Этот тест получается при сравнении статистики отклонения, когда данный член исключен, с отклонением в полной модели. В строке All рассматривается модель, включающая только свободный член.

DF – число степеней свободы для статистики $\chi 2$, приведенной в этой строке.

Отклонение (Deviance) – это (-2)*логарифм правдоподобия, достигаемый в текущей модели.

Увеличение в отклонении модели (Increase From Model Deviance (Chi Square)) – разница отклонений модели данной строки и полной модели. Распределена эта величина для больших выборок приблизительно как $\chi 2$.

Уровень значимости (Prob Level) – для критерия $\chi 2$. Это вероятность того, что значение функции $\chi 2$ с числом степеней свободы DF равно или больше этой величины. Если уровень значимости меньше 0.05, данный член следует считать статистически значимым для модели. В данном случае незначимой оказалась продолжительность работ.

Логарифм правдоподобия и R-квадрат (Log Likelihood & R-Squared Section)

Term Omitted	DF	Log Likelihood	R-Squared of Remaining Term(s)	Reduction From Model R-Squared	Reduction From Saturated R-Squared
All	1	-4911.375	0.0000		
Age_group_input	2	-4621.560	0.3230	0.5596	0.6770
dose_20_	1	-4124.448	0.8770	0.0055	0.1230
period_4	1	-4148.220	0.8505	0.0320	0.1495
REGION	4	-4308.769	0.6716	0.2110	0.3284
srok_1_month	1	-4121.040	0.8808	0.0017	0.1192
None(Model)	9	-4119.489	0.8826	0.0000	0.1174
None(Saturated)	171	-4014.125	1.0000		0.0000

Исключенный член (Term Omitted) – так же, как в предыдущем разделе. Только в строке “None(Saturated)” приведены результаты для насыщенной модели.

DF – число степеней свободы.

Логарифм правдоподобия (Log Likelihood) – для модели, проверяемой в данной строке. Это логарифм правдоподобия логистической регрессии без того члена, который приведен в списке.

R-квадрат для оставшихся членов (R-Squared of Remaining Term(s)) – для модели, проверяемой в текущей строке. Эта величина аналогична R2 для

множественной регрессии, но не одно и то же: в случае, когда величина R-квадрат составляет 1.0, это показывает, что логистическая регрессионная модель достигла того же правдоподобия, что и насыщенная модель. Это не означает, что данные будут описаны моделью точно, а лишь максимально возможным образом.

Уменьшение от R-квадрат модели (Reduction From Model R-Squared) – уменьшение R-квадрат модели из-за исключения текущего члена.

Уменьшение от R-квадрат насыщенной модели (Reduction From Saturated R-Squared) – вычисляется с помощью R-квадрат, достигаемого в насыщенной модели. Показывает, насколько существенно влияет удаление этого члена на наилучшие из возможных значения R-квадрат.

Эти таблицы показывают, что наиболее существенным параметром для предсказания наличия ИБС среди ликвидаторов Северо-Запада является возрастная группа, а следующий по важности фактор – место проживания. Все остальные значимые показатели существенно уступают первым двум по важности при описании зависимой переменной (отклика) – на порядок (4 период) и даже на два порядка (доза выше 20 сЗв и продолжительность работ менее месяца).

Для того, чтобы получить максимально возможный результат классификации, имеет смысл наиболее полно учесть информацию о возрасте, как о наиболее важном факторе. Рассмотрим в качестве переменной, задающей возраст, (а) непрерывную переменную «год рождения» и (б) дискретную переменную «возраст участия» с 6 уровнями: (1) 18-29 лет, (2) 30-34 года, (3) 35-39 лет, (4) 40-44 года, (5) 45-49 лет, (6) 50 и более лет. При использовании непрерывной переменной информация учитывается наиболее полно, однако интерпретация полученных результатов логистической регрессии менее понятна, в отличие от дискретной переменной, для которой вычисляются отношения шансов по каждому уровню.

Кроме того, можно исключить из модели незначимую теперь переменную, задающую продолжительность работ, а в переменной «регион» объединить 1 и 4 уровни (Калининградская и Новгородская области), поскольку они не отличаются по отношению к переменной-отклику (ИНД_ИБС). Окончательно получим следующие два варианта.

(а) **Model** Y_BIRTH + period_4 + dose_20_ + region_1_2_3_5

Model R-Squared 0.7997

Анализ отклика

ИНД_ИБС Categories	**Count**	**Unique Rows**	**Prior**	**Act vs Pred R-Squared**	**% Correctly Classified**
0	3883	336	0.5000	0.2340	71.465
1	3244	319	0.5000	0.2340	70.037
Total	7127	655			**70.815**

Раздел проверки значимости параметров (Reference Group: ИНД_ИБС = 0)

Parameter	**Regression Coefficient (B or Beta)**	**Standard Error**	**Wald Z-Value (Beta=0)**	**Wald Prob Level**	**Odds Ratio Exp(B)**
B0: Intercept	251.3507	8.1569	30.814	**0.0000**	10000+
B1: **(dose_20_=2) «20+ сЗв»**	0.2144	0.0667	3.213	**0.0013**	1.2392
B2: **(period_4=2) «700+ дни приезда»**	-0.5537	0.0816	-6.789	**0.0000**	0.5748
B3: (region_1_2_3_5=2) **Ленобласть**	1.1158	0.0845	13.210	**0.0000**	3.0519
B4: (region_1_2_3_5=3) **СПб**	0.9990	0.0726	13.757	**0.0000**	2.7155
B5: (region_1_2_3_5=5) **Псковская обл.**	-0.5607	0.1390	-4.033	**0.0001**	0.5708
B6: Y_BIRTH	-0.1292	0.0042	-30.908	**0.0000**	0.8788

(б) **Model** period_4 + dose_20_ + region_1_2_3_5 + Age_input_gr_6

Model R-Squared 0.914

Анализ отклика

ИНД_ИБС Categories	**Count**	**Unique Rows**	**Prior**	**Act vs Pred R-Squared**	**% Correctly Classified**
0	3883	67	0.5000	0.2217	67.654
1	3244	64	0.5000	0.2217	73.890
Total	7127	131			**70.492**

Раздел проверки значимости параметров
(Reference Group: ИНД_ИБС = 0)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-1.9490	0.0808	-24.115	**0.0000**	0.1424
B1: (Age_input_gr_6=2) **30-34 года**	0.9928	0.0795	12.485	**0.0000**	2.6989
B2: (Age_input_gr_6=3) **35-39 лет**	1.4909	0.0754	19.781	**0.0000**	4.4409
B3: (Age_input_gr_6=4) **40-44 года**	2.0389	0.0981	20.781	**0.0000**	7.6821
B4: (Age_input_gr_6=5) **45-49 лет**	2.8415	0.1292	21.996	**0.0000**	17.1412
B5: (Age_input_gr_6=6) **50 и более лет**	3.7886	0.2471	15.332	**0.0000**	44.1934
B6: (dose_20_=2) **«20+ сЗв»**	0.2229	0.0664	3.357	**0.0008**	1.2497
B7: (period_4=2) **«700+ дни приезда»**	-0.5488	0.0824	-6.663	**0.0000**	0.5777
B8: (region_1_2_3_5=2) **Ленобласть**	1.1101	0.0840	13.214	**0.0000**	3.0348
B9: (region_1_2_3_5=3) **СПб**	0.9961	0.0726	13.725	**0.0000**	2.7077
B10: (region_1_2_3_5=5) **Псковская обл.**	-0.5283	0.1390	-3.801	**0.0001**	0.5896

Для того, чтобы показать способ использования полученных результатов и сравнить логистические регрессионные оценки в моделях (а) и (б), оценим с их помощью вероятности наличия ИБС и шансы иметь этот диагноз у нескольких ликвидаторов.

№	Возраст участия	Год рождения	Год участия	Доза	Место жительства
1	37	1949	1986	20.0	Калинингр.
2	38	1948	1986	15.6	Лен.обл.
3	25	1961	1986	17.6	СПб
4	49	1937	1986	18.0	СПб

Модель (б)

Parameter	Regression Coefficient (B or Beta)	№ 1		№2		№3		№4	
		Ind	ind*B	Ind	ind*B	Ind	ind*B	Ind	ind*B
B0: Intercept	-1.949	1	-1.949	1	-1.949	1	-1.949	1	-1.949
B1: **Возраст участия 30-34 года**	0.993	0	0	0	0	0	0	0	0
B2: **Возраст участия 35-39 лет**	1.491	1	1.491	1	1.491	0	0	0	0
B3: **Возраст участия 40-44 года**	2.039	0	0	0	0	0	0	0	0
B4: **Возраст участия 45-49 лет**	2.842	0	0	0	0	0	0	1	2.842
B5: **Возраст участия 50 и более лет**	3.789	0	0	0	0	0	0	0	0
B6: **Доза «20+ сЗв»**	0.223	1	0.223	0	0	0	0	0	0
B7: **Время участия «700+ дни приезда»**	-0.549	0	0	0	0	0	0	0	0
B8: **Место жительства Ленобласть**	1.110	0	0	1	1.110	0	0	0	0
B9: **Место жительства СПб**	0.996	0	0	0	0	1	0.996	1	0.996
B10: **Место жительства Псковская обл.**	-0.528	0	0	0	0	0	0	0	0
L= Ln(отношения шансов) = сумма по столбцу			-0.235		0.652		-0.953		1.889
P(1) = Вероятность (ИНД_ИБС=1) = 1/(1+exp(-L))			0.441		0.657		0.278		0.869
P(1)/P(0) = Шансы (Инд_ИБС=1) / (Инд_ИБС=0) = exp(L)			0.790		1.919		0.386		6.610

В столбце ind содержатся 0 или 1 в зависимости от того, присутствует ли данный уровень в текущем наблюдении. Например, для наблюдения №1 возраст участия (37 лет) соответствует параметру B2, поэтому в строке B2 стоит «1», а в строках B1, B3, B4 и B5 стоят «0». Также «1» стоит в строке B6: Доза «20+ сЗв». Остальные строки содержат нули, поскольку для данного наблюдения указанные значения переменных «Время участия» и «Место жительства» не имеют места.

В строке B0: Intercept для всех наблюдений ind = 1.

Модель (a)

Parameter	Regression Coefficient (B or Beta)	№ 1		№2		№3		№4	
		Ind	ind*B	Ind	ind*B	Ind	ind*B	Ind	ind*B
B0: Intercept	251.4	1	251.4	1	251.4	1	251.4	1	251.4
B1: **доза «20+ сЗв»**	0.214	1	0.214	0	0	0	0	0	0
B2: **Время участия «700+ дни приезда»**	-0.554	0	0	0	0	0	0	0	0
B3: **Место жительства Ленобласть**	1.116	0	0	1	1.116	0	0	0	0
B4: **Место жительства СПб**	0.999	0	0	0	0	1	0.999	1	0.999
B5: **Место жительства Псковская обл.**	-0.561	0	0	0	0	0	0	0	0
B6: **Год рождения**	-0.129	1949	-251.8	1948	-251.7	1961	-253.4	1937	-250.3
L= Ln(отношения шансов) = сумма по столбцу			-0.246		0.785		-1.012		2.089
P(1) = Вероятность (ИНД_ИБС=1) = 1/(1+exp(-L))			0.439		0.687		0.267		0.890
P(1)/P(0) = Шансы (Инд_ИБС=1)/(Инд_ИБС=0)			0.782		2.192		0.364		8.079

Отличие этой таблицы от предыдущей состоит в том, что в строке B6 вносятся реальные данные о годе рождения, а не индексы, как в остальных строках. При этом заметим, что выводы относительно каждого из наблюдений в рамках двух моделей совпадают: наблюдаемые №1 и №3 имеют шансы ИБС менее 1, т.е. скорее не имеют этого диагноза, в отличие от наблюдаемых №2 и №4. Наибольшие шансы иметь ИБС у наблюдаемого №4. Это относится к обеим моделям и связано, прежде всего, с его возрастом. Дополнительный фактор, увеличивающий его шансы на заболевание – место жительства, СПб. Наблюдаемые №1 и №2 относятся к одной возрастной группе, у первого имеется фактор повышенной дозы, а у второго – фактор места жительства в Ленобласти. И этот фактор оказывается более значимым – наблюдаемый №1 относится к лицам с шансами не иметь ИБС, а наблюдаемый №2 – к группе, в которой шансы иметь это заболевание превышают 1. Наблюдаемый №3 – самый молодой, поэтому, несмотря на место жительства в СПб, его шансы наличия ИБС минимальны.

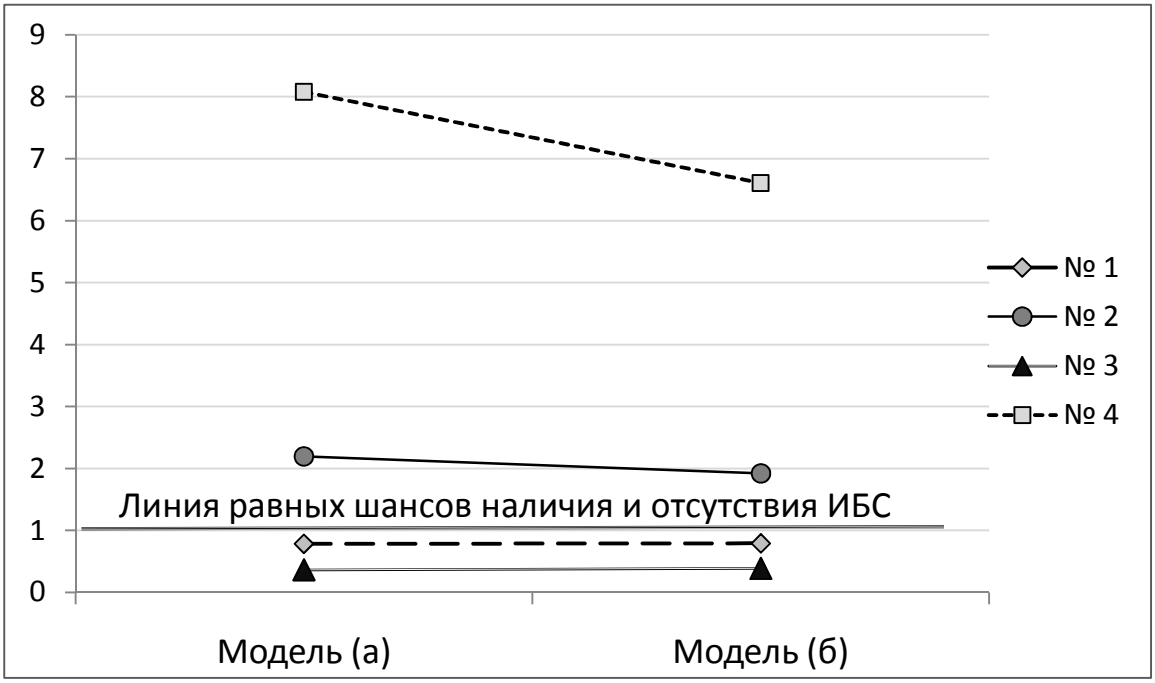


Рис. П15-1. Шансы наличия ИБС для четырех наблюдаемых в соответствии с моделями (а) и (б)

5.6. Выбор подмножества независимых переменных

Задача выбора подмножества состоит в отборе из всего множества независимых переменных небольшого их количества, обеспечивающего, тем не менее, хорошее предсказание зависимой переменной. Обычно используется техника пошагового добавления и исключения переменных, поскольку перебор всех возможных подмножеств для логистической регрессии требует очень больших вычислений. Используются два алгоритма: пошаговый отбор и пошаговый отбор с переключением.

В случаях, когда классов более двух, используется общее отношение правдоподобия для оценки переменных, поэтому включаются все независимые переменные, которые важны для описания хотя бы одного класса.

Иерархические модели

Обычно взаимодействие включается в модель только после того, как все его составляющие вошли в модель: A*B*C будет включено только после включения в модель A, B, C, A*B, A*C, B*C. Такие модели называются иерархическими.

Пошаговый отбор (Forward Selection)

1. На первом шаге в модели нет ни одного члена.

2. В модель включается тот член, для которого достигается максимум логарифма правдоподобия.
3. Процесс продолжается, пока не будет достигнуто критическое значение критерия или будет включено максимально возможное количество членов.

Этот метод не обязательно дает наилучший выбор модели и применяется, когда число наблюдений и переменных очень велико.

Пошаговый отбор с переключением (Forward Selection with Switching)

Алгоритм работает аналогично предыдущему, но после включения каждого члена в модель происходит процесс тестирования связок из включенных и не включенных в модель членов – насколько они увеличивают величину критерия (логарифма правдоподобия). Если связка найдена, она фиксируется, и проверяется результат присоединения к ней еще одного члена (величина критерия).

После окончания процесса увеличивается на единицу количество возможных членов в модели, и процесс повторяется. Алгоритм завершается, если достигнут максимум возможного числа членов или включены все переменные.

ГЛАВА 6. ЛОГЛИНЕЙНАЯ МОДЕЛЬ (LLM)

Логлинейные модели позволяют изучать связи между двумя и более дискретными переменными. На него часто ссылаются как на многомерный анализ частот, поскольку этот метод является расширением аналогичного теста χ^2 для проверки независимости таблиц сопряженности с двумя входами (глава 2, проверка гипотезы о независимости $H_н$).

Этот метод часто используется для анализа обзоров, анкет, исследований, где присутствуют сложные внутренние взаимосвязи между откликами (переменными). Обычно исследуются только двумерные таблицы откликов, что исключает из рассмотрения трехмерные и большей размерности связи. Использование LLM для анализа данных такого типа аналогично применению множественной регрессии по сравнению с использованием простых корреляций для непрерывных данных, с той разницей, что использование LLM не предполагает выделение одной какой-либо переменной в качестве зависимой.

6.1. Ограничения и предположения

Использование LLM предполагает очень малое количество ограничений. Метод может применяться практически во всех случаях, когда переменные дискретны (или могут быть дискретизированы).

LLM основан на трех основных предположениях.

1. Наблюдения независимы. Практически это означает, что все наблюдения соответствуют разным субъектам и получены случайным образом из популяции, без специфических групп субъектов.
2. Все наблюдения распределены одинаково. Это означает, что они получены одним и тем же способом.
3. Количество наблюдений достаточно велико. Это связано с тем, что в LLM используется аппроксимация, применимая для больших выборок. Алгоритм LLM начинается с логарифмирования всех частот в ячейках таблицы сопряженности, поэтому нулевые частоты недопустимы.

Ограничения LLM менее строги, чем при использовании обычного теста χ^2 для проверки независимости, поэтому, если применим этот тест, то можно использовать и LLM.

6.2. Основные принципы

Применение LLM предполагает осуществление двух этапов. Важно помнить о целях, которые ставятся при выполнении каждого из них.

(а) Выбор соответствующей модели

Первым этапом является выбор модели, соответствующей данным. Есть несколько методов выбора. Одним из наиболее популярных является пошаговый метод, при котором сложные составные элементы модели постепенно исключаются до тех пор, пока в модели не останутся только значимые элементы. Такой поиск подходящей модели применим только к иерархическим моделям. Иерархическими называются модели, которые наряду с каждым членом включают и все его компоненты. Например, если модель включает взаимодействие AB, то она должна также включать члены A и B.

В процессе выбора модели следует исследовать остатки для определения качества описания данных с помощью текущей модели.

(б) Интерпретация выбранной модели

На этом этапе требуется определить, что означает выбор модели для объяснения связей в данных.

6.3. Обозначения

Рассмотрим таблицу с двумя входами, у которой переменная строк **A** имеет I категорий (уровней) i=1,…, I, а переменная столбцов **B** имеет J категорий j=1,…, J. Точная мультипликативная модель, определяющая частоты в ячейках f_{ij}, записывается как

$$m_{ij} = N\alpha_i\beta_j\gamma_{ij} \quad (6.1),$$

где $m_{ij} = E(f_{ij})$ – ожидаемая частота в строке i и столбце j. Если m_{ij} оцениваются с использованием метода максимального правдоподобия, результат обозначается $\tilde{m}_{ij}$. Также заметим, что $N = \Sigma_{i,j}f_{ij}$.

В этой таблице интерес представляет только один момент: являются ли независимыми **A** и **B**. Это можно проверить с помощью соответствующего теста χ^2. В модели (6.1) независимость будет установлена, если все γ_{ij} будут равны 1.

После логарифмирования выражения (6.1) получим линейный относительно неизвестных коэффициентов вид модели

$$\ln(m_{ij}) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad (6.2)$$

Слагаемые λ. называются эффектами. Верхние индексы обозначают переменные, нижние индексы – категории этих переменных. Порядок эффекта равен числу переменных в верхнем индексе.

Поскольку полученная формула аддитивна, она называется логлинейной моделью. Из-за логарифмирования в данной модели присутствует ограничение: ни один из m_{ij} не равен 0.

Заметим, что в данной модели общее количество коэффициентов λ. составляет 1 + I + J + I*J, что превышает количество частот в ячейках (которое составляет I*J). Если число параметров модели превышает или равно количеству ячеек, такая модель называется насыщенной (saturated). Насыщенная модель точно воспроизводит наблюдаемые частоты.

Проверяя, равны ли определенные параметры λ нулю, мы проверяем различные связи между переменными. Например, проверяя, являются ли все коэффициенты $\{\lambda_{ij}^{AB}\}_{i,\ j}$ нулевыми, мы проверяем независимость переменных **A** и **B**. Проверяя, равны ли 0 все $\{\lambda_i^A\}$, мы проверяем, равны ли между собой все вероятности появления категорий **A**. Таким образом, эта модель позволяет ответить на многие вопросы относительно факторов **A** и **B**.

6.4. Качество подгонки

В случае, когда рассматривается несколько вариантов моделей, следует оценить качество каждой из них. Качество модели определяется качеством подгонки данных и проверяется с использованием одной из двух статистик χ^2:

- статистики Пирсона χ^2

$$\chi^2 = 2 \sum_{i,j,k} \frac{\left(f_{ijk} - \tilde{m}_{ijk}\right)^2}{\tilde{m}_{ijk}} \qquad (6.3)$$

- и статистики максимального правдоподобия

$$G^2 = 2\Sigma_{i,j,k}\, f_{i\,j\,k} \ln\left(f_{i\,j\,k} / \tilde{m}_{i\,j\,k}\right) \qquad (6.4)$$

Обе эти статистики распределены как χ^2, когда N велико и ни одна из частот $\tilde{m}_{ijk}$ не является малой. Обе статистики имеют (n – p) степени свободы, где n – количество ячеек таблицы, p – количество параметров в модели, для которой вычислены $\tilde{m}_{i\,j\,k}$.

С помощью этих статистик проверяется следующее утверждение: отличаются ли статистически значимо от 0 те члены насыщенной модели, которые не включены в текущую модель?

Например, пусть текущая иерархическая модель имеет вид {AB, BC}. Тогда раскрытая версия этой модели представляет собой A+B+C+AB+BC. Заметим, что члены AC и ABC насыщенной модели исключены. Вычислим статистики χ^2 и G^2, используя $\tilde{m}_{ijk}$, вычисленные по этой модели. Значения этих статистик позволяют определить, действительно ли эффекты AC и ABC нулевые. Другими словами, эти статистики проверяют, не пропущены ли в текущей модели какие-либо важные эффекты.

В отличие от статистики χ^2 Пирсона, отношение правдоподобия G^2 имеет одно важное свойство – оно является аддитивным для частичных связанных моделей. Поясним это утверждение на примере. Пусть оценивается модель (1) {AB, AC, BC}, соответствующее значение $G^2(1)$ получено равным 17.8 с 8 степенями свободы. При проверке модели (2) {A, B, C} получено значение $G^2(2)$, равное 69.9 с 24 степенями свободы. Если мы рассмотрим расширенные варианты записей обеих моделей, то обнаружим, что члены AB, AC и BC входят в первую модель и не входят во вторую. Далее, заметим, что вторая модель связана с первой моделью (целиком вложена). Разность $G^2(2) - G^2(1) = 52.1$. Это тоже значение статистики χ^2 с $24 – 8 = 16$ степенями свободы. Она позволяет проверить, значимы ли эффекты AB, AC, BC.

Это свойство аддитивности чрезвычайно важно. Оно позволяет проверять значимость отдельных членов модели. Предположим, при проверке модели {AB, BC, AC} мы обнаружили, что значение теста качества подгонки незначимо. Это означает, что эта модель, в полной записи имеющая вид A + B + C + AB + BC + AC, адекватно описывает данные. Следующий возникающий вопрос: все ли из 6 членов в этой модели необходимы? Для проверки значимости BC следует проверить модель A + B + C + AB + AC и вычислить разность между получившимися значениями статистики G^2. Эта разность и будет тестом для определения значимости BC.

Предостережение: разница между двумя значениями G^2 распределена как χ^2 только в том случае, когда более полная модель описывает данные адекватно. Другими словами – если значение статистики G^2 для большей модели незначимо.

Благодаря свойству аддитивности применение статистики G^2 широко распространено в LLM. Поскольку статистика χ^2 Пирсона не обладает этим свойством, то возникает вопрос – зачем ее вообще

вычислять? По двум причинам: во-первых, ряд исследований показывает, что статистика Пирсона дает более точную оценку качества подгонки. Во-вторых, поскольку обе статистики являются асимптотическими, при не очень больших размерах выборки разумно провести вычисления двумя методами для уверенности в полученном выводе.

6.5. Техника выбора модели в программах STATISTICA и NCSS

При работе с LLM одной из наиболее важных является задача выбора модели среди большого числа возможных. Количество членов в насыщенной модели удваивается при добавлении нового фактора. Например, для четырехфакторного исследования насыщенная модель содержит 16 членов, а для пятифакторного – 32 члена. Соответственно, количество различных иерархических моделей для четырех факторов составляет более 100, для пяти факторов – более 1000. Поскольку перед исследователем прежде всего стоит задача выбора модели, адекватно описывающей данные и содержащей наименьшее возможное количество членов, требуется метод для ограничения перебора рассматриваемых моделей.

Программа STATISTICA: модуль *Логлинейный анализ* содержит команду автоматической подгонки модели с целью облегчения поиска "хорошей модели" по имеющимся данным. Общая логика этого алгоритма следующая. Сначала STATISTICA подгоняет модель, в которой нет связей между факторами. Если она отвергается (т.е. соответствующая статистика *хи*-квадрат имеет значимую величину), то подгоняется модель со всеми возможными взаимодействиями двух факторов. Если эта модель тоже не принимается, то STATISTICA проверит модель со всеми трехфакторными взаимодействиями и т.д. Теперь предположим, что в ходе этого процесса установлено, что модель со всеми двухфакторными взаимодействиями подходит для имеющихся данных. Тогда STATISTICA начнет устранять двухфакторные взаимодействия, которые не являются статистически значимыми. Результирующей моделью станет такая модель, которая включает наименьшее необходимое для согласия число взаимодействующих факторов.

В программе NCSS предусмотрено использование нескольких методов выбора модели, приведенных ниже. Конечная модель будет результатом их применения для предлагаемых данных.

Методы выбора модели - стандартизованные оценки параметров

Этот метод следующим способом просматривает модели. Прежде всего вычисляются стандартизованные оценки всех λ для насыщенной модели. Далее составляется список наибольших эффектов (превышающих некоторый порог, например 2.0 или 3.0). Наконец, выбирается иерархическая модель, включающая наименьшее возможное число членов, каждый из которых входит в список со значимыми эффектами. Эта модель проверяется на адекватность данным с использованием теста χ^2. Если значение статистики незначимо, эта модель принимается. В противном случае в текущую модель добавляются эффекты из списка, пока не получится адекватная модель.

Методы выбора модели - проверка маргинальных и частных ассоциаций

При применении этого метода вычисляются два теста для каждого члена (вплоть до членов четвертого порядка). Предполагается, что членами большего порядка можно пренебречь. Эти тесты измеряют частные и маргинальные ассоциации. Частная ассоциация рассматривает значимость одного члена после рассмотрения всех остальных членов того же порядка. Маргинальная ассоциация проверяет значимость одного члена при исключении влияния других факторов в модели.

Тест частной ассоциации строится следующим образом. Рассмотрим две модели: первая содержит все члены того же порядка, что и оцениваемый член. Вторая – содержит все члены первой, кроме оцениваемого. Вычисляем разность $G^2(2) - G^2(1)$ и разность степеней свободы для $G^2(2)$ и $G^2(1)$.

Например, пусть требуется проверить, что частная ассоциация факторов **A** и **B** в четырехфакторной таблице – нулевая. Вычисляем G^2 для моделей (1) {AB, AC, AD, BC, BD, CD} и (2) {AC, AD, BC, BD, CD}. Разница статистик позволяет проверить значимость частной ассоциации.

Тест маргинальной ассоциации строится с помощью сворачивания таблиц, пока интересующий нас член не окажется взаимодействием наивысшего порядка и в модели не останется ни одного члена того же порядка. Затем этот член исключается и оценивается модель более низкого порядка. Значение G^2 оценивает маргинальную ассоциацию между факторами в оцениваемом члене.

Например, для того, чтобы проверить, что маргинальная ассоциация между **A** и **B** в четырехфакторной модели равна 0, прежде всего сворачиваем исходную таблицу к двухвходовой таблице, задаваемой факторами **A** и **B.** Далее для этой свернутой таблицы оцениваем модель {A, B} (без взаимодействия AB) с помощью статистики G^2. Это значение статистики и является маргинальной ассоциацией A и B.

Используя результаты этих двух тестов, можно получить хороший индикатор того, является ли рассматриваемый член значимым или нет. Как и раньше, для получения конечной модели формируется список всех значимых

членов. Далее следует составить минимальную иерархическую модель, которая включает эти члены.

Методы выбора модели - одновременные порядковые тесты

Эта программа дает список одновременных тестов для всех членов заданного порядка и всех членов заданного порядка и выше. Эти тесты позволяют сразу же уменьшить количество рассматриваемых моделей. Например, если тест для моделей второго порядка и выше значим, а для моделей третьего порядка и выше – незначим, следовательно, следует ограничиться рассмотрением моделей второго и менее порядков. Это сокращает поиск оптимальной модели.

Методы выбора модели - пошаговая процедура отбора

Это наиболее популярный метод выбора модели. Он по умолчанию используется в программе. Процедура начинается с определенной модели (часто с насыщенной модели, поскольку она описывает данные заведомо хорошо) и ищет модели с членами меньших порядков, которые также хорошо описывают данные. В программе используется техника обратного исключения, поскольку она работает лучше, чем техника прямого включения.

Для работы процедуры прежде всего следует задать уровень значимости (α) для того, чтобы тест качества подгонки сообщал о значимости модели (модель не описывает данные удовлетворительно). Далее исключается каждый из членов наивысшего порядка в иерархической модели и рассматривается расширенная модель, отличающаяся только этим членом. Тогда разница между статистиками G^2 начальной и полученной моделей позволяют оценить исключенный член отдельно. Отбирается для дальнейшей работы та подмодель, которая имеет наибольшую значимость. Процедура заканчивается, когда ни одна из подмоделей не обладает значимостью выше α.

6.6. Анализ остатков

Когда получена возможная модель, следует оценить ее адекватность. Кроме статистики для оценки качества подгонки следует изучить остатки между оцененными и действительными частотами. Если какая-либо ячейка дает существенное отклонение оценки, следует модифицировать модель. (возможно, вернуть исключенный член в модель). После получения удовлетворительных остатков производится интерпретация отдельных членов модели. Она связана со сворачиванием таблиц и вычислением соответствующих процентов.

6.7. Структура данных

Пример 16. Приведены табулированные данные о частотном распределении 150 обследованных во ВЦЭРМ ликвидаторов в зависимости от их генотипа ACE (получен в НИО генетической диагностики, начальник Слозина Н.М.), возраста участия в ликвидации и наличия стенокардии (I20 по МКБ10) - по данным НРЭР.

Таблица П15-1.

Частота	**Возраст участия, лет (Age_gr)**	**I20**	**ACE**
6	18-29	0 (нет)	II
11	18-29	0	ID
6	18-29	0	DD
1	18-29	1 (есть)	II
4	18-29	1	ID
4	18-29	1	DD
10	30-39	0	II
27	30-39	0	ID
7	30-39	0	DD
7	30-39	1	II
24	30-39	1	ID
9	30-39	1	DD
2	40+	0	II
1	40+	0	ID
3	40+	0	DD
8	40+	1	II
12	40+	1	ID
8	40+	1	DD

При работе с программой NCSS может быть введено до 7 факторов, но как минимум 2. Кроме того, при работе с табличными данными еще должна быть введена частота. В данном примере учитываются 3 фактора: **Возраст участия, I20 и ACE.** Они обозначаются последовательно символами A, B, C.

При работе с программой STATISTICA возможно использование таких же таблиц, а также исходного файла данных. Модуль *Логлинейный анализ* блока *Углубленные методы анализа* содержит полную реализацию процедур логлинейного анализа многовходовых таблиц частот. Могут анализироваться таблицы с числом измерений от 2 до 7. Таблицы могут содержать структурные нули. Частотные таблицы могут быть вычислены по исходным данным либо введены

непосредственно. В данной программе факторы обозначаются последовательными номерами: 1, 2, 3, и т.д.

6.8. Задание параметров LLM для программы NCSS

Модель. Эта опция позволяет определить иерархическую модель для оценивания. Если применяется пошаговая процедура отбора, эта модель будет стартовой.

Полная модель (Full Model). Эта опция определяет в качестве модели для оценивания насыщенную модель.

До (1, 2, 3) – входов (Up to (1, 2, 3) – Way). Эта опция устанавливает, что в модель включаются члены вплоть до указанного порядка. Например, если указаны «2 входа», при трех факторах это означает, что будет проанализирована иерархическая модель {AB, AC, BC}.

Пользовательская модель (Custom Model). Используя данную опцию, можно определить нужную модель (иерархическую), руководствуясь следующими правилами.

Каждый иерархический член предполагает включение в модель и всех комбинаций составляющих его факторов меньшего порядка. Например, если для 5-факторной модели (A – E) определена иерархическая модель {ABC, DE}, это означает, что в модель включены следующие члены: A, B, C, AB, AC, BC, ABC, D, E, DE.

Δ (Delta Value). Это число, обычно из интервала (0.1, 0.9), которое прибавляется к числу наблюдений в каждой ячейке таблицы, если там присутствуют нули. Это позволяет анализировать таблицы с нулями (поскольку процедура предполагает логарифмирование). При использовании этой опции лучше провести анализ при 2 – 3 значениях параметра, чтобы определить, насколько его значение влияет на результат анализа.

Опция процедуры максимального правдоподобия – максимальное число итераций (Max Iterations). В этой опции определяется максимальное число итераций. Обычно алгоритм сходится менее чем за 5 шагов, поэтому 25 итераций будет более чем достаточно.

Опция процедуры максимального правдоподобия – максимальная разность (Max Difference). Эта опция определяет максимальную разность между наблюдаемыми и предсказанными частотами таблицы. Как только максимум становится меньше этого числа, процедура максимального правдоподобия прекращается (сходится).

Осуществить пошаговый поиск (Perform Step-Down Search). Опция определяет, используется ли данная процедура. Процедура начинается с той модели, которая определена в опции «Модель». Осуществление процедуры определяется двумя параметрами – «Максимальное число моделей» и «α для остановки».

Максимальное число моделей (Max Models). Эта опция определяет максимальное количество моделей, которые могут быть протестированы до окончания работы процедуры.

α для остановки (Stopping Alpha). Эта опция задает значение α, которое является уровнем значимости для оценки качества проверяемой модели. Если в процессе поиска не будет найдено ни одной модели, для которой p-значение выше заданного α, поиск заканчивается. Напомним, что мы ищем модель, которая хорошо описывает данные, и прекращение процесса означает, что не получено достаточного качества описания.

Хотя вы, возможно, привыкли всегда использовать уровень α, равный 0.05, следует использовать и большие значения (например, 0.15 или 0.25), поскольку требуется модель, хорошо описывающая данные, а это не всегда связано со значимостью. Модель, которая «почти значима» (с α=0.06 или α=0.08), может не включать в себя важные члены. Если же вы выбираете значение α, равное 0.25, то можете быть уверены, что модель хорошо описывает данные.

К сожалению, соответствующее значение α также связано с объемом выборки. Для малых выборок уровень значимости 0.25 может привести к отклонению всех гипотез и отсутствию согласованности между моделью и данными. Поэтому для малых выборок можно получить плохую подгонку и большое α. С другой стороны, для больших выборок даже уровень 0.05 может оказаться чрезмерно большим, и его следует уменьшить.

6.9. Содержание отчетов программ NCSS и STATISTICA при реализации алгоритма LLM

При работе с программой NCSS

Раздел проверки нескольких членов (Multiple-Term Test Section)

K-Terms	DF	Like. Ratio Chi-Square	Prob Level	Pearson Chi-Square	Prob Level
1WAY & Higher	17	82.45	**0.0000**	95.77	**0.0000**
2WAY & Higher	12	29.28	**0.0036**	26.24	**0.0099**
3WAY & Higher	4	2.42	0.6597	2.48	0.6489

K-Terms	DF	Like. Ratio Chi-Square	Prob Level
1WAY Only	5	53.17	**0.0000**
2WAY Only	8	26.86	**0.0007**
3WAY Only	4	2.42	0.6597

Note: Simultaneous test that all interactions of order k are zero. These Chi-Squares are differences in the above table.

Этот отчет помогает в процессе выбора модели – изолируя члены высших порядков, можно оценить, какие из них следует включить в окончательную модель.

Верхняя таблица показывает значимость всех членов данного порядка и более высоких порядков. Например, 29.28 дает значимость всех членов 2 и 3 порядка, 2.42 – значимость взаимодействия 3-го порядка (членов более высокого порядка здесь нет).

Просматривая уровни значимости в этой таблице (**Prob Level**), можно сразу же определить наивысший порядок значимых членов. В данном примере значимыми являются члены 1 и 2 порядков.

Вторая таблица получена из первой с помощью вычитания. Это касается только критерия максимального правдоподобия (**Like. Ratio Chi-Square**), так как критерий χ^2 Пирсона свойством аддитивности не обладает.

При работе с программой STATISTICA

Критерии маргинальных и частных связей					
	Число	**Част.св.**	**Част.св.**	**Марг.св.**	**Марг.св.**
1	2	30.60	0.000	30.60	0.000
2	1	0.10	0.751	0.10	0.751
3	2	22.47	0.000	22.47	0.000
12	2	20.13	0.000	19.83	0.000
13	4	6.39	0.172	6.09	0.193
23	2	0.95	0.623	0.65	0.724

Результаты подгонки К-факторн.взаимодействий Это одновременная проверка того, что все К-факторные взаимодействия равны нулю					
	Число	МП	Вероятн.	Пирсона	Вероятн.
1	5	53.17	**0.000**	69.53	**0.000**
2	8	26.86	**0.001**	23.77	**0.003**
3	4	2.42	0.660	2.48	0.649

Раздел проверки отдельных членов (Single -Term Test Section)

При работе с программой NCSS

Single-Term Test Section

Effect	**DF**	**Partial Chi-Square**	**Prob Level**	**Marginal Chi-Square**	**Prob Level**
A (Age_gr)	2	30.60	**0.0000**	30.60	**0.0000**
B (I20)	1	0.10	0.7511	0.10	0.7511
C (ACE)	2	22.47	**0.0000**	22.47	**0.0000**
AB	2	20.13	**0.0000**	19.83	**0.0000**
AC	4	6.39	0.1721	6.09	0.1928
BC	2	0.95	0.6226	0.65	0.7238
ABC	4	2.42	0.6597	2.42	0.6597

В этом отчете приведены значения тестов для оценки частных и маргинальных ассоциаций для членов 1 – 3 порядка. Следует отметить, что частный критерий χ^2 позволяет проверить, является ли значимым влияние данного члена после учета всех других членов того же порядка. А маргинальный критерий χ^2 дает значимость данного члена, если не учитываются все остальные члены того же порядка. Следовательно, если результаты обоих тестов совпадают, это позволяет отобрать значимые члены модели (в данном случае это A, C, AB).

Раздел пошаговой процедуры отбора (Step-Down Model-Search Section**)**

При работе с программой NCSS

Step No	Best No	DF	Chi-Square	Prob Level	Term Deleted	DF	Chi-Square	Prob Level	Hierarchical Model
1	1	4	2.4	0.6597	None	0	0.0	**0.0000**	AB,AC,BC
2	1	6	22.5	0.0010	AB	2	20.1	**0.0000**	AC,BC
3	1	8	8.8	0.3592	AC	4	6.4	0.1721	AB,BC
4	1	6	3.4	0.7620	BC	2	0.9	0.6226	AB,AC
5	4	8	23.2	0.0031	AB	2	19.8	**0.0000**	AC,B
6	4	10	9.4	0.4900	AC	4	6.1	0.1928	AB,C
7	6	12	29.3	0.0036	AB	2	19.8	**0.0000**	C,B,A
8	6	12	31.9	0.0014	C	2	22.5	**0.0000**	AB
Best model found: AB,C									
6	6	10	9.4	0.4900	AC	4	6.1	0.1928	AB,C

В данном разделе показана процедура отбора на каждом шаге. На первом шаге рассматривается насыщенная модель, затем начинается процесс исключения. Иногда интересно просмотреть весь процесс, а не только финальную модель.

Приводится номер шага (1-й столбец таблицы). На этот номер ссылается 2-й столбец (**Best No**). В 3-ем столбце DF – число степеней свободы – приведено для тех членов, которые на данном шаге НЕ включены в модель, по сравнению с исходной.

4-й столбец (**Chi- Square**) – значение отношения правдоподобия G^2 для членов, исключенных из исходной модели на этом этапе. Оно позволяет проверить качество подгонки данных текущей моделью: если тест незначим, можно предположить, что все существенные члены присутствуют в модели.

Следующий столбец (**Prob Level**) – p-значение для приведенного выше значения статистики.

6-й столбец (**Term Deleted**) – тот член, который был исключен на данном шаге из текущей модели. Заметим, что на каждом шаге

исключается ровно один член. В 7-ом столбце DF – число степеней свободы – приведено для того члена, который исключен на данном шаге.

Следующий столбец - (**Chi- Square**) – значение отношения правдоподобия G^2 для того члена, который исключен из модели на этом этапе. Далее - (**Prob Level**) – p-значение для этого значения статистики.

Последний столбец (**Hierarchical Model**) – запись той модели, которая оценивается на данном шаге. Лучшая из полученных моделей (в данном случае это полученная на 6-м шаге иерархическая модель {AB, C}) приведена в последней строке данного раздела.

В программе STATISTICA приводятся только финальные результаты, сама процедура отбора скрыта:

Лучш.начал.модель: 21,31,32	**хи-квадрат = 2.4161**	**сс = 4**	**p = 0.6597**
Лучшая мод. 21,3	**хи-квадрат = 9.4495**	**сс = 10**	**p = 0.4900**

Раздел описания модели (Model Section**)**

При работе с программой NCSS

Model Section
Hierarchical Model: AB,C

Model Term	Individual DF	Cumulative DF
Mean	1	1
A	2	3
B	1	4
AB	2	6
C	2	8
Error	10	18

В данном разделе перечислены все члены выбранной модели и соответствующее им число степеней свободы.

Раздел проверки модели с помощью критерия χ2

При работе с программой NCSS

Chi-Square Tests Section

DF	Like. Ratio Chi-Square	Prob Level	Pearson Chi-Square	Prob Level	Model
10	9.45	0.4900	9.31	0.5026	AB,C

В этом разделе приведены значения обеих статистик (χ^2 Пирсона и G^2 максимального правдоподобия) для выбранной модели.

Раздел оценки параметров

При работе с программой NCSS

Parameter Estimation Section

Model Term	Number Cells	Percent Count	Count	Average Log (Count)	Effect (Lambda)	Effect Std. Error	Effect Z-Value
Mean	18	159	100.00	1.9154	1.9154	0.1049	18.25
A: Age_gr							
18-29	6	35	22.01	1.6083	-0.3071	0.1545	-1.99
30-39	6	87	54.72	2.6050	0.6896	0.1240	5.56
40+	6	37	23.27	1.5329	-0.3825	0.1638	-2.34
B: I20							
нет	9	78	48.74	1.8437	-0.0717	0.1049	-0.68
есть	9	82	51.26	1.9871	0.0717	0.1049	0.68
C: ACE							
DD	6	40	25.16	1.7021	-0.2133	0.1545	-1.38
ID	6	82	51.57	2.4199	0.5045	0.1315	3.84
II	6	37	23.27	1.6241	-0.2913	0.1578	-1.85
AB: Age_gr, I20							
18-29, нет	3	25	15.41	2.0319	0.4953	0.1545	3.21
18-29, есть	3	11	6.60	1.1846	-0.4953	0.1545	-3.21
30-39, нет	3	46	28.62	2.6510	0.1177	0.1240	0.95
30-39, есть	3	42	26.10	2.5590	-0.1177	0.1240	-0.95
40+, нет	3	8	4.72	0.8482	-0.6130	0.1638	-3.74
40+, есть	3	30	18.55	2.2177	0.6130	0.1638	3.74

В таблице приведены детали логлинейной оценки выбранной модели. Она и является целью LLM-анализа. Столбцы означают следующее:

1. Model Term - отдельные члены модели и все их уровни.
2. Number Cells – количество ячеек, включенных в данный член
3. Count – общее число объектов в ячейках, относящихся к данному уровню
4. Percent Count – общее число объектов в ячейках, выраженное в процентах по отношению к общему количеству объектов. Эти

проценты также используются для того, чтобы понять, почему этот член оказался значимым

5. Average Log(Count) – среднее значение Log(Count+Δ) всех ячеек с указанными уровнями.
6. Effect (Lambda) – оцененное значение λ для данного члена. Эти параметры описаны выше. Они оценены с помощью процедуры Хабермана (Haberman).
7. Effect Std. Error – асимптотическая стандартная ошибка для приведенного выше эффекта λ. Когда оценивается насыщенная модель, стандартная ошибка вычисляется как квадратный корень из дисперсии эффекта. А дисперсия оценивается по формулам Ли (Lee). При оценке ненасыщенной модели программа использует при вычислениях оценки насыщенной модели в соответствии с аппроксимационным методом Ли.
8. Effect Z-Value – это эффект, деленный на стандартную ошибку. Поскольку количество ячеек для разных членов модели различно, точность оценки также отличается. Z-значение позволяет сравнивать относительную величину эффектов первого порядка и взаимодействий. Эти значения представляют собой относительную важность данного члена в логлинейной модели. Используются именно z-значения для членов модели, поскольку они распределены асимптотически нормально. Они называются стандартизованными оценками параметров в разделе «Техника выбора модели».

При работе с программой STATISTICA

Марг.Табл.(част+дельта): age по I 20 в перем.: ACE:II				
I 20	**Age 18-29**	**Age 30-39**	**Age 40+**	**Сумма**
0	6.50	10.5	2.5	19.5
1	1.50	7.5	8.5	17.5
Сумма	8.00	18.0	11.0	37.0

Подогн.част.: I 20 по age в перем.: ACE:II				
I 20	**Age 18-29**	**Age 30-39**	**Age 40+**	**Сумма**
0	5.701	10.59	1.745	18.03
1	2.443	9.66	6.865	18.97
Сумма	8.145	20.25	8.610	37.00

Откл. Фримена-Тьюки: I 20 по age в перем.: ACE:II				
I 20	**Age 18-29**	**Age 30-39**	**Age 40+**	**Сумма**
0	0.409	0.047	0.627	1.083
1	-0.476	-0.641	0.663	-0.455
Сумма	-0.067	-0.594	1.290	0.629

Комп. МП хи-кв.: I 20 по age в перем.: ACE:II				
I 20	**Age 18-29**	**Age 30-39**	**Age 40+**	**Сумма**
0	1.70	-0.18	1.797	3.33
1	-1.46	-3.79	3.632	-1.62
Сумма	0.24	-3.97	5.429	1.70

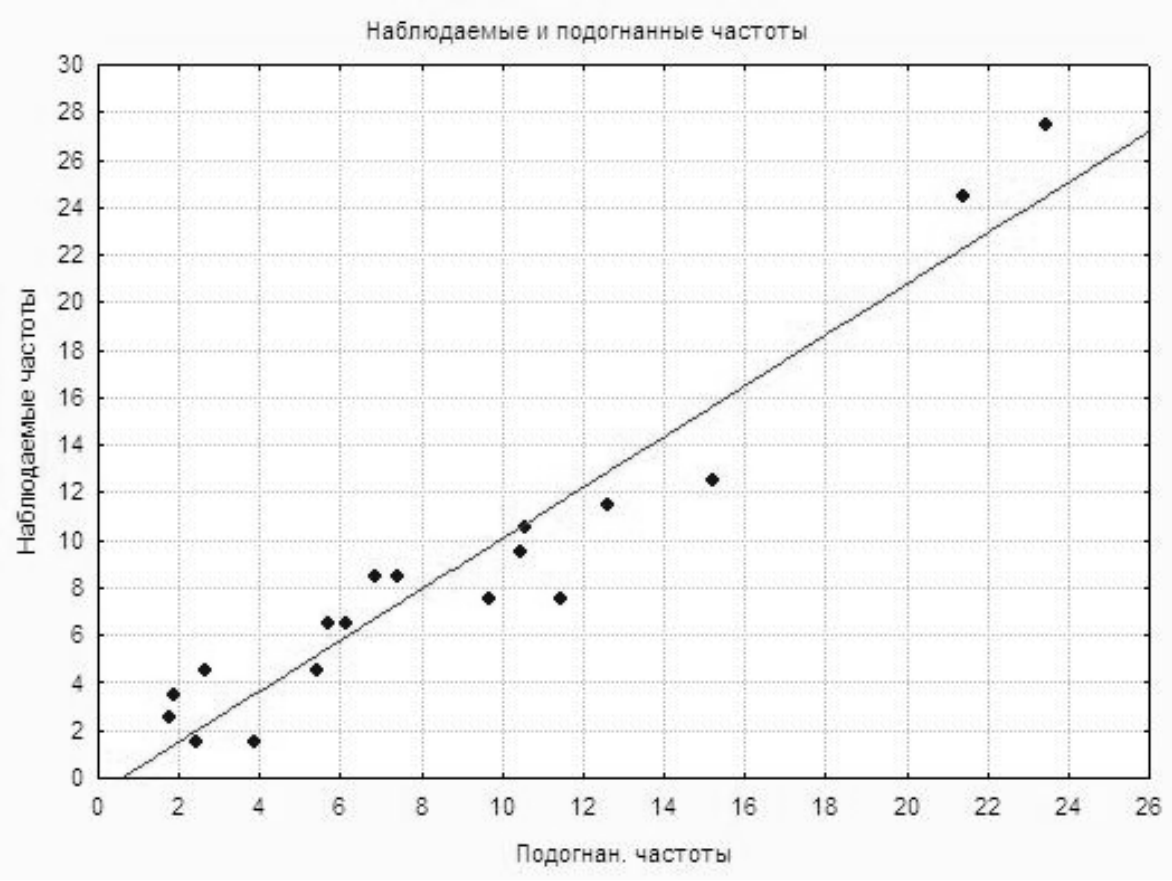


Рис. П16-1.

Интерпретация значимых эффектов

Последний этап логлинейного анализа – интерпретация полученных результатов. Для этого осуществляется сворачивание общей процентной таблицы по всем значимым эффектам. Например,

для значимого взаимодействия AB получим следующую свернутую таблицу:

Возрастная группа	**Наличие I 20**		**Сумма**
	нет	**есть**	
18-29 лет	70.0% (=100*15.41/22.01)	30.0% (=100*6.60/22.01)	100%
30-39 лет	52.3% (=100*28.62/54.72)	47.7% (=100*26.10/54.72)	100%
40+ лет	20.3% (=100*4.72/23.27)	79.7% (=100*18.55/23.27)	100%

Разница в распределении процентов по каждой из строк и обусловила значимость данного взаимодействия.

Таким образом, при логлинейном анализе взаимодействия стенокардии и ACE не обнаружено.

Раздел «таблица данных» (Data Table Section**)**

При работе с программой NCSS

ACE	I20	Age_gr	Actual	Pred	Diff	Chi	FT-SR
DD	0	18-29	6.5	6.2	0.3	0.14	0.22
DD	0	30-39	7.5	11.4	-3.9	-1.17	-1.19
DD	0	40+	3.5	1.9	1.6	1.17	1.07
DD	1	18-29	4.5	2.6	1.9	1.14	1.07
DD	1	30-39	9.5	10.4	-0.9	-0.29	-0.22
DD	1	40+	8.5	7.4	1.1	0.40	0.46
ID	0	18-29	11.5	12.6	-1.1	-0.32	-0.25
ID	0	30-39	27.5	23.5	4.0	0.83	0.84
ID	0	40+	1.5	3.9	-2.4	-1.20	-1.25
ID	1	18-29	4.5	5.4	-0.9	-0.39	-0.29
ID	1	30-39	24.5	21.4	3.1	0.67	0.69
ID	1	40+	12.5	15.2	-2.7	-0.70	-0.66
II	0	18-29	6.5	5.7	0.8	0.33	0.41
II	0	30-39	10.5	10.6	-0.1	-0.03	0.05
II	0	40+	2.5	1.7	0.8	0.57	0.63
II	1	18-29	1.5	2.4	-0.9	-0.60	-0.48
II	1	30-39	7.5	9.7	-2.2	-0.69	-0.64
II	1	40+	8.5	6.9	1.6	0.62	0.66

Данная таблица позволяет найти большие разности – т.е. ячейки, которые неудовлетворительно описаны LLM.

Actual – частоты в ячейках f_{ijk} , полученные из исходных данных.

Predicted – предсказанные на основании выбранной модели частоты m_{ijk}. Уравнение для этих частот имеет следующий вид:

$$\ln(m_{ijk}) = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \ ,$$

где оценки параметров получены с помощью алгоритма Хабермана.

Difference – это остатки: исходные частоты – предсказанные частоты. Важно найти не просто большие отклонения, а большие стандартизованные отклонения – они в следующей колонке.

Chi - это стандартизованные остатки. Они вычисляются по формуле:

$$Chi = (f_{ijk} - \tilde{m}_{ijk}) / \sqrt{\tilde{m}_{ijk}}.$$

Это корень квадратный из компонента, соответствующего данной ячейке, в общем выражении статистики χ^2 Пирсона для оценки качества подгонки. Эти стандартизованные остатки позволяют непосредственно сравнивать подгонку отдельных ячеек. Если значение |Chi| > 1.96, такой остаток следует рассматривать как большой.

FT-SR – стандартизованные остатки Фримена-Тьюки.

$$FTSR = \sqrt{f_{ijk}} + \sqrt{1 + f_{ijk}} - \sqrt{1 + 4\tilde{m}_{ijk}}$$

Эти значения также могут рассматриваться как полученные из N(0,1). Соответственно, так же, как в предыдущем случае, значения, превышающие по абсолютной величине 1.96, следует считать большими.

В разобранном примере больших отклонений предсказанных и наблюдаемых частот нет, поэтому модель следует признать удовлетворительной.

Литература

1. С. А. Айвазян. Прикладная статистика. Исследование зависимостей.: Справ.изд./ Айвазян С. А., Енюков И. С., Мешалкин Л. Ш. – М.: Финансы и статистика, 1985. – 487 с.
2. А. Альбом. Введение в современную эпидемиологию / Альбом А., Норелл С.; пер.с англ. И. Боня. – Таллинн, 1996. – 122 с.
3. Анализ медицинских данных государственного статистического наблюдения. Сборник Комитета по здравоохранению Администрации Санкт-Петербурга / В.М.Дорофеев и др.,- СПб, 2003.
4. А. Банержи. Медицинская статистика понятным языком: вводный курс / Банержи А.; пер.с англ.под ред. В.П.Леонова. – М.: Практическая медицина, 2007. – 287 с.
5. В. Боровиков. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд./ Боровиков В. - СПб: Питер, 2003. – 688 с.
6. А. Бююль. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. / Бююль А., Цефель П.; пер.с нем./ под ред. В. Е. Момота. СПб: ООО «ДиаСофтЮП», 2002. – 608 с.
7. И. Гайдышев. Анализ и обработка данных. Специальный справочник./ Гайдышев. И. – СПб: Питер, 2001. – 752 с.
8. С. Гланц. Медико-биологическая статистика./ Гланц С.; пер.с англ. Ю.А. Данилова; под ред. Н. Е. Бузикашвили и Д. В. Самойлова. - М., Практика, 1999. – 459 с.
9. Государственный доклад о состоянии здоровья населения Российской Федерации в 1999 году.: офиц. Изд. –-М.:ГЭОТАР-МЕД, 2000. – 104 с.
10. В. А. Медик. Статистика в медицине и биологии: Руководство. В 2-х томах / Медик В. А., Токмачев М. С., Фишман Б. Б.; под ред. Ю.М. Комарова. – М.: Медицина. - 2000. - Т.1. Теоретическая статистика.– 412 с.; 2001 - Т.2. Прикладная статистика здоровья. – 352 с.
11. В. М. Медков. Демография: Учебник./ Медков В. М. – М.: ИНФРА - М, 2004. – 576 с.
12. А. Петри. Наглядная медицинская статистика./ Петри А., Сэбин К.; пер.с англ.под ред. В.П.Леонова, 2-е изд. – М:Издат.группа «ГЭОТАР-Медиа», 2009. – 165 с.
13. Справочник по прикладной статистике. Т.1.: пер.с англ. / под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. – М.: Финансы и статистика, 1989. – 510 с.

14. Справочник по прикладной статистике. Т.2.: пер.с англ. / под ред. Э. Ллойда, У. Ледермана, С. А. Айвазяна, Ю. Н. Тюрина. – М.: Финансы и статистика, 1990. – 526 с.

15. BMDP Statistical Software Manuel. Volume 1, 2. / W. J. Dixon, chief editor – University of California Press, Berkeley –Los Angeles - Oxford, 1990. – 1380 p.

16. N. E. Breslow. Statistical Methods in Cancer Research. V.1: The Analysis of Case-Control Studies. / Breslow N. E., Day N. E. - IARC Scientific Pub. № 32. Lion: IARC, 1980.

17. NCSS Help System. Copyright © 2007 Dr. Jerry L. Hintze, Kaysville, Utah 84037

18. Osborn J. F. Basic Statistical Methods for Epidemiological Studies.

ПРИЛОЖЕНИЕ: СЛОВАРЬ И ФОРМУЛЫ

Распределения случайных величин и статистические характеристики выборки

Закон распределения случайной величины (с.в.)

Нормальный закон N(μ, σ). С.в. непрерывного типа распределена по нормальному (гауссовскому) закону с параметрами μ и σ, если плотность распределения вероятностей этой с.в. задается формулой

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \text{ где } -\infty < x < \infty$$

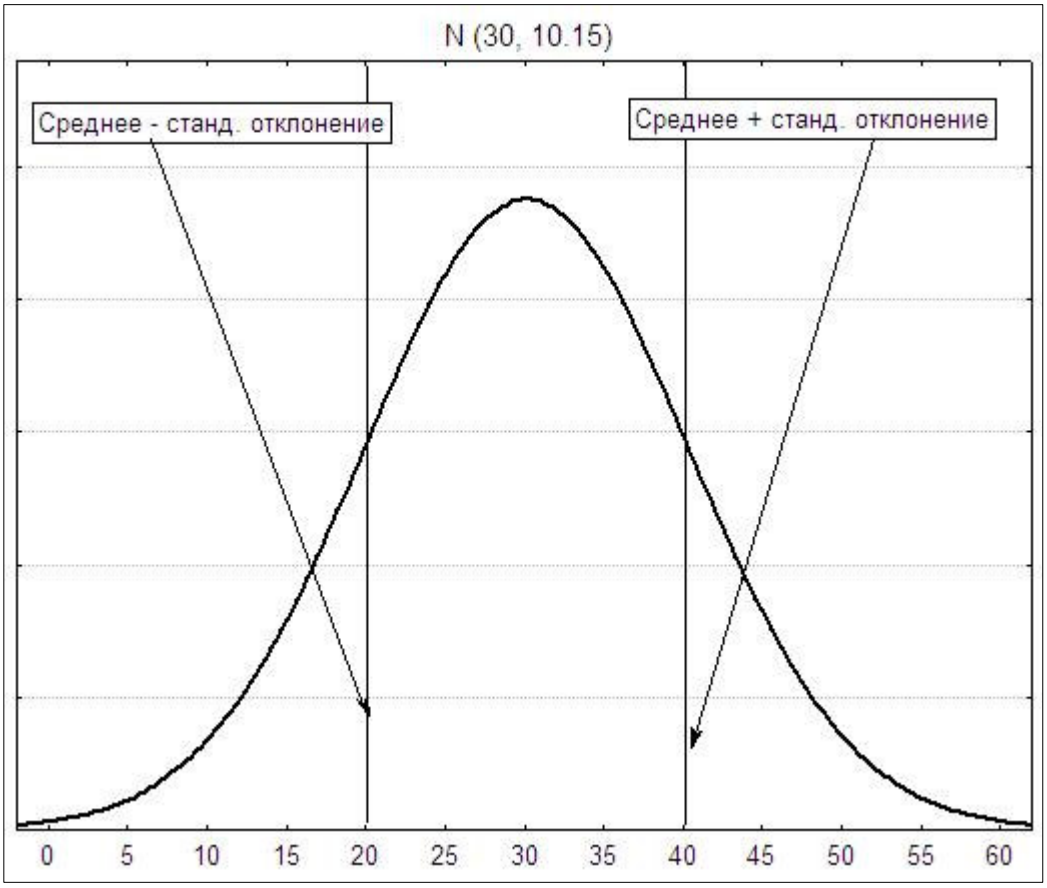


Рис.1. График плотности распределения вероятностей нормальной с.в. с параметрами μ =30, σ =10.15

Для нормально распределенной с.в. верно следующее утверждение. Вероятность того, что отклонение с.в. x от ее математического ожидания не превзойдет kσ, где k = 1, 2,3 а σ - стандартное отклонение, составляет:

$$P(|x - \mu| < \sigma) \approx 0.683 \ (k=1); \quad P(|x - \mu| < 2\sigma) \approx 0.954 \ (k=2);$$

$$P(|x-\mu|<3\sigma) \approx 0.997 \ (k=3)$$

(2) ***Биномиальный*** *(Бернулли)* B(n,p). Используется для моделирования дихотомических данных – признаков, которые могут иметь только два значения. Случайная величина x распределена по биномиальному закону, если x – количество успехов в серии из n независимых испытаний с двумя исходами («успех» и «неуспех») при том, что вероятность успеха в каждом испытании одинакова и равна p. Вероятность того, что в серии из n испытаний количество успехов будет равно k, задается формулой Бернулли:

$$P(x=k) = C_n^k \bullet p^k \bullet (1-p)^{n-k}$$

(3) ***Полиномиальный*** – обобщение биномиального закона для схемы, когда в каждом из n независимых испытаний имеется r взаимоисключающих исходов A_1, A_2, …, A_r соответственно с вероятностями $p_1, p_2, \ldots, p_r$; $\sum_{i=1}^{n} p_i = 1$. Вероятности полиномиального распределения задаются формулой:

$$P(x_1 = n_1, x_2 = n_2, ..., x_r = n_r) = \frac{n!}{n_1! n_2! ... n_r!} p_1^{n_1} p_2^{n_2} ... p_r^{n_r}$$

Это вероятность того, что в серии из n испытаний событие A_1 появится ровно n_1 раз, событие A_2 появится ровно n_2 раз, …, событие A_r появится ровно n_r раз, причем $\sum_{i=1}^{r} n_i = n$

(4) С.в. x распределена по ***закону Пуассона*** с параметром λ ($\lambda > 0$), если она может принимать только целочисленные значения 0, 1, 2, … , а вероятности этих значений определяются формулой

$$P(x=k) = \frac{\lambda^k}{k!} e^{-\lambda} .$$

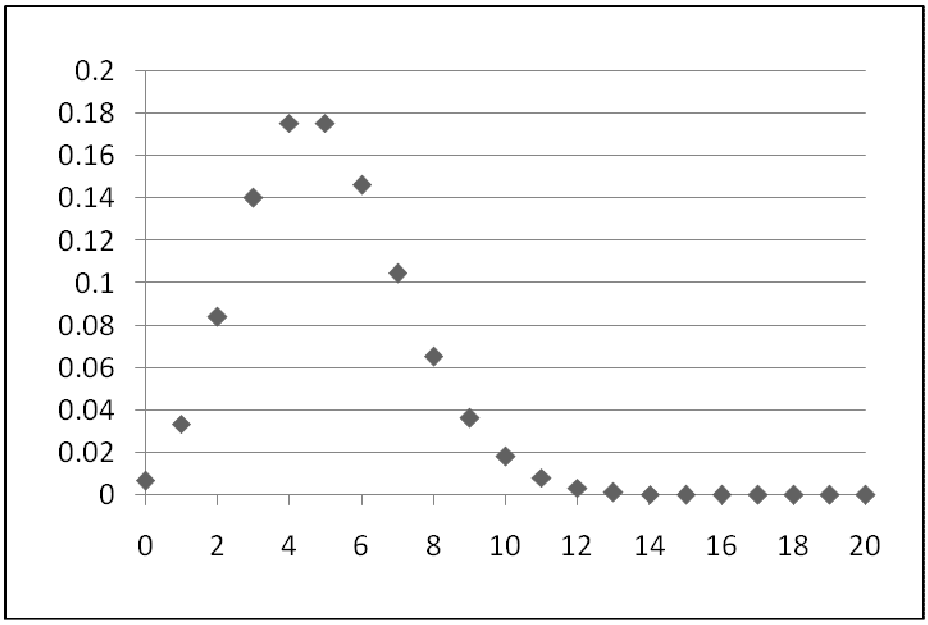


Рис.2. График закона распределения вероятностей Пуассона с параметром $\lambda=5$

Распределение Пуассона используют в качестве удобного приближения биномиального распределения в случае, когда p мало ($p << 1$), а n велико ($n >> 100$). В этом случае распределение Пуассона интерпретируется как «закон редких явлений». Параметр λ принимается равным np.

Выборка объема n – $\{x_1,x_2,\ldots,x_n\}$

ряд значений (реализаций) с.в. x, подчиняющейся некоторому закону распределения (например, нормальному $N(\mu, \sigma)$, биномиальному B(n, p), Пуассона и т.д. или не имеющему параметрического вида).

Характеристики положения с.в.

Среднее M с.в. x (expected value, mean)

$$M = E(x)$$

математическое ожидание (для любой с.в., определение среднего).

Для частных случаев распределений с.в. среднее вычисляется по формулам:

- для количества ответов «да» биномиального закона B(n,p):

$$M = np$$

- для нормального закона $N(\mu, \sigma)$:

$$M = \mu$$

- для распределения Пуассона:

$M = \lambda$ (параметр распределения Пуассона)

Выборочное среднее (sample mean)

$M_x = (x_1+x_2+\ldots+x_n) / n$ – среднее арифметическое

$M_x = x$ – количество ответов «да» (успехов) для биномиального закона B(n,p)

Оценка параметра p биномиального распределения

- это относительная частота x – количества ответов «да» в выборке объема n.

$$h = x / n$$

Если ответы «да» закодированы 1, а ответы «нет» - 0, то

$$h = M_x.$$

Мода (mode) $M_o = x_0$

- это наиболее вероятное значение с.в. x (значение, для которого его вероятность p_0 или плотность вероятности $p(x_0)$ достигает максимума). Распределение может иметь несколько максимумов. В этом случае оно называется многомодальным. Нормальное распределение, распределение Пуассона – унимодальные. Пример многомодальности на следующем рисунке.

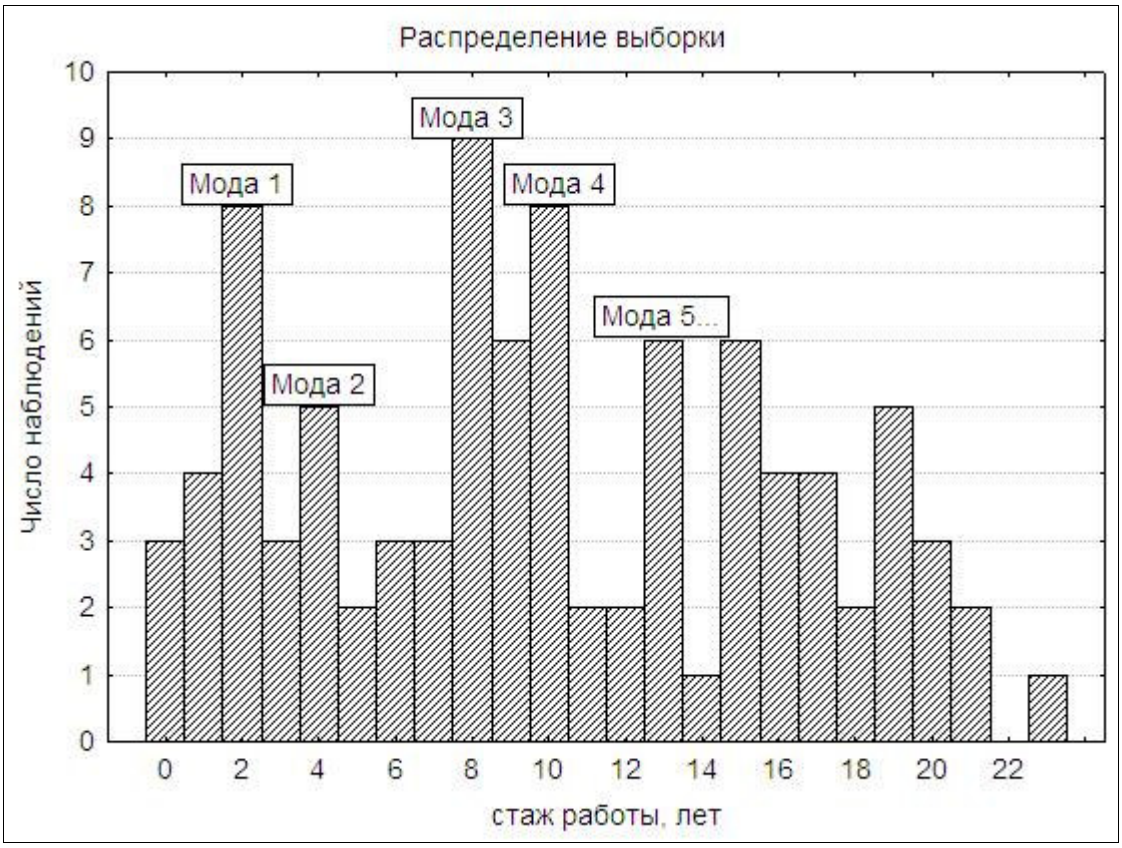


Рис.3. Многомодальность выборочного распределения

Квантиль (Percentile)

x_p – квантиль с.в., имеющей функцию распределения F(x), если x_p является решением уравнения F(x) = p (квантиль уровня p).

Децили

квантили уровней 0.1, 0.2, …, 0.9.

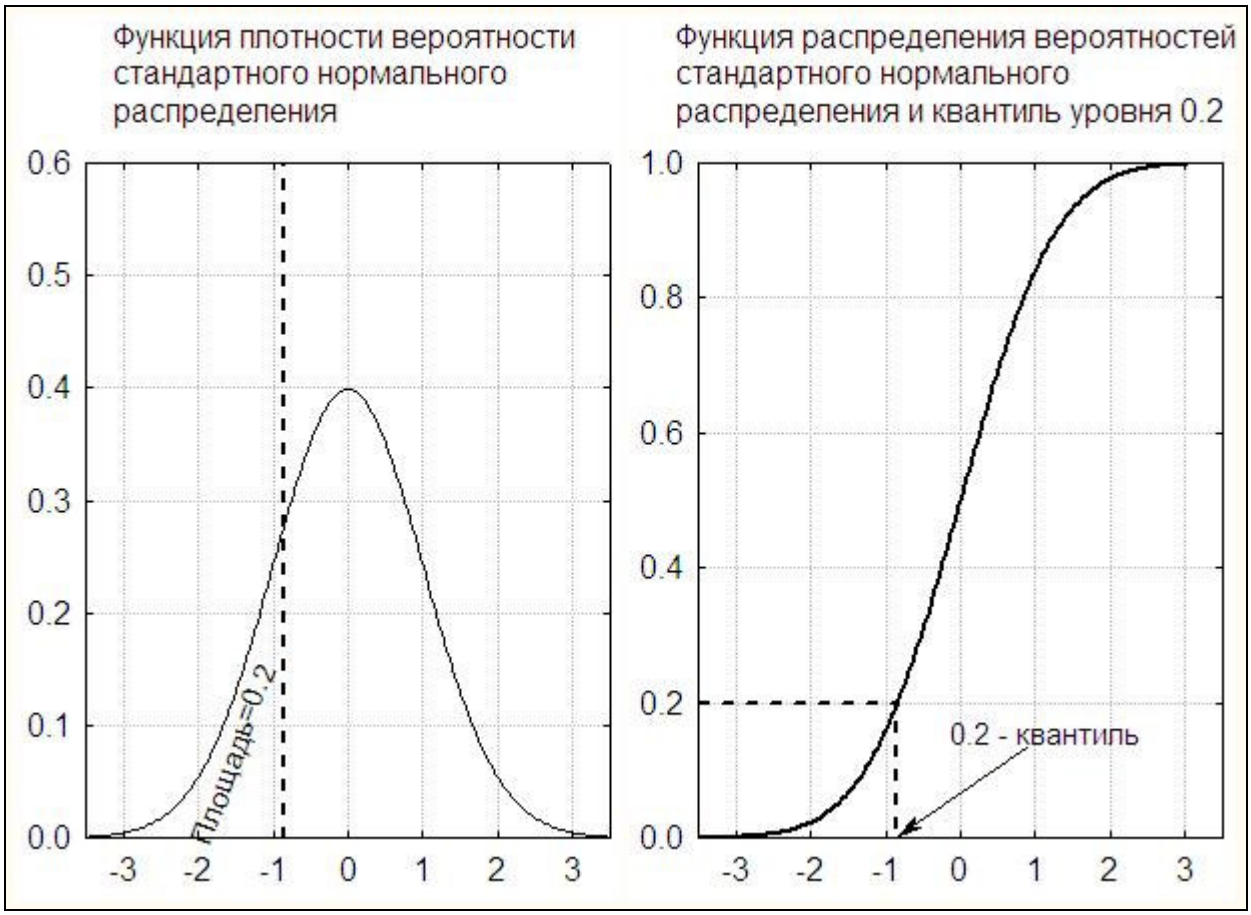


Рис.4. Квантиль уровня 0.2 для стандартного нормального распределения N(0,1)

Медиана (median)

$M_e = x_e$ – квантиль, соответствующая значению p=0.5: решение уравнения $F(x_e) = 0.5$

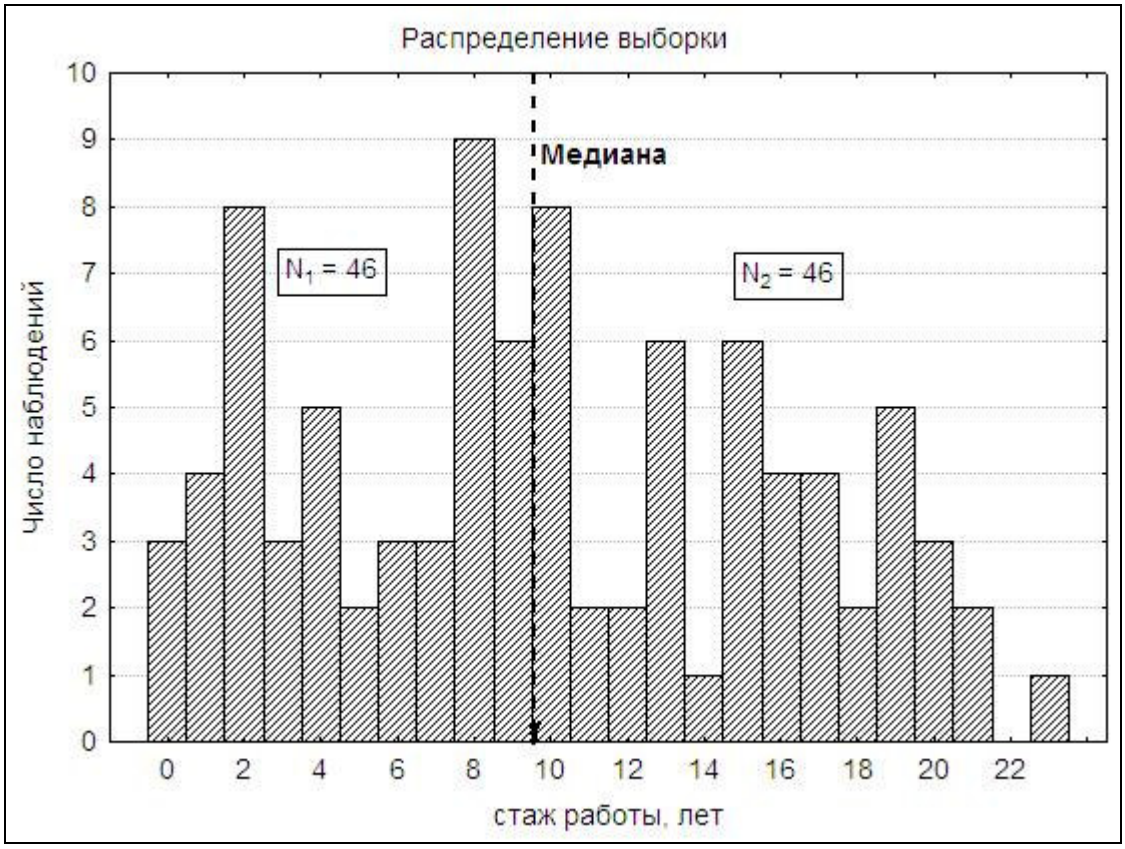


Рис.5. Медиана выборочного распределения

Выборочная медиана

$\hat{x}_e$ - решение уравнения $F_n(\hat{x}_e) = 0.5$, где $F_n(x)$ – выборочная функция распределения.

Характеристики формы с.в.

Дисперсия (variance)

- это математическое ожидание квадрата отклонения от среднего (для любой с.в., определение дисперсии):

$$D = E(x-M)^2$$

дисперсия x – количества ответов «да» для биномиального закона, q = 1-p:

$$D = npq$$

для нормального закона $N(\mu, \sigma)$:

$$D = \sigma^2$$

для распределения Пуассона:

$$D = \lambda$$

Асимметрия (skewness)

A – коэффициент асимметрии - третий нормированный центральный момент. Характеризует несимметричность распределения с.в. Для нормального и любого другого симметричного распределения A=0. Если A<0, кривая распределения скошена влево, а если A>0, то вправо.

$$A = \frac{E(x-M)^3}{D^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$$

Эксцесс (kurtosis)

E – коэффициент эксцесса, - четвертый нормированный центральный момент минус 3. Характеризует выраженность «хвостов» распределения с.в. Для нормального распределения E=0. Для распределений, более размазанных вдоль OX, нежели нормальное, E<0.

$$E = \frac{E(x-M)^4}{D^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$$

Иногда используется значение эксцесса без вычитания 3. Тогда для нормального распределения $E^*=3$, для более островершинных распределений E^* больше 3, для более пологих – меньше 3.

Выборочная дисперсия

Общие определения:

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2$ - дисперсия оценки x, если известно среднее M генеральной совокупности.

$S_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - M_x)^2$ - несмещенная оценка дисперсии оценки x, если неизвестно генеральное среднее M, а M_x оценено по выборке.

Для биномиального распределения:

$S_x^{\,2} = \frac{x(n-x)}{n}$ – дисперсия оценки x – количества ответов «да» - для B(n,p)

$S_p^{\,2} = \frac{x(n-x)}{n^3} = \frac{h(1-h)}{n}$ - дисперсия h - оценки p для B(n,p).

Стандартное отклонение σ

$$\sigma = D^{1/2}$$

Выборочное стандартное отклонение s или s_x. (standart deviation)

$s = (S^2)^{1/2}$; $s_x = (S_x^2)^{1/2}$; $s_p = (S_p^2)^{1/2}$

Коэффициент вариации

Характеристика рассеяния распределения случайной величины. Выражается в долях или процентах и показывает, какую часть среднего составляет среднеквадратичное отклонение.

$$\vartheta = \frac{\sigma}{M} \times 100\% \,,\; C_v = \frac{\sigma}{M}$$

Стандартная ошибка

Стандартное отклонение среднего (стандартная или основная ошибка) m

$$m = \sigma / \sqrt{n}$$

(из предельной теоремы, распределение выборочного среднего при независимых испытаниях, при любом распределении с.в., имеет асимптотически нормальное распределение со средним M генеральной совокупности и стандартным отклонением m)

Стандартная (основная) ошибка выборочного среднего

$$m_x = s / \sqrt{n}; \qquad m_x = s_x / \sqrt{n}$$

Стандартная ошибка h - оценки p для B(n,p)

$$m_h = \sqrt{\frac{h \times (1-h)}{n}}$$

Стандартная ошибка разности $h_1 - h_2$ для B(n,p)

При сравнении двух независимых выборок с параметрами (n_1, h_1) и (n_2, h_2)

$$m_{h_1 - h_2} = \sqrt{h \times (1-h) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$, при оценке общей частоты

$$h = \frac{n_1 h_1 + n_2 h_2}{n_1 + n_2}.$$

Стандартная (основная) ошибка (выборочного) стандартного отклонения

$$m_\sigma = \frac{\sigma}{\sqrt{2n}}; \quad m_s = \frac{s}{\sqrt{2n}}$$

Стандартная (основная) ошибка (выборочного) коэффициента вариации

$$m_\vartheta = \vartheta \times \sqrt{\frac{0.5 + 0.0001 \times \vartheta^2}{n}}$$

Стандартная (основная) ошибка (выборочного) коэффициента корреляции

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}$$

Предельная ошибка выборки Δ (при доверительной вероятности 1-α) то же, что

Доверительный интервал

Доверительный интервал (confidence interval, confidence limit) - статистический показатель, позволяющий оценить, в каких пределах может находиться истинное значение параметра в популяции (генеральной совокупности). 95% доверительный интервал означает, что истинное значение параметра с вероятностью 95% лежит в его пределах.

Доверительный интервал для среднего

(1) Для непрерывных случайных величин (с.в.) (т.е. признаков, измеряемых в шкале интервалов), если распределение признака в генеральной совокупности – нормальное.
(1.1) Если дисперсия (D или σ) считается известной.

$$\Delta = M \pm u_{1-\alpha/2} * m,$$

где $u_{1-\alpha/2}$ – (1 – α/2) квантиль стандартизованного нормального распределения N(0,1).

При α = 0.05: $u_{0.975} = 1.96$

(1.2) Дисперсия неизвестна и оценивается по выборке.

$$\Delta = M_x \pm t_{1-\alpha/2}(n-1)*m_x,$$

где $t_{1-\alpha/2}(n-1)$ – это (1 – α/2) квантиль распределения Стьюдента с (n-1) степенью свободы.
При α = 0.05:

$$t_{0.975}(20) = 2.09;\ t_{0.975}(60) = 2.00;\ t_{0.975}(\infty) = 1.96\ (n \geq 100)$$

Доверительный интервал для параметра биномиального распределения p

(2) Для биномиальных с.в. (признаков типа «да»-«нет», причем в генеральной совокупности ответ «да» дается с вероятностью p, а оценка этой вероятности h = x/n) Используются точные и приближенные формулы.

Приближенные формулы допустимы, если nh(1-h) > 9.

(2.1) Большое число наблюдений (n > 100)
– приближенные границы, $u_{1-\alpha/2}$ – квантиль стандартного нормального распределения

$$\Delta = h \pm u_{1-\alpha/2}\cdot[h \cdot (1-h)/n]^{1/2}$$

(2.2) Не очень много наблюдений (50 < n ≤ 100)
– приближенные границы с поправочным слагаемым.

$$\Delta = \{h + 0.5\cdot(u_{1-\alpha/2})^2/n \pm u_{1-\alpha/2}\cdot[h(1-h)/n + (u_{1-\alpha/2}/2n)^2]^{1/2}\}\cdot n/[n + (u_{1-\alpha/2})^2]$$

(2.3) Малое число наблюдений или редкие события (n ≤ 50 или x < 5 или n-x < 5)
– точные границы

$$h_1 = \frac{h\times F_{\alpha/2}(m_1, m_2+2)}{n-hn+1/n+hn\times F_{\alpha/2}(m_1, m_2+2)}$$

$$h_2 = \frac{(hn+1)\times F_{1-\alpha/2}(m_1+2, m_2)}{n-hn+(1+hn)\times F_{1-\alpha/2}(m_1+2, m_2)}$$

где $m_1=2hn$, $m_2=2(n-hn)$,

$F_\alpha(k_1,k_2)$ – квантиль распределения Фишера с k_1 и k_2 степенями свободы. Квантили распределения Фишера связаны между собой соотношением

$$F_\alpha(k1, k2)=1/F_{1-\alpha}(k2, k1)$$

$\Delta = (h_1, h_2)$ – доверительный интервал.

Доверительный интервал для разности параметров 2-х биномиальных распределений p_1 и p_2

При достаточно больших n_1 и n_2 для доверительной вероятности 0.95 доверительный интервал для разности $(h_1 - h_2)$ (отличие этой разности от 0) составляет

$$\Delta = \left(-2\sigma_{h_1-h_2}, 2\sigma_{h_1-h_2}\right) = \mp\{2\times\sqrt{h\times(1-h)\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\}, \text{ где}$$

$$h = \frac{n_1h_1 + n_2h_2}{n_1 + n_2}$$

Критерии согласия

Критерий Колмогорова (статистика Колмогорова-Смирнова)

Предназначен для сравнения выборочной (эмпирической) функции распределения с теоретической (для непрерывных распределений). Статистика критерия

$$D_n = \max |F_n(x) - F(x)|$$

Распределение этой статистики одно и то же для всех непрерывных распределений.

При $n\to\infty$: $D_n\sqrt{n} \to$ функции распределения Колмогорова.

Это утверждение верно, если параметры теоретической функции распределения известны, а не оцениваются по выборке.

В случае, когда используются выборочные оценки параметров, предельные функции для различных семейств распределений отличаются. Для наиболее распространенных распределений они определены и табулированы.

В случае нормального распределения и выборочных оценок параметров следует использовать «вероятности Лиллиефорса (Lilliefors)» для определения значимости D_n.

В отличие от критерия χ^2, критерий КС неприменим для дискретных распределений.

Критерий Шапиро-Уилка (W)

Данный критерий более предпочтителен для проверки нормальности, нежели статистика Колмогорова-Смирнова. Этот тест рекомендуется использовать при объеме выборки от 3 до 5000. В большинстве ситуаций он является наиболее мощным. Это отношение двух оценок дисперсии нормального распределения, основанное на случайной выборке из n наблюдений. Числитель пропорционален квадрату наилучшей линейной оценки стандартного отклонения. Знаменатель – сумма квадратов отклонений наблюдений от выборочного среднего.

Мощность теста снижается при наличии выбросов в выборке.

Критерии Шапиро-Уилка и Колмогорова вычисляются в большинстве статистических пакетов. Более редко используются критерии Андерсона-Дарлинга (Anderson-Darling Test), Мартинеса-Иглевица (Martinez-Iglewicz Test) и тесты, разработанные Д'Агостиньо (D'Agostino): тест, основанный на коэффициенте асимметрии, тест, основанный на коэффициенте косости и тест, основанный на комбинации этих коэффициентов.

Простейшая проверка нормальности

В случае, когда $|E| < 0.1$ (E – эксцесс), распределение можно считать близким к нормальному.

Если $|E| > 0.5$, отклонения от нормальности значительные;

В случае, когда $|A| < 0.1$ (A – асимметрия), распределение симметрично.

Если $|A| > 0.5$, распределение сильно асимметрично.

Критерий χ^2

Критерий χ^2 в форме критерия согласия предназначен для сравнения эмпирической функции распределения с теоретической функцией распределения и вычисляется по формуле

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

где n_i , i = 1, 2, … , k, - частоты наблюдаемых случаев в k интервалах,

p_i , i = 1, 2, … , k, - теоретические вероятности выбранного распределения,
k – число интервалов значений показателя или, если показатель имеет конечное число значений, количество возможных значений;
n – общее число наблюдений.
Число степеней свободы для определения критического значения равно (k – r – 1), где r – количество параметров теоретического закона, которые были вычислены с помощью выборки. Для биномиального закона r = 1, для нормального закона r = 2, если все параметры оценивались по выборке.

Особенности применения критерия χ^2 изложены в главе «Использование критерия χ^2».

Точный метод Фишера (Фишера-Ирвина)

Проведена серия из n испытаний, в которой событие A появилось x раз. Согласуется ли на уровне α частота появления события в данной серии с заранее известной частотой (параметром биномиального распределения) p?

Оценкой параметра распределения p является x/n. Точный метод предполагает проверку согласия с учетом дискретности распределения x. В этом случае мы получаем точные границы для значений x, которые согласуются с известным параметром p на уровне, не меньшем, чем α. Именно, для ситуации, когда событие A в каждом испытании может появиться с одной и той же вероятностью p, для серии из n испытаний определяются числа x_1 и x_2 такие, что

I. {вероятность того, что событие A появится в серии из n испытаний менее x_1 раз}={ $P(x < x_1)$} ≤ α/2,

при этом x_1 - это наибольшее такое число, так что для x_1+1:

{ $P(x < x_1+1)$} > α/2;

II. {вероятность того, что событие A появится в серии из n испытаний более x_2 раз} = { $P(x > x_2)$} ≤ α/2,

при этом x_2 – это наименьшее такое число, так что для x_2-1:

{ $P(x > x_2–1)$} > α/2;

Формулы для вычисления этих вероятностей

$$P(x < x_1) = \sum_{k=0}^{x_1-1} C_n^k \times p^k \times (1-p)^{n-k} \le \alpha/2$$

$$P(x > x_2) = \sum_{k=x_2+1}^{n} C_n^k \times p^k \times (1-p)^{n-k} \le \alpha/2$$

В точности уровень значимости может оказаться существенно меньше α из-за дискретности распределения. Интервал $[x_1, x_2]$ образует область принятия гипотезы H_c на уровне α: наблюдаемое количество x появлений события A в серии из n испытаний согласуется с предположением о том, что вероятность появления события A в одном испытании равна p, если $x_1 \le x \le x_2$. Если $x < x_1$ или $x > x_2$, мы отвергаем гипотезу H_c . При этом вероятность ошибочно отвергнуть эту гипотезу равна точному уровню значимости и не превосходит α .

Характеристики связи (зависимости) случайных величин

Если с.в. X и Y характеризуются следующими параметрами: математическими ожиданиями m_X, m_Y и дисперсиями σ^2_X, σ^2_Y, то

Ковариация с.в. X и Y (корреляционный момент) k_{XY} –

числовая характеристика распределения случайного вектора (X,Y).

ковариация $k_{XY} = M[(X-M_X)\cdot(Y-M_Y)]$

Коэффициент корреляции с.в. X и Y (ρ, характеристика линейной связи)

$$\rho = \frac{k_{XY}}{\sigma_X \sigma_Y}$$

Выборочная ковариация

$$k_{XY} = \frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}$$

Выборочный коэффициент корреляции

Шкала Чеддока – качественная оценка зависимости переменных.

Величина R	**< 0.3**	**0.3 – 0.5**	**0.5 – 0.7**	**0.7 – 0.9**	**> 0.9**
Характеристика связи	слабая	умеренная	заметная	высокая	очень высокая

Непараметрические меры связи.

Коэффициенты ранговой корреляции

Эти показатели предназначены для измерения силы связи между рангами (порядковыми местами в вариационном ряду) двух показателей. При этом даже те показатели, которые были измерены в количественной шкале, приводятся к порядковым. Предполагается, что оба показателя измеряются по крайней мере в порядковой шкале.

Коэффициент Спирмена ρ_S

Для выборки объема n коэффициент ρ_S между переменными (показателями) x и y вычисляется с помощью меры различия S_ρ.

$$S_\rho = \sum_{i=1}^{n} (r_i - s_i)^2 ,$$

где r_i и s_i – порядковые номера (ранги) i-го элемента в вариационных рядах каждого показателя в отдельности.

$$\rho_S = 1 - \frac{6 \times (S_\rho + B_x + B_y)}{n^3 - n} ,$$

где B_x и B_y – поправки на объединение рангов (при совпадении значений в вариационном ряду).

$$B_x = \frac{1}{12} \times \sum_{i=1}^{m} n_i \times (n^2{}_i - 1) ,$$

где m – количество различных вариант в вариационном ряду (групп объединенных рангов), n_i – количество наблюдений, соответствующих i-му варианту (рангу). Аналогично вычисляется B_y. Коэффициент ранговой корреляции Спирмена может принимать значения от –1 до 1.

Коэффициент Кендалла τ

Для выборки объема n коэффициент τ между переменными (показателями) x и y вычисляется с помощью меры различия S_τ.

$$S_\tau = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \operatorname{sgn}(r_j - s_i) ,$$

где r_i и s_i – порядковые номера (ранги) i-го элемента в вариационных рядах каждого показателя в отдельности.

$$\tau = \frac{S_\tau}{\sqrt{(\frac{n(n-1)}{2} - B_x)(\frac{n(n-1)}{2} - B_y)}},$$

B_x и B_y – поправки на объединение рангов (при совпадении значений в вариационном ряду).

$$B_x = \frac{1}{2}\sum_{i=1}^{m} n_i (n_i - 1),$$

m – количество различных вариант в вариационном ряду (групп объединенных рангов), n_i – количество наблюдений, соответствующих i-му варианту (рангу). Аналогично вычисляется B_y. Коэффициента ранговой корреляции Кендалла может принимать значения от –1 до 1.

Для умеренно больших значений n ($n > 10$) и коэффициентов τ и ρ_S, не слишком близких к 1, верно приближенное соотношение: $\rho_S \cong 1.5\tau$.

Коэффициент γ

вычисляется по формуле

$$\gamma = \frac{\Pi_s - \Pi_d}{1 - \Pi_t},$$

где Π_s – частота пар значений показателей, у которых ранги согласованы (оба больше или оба меньше); Π_d – частота пар, у которых ранги рассогласованы; Π_t – частота пар с совпадающими рангами хотя бы по одному из показателей.

Коэффициент γ особенно рекомендуется использовать в случае, когда по каждой переменной имеется значительное количество совпадающих значений. Вообще говоря, при его применении предполагается, что показатели x и y дискретны по существу (измерены в шкале порядка, а не интервалов).

При вычислении τ, γ и ρ используются поправки на повторяемость значений в вариационном ряду. Большое количество повторяющихся значений переменной может сильно исказить получаемые результаты, особенно для коэффициентов Спирмена и Кендалла. В общем случае коэффициент Кендалла считается более строгой оценкой связи показателей, нежели коэффициент Спирмена.

При вычислении ранговых статистик можно использовать как вариационные ряды, так и таблицы сопряженности. При использовании

таблиц сопряженности формулы для вычислений записывают с использованием матричных обозначений, и они имеют другой вид.

Коэффициенты связи между качественными переменными

Эти коэффициенты предназначены для измерения силы связи между показателями, для которых не определен порядок на множестве их значений. При этом даже те показатели, которые были измерены в порядковой или интервальной шкалах, приводятся к номинальным. Коэффициенты вычисляются с использованием статистики Пирсона X^2 (χ^2) и статистическая значимость этих коэффициентов также определяется статистикой Пирсона.

Проверка независимости двух показателей с помощью критерия χ^2

Проверка гипотезы о независимости двух показателей, где один имеет r значений, а второй – l значений, производится по таблице сопряженности r×l.

Статистика критерия – мера отклонения наблюдаемых частот от ожидаемых

$$\chi^2_{в}=\Sigma_{i=1}^{r}\Sigma_{j=1}^{l}(n_{ij}-\tilde{n}_{ij})^2/\tilde{n}_{ij} \quad ,$$

где n_{ij} – элемент i-ой строки и j-го столбца таблицы сопряженности,

$\tilde{n}_{ij} = n_{i\bullet} \times n_{\bullet j} / n$, n – объем выборки,

$n_{i\bullet}$ - сумма i-ой строки таблицы сопряженности,

$n_{\bullet j}$ - сумма j-ого столбца таблицы сопряженности,

Число степеней свободы $d = (r-1) \times (l-1)$

Гипотеза о независимости показателей принимается на уровне α, если

$$\chi^2_{в} < \chi^2_{1-\alpha}(d)$$

В случае, если гипотеза о независимости отвергается, характеристикой величины связи между показателями может быть один из следующих коэффициентов

Коэффициент Ф

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$ - коэффициент связи показателей.

Коэффициент Ф, в основном, принимает значения из интервала [0, 1]. 0 – нет связи, 1 – сильная связь. Коэффициент может превышать 1.

Коэффициент С

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$ - коэффициент сопряженности (контингации).

Коэффициент С принимает значения из интервала [0, 1]. 0 – нет связи, 1 – сильная связь. 1 достигается только асимптотически. Фактическое максимальное значение коэффициента меньше 1 и зависит от размера и распределения частот таблицы.

Коэффициент V Крамера

$$V = \sqrt{\frac{\chi^2}{n * (q - 1)}}$$, где q = min (r, l)

Коэффициент связанности Крамера V принимает значения из интервала [0, 1]. 0 – нет связи, 1 – сильная связь. 1 достигается только асимптотически.

Коэффициенты связи между дихотомическими (бинарными) переменными

При перекрестной табуляции дихотомических показателей получается квадратная таблица 2x2.

Переменная B / Переменная A	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

Сила связи между бинарными переменными может быть измерена с помощью **относительных и атрибутивных рисков**, а также **отношения шансов** (см.)

Непараметрические критерии однородности выборок

Однородность – это принадлежность двух или более выборок одной и той же генеральной совокупности, т.е. проверка однородности – это проверка гипотезы о совпадении функций распределения. Необходимые условия однородности – равенство характеристик положения и рассеивания, таких, как средние, медианы, дисперсии.

Критерий серий Вальда-Вольфовица

Применяется для сравнения двух независимых выборок. Проверяется нулевая гипотеза об их однородности и, соответственно, совпадении

параметров выборок: средних, медиан, коэффициентов асимметрии и т.д.

Две выборки объемом n1 и n2 объединяются в одну, объединенная выборка сортируется по возрастанию. В отсортированной выборке подсчитывается число серий элементов, относящихся к первой и второй выборкам. При достаточном объеме выборки для определения p-значений используется нормальная аппроксимация.

Критерий Вилкоксона, Манна и Уитни (U или W)

Две статистики, связанные между собой. Применяются для сравнения двух независимых выборок. Проверяется однородность выборок и, в частности, совпадение средних и медиан.

Статистика критерия вычисляется с помощью ранговых сумм каждой из выборок в общем вариационном ряду. Одна из возможных формул для вычисления имеет следующий вид. Пусть R_1 и R_2 – суммы рангов каждой выборки в общем вариационном ряду, n_1 и n_2 – объемы выборок.

$$U_1 = n_1 n_2 + n_1(n_1+1)/2 - R_1$$

$$U_2 = n_1 n_2 + n_2(n_2+1)/2 - R_2$$

$$U = \max(U_1, U_2) \text{ – статистика критерия.}$$

Для определения критических значений при малых объемах выборки ($n_1 + n_2 < 20$) используются таблицы. Если n_1, $n_2 > 8$, может применяться нормальная аппроксимация.

Критерий Вилкоксона для связанных выборок (T)

Применяется для попарно связанных выборок.

Критерием проверяется статистическая значимость нулевой гипотезы о том, что распределение разностей двух выборок симметрично относительно нуля. Вычисления производятся с рангами, а не с самими величинами. При этом предполагается, что величина разностей x_i-y_i имеет смысл, т.е. исследуемый показатель измеряется, по крайней мере, в метрической порядковой шкале.

В статистических пакетах для определения p-значений используется нормальная аппроксимация, поэтому для малых выборок ($n < 10$) не следует его использовать.

Критерий знаков

Применяется для попарно связанных выборок.

Основное предположение об однородности связанных выборок $\{x_i\}$, $\{y_i\}$, i =1,…,n :

$$P(x_i-y_i>0) = P(x_i-y_i<0) = ½,\ i=1,\ldots,n.$$

Нулевые разности имеют нулевую вероятность, поскольку распределения предполагаются непрерывными. Если нули все же встречаются, то соответствующие наблюдения исключаются из рассмотрения.

Статистикой критерия является число положительных разностей среди всех ненулевых (r из k). Проверяется нулевая гипотеза H_0: "r «плюсов» из k наблюдений согласуется с биномиальным распределением с параметром p =1/2."

Критерий не следует использовать при частых совпадениях парных значений x_i, y_i.

Критерий χ^2

Проверка гипотезы об однородности l показателей, каждый из которых имеет r значений, производится по таблице сопряженности r×l.

Статистика критерия – мера отклонения наблюдаемых частот от ожидаемых

$$\chi^2_в=\Sigma_{i=1}^{r}\Sigma_{j=1}^{l}(n_{ij}-\tilde{n}_{ij})^2/\tilde{n}_{ij}\ ,$$

где n_{ij} – элемент i-ой строки и j-го столбца таблицы сопряженности,

$\tilde{n}_{ij} = n_{i\bullet}\times n_{\bullet j}/n$, n – объем выборки,

$n_{i\bullet}$ - сумма i-ой строки таблицы сопряженности,

$n_{\bullet j}$ - сумма j-ого столбца таблицы сопряженности,

Число степеней свободы d = (r–1) × (l–1)

Гипотеза об однородности показателей принимается на уровне α, если

$$\chi^2_в < \chi^2_{1-\alpha}(d)$$

X-критерий Ван дер Вардена

Применяется для сравнения двух независимых выборок. Проверяется однородность выборок. X-критерий сравнивает ранжированные ряды вариант по их центральной тенденции. Предполагается отсутствие совпадающих значений.

Критерий Колмогорова-Смирнова

Применяется для сравнения двух независимых выборок. Проверяется однородность выборок при условии непрерывности их функций распределения, что означает отсутствие совпадающих значений.

Статистика критерия аналогична статистике критерия КС при проверке согласия с известным законом распределения и имеет вид:

$$D_{m,n} = \sup_{-\infty<x<\infty} | F_n(x)-G_m(x) | ,$$

где $D_{m,n}$ – максимальная разность между частотами выборочных рядов объемом m и n.

Параметрические критерии однородности выборок

Точный метод Фишера (Фишера-Ирвина)

Применяется для сравнения двух *дихотомических* независимых выборок. Проверяется нулевая гипотеза о том, что выборки извлечены из генеральных совокупностей, распределенных биномиально, с одинаковой частотой встречаемости изучаемого эффекта.

В первой серии из n_1 испытаний событие А появилось x_1 раз, во второй серии было n_2 испытаний, и событие А появилось x_2 раз.

Нулевая гипотеза: частота появления события в первой серии не отличается от частоты его появления во второй серии.

Точный уровень значимости нулевой гипотезы вычисляется по формуле:

$$\alpha = \frac{n_1!*n_2!*(x_1+x_2)!*(n_1-x_1+n_2-x_2)!}{(n_1+n_2)!*x_1!*x_2!*(n_1-x_1)!*(n_2-x_2)!}$$

Если $\alpha < \alpha_0$, выбранного уровня значимости, то нулевая гипотеза отвергается.

Критерии наличия линейного тренда

Критерий χ^2

Критерий χ^2 применяется к таблице сопряженности 2×k, причем каждой серии поставлена в соответствие дозовая нагрузка x_1, x_2, …, x_k. Линейный тренд – это регрессия пропорций $\{n_{1i} / n_{\bullet i}\}$ на дозы $\{x_i\}$.

Серия / Значение	1	2	…	k	Сумма
А	n_{11}	n_{12}	…	n_{1k}	$n_{1\bullet}$
не А	n_{21}	n_{22}	…	n_{2k}	$n_{2\bullet}$
Сумма	$\mathbf{n_{\bullet 1}}$	$\mathbf{n_{\bullet 2}}$	…	$\mathbf{n_{\bullet k}}$	**N**
Дозовая нагрузка	x_1	x_2	…	x_k	

Для проверки нулевой гипотезы H_0: в пропорциях отсутствует линейный тренд против альтернативной гипотезы $H_т$: есть линейный тренд - вычисляется статистика критерия

$$\chi^2_{в} = \frac{(\sum_{i=1}^{k} x_i * n_{1i} - n_{1\bullet} * \sum_{i=1}^{k} \frac{x_i * n_{\bullet i}}{N})^2}{\frac{n_{1\bullet}}{N} * (1 - \frac{n_{1\bullet}}{N}) * (\sum_{i=1}^{k} n_{\bullet i} * x_i^{\,2} - N * (\sum_{i=1}^{k} \frac{x_i * n_{\bullet i}}{N})^2)}$$

В предположении нулевой гипотезы критерий распределен как $\chi^2(1)$.

Гипотеза об отсутствии тренда принимается на уровне α, если

$$\chi^2_{в} < \chi^2_{1-\alpha}(1)$$

В противном случае гипотеза отклоняется (принимается альтернативная гипотеза о наличии линейного тренда).

Разность $\chi^2_{в}(d-1)$ и $\chi^2_{в}(1)$, распределенная как $\chi^2(d-2)$, используется для проверки значимости отклонения пропорций от линейного тренда.

Тест Кокрана (Cochran) – аналогичный предыдущему тест

В обозначениях предыдущей таблицы

$$\chi^2 = \frac{[\sum_{j=1}^{k} n_{\cdot j}\left(\frac{n_{1j}}{n_{\cdot j}} - \frac{n_{1\cdot}}{N}\right)\cdot(x_j - \tilde{x})]^2}{\sum_{j=1}^{k} n_{\cdot j}(x_j - \tilde{x})} \cdot \frac{N^2}{n_{1\cdot} n_{2\cdot}}, \text{ где } \tilde{x} = \frac{\sum_{j=1}^{k} n_{\cdot j} x_j}{N}$$

В предположении нулевой гипотезы критерий тоже распределен как $\chi^2(1)$.

Тест Армитэйджа (Armitage) проверки тренда в пропорциях

Критерий применяется к таблице сопряженности 2×k. Он позволяет проверить, есть ли линейный тренд в пропорциях осуществления события (безотносительно к дозовым нагрузкам). В обозначениях предыдущей таблицы определим

$$A = \sum_{i=1}^{k-1} n_{2i} \cdot \sum_{j=i+1}^{k} n_{1j}\ ,\ B = \sum_{i=1}^{k-1} n_{1i} \cdot \sum_{j=i+1}^{k} n_{2j}$$

Тогда статистика критерия S = A – B, а оценка стандартной ошибки S: $V = \frac{n_{1\cdot} n_{2\cdot} (N^3 - \sum_{i=1}^{k} n_{\cdot i}^{\,3})}{3N(N-1)}$

Статистика теста S стандартизуется к нормальному распределению:

$$z = \frac{S}{\sqrt{V}}$$

Полученное z-значение далее проверяется с помощью стандартного нормального распределения.

Риски

Риском называется вероятность возникновения неблагоприятного исхода, она принимает значения в интервале от 0 (риск отсутствует) до 1 (неблагоприятный исход наступит наверняка). В качестве меры связи некоторого фактора с риском возникновения события (частотой) используют относительный риск, атрибутивный риск или отношение шансов. Основой для вычисления этих мер связи является таблица 2x2:

	Уровни изучаемого фактора F	
	F_1	**F_2**
Событий (случаев)	a_1	a_2
Наблюдений	c_1	c_2
Частота события	p_1	p_2

Риск наступления события на каждом уровне фактора F обозначается p_i и может вычисляться тремя способами.

1) $p_i = a_i/c_i$, и при этом во второй строке таблицы – количество наблюдений. Тогда p_i называется **пропорцией**

2) $p_i = a_i/c_i$, и при этом во второй строке таблицы – количество «человеко-лет наблюдений». Тогда p_i называется **уровнем**

3) $\tilde{p}_i = \frac{a_i}{c_i - a_i}$, во второй строке таблицы – количество наблюдений. Тогда $\tilde{p}_i$ называется **шансами события** - отношение числа «случаев» к числу «не случаев». Эта характеристика чаще всего используется при вычислении рисков для редких событий.

Относительный риск RR (relative risk) – это отношение $\frac{p_2}{p_1}$, где p_i – пропорции или уровни.

$$RR = R_{21} = \frac{a_2 \times c_1}{c_2 \times a_1}$$

Атрибутивный риск (attributable risk, **AR**) – разность пропорций или уровней

$$\mathrm{AR} = \mathrm{p}_2 - \mathrm{p}_1 = \frac{a_2}{c_2} - \frac{a_1}{c_1}$$

RR и AR связаны соотношением:

$$\mathrm{AR} = \mathrm{p}_1 \times (\mathrm{RR}-1)$$

Отношение шансов (odds ratio) обычно обозначается символом **OR** и вычисляется как

$$OR = OR_{21} = \frac{a_2 \times (c_1 - a_1)}{(c_2 - a_2) \times a_1}$$

Стандартные ошибки и доверительные интервалы для рисков

p_i - пропорция

Если строка «наблюдений» означает количество объектов наблюдения, то есть p_i – *пропорция (частота)*, то распределение числа событий моделируется биномиальным распределением, и дисперсия p_i определяется формулой

$$VAR(p_i) = S^2(p_i) = \frac{p_i \times (1 - p_i)}{c_i},$$

тогда стандартная ошибка пропорции p_i

$$\mathrm{SE}(\mathrm{p_i}) = S(p_i) = \sqrt{\frac{p_i \times (1 - p_i)}{c_i}},$$

$$\ln(R_{21}) = \ln(p_2) - \ln(p_1),$$

Поэтому дисперсия $\ln(R_{21})$ вычисляется как сумма дисперсий

$$\mathrm{VAR}(\ln R_{21}) = \mathrm{S}^2(\ln R_{21}) = \frac{1 - \mathrm{p}_2}{\mathrm{a}_2} + \frac{1 - \mathrm{p}_1}{\mathrm{a}_1}.$$

Соответственно, стандартная ошибка логарифма относительного риска в этом случае

$$\mathrm{SE}(\ln R_{21}) = \mathrm{S}(\ln R_{21}) = \sqrt{\frac{1-p_2}{a_2} + \frac{1-p_1}{a_1}} = \sqrt{\frac{1}{a_2} + \frac{1}{a_1} - \frac{1}{c_2} - \frac{1}{c_1}}$$

Так как $\ln R_{21}$ распределен асимптотически нормально, то

95% доверительный интервал для $\ln R_{21}$

$$\{\ln R_{21} - 1.96 \bullet \mathrm{SE}(\ln R_{21}),\ \ln R_{21} + 1.96 \bullet \mathrm{SE}(\ln R_{21})\}$$

Тогда **95% доверительный интервал для R_{21}** имеет вид:

$$\left(R_{21}^{1-1.96\times\sqrt{\frac{1-p_2}{a_2}+\frac{1-p_1}{a_1}}},\ R_{21}^{1+1.96\times\sqrt{\frac{1-p_2}{a_2}+\frac{1-p_1}{a_1}}}\right) \qquad (\mathrm{I})$$

p_i - уровень

В строке «наблюдений» - человеко-годы наблюдения за период (общее время под риском). Тогда p_i – *уровень* по содержанию, распределение числа событий моделируется распределением Пуассона, и

$$SE(p_i) = S(p_i) = \frac{p_i}{\sqrt{a_i}}.$$

$$VAR(\ln p_i) = S^2(\ln p_i) = \frac{1}{p_i^{\,2}} \times \mathrm{var}(p_i) = \frac{1}{p_i^{\,2}} \times \frac{p_i}{c_i} = \frac{1}{a_i},$$

отсюда

$$VAR(\ln R_{21}) = S^2(\ln R_{21}) = \frac{1}{a_2} + \frac{1}{a_1},$$

$$SE(\ln R_{21}) = S(\ln R_{21}) = \sqrt{\frac{1}{a_2} + \frac{1}{a_1}}.$$

95% доверительный интервал для R_{21} имеет вид:

$$\left(R_{21}^{1-1.96\times\sqrt{\frac{1}{a_2}+\frac{1}{a_1}}},\ R_{21}^{1+1.96\times\sqrt{\frac{1}{a_2}+\frac{1}{a_1}}}\right) \qquad (\mathrm{II})$$

p_i - шансы события

В строке «наблюдений» количество объектов наблюдения. Шансы используются при исследованиях «случай – контроль» или при

изучении редких событий. Вместо частоты p_i в этом случае вычисляются *шансы* (осуществления события в группе):

$$\tilde{p}_i = \frac{a_i}{c_i - a_i}, \tilde{p}_i = \frac{p_i}{1 - p_i}$$ Это выражение называется *логитом* p_i.

Тогда отношение шансов OR_{21} есть отношение логитов: $OR_{21} = \frac{\tilde{p}_2}{\tilde{p}_1}$.

Для редких событий величина $\tilde{p}_i$ практически не отличается от пропорции p_i, но ошибка в этом случае вычисляется иначе:

$$VAR(\ln \frac{p_i}{1 - p_i}) = \frac{1}{c_i \times p_i \times (1 - p_i)},$$

$$VAR(\ln OR_{21}) = \frac{1}{c_2 \times p_2 \times (1 - p_2)} + \frac{1}{c_1 \times p_1 \times (1 - p_1)} =$$

$$\frac{1}{a_2} + \frac{1}{c_2 - a_2} + \frac{1}{a_1} + \frac{1}{c_1 - a_1}$$

$$SE(\ln OR_{21}) = \sqrt{\frac{1}{a_2} + \frac{1}{c_2 - a_2} + \frac{1}{a_1} + \frac{1}{c_1 - a_1}}$$

95% доверительный интервал для OR_{21} имеет вид:

$$(OR_{21}^{1-1.96\times\sqrt{\frac{1}{a_2}+\frac{1}{c_2-a_2}+\frac{1}{a_1}+\frac{1}{c_1-a_1}}}, OR_{21}^{1+1.96\times\sqrt{\frac{1}{a_2}+\frac{1}{c_2-a_2}+\frac{1}{a_1}+\frac{1}{c_1-a_1}}})$$ (III)

Сравнивая формулы для доверительных интервалов (I), (II) и (III), можно отметить, что самый «узкий интервал» соответствует выражению (I), шире интервал для случая (II), и еще шире для отношения шансов (III).

Объединенные риски при наличии мешающих факторов

Весь анализ проводится для двух уровней изучаемого фактора и К (K>=2) уровней мешающего фактора. Если изучаемый фактор имеет более двух уровней, весь анализ нужно повторять для каждой пары уровней.

В случае, когда изучаемый фактор имеет ровно два уровня, предполагается, что уровень **F_1** соответствует экспонированности

изучаемым фактором (есть влияние фактора F), $\mathbf{F_2}$ – отсутствие влияния фактора F. Если уровней более двух, то самый высокий уровень экспонированности обозначается $\mathbf{F_1}$, и далее в порядке убывания.

Уровни связанного фактора U	**Содержание таблиц**	**Уровни изучаемого фактора F**	
		$\mathbf{F_1}$	$\mathbf{F_2}$
$\mathbf{U_1}$	Событий	a_{11}	a_{12}
	Наблюдений	c_{11}	c_{12}
$\mathbf{U_2}$	Событий	a_{21}	a_{22}
	Наблюдений	c_{21}	c_{22}
…			
$\mathbf{U_K}$	Событий	a_{K1}	a_{K2}
	Наблюдений	c_{K1}	c_{K2}

Каждому уровню i фактора U: U_1, U_2, …, U_K, - соответствует таблица T_i.

Таблица T_i :

a_{i1}	a_{i2}
c_{i1}	ci_2

Статистика (риск) Мантеля-Ханзела для сравнения уровней факторов F_1 и F_2 вычисляется по формуле:

$$R_{MH} = \frac{\sum_{i=1}^{K} \frac{a_{i1} \times (c_{2i} - a_{2i})}{n(i)}}{\sum_{i=1}^{K} \frac{a_{2i} \times (c_{1i} - a_{1i})}{n(i)}}$$

Для проверки гипотезы H_0: $R_{MH}=1$ (все составляющие риски OR(i)=1 против альтернативной гипотезы H_1: хотя бы один из этих рисков отличен от 1) используется статистика

$$\chi^2{}_{в} = \chi^2{}_{mhcc} = \frac{(l-0.5)^2}{m}, \text{ где}$$

$$l = \sum_{i=1}^{K} \frac{a_{i1} \times c_{i2} - a_{i2} \times c_{1i}}{n(i)},$$

$$m = \sum_{i=1}^{K} \frac{(a_{i1} + a_{i2}) \times (c_{i1} - a_{i1} + c_{i2} - a_{i2}) \times c_{i1} \times c_{i2}}{n^2(i) \times (n(i) - 1)}$$

Эта статистика аппроксимируется распределением χ^2 (1). Она называется статистикой Мантеля-Ханзела с поправкой на непрерывность (χ^2_{mhcc}), в отличие от статистики Мантеля-Ханзела χ^2_{mh}, которая отличается от χ^2_{mhcc} отсутствием поправки на непрерывность 0.5:

$$\chi^2_{в} = \chi^2_{mh} = \frac{l^2}{m}$$

95% доверительный интервал для R_{mh}

(a) На основе статистики Мантеля-Ханзела χ^2_{mh}

$$R_{mh,нижн.} = \exp[(1 - \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mh}}}) \times \ln(R_{mh})]$$

$$R_{mh,верхн.} = \exp[(1 + \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mh}}}) \times \ln(R_{mh})],$$

где α – уровень значимости, в данном случае 0.05, $z_{\alpha/2}$ – соответствующая процентная точка стандартного нормального распределения.

(b) С учетом поправки на непрерывность доверительный интервал имеет вид:

$$R_{mhc,нижн.} = \exp[(1 - \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mhcc}}}) \times \ln(R_{mh})]$$

$$R_{mhc,верхн.} = \exp[(1 + \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mhcc}}}) \times \ln(R_{mh})]$$

(c) Вычисление доверительных интервалов методом Робинса.

$$R_{mhR,нижн.} = \exp(\ln(R_{mh}) - z_{\alpha/2} \times \sqrt{V})$$

$$R_{mhR,верхн.} = \exp(\ln(R_{mh}) - z_{\alpha/2} \times \sqrt{V}),$$

Где $$V = \frac{\sum_{i=1}^{K} P_i \times R_i}{2 \times (\sum_{i=1}^{K} R_i)^2} + \frac{\sum_{i=1}^{K} (P_i \times S_i + Q_i \times R_i)}{2 \times \sum_{i=1}^{K} R_i \times \sum_{i=1}^{K} S_i} + \frac{\sum_{i=1}^{K} Q_i \times S_i}{2 \times (\sum_{i=1}^{K} S_i)^2},$$

$$P_i = \frac{a_{i1} + c_{i2} - a_{i2}}{n(i)}, R_i = \frac{a_{i1} \times (c_{i2} - a_{i2})}{n(i)},$$

$$Q_i = \frac{a_{i2} + c_{i1} - a_{i1}}{n(i)}, S_i = \frac{a_{i2} \times (c_{i1} - a_{i1})}{n(i)}$$

Еще один способ вычисления объединенного риска был предложен Вульфом. С помощью объединенного относительного риска Вульфа вначале проверяется взаимодействие факторов U и F (анализ однородности таблиц). Для этого используется критерий χ^2 для статистики, связывающей риски в стратах и объединенный риск Вульфа.

Предполагая, что в строке «Наблюдений» у нас количество наблюдений, а не человеко-годы наблюдения, т.е. риск возникновения события оценивается с помощью пропорций или шансов, но не уровней, для каждой из таблиц вычислим следующие величины:

4. Отношение шансов OR(i), или перекрестное произведение.
5. Весовые коэффициенты таблиц для вычисления взвешенного риска W(i). Эти коэффициенты обратно пропорциональны дисперсиям ошибок для каждой таблицы.
6. Объем выборки n(i).

Объем вычисляется как сумма числа наблюдений по столбцам: $n(i) = c_{i1} + c_{i2}$

Если выполнено следующее условие:

► $a_{i1}>0$, $a_{i2}>0$, $c_{i1}-a_{i1}>0$, $c_{i2}-a_{i2}>0$, то

$$OR(i) = \frac{a_{i2} \times (c_{i1} - a_{i1})}{a_{i1} \times (c_{i2} - a_{i2})},$$

$$W(i): \frac{1}{W(i)} = \frac{1}{a_{i1}} + \frac{1}{c_{i1} - a_{i1}} + \frac{1}{a_{i2}} + \frac{1}{c_{i2} - a_{i2}}$$

В правой части – квадрат стандартной ошибки логарифма отношения шансов - $SE^2(\ln OR(i))$, т.е. веса обратно пропорциональны квадратам стандартных ошибок.

Если же хотя бы одно из чисел a_{i1}, a_{i2}, $c_{i1} - a_{i1}$, $c_{i2} - a_{i2}$ равно 0, то все значения в ячейках увеличиваются для вычислений на 0.5 (в некоторых программах допускается увеличение на другое малое число, например, 0.25):

$$OR(i) = \frac{(a_{i2} + 0.5) \times (c_{i1} - a_{i1} + 0.5)}{(a_{i1} + 0.5) \times (c_{i2} - a_{i2} + 0.5)},$$

$$\frac{1}{W(i)} = \frac{1}{a_{i1} + 0.5} + \frac{1}{c_{i1} - a_{i1} + 0.5} + \frac{1}{a_{i2} + 0.5} + \frac{1}{c_{i2} - a_{i2} + 0.5}$$

Для вычисления объединенного взвешенного риска R_w (Вульфа) сначала вычисляется логарифм R_w как взвешенная комбинация логарифмов рисков OR(i):

$$\ln(R_w) = \frac{\sum_{i=1}^{K} W(i) \times \ln OR(i)}{\sum_{i=1}^{K} W(i)}$$

Логарифм объединенного риска распределен асимптотически нормально. Стандартная ошибка логарифма объединенного взвешенного риска R_w определяется весовыми коэффициентами W(i).

$$\mathrm{SE}\ln(\mathrm{R_w}) = \sqrt{\frac{1}{\sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{W(i)}}}$$

Поэтому 95% доверительный интервал для объединенного риска выглядит следующим образом.

$$\{\exp[\ln(R_w) - \frac{1.96}{\sqrt{\sum_{i=1}^{K} W(i)}}], \exp[\ln(R_w) + \frac{1.96}{\sqrt{\sum_{i=1}^{K} W(i)}}]\}$$

В некоторых работах (в частности, J.F.Osborn. Basic Statistical Methods for Epidemiological Studies) использование взвешенного риска предполагается в случае, когда для измерения взаимодействия факторов

используется относительный риск RR, т.е. используются пропорции или уровни для оценки частоты наблюдаемого явления.

Для проверки однородности таблиц используется статистика χ^2, ее выборочное значение вычисляется как

$$\chi^2_в = \sum_{i=1}^{K} W(i) \times (\ln OR(i))^2 - (\sum_{i=1}^{K} W(i)) \times (\ln R_w)^2,$$ и она распределена асимптотически как χ^2(K-1).

Стандартизация

Группы мешающего параметра	**Изучаемая популяция**			**Стандартная популяция**		
	Кол-во объектов под риском	Кол-во случаев	Уровень	Кол-во объектов под риском	Кол-во случаев	Уровень
1	n_1	r_1	p_1	N_1	R_1	P_1
2	n_2	r_1	p_2	N_2	R_2	P_2
…						
k	n_k	r_k	p_k	N_k	R_k	P_k
Всего	n	r	p	N	R	P

Прямая стандартизация

С использованием обозначений таблицы стандартизованный уровень

$$p_{n.cm.} = \sum_{i=1}^{k} \frac{N_i}{N} \times p_i$$

Стандартная ошибка прямого стандартизованного уровня

$$\text{Ст.ош.}(p_{\text{п.ст.}}) = \sqrt{\sum_{i=1}^{k} \frac{N_i^{\,2} p_i q_i}{N^2 n_i}}$$

Сравнительный индекс (CMI для смертности, CII для первичной заболеваемости)

$$\text{CMI} = \frac{\sum_{i=1}^{k} p_i N_i}{NP}$$

Стандартная ошибка сравнительного индекса (относительного риска)

$$\text{Ст.ош.(CMI)} = \sqrt{\sum_{i=1}^{k} \frac{N_i^{\,2} p_i q_i}{P^2 n_i}}$$

Непрямая стандартизация

С использованием обозначений таблицы стандартизованный уровень

$$p_{н.ст.} = \frac{rP}{\sum_{i=1}^{k} P_i n_i}$$

Стандартная ошибка непрямого стандартизованного уровня

$$\text{Ст.ош.}(p_{\text{н.ст.}}) = \sqrt{\frac{P^2 \sum_{i=1}^{k} n_i p_i q_i}{(\sum_{i=1}^{k} P_i n_i)^2}}$$

При малых p_i можно использовать приближенное выражение

$$\text{Ст.ош.}(p_{\text{н.ст.}}) \approx \frac{стандартиз.уровень}{\sqrt{r}} = \frac{p_{н.ст.}}{\sqrt{r}}$$

Стандартизованное отношение смертности SMR:

$$\text{SMR} = \frac{r}{\sum_{i=1}^{k} P_i n_i}$$

Стандартная ошибка стандартизованного отношения

$$\text{Ст.ош.(SMR)} = \sqrt{\frac{\sum_{i=1}^{k} n_i p_i q_i}{(\sum_{i=1}^{k} P_i n_i)^2}} \approx \frac{SMR}{\sqrt{r}},$$

приближенная формула для малых значений p_i .

Логистическая регрессия

Логистическая регрессия для бинарного отклика

Линейная логистическая модель для зависимой переменной Y, имеющей два значения (y_1=0 и y_2=1), и независимых переменных X_1, … , X_p произвольной природы имеет вид: вероятность (Prob) того, что Y принимает значение 1,

$$\text{Prob}\,(Y=1) = \frac{1}{1+\text{Exp}(-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p))}$$

Здесь $\{\beta_i\}$ – логистические регрессионные коэффициенты. Их оценки обозначаются $\{b_i\}$.

Поскольку Prob (Y = 0) = 1 – Prob (Y = 1), эта модель может быть записана также и следующим – линейным – образом:

$$\text{Ln}\left(\frac{\text{Prob}\,(Y=1)}{\text{Prob}\,(Y=0)}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p$$

Левая часть этого выражения называется *логит-преобразованием* вероятности или еще *логарифмом отношения шанс*ов.

Обозначим Prob(Y=1) = p. Тогда 1-p – вероятность отрицательного отклика (Y=0). Отношение p/(1-p) есть шансы события, а логитом называется логарифм шансов.

$$l = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Логистическое преобразование является обратным к логит-преобразованию и позволяет определить значение p по значению l.

$$p = \text{logistic}(l) = \frac{e^l}{1+e^l}$$

Общая логистическая регрессия и логит-модели

В общем виде рассматривают множественную логистическую регрессию для описания дискретной зависимой переменной с конечным числом (2 и более) значений. Множественная логистическая регрессия представляет дискретную переменную Y, имеющую G ($G \geq 2$) значений$\{Y_1, Y_2,\ldots,Y_G\}$ через набор из p независимых переменных X_1, X_2, …, X_p.

Обозначим множество независимых переменных $X = (X_1, X_2, \ldots, X_p)$, а наборы соответствующих всем значениям зависимой переменной параметров β обозначим

$$B_g = \begin{pmatrix} \beta_{g1} \\ \dots \\ \beta_{gp} \end{pmatrix}$$

Логистическая регрессионная модель определяется фактически G-1 уравнением (g = 2,…, G):

$$\ln\left(\frac{p_g}{p_1}\right) = \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p =$$

$$= \ln\left(\frac{P_g}{P_1}\right) + XB_g \quad (1)$$

p_g – это вероятность того, что наблюдение, для которого независимые переменные имеют значения X_1, X_2, …, X_p,, относится к группе g, т.е. зависимая переменная Y принимает значение Y_g

$$p_g = \text{Prob}(Y = Y_g \mid X)$$

Обычно $X_1 \equiv 1$ (т.е. в модель включено пересечение, или свободный член), но это не обязательно. Величины P_1, P_2, … , P_G - это априорные вероятности групп. Если они предполагаются равными, тогда первый член в уравнениях $\ln(P_g / P_1)$ равен 0 и исключается из уравнения. Если эти вероятности не предполагаются равными, они изменяют значения свободного члена в логистическом регрессионном уравнении.

Первая группа называется референтной (reference). Выбор референтной группы произвольный. Обычно это наибольшая группа или контрольная группа, с которой сравниваются все остальные группы. Регрессионные коэффициенты β_{11}, β_{12},…, β_{1p} для референтной группы равны 0.

$\{\beta_{ij}\}$ – это множество регрессионных коэффициентов (неизвестных), которые требуется оценить по имеющимся данным. Эти оценки обозначаются $\{b_{ij}\}$.

Уравнения (1) линейны относительно логитов p. Но в терминах вероятностей они не являются линейными. Эта форма уравнений выглядит следующим образом

$$p_g = \text{Prob}\left(Y = Y_g \mid X\right) = \frac{e^{XB_g}}{1 + e^{XB_2} + \dots + e^{XB_G}} \quad (2)$$

В этом уравнении учтено, что $e^{XB_1}=1$, поскольку все регрессионные коэффициенты здесь нулевые.

Решение уравнений правдоподобия

Перепишем уравнения (2) в виде

$$\pi_{gj} = \mathbf{Prob}\left(Y = Y_g \mid X_j\right) = \frac{e^{X_j B_g}}{e^{X_j B_1} + e^{X_j B_2} + \ldots + e^{XB_G}} = \frac{e^{X_j B_g}}{\sum_{j=1}^{G} e^{X_j B_g}}$$

Тогда для выборки из N наблюдений отношение правдоподобия имеет вид

$$l = \prod_{j=1}^{N} \prod_{g=1}^{G} \pi_{gj}^{y_{gj}} \quad (3)$$

Здесь y_{gj} равно 1, если j–е наблюдение относится к группе g, и 0 в противном случае.

Оценки максимального правдоподобия параметров $\{\beta_{ij}\}$ получаются с помощью нахождения точки экстремума логарифма этого выражения. Решение системы получившихся уравнений производится итерационным методом Ньютона – Рапсона.

Статистические критерии и доверительные интервалы

Для проверки значимости одной и более независимых переменных в логистической регрессии используются две процедуры: тест отношения правдоподобия и тест Вальда. Как правило, первая процедура более адекватна. Тест Вальда используется, в основном, из-за простоты вычислений.

Отношение правдоподобия и отклонение (deviance)

Статистика теста отношения правдоподобия (LR) представляет собой разность отношений правдоподобия для двух моделей (полной и частичной), умноженную на число (-2) для того, чтобы распределение статистики аппроксимировалось распределением χ^2.

$$LR = -2[L_{частичн.} - L_{полная}]$$

Отклонение (D) – это статистика LR для случая, когда полная модель является насыщенной (включает члены всех порядков). ΔD - изменение отклонения из-за включения или исключения одной или нескольких переменных, - в логистической регрессии используется так же, как F-статистика в множественной регрессии. Оно распределено асимптотически как χ^2. Используется для тестирования значимости регрессионных коэффициентов, связанных с отдельной независимой переменной.

Статистика Вальда используется для проверки значимости отдельных регрессионных коэффициентов. Формула для статистики Вальда:

$$z_j = b_j / s_{bj},$$

где s_{bj} – оценка стандартной ошибки b_j , она задается корнем квадратным из соответствующего диагонального элемента ковариационной матрицы $V(\tilde{\beta})$.

При больших объемах выборки эта статистика хорошо аппроксимируется нормальным распределением, при малых или средних объемах – «адекватно».

Доверительные интервалы

Доверительные интервалы для регрессионных коэффициентов основаны на статистике Вальда. Формула для границ $100(1-\alpha)\%$ двустороннего доверительного интервала $b_j \pm |z_{\alpha/2}| s_{bj}$

R^2

$$R_L^2 = (L_p - L_0)/(L_0 - L_S),$$

где L_0 – логарифм правдоподобия для модели, в которую включен только свободный член, L_p – логарифм правдоподобия для модели, включающей независимые переменные, и L_S – логарифм правдоподобия для насыщенной модели. Введение параметра L_S необходимо для того, чтобы величина R_L^2 была в пределах от 0 до 1, поскольку L_p варьируется в пределах от L_0 до L_S, однако это вносит некоторую неясность в данную характеристику, поскольку величина L_S зависит от конфигурации независимых переменных. Если $R_L^2 = 1$, это означает лишь, что получена максимально возможная подгонка данных при использовании выбранных независимых переменных. R^2 в логистической регрессии меняется в зависимости от того, какие переменные были включены в насыщенную модель. Поэтому ряд исследователей предлагает отказаться от этого показателя при оценке качества модели.

Анализ остатков

Анализ остатков позволяет найти выбросы, определить соответствие данным выбранной логистической модели.

Для анализа остатков используются: остатки Пирсона, остатки отклонения, диагональ матрицы h.

Статистика DFбэта (DFbeta)

Применяется для изучения влияния отдельных наблюдений на каждый регрессионный коэффициент. Эта статистика – стандартизованная разность регрессионного коэффициента до и после исключения j-го наблюдения.

Расстояние Кука: C и Cbar

Это расширение расстояния Кука для логистической регрессии. Так же измеряет влияние отдельных наблюдений на регрессионные коэффициенты.

Статистики DFотклон. и DFχ^2 (DFDEV и DFCHI2)

Эти статистики измеряют изменения отклонения и статистики хи-квадрат Пирсона, соответственно, при удалении одного наблюдения. Большие значения статистик позволяют обнаружить наблюдения, которые недостаточно хорошо описаны моделью.

Предсказанные вероятности

Здесь описано, как вычислить предсказанные вероятности принадлежности к группе и соответствующие доверительные интервалы.

Основное уравнение (1) может быть представлено, если включить априорные вероятности в свободный член, следующим образом:

$$ln\left(\frac{p_g}{p_1}\right) = \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p = XB_g$$

После получения оценок регрессионных коэффициентов, $\hat{B}_g$, оценкой левой части уравнения для набора значений независимых переменных X будет $l_j = ln\left(\frac{p_g}{p_1}|X_j\right) = X_j\hat{B}_g$.

Доверительные интервалы для логитов можно получить в предположении, что регрессионные коэффициенты имеют асимптотически нормальное многомерное распределение. Однако их невозможно корректно преобразовать в доверительные интервалы для предсказанных вероятностей.

Логлинейная модель (LLM)

Логлинейные модели позволяют изучать соотношения между двумя и более дискретными переменными. Это метод многомерного анализа частот.

Обозначения

Для таблицы с двумя входами, у которой переменная строк A имеет I категорий (уровней) i=1,…, I, а переменная столбцов B имеет J категорий j=1,…, J, точная мультипликативная модель, определяющая частоты в ячейках f_{ij}, записывается как

$$m_{ij} = N\alpha_i\beta_j\gamma_{ij} \quad (1)$$

Здесь $m_{ij} = E(f_{ij})$ – ожидаемая частота в строке i и столбце j. Если m_{ij} оцениваются с использованием метода максимального правдоподобия, результат обозначается $\tilde{m}_{ij}$. Заметим, что $N = \Sigma_{i,j}f_{ij}$.

Основной вопрос, связанный с таблицей: являются ли независимыми **A** и **B**. Это можно проверить с помощью соответствующего теста χ^2. В модели (1) независимость будет установлена, если все γ_{ij} будут равны 1.

Для приведения формулы (1) к аддитивному виду проводится логарифмирование, после чего получим

$$\ln(m_{ij}) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad (2)$$

Слагаемые $\lambda_{\cdot}$ называются эффектами. Верхние индексы обозначают переменные, нижние индексы – категории этих переменных. Порядок эффекта равен числу переменных в верхнем индексе.

Поскольку полученная формула аддитивна, она называется логлинейной моделью. Из-за логарифмирования в данной модели присутствует ограничение: ни один из m_{ij} не равен 0.

В данной модели общее количество коэффициентов $\lambda_{\cdot}$ составляет 1 + I + J + I*J, что превышает количество частот в ячейках (которое составляет I*J). Если число параметров модели превышает или равно количеству ячеек, такая модель называется насыщенной (saturated). Насыщенная модель точно воспроизводит наблюдаемые частоты.

Проверяя, равны ли определенные параметры $\lambda_{\cdot}$ нулю, мы проверяем различные связи между переменными. Например, проверяя, являются ли все коэффициенты $\{\lambda_{ij}^{AB}\}_{i,\ j}$ нулевыми, мы проверяем независимость переменных **A** и **B**.

Качество подгонки

При выборе из нескольких вариантов моделей следует оценить качество каждой из них. Качество модели определяется качеством подгонки данных и проверяется с использованием одной из двух статистик χ^2:

статистика Пирсона χ^2 : $\chi^2 = 2\,\Sigma_{i,j,k}\,[(f_{ijk} - \tilde{m}_{ijk})^2 / \tilde{m}_{ijk}]$

и статистика максимального правдоподобия

$$G^2 = 2\Sigma_{i,j,k} f_{ijk} \ln (f_{ijk} / \tilde{m}_{ijk})$$

Обе эти статистики распределены как χ^2, когда N велико и ни одна из частот $\tilde{m}_{ij}$ не является малой. Обе статистики имеют n-p степени свободы, где n – количество ячеек таблицы, p – количество параметров в модели, для которой вычислены $\tilde{m}_{ijk}$. В отличие от статистики χ^2 Пирсона, отношение правдоподобия G^2 имеет одно важное свойство – оно является аддитивным для частичных связанных моделей. Это позволяет проверять значимость отдельных членов модели.

С помощью этих статистик проверяется следующее утверждение: отличаются ли статистически значимо от 0 те члены насыщенной модели, которые не включены в текущую модель?