ГБОУ ВПО Саратовский ГМУ им. В.И.Разумовского
Минздрава России

Р.Н. Каримов, Ю.Г. Шварц

СТАТИСТИКА ДЛЯ ВРАЧЕЙ В ПОНЯТНОМ ИЗЛОЖЕНИИ

Руководство

Издательство Саратовского медицинского университета
2014

УДК 61:311(035). 58
ББК51.1(251.1(2)9
К43

В книге рассматриваются основные методы статистической обработки биомедицинских данных нечисловой и числовой природы. Приводится содержательная интерпретация методов одномерной, многомерной статистики, часто применяемых для исследований в области медицины и биологии.

Материал книги изложен в форме, доступной для исследователей, не владеющих сложным математическим аппаратом статистического анализа; иллюстрирован тщательно подобранными примерами, взятыми авторами из многолетней собственной практики решения реальных задач в области медицины и биологии.

Книга рассчитана на аспирантов, врачей и биологов, занимающихся научными исследованиями. Может быть также полезна для анализа социологических, эконометрических и данных из других областей.

Рецензенты:

Н.П.Лямина д-р мед. наук, профессор, заместитель директора ФГБУ «СарНИИК» Минздрава России;

В.Э.Олейников д-р мед.наук., профессор, заведующий кафедрой «Терапия» Медицинского института ФГБОУ ВПО «Пензенский государственный университет».

Одобрено к изданию Редакционно-издательским Советом СГМУ.

JSBN 978-5-7213-0526-9

© Р.Н. Каримов, Ю.Г. Шварц, 2014.
© Саратовский медицинский университет, 2014.

ПРЕДИСЛОВИЕ

Главное знать, что считать, как считать и как интерпретировать результаты.

Дж. Гэллап

Большинство основных положений и практических рекомендаций, на которые мы опираемся в своей врачебной деятельности, получено с использованием тех или иных статистических методов. Реальных успехов в медицинской практике можно достичь только тогда, когда мы опираемся на верные выводы медицинской науки, а эти выводы будут точными лишь тогда, когда наука правильно использует свои инструменты, в том числе один из важнейших – математическую статистику. Однако сохраняется парадоксальная ситуация. Знать патологическую физиологию, биохимию, лабораторную диагностику и т.п. при проведении научных изысканий в медицине считается просто обязательным, так как это – методологическая и инструментальная основа исследования, а знать статистику, которая позволяет сделать выводы, многие считают излишним.

С появлением доказательной медицины возросло понимание актуальности проблем [12, 16, 85], связанных с качественным анализом статистического материала. Очевидно, одним из результатов этого стало то, что вы открыли эту книгу.

Под термином «статистика» в медицине существует так называемая медицинская статистика, описывающая деятельность медицинских учреждений, показателей здоровья населения. Эти вопросы в книге не рассматриваются. В ней идет речь о применении методов математической статистики в медико-биологических исследованиях. Статистика – это наука, изучающая методы сбора, обработки и интерпретации данных числовой и нечисловой природы. Цель статистики – получение осмысленных заключений из подверженных случайному разбросу данных. В книге интерпретация данных рассматривается как существенный аспект.

В медицине можно выделить две категории врачей: практикующие медики (врачи) и ученые медики-исследователи [47].

Врач должен обладать минимальным объемом математической подготовки, необходимой для свободного понимания основ теории вероятностей, а также знаниями элементарной алгебры. Эти знания ему необходимы для оценки публикаций статистических методов в своей узкой специализации. Полученные оценки он может применить в своей практической деятельности для повышения качества работы с пациентами.

Познания учёного-медика (работники кафедр вузов, НИИ, издательств, диссертационных советов т.п.) шире, чем у врача-практика. Свои знания он использует для оценки истинности публикаций и для проведения собственных исследований.

Сравнительно недавно основное применение статистики при обработке медико-биологических данных заключалось в оценке средних и их погрешностей, но реальные многомерные зависимости между наблюдениями требуют применения более сложных разделов математической статистики. Для врача-практика, до предела загруженного работой по своей специальности, всестороннее освоение этих разделов практически нереально.

Главной причиной появления этой книги было желание специалиста по статистике помочь врачам и биологам сделать свои исследования интереснее и получать по-настоящему новые и значимые выводы. Соавтор-врач был привлечен для того, чтобы сделать руководство более понятным не только для специалистов, но и для исследователей с медицинским и биологическим образованием.

Авторы уверены в том, что, овладев лишь основами прикладной статистики, можно получить гораздо более интересные и полезные во всех отношениях заключения, чем при использовании традиционных подходов к обработке данных. При этом, как правило, считается, что данные измеряются в количественных шкалах и имеют нормальное распределение и однородную структуру.

Врачи и биологи зачастую и не подозревают о глубине той бездны, которая отделяет настоящую статистику от реального положения дел с обработкой данных. Об этом можно судить как по большинству публикаций в научных журналах, так и по доступным

нам авторефератам диссертаций. Это не вина врачей, а десятилетиями созревавшая проблема. Для того чтобы ее решить, можно постоянно приглашать специалистов по статистике и поручать им всю работу. В этом случае вы доверяетесь людям, ничего не понимающим в сути изучаемого явления (медицине). Здесь очень высок риск ухода от реального решения задачи. Можно самому (самой) овладеть биомедицинской статистикой либо овладеть ею в той мере, чтобы работать совместно со статистиком. Последний вариант наиболее практичен. Мы все же полагаем, что в любом случае от медика требуются знания хотя бы принципов статистики, без которых даже планирование исследования и сбор данных существенно пострадают.

Однако основные доступные руководства по представлению результатов статистических исследований содержат сведения только наиболее распространенных статистических процедур поперечного анализа и загромождены формулами.

В работах [1, 8, 25, 44–49, 51, 64, 84] отмечается, что диссертации и статьи по статистическому анализу биомедицинских данных содержат большое количество ошибок (врачи плохо владеют методами математической статистики). Причин здесь несколько.

Во-первых, литература по статистике врачам недоступна из-за их низкого уровня математических знаний.

Во-вторых, книги по математической статистике для нематематиков имеют большой объем и написаны по принципу «ничего обо всем», подчас жертвуя корректностью изложения в пользу его доступности.

Третьей причиной является представление о том, что диссертации на соискание ученой степени обязательно должны содержать элементы статистической обработки не потому, что они необходимы, а потому что так принято. Соискатели ученых степеней статистическую обработку имитируют. Такие диссертации являются примером для врачей, начинающих свои собственные исследования, и отсюда некорректное использование статистических методов.

В четвертых, применение пакетов прикладных программ (ППП) в значительной степени облегчает обработку широкого класса экс-

периментальных данных. Простота получения результата таит в себе следующие опасности:

1. Руководства пользователя ППП, как правило, не содержат доступного описания методов предобработки данных, таких как восстановление пропусков, выбор требуемого объема выборки, обнаружение аномальных наблюдений, расслоения неоднородных выборок, особенности проверки адекватности полученной модели.

2. Не уделяется достаточного внимания данным, измеряемым в различных шкалах. Часто можно увидеть книги ППП, в которых методы обработки количественных наблюдений применяются к нечисловым (номинальным, порядковым) выборкам.

Данное руководство не претендует на оригинальность и абсолютную полноту изложения проблем статистических методов. В книге использовались материалы из предыдущих работ авторов, из руководств для пользователей ППП Statgraphics, Statistica [89] и некоторых других. В нем не рассматривается подробно статистический аппарат, используемый в крупных клинических испытаниях, поскольку в этой области работают профессиональные специалисты в области статистики и существует достаточное количество работ.

Желающим пополнить свои знания по математике, теории вероятностей и по теории информации мы рекомендуем замечательную трилогию о математике [71], написанную простым понятным языком, без формул, а также учебное пособие по теории вероятностей [79]. Для детального ознакомления с основными методами биомедицинских исследований рекомендуем книгу [15], а для владеющих работой с прикладными программами - учебник [89].

Авторы

СПИСОК СОКРАЩЕНИЙ

ANOVA	—	дисперсионный анализ
H	—	критерий Крускала-Уоллиса
MANOVA	—	многомерный дисперсионный анализ
R	—	коэффициент корреляции Пирсона
SSE	—	сумма квадратов остатков
SSR	—	сумма квадратов модели общая
SST	—	общая сумма квадратов
ДИ	—	доверительный интервал
ККК	—	коэффициент канонической корреляции
КДФ	—	каноническая дискриминантная функция
МГК	—	метод главных компонент
МНК	—	метод наименьших квадратов
МНМ	—	метод наименьших модулей
ММП	—	метод максимума правдоподобия
МЦВ	—	многократно цензурированная выборка
ППП	—	пакет прикладных программ
ОШ	—	отношение шансов
ОЦВ	—	однократно цензурированная выборка
СВР	—	случайный временной ряд
СКО	—	среднеквадратическое отклонение
СНУ	—	система нормальных уравнений
ХСН	—	хроническая сердечная недостаточность
ТСП	—	таблица сопряженности признаков
ЦВ	—	цензурированная выборка
ЦПТ	—	центральная предельная теорема
Ч.с.с. (df)	—	число степеней свободы

ВВЕДЕНИЕ

Математизация является одной из важнейших тенденций современного развития медицинской науки. Объекты медицинских исследований отличаются большой сложностью и разнообразием протекающих в них процессов, поэтому не представляется возможным указать какой-то компактный математический аппарат, который бы достаточно полно их описывал. Однако можно выделить одно общее свойство, присущее многим медицинским задачам, – это их вероятностно-статистический характер.

Вероятностно-статистический характер имеют как сами фундаментальные биологические законы, так и возникновение различных заболеваний. Все это диктует применение статистических представлений и методов их использования при обработке эмпирического материала.

В последние годы по мере роста возможностей вычислительной техники многомерный статистический анализ постепенно превращается из важного теоретического раздела математической статистики в мощный инструмент исследования в медицине. Методы математической статистики разработаны и изложены во многих фундаментальных работах [1–3, 33, 34, 38, 39, 81]. В силу характера изложения они требуют от врача или биолога значительных затрат времени и сил. Такое положение дел «отпугивает» врачей от изучения и применения современных методов прикладной статистики.

Для успешного претворения статистических методов в медицинскую практику необходимо, чтобы исследователь овладел не только словарем математической статистики, но и знал основные методы представления и предварительной обработки данных. Последние нужны не столько для общения и постановки задач математикам, сколько для выработки взглядов и подходов к обработке данных.

Авторы ставили своей целью изложить подобный материал, с одной стороны, понятным для врачей и биологов языком, а с другой,- не допуская вульгаризации и искажения.

Математические методы расширяют возможность получения наибольшей и исчерпывающей информации о больном, но будет глубоко ошибочным надеяться только на помощь и знания специалистов прикладной математики. Знания основ статистики необходимы врачу

для грамотной формулировки цели и постановки задач как в самом медицинском исследовании, так и для постановки задач перед математиками-статистиками.

Отметим, что применение статистики требует более высокого уровня клинического мышления врача, а изучение статистических методов, безусловно, повышает этот уровень.

В первой главе учебного пособия подробно рассматриваются шкалы измерений, вопросы сбора и виды представления медико-биологических данных в задачах поперечного и лонгитюдального (продольного) анализов [68, 102]. В настоящее время статистические методы широко используются для обработки многомерных данных, измеренных в смешанных шкалах; часто эти данные собраны в различные социологические, биологические и хронологические временные моменты, поэтому в пособии подробно рассматриваются типы представления многомерных исходных данных матрицей типа «объект-признак», случайной векторной переменной, ковариационными и корреляционными матрицами, таблицами сопряженности признаков, матрицами близостей и нечеткими методами представления данных.

Вторая глава посвящена методам классической одномерной статистики и основным задачам теории выводов с четким разделением последней на задачи оценивания и проверки статистических гипотез. На большом числе примеров показано, что для реальных данных условия классической статистики не выполняются. Нужен тщательный анализ входных данных с целью выявления соответствия наблюдений требуемым условиям теории прикладной статистики.

В третьей главе изложены методы предварительного анализа исходных данных: восстановление неигнорируемых пропусков, обнаружение и редактирование аномальных наблюдений, преобразование данных с целью нормализации и стабилизации дисперсии, линеаризации связей и выбора типа шкалы. В заключении главы рассмотрены методы оценки валидности результатов в зависимости от поставленной цели исследования.

Четвертая глава посвящена раскрытию связей между интересующими нас качественными признаками. Принципиально различают два вида связи - причинно-следственную и статистическую. Мы обсуждаем лишь последнюю.

В главе рассматриваются два подхода. Первый – для вычисления меры связи между признаками. Второй метод связан с так называемыми *логлинейными моделями*, которые основаны на представлении распределений с системами «*вкладов*» (*эффектов*), даваемых теми или иными множествами признаков.

В реальной ситуации, кроме ответа на вопрос: «есть или нет прямой связи между исследуемыми переменными», нужна содержательная интерпретация выявленной связи. В решении этой задачи нам помогут медицинские знания и использование ТСП с тремя и более входами. Но еще лучше – многомерный *логлинейный* анализ. Он более эффективен, чем детальная разработка таблиц сопряженности.

Пятая глава посвящена анализу связей между двумя количественными и неоднотипными переменными. Связь между количественными переменными характеризуется одномерной линейной регрессией и коэффициентами парной и частной корреляции Пирсона. Частная корреляция позволяет выявить поведение исследуемой парной связи в присутствии третьей контролируемой переменной, когда связь может оказаться кажущейся, опосредованной. Эти методы тесно связаны между собой, но показатели корреляции и регрессии отражают разные свойства связи переменных. Для проверки наличия связи между двумя переменными используют регрессионный анализ и F критерий. Сам корреляционный анализ больше используется в разведочных целях и менее надежен. Обобщением парной корреляции Пирсона является множественная корреляция и коэффициент конкордации.

Далее исследуется связь между переменными, измеренными в разных шкалах, определяемые следующими коэффициентами: точечно-бисериальная корреляция, бисериальная корреляция, ранговобисериальная корреляция, тетрахорический коэффициент корреляции.

В шестой главе представлен метод множественной регрессии, который в основном используется для тех случаев, когда изучается влияние сразу нескольких факторов на одну зависимую переменную [106]. Подчеркивается то, что метод приспособлен именно для количественных нормально распределенных данных. Предлагают-

ся варианты решения проблемы для ненормально распределенных данных. Описаны шесть постулатов регрессионного анализа.

Значительное внимание уделено возможности отклонения от этих постулатов, обсуждаются наиболее типичные медицинские ситуации, когда эти отклонения есть и нуждаются в анализе и принятии адекватных решений. Однако здесь не все проблемы удается разрешить аналитическими методами, поэтому значительное внимание уделено визуальному анализу графиков остатков.

Рассматриваются множественный коэффициент корреляции, его значение в оценке адекватности регрессионной модели и определении возможности использования уравнения регрессии в прогнозировании. Особое внимание уделено проверке гипотез о наличии связи и роли различных критериев на данном этапе статистического анализа.

В седьмой главе излагается математический аппарат, позволяющий строить модели в медицине и решать задачи логистической регрессии и классификации. Рассматривается ROC-анализ – инструмент для оценки качества построенных логистических моделей.

Метод весьма полезен для принятия клинических решений, когда надо ответить на вопрос «да» или «нет» [103]. При этом мы можем определить значение количественной меры связи и оценить шансы или риски, а при наличии бинарных предикторов отношение шансов удобно не только для оценки силы влияния, но и для сравнения силы и характера влияния различных предикторов. Иными словами, можно оценить, какой из изучаемых предикторов влияет на исход сильнее, и в какой мере сильнее.

В восьмой главе описываются параметрический одно- и многофакторные дисперсионные анализы (ANOVA – MANOVA). ANOVA позволяет ответить на очень важные и частые вопросы. Влияет ли интересующий нас фактор (факторы) на изучаемую количественную переменную? Является ли влияние фактора независимым и не опосредованным? Есть ли взаимодействие между факторами? Непараметрические ANOVA можно широко использовать как для ранговых данных, так и для количественных с любыми видами распределения. Существуют непараметрические критерии для решения почти всех задач, для которых существует классический дисперсионный анализ.

В девятой главе обсуждаются свободные от распределения методы, в том числе и непараметрический дисперсионный анализ, которые можно с уверенностью использовать на практике в силу их применимости в широком диапазоне изменения условий, но, кроме того, в некоторых ситуациях только они и могут быть применены. Рассматриваемые методы хороши еще и тем, что работать с ними часто оказывается проще, чем со стандартными методами, связанными со специальными распределениями. Главное же, чего невозможно избежать, – это потеря информации, заключенной в числовых значениях случайных величин, которые умышленно игнорируют. Непараметрический ANOVA можно широко использовать как для ранговых данных, так и для количественных с любыми видами распределения. Разработаны непараметрические критерии для решения почти всех задач, для которых существует классический дисперсионный анализ.

В десятой главе рассматривается метод канонической корреляции, теория которого базируется на методе главных компонент. Цель этого метода – получение канонических переменных, которые позволяют сжато описать зависимость между двумя векторными случайными переменными, то есть между группами переменных в целом. Основное преимущество метода заключается в переходе от анализа множества парных коэффициентов корреляции между множеством переменных к изучению связи между гораздо меньшим количеством новых переменных, обобщающих в себе исходную информацию.

В главе 11 излагаются основные свойства метода главных компонент (МГК). Из обширного перечня задач, решаемых с помощью МГК, рассматриваются задачи сжатия временных рядов и выбора пространства признаков при классификации многомерных объектов.

Главные задачи метода, которые рассматриваются в МГК, - анализ факторной структуры явления и переход к новым латентным (интегральным) переменным. Иными словами, МГК позволяет произвести некую классификацию и уменьшение пространства признаков. В результате из множества переменных мы получаем небольшое количество новых величин (скрытых, латентных переменных), которые отражают суть изучаемого явления

В двенадцатой главе изложены основные положения факторного анализа. Из всего многообразия методов факторного анализа изложены только современные методы, тесно связанные с МГК. Такой подход не наносит большого ущерба, так как применение остальных методов приводит практически к сходным результатам. Факторный анализ позволяет решать следующие задачи: сокращение числа переменных (редукция данных) признакового пространства в данных типа «объект-признак», преобразование исходных переменных к более удобному для интерпретации виду, классификацию объектов на основе сжатого признакового пространства, косвенную оценку признаков, не поддающихся непосредственному измерению, создание структурных моделей многомерных данных.

Тринадцатая глава посвящена многомерному дискриминантному анализу, который представляет собой статистический метод, предназначенный для изучения различий между двумя и более группами объектов, описываемых ($n \times p$)-матрицей количественных данных типа «объект-признак». При этом для обозначения двух групп используется дихотомическая переменная (два класса), а в случае k групп – полихотомическая несовместная переменная (много классов).

Дискриминантный анализ используется в социологии при решении задач классификации; в медицине для целей дифференциальной диагностики; в психологии для изучения стереотипов в поведении людей; в учебном процессе для определения ожидаемой успеваемости студентов до его прослушивания.

В четырнадцатой главе рассматриваются методы анализа соответствий, которые являются дальнейшим развитием МГК для данных нечисловой природы, представленных в виде таблицы сопряженности признаков (ТСП). В результате применения анализа соответствий мы получаем возможность представить данные строк и столбцов ТСП точками в совместном пространстве и наглядно изучить структуру связей между переменными.

Особенностью пособия является то, что теоретические изложение материала сопровождается рассмотрением тщательно подобранных примеров. В конце каждой главы приводятся задачи, упражнения и темы рефератов для более углубленного изучения

наиболее трудных разделов. Для трудных задач даются в помощь читателю даются краткие указания.

Как читать эту книгу. Это можно делать тремя способами. Для предварительного или поверхностного знакомства со статистикой достаточно прочитать то, что выделяется полужирным и курсивным текстом и несколько углубиться в те разделы, которые для вас представляют особый интерес. Второй способ позволит врачу или биологу изучить основные принципы начальных этапов статистического анализа достаточно подробно. Для этого можно читать все, за исключением формул и текста, напечатанного мелким шрифтом. Эти части рассчитаны в большей мере на тех, кто уже имеет специальную подготовку и собирается изучать статистику на профессиональном уровне. Следовательно, третий способ – читать все – предназначен для будущих специалистов по статистике.

Глава 1

ВИДЫ ИСХОДНЫХ ДАННЫХ

При исследовании любого реального явления, например, организма человека, развития болезни, эффекта от лечения, необходимо выполнить простую, но крайне важную последовательность действий. Прежде всего, нужно построить модель этого явления, а уже потом провести его последующий анализ и оценку. С точки зрения практики познания такой подход является естественным. Действительно, прежде чем произвести первое наблюдение простейшей одномерной величины, например, длительности ремиссии заболевания, вязкости крови или появления мутации гена, нужно указать, каковы природа и свойства этой величины, то есть использовать априорную информацию. Чем полнее априорная информация, тем точнее и с меньшими затратами труда можно получить необходимые данные, поэтому большое значение имеет предварительная формализация методов сбора, предобработки и использования априорной информации. К сожалению, именно этому начальному этапу редко уделяется должное внимание, и именно поэтому дальнейший ход исследования может пойти по ложному пути.

На основе анализа этой априорной информации строится модель изучаемого процесса, определяются типы шкал измеряемых переменных, что не только закладывает основу успеха последующего статистического анализа, но и влияет на дизайн исследования в целом. Кроме того, уже на этапе планирования работы разрабатывается методика обработки экспериментальных данных.

Попробуйте сделать это именно так, а не как обычно, то есть сначала что-то сделать, кого-то полечить, обследовать, а потом вспомнить, что существует статистика.

Основное понятие в системах обработки информации – *данные,* которые подразделяются на два вида: *нечисловые* (*категоризованные, интервальные*) и *числовые* (*количественные*) [6, 61, 66]. От того, насколько сам исследователь это понимает, в каком виде представляется, как все это вносится в компьютер и по каким правилам обрабатывается, зависит не только качество анализа медицин-

ских данных, но также и то, как правильно вы спланируете свое исследование, сколько лишней работы проделаете или не проделаете и сколько раздражения испытаете в конце исследования оттого, что натворили в его начале.

Примерами различного вида данных могут служить: пол, социальное положение, исход операции, температура, вес и рост человека. Данные получаются в результате измерений признаков индивидуумов или подопытных образцов из исследуемой совокупности. Измерение – операция, посредством которой определяется отношение одной (измеряемой) величины к другой однородной величине, принимаемой за единицу [6, 10, 64, 77], то есть тип шкалы измерения. Выбор или оценка используемых в исследовании шкал измерений – одна из основ успеха в науке [100].

1.1.Шкалы измерений

В чем лучше измерять?

Перед тем как исследовать и измерять что-либо или анализировать результаты измерений, надо абсолютно точно понять и решить, какой шкалой вы пользуетесь. Прежде чем оценить окружность талии девушке или больному с ожирением, мы четко решаем, используем ли мы сантиметры, дюймы или соотношение окружности талии к окружности бедер. По сути, мы выбираем шкалу измерений. В зависимости от того, что мы выбрали, будет зависеть то, как мы зафиксируем, как представим и как будем анализировать результаты измерений. Что такое шкалы и что такое тип шкалы – базисные вопросы статистики, и не только.

Тип шкалы определяется группой допустимых преобразований $\mathbf{\Phi} = \{\varphi(x)\}$, переводящих одну систему измерений, являющуюся гомоморфным образом эмпирической системы, в другую, также являющуюся гомоморфным образом этой же эмпирической системы [10, 61]. Для основных типов шкал допустимы преобразования: взаимнооднозначные, монотонные, тождественные, подобия, сдвига и линейные. Чем меньше множество систем, в которых отображается (без потери внутренней структуры и смысла) гомоморфно рассматриваемая эмпирическая система с отношением, тем сильнее шкала.

Тип шкалы также определяет возможности применения к измерениям операций сравнения и арифметических действий. А это значит, что от типа данных и от шкалы, которую вы используете, тесно зависит не только выбор статистических методов, но и характер выводов, которые вы сможете сделать. Кстати, именно в этой области допускается большое количество грубых ошибок. Так, например, определять взаимосвязи между районом проживания и распространенностью анемии с использованием метода корреляции или оценивать влияние нового средства для обезболивания на риск возникновения послеоперационных осложнений, опираясь на коэффициент Стьюдента, просто недопустимо, хотя это делается. Это почти то же самое, что изучать состояние противоинфекционной защиты с помощью электрокардиограммы или определять функциональные резервы организма на основании патологоанатомического вскрытия.

1.1.1. Качественные данные

Как учитывать цвет волос? Жизнь или смерть? Быть или не быть?

Качественные (синоним: категориальные, ординарные, ординальные, номинальные, дихотомические, бинарные, порядковые, ранговые, альтернативные, полихотомические, экзистенциальные) данные в лечебной практике чаще всего получают путем проведения беседы врача с пациентом. Врач задает главным образом открытые, зондирующие вопросы, побуждающие больного к высказыванию и обсуждению своих мыслей и чувств по отношению к своей болезни. Такие данные позволяют понять основные связи между чувствами, мнениями пациентов и их поведением, увидеть и услышать, как больной высказывает свои мысли. Понимание, которое достигается врачом при общении с пациентом, позволяет получить ту информацию, которой недостает числовым результатам, полученным в лабораторных исследованиях. Наконец, получение качественных данных быстрее и дешевле и позволяет врачу выбрать минимальное число дальнейших высоко формализованных исследований.

Для представления качественных данных существуют две шкалы: *номинальная* и *порядковая*.

Номинальная шкала

Эта шкала называется также шкалой *наименований, дихотомической, ординарной,* (не перепутайте с ординальной), *категориальной*. Номинальная шкала строится разбиением данных на классы по тому или иному признаку [6, 7, 10, 47, 61, 66]. Данным, попавшим в один и тот же класс, присваивается одно и то же обозначение, которое используется как метка для различения объектов.

В номинальной шкале измерены, например, пол людей, социальное положение, исходы заболевания (выздоровление, смерть), прием лекарства или отсутствие приема лекарства, вид операции, место проживания больного, вид микроорганизма, наличие или отсутствие возбудителя в крови, характер заболевания (диагноз), вредной привычки, мутации гена, тип наследования признака и т.п.

Разбиение на классы нужно производить так, чтобы они носили *неупорядоченный характер* и не перекрывали друг друга: каждый должен быть либо мужчиной, либо женщиной, других вариантов нет. Номинальная шкала предполагает внутреннюю эквивалентность категорий – все измерения, отнесенные к одной и той же категории и представленные одним и тем же обозначением, одинаковы.

Представление данных в номинальной шкале не должно изменять семантический смысл, то есть медицинское содержание информации [6, 31.ч.1]. Если имеются два или более класса заболеваний, то кодовые значения этим заболеваниям могут быть присвоены произвольно. Например, мужской и женский пол можно обозначить М и Ж, или 1 и 2 (или 2 и 1), или знаками зодиака. Благоприятный и неблагоприятный исходы заболевания (например, смерть) можно обозначить 1 и 2, Б и Н и т.п. Для удобства при работе с различными статистическими программами используются цифровые обозначения. Виды оперативного вмешательства при язве желудка также можно кодировать цифрами: 1, 2, 3, 4 … Следует отметить, что ***цифра здесь – условный символ, который выбирается произвольно, но вовсе не обозначает, что операция «5» в пять раз лучше, чем операция «1».***

Выбор шкалы наименований означает, что измерения разбиваются на классы по тому или иному признаку. Одна шкала не может выглядеть так: 1 (местное обезболивание), 2 (перидуральная анесте-

зия), 3 (внутривенная анестезия), 4 (адекватная анестезия), 5 (неадекватная анестезия). Здесь первые три класса разделены по признаку «вид обезболивания», а два последних – по глубине наркоза, а это некорректно. Правильно в данном случае использовать не одну, а две шкалы: по виду обезболивания (1, 2, 3…) и отдельно по адекватности обезболивания, например: 1 – адекватное, 2 – неадекватное.

Для этих двух признаков в таблице, где представлены ваши данные, необходимо выделение двух отдельных столбцов, куда против фамилии или условного номера каждого пациента следует вставлять символы (цифры), соответствующие виду обезболивания в одном столбце и наличию или отсутствию адекватного обезболивания в другом столбце. Мы так подробно и назойливо пишем об этом, потому что еще одной частой ошибкой является попытка в одну ячейку вставить характеристики двух или более признаков.

***Пример* 1.1.** У больного – передний инфаркт миокарда (ученый обозначает этот вариант цифрой «1») и острая сердечная недостаточность III степени, по Killip (обозначение «3»). Ни в коем случае не пишите в одну ячейку 1, 3. Статистические программы такого представления могут вообще не понять. В вашей таблице должен быть один столбец для локализации инфаркта, другой – для степени сердечной недостаточности. ■

Обратите внимание, что хотя при вводе в компьютер номинальные данные чаще обозначают числами, например, пол мужской и женский можно обозначить* 0 *и* 1*, но эти числа (как и номера телефонов, истории болезней) нельзя складывать или умножать.

Группа допустимых преобразований в номинальной шкале состоит из *всех взаимно однозначных преобразований*. Арифметические операции (сложение, вычитание) не имеют смысла. Не существуют арифметическое среднее, медиана. Оценкой среднего может служить мода распределения, которая не зависит от однозначных преобразований измерений. Например, гипертоников среди больных с инфарктом миокарда больше, чем гипотоников. Мода является параметром сдвига для гипертоников и гипотоников вне зависимости от обозначений: ГП и ГИП, 1 и 2. Хотя измерения качественны, но можно сосчитать количество объектов каждого класса и определить частоты.

***Методы анализа*:** пригодны только методы категориального анализа – χ^2 (Хи-квадрат) критерий для полиномиального распределения; для проверки гипотезы о связанности двух и более номинальных переменных; выводы относительно биномиального распределения; операции над структурными функциями от дихотомических переменных.

Порядковая шкала

В порядковой или *ординальной* (в отличие от ординарной) шкале качественные данные не только разбиваются на классы, но и упорядочиваются сами классы. Каждому классу присваивается свой *символ,* и порядок символов соответствует порядку класса по правилу «больше, чем», «более предпочтителен, чем», «сильнее». Любое множество A называется упорядоченным, если для любых двух его элементов A^* и B^* установлено, что либо A^* предшествует B^*, либо B^* предшествует A^*. Если не удается установить строгое предшествование для всех элементов множества, но можно произвести групповое упорядочение, тогда упорядочиваются подмножества равноценных элементов.

***Пример* 1.2**. Рассмотрим несколько случаев приписывания кодов переменным. Степень артериальной гипертонии по классификации, принятой ВОЗ, можно расположить в следующем порядке: легкая, средняя и тяжелая (уровни стадии обычно обозначают: 1, 2, 3). Индивидуумов можно упорядочить по цвету волос (x – светлый, y – серый, z – черный). Социальное положение – по уровню доходов: – низкий, средний, высокий; страны, где проживают пациенты, – по уровню развития медицинской помощи: низкий (Россия, Украина, Грузия и др.), средний (Польша, Чехия, Португалия, Турция и т.п.), и высокий (США, Израиль, Германия и др.). Можно эти классы обозначить: 1, 2, 3 (так удобнее), или 0, 1, 10, или 10, 1, 0. Главное – наличие определенного порядка. ■

Классическим примером такой шкалы является распределение больных по классам (стадиям, степеням тяжести), когда используется комплекс клинических характеристик – например, классы или стадии недостаточности кровообращения, шока, злокачественного образования, гестоза, метаплазии. Исследователь может самостоятельно создавать шкалы. Например, интенсивность боли разделить на 3, или 5, или 10 степеней по своему усмотрению, вид операций – по степени травматичности или степени кровопотери.

Кодирование данных порядковой шкалы числами нужно произвести так, чтобы сохранить максимальное отличие. В этой шкале можно отличить число градаций признака в баллах по 7-балльной шкале. **Здесь б*аллы – не числа, а порядок*** [66].

Такое число градаций соответствует возможностям человека по переработке информации, что сформулировано специалистами в области инженерной психологии в виде известного правила «семь плюс-минус два» [31, ч.1]. При этом следует помнить, что по большинству наиболее важных признаков подобные шкалы уже созданы, и лучше поискать их, как следует, в литературе.

В порядковой шкале можно различать высокие значения от низких, но нельзя описать различия между объектами в точных единицах измерения. Хотя известно, что степень артериальной гипертонии 3 больше чем степень 1, нельзя утверждать, что она в три раза больше, в три раза тяжелее. На порядковом уровне присвоенные числа – это символы, обозначающие место объекта в упорядоченном ряду.

В ряде ситуаций не всегда бывает, ясен принцип сравнения, и возникают случаи нетранзитивности. Например, при психологическом тестировании и необходимости выбора одной фигуры из двух ребенок предпочитает фигурку слона фигурке кошки, фигурку кошки фигурке лошади, а фигурку лошади фигурке слона. Здесь попытка выстроить фигурки по популярности будет неудачной. То же может касаться выбора хирургом вида оперативного вмешательства или доминирования аллельных генов при множественном полиморфизме.

В случае, когда некоторые элементы измеряемого множества несравнимы по упорядочивающему отношению, а остальное подмножество элементов допускает сравнение, имеем так называемый частичный порядок.

Недостатком порядковой шкалы является то, что ничего не можем сказать о равномерности или неравномерности интервалов между соседними классами. Очевидно, в двух последних случаях стоит задуматься о принципе упорядочивания и возможности его модификации или замены. В противном случае придется использовать методы статистики для качественных переменных.

Преобразования в шкале порядка – *монотонные функции* $\varphi(x)$. Порядковая шкала позволяет производить все операции номинальной шкалы и преобразования полученных оценок (например, от экспер-

тов), отвечающие всем монотонно возрастающим функциям. Положительные оценки могут быть заменены их квадратами, или логарифмами, или любой другой монотонно возрастающей функцией.

Арифметические операции не имеют смысла. Нельзя вычислить среднее среднеарифметическим. Среднее можно оценить медианой (*median*) или, с некоторой потерей информации, модой.

Методы анализа. Пригодны *непараметрические методы* (nonparametric methods), свободные от вида распределений; используются знаки, ранги абсолютных значений и знаки их разностей, число инверсий. Применяются процедуры проверки гипотезы равенства медианы заданному значению и равенства двух медиан, а также методы дисперсионного анализа Крускала–Уоллиса и Фридмана, методы ранговой корреляции Кендалла и Спирмена.

1.1.2.Количественные измерения

Как не запутаться в цифрах и единицах измерения?

Для количественных измерений различают шкалы: *интервальную, отношений, разностей и абсолютную* [6, 10, 31, ч. 1, 54].

С одной стороны, для всех количественных шкал приемлемы арифметические действия, а значит, и большинство критериев, предназначенных для количественных данных. Знать различия этих шкал следует в большей мере для того, чтобы правильно использовать преобразование данных и при этом не попасть впросак. Например, переводить значения лабораторных показателей из одной системы измерения в другую.

***Пример* 1.3.** Один из авторов этой книги участвовал в крупном международном клиническом испытании, которое было спланировано и проводилось очень серьезными и авторитетными компаниями. В протокол была заложена серьезная ошибка, внесшая некоторую сумятицу и выявленная в ходе испытания.

Основой ошибки было то, что не учитывались различия между разными видами шкал для количественных данных. Допускалась оценка одного из ключевых биохимических показателей в мг, % и моль/л. Одним из важнейших показателей для включения больного в исследование была разность между значениями этого показателя, определенного во время двух последовательных визитов пациентов (обследований), причем разница, выраженная в процентах. Организаторами исследования предлагалась специальная формула для подсчета этой разности. Случалось так, что данные отдельных больных при подсчете в мг, % укладывались в допустимые границы этой разности, а значит, могли быть включены в исследование, а при

подсчете этой же разности, но с использованием моль/л – не соответствовали критериям включения. ■

Опишем разновидности шкал для количественных данных.

Интервальная шкала

В этой шкале измеряют значения величин на числовой оси с точностью до линейных преобразований:

$$\varphi(x) = \alpha x + \beta, \quad \alpha > 0.$$

Здесь должны быть заданы единицы измерения, точка начала отсчета, и при переходе от одной системы измерений к другой они должны быть определены. Шкала позволяет не только определить, что одно измерение больше другого, но и утверждать, насколько велики расстояния между двумя точками в принятой шкале. Однако, если мы хотим определить, во сколько раз одно измерение больше другого, или выразить это соотношение в процентах, следует понимать, что при переходе от одной системы измерений к другой результаты могут быть разными (как в вышеприведенном примере).

***Пример* 1.4.** Измерение температуры по Цельсию, Фаренгейту. Если мы хотим перейти от Фаренгейта к Цельсию, где x – температура по Фаренгейту, y – по Цельсию, то:

$$y = (5/9)(x - 32) = 0{,}55x - 17{,}8.$$

Кстати, формула подобного вида для пересчета одних единиц в другие, где, условно говоря, есть и умножение (деление) на какой-либо коэффициент, и сложение (вычитание), является индикатором интервальной шкалы.

Здесь можно не только утверждать, что температура 50 градусов выше, чем 20 градусов, но и то, что увеличение температуры с 10 до 50 градусов в 5 раз больше увеличения температуры с 40 до 50 градусов. Вместе с тем, если эти измерения были по Цельсию и мы захотим их выразить по Фаренгейту, это соотношение (в 5 раз) не сохранится. В этой шкале отношение оценок не сохраняется, и отношение одного измерения к другому меняется вместе с типом используемой системы измерения. Например, температурам 100^o и 0^o в шкале по Цельсию соответствуют температуры 212^o и 32^o по Фаренгейту. Однако в шкале интервалов сохраняются отношения разностей численных оценок.

Действительно, пусть x_1, x_2, x_3, x_4 – измерения признака в некоторой числовой системе, а $\varphi(x_1)$, $\varphi(x_2)$, $\varphi(x_3)$, $\varphi(x_4)$ – соответствующие им измерения этого же признака в другой числовой системе. Тогда [10, 61]:

$$\frac{\varphi(x_1) - \varphi(x_2)}{\varphi(x_3) - \varphi(x_4)} = \frac{\alpha x_1 - \alpha x_2}{\alpha x_3 - \alpha x_4} = \frac{x_1 - x_2}{x_3 - x_4}. \blacksquare$$

В интервальной шкале применимы оценки среднеарифметического, дисперсии, высших и смешанных моментов, процедуры сравнения средних арифметических для двух совокупностей объектов. Однако отношение среднеарифметических величин не имеет смысла. Корректно лишь использование отношения разностей средних арифметических. Отсюда следует, что не имеет смысла коэффициент вариации – отношение к математическому ожиданию, так как обе эти величины зависят от начала отсчета.

Шкала отношений

Значения в измеряемой системе определяются с точностью до преобразований подобия [10, 54, 61]:

$$\varphi(x) = \alpha x,\ \alpha > 0.$$

Здесь, в отличие от интервальной шкалы, фиксировано начало отсчета, но не единица измерения, и отношение измеренных значений одинаково при переходе от одной системы к другой. То есть в этой ситуации если при измерении роста в сантиметрах вы в 1,2 раза выше своей жены (или своего мужа), то и при измерении вашего роста в футах, или метрах, или верстах вы все равно выше в 1,2 раза.

Пусть измерения x_1 и x_2 – в одной системе, а $y_1 = \alpha x_1$, $y_2 = \alpha x_2$ – в другой. Тогда $x_1 / x_2 = y_1 / y_2$, следовательно, отношение любых двух точек шкалы не зависит от единицы измерения.

Шкала имеет все свойства шкал количественных измерений, для нее справедливы свойства аддитивности, можно использовать все критерии, приемлемые для количественных данных, она обычная в технических и физических системах, но редко встречается в обществоведении и гуманитарных дисциплинах.

***Пример* 1.5.** Примеров использования шкалы отношений в медицине немало. Лимфоциты в лейкоцитарной формуле крови измеряются в относительной шкале: x – число клеток в тыс. в 1 мкл крови или $y = 10^6 x$ в единицах СИ, где $\alpha = 10^6$. То же самое касается измерения длины, веса и т.п. ■

Шкала разностей

Единица измерения фиксирована, а начало отсчета – нет, при переходе от одной числовой системы к другой меняется лишь начало отсчета. Преобразование [10, 54, 61]:

$$\varphi(x) = x + \beta.$$

Шкала имеет все свойства шкал количественных измерений, для нее можно использовать все критерии для количественных данных. При переходе от одних единиц измерения к другим соотношения (например, в 2 раза быстрее) сохраняются.

***Пример* 1.6.** Солнечные календари: юлианский, григорианский, древнерусский, Великой Французской революции, всемирный. В медицине чаще всего это длительность наблюдения, время до возникновения какого-либо события. Как мы знаем, в этих случаях началом отсчета может выбираться и время поступления, и время приема первой дозы препарата, и день операции, и день выписки из больницы и т.п. ■

Абсолютная шкала

Известны начало отсчета и единица измерения. Численные значения измерений тождественно равны значениям числовой оси. Например, количество детей в семье, число случаев заболевания дизентерией в городе, число зубов во рту, число окрашенных гранул в клетке и т.п. Преобразование – тождественное: $\varphi(x) = x$. Для этой шкалы, равно как и для любых шкал количественных измерений, имеют смысл все арифметические операции. Оценкой параметра сдвига могут служить среднее, медиана и мода распределения.

1.1.3. Общая характеристика шкал

В чем и как лучше измерять* (*оценивать*) *изучаемые признаки?

Сила шкалы

Сильные шкалы позволяют провести наиболее тонкий анализ и найти различия, взаимосвязи и т.п. с наибольшей точностью, а также в максимальной степени использовать имеющуюся у исследователя информацию. Это положительная сторона «силы». Сила шкалы определяется группой допустимых преобразований: чем шире эти преобразования, тем слабее шкала [10, 54]. По мере увеличения силы шкалы располагаются в таком порядке: номинальная, порядковая, интервальная, отношения, разностей и абсолютная.

Хотя качественные шкалы (номинальная и порядковая) – самые слабые, но они более помехоустойчивы, если помеха находится в рамках допустимых преобразований. Именно эти шкалы, на наш взгляд,

позволяют хоть как-то нивелировать те погрешности, которые неизбежно допускаются при сборе данных в медицинских исследованиях.

Использование номинальной или порядковой шкалы обусловливает существенное огрубление данных. Например, когда мы отмечаем, что у больного артериальная гипертония III степени это гораздо менее точно, чем зафиксировать артериальное давление 190 и 115 мм рт.ст. Напротив, когда мы имеем дело с количественными данными, то есть сильными шкалами, эту «силу» мы можем хорошо использовать, только когда сбор данных и сами данные удовлетворяют очень жестким условиям, каковые в реальной практике соблюсти крайне сложно, а подчас и невозможно. При этом, как только условия нарушаются, возникают такие отклонения в результатах статистической обработки, которые могут катастрофически исказить реальное положение вещей и привести врача к абсолютно ложным выводам.

Следует помнить, что методы вычислений, справедливые в слабой шкале, применимы с некоторой потерей информации и для измерений в сильных шкалах. Обратное утверждение не имеет смысла, так как данные слабой шкалы без специальных преобразований невозможно обработать методами сильных шкал. Например, некорректно использовать критерий Стьюдента для работы с номинальной и порядковой шкалами.

Квазиквантитативные переменные

Как и зачем качественные данные преобразовывать в количественные?

Нередко у обследуемых пациентов мы измеряем и анализируем одновременно и количественные, и качественные показатели. Например, данные эхокардиографического обследования содержат множество величин, измеряемых в миллиметрах, процентах, а ряд показателей – отвечающих на вопрос «да» или «нет» (нарушение локальной сократимости, диастолическая дисфункция и т.п). Нам очень хочется провести многофакторный анализ с учетом всей полученной информации, а не отделять и порознь обрабатывать количественные и качественные показатели, как того требуют статистики. Одним из возможных выходов из этого положения является замена качественных

характеристик на *квазиквантитативные* [9, 10] с последующим использованием методов классической статистики, например, многофакторного линейного регрессионного анализа.

Полихотомические данные – многозначные (более 2 классов) номинальные (цвет волос, вид операции, штамм вибриона и т.п.), порядковые и данные, построенные на субъективных критериях – удобно заменять переменными, называемыми ***квазиквантитативными*** (квантитативное – от лат. *quantitas* – количество). Такая замена переменных приводит к дихотомизации данных с некоторой потерей информации, но зато каждый качественный признак можно представить в виде квазиколичественного признака.

Пусть имеется измерение x качественного признака X, принимающего k взаимно исключающих значений (классов, альтернатив), упорядоченных по некоторому правилу, например, путем соглашения. Измерение x можно представить как квазиквантитативную переменную (или k новых дихотомических переменных) с компонентами, принимающими только значения 0 и 1, причем одна компонента вектора, соответствующая наблюдаемому классу, равна 1, а остальные компоненты равны 0.

Предположим, что номинальный признак X_j отдельно взятого i-го объекта из некоторой совокупности может принимать одно из k возможных значений (классов, альтернатив). Пусть $k = 3$ и каждому значению j-й переменной ($j = 1, 2, 3$) i-го объекта поставим в соответствие свой дихотомический признак, задаваемый правилом [9]:

$$x_{i1} = \begin{cases} 1, \text{если измеренное значение равно 1-й альтернативе;} \\ 0 \quad \text{в противном случае.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{если измеренное значение равно 2-й альтернативе;} \\ 0 \quad \text{в противном случае.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{если измеренное значение равно 2-й альтернативе;} \\ 0 \quad \text{в противном случае.} \end{cases}$$

$$x_{i3} = \begin{cases} 1, \text{если измеренное значение равно 3-й альтернативе;} \\ 0 \quad \text{в противном случае.} \end{cases}$$

Если значения некоторого признака, имеющего k классов, измерены у n объектов (n-подвекторов), то получим *матрицу квазиквантитативной* (*дихотомизированной*) *переменной* $\mathbf{X}$ размерности $n \times k$ с элементами 0 и 1. В матрице $\mathbf{X}$ каждый объект отображается своей строкой как точка k-мерного пространства, а признак отображается k-столбцами:

$$\mathbf{X} = [(x_{11}, \ldots, x_{1k}), (x_{21}, \ldots, x_{2k}), \ldots, (x_{n1}, \ldots, x_{nk})]',$$

$$x_{ij} = \begin{cases} 1, \text{если событие } s_j \ (j = 1, \ldots, k) \text{ наступило в } i-\text{м подвекторе;} \\ 0 - \text{в противном случае.} \end{cases}$$

***Пример* 1.7.** Пусть задан вектор качественной оценки стадий заболевания у 5 онкологических больных по шкале: первая (1), вторая (2), третья (3), четвертая (4), т.е. имеем $\mathbf{x} = [4\ 2\ 1\ 3\ 1]'$, $k = 4$, $n = 5$. Порядок классов приравняем шкалам стадий. Преобразовав эти данные в квазиквантитативную переменную, получим матрицу размерности (5×4):

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Так как для каждого подвектора (строки) матрицы $\mathbf{X}$ справедливо соотношение $\Sigma_{s=1}^{k} x_{js} = 1$, то представление качественных данных с помощью квазиквантитативных переменных избыточно.

Наблюдение *дихотомической* (*булевской*, *бинарной*) переменной, состоящей из двух элементов «0 или 1», соподчиненных постулату классического принципа исключенного третьего «*A* или *не-A*», является частным случаем квазиквантитативной переменной при $k = 2$ для всех j. ■

***Пример* 1.8.** Квазиквантитативные переменные имеют большое значение в медико-биологических, социологических и эконометрических исследованиях, в которых большинство переменных, интересующих специалистов, не может быть измерено в количественных шкалах. В медицине это случаи типа умер или выжил, случился перелом или нет, произошел ретромбоз или не произошел и т.п. При этом квазиквантитативные переменные с небольшим числом состояний зачастую являются более адекватными, чем результаты измерений по методикам, использующим большее число градаций. В этом отношении достаточно нагляден переход кардиологов к ЭКГ-классификации инфаркта «с зубцом Q» и «без зубца Q», вместо того, что было 20–30 лет назад, когда была градация на несколько степеней по объему – от микроинфаркта до обширного. ■

Квазиквантитативный признак X качественных измерений с n классами задается k булевскими столбцами, который в силу избыточности можно задать $k-1$ столбцами. Такое представление квазиквантитативной переменной с матрицей с $(k-1)$ столбцами носит название *индикаторной переменной* (*IND*) и широко используется в ППП при преобразовании качественных переменных.

Таким образом, в ряде случаев в статистических программах заложена возможность при помощи команды «*IND*» преобразовать качественные в «почти» количественные (квазиквантитативные) переменные и использовать после этого их вместе с количественными показателями, например, для включения в модель многофакторного регрессионного анализа при прогнозировании. При этом, если вы воспользовались такой процедурой, целесообразно обязательно на это сослаться в описании методов исследования.

1.1.4. Выбор шкалы измерений

Выбор типа шкалы производят конечно же, исходя из того, какую характеристику вы измеряете. Кроме того, учитывают свойство и степень необходимой детализации. Чем детальнее и тоньше вы хотите проанализировать данные, чем меньшие сдвиги вы хотите заметить или тонкие взаимосвязи выявить, тем более сильную шкалу следует выбирать, поскольку измерения в более сильной шкале обеспечивают получение той же информации во всех более слабых шкалах.

Количественные данные можно представить при желании с использованием номинальной и порядковой шкал, но нужно стараться собирать информацию по измерениям на уровне самой сильной шкалы. Например, возраст человека, зафиксированный в годах (шкала отношений), в дальнейшем можно отнести к одной из возрастных групп: молодых, средних лет, пожилых и старческих лет, которые можем представить в порядковой шкале – 1, 2, 3, 4.

Если вы зафиксировали только возрастную группу, 0будет трудно установить точный возраст человека, который себя отнес к возрастной группе пожилых. То же самое касается клиренса креатинина и стадии почечной недостаточности, уровня гемоглобина в крови и наличия или отсутствия анемии, числа заболевших грип-

пом за год и эпидемии гриппа, поэтому при возможности следует регистрировать данные в рамках шкал абсолютной, отношений и т.п. Иными словами, лучше использовать количественные показатели, для того чтобы иметь возможность использовать методы и критерии для подобного рода данных. Однако нам кажется, что одновременно с этим в соседнем столбце или столбцах полезно вносить и данные порядковой или номинальной шкале. Например, в столбце «гемоглобин» вы записываете «143», а в соседнем столбце «анемия» – цифру, обозначающую «нет», например «2». (Напоминаем, цифры, обозначающие «да» и «нет», вы выбираете произвольно.) Такой подход занесения данных очень практичен. Он позволит вам в зависимости от постановки задачи, которая в процессе исследования может изменяться, а также для проверки ваших новых гипотез, использовать максимально возможное число методов и ответить на наибольшее число вопросов.

Кроме того, когда вы проводите какое либо исследование, вы постепенно осознаете наличие все большего количества технических недостатков и несоответствие «идеалу» условий проведения работы. Это значит, что методы статистики для количественных данных применить вы можете лишь с очень большими оговорками. Также вы можете осознать, что для того чтобы узнать правду, будет лучше использовать качественные шкалы и соответственно непараметрические методы.

Роль качественных шкал в медицине огромна. Это можно объяснить, во-первых, простотой их получения и интерпретации. Во-вторых, вопросы, суждения, мнения и т.д. на качественном уровне адекватно воспринимаются пациентами, что позволяет выдвигать легко проверяемые модели. В третьих, методы анализа качественных данных используют простой и доступный для врача математический аппарат.

Таким образом, на этапе планирования и сбора информации, старайтесь при возможности делать так, чтобы в дальнейшем при обработке данных вы могли использовать и сильные, но чувствительные к помехам шкалы, и слабые, но устойчивые к ошибкам измерений. Фиксируйте результаты исследований сразу в нескольких видах, количе-

ственных, качественных и др. Однако понятно, что умерших и выживших можно записать только одним способом.

1.1.5. Нечисловые данные

Что можно, а чего нельзя делать с «да» или «нет»

Продолжим обсуждение качественных измерений в более широком контексте статистики нечисловых данных. Объектами нечисловой природы [61] называются такие объекты, переменные которых невозможно описать числами. «Объект» используется в качестве общего термина, которым можно заменить термины «вопрос», «обработка», «стимул» и им подобные. Примерами являются измерения в качественных шкалах, результаты парных сравнений, множества, нечеткие множества, экспертные оценки. Приведем ряд примеров нечисловых данных.

1. ***Дихотомические*** (***бинарные***) ***данные.*** Эти данные являются результатами измерений альтернативного признака и могут принимать только одно из двух значений (0 или 1). Например, психологические тесты Миннесотского многофакторного личностного исследования (MMPI) построены на дихотомических данных.

Результат *парного сравнения* [19] также является дихотомической переменной. В парном сравнении эксперт должен в одних случаях выбрать из двух объектов лучший по качеству, в других – ответить, похожи объекты или нет. В обоих случаях ответом является дихотомическая переменная – 0 или 1. Парное сравнение позволяет решить задачи упорядочения, оценивания и принятия решений. Примерами являются: упорядочение стадий гипертонической болезни по Мясникову; сравнения в суждениях типа «какой метод лечения лучше». Если сравнение выступает как инструмент выбора, то оно включается в процедуру принятия решений как средство сопоставления или упорядочения альтернатив, то есть в сравнении заложена идея предпочтения. Рассматривая данные парных сравнений, как категоризованные данные и представляя их в виде таблиц сопряженности признаков (ТСП), можно облегчить анализ результатов парных сравнений [19].

Дихотомические данные используются при анализе медико-социологических анкет с ответами типа «да» – «нет».

2. ***Результаты анализа количественных данных как объекты нечисловой природы***. Выбор числа регрессоров в регрессионном анализе, числа канонических дискриминантных функций в дискриминантном анализе, размерности факторного пространства связаны с дихотомической переменной – принять или отвергнуть ту или иную гипотезу.

То же касается и медицинских ситуаций. Вводить добутамин на основании анализа цифр артериального давления, переливать эритромассу или ограничиться кристаллоидными растворами, исходя из цифр гематокрита у больного с кровотечением, какую стадию рака установить исходя из размеров опухоли.

Факты введения добутамина и эритромассы, стадия рака хоть и являются результатом анализа каких-то цифр, но при этом уже становятся объектами нечисловой природы, так как здесь также принимаются альтернативные решения. В этой связи для обработки значений величины опухоли и стадии ее развития используются совсем разные подходы, ибо методы статистического анализа данных объектов нечисловой природы принципиально иные, чем количественных данных. Разумеется, что нечисловые данные в компьютере представлены числами, но эти числа нельзя складывать и умножать на другое число.

В строго математической терминологии – «***нечисловые данные не являются структурой линейного пространства***». Можно «насильно» ввести эти данные в это самое пространство, например, посчитать коэффициент линейной корреляции между степенью тяжести гестоза и инфицированностью женщин хламидиями. Компьютер при этом ничего не «заметит» и все подсчитает. Только компетентный человек заметит, что произошло насилие над статистикой и здравым смыслом.

Таким образом, нечисловые данные нельзя непосредственно обрабатывать с использованием методов статистики, предназначенных для числовых данных. Следует использовать только специальные методы или заниматься преобразованиями: из нечисловых в «квазиквантитативные». Хотя чаще делают наоборот – из числовых данные переводят в нечисловые.

1.1.6. Интервальные данные

Почему точность измерений важна и для статистики?

В статистике интервальных данных элементами выборки являются не числа, а интервалы [61, 66]. Таких данных в медицине довольно много. Это все то, что измеряется приборами в каких-либо единицах с определенной точностью, начиная от артериального давления в мм рт. ст. и заканчивая раскрытием шейки матки при родах, измеряемого в «пальцах».

Интервальные данные – это результат измерения выборочных данных с погрешностью, присущей каждому измерительному прибору. Такие данные получаются при дискретном представлении непрерывных одинаково распределенных случайных величин. Измеряемая величина на выходе измерительной системы представляется в виде конечного числа разрешенных уровней, отстоящих друг от друга на конечный интервал, который в теории сигналов называется квантованием по уровню. Если истинное значение наблюдения находится внутри этого интервала, то вместо него передается ближайший разрешенный уровень. Например, реальное систолическое артериальное давление у пациента равно 122,3 мм рт. ст., тогда как записанный врачом результат составит 122, а чаще всего 120 мм рт. ст. Это же касается всех инструментальных и лабораторных показателей, при измерении которых всегда есть определенная погрешность. Эта погрешность, как правило, указывается в инструкции к прибору.

Погрешность измерений неизбежна, так как любое наблюдение представляется конечным числом значащих цифр.

Величина максимально возможного (по абсолютной величине) отклонения, вызванного погрешностью ε, известного исследователю значения $f(y)$ от истинного $f(x)$ [66],

$$Nf(x) = \sup | f(y) - f(x) |,$$

называется нотной.

Нотна – неустранимая погрешность, поэтому оценки статистик интервальных данных имеют погрешность неубывающую при неограниченном увеличении объема выборки.

***Пример* 1.9.** Истинные скорости оседания эритроцитов (СОЭ) и их интервальные значения с уровнями шкалы 1 мм/ч приведены в табл. 1.1.

Таблица 1.1

Скорость оседания эритроцитов

Пациенты	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
СОЭ ист.	2,4	9,6	8,1	11,6	9,4	3,8	13,7	8,4	10	4,9
СОЭ инт.	2	10	8	12	9	4	14	8	10	5

Оценка среднего истинных данных, не известных исследователю, равна:

$$\bar{x} = f(x) = \frac{x_1 + x_2 + \ldots + x_{10}}{n} = 8{,}19, \; S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 12{,}49.$$

Если случайная величина X распределена нормально, то дисперсия оценки среднего равна:

$$S_{\bar{x}}^2 = \frac{S_x^2}{n} = \frac{12{,}49}{10} = 1{,}249,$$

доверительный интервал (ДИ) на уровне значимости $\alpha = 0.05$ определяется выражением:

$$(\bar{x} \pm t(\alpha/2, v)S_{\bar{x}}) = 8{,}19 \pm 2{,}262 \cdot 1{,}12 = 8{,}19 \pm 2{,}53 = [5{,}66; 10{,}72].$$

Так как каждое значение x_i в интервальной шкале известно с точностью до $\Delta = 0{,}5$, то и среднеарифметическое известно с той же точностью, поэтому вместо X нужно подставить Y. Тогда имеем:

$$\bar{y} = 8{,}2; \; S_y^2 = 13{,}5; \; S_{\bar{y}} = 1{,}16.$$

ДИ на уровне значимости $\alpha = 0{,}05$ равен

$$(\bar{y} - t(\alpha/2, v)S_{\bar{y}} - \Delta; \; \bar{y} + t(\alpha/2, v)S_{\bar{y}} + \Delta] = 8{,}2 - (2{,}262 \cdot 1{,}16 - 0{,}5);$$

$$8{,}2 + (2{,}262 \cdot 1{,}16 + 0{,}5) = [5{,}06; 11{,}32].$$

Отсюда видно, что ДИ с учетом нотны шире, чем в классическом случае. ■

Что, исходя из этого, следует делать и не делать в реальной практике? Помнить, что, измеряя что-либо с какой-либо точностью, вы обречены иметь погрешность, которую в дальнейшем нельзя никак устранить и сгладить, в том числе увеличивая число обследований. Эту погрешность всегда следует знать: задумывайтесь над характеристикой метода измерения или прибора, единицами измерения, условиями сбора данных.

1.1.7. Производные данные

В медицине мы можем столкнуться со множеством других типов данных – такими, как проценты и пропорции, отношения и удельные показатели, интенсивность и индексы [26, 66, 68].

Проценты и пропорции

Рассмотрим утверждение: «Из 69 пациентов кардиологического отделения у 41 установлен инфаркт миокарда». Этот факт можно выразить более четко, если привести процентное содержание: «у 59% пациентов – инфаркт миокарда».

Процентные показатели и пропорции стандартизуют: процентные показатели приводят к основанию 100, а пропорции – к основанию 1,00. Пропорция и процентный показатель определяются по формулам:

пропорция: $\hat{P} = \frac{x}{n}$;

процентный показатель: $\hat{P} = \frac{x}{n}100$,

где x – частота, количество наблюдений в какой-либо категории; n – общее количество наблюдений во всех категориях.

Для нашего примера получаем:

$$\hat{P} = \frac{x}{n} = \frac{41}{69} = 0{,}5913; \quad \hat{P} = \frac{x}{n}100 = \frac{41}{69} \cdot 100 = 59{,}13\%.$$

Особенности использования процентных показателей и пропорций.

1. При небольшом количестве наблюдений (менее 20) предпочтительнее указывать значения частот, а не процентные показатели или пропорции, так как в этом случае процентные показатели изменяются значительнее при сравнительно малых изменениях в данных. Например, наблюдаются 16 пациентов, по 8 человек каждого пола, то есть по 50%. Выборка уменьшилась на одну женщину. Распределение процентных показателей станет равным 46,66% женщин и 53,33% мужчин. Если выборка содержала бы 400 мужчин и женщин, и из выборки исключили одну женщину, процентный показатель женщин будет равняться 49,87%, т.е. изменится на величину менее 0,1%.

2. Пропорции и процентные показатели отражают степень изменения показателя, а не абсолютные данные, поэтому наряду со значениями пропорций и процентов следует указывать общее количество наблюдений.

3. Процентные показатели и пропорции можно вычислять для переменных, измеряемых по *любой* шкале измерений, так как для их вычисления нужно знать лишь *количество наблюдений,* относящихся к отдельной категории переменных, а не значения самих переменных.

4. Понятие «пропорции» тесно связано с биномиальным распределением (см. п. 2.5.1) и его точечными и интервальными оценками (см. п. 2.7.2 и п. 2.7.3).

Отношения и удельные показатели

Отношения и удельные показатели особенно полезны при сравнении категорий с точки зрения относительной частоты [68].

Отношения определяют делением частоты наблюдений одной категории на частоту наблюдений другой категории:

$$отношение = \frac{n_1}{n_2},$$

где n_1 – количество наблюдений в первой категории; n_2 – количество наблюдений во второй категории.

Например, при определении индекса массы тела (индекс Кетле) массу тела (кг) делят на квадрат его/ее роста ($м^2$). Таким образом, составляют суждение превышает ли его/ее масса тела норму или, наоборот, имеется ее недостаток.

Удельные показатели определяются как количество действительных появлений определенного явления, разделенное на количество возможных появлений за некоторую единицу времени. Удельные показатели обычно умножаются на некоторую степень числа 10 для получения целых чисел в десятичной дроби. Например, общий удельный показатель смертности в популяции из 8000 человек определяется как число смертей 150 в данной популяции за год:

$$\frac{\text{Количество смертей за год}}{\text{Общая численность популяции}} \times 1000 = \frac{150}{8000} \times 1000 = 17{,}5,$$

т.е. на каждую тысячу человек пришлось 17,5 смертельных исходов.

Интенсивность. Это относительная частота заболеваний, равная частному от деления числа заболеваний на длительность рассматриваемого периода. Эти данные являются обычными при эпидемиологическом исследовании.

Метки, оценки. Это произвольные данные, применяемые тогда, когда мы не можем измерить количество. Например, ответы на вопросы относительно качества жизни можно обобщить, чтобы получить оценку качества жизни каждого индивидуума.

1.2. Многомерные данные

Анализировать только одну характеристику опасно

Как уже упоминалось, объекты с единственной характеристикой в реальности отсутствуют. Это значит, что даже при самом корректном использовании самых замечательных методов одномерного анализа мы постоянно рискуем. В этом случае мы делаем существенное допущение о том, что изучаемая характеристика важна и обладает самостоятельной ценностью, то есть находится вне влияния других свойств изучаемого объекта, например, больного, микроорганизма, генома.

На практике объекты характеризуются большим количеством различных свойств и связей между ними, примерами являются объекты в социологии, эконометрике, биологии, медицине, экологии, сложных технических системах. Рассмотрим известный пример. Когда по результатам одномерного анализа делался вывод, что использование заместительной гормональной терапии у женщин после менопаузы снижает риск инфаркта, то не учитывали важный момент. Женщины, принимающие эти лекарства, имели существенные отличия в социальном статусе, привычках и т.д. и что, скорее всего именно эти особенности и определяли меньший риск инфаркта. Поэтому, какую бы мы узкую задачу не решали, какой бы важный параметр не изучали, всегда следует регистрировать и анализировать максимально возможное число показателей изучаемого объекта.

Если каждый объект в выборке определяется значениями двух или более переменных, то имеем дело с *многомерными данными*, которые подразделяются на два класса – числовые и нечисловые.

1.2.1. Матрица «объект-признак»

Как должны выглядеть данные? Как их заносить в компьютер?

Многомерные данные можно представить в виде матрицы «*объект-признак*» [31, ч.1, 61], строки которой соотнесены с анализируемыми объектами или номерами опытов, а столбцы – со значениями изучаемых признаков. Следовательно, матрица – это таблица чисел, расположенных по строкам и столбцам. Обычный лабораторный журнал почти идентичен такой матрице. В нем обычно под номерами следуют фамилии больных, на каждого больного – одна строчка, а напротив – лабораторные показатели, для каждого из которых выделен свой столбец.

Основные схемы данных типа «объект-признак»

Характерной особенностью данных типа «объект-признак» является то, что формирование матрицы связано со временем. В реальной жизни характеристики у больного Иванова вы измеряли вчера, Петрова – сегодня, Сидорова – будете обследовать завтра, Исмаилова – через неделю, а Рабиновича – через месяц и т.д. по мере поступления пациентов в клинику. У вас нет никакой гарантии, что за время проведения вами исследования на изучаемые вами показатели не повлияют факторы, связанные со временем. Такими факторами являются: изменения погоды, магнитного поля, социальной обстановки, питания (осенью и зимой многие питаются по-разному), температуры воздуха в лаборатории, эпидемии вирусных инфекций, смена персонала лаборатории и многое другое, не учитываемое в рутинной практике. В связи с этим различают несколько схем матрицы «объект-признак». Если наблюдения признаков осуществляются с неизменным сдвигом по времени, например, каждый день исследуется по одному больному или по одному штамму вибриона, то получаем *синхронную схему*. Частным случаем синхронной схемы являются данные, полученные «одномоментно».

Если пациент или группа пациентов исследуется по одномоментным данным, то такие исследования называются также *поперечными*.

Поперечные исследования в простейшем случае дают описание заболевания, но не его развития. В более сложном случае учитываются связи признаков с вариантом заболевания [66].

Данные, полученные одномоментно

Почему нецелесообразно затягивать время сбора данных?

Термин «одномоментный» означает такой отрезок времени, в течение которого не могут произойти существенные изменения в объекте. В подавляющем большинстве случаев в биологии и медицине исходят из того, что исследование проходит «одномоментно» (поперечно), хотя очень часто это совсем не так. Мы представляем данные в матрице, а затем обрабатываем их так, как будто все больные обследованы в один момент. Например, у всех наших ста или тысячи пациентов в один момент взяли кровь на анализ, затем одновременно прооперировали, или они одновременно проглотили исследуемый препарат, после чего мы сразу или через год наблюдения одномоментно взяли повторные анализы.

Фактор времени, то есть неодномоментности поступления больных и завершения исследований, как правило, не учитывается. А это совершенно неправильно. Различные влияния, связанные со временем, могут быть так сильны, что способны, как минимум, значительно увеличить разброс полученных данных, а в ряде случаев и серьезно исказить результаты исследования. На практике даже относительную одномоментность получить крайне сложно, хотя надо к этому стремиться. При этом довольно обременительно точно учитывать «неодномоментность» с помощью описанных в следующих разделах процедур. Есть и выходы из такого сложного положения.

Не проводите слишком длительных исследований для включения наибольшего числа больных. В этих случаях статистическая «мощность» не улучшится из-за увеличения выборки, как мы того хотим, а уменьшится из-за влияния описанных выше причин. Старайтесь проводить рандомизацию объектов (пациентов и т.п.) при формировании основной и контрольной группы, а не набирать сначала группу контроля, а потом основную, как это нередко делается.

При вынужденных продлениях срока набора данных проверяйте полученные результаты на «случайность», то есть на отсут-

ствие «тренда» или зависимости последующих измерений от величины предыдущих. Если эти зависимости обнаружены, то одномоментные (поперечные) исследования некорректны.

В таких случаях нужно применить ***метод продольных*** (*проспективных, лонгитудиальных*) ***исследований*** [10, 68, 102], в которых в течение некоторого времени проводят наблюдения за определенной, заранее отобранной группой больных (*когортной*). Длительность наблюдения зависит от поставленных целей исследования. За время наблюдения должны проводиться более одного оценивания состояния больного. Продольные исследования являются самыми доказательными, но требуют длительных наблюдений и больших материальных затрат.

Формирование матрицы синхронных данных «объект-признак» для статистического анализа в виде файла Excel

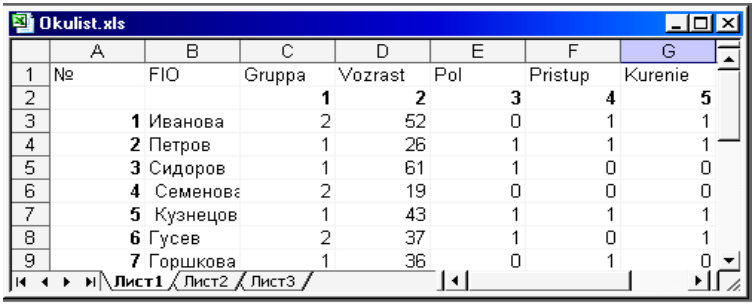
1. Каждому пациенту отводится одна и только одна строка.
2. Каждый признак записывается в одном столбце.
3. Все наблюдения приводятся на одном листе Excel.
4. Идентификация данных, относящихся к разным группам, задается значениями столбца «Gruppa».
5. В первой строке записывают названия признаков латиницей.
6. Вторая строка содержит номера признаков.
7. Первый столбец – порядковый номер больного.
8. Второй столбец – фамилии, записанные на кириллице.
9. Третий столбец – Gruppa (коды групп).
10. Далее, во избежание путаницы нечисловые и числовые данные приводите блочно:

а) после столбца «Gruppa» приводим блок номинальных данных («есть» – 2, «нет» – 1),

б) затем записываем коды порядковых данных (например, «тяжесть приступа» числами – 1, 2, 3),

в) после блока качественных данных приводится блок количественных данных.

Вид электронной таблицы данных приведен на рис. 1.3.



Okulist.xls

	A	B	C	D	E	F	G
1	№	FIO	Gruppa	Vozrast	Pol	Pristup	Kurenie
2			**1**	**2**	**3**	**4**	**5**
3	**1**	Иванова	2	52	0	1	1
4	**2**	Петров	1	26	1	1	1
5	**3**	Сидоров	1	61	1	0	0
6	**4**	Семенова	2	19	0	0	0
7	**5**	Кузнецов	1	43	1	1	1
8	**6**	Гусев	2	37	1	0	1
9	**7**	Горшкова	1	36	0	1	0

Лист1 Лист2 Лист3

Рис. 1.3. Запись данных типа «объект-признак» в виде файла Excel

Так как все качественные данные записываются числами, то в отдельном файле нужно привести расшифровку всех кодировок номинальных и порядковых символов.

1.2.2. Ковариационная и корреляционная матрицы

Корреляции бывают разные

Большинство методов многомерного анализа данных, включая множественную линейную регрессию, анализ главных компонент, факторный анализ, дискриминантный и канонический корреляционный анализы построены на преобразованиях исходной матрицы «объект-признак» или СВП в *ковариационную* или *корреляционную матрицу*.

Ковариационная матрица $\mathbf{K}_x$ – это квадратная таблица типа «признак-признак» размера $m \times m$, образованная из попарных ковариаций m случайных признаков $X_1, \ldots, X_m$ матрицы «объект-признак» $\mathbf{X}$.

$$\mathbf{K}_x = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ k_{m1} & k_{m1} & \cdots & k_{mm} \end{bmatrix}$$

Компоненты ковариационной матрицы при $i = j$ совпадают с дисперсией случайного признака X_i. Матрица $\mathbf{K}_x$ представляет собой симметричную *неотрицательно определенную матрицу*.

Если дисперсии величин $X_1, \ldots, X_m$ равны 1, то матрица $\mathbf{K}_x$ называется *корреляционной матрицей и обозначается как* $\mathbf{R}_x$.

***Пример* 1.10.** Рассмотрим чисто демонстрационный пример построения ковариационной и корреляционной матриц 6 параметров для 14 наблюдений у некоторого объекта исследования.

Вычисленные по исходным данным элементы ковариационной матрицы представлены в таблице 1.2.

Таблица 1.2

Ковариационная матрица

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	282,132	−199,725	−141,923	−4,039	−132,549	15,246
X_2	−199,725	499,412	190,577	−1,250	−27,593	−19,223
X_3	−141,923	190,577	125,962	−0,546	−25,000	−21,654
X_4	−4,039	−1,250	−0,546	0,804	25,522	5,201
X_5	−132,549	−27,593	−25,0	25,522	883,187	184,108
X_6	15,246	−19,223	−21,654	5,201	184,108	53,425

Из таблицы 1.2 видно, что ковариационная матрица представляет собой матрицу вида «признак-признак», она симметрична относительно главной диагонали, элементы которой представляют собой дисперсии соответствующих переменных.

Элементы ковариационной матрицы зависят от размерности измеряемых величин, поэтому непосредственно по этой матрице невозможно определить меру связи между переменными.

Для определения меры связи между переменными нужно перейти к стандартизованным переменным с нулевым средним и единичной дисперсией, предварительно вычислив средние и среднеквадратические значения таблица 1.3).

Таблица 1.3

Таблица средних, дисперсий и СКО исходных данных

	X_1	X_2	X_3	X_4	X_5	X_6
$\bar{x}_i$	88,143	114,214	72,500	3,531	107,571	33,100
$S^2_{x_i}$	282,132	499,412	125,962	0,804	883,187	53,425
S_{x_i}	16,797	22,347	11,223	0,896	29,719	7,309

Таблица 1.4

Корреляционная матрица

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1	**−0,532**	**−0,753**	−0,268	−0,266	0,124
X_2		1	**0,760**	−0,062	−0,042	−0,118
X_3			1	−0,054	−0,075	−0,264
X_4				1	**0,958**	**0,794**
X_5					1	**0,848**
X_6						1

Замечания

1.Так как корреляционная матрица симметричная, то общепринято матрицу изображать только верхней треугольной частью (табл. 1.4).

2.Элементы, значимые на уровне $p < 0,05$, отмечены полужирным шрифтом.

3.Свойства корреляции будут подробно изложены в главе 4. ■

1.2.3. Двухвходовые таблицы сопряженности признаков

Есть ли взаимосвязь между полом и цветом глаз?

Таблица сопряженности признаков (ТСП) (*contingency tables*) – одна из наиболее часто используемых в медицине таблиц. Во-первых, потому что в медико-биологических исследованиях часто анализируются сочетания качественных или «ранговых» показателей. Во-вторых, эти таблицы наглядны и позволяют исследователю лучше осознать полученные результаты. ТСП применяется в случае, когда каждый индивидуум (или экспериментальная единица) в популяции описывается двумя различными признаками (факторами, переменными) A и B.

Кроме упомянутых пола и цвета глаз, в качестве примера можно привести анализ сопряженности локализации инфаркта (передний, нижний, боковой) и тяжести острой сердечной недостаточности по Киллип (I – IV степени), вид родоразрешения и исходы родов, наличие гипертонии у пациента и вариант гена ангиотензин-превращающего фермента.

Пусть признак A имеет $r \geq 2$ *классов* или *уровней*, а B состоит из $c \geq 2$ классов или уровней, измеренных в номинальной или порядковых шкалах. Если данные измерены в количественных шкалах, то они должны быть сгруппированы в классы. Рассматривая r классов

признака *A* как строки (rows), а с классов признака *B* как столбцы (column) получим двухфакторную $r \times c$-ТСП (табл. 1.5).

Таблица 1.5

Таблица сопряженности признаков

Признак A	***Признак B***				***Всего по строкам***
	1	**2**	**- - -**	***c***	
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\bullet}$
Всего по столбцам	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet c}$	$n_{\bullet\bullet}$

Пересечение *i*-й строки и *j*-го столбца образуют *ij*-ячейку (*cell*). Число объектов n_{ij} в *ij*-ячейке называют *наблюдаемой частотой* (*observed frequencies*) ячейки. В таблице $n_{i\bullet}$ – это сумма объектов опр $n_{i\bullet}$еделенной *i*-й строки, $n_{\bullet j}$ – определенного *j*-го столбца, $n = n_{i\bullet\bullet}$ – объем выборки, равный сумме итогов строк и столбцов.

Построение ТСП

Пусть исследуется связь между признаками *A* и *B*. Для построения ТСП нужно указать для каждого объекта уровни признаков *A* и *B*. По исходным данным составляем *бинарную* (*индикаторную*) *матрицу* (табл. 1.6), в которой для каждого объекта (строки) и соответствующего ему уровня признака *B* (столбца) ставят 1, а на пересечениях с другими уровнями – 0. Каждая строка содержит только две 1. Если дополнить бинарную матрицу столбцом, то получим *состояние* конкретного объекта на *микроуровне* (табл. 1.6).

Таблица 1.6

Бинарная матрица

№ или код объекта	**Уровни признака *A***				**Уровни признака *B***				**Состояние: ячейка**
	1	**2**	**…**	***R***	**1**	**2**	**…**	***C***	***Ij***
1	0	1	…	0	1	0	…	0	21
2	0	0	…	1	0	1	…	0	r1
⋮	⋮	⋮	…	⋮	⋮	⋮	…	⋮	⋮
$\boldsymbol{N = n_{\bullet\bullet}}$	1	0	…	0	1	0	…	0	11

Подсчитывают количество n_{ij} одинаковых (состояний) ячеек ij, которыми заполняют ячейки ТСП. Если исследуются выборки объектов с более чем двух-трех признаков одновременно, то приходится строить ТСП, состоящую из нескольких частных ТСП. С увеличением количества таких таблиц, имеющегося объема выборки может оказаться недостаточно, чтобы заполнить все ячейки таблицы.

Пустые ячейки или недостаточное количество наблюдений в ячейках могут стать причиной серьезных проблем при интерпретации результатов или вызвать сомнения в их достоверности.

***Пример* 1.11.** Исследуется связь дыхательной функции по показателю легочной пробы: нормальная B – 0, ненормальная NE B – 1 и привычке к курению: курящие A – 0, некурящие NE A – 1) в популяции из 20 студентов. Легочная проба измеряется в литрах выдохнутого воздуха через 1 с после начала форсированного выдоха.

Таблица 1.7

Зависимость показателя легочной пробы от привычки к курению

	B	**NE *B***	**Всего**
A	1	6	7
NE *A*	10	3	13
Всего	11	9	20

Из табл. 1.7 видно, что среди 13 некурящих на 10 студентов с нормальной функцией есть всего 3 студента с ненормальной функцией легких, при этом в группе курящих на 1 здорового приходится 6 с нарушенной функцией легких. При визуальном анализе складывается впечатление, что курение ассоциируется с увеличением случаев нарушения функции легких. О том, как убедиться, что эта ассоциация существенна и достоверна, – в последующих главах.

Замечания. *Приведенный пример является не более чем демонстрацией построения ТСП, и она не удовлетворяет основному требованию: минимальное значение частоты в ячейке должно быть не менее* 5. ■

***Пример* 1.12.** В качестве классических примеров ТСП можно привести таблицы с результатами определения чувствительности и специфичности какого-либо нового метода диагностики (теста), для чего определяются соотношения истинно и ложно положительных и отрицательных результатов данного теста, т.е. сопоставляются результаты нового метода с золотым стандартом (табл. 1.8). ■

Таблица 1.8

Зависимость результатов применения нового диагностического метода от наличия заболевания

	Положительный результат	**Отрицательный Результат**	**Всего**
Болезнь есть	73	27	100
Болезни нет	15	85	100
Всего	88	112	200

***Пример* 1.13.** Традиционным примером ТСП является представление положительных и отрицательных результатов лечения в основной и контрольной группах – либо еще более частый вариант число исходов в группе лечения и группе контроля. Результаты клинического исследования «EUROPA» приведены в табл. 1.9.

Таблица 1.9

Первичные и некоторые вторичные исходы лечения периндоприлом по сравнению с плацебо у больных со стабильной ишемической болезнью сердца

	Плацебо (*n* = 6108) No. (%)	**Периндоприл (*n* = 6110) No. (%)**	**P**
Первичная конечная точка: сердечно-сосудистая смерть, острый инфаркт миокарда или остановка сердца	603 (9,9)	488 (8,0)	0,0003
Сердечно-сосудистая смерть	249 (4,1)	215 (3,5)	0,107
Несмертельный инфаркт миокарда	378 (6,2)	295 (4,8)	0,001
Общая смертность, несмертельный инфаркт миокарда, нестабильная стенокардия, остановка сердца	1043 (17,1)	904 (14,8)	0,0009
Общая смертность	**420 (6,9)**	**375 (6,1)**	**0,1**

The EUROPA Investigators. *Lancet*, 2003; **362** : 782−8

Практически все статистические ППП делают возможным построение ТСП обычно в разделе «*Contingency Tables»* с представлением данных в абсолютных значениях и в процентах.

При построении ТСП очень важно учитывать ту задачу, которую вы ставите, в частности, для определения, что формирует строки, а что столбцы, как подсчитывать проценты: по строкам, по столбцам или от общего числа. Кроме того, после анализа ТСП по ряду причин нередко приходится, если не менять шкалу измерения, то укрупнять выделяемые классы объектов. ■

***Пример* 1.14.** Данные обследования пациентов с ишемической болезнью сердца в сочетании с мерцательной аритмией и с нормальным ритмом приведены в ТСП (табл. 1.10). Регистрировалось наличие или отсутствие синдрома гипермобильности суставов. Имелся в виду «мягкий» вариант синдрома.

Таблица 1.10

Зависимость гипермобильности суставов от нарушений ритма

Нарушение ритма	**Гипермобильность суставов**	
	есть (%)	**нет (%)**
Нет нарушений ритма	15 (8,6%)	159 (91,4%)
Персистирующая мерцательная аритмия	7 (15,6%)	38 (84,4%)
Постоянная мерцательная аритмия	11 (31,4%)	24 (68,6%)

Из таблицы видно, что у больных с персистирующей мерцательной аритмией этот синдром выявлялся чаще чем у пациентов с синусовым ритмом, а в группе с постоянной формой аритмии – еще чаще. Однако, забегая вперед, скажем, что для статистической обработки здесь желательно использовать критерий Хи-квадрат, а для этого метода не желательны ячейки со значением менее 10. Вероятно можно, и с медицинской точки зрения в том числе, объединить пациентов с постоянной и персистирующей формой аритмии в одну подгруппу «мерцательная аритмия». В этом случае можно будет использовать указанный статистический метод и получить достоверные, хотя и несколько «обобщенные» выводы, в частности о взаимосвязи отмечаемых признаков.

Подобные укрупнения нередко делаются при анализе таких признаков как класс сердечной недостаточности (объединяют I и II классы, а также III с IV), степень тяжести пародонтита объединяют легкую и среднюю степень тяжести, и т.п. В данных случаях безусловно необходимо учитывать и медицинскую целесообразность таких изменений. ■

ТСП используются для решения следующих задач.

1. Выявления *статистически значимой связи* между признаками [31, ч.1, гл. 5]. В приведенном выше примере можно было исследовать связь между наличием мерцательной аритмии и синдрома гипермобильности суставов.

2. Построения *логарифически-линейных* (логлинейных) *моделей* [31,ч.1, п. 5.6]), которые позволяют решать задачи конструирования факторов, наилучшим образом соответствующих исходным данным.

3. *Анализа соответствий* (АС) [10, гл. 9], основной целью которого является переход от ТСП к *числовым* матрицам типа «объект-объект» при исследовании пространства объектов-точек, или «признак-признак» при исследовании признаков-точек в координатном пространстве малой размерности.

4. Анализа сложных эконометрических задач и задач оптимального размещения объектов в некотором пространстве при заданных ограничениях с помощью энтропийных и гравитационных моделей.

Эти задачи будут обсуждаться в последующих главах.

1.2.4. Матрицы близостей

Насколько похожи наши профили?

Ряд процедур статистической обработки – такие, как Q-факторный и кластерный анализы, многомерное шкалирование, анализ соответствий, – использует в качестве исходных данных квадратную матрицу Δ «объект-объект», а не привычную матрицу «признак-признак».

Матрицы «объект-объект» размерности $n \times n$, элементами которых являются меры сходства или различий δ_{ij} между объектами X_j и X_j, i, $j = 1, \ldots, n$ [10, 20]. Каждая строка и каждый столбец матрицы Δ соответствует одному объекту. Элементом δ_{ij} в i-й строке и j-м столбце матрицы Δ является мера сходства между объектами i и j. Рассмотрим некоторые меры сходства.

Корреляционная матрица «объект-объект»

Если признаки в матрице «объект-признак» измерены в количественной шкале, то меру сходства двух объектов в этой матрице определяют коэффициентом корреляции:

$$\hat{r}_{ik} = \frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2 \sum_{j=1}^{m}(x_{kj} - \bar{x}_k)^2\}^{1/2}}, \; i, \; k = 1, \ldots, n,$$

где x_{ij} – значение j-переменной для i-объекта, $\bar{x}_i$ – среднее для всех переменных i-объекта, n – число объектов, m – число признаков.

Коэффициенты корреляции образуют корреляционную матрицу типа «объект-объект» размерности $n \times n$. Если признаки измерены в номинальной или в порядковой шкале, то вместо r_{ik} используются соответственно *коэффициенты сопряженности* и *ранговой корреляции Спирмена* или *Кендалла* [6, 35].

Коэффициенты корреляционной матрицы «объект-объект» в отличие от корреляционной матрицы «признак-признак» не имеют ясного физического смысла, так как средние $\bar{x}_j$ определяются по всем признакам одного объекта.

Поэтому полученный коэффициент отражает форму, но не чувствителен к различиям в величине переменных, используемых при вычислении коэффициента.

Чувствительность к форме особенно важна в таких науках как социология, психология, антропология и во всех других, в которых важны профили. Профиль – это вектор значений признаков объекта, изображенный в виде ломаной линии. Два профиля могут иметь корреляцию, равную 1, но не будут идентичными (рис. 1.4).

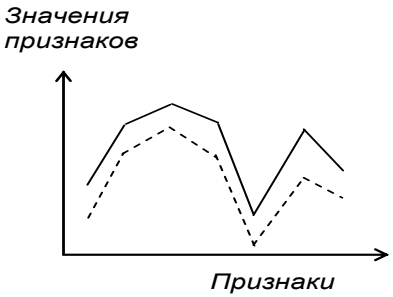


Рис. 1.4. Профили двух объектов

В данном контексте может изучаться, например, свойства гипотензивных препаратов в амбулаторных условиях. Измеряемыми характеристиками могут быть: число упаковок, проданных в аптеках города за месяц, частота назначений в амбулаторных картах, оценки врачами и пациентами эффективности и безопасности, цена, число необходимых приемов в течение дня и т.п.

Для наглядного представления профилей нередко используют специальные виды диаграмм: «лепестковые», «звездчатые» и т.п.

Недостатком корреляционной матрицы «объект-объект» как меры сходства является то, что элементы этой матрицы не удовлетворяют аксиоме треугольника. При этом мера сходства является полезной в приложениях кластерного анализа и многомерного метрического и неметрического шкалирования, где важна форма, а не сдвиг и масштаб. На оценке профиля основывается множество исследований в микробиологии, судебной медицине, морфологии.

1.3. Случайные временные ряды

Если важны различия биологического, социального и хронологического (астрономического) времен, то по каким часам жить и измерять данные?

В предыдущих разделах рассматривались данные, для которых порядок их расположения не принимался в расчет. Не имело значения, какое измерение сделано сначала, а какое потом, как, например, при исследовании артериального давления у пациентов, поступивших с мозговым инсультом, или определении частоты инфицированности *Helicobacter pylori* у больных диспепсией и т.п. При этом всем понятно,

что обследования проводились не в одно мгновение, а в течение иногда довольно длительного времени, однако условно считалось, что в данной ситуации фактор времени не важен, и его можно никак не учитывать. Иными словами, постулировалось, что случайные переменные x_t фиксируются в один и тот же момент времени t и что не существует явной связи между вектором x_t и вектором x_{t-1}, x_{t-2} … для предыдущих моментов времени. Если эти постулаты не удовлетворяются, то есть если влияние времени и/или последовательности измерения результатов (взятия проб крови, измерения пульса и т.п.) нельзя игнорировать, то мы имеем дело со *случайным временным рядом* (СВР), представляющим собой множество наблюдений, упорядоченных во времени. Если время непрерывно, то СВР называется случайным процессом [32, 33].

Во временном ряду каждому из моментов времени t_1, t_2, …, t_p соответствует фиксированное значение переменной X для фиксированного объекта U. Самым известным отображением этого является температурный лист, прикрепленный к спинке кровати больного. Еще одним наглядным примером СВР являются ритмограммы – последовательность, элементами которой являются промежутки времени между двумя соседними сердечными сокращениями обследуемого индивидуума. Эти промежутки определяются по R-зубцам кардиограммы и обозначаются RR интервалами. Другим примером СВР могут служить данные типа времени жизни [31, ч.1, 40]. В анализе таких данных особый интерес представляют группы объектов (индивидуумов), для каждого из которых определено точечное событие, условно называемое *отказом*.

Примерами *наработок до отказа* могут служить продолжительность жизни больных, период с начала лечения до рецидива заболевания или развития определенного осложнения при клинических испытаниях; время, затраченное индивидуумами для выполнения определенных задач в психологических экспериментах.

Результаты анализа по одной переменной t для разных объектов не всегда можно сравнивать непосредственно, потому, что процессы по-разному соотнесены со шкалой времени. Например, в социальной сфере и в биологии некоторые объекты изменяются быстрее, чем другие [59]. *Более того, в разный период болезни или выздоровления у одного и того же больного как патологические процессы, так*

и процессы выздоровления (*заживления, репарации*), *имеют совершенно различные темп и периодическую структуру*.

Таким образом, ***социальное или биологическое время – не то же самое, что и хронологическое*** (*астрономическое*) [59]. *Следовательно, первой задачей может явиться построение временных рядов для объектов так, чтобы ряды были сравнимыми, с рядами по социальному или биологическому, а не хронологическому времени.*

***Пример* 1.15.** Исследуя рост растений, Торнвейт[1] пришел к выводу о непригодности астрономического времени к измерению скорости роста. Биологические часы он обнаружил в горохе. В качестве единицы измерения времени он использовал промежуток между появлениями соседних узлов на стебле гороха. Эти промежутки имели различную длительность в астрономических единицах времени, но с их помощью удавалось лучше предсказывать урожай и управлять его сбором, чем при использовании астрономического времени. ■

В рассмотренных примерах ***время – это функциональное понятие,*** поэтому при совместном рассмотрении различных объектов нужно исследуемые переменные привести к общему временному базису, и только после этого они пригодны для статистического анализа. Эти преобразования нетривиальны и во многом зависят от типа шкалы измерения и решаемой задачи.

Анализу СВР посвящено огромное количество работ как теоретического, так и прикладного характера, и трудно перечислить все классы задач, решаемых с помощью СВР. Из всего многообразия медицинских проблем, решаемых в рамках теории СВР, в этой книге рассматриваются две задачи: первая из них связана с вариабельностью ритма сердца, а вторая − с упомянутой выше задачей типа *времени жизни*.[40].

1.3.1. Данные типа времени жизни

Как определить время жизни демографам и страховым агентам?

В анализе *данных типа времени жизни* (*дожития, выживания*[2]) особый интерес представляют группы индивидуумов, для каждого из которых определено точечное событие, называемое отказом.

Отказ наступает после некоторого времени, называемом временем наработки до отказа, для каждого индивидуума только один

[1]Thorntwaite C.W. Operations Research in Agriculture // Jornal of the Operations Research Society of America. – 1 (1953). – № 1. – P. 33–38.
[2]От англ. survival analysis.

раз. В медицинской практике отказом можно считать смерть больного, смерть больного от определенной причины (например, рак груди), первое повторное проявление болезни после стационарного лечения (повторный инфаркт) или случай нового заболевания, попадание в больницу, новые случаи диабета и т.п. Такие отказы, как правило, называют конечными точками, которые определяются при планировании исследования.

Начало отсчета времени должно быть точно установлено для каждого индивидуума, но оно, как правило, не совпадает с *календарным временем* (рис. 1.5).

Чаще всего началом (нулевым временем) условно считаются такие моменты, как поступление или выписка из стационара, начало лечения, день операции, первое обследование, время зачатия при беременности и т.п. Для каждого наработка до отказа (время до наступления конечной точки) измеряется от этого момента (рис. 1.5).

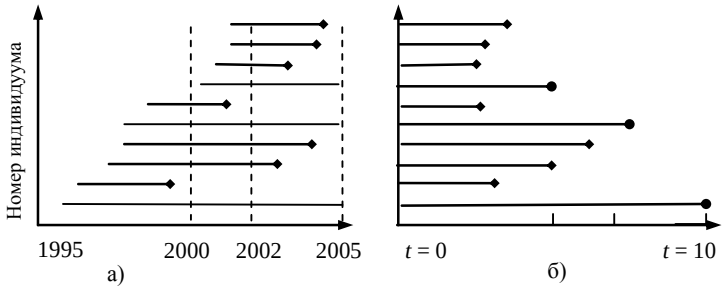


Рис. 1.5. Распределение значений наработок распределенными по временам моментами поступления до 2005 г.:

◆ – отказ; ● – цензурирование; а) календарное время; б) время от начала поступления больного на исследование

Виды выживаемости[1]

- *Бессобытийная выживаемость* (event free survival). Начало мониторинга выбирается и обосновывается исследователем (например, момент трансплантации костного мозга, ТКМ.

[1]Приводится по работе Гемджан Э. Статистическая обработка данных в медицине. Доказательная медицина. Evidence Based Medicine (с изменениями). E-mail: edstat@mail.ru. http://www.ed.blood.ru.com.

Событие: все события, значимые для целей исследования. Нужно указать, какие конкретные события рассматриваются.

В каждом конкретном исследовании значимые события определяются в соответствии со стандартами. Если нет стандарта, то исследователь составляет список и обосновывает свой выбор.

- *Общая выживаемость* (overall observed survival, os).

Начало мониторинга: обычно, это начало лечения (например, момент ТКМ).

Событие: смерть от любой причины.

- *Безрецидивная выживаемость* (reccurence relapce free survival, rfs).

Начало мониторинга: достижение ремиссии после окончания лечения (в гематологии это *может быть* момент ТКМ).

Событие: рецидив.

- *Общая выживаемость скорректированная*.

Начало мониторинга: начало лечения.

Событие: смерть от основного заболевания).

- *Progression free survival* (*pfs*).

Начало мониторинга: конец лечения (причем ремиссия *не обязательно* достигнута).

Событие: обострение заболевания.

- *Время наработки до отказа от начала лечения* (*time to treatment failure, ttf*).

Начало мониторинга: начало лечения.

Событие: прогрессирование заболевания или смерть *в процессе лечения*.

1.3.2. Цензурирование

Что делать с потерянными пациентами?

Особым источником затруднений при анализе данных типа времени жизни является то, что некоторые объекты могут не наблюдаться в течение полного времени до отказа. Основная трудность оценивания состоит в возможности «исчезновения» некоторых больных из поля зрения исследователя до окончания периода наблюдения, некоторые из них будут продолжать жить, например, уехав в другой город.

Такое неполное наблюдение наработки до отказа называется ***цензурированием***, а получаемые в результате выборки – ***цензурированными выборками*** (ЦВ). Например, при исследовании некоторой популяции больных в течение определенного интервала времени часть пациентов по самым различным причинам «исчезает из-под наблюдения» или популяция пополняется другими индивидуумами.

Причинами цензурирования могут быть отказ больного от дальнейших диспансерных наблюдений, смерть, смена места проживания, полное выздоровление, желание «новых» больных пополнить группу исследуемых и т.д. Возможны случаи, когда отказ наступает от причин, которые желательно исключить из рассмотрения.

***Несколько примеров цензурирования*:**

- Оценка показателей ведения для группы больных с одним диагнозом (колоректальный рак), а отказ наступает от другого заболевания (смерть в автокатастрофе).
- Выбранный интервал наблюдения слишком мал для регистрации продолжительности жизни больных в исследуемой группе: 6 месяцев после радикальной операции у онкологических больных.
- Разное по сути время начала наблюдений за конкретными больными исследуемой группы; наблюдая с момента поступления в клинику больных с инсультом, наступившим в день поступления и существенно более (5–7 дней), симптомы отека головного мозга будут зафиксированы раньше у тех, что поступили поздно. Очевидно, что этих пациентов нельзя включать в одну группу, или, на худой конец, за время отсчета следует принять день возникновения инсульта, а не день поступления. ■

Изменение объема выборки в случайные моменты времени не позволяет обработать цензурированные выборки классическими статистическими методами. Этим обусловлена необходимость применения специальных методов анализа цензурированных выборок.

Отметим, что цензурированные выборки образуются из цензурированных данных, причем эти понятия нельзя отождествлять, так как цензурированные данные могут быть неоднородными и содержать несколько выборок, то есть цензурированные данные являются более общим понятием по сравнению с цензурированными выборками.

К сожалению, статистические оценки показателей выживаемости требуемой точности могут быть получены только к концу жизни объектов, то есть через много лет, когда их практическое использование уже не имеет ценности. Именно поэтому важное значение приобретает

программа сокращения длительности наблюдений при оценке показателей выживаемости. Сокращение продолжительности наблюдений для всех или части объектов приводит к ЦВ, к которым классические методы теории надежности, использующие полные выборки, как правило, не применимы.

Двумя составляющими элементами ЦВ являются значения наработки до отказа и наработки до цензурирования. Например, в клиническом испытании, которое длилось 10 лет, деменция развилась у 15% пожилых людей, а 23% умерли. Мы предполагаем, что в дальнейшем еще у части из них разовьется деменция, только не знаем, у кого именно, также понятно, что когда-либо все они умрут, только неизвестно кто и когда. Другими словами, *цензурирование – это процесс возникновения неопределенности момента отказа объекта*, причем интервал неопределенности известен исследователю.

Интервал неопределенности – интервал наработки, внутри которого произошел либо произойдет отказ объекта, причем точное значение наработки до отказа неизвестно. Если интервал неограничен справа (рис. 1.7 б), то имеем цензурирование слева, если ограничен справа, то цензурирование слева. Если интервал неопределенности момента отказа ограничен слева и справа, то говорят о цензурировании интервалом. Если в ЦВ отсутствуют условные и неполные наработки, то такая выборка превращается в полную, состоящую только из полных наработок.

1.3.3. Модели формирования данных ЦВ

Какие бывают виды потерь пациентов

Для обоснованного выбора того или иного метода анализа ЦВ из числа существующих к настоящему времени необходимо четко представлять статистическую структуру анализируемых данных, поэтому предварительный анализ процесса формирования наблюдаемых данных является, как правило, важным этапом решения любой прикладной статистической задачи.

При анализе ЦВ роль этого этапа обычно еще более важна, чем при анализе полных выборок, так как степень неопределенности информации, заключенной в выборке, при прочих равных условиях,

как правило, выше при наличии цензурирования. В общем виде динамика процесса формирования данных ЦВ может быть описана схемой, представленной на рис. 1.6[1] .

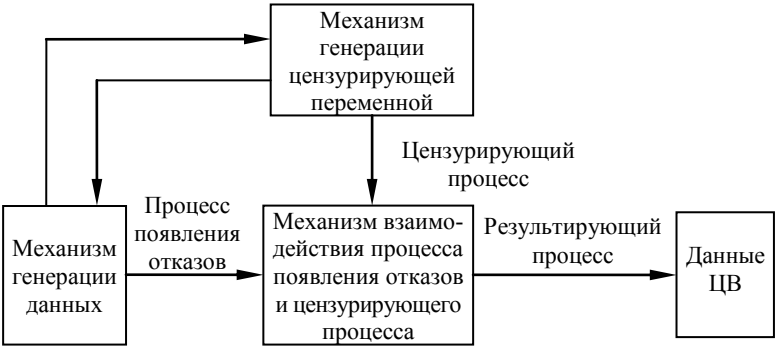


Рис. 1.6. Формирование данных ЦВ

При рассмотрении различных моделей будем считать, что: характеристика (например, размер опухоли) начального состояния объекта, участвующего в испытаниях, представляет собой неслучайную величину; отказ объекта может быть единственного типа (например, гибель опухоленосителя); априорная информация о множестве переходных состояний объекта на протяжении интервала наблюдения не включается в рассмотрение. При таких упрощениях полное наблюдение является реализацией изучаемой случайной величины, представляющей собой интервал времени от момента фиксации начального состояния до момента отказа.

Выборочные данные клинических испытаний могут содержать неполные наблюдения, возникающие в результате завершения испытаний, либо по той причине, что наблюдение за объектом началось позднее момента наступления начального испытания. Такие неполные наблюдения относятся к группе *A*.

Неполные наблюдения могут возникать в результате случайной потери наблюдений, либо по причине отказа, либо вследствие несвоевременного прекращения наблюдения за данным конкретным

[1] Анализ надежности технических систем по цензурированным выборкам / В.М. Скрипник, А.Е. Назин, Ю.Г. Приходько, Ю.Н. Благовещенский. М.: Радио и связь, 1988. – 184 с.

объектом (например, в случае самовольного выхода пациента из испытаний или нецелесообразности дальнейших наблюдений за ним). Неполные наблюдения такого типа отнесем к группе *B*.

1.3.4. Классификация ЦВ

При воздействии цензурирующего процесса исследователь может наблюдать реализации случайных наработок, цензурированных либо в одной, либо во многих точках, поэтому все многообразие структур представления данных ЦВ можно разбить на два больших класса: однократно цензурированные выборки (ОЦВ) и многократно цензурированные выборки (МЦВ).

Вначале рассмотрим ОЦВ. Для случая, когда интервал неопределенности неограничен справа у ОЦВ, известно только то, что часть наработок до отказа в этом случае находится в интервале $[0,T]$ (условные реализации), а моменты других отказов известны (полные наработки) и находятся в интервале $[T, \infty]$, имеет место ОЦВ слева (рис. 1.7 а).Если интервал неопределенности неограничен слева, то имеем случайное цензурирование справа (рис. 1.7 б).

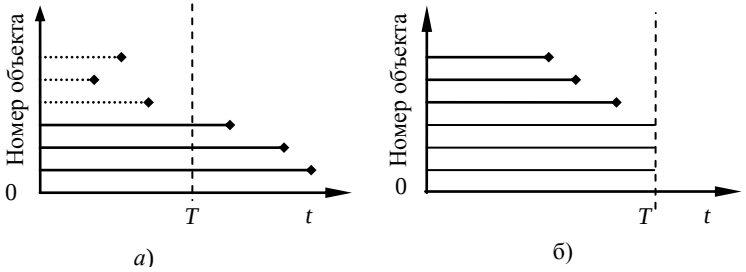


Рис. 1.7. Распределение наработок при ОЦВ: а) слева, б) справа

На практике при оценке надежности изделий имеют место случаи, когда о части отказов известно только то, что их моменты находятся в каком-то временном интервале $[T_1, T_2]$ (то есть имеют место условные наработки при цензурировании интервалом), а моменты других отказов известны (полные наработки) и находятся правее данного интервала или левее его.

Если структура данных ЦВ устроена так, что значения неполных наработок или условных наработок не равны между собой, то такие выборки называются МЦВ. МЦВ могут быть как одного вида

(слева, справа, интервалом), так и комбинированными при различных комбинациях видов цензурирования и расположения полных, неполных и условных наработок. Распределение реализаций случайных наработок в случае МЦВ приведено на рис. 1.8.

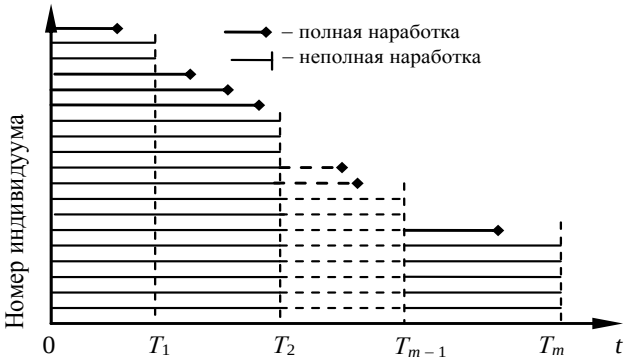


Рис. 1.8. Распределение наработок при многократном цензурировании

Пусть в случае МЦВ, изображенном на рис. 1.8, число интервалов равно $m = 4$, число объектов в начале T_i-го интервала исследования обозначим n_i, число отказов – d_i и число цензурированных объектов – w_i. Тогда данные МЦВ можно представить таблицей (табл. 1.13), которую часто называют «таблицей времени жизни (наработок на отказ)», они являются исходными данными для непараметрического оценивания функции выживаемости по МЦВ [31, 40, 68].

Данные МЦВ Таблица 1.13

Номер Интервала	**Число объектов n_i**	**Число отказов d_i**	**Число цензурированных w_i**
0	19	0	0
1	19	1	2
2	16	3	3
3	13	2	2
4	9	1	5

Заключение

Согласно теории классической математической статистики изложение учебного пособия следовало бы начинать с описательной статистики одномерных данных числовой природы. Однако реальные

данные далеки от идеальных наблюдений, они отличаются многообразием шкал, содержат сингулярности типа аномальных выбросов, трендов, помех, неигнорируемых пропусков и т.д.

Поэтому сначала рассматривается широкий спектр видов представления исходных данных, которые характеризуют свойства целого и неделимого организма. Основной проблемой обработки таких данных является то, что они принадлежат к разным пространствам и без соответствующих преобразований к ним нельзя применить классические методы многомерного статистического анализа.

Для совместной обработки данных числовой и нечисловой природы приходится все рассматриваемые признаки «рассыпать» на отдельные категории (альтернативы) и затем снова «склеивать», используя адекватные методы (логлинейный анализ многомерных ТСП, анализ соответствий и т.д.).

Методы описательной статистики для идеальных данных изложены во 2-й главе, а особенности статистического анализа реальных наблюдений – в главе 3.

Упражнения и задачи

1.Докажите, что абсолютная шкала является самой сильной.

2.К какой шкале относится оценка знаний в баллах?

3.В каких шкалах измерены: номера автомашин, телефонов, студенческих билетов; время, если единицей измерения является год?

4.Докажите, что результаты качественных измерений не являются элементами линейного (векторного) пространства.

5.Не используя примеров, приведенных в пособии, дайте по одному физически реальному примеру: а) одномерной шкалы, наименований, шкалы частичного порядка, шкалы интервалов и шкалы отношений; б) двумерной шкалы, один компонент которой есть шкала частичного порядка, а другой – шкала отношений.

6.Какие виды шкал можно применять для измерения целей?

7.Приведите полный анализ применяемых шкал в отчетном сообщении: «Врач-кардиолог 1-й категории Иванов в течение рабочего времени принял 5 пациентов, из них 1 пациента – студентку 5-го курса педиатрического факультета с признаками воспаления легких направил на снимок легких, 3 направил к гастрологу с подозрением на язвенную болезнь желудка, у 2 объяснил «боли в области сердца» синдромом межре-

берной невралгии, и только у одного – ветерана войны преклонного возраста – установил признаки предсердной мерцательной аритмии».

8.Приведите доказательство отличия номинальных и квазиквантитативных переменных.

9. Исследованы 42 ребенка, из них 20 вскормлены грудью, а остальные – «искусственники». Из вскормленных грудью 16 имели неправильный прикус зубов. У «искусственников» только у 1 ребенка были правильные зубы.

Постройте ТСП зависимости прикуса зубов от кормления грудью.

10.Фирма хотела бы знать, можно ли отправлять заказчикам индийский и грузинский чай в одной упаковке. Не ухудшится ли качество и аромат индийского чая? Для проведения исследования изготовили 400 картонных коробок и в 250 из них положили оба вида чая, а в 150 положили только индийский. Через месяц коробки вскрыли и определили изменение аромата индийского чая. Оказалось, аромат не изменился в 190 коробках, из них в 72 находились оба чая.

Постройте ТСП зависимости между способом упаковки и изменением аромата индийского чая.

11.Опросили 900 студентов с целью выяснения: имеется ли связь между полом студентов и специализацией. Из 350 студентов отделения искусствоведения – 185 студенток, естественные науки предпочитают 168 студентов, в направлении социально-экономических наук всего 220 студентов и студенток, в отделении музыки обучаются 70, из них 32 студента и 38 студенток. Всего студенток – 420.

Постройте ТСП связи между полом студентов и выбранным направлением обучения.